UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELOS DE DECISÃO E SAÚDE

DIFICULDADE NO JULGAMENTO PERCEPTIVO-AUDITIVO DA QUALIDADE VOCAL: UMA ABORDAGEM BASEADA EM APRENDIZADO DE MÁQUINA

MAXSUEL ALVES AVELINO DE PAIVA

JOÃO PESSOA – PB

2025

MAXSUEL ALVES AVELINO DE PAIVA

DIFICULDADE NO JULGAMENTO PERCEPTIVO-AUDITIVO DA QUALIDADE VOCAL: UMA ABORDAGEM BASEADA EM *APRENDIZADO DE MÁQUINA*

Tese apresentada ao Programa de Pós-Graduação em Modelos de Decisão e Saúde – Nível Doutorado, do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba.

Linha de Pesquisa: Modelos em Saúde.

Orientadores:
Prof. Dr. Leonardo Wanderley Lopes
Prof. Dr. Luiz de Medeiros de Araújo Lima Filho
Coorientadora:
Profa. Dra. Bruna Gregory Palm

JOÃO PESSOA – PB
2025

Catalogação na publicação
Seção de Catalogação e Classificação

P149d Paiva, Maxsuel Alves Avelino de.
Dificuldade no julgamento perceptivo-auditivo da qualidade vocal : uma abordagem baseada em aprendizado de máquina / Maxsuel Alves Avelino de Paiva. - João Pessoa, 2025.
87 f. : il.

Orientação: Leonardo Wanderley Lopes, Luiz de Medeiros de Araújo Lima Filho.
Coorientação: Bruna Gregory Palm.
Tese (Doutorado) - UFPB/CCEN.

1. Percepção auditiva. 2. Acústica. 3. Voz. 4. Aprendizado de máquina. 5. Fonoaudiologia. I. Lopes, Leonardo Wanderley. II. Lima Filho, Luiz de Medeiros de Araújo. III. Palm, Bruna Gregory. IV. Título.

UFPB/BC CDU 612.858.7(043)

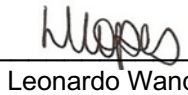
Elaborado por RUSTON SAMMEVILLE ALEXANDRE MARQUES DA SILVA - CRB-15/0386

MAXSUEL ALVES AVELINO DE PAIVA

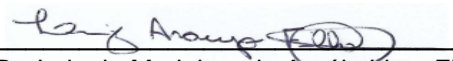
DIFICULDADE NO JULGAMENTO PERCEPTIVO-AUDITIVO DA QUALIDADE VOCAL: UMA ABORDAGEM BASEADA EM *APRENDIZADO DE MÁQUINA*

João Pessoa, 21 de julho de 2025.

BANCA EXAMINADORA



Prof. Dr. Leonardo Wanderley Lopes
Orientador (UFPB)



Profa. Dr. Luiz de Medeiros de Araújo Lima Filho
Orientadora (UFPB)

Profa. Dra. Bruna Gregory Palm
Coorientadora (Blekinge Institute of Technology)

Profa. Dra. Anna Alice Figueirêdo de Almeida
Membro Interna (UFPB)

Prof. Dr. Marcelo Rodrigo Portela Ferreira
Membro Interno (UFPB)

Profa. Dra. Ana Carolina Constantini
Membro Externa (UNICAMP)

Profa. Dr. Samuel Ribeiro de Abreu
Membro Externo (UFR)

“É justo que muito custe o que muito vale.”

Santa Teresa d'Ávila

Dedico este trabalho à minha mãe,
pelo amor incondicional, pela força e pelo apoio em cada etapa da minha caminhada.
Sua coragem, generosidade e fé foram o alicerce que me sustentou até aqui.
Tudo o que conquistei carrega um pouco de você.

AGRADECIMENTOS

Chegar até aqui não foi uma jornada solitária. Cada passo foi acompanhado, incentivado e sustentado por pessoas que fizeram toda a diferença no meu caminho.

À minha família, meu porto seguro. À minha mãe, Edvalda, e ao meu pai, Marcelo, por todo o amor e apoio incondicional. Um agradecimento especial ao meu irmão, Marcel, cuja presença constante, incentivo e força nos bastidores foram fundamentais para que eu me mantivesse firme nos momentos mais desafiadores. Vocês me ensinaram o verdadeiro significado de resiliência e me mostraram, todos os dias, que vale a pena seguir em frente.

À Dona Zélia, minha avó amada, que sempre se emocionava com minhas conquistas, mesmo na sua simplicidade. Sua ausência é sentida, mas sua memória me fortalece.

À Karina, minha amiga de tantas jornadas, obrigado pela escuta, pela lealdade e pela cumplicidade que sempre me abraçou com suas palavras nos momentos certos.

Ao Alex, meu namorado. Obrigado por ser abrigo nos dias difíceis e companhia nos dias leves. Seu carinho, paciência e presença foram fundamentais para que eu pudesse seguir firme até aqui. Ter você ao meu lado tornou essa caminhada mais feliz.

Aos meus amigos, Ruama, Taynara e Tiago, por se fazerem presentes e por todo apoio na minha jornada acadêmica.

Aos amigos do Programa de Pós-graduação em Modelos de Decisão e Saúde, por cada troca, conversa e apoio compartilhado ao longo dessa trajetória.

Aos meus alunos da graduação em Fonoaudiologia da UNINASSAU e da FASP, que, com sua curiosidade, energia e presença, renovam meu entusiasmo pela docência e me lembram diariamente do porquê escolhi ensinar.

Ao professor Leonardo Wanderley, meu orientador desde os primeiros passos na graduação. Obrigado por acreditar em mim, por orientar com sabedoria e humanidade, e por ser referência de profissional e ser humano. Tê-lo ao meu lado em todas essas fases é um privilégio raro e precioso.

Ao professor Luiz, que acolheu este projeto com generosidade, atenção e imensa disponibilidade. Obrigado por cada orientação oferecida com tanto cuidado e por cada sugestão precisa que ajudaram a tornar este trabalho mais consistente e significativo. Seu compromisso e dedicação foram essenciais ao longo de todo o processo.

Aos professores Anna Alice, Ana Carolina, Bruna Gregory e Samuel Ribeiro, meu sincero agradecimento pela disponibilidade e pelas valiosas contribuições que enriqueceram essa pesquisa. Ter suas leituras foi uma grande honra.

À CAPES, pelo suporte essencial à viabilidade deste trabalho.

RESUMO

O Julgamento Perceptivo-auditiva da voz (JPA) é amplamente utilizada na prática clínica e na formação em Fonoaudiologia, porém apresenta elevada variabilidade entre juízes, o que compromete sua confiabilidade. Tradicionalmente, os estudos focam em fatores ligados ao juiz ou à tarefa de fala, negligenciando a possibilidade de que a própria voz apresente graus distintos de dificuldade para julgamento. O objetivo deste trabalho é desenvolver um sistema de classificação de padrões para predizer a dificuldade do JPA da voz com base em modelos de AM treinados com dados de concordância e medidas acústicas. As vozes analisadas neste estudo pertencem ao banco de dados do Laboratório Integrado de Estudos da Voz (LIEV), da instituição de origem, sendo frequentemente utilizadas em pesquisas desenvolvidas pelo grupo. Inicialmente, técnicas de agrupamentos foram aplicadas para agrupar 295 vozes segundo o grau de dificuldade do JPA, considerando o coeficiente de variação do JPA e concordância fuzzy entre cinco juízes especialistas em voz. O algoritmo K-means foi selecionado por apresentar o melhor índice interno de validade (Calinski-Harabasz) e melhor separação visual entre os grupos. As vozes foram classificadas em três níveis de dificuldade: fácil, médio e difícil. Posteriormente, quatorze modelos de aprendizado de classificação supervisionada foram treinados para classificar, com base em medidas acústicas, o grau de dificuldade de julgamento das vozes. Os modelos de Regressão Multinomial, Multilayer Perceptron e Random Forest apresentaram os melhores desempenhos para diferentes parâmetros vocais. Os achados demonstram o potencial da combinação de métodos não supervisionados e supervisionados para classificar a dificuldade do JPA, contribuindo para a criação de sistemas de treinamento mais objetivos e fundamentados em evidência. A proposta fornece subsídios para o desenvolvimento de simuladores de treinamento com progressão baseada na dificuldade perceptiva das vozes, com impacto direto na formação e na prática clínica em Fonoaudiologia.

Descritores: percepção auditiva; aprendizado de máquina; acústica; treinamento auditivo; voz; Fonoaudiologia.

ABSTRACT

Auditory-Perceptual Judgment (APJ) of voice is widely used in clinical practice and speech-language pathology training, but it presents high variability among judges, which compromises its reliability. Traditionally, studies focus on factors related to the judge or the speech task, neglecting the possibility that the voice itself presents different degrees of difficulty for judgment. The objective of this work is to develop a pattern classification system to predict the difficulty of the APJ of voice based on ML models trained with concordance data and acoustic measurements. The voices analyzed in this study belong to the database of the Integrated Laboratory of Voice Studies (LIEV) at the home institution and are frequently used in research developed by the group. Initially, clustering techniques were applied to group 295 voices according to the degree of APJ difficulty, considering the APJ coefficient of variation and fuzzy concordance among five expert voice judges. The K-means algorithm was selected because it presented the best internal validity index (Calinski-Harabasz) and the best visual separation between groups. The voices were classified into three difficulty levels: easy, medium, and difficult. Subsequently, fourteen supervised classification learning models were trained to classify, based on acoustic measurements, the degree of voice judgment difficulty. The Multinomial Regression, Multilayer Perceptron, and Random Forest models performed best for different vocal parameters. The findings demonstrate the potential of combining unsupervised and supervised methods to classify JPA difficulty, contributing to the creation of more objective and evidence-based training systems. The proposal provides support for the development of training simulators with progression based on the perceptual difficulty of voices, with a direct impact on training and clinical practice in Speech-Language Pathology.

Keywords: auditory perception; machine learning; acoustics; auditory training; voice; speech-language pathology

LISTA DE QUADROS, TABELAS, FIGURAS e GRÁFICOS

Quadro 1 – Literatura sobre variáveis relacionadas ao TPA 27

Quadro 2 – Aplicação de AM na área da voz.....37

Quadro 3 – Definição de variáveis.....49

Quadro 4 – Classificação para os valores do Coeficiente Kappa (LANDIS et al., 1977), acurácia, sensibilidade e especificidade (HOSMER E LEMESHOW, 2000) 53

Quadro 5 – Índice interno de avaliação dos modelos de agrupamento para o JPA.....55

Quadro 6 – Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GG.....58

Quadro 7 – Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GR.....59

Quadro 8 – Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GS.....60

Quadro 9 – Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GT.....61

Quadro 10 – Matriz de confusão do modelo de Regressão Multinomial para os níveis de dificuldade do JPA do GG.....66

Quadro 11 – Matriz de confusão do modelo de Regressão Multinomial para os níveis de dificuldade do JPA do GR.....68

Quadro 12 – Matriz de confusão do modelo MLP para os níveis de dificuldade do JPA do GS.....70

Quadro 13 – Matriz de confusão do modelo Random Forest para os níveis de dificuldade do JPA do GT.....72

Tabela 1 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GG no conjunto de treinamento..... 61

Tabela 2 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GR no conjunto de treinamento.....62

Tabela 3 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GS no conjunto de treinamento.....63

Tabela 4 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GT no conjunto de treinamento.....63

Tabela 5 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GG no conjunto de teste.....64

Tabela 6 – Métricas de desempenho do modelo de Regressão Multinomial baseado na matriz de confusão para os níveis de dificuldade do JPA do GG..66

Tabela 7 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GR no conjunto de teste..... 67

Tabela 8 – Métricas de desempenho do modelo de Regressão Multinomial baseado na matriz de confusão para os níveis de dificuldade do JPA do GR..68

Tabela 9 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GS no conjunto de teste..... 69

Tabela 10 – Métricas de desempenho do modelo de Regressão Multinomial baseado na matriz de confusão para os níveis de dificuldade do JPA do GS..70

Tabela 11 – Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GT no conjunto de teste.....71

Tabela 12 – Métricas de desempenho do modelo de Regressão Multinomial baseado na matriz de confusão para os níveis de dificuldade do JPA do GT...72

Figura 1 – Oscilograma da vogal /a/, silêncio e contagem.....41

Figura 2 – Aplicação da lógica fuzzy para obtenção do valor final da EAV fuzzy a partir dos julgamentos de cinco especialistas.....46

Gráfico 1 – Método do cotovelo para identificação do número ideal de clusters do grau de dificuldade do JPA segundo o modelo K-means.....56

Gráfico 2 – Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GG.....57

Gráfico 3 – Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GR.....58

Gráfico 4 – Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GS.....59

Gráfico 5 – Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GT.....60

LISTA DE ABREVIATURAS E SIGLAS

AM – Aprendizado de Máquina
JPA – Julgamento Perceptivo-Auditiva
TPA – Treinamento Perceptivo-Auditivo
ST – Simulador de Treinamento
EDV – Escala de Desvio Vocal
GRBAS – Escala numérica para avaliação da voz
G – *Grade*
R – *Roughness*
B – *Breathness*
A – *Asteny*
S – *Strain*
I – *Instability*
GG – Grau geral de desvio vocal
GR – Grau de Rugosidade
GS – Grau de Soprosidade
GT – Grau de Tensão
CAPE-V – Protocolo de avaliação vocal
ASHA – *American Speech-Language na Hearing Association*
EAV – Escala Analógica-Visual
CoDAS – *Communication Disorders, Audiology and Swallowing*
LIEV – Laboratório Integrado de Estudo das Voz
fo – Frequência fundamental
fo_DP – Desvio padrão da frequência fundamental
fo_q1 – Primeiro quartil da frequência fundamental
fo_mediana – Mediana da frequência fundamental
fo_q3 – Terceiro quartil da frequência fundamental
fo_CV – Coeficiente de variação da frequência fundamental
PSD – Desvio padrão do período (do inglês *Period Standard Deviation*)
LNPSD – Logaritmo natural do desvio padrão do período
RAP – *Relative Average Perturbation*

PPQ5 – *Pitch Period Perturbation Quotient (5 ciclos)*

DDP – *Difference of Differences of Periods*

APQ3 – *Amplitude Perturbation Quotient (3 ciclos)*

APQ5 – *Amplitude Perturbation Quotient (5 ciclos)*

APQ11 – *Amplitude Perturbation Quotient (11 ciclos)*

AVI – *Amplitude Variation Index*

SNL1 a SNL6 – Bandas de energia espectral (níveis de ruído espectral em diferentes faixas)

H1-H2 – Diferença entre os harmônicos 1 e 2

H1-A1 – Diferença entre o primeiro harmônico e a primeira região formântica

H1-A3 – Diferença entre o primeiro harmônico e a terceira região formântica

HNR – *Harmonics-to-Noise Ratio* (Relação harmônico-ruído)

HNR_DP – Desvio padrão do HNR

HNRD – Desvio absoluto do HNR

Hfno – Frequência harmônica não periódica

PA – *Prediction of Aperiodicity* (Previsão de aperiodicidade)

SFR – *Strength of Fundamental Frequency* (Força da frequência fundamental)

GNE – *Glottal-to-Noise Excitation Ratio* (Razão entre excitação glotal e ruído)

GNE1 – Excitação do ruído glotal na banda de 1.000 Hz

GNE2 – Excitação do ruído glotal na banda de 2.000 Hz

GNE3 – Excitação do ruído glotal na banda de 3.000 Hz

CPP – *Cepstral Peak Prominence*

CPPS – *Cepstral Peak Prominence-Smoothed* (Proeminência do Pico Cepstral Suavizada)

CPPS_fo – Frequência fundamental associada ao CPPS

CPPS_média – Média do CPPS

CPPS_DP – Desvio padrão do CPPS

CPPS_q1 – Primeiro quartil do CPPS

CPPS_mediana – Mediana do CPPS

CPPS_q3 – Terceiro quartil do CPPS

CPPS_DAM – Desvio absoluto da mediana do CPPS

SUMÁRIO

Resumo

1 Introdução 16

1.1 Motivação e Relevância 20
1.2 Objetivos 22
1.3 Estrutura do Trabalho 23

2 Fundamentação Teórica 24

2.1 Julgamento e Treinamento Perceptivo-auditivo 24

2.2 Medidas acústicas 30

2.3 Modelos de Aprendizado de Máquina em voz 33

3 Metodologia 40

3.1 Delineamento, período de referência e local do estudo.... 40

3.2 Amostra do estudo e procedimentos de coleta 40

4 Resultados 55

5 Discussão 73

6 Limitações e perspectivas futuras 79

7 Conclusão 81

8 Publicações 82

Referências 83

1 INTRODUÇÃO

A avaliação vocal é multidimensional porque envolve várias medidas, como as acústicas, aerodinâmicas, exame visual da laringe, autoavaliação com protocolos validados e o julgamento perceptivo-auditivo da voz (JPA) (Dejonckere, 2001; Roy, 2013; Van Stan; Mehta; Hillman, 2017; Patel *et al.,*2018). A análise acústica oferece informações quantitativas e qualitativas extraídas do sinal vocal; a aerodinâmica fornece dados sobre o controle do fluxo aéreo durante a fonação; o exame visual da laringe avalia as estruturas e a funcionalidade da laringe; a autoavaliação revela o impacto do problema vocal na qualidade de vida, a desvantagem percebida pelo indivíduo e a frequência/severidade dos sintomas; e o JPA identifica a presença, tipo e intensidade do desvio vocal, além de analisar outros parâmetros como *pitch*, *loudness*, ressonância e articulação (Dejonckere, 2001; Roy, 2013).

O JPA é o principal método de avaliação vocal utilizado pelos fonoaudiólogos, devido ao seu baixo custo, rapidez, conforto para o paciente e poucos requisitos técnicos. Além de ser amplamente usado na prática clínica, o JPA é muito empregado em pesquisas na área de voz (Eadie; Baylor, 2006; Oates, 2009; Roy, 2013; Van Stan; Mehta; Hillman, 2017; Patel *et al.,*2018).

Por ser baseado na percepção do clínico, o JPA é considerado subjetivo, pois pode ser influenciado por diversos fatores como o treinamento, experiência do avaliador, habilidades perceptivas, preferências pessoais, tarefa de fala, habilidades de processamento auditivo, entre outros fatores (Oates, 2009; Paiva *et al.,* 2019). Mesmo com essa subjetividade, o JPA é tradicional na clínica fonoaudiológica e é considerado o padrão-ouro para a avaliação da qualidade vocal, pois é a melhor forma de identificar a presença de desvios vocais (Eadie; Baylor, 2006). A subjetividade do JPA não é motivo suficiente para rejeitar seu uso, pois, sendo a voz um fenômeno fundamentalmente perceptivo em resposta a um estímulo acústico, o JPA é o melhor método para avaliar esse fenômeno (Oates, 2009; Shrivastav, 2005).

Um dos principais fatores que afetam a confiabilidade do JPA é o fato de ser baseado na experiência do juiz e, consequentemente, nas suas habilidades de percepção auditiva, preferências pessoais, tarefa de fala, treinamento prévio, entre outros (Oates, 2009; Eadie *et al*.,2010). Estar ciente dessas

fragilidades à validade e confiabilidade do JPA é um passo importante para criar modelos de treinamento perceptivo-auditivo (TPA) mais eficientes, que contribuam para o uso do JPA de maneira mais assertiva e refinada no contexto clínico e na formação de novos fonoaudiólogos (Kent, 1996; Chartrand; Belin, 2006).

A literatura aponta que as principais variáveis associadas à confiabilidade do JPA estão relacionadas ao perfil do juiz e ao tipo de tarefa de fala julgada (Oates, 2009; Eadie et al., 2010). No entanto, há pouca discussão sobre o papel da qualidade vocal e da complexidade dos estímulos utilizados no TPA e no JPA. Considerar que determinadas características da voz podem influenciar a facilidade ou dificuldade de identificação dos parâmetros desviantes pode contribuir para o desenvolvimento de modelos de TPA que incorporem essas características como variáveis relevantes para a confiabilidade da JPA.

O estudo de Bispo *et al*., (2022) observou que a complexidade dos estímulos sonoros interfere na acurácia do JPA. Quanto mais simples o estímulo, maior a acurácia na identificação do desvio vocal predominante. Além disso, o tipo de desvio vocal também interfere na acurácia, sendo que as vozes soprosas possuem maior taxa de acertos no JPA.

Diversas estratégias têm sido empregadas no TPA com o objetivo de minimizar a subjetividade desse processo. Entre as mais recorrentes, destacam-se: o uso de estímulos âncora — escuta de vozes de referência antes do julgamento; o fornecimento de feedback imediato — informação ao juiz sobre a correspondência entre sua avaliação e a de um juiz de referência; e a adoção de escalas padronizadas — que possibilitam o registro sistemático e a comparação entre avaliações (Zraick et al., 2011; Paz, 2023; Paiva, 2023).

Em revisão, Paz et al. (2023) identificaram que as estratégias mais utilizadas nos TPA são o uso de âncoras auditivas, com vozes naturais ou sintetizadas, e o feedback. No entanto, não foram encontrados estudos que abordem a classificação da dificuldade das vozes nem a aplicação de níveis graduais de dificuldade como variável estratégica nos TPA.

Embora essas abordagens sejam amplamente utilizadas na formação de fonoaudiólogos, elas apresentam algumas limitações práticas, como o tempo demandado para a tabulação das respostas, a necessidade de encontros

presenciais com disponibilidade dos juízes, além da exigência de conhecimentos em análise de dados para a realização de comparações pré e pós-treinamento (Walden, Khayumov, 2020; Paiva et al., 2022).

Com o avanço tecnológico, as estratégias tradicionais de ensino estão sendo gradativamente complementadas por abordagens interativas. Nesse contexto, o uso de recursos como simuladores de treinamento (ST) proporciona experiências controladas de diversas situações no ambiente profissional, sejam elas comuns ou atípicas (Machado, Costa, Moraes, 2018, Paiva, 2022). Os ST's são aplicações que simulam atividades cotidianas, como treinamento e ensino de habilidades, por meio de sistemas computacionais que realizam a atividade proposta com interação entre usuário e máquina (Von Ahn, 2006).

Na área da saúde, esses recursos são aplicados para o tratamento de pacientes, treinamento médico para procedimentos cirúrgicos, simulação de situações de risco e emergência, entre outras aplicações. O uso de um ST permite cometer e corrigir erros sem prejudicar os pacientes e oferece avaliação de desempenho por meio de medidas objetivas imediatas (avaliação *online*) resultantes da interação entre o usuário e o ambiente virtual (Machado *et al*., 2011; Machado, Costa, Moraes, 2018).

A implementação do ST para o TPA, chamado Ouvindo Vozes, como o proposto por Paiva (2022) pode representar um avanço significativo tanto na formação de novos juízes quanto no aperfeiçoamento daqueles com experiência prévia, ao favorecer o refinamento contínuo de suas habilidades avaliativas. O TPA simulado apresenta vantagens como a redução de custos logísticos, a possibilidade de análise objetiva do desempenho dos juízes com base em métricas predefinidas e a flexibilidade de tempo e espaço, uma vez que elimina a necessidade de encontros presenciais. Além disso, um ST pode oferecer maior controle sobre o número e o tipo de estímulos apresentados, o tempo dedicado a cada etapa do treinamento e a estratégia adotada para alcançar níveis aceitáveis de confiabilidade interjuízes.

Um diferencial importante dessa abordagem é a possibilidade de incorporar níveis graduais e balanceados de complexidade nas vozes avaliadas. Essa estrutura favorece a calibração dos estímulos em termos de sua capacidade discriminativa e nível de dificuldade, permitindo que o treinamento seja mais responsivo ao desempenho individual dos usuários.

Essa estratégia segue o princípio do estado de fluxo descrito por Dörner et al. (2016), que propõe a manutenção de um equilíbrio dinâmico entre desafio e habilidade. À medida que o usuário aprimora suas competências, os estímulos tornam-se gradualmente mais exigentes, sem ultrapassar o limiar de frustração ou desmotivação. Isso pode tornar o processo de aprendizagem mais adaptativo, engajador e eficaz para o desenvolvimento das habilidades perceptivo-auditivas.

Para que um ST alcance níveis personalização e eficácia, é essencial que ele incorpore mecanismos capazes de interpretar dados complexos e ajustar o processo de aprendizagem em tempo real. É nesse contexto que os recursos de inteligência artificial, especialmente o Aprendizado de Máquina (AM), tornam-se ferramentas estratégicas. Por meio da análise de padrões os algoritmos de AM permitem automatizar tarefas que tradicionalmente dependeriam da intervenção humana.

O AM é um campo da inteligência artificial que envolve o desenvolvimento de algoritmos capazes de aprender a partir de dados e realizar previsões ou classificações sem serem explicitamente programados para cada tarefa específica. Entre as suas aplicações mais comuns está a classificação de sinais, em que os modelos são treinados para reconhecer padrões em dados complexos e, a partir disso, atribuir rótulos a novos exemplos com base em características previamente aprendidas (Malik *et al*., 2019; Shi; Iyengar, 2020). Esses sinais podem ser de natureza diversa — como imagens, sons, textos ou sinais biomédicos — e os modelos são especialmente úteis quando os padrões relevantes são sutis ou difíceis de serem definidos por regras fixas. Em contextos envolvendo sinais de fala ou voz, os algoritmos de AM têm se mostrado eficazes na detecção de patologias, análise emocional, identificação de locutores e outras tarefas que exigem sensibilidade a variações acústicas.

Modelos baseados em AM oferecem uma abordagem promissora e podem ser utilizados para a classificação da dificuldade do JPA de vozes, ao permitir a identificação de padrões complexos de dados acústicos e perceptivos. Ao serem treinados, esses modelos conseguem aprender quais combinações de características vocais estão associadas a maior ou menor variabilidade nas avaliações, ou seja, a dificuldade da classificação. Com isso,

o algoritmo torna-se capaz de prever, a partir de novas amostras de voz, se um determinado estímulo tende a ser julgado de forma mais consensual entre os juízes — classificando-o como “fácil” — ou se há maior divergência entre os julgamentos, caracterizando-o como “difícil”. Essa classificação automatizada oferece uma contribuição significativa para o desenvolvimento de ST, pois possibilita a seleção de estímulos conforme o nível de dificuldade desejado, facilitando a construção de progressões didáticas mais adaptadas e eficazes para o aprimoramento do JPA.

O objetivo deste estudo é desenvolver e avaliar modelos de AM para rotular a dificuldade do JPA da voz com base no coeficiente de variação do JPA e na concordância Fuzzy de cinco especialistas em voz e, em seguida, classificar a dificuldade do JPA a partir de medidas acústicas. A hipótese é que modelos não supervisionados de AM poderão agrupar as vozes conforme os níveis de dificuldade do JPA e modelos supervisionados de AM, treinados a partir de medidas acústicas extraídas das vozes, serão capazes de prever com acurácia se uma voz será avaliada de forma mais consensual ou com maior divergência entre os juízes. Espera-se que tais modelos identifiquem padrões objetivos que estejam associados à variabilidade perceptiva, permitindo assim automatizar essa classificação e contribuir para o aprimoramento de estratégias de TPA.

1.1 MOTIVAÇÃO E RELEVÂNCIA

O JPA é amplamente utilizado na prática clínica e em pesquisas por sua sensibilidade na identificação de desvios vocais. No entanto, essa técnica está sujeita a uma considerável variabilidade entre juízes, o que compromete a confiabilidade dos resultados e dificulta a padronização do julgamento. Tradicionalmente, os estudos que investigam a confiabilidade do JPA concentram-se em fatores relacionados ao juiz — como experiência e treinamento — ou ao tipo de tarefa de fala, negligenciando a possibilidade de que a própria voz avaliada possa apresentar graus distintos de dificuldade para julgamento. Essa lacuna sugere a necessidade de uma nova abordagem: considerar que a dificuldade perceptiva pode ser uma característica inerente à voz, e não apenas uma limitação do avaliador.

A proposta de classificar a dificuldade perceptiva das vozes representa uma inovação teórica e metodológica significativa na área. Ao empregar modelos de AM para identificar padrões acústicos relacionados à variabilidade dos julgamentos, o projeto oferece uma ferramenta objetiva para estimar a complexidade de uma voz. Essa estimativa é essencial para o desenvolvimento de programas de TPA mais eficazes, nos quais os estímulos vocais devem ser organizados de forma hierárquica, respeitando uma progressão gradual de dificuldade. Estruturar o treinamento com base em níveis crescentes de complexidade é fundamental para promover uma aprendizagem mais consistente, facilitar a adaptação e garantir que os juízes desenvolvam suas habilidades de forma sólida e sistemática (Dörner *et al*., 2016).

Além disso, o presente trabalho representa uma etapa fundamental no processo de aprimoramento do ST Ouvindo Vozes, conforme proposto por Paiva (2022). Uma das principais limitações enfrentadas pelo nosso laboratório é a ausência de um banco de vozes previamente avaliadas por múltiplos juízes, o que inviabiliza, até o momento, a aplicação de modelos baseados na Teoria de Resposta ao Item (TRI). Essa ausência impede a ponderação objetiva do nível de dificuldade das vozes em relação ao JPA, bem como a adoção de estratégias mais robustas para avaliação do desempenho dos usuários em contextos de treinamento. O presente estudo busca suprir essa lacuna ao propor uma metodologia alternativa para estimar a dificuldade das vozes com base em dados de concordância e medidas acústicas.

Portanto, ao fornecer rótulos objetivos para as vozes — derivados de agrupamentos baseados em variabilidade de julgamento e concordância — esta pesquisa contribui diretamente para a construção de um banco de dados adequado e estruturado, que poderá ser integrado a futuros ST’s com base em TRI, ampliando a eficácia, a validade e a progressão das estratégias de capacitação na área da voz.

Portanto, este projeto se justifica pelo potencial de aprimorar estratégias de formação de juízes e pela aplicabilidade prática no desenvolvimento de ferramentas educacionais baseadas em inteligência artificial. Trata-se de uma proposta com relevância científica, clínica e pedagógica, que dialoga com as

demandas contemporâneas da Fonoaudiologia e da tecnologia aplicada à saúde.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Desenvolver um sistema de classificação de padrões para predizer a dificuldade do JPA da voz com base em modelos de AM treinados com dados de concordância e medidas acústicas.

1.2.2 Objetivos Específicos

I. Utilizar modelos de agrupamento para categorizar as vozes segundo o nível de dificuldade de JPA, com base no coeficiente de variação da EAV Fuzzy e na concordância Fuzzy de cinco juízes especialistas em voz;

II. Treinar modelos de AM para classificação do nível dificuldade do julgamento do grau geral, a partir de medidas acústicas;

III. Treinar modelos de AM para classificação do nível dificuldade do julgamento do grau de rugosidade, a partir de medidas acústicas;

IV. Ajustar modelos de AM para classificação do nível dificuldade do julgamento do grau de soprosidade, a partir de medidas acústicas; e

V. Treinar modelos de AM para classificação do nível dificuldade do julgamento do grau de tensão, a partir de medidas acústicas.

1.3 ESTRUTURA DO TRABALHO

A tese está estruturada em 5 capítulos. Sendo eles:1) Introdução (Motivação e Relevância; Objetivos; 2) Fundamentação Teórica; 3) Metodologia; 4) Resultados; 5) Discussão; e 6) Limitações e perspectivas futuras; 7) Conclusão; e 8) Publicações

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste trabalho será dividida em três tópicos, a saber:

1) Julgamento e treinamento perceptivo-auditivo
2) Medidas acústicas
3) Modelos de *Aprendizado de Máquina* em voz

2.1 Julgamento e treinamento perceptivo-auditivo

O JPA da voz é a principal ferramenta para avaliação vocal na prática clínica, permitindo analisar os resultados de intervenções fonoaudiológicas e médicas (Roy *et al*.,2013). Esse procedimento possibilita classificar a qualidade vocal, medir a magnitude do desvio e entender os ajustes laríngeos e de trato vocal durante a emissão da voz. O JPA é uma avaliação de baixo custo, não invasiva, fácil de implementar e acessível a todos os profissionais da área (Ghio *et al*.,2015).

Apesar de seu uso amplo na clínica e em pesquisas, essa avaliação é considerada subjetiva, podendo resultar em variações nas concordâncias intra e interjuízes. Diversas variáveis podem afetar a confiabilidade dessa avaliação, como os padrões internos do juiz (experiência e treinamento auditivo), a tarefa de fala (vogal sustentada, fala espontânea, frases padronizadas, leitura de texto e fala automática) e a escala utilizada (numérica, nominal e escala analógico-visual) (Oates, 2009; Yamasaki; Gama, 2019). Compreender e controlar o efeito dessas variáveis pode aumentar a confiabilidade do JPA, melhorando a concordância intra e interjuízes (Kent, 1996; Chartrand; Belin, 2006).

Um princípio fundamental do JPA, baseado na ciência cognitiva, é a utilização de padrões internos. Quando o juiz faz o julgamento de uma voz, ele acessa esses padrões internos de referência e, pela similaridade das características vocais julgadas e seus padrões internos, realiza a classificação (Ghio *et al*.,2015). A formação dos padrões internos é gerada pelo treinamento e pelas experiências de avaliação que o juiz adquire ao longo de sua trajetória acadêmica e profissional. Juízes mais experientes tendem a ter maior

concordância no JPA do que juízes inexperientes, o que é uma interferência positiva. Contudo, existem algumas críticas sobre a falta de definição do que é considerado experiência. Algumas pesquisas consideram a formação profissional como sinônimo de experiência (fonoaudiólogos e professores de canto) (Eadie; Boven; Stubbs; Giannini, 2010; Sofranko; Prosek, 2012; Yamasaki; Gama, 2019), enquanto outras consideram a experiência como uma característica temporal (juízes com dois ou mais anos de treinamento no JPA) (Sofranko; Prosek, 2014; Yamasaki; Gama, 2019).

O treinamento dos juízes é composto por protocolos específicos e/ou pela prática clínica, que proporciona contato com modelos vocais disfônicos e neutros, favorecendo a consolidação de seus padrões internos. Algumas estratégias que auxiliam o TPA descritas na literatura incluem a apresentação das definições dos parâmetros perceptivo-auditivos, o uso de estímulos âncoras como referências externas e programas específicos de treinamento (Brinca *et al*., 2015). O uso de âncoras auditivas melhora a concordância dos juízes, pois substitui os padrões internos individuais por uma referência comum a todos os juízes. Além disso, entende-se que usar vozes sintetizadas como âncoras nos momentos iniciais do TPA seja mais eficiente do que vozes naturais. Isso ocorre porque, em vozes sintetizadas, os parâmetros vocais podem ser manipulados para gerar vozes unidimensionais (com apenas um parâmetro alterado), simplificando o julgamento perceptivo-auditivo (Chan; Yiu, 2002; Gurlekian; Torres; Vaccari, 2016).

Existem diferentes protocolos para a realização do JPA, desde os mais minimalistas, como a Escala de Desvio Vocal (EDV) e a GRBAS, até os mais complexos, como o CAPE-V, sendo estes os mais aceitos e utilizados mundialmente (Yamasaki; Gama, 2019). A GRBAS, proposta por Hirano (1981), é uma escala numérica composta por quatro pontos (0, 1, 2 e 3) que representam o grau de alteração dos parâmetros avaliados. Esta escala avalia os parâmetros G (*Grade*) – Grau geral de alteração vocal, R (*Roughness*) – Rugosidade, B (*Breathiness*) – Soprosidade, A (*Astheny*) – Astenia e S (*Strain*) – Tensão. Em 1996, Dejonckere e colaboradores sugeriram a adição do parâmetro I (*Instability*) – Instabilidade. Esta escala pode ser utilizada com qualquer tarefa de fala e sua avaliação é focada na laringe. O protocolo CAPE-V, proposto pela American Speech-Language na Hearing Association (ASHA)

(Kempster *et al*.,2009), avalia os mesmos parâmetros da GRBAS (exceto a Astenia), inclui a avaliação do Pitch e da Loudness, além da adição de dois parâmetros, caso necessário. Este protocolo utiliza uma Escala Analógico-Visual (EAV) de 100mm para cada parâmetro avaliado e pontos de corte estabelecidos para os graus leve, moderado e intenso. Neste protocolo, as tarefas de fala são específicas, como o uso de frases padronizadas, vogais sustentadas e conversa espontânea.

A utilização de protocolos padronizados no JPA torna possível a comparação de estudos na área, comparação pré e pós-intervenção e a realização de treinamentos auditivos padronizados. Apesar dos benefícios, o JPA não pode ser restrito aos protocolos porque eles limitam a identificação dos ajustes musculares utilizados pelo indivíduo de acordo com a tarefa de fala utilizada. Assim, deve-se utilizar os protocolos padronizados associados a outras tarefas de fala para a melhor compreensão da funcionalidade vocal do paciente (Yamasaki; Gama, 2019).

As tarefas de fala usadas no JPA, em sua maioria, são de dois tipos: vogais sustentadas ou fala encadeada. A vogal sustentada permite que o juiz tenha informações sobre a fonte glótica, sem a interferência dos ajustes supra glóticos (trato vocal) (Brinca *et al*., 2015). Já com a fala encadeada, é possível extrair informações sobre os ajustes musculares e os padrões de voz usados em situações comunicativas usuais. O TPA deve incluir tarefas de fala dos dois tipos, porque é importante que os juízes iniciantes conheçam aspectos de fonte glótica somados aos ajustes musculares e seus correlatos auditivos, para que sejam internalizados aos seus padrões internos (Yamasaki; Gama, 2019).

Segundo Yamasaki e Gama (2019), o uso e referências externas nos treinamentos perceptivo-auditivos parece ser o melhor caminho a ser seguido para a construção de um modelo de treino para alunos e profissionais da Fonoaudiologia. Além disso, vozes sintetizadas podem ser a melhor opção para âncoras auditivas devido ao controle sobre os parâmetros na geração das vozes (Bispo; Yamasaki; Padovani; Behlau, 2022). Com isso, faz-se necessário o desenvolvimento de ferramentas para o TPA que utilizem âncoras sintetizadas, combinações de tarefa de fala e que compreendam quais as estratégias mais efetivas para o treinamento de juízes para o JPA.

Algumas variáveis importantes para TPA são: tempo de treinamento, tipo de estímulo, parâmetros vocais, tarefa de fala, escalas de avaliação e nível de experiência do juiz (Yamasaki; Gama, 2019). Estas variáveis vêm sendo estudadas, e os resultados encontrados podem justificar algumas escolhas na construção de novos modelos de TPA (Wiet *et al*., 2012). O Quadro 1 apresenta alguns estudos sobre essas variáveis, aponta seus objetivos e os principais resultados encontrados.

Quadro 1: Literatura sobre variáveis relacionadas ao TPA.

Autor	Objetivo	Variável	Resultados
Bassich e Ludlow (1986)	Determinar a validade e confiabilidade do uso de classificações perceptivas para avaliar a qualidade da voz em pacientes com nódulos ou pólipos nas pregas vocais.	Tempo	Verificou que ouvintes inexperientes necessitaram de 8 horas de treinamento para atingir 80% de confiabilidade interjuízes.
Eadie e Baylor (2006)	Determinar se ocorrem alterações na confiabilidade intra e interjuízes para julgamentos de ouvintes inexperientes sobre falantes disfônicos e normais após 2 horas de treinamento do ouvinte.	Tempo	Observou aumento da confiabilidade intra e interjuízes após 2 horas de treinamento.
Barsties *et al.,*(2015)	Avaliar o efeito do feedback visual na classificação do grau geral, rugosidade e soprosidade e na confiabilidade do julgamento da qualidade da voz por	*Feedback*	O treinamento com feedback visual e auditivo pode influenciar minimamente a confiabilidade no julgamento da qualidade de voz,

	ouvintes inexperientes		mas mostrou influência significativa na classificação do grau dos parâmetros de grau geral, rugosidade e soprosidade.
Alves (2019)	Analisar se existe associação entre a experiência do ouvinte, o tipo de tarefa de fala e o JPA da intensidade do desvio vocal e da qualidade vocal predominante.	Tarefa de fala, Nível de experiência, Escala de avaliação	As frases do CAPE-V apresentaram a melhor confiabilidade entre todas as tarefas. A escala analógica visual obteve maior confiabilidade entre os juízes quando comparada a escala numérica. O grupo composto por fonoaudiólogos especialistas foi o que apresentou as melhores taxas de acurácia. A vogal /ɛ/ foi a que apresentou os maiores valores de acurácia em todos os grupos em relação ao juiz de referência.
Santos, Vieira, Sansão e Gama (2019)	Analisar os efeitos do TPA com estímulos âncora de vozes naturais na concordância interjuízes durante a avaliação da qualidade vocal.	Uso de âncoras	Foi observada melhor concordância com o treinamento auditivo com âncoras quando comparado ao sem uso de âncoras, mas essa melhora não alcançou significância

			estatística.
Walden e Khayumov (2020)	Documentar o estado atual do treinamento para realizar a análise perceptivo-auditiva conforme relatado na literatura sobre voz.	Tempo, Tipos de estímulos, Tarefa de fala, Nível de experiência, Parâmetros treinados	Existe uma grande variação nos procedimentos de treinamento utilizados em pesquisas que incluem a avaliação perceptivo-auditiva da qualidade da voz.
Santos, Vieira, Sansão e Gama (2021)	Analisar se a utilização de emissões âncoras de vozes sintetizadas na avaliação perceptivo-auditiva melhora a concordância intra e interavaliador.	Uso de âncoras	A concordância interjuízes foi maior para o grau intenso do parâmetro soprosidade com uso de âncoras quando comparada sem o uso de assim como a concordância intra-juízes do parâmetro rugosidade. O uso de emissões âncoras de vozes sintetizadas diretamente na avaliação melhora a concordância intra e interjuízes na análise perceptivo-auditiva da voz.
Bispo, Yamasaki, Padovani e Behlau (2022)	Verificar o efeito da repetição de estímulos âncoras no JPA realizada por estudantes; analisar a relação entre o número de dimensões vocais dos estímulos	Uso de âncoras, complexidade do estímulo	É importante realizar o TPA de alunos da graduação utilizando estímulos âncoras, iniciar com estímulos sonoros mais simples e com dimensões vocais

	sonoros e a acurácia; e investigar a relação entre o desvio vocal predominante e a acurácia		mais concordantes.
Paz *et al.*,(2023)	Sintetizar o estado do conhecimento científico sobre treinamento para o julgamento perceptivo-auditivo da voz	Âncoras auditivas, *feedback*, Tipos de estímulos, Parâmetros treinados, Tempo	As estratégias mais usuais incluem o uso de âncoras auditivas, *feedback*, vozes naturais com diferentes graus de desvios, sendo os parâmetros soprosidade e/ou rugosidade os mais usuais e breve tempo de treinamento. Há escassez na literatura quanto à aplicação do treino de habilidades de processamento auditivo central no treinamento para o JPA da voz.

Fonte: Elaborado pelo próprio autor, 2024.

2.2 Medidas acústicas

As medidas acústicas da voz permitem quantificar características perceptivas e inferir aspectos do funcionamento laríngeo, integrando os componentes fisiológicos da produção vocal com os aspectos perceptivos da audição. Dessa forma, a voz é compreendida como um fenômeno físico sujeito às influências do desempenho neurofisiológico (Lopes et al., 2013).

Diversas técnicas de análise acústica quantitativa têm sido empregadas na avaliação vocal, com destaque para as abordagens tradicionais (lineares), cepstrais e não-lineares. Cada uma dessas técnicas contribui de maneira específica para a compreensão da qualidade vocal (Brockmann-Bauser; Drinnan, 2011; Titze, 1995). As medidas lineares baseiam-se no modelo fonte-filtro da produção vocal, enquanto as medidas cepstrais avaliam a proeminência dos harmônicos em relação ao ruído do sinal. Já as abordagens não-lineares se fundamentam em modelos caóticos de produção vocal, priorizando aspectos biomecânicos e aerodinâmicos do funcionamento laríngeo (Florencio et al., 2021).

As medidas lineares, como *jitter* e *shimmer*, são amplamente utilizadas por sua extração simples e relação direta com parâmetros fisiológicos da voz. O *jitter* reflete a variação da frequência fundamental (F_0) a curto prazo, enquanto o *shimmer* expressa a variabilidade da amplitude do sinal. Essas medidas são úteis no monitoramento terapêutico e complementam o JPA (Lopes et al., 2020). Contudo, sua confiabilidade é limitada em vozes altamente desviadas, pois a elevada aperiodicidade prejudica a estimativa precisa da F_0 (Lopes; Cavalcante; Costa, 2014; Lopes et al., 2013).

Por outro lado, as medidas cepstrais não exigem a identificação dos ciclos glóticos para estimar a F_0 e avaliar a presença de ruído ou aperiodicidade. Isso as torna mais robustas e confiáveis para a análise de vozes com maior grau de desvio, uma vez que são menos suscetíveis às falhas que comprometem as medidas lineares (LOPES et al., 2020).

As técnicas não-lineares oferecem uma abordagem complementar à análise vocal, baseando-se em modelos que consideram a complexidade e a imprevisibilidade do sistema vocal. Por não dependerem da estimativa da F_0, essas técnicas se concentram em fatores como as propriedades viscoelásticas das pregas vocais, o atrito na coaptação glótica e o fluxo aéreo subglótico, ampliando a compreensão do comportamento vocal em condições patológicas (Florencio et al., 2021; Lopes et al., 2020).

Nas últimas décadas, diversas medidas acústicas foram desenvolvidas com o objetivo de avaliar a qualidade vocal a partir do sinal de voz (Buder, 2000; Maryn et al., 2009; Latoszek et al., 2018). Diante da ampla variedade de medidas existentes, permanece o desafio de identificar aquelas que

apresentam maior correlação com o JPA, com o intuito de detectar a presença e estimar a intensidade do desvio vocal. Além disso, a padronização de métodos analíticos e condutas clínicas ainda é uma demanda constante na área (Maryn et al., 2009; Brockmann-Bauser; Drinnnan, 2011; Latoszek et al., 2018).

Nesse cenário, duas revisões sistemáticas com metanálise foram conduzidas para apontar as medidas acústicas mais fortemente correlacionadas com os escores perceptivos da voz. Maryn et al. (2009) identificaram medidas relacionadas ao grau geral do desvio vocal, como o pico de autocorrelação (RPK), a amplitude de pitch (PA), o nivelamento espectral do resíduo (SFR) e o pico cepstral proeminente suavizado (CPPS) para a tarefa de vogal sustentada, além do CPP, CPPS e relação sinal-ruído Qi para a fala contínua. Já Latoszek et al. (2018) indicaram um conjunto de 13 medidas para predizer rugosidade, incluindo diferentes níveis de ruído espectral (SNL), H2A, RPK, JF, GNE, AVI, SPPQ e CPPS. Para soprosidade, destacaram 12 medidas, como LNPSD, GNE 3000 Hz, H1A1, Hfno, H1H2, SPPQ, HNRD, APQ5, NNE, SAPQ, CPP e CPPS.

Esses achados evidenciam a robustez da análise acústica como ferramenta complementar à avaliação clínica da voz. Contudo, a ampla variabilidade dos desvios vocais – como rugosidade, soprosidade, tensão e instabilidade – e as distintas abordagens metodológicas para rotular os dados trazem desafios à consistência dos resultados entre estudos. Nota-se, por exemplo, uma predominância do uso de escalas numéricas discretas e uma escassez de métodos mais refinados de rotulação, como a escala analógico-visual (EAV) ou modelos fuzzy.

Avançando nesse campo, o estudo de Lima-Filho *et al*. (2024) propôs um modelo preditivo baseado em AM para classificar o grau geral de desvio vocal (GG). Utilizando 47 medidas acústicas extraídas de duas tarefas de fala (vogal sustentada /a/ e fala contínua), combinadas aos escores de rugosidade, soprosidade, tensão e instabilidade atribuídos por cinco fonoaudiólogos, o modelo de melhor desempenho foi o XGBoost. O modelo final, denominado *Integrated Vocal Deviation Index* (IVDI), foi composto por oito variáveis acústicas e perceptivas, apresentando excelente desempenho preditivo, com acurácia de 93,75% e *kappa* ponderado de 0,9374.

Esse trabalho exemplifica de forma consistente o potencial do uso de algoritmos de AM para integrar medidas acústicas objetivas com julgamentos clínicos. A abordagem adotada por Lima-Filho e colaboradores fornece uma ferramenta robusta para apoiar o diagnóstico vocal, otimizando a tomada de decisão terapêutica e favorecendo a personalização do cuidado. O modelo IVDI reforça a relevância da combinação entre tecnologia e expertise clínica e aponta caminhos promissores para novas investigações na área.

Apesar dos avanços, observa-se que os modelos desenvolvidos até o momento têm se concentrado na predição do grau de desvio vocal e na criação de índices acústicos validados, mas não há registros de estudos que explorem a possibilidade de classificar a dificuldade do JPA a partir das características acústicas da voz. Nesse contexto, estudos que buscam preencher essa lacuna, como o presente trabalho, tornam-se especialmente relevantes para ampliar o entendimento sobre a percepção vocal e sua relação com os dados objetivos da voz. Além disso, a classificação do nível de dificuldade das vozes poderia representar uma contribuição significativa para o treinamento de estudantes e fonoaudiólogos, permitindo a organização de bancos de vozes de forma progressiva, com base no grau de complexidade do JPA, favorecendo assim o desenvolvimento de habilidades perceptivo-auditivas de forma mais sistemática e eficaz.

2.3 Modelos de *Aprendizado de Máquina* em voz

O AM é atualmente compreendido como um conjunto de métodos capazes de identificar automaticamente padrões em dados e utilizá-los para prever situações futuras ou apoiar decisões em contextos incertos (Malik *et al*., 2019; Shi; Iyengar, 2020). Sua aplicação se estende a diversas áreas, como a medicina, onde pode auxiliar no diagnóstico e na tomada de decisões. A principal vantagem do AM está na capacidade de classificar informações, reconhecer padrões e gerar respostas com base nos dados fornecidos(Albon, 2018; Wernick *et al*., 2010). Em essência, espera-se que a máquina aprenda com os dados e seja capaz de prever resultados futuros a partir das relações observadas.

O AM refere-se a um conjunto de técnicas computacionais que possibilitam que sistemas aprendam padrões e realizem tarefas específicas a partir de dados, sem que tenham sido explicitamente programados para isso. Esses métodos são particularmente úteis em domínios como o da voz, onde os sinais carregam variabilidades sutis e multifatoriais que muitas vezes escapam à análise estatística tradicional. Com o avanço da capacidade computacional e a disponibilidade de grandes bases de dados, o AM tornou-se uma ferramenta fundamental para a extração de conhecimento a partir de sinais acústicos complexos (Brink *et al.,* 2017; Abreu; 2024; Lima-Filho, Lopes, Silva-Filho, 2024).

No âmbito do AM, os algoritmos são geralmente organizados em duas categorias principais: aprendizado não supervisionado e aprendizado supervisionado. O aprendizado não supervisionado é aplicado quando não se tem rótulos nos dados. Nesse caso, os algoritmos buscam estruturas ou agrupamentos naturais com base em similaridades entre os exemplos. Um dos métodos mais conhecidos é o K-Means, utilizado para agrupar dados em clusters com base na proximidade dos seus atributos. Outros exemplos incluem DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), que identifica agrupamentos com base na densidade dos dados, e modelos de agrupamento hierárquico, úteis para construir relações de similaridade em múltiplos níveis (Jo, 2021; Leite, 2022).

Por outro lado, o aprendizado supervisionado requer um conjunto de dados rotulado, isto é, cada amostra do conjunto de treinamento possui uma entrada (ex: medidas acústicas como *jitter*, shimmer, HNR, formantes etc.) e uma saída conhecida (ex: grau de alteração vocal, diagnóstico clínico etc.). Com base nesses dados, o modelo aprende uma função que mapeia as entradas para as saídas, e pode então aplicar esse conhecimento a novas amostras. Entre os algoritmos mais utilizados nessa categoria estão: *Random Forest* – modelo baseado em múltiplas árvores de decisão, que combina os resultados de várias "árvores" para melhorar a precisão e reduzir o risco de *overfitting*; *Support Vector Machine* (SVM) – método eficaz para separar classes com margens máximas de decisão, útil em problemas de classificação binária e multiclasse; Redes Neurais Artificiais (ANNs) – modelos inspirados no funcionamento do cérebro humano, que são capazes de aprender

representações complexas e não lineares; e *K-Nearest Neighbors* (KNN) – algoritmo simples baseado na ideia de que amostras semelhantes tendem a pertencer à mesmo grupo (Albon, 2018; Leite, 2022).

Enquanto os algoritmos de aprendizado de máquina supervisionado utilizam dados rotulados, com respostas conhecidas, os modelos não supervisionados operam sem a necessidade de rótulos ou resultados previamente definidos. Essa distinção é fundamental para compreender por que os métodos não supervisionados não são aplicáveis diretamente a problemas de regressão ou classificação, pois não possuem os valores de saída como referência. Em geral, os algoritmos de aprendizado não supervisionado são utilizados em três tarefas principais: agrupamento (*clustering*), que organiza instâncias com características semelhantes em grupos; associação, que identifica relações entre dados por meio de regras; e redução de dimensionalidade, que simplifica conjuntos de dados com muitos atributos, mantendo o máximo de informação possível (Jo, 2021; Leite, 2022).

Comparados aos supervisionados, os modelos não supervisionados são mais indicados para tarefas exploratórias e complexas em contextos em que não há dados rotulados disponíveis. No entanto, os algoritmos supervisionados tendem a ser mais precisos, pois contam com a orientação explícita de um especialista sobre os padrões que devem ser aprendidos, enquanto os modelos não supervisionados podem produzir resultados mais imprevisíveis (Jo, 2021).

Na área da voz, o aprendizado de máquina tem sido amplamente utilizado para detectar e classificar patologias vocais (Al-Dhief *et al*, 2019), detectar distúrbios da voz (Hegde, Shetty, Rai, Dodderi, 2018; Ribas *et al.,* 2023; Lima-Filho, Lopes, Silva-Filho, 2024), classificar vozes disfônicas e não-disfônicas (Sabir, *et al.,* 2017; Leite, Moraes, Lopes, 2020), reconhecer emoções por meio da voz (Chen *et al.,* 2020). Esses avanços demonstram o potencial do AM para lidar com tarefas que envolvem alta complexidade e subjetividade, como é o caso do julgamento perceptivo-auditivo da voz.

O uso de modelos de AM em estudos sobre a voz representa um avanço metodológico importante. Isso permite não apenas compreender melhor os fatores que afetam a confiabilidade da avaliação, como também aplicar esse conhecimento no desenvolvimento de ferramentas de TPA mais eficazes. As

técnicas de AM proporcionam maior flexibilidade e robustez, especialmente ao lidar com dados de alta dimensão e interações complexas (Sanchez-Pinto *et al.,* 2018; D'souza, Prema, Balaji; 2020). Os métodos de AM também tendem a melhorar o poder preditivo e são mais eficazes em grandes conjuntos de dados (Sanchez-Pinto *et al.,* 2018; D'souza; Prema; Balaj, 2020; Kaneko, 2021; Charilaou; Battat, 2022). Embora os modelos estatísticos ainda possam ser úteis em conjuntos de dados menores e por sua eficiência computacional, os benefícios gerais das técnicas de AM as tornam uma escolha superior para seleção de variáveis em muitos cenários (Sanchez-Pinto *et al.,* 2018). Além disso, as técnicas de AM são menos suscetíveis à especificação incorreta em cenários com interações não lineares e estruturas de dados complexas (como no caso de vozes disfônicas). Neste contexto, as técnicas de AM surgiram como soluções promissoras para melhorar a eficácia e a confiabilidade do JPA na classificação da gravidade dos distúrbios da voz. A aplicação de AM pode ajudar a padronizar e objetivar o processo de avaliação vocal, reduzindo a subjetividade inerente ao JPA.

O uso de AM não apenas melhora a classificação da gravidade dos distúrbios vocais, como também fornece suporte significativo para a tomada de decisões clínicas. Sistemas baseados em AM podem fornecer feedback em tempo real, auxiliando os médicos a identificar padrões que podem não ser facilmente discerníveis por meio da JPA tradicional (Sabir, *et al.,* 2017; Hegde, Shetty, Rai, Dodderi, 2018; Al-Dhief *et al*, 2019; Leite, Moraes, Lopes, 2020; Ribas *et al.,* 2023). Além disso, esses sistemas podem ser usados para monitorar o progresso do paciente durante o tratamento, oferecendo uma ferramenta objetiva para avaliar a eficácia das intervenções terapêuticas (Melley; Sataloff, 2022).

O Quadro 2 apresenta uma síntese de estudos que aplicaram modelos de AM na área da voz, destacando seus objetivos, os métodos de AM utilizados e os principais resultados obtidos. Esses trabalhos ilustram a diversidade de abordagens empregadas para a detecção, classificação ou análise de alterações vocais, evidenciando a crescente relevância das técnicas de aprendizado de máquina nesse campo. A tabela permite observar tanto a aplicação de algoritmos clássicos, como SVM e MLP, quanto o uso de métodos mais recentes, como aprendizado auto supervisionado e modelos sequenciais,

além de refletir os avanços na acurácia e eficiência dos sistemas desenvolvidos.

Quadro 2: Aplicações de AM na área da voz.

Autor	Objetivo	Método de AM	Principais Resultados
Sabir *et al.*, 2017	Melhorar a acurácia na identificação de vozes patológicas e normais.	Algoritmo modificado baseado em energia e entropia de sinais, com AM (SVM, MLP)	O algoritmo proposto alcançou acurácia de até 97% com MLP, demonstrando eficácia na diferenciação entre vozes normais e patológicas.
Hegde *et al.*, 2019	Revisar abordagens de aprendizado de máquina para detecção automática de distúrbios vocais.	SVM, KNN, *Random Forest,* ANN, HMM, GMM	Identificou que SVM e Random Forest estão entre os modelos mais eficazes. Ressaltou a importância da escolha dos atributos acústicos e da base de dados usada.
Chen *et al.*, 2020	Reconhecer emoções na fala através de técnicas de aprendizado de máquina	*Random forest* multifuzzy em duas camadas	Demonstrou eficácia na extração de representações acústicas relevantes para classificação emocional
Al-Dhief et al., 2021	Detectar e classificar patologias vocais com alto desempenho e rapidez utilizando dados acústicos.	*Online Sequential Extreme Learning Machine* (OS-ELM)	O modelo OS-ELM apresentou alta acurácia (>95%) na detecção de vozes patológicas, com tempo computacional reduzido, mostrando ser eficiente para aplicações em tempo real.
Leite; Moraes; Lopes, 2022	Comparar o desempenho de diferentes modelos de AM na classificação	SVM, *Decision Trees*, KNN, *Random Forest, Logistic*	SVM e MLP apresentaram os melhores desempenhos, com acurácia superior a 90%. O estudo reforça o potencial de AM na triagem

	de vozes disfônicas e normais.	*Regression*, MLP	vocal.
Melley; Sataloff, 2022	Discutir o uso atual e futuro da inteligência artificial na laringologia, além da aplicação prática em voz.	Não especifica um modelo; trata-se de um artigo de opinião/reflexão	Enfatiza o potencial da IA para diagnóstico, triagem e suporte clínico, destacando desafios éticos, de validação clínica e integração com a prática médica.
Ribas *et al*., 2023	Utilizar representações auto-supervisionadas para detecção automática de distúrbios vocais.	*Self-supervised learning* com modelos baseados em *embeddings* acústicos	As representações auto supervisionadas superaram os métodos clássicos em sensibilidade e especificidade, especialmente em contextos com poucos dados rotulados.
Lima-Filho; Lopes; Silva-Filho, 2024	Desenvolver um modelo de AM para classificar automaticamente o grau geral de desvio vocal com base em medidas acústicas.	Regressão linear, árvore de decisão, Random Forest, k-vizinhos mais próximos (KNN), máquinas de vetores de suporte linear (SVML), máquinas de vetores de suporte de kernel gaussiano (SVMGK), máquinas de vetores de suporte de kernel polinomial (SVMPK),	O modelo XGBoost previu e classificou melhor a GG. O modelo final foi denominado IVDI e incluiu oito medidas acústicas e perceptivo-auditivas. O IVDI apresentou excelente desempenho na predição e classificação da GG.

		Multilayer Perceptron (MLP), redes neurais ensacadas (BNNs) e XGBoost.	

Fonte: Elaborado pelo próprio autor, 2025.

Apesar dos avanços observados nos estudos apresentados, nenhum deles teve como foco a classificação da dificuldade do JPA das vozes. As pesquisas concentram-se, predominantemente, na detecção de distúrbios vocais ou na diferenciação entre vozes normais e patológicas, sem considerar o grau de variabilidade entre juízes ou a complexidade perceptiva dos estímulos vocais. Essa lacuna evidencia a originalidade e a relevância do presente estudo, que propõe investigar um aspecto ainda pouco explorado: a dificuldade inerente ao JPA de determinadas vozes, utilizando modelos de AM para classificá-las com base nesse critério.

3 METODOLOGIA

3.1 Delineamento, período de referência e local do estudo

Esta pesquisa caracteriza-se como um estudo retrospectivo, uma vez que parte da análise de um desfecho já ocorrido (FREITAS, 2017). Trata-se de uma investigação documental, de delineamento transversal e com enfoque em acurácia diagnóstica de modelos baseados em AM, utilizando registros previamente coletados e armazenados. O estudo foi avaliado e aprovado pelo Comitê de Ética em Pesquisa da instituição de origem, sob o parecer consubstanciado nº 52492/12.

A pesquisa foi conduzida entre janeiro de 2024 e julho de 2025, no Laboratório Integrado de Estudos da Voz (LIEV) da Universidade Federal da Paraíba (UFPB). O estudo utilizou um banco de vozes pertencente ao próprio laboratório, amplamente utilizado em investigações anteriores desenvolvidas no mesmo contexto (Abreu, 2024).

3.2 Amostra do estudo e procedimentos de coleta de dados

O banco de vozes utilizado contém amostras de 295 indivíduos com e sem queixa vocal, a média de idade dos pacientes é de 36,47 ± 12,07 anos, sendo 230 do gênero feminino e 65 do gênero masculino. Esses indivíduos apresentaram os seguintes diagnósticos otorrinolaringológicos: 77 (26,10%) sem alteração estrutural ou funcional na laringe, 67 (22,71%) pacientes com nódulos vocais, 55 (18,65%) com distúrbio de voz secundário a refluxo laringofaríngeo, 28 (9,49%) com cisto vocal, 25 (8,47%) com fenda triangular médio-posterior, 17 (5,76%) com paralisia unilateral de prega vocal, 15 (5,09%) com pólipo de prega vocal, 6 (2,03%) com sulco vocal e 5 (1,70%) com edema de Reinke.

As vozes foram coletadas em uma cabine acústica com tratamento sonoro adequado e nível de ruído inferior a 50 dB NPS. Utilizou-se uma taxa de amostragem de 44,1 kHz, com resolução de 16 bits por amostra, mantendo-se uma distância de 10 cm entre o microfone e os lábios do participante. O equipamento utilizado incluiu o *software Fonoview* (versão 4.5, CTS Informática), computador *desktop Dell All-in-One*, microfone cardioide

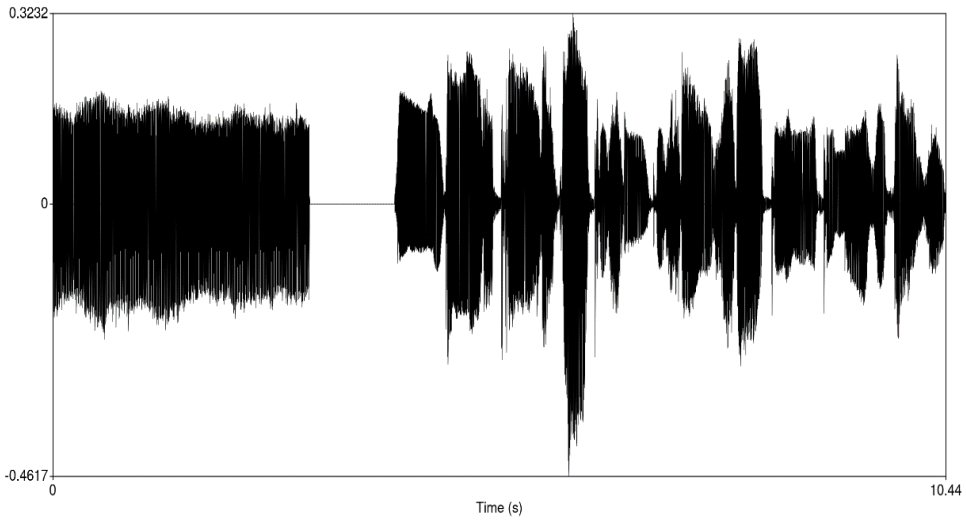
unidirecional Sennheiser modelo E-835, posicionado em pedestal e conectado a um pré-amplificador *Behringer* modelo U-Phoria UMC 204.

Durante a coleta, o participante permaneceu em pé, posicionado de frente para o pedestal, respeitando a distância previamente estabelecida. Após receber instruções, realizou-se a gravação, que incluiu a emissão sustentada da vogal /a/ e a contagem de um a dez, ambas em frequência e intensidade autorreferidas como confortáveis e habituais.

Após a coleta, os sinais vocais foram editados no *software SoundForge* (versão 10.0). Nos registros da vogal sustentada, foram removidos os dois segundos iniciais e finais, devido à maior instabilidade acústica nesses trechos, assegurando-se um tempo mínimo de três segundos de emissão contínua. Para a contagem numérica, foram eliminadas pausas de silêncio no início e no fim da gravação. Em seguida, todos os sinais foram normalizados, a fim de padronizar o volume de saída entre -6 dB e +6 dB.

Para a extração de medidas acústicas e realização do JPA, foi utilizado um *script* no *software* Praat que automatizou a concatenação das tarefas de fala de cada paciente: vogal sustentada, um segundo de silêncio e contagem. Essa organização padronizada das amostras foi aplicada uniformemente a todos os participantes, conforme ilustrado na Figura 1 (Abreu, 2024).

Figura 1: Oscilograma da vogal sustentada /a/, silêncio e contagem.



Fonte: Abreu, 2024.

A vogal sustentada /a/ foi escolhida por ser amplamente utilizada em pesquisas sobre qualidade vocal (Kempster, 2009), uma vez que permite

melhor análise das características acústicas da voz por apresentar um sinal quase periódico e menor variabilidade em comparação à fala encadeada. Essa tarefa facilita a identificação de alterações relacionadas à função vocal e à biomecânica da produção vocal (Brinca *et al*., 2015; Costa *et al*., 2023). Por outro lado, a fala conectada — como contagens numéricas, leitura de frases foneticamente balanceadas ou textos — também é comum em estudos na área de voz, por refletir um pouco mais o uso cotidiano da voz.

Embora distintas, as tarefas de vogal sustentada e fala encadeada são complementares. A vogal sustentada favorece a detecção direta de desvios na qualidade vocal relacionados a fonte glótica, enquanto a fala conectada fornece uma perspectiva mais funcional do uso vocal do indivíduo (Yamasaki; Gama, 2019).

I. Extração de medidas acústicas

Todas as medidas acústicas foram extraídas por meio do *software* Praat, no *plugin* "VoxMore: Análise Acústica", desenvolvido por Abreu et al. (2023). Um total de 50 medidas foram extraídas a partir do sinal de fala de cada voz, sendo elas:

- Autocorrelação;
- Medidas da frequência fundamental (f_o): média (f_o_média), desvio padrão (f_o_DP), primeiro quartil (f_o_q1), mediana (f_o_mediana), terceiro quartil (f_o_q3) e coeficiente de variação (f_o_CV);
- Medidas do período: média, desvio padrão (PSD) e logaritmo natural do desvio padrão (LNPSD);
- Medidas de perturbação de curto termo da f_o (*jitter*): *jitter* total, *jitter* local, *jitter* RAP, *jitter* PPQ5 e *jitter* DDP;
- Medidas de perturbação de curto termo da amplitude (*shimmer*): *shimmer* local, *shimmer* em dB, APQ3, APQ5, APQ11, *shimmer* DDP e AVI;
- Medidas espectrais: SNL1 a SNL6, H1-H2, H1-A1, H1-A3, HNR (relação harmônico-ruído), desvio padrão do HNR (HNR_DP), HNRD, Hfno, declínio espectral e tilt;
- Medidas de sinal residual: PA e SFR;

- Medidas de ruído glotal (GNE): excitação do ruído glotal nas bandas de 1.000 Hz (GNE1), 2.000 Hz (GNE2) e 3.000 Hz (GNE3);
- Medidas cepstrais: CPP, f_o do CPPS (CPPS_ f_o), média do CPPS (CPPS_média), desvio padrão (CPPS_DP), primeiro quartil (CPPS_q1), mediana (CPPS_mediana), terceiro quartil (CPPS_q3) e desvio absoluto da mediana (CPPS_DAM).

A seleção dessas medidas foi fundamentada em evidências da literatura que apontam sua correlação com o JPA (MARYN *et al*., 2009; LATOSZEK *et al*., 2018), sua presença em índices acústicos consolidados (WUYTS *et al*., 2000; MARYN *et al*., 2010; PETERSON *et al*., 2013; LATOSZEK *et al*., 2017), além de serem amplamente utilizadas em pesquisas com análise acústica da voz (BUDER, 2000).

II. Realização do julgamento perceptivo-auditivo

A aplicação do JPA foi realizada por meio do programa VoxMore: Julgamento Perceptivo-Auditivo, desenvolvido por Abreu (2024). Optou-se pela utilização da escala analógico-visual (EAV) por apresentar, de modo geral, melhores índices de concordância inter e intrajuiz (Martins, Couto, Gama, 2015; Contreras-Ruston *et al*., 2021). A EAV também oferece maior sensibilidade ao julgamento do avaliador, permitindo marcações contínuas ao longo de uma linha de 100 mm, cuja extremidade esquerda representa “sem desvio” e a direita, “desvio intenso”, superando a limitação de escalas categóricas frequentemente usadas na prática clínica e em pesquisas.

Diferentemente da versão impressa da EAV, que pode estar sujeita a perdas físicas, erros de escala, digitação ou arquivamento, o programa utilizado no presente estudo viabiliza o registo digital padronizado e seguro dos julgamentos. Além disso, permite a escuta das tarefas de fala, a marcação da intensidade de cinco qualidades vocais — grau geral, rugosidade, soprosidade, tensão e instabilidade.

O *VoxMore: Julgamento Perceptivo-Auditivo* foi construído para que o juiz possa selecionar a tarefa de fala, escutar a amostra de fala da tarefa selecionada, avaliar a intensidade de cinco qualidade vocais como grau geral, rugosidade, soprosidade, tensão e instabilidade, se necessário escutar estímulos âncora dessas qualidades, acompanhar o quantitativo de amostras

julgadas em relação ao total através de uma barra de evolução, e salvar o julgamento.

Para o JPA, foi convidado um juiz vinculado à instituição onde o presente estudo foi conduzido, em razão de sua expertise na área de análise acústica aplicada à clínica vocal. Por conveniência, foram indicados outros quatro juízes com perfil semelhante, que aceitaram participar voluntariamente do estudo, totalizando cinco juízes.

Os critérios de inclusão para a seleção dos juízes foram: formação em Fonoaudiologia; titulação de doutorado; especialização em voz; experiência mínima de dez anos em análise acústica vocal na prática clínica; produção científica relevante na área; vinculação a grupo de pesquisa; e prática profissional baseada em evidências científicas. Essas informações foram obtidas por meio da análise do Currículo *Lattes* e confirmadas durante o processo de recrutamento. Os juízes foram contatados por *e-mail* ou telefone e, após aceitarem participar, receberam uma carta convite e um tutorial de uso do programa *VoxMore*: Julgamento Perceptivo-Auditivo (Abreu, 2024).

Antes do início dos julgamentos, os juízes foram orientados a escutar os estímulos-âncora previamente selecionados. Essas amostras haviam sido analisadas anteriormente por fonoaudiólogos experientes e são rotineiramente utilizadas para fins de treinamento e padronização perceptiva no LIEV.

Durante a aplicação do JPA, os juízes foram instruídos a avaliar os seguintes parâmetros vocais: Grau geral do desvio vocal (GG) – impressão auditiva global do desvio da qualidade vocal; Grau de Rugosidade (GR) – percepção auditiva de irregularidade e crepitação na emissão;Grau de Soprosidade (GS) – percepção de escape de ar audível durante a fala; Grau de Tensão (GT) – impressão auditiva de esforço vocal e constrição glótica/supraglótica; e Grau de Instabilidade – percepção de flutuações na frequência (*pitch*) e na intensidade (*loudness*) da emissão vocal. Para o presente estudo, foram utilizados, apenas, os parâmetros GG, GR, GS e GT. Isso porque esses parâmetros são os mais utilizados durante o TPA e nas pesquisas a área de voz.

Para evitar a fadiga auditiva, foi solicitado que os juízes realizassem pausas a cada bloco de até 50 amostras. Cada juiz avaliou um total de 354 amostras, incluindo 295 vozes únicas e uma replicação aleatória de 20%

dessas (59 amostras), com o objetivo de calcular a concordância intrajuiz por meio do Coeficiente de Correlação Intraclasse (CCI).

Além de verificar a consistência intrajuiz e interjuiz nos julgamentos, o CCI também foi utilizado como um parâmetro para avaliar a adaptação dos juízes ao uso do programa *VoxMore*. Esperava-se que os valores de concordância obtidos com a versão digital fossem comparáveis aos tradicionalmente alcançados com a escala impressa em papel. Após a finalização dos julgamentos, cada juiz enviou seu arquivo com as avaliações por e-mail à equipe responsável pela análise dos dados (Abreu, 2024).

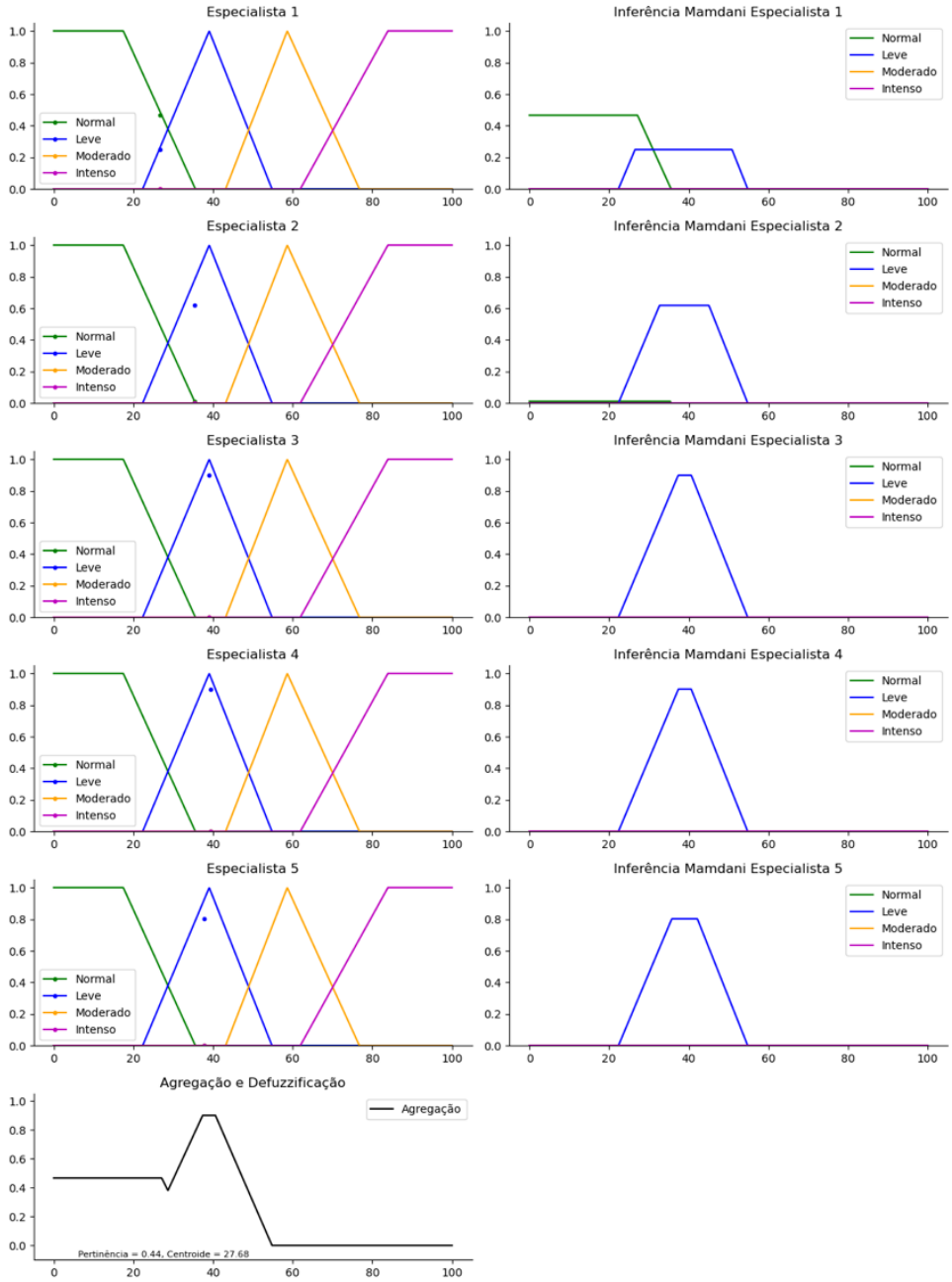
Inicialmente, os resultados do JPA realizados por cada juiz foram reunidos em um único arquivo. A partir desses dados, foi criado um vetor contendo os valores agregados, que serviu como base para a segmentação das amostras em grupos utilizando o algoritmo de agrupamento *k*-médias (*k*-means). De forma empírica, foram definidos quatro clusters, nomeados como: normal, leve, moderado e intenso. Para cada um desses clusters, foram calculados o valor do centróide e a dispersão dos dados (Abreu, 2024).

Com base nesses agrupamentos, foi construída uma EAV com lógica *fuzzy*. Essa lógica é uma abordagem matemática que permite lidar com incertezas e variações graduais, útil em contextos subjetivos como o JPA. Em vez de trabalhar com categorias rígidas, a lógica fuzzy atribui graus de pertinência a diferentes valores, permitindo que uma mesma voz seja parcialmente classificada em mais de uma categoria. Isso possibilita representar de forma mais realista a natureza contínua e ambígua das avaliações humanas, contribuindo para uma análise mais refinada da concordância entre juízes. O domínio da escala compreendeu o intervalo de 0 a 100, conforme o padrão da EAV tradicional. Os conjuntos linguísticos foram representados pelos quatro níveis de desvio vocal identificados anteriormente, e suas respectivas funções de pertinência foram modeladas a partir dos valores de centróide e dispersão. Para os conjuntos *normal* e *intenso*, utilizaram-se funções trapezoidais, enquanto os conjuntos *leve* e *moderado* foram modelados por funções triangulares (Abreu, 2024).

A Figura 2 ilustra o processo de aplicação da lógica *fuzzy* para obtenção do valor final EAV a partir dos julgamentos de cinco especialistas, com base no GG de uma amostra. Nos primeiros cinco gráficos à esquerda, são exibidas as

funções de pertinência *fuzzy* utilizadas para cada juiz, contendo os conjuntos linguísticos "normal", "leve", "moderado" e "intenso". Cada ponto azul indica o valor de julgamento dado por cada especialista para a amostra, e sua localização em relação às curvas de pertinência demonstra em quais categorias linguísticas o valor se insere com maior ou menor grau de pertencimento. Conforme essa figura o GG de cada juiz foi: juiz 1 (26,7) juiz 2 (35,7) juiz 3 (39), juiz 4 (39,4) e do juiz 5 (38,8) e a EAV *fuzzy* calculou o grau geral do desvio como 56,22.

Figura 2: Aplicação da lógica fuzzy para obtenção do valor final da EAV fuzzy a partir dos julgamentos de cinco especialistas.



Fonte: Elaborado por Abreu, 2024.

Na sequência, os cinco gráficos à direita representam o processo de inferência Mamdani para cada juiz, demonstrando quais conjuntos linguísticos foram ativados com base nos valores julgados. Essa etapa aplica a T-norma mínima para combinar a pertinência do valor com as funções correspondentes, resultando em áreas de ativação mais ou menos intensas em cada conjunto linguístico.

Por fim, o gráfico inferior mostra a etapa de agregação e defuzzificação. As inferências dos cinco juízes são combinadas utilizando a T-conorma do tipo máximo, formando um único conjunto agregado. Esse conjunto representa o consenso ponderado entre os juízes. A defuzzificação é realizada pelo método do centróide, gerando o valor numérico final da EAV fuzzy, que, nesse exemplo, corresponde a 27,68. A pertinência máxima da curva agregada foi de 0,44, refletindo o grau de confiança do sistema *fuzzy* em relação à classificação obtida. Essa abordagem permite considerar simultaneamente os julgamentos individuais e suas variações, resultando em uma rotulação contínua e mais robusta das amostras vocais.

III. Variáveis utilizadas nos modelos de AM

A partir do sistema de inferência *fuzzy*, foi criada uma variável denominada Concordância EAV Fuzzy, com o objetivo de quantificar o nível de concordância entre os juízes com base nas áreas de maior ativação obtidas nas inferências Mamdani. Para cada amostra, observou-se a área de maior ativação em cada conjunto linguístico (normal, leve, moderado ou intenso), conforme os resultados da inferência *fuzzy* de cada juiz. Em seguida, foi contabilizado quantos juízes apresentaram a mesma área com maior grau de ativação. O valor da variável Concordância EAV *Fuzzy* poderia variar de 1 a 5, indicando, respectivamente, se não houve concordância entre nenhum dos juízes ou até todos os cinco juízes atribuíram a maior ativação ao mesmo conjunto linguístico.

Esse procedimento possibilita uma análise complementar sobre a coerência entre os julgamentos, considerando não apenas os valores absolutos atribuídos, mas também o grau de sobreposição entre as classificações *fuzzy* de cada especialista.

Além disso, foi calculado um coeficiente de variação (CV) do valor numérico da EAV inicial dos cinco juízes para cada uma das amostras vocais. Esse coeficiente foi obtido a partir dos valores do JPA atribuídos por cada juiz, permitindo quantificar a dispersão relativa dos julgamentos em relação à média. O uso do CV possibilita identificar quais vozes apresentaram maior ou menor variabilidade entre os juízes, fornecendo um indicativo adicional da consistência perceptiva entre os juízes. Assim, quanto menor o coeficiente de variação, maior a homogeneidade dos julgamentos atribuídos àquela amostra vocal, refletindo maior concordância na percepção da intensidade do desvio vocal.

Todos os procedimentos descritos para o GG também foram aplicados aos GR, GS e GT. Isso inclui a utilização da lógica *fuzzy* para combinar os julgamentos dos juízes, a construção da EAV *fuzzy* a partir dos clusters definidos pelo k-médias, a aplicação da inferência de Mamdani, a defuzzificação para obtenção do valor numérico, a criação da variável de concordância EAV *fuzzy* com base na área de maior ativação comum entre os

juízes, e o cálculo do coeficiente de variação para cada amostra. Dessa forma, foi mantida a padronização metodológica na análise de todos os parâmetros vocais.

Todas as informações sobre as variáveis dependentes e independentes inerentes as modelagens de AM para o presente estudo são apresentadas no Quadro 3.

Quadro 3: Definição de variáveis.

Nome	Descrição	Natureza	Categorias/Escala de mensuração
Grau de Dificuldade	Rótulo da dificuldade do JPA da voz, obtida por agrupamento de cada parâmetro (GG, GR, GS e GT)	Categórica ordinal	Fácil, Médio, Difícil
Medidas Acústicas	Conjunto de 50 medidas acústicas extraídas das tarefas de fala	Quantitativa contínua	Intervalar (valores contínuos em escalas específicas de cada medida)
Coeficiente de Variação	Coeficiente de variação do JPA entre os juízes para cada parâmetro (GG, GR, GS e GT)	Quantitativa contínua	Razão (0 a ∞)
Concordância EAV Fuzzy	Número de juízes que apresentaram a mesma área com maior ativação na inferência Fuzzy para cada parâmetro (GG, GR, GS e GT)	Quantitativa discreta	1 a 5

Fonte: Elaborado pelo próprio autor, 2024.

IV. Modelos de Aprendizado de Máquina

Foram utilizadas duas abordagens de Aprendizado de Máquina (AM), a saber: não supervisionada e supervisionada. No campo da voz, a maioria das inovações envolvendo AM concentra-se principalmente em técnicas supervisionadas (Hegde et al., 2019), devido à disponibilidade, mesmo que restrita, de bases de dados previamente rotuladas. O aprendizado supervisionado tende a apresentar melhor desempenho em relação ao não supervisionado, uma vez que os dados utilizados para treinar o modelo já estão associados aos desfechos esperados, o que facilita o processo de aprendizagem (Leite, 2022).

A primeira etapa foi de modelagem não supervisionada, com o objetivo de agrupar automaticamente as vozes em grupos representativos do grau de dificuldade de JPA, utilizando técnicas de agrupamento. Para isso, aplicou-se a agrupamento sobre duas variáveis principais: o coeficiente de variação (CV)

dos julgamentos dos juízes, que reflete a variabilidade perceptiva quanto ao grau geral de desvio vocal, e a variável de Concordância EAV Fuzzy, que representa o número de juízes que atribuíram à amostra a mesma área de maior ativação na inferência fuzzy. Foram comparados três algoritmos de agrupamento: K-means, Método Hierárquico e Fuzzy C-means. Cada modelo foi avaliado por meio de critérios internos de qualidade de cluster, incluindo o índice de Calinski-Harabasz. Esse índice é uma métrica utilizada para avaliar a qualidade de agrupamentos gerados por algoritmos de agrupamento. Ele considera a razão entre a variabilidade entre os grupos e a variabilidade dentro dos grupos. Valores mais altos indicam agrupamentos mais bem definidos, com maior separação entre os grupos e maior coesão interna, sendo, portanto, um critério útil para a escolha do número ideal de clusters (Calinski, Harabasz, 1974).

A escolha do número ideal de clusters foi baseada no índice de Calinski-Harabasz, no método do cotovelo e na visualização dos clusters. O método do cotovelo é uma abordagem visual amplamente utilizada para determinar o número ideal de clusters em algoritmos de agrupamento. Ele consiste em calcular uma métrica de variabilidade intra-grupo para diferentes números de grupos. Em seguida, os valores são plotados em um gráfico. O ponto em que ocorre uma redução abrupta na taxa de diminuição da variabilidade, formando um “cotovelo” na curva, é considerado o número ideal de agrupamentos, pois a partir desse ponto os ganhos adicionais de divisão se tornam marginais.

Para fins de interpretação e posterior aplicação de modelos supervisionados, os rótulos gerados pelo método de agrupamento selecionado foram incorporados ao conjunto de dados original. Cada grupos foi nomeado com categorias interpretáveis: "Fácil", "Médio" e "Difícil". A partir dessa rotulação, foi elaborada uma tabela resumo apresentando a frequência absoluta e o percentual de indivíduos em cada grupo, o que permite visualizar de forma clara e objetiva a distribuição das amostras de voz entre os diferentes níveis de dificuldade perceptiva.

Após a definição dos rótulos de dificuldade obtidos por meio de algoritmos de agrupamento, foi realizada a etapa de modelagem supervisionada com o objetivo de prever automaticamente o grau de dificuldade do JPA das vozes, utilizando medidas acústicas como preditores.

No entanto, neste estudo, optamos por incluir medidas acústicas com o objetivo de validar os agrupamentos gerados na etapa de aprendizado não supervisionado, considerando que esses grupos foram formados a partir de informações extraídas apenas do JPA. Importante ressaltar que não é objetivo deste trabalho inferir quais medidas acústicas são preditoras da classificação, mas sim verificar a consistência dos grupos gerados com base em variáveis objetivas.

O conjunto de dados foi particionado em 70% para treinamento e 30% para teste, com amostragem estratificada pelo nível de dificuldade do JPA para manter o equilíbrio entre os grupos. O pré-processamento incluiu a padronização das variáveis quantitativas, transformação de variáveis categóricas em *dummies* e eliminação de preditores com variância próxima de zero. Para ajuste dos hiperparâmetros, foi empregada validação cruzada estratificada com cinco *folds*, repetida 30 vezes. Nesse processo foi considerado a métrica acurácia balanceada. A seleção dos melhores hiperparâmetros foi realizada por busca aleatória com base em um hipercubo latino. Foram avaliados 14 algoritmos de classificação, a saber: Discriminante Regularizada (RDA), Discriminante Flexível (FDA), Naive Bayes, SVM Linear, SVM Polinomial, Regressão Logística Multinomial, Discriminante Quadrática, Rede Neural MLP, Discriminante Linear, Floresta Aleatória, K-Vizinhos Mais Próximos, SVM com Kernel RBF, Árvore de Decisão e Árvore Boosting (XGBoost).

A Análise Discriminante Linear (LDA) é um método clássico de classificação que projeta os dados em um espaço de menor dimensão, maximizando a separação entre classes com base na variância entre e dentro dos grupos. A Análise Discriminante Quadrática (QDA) permite fronteiras de decisão não lineares ao estimar uma matriz de covariância específica para cada classe, diferentemente da LDA, que assume uma mesma matriz para todas.

A RDA combina aspectos das abordagens de análise discriminante linear e quadrática, aplicando penalizações para lidar com problemas de multicolinearidade e superajuste. Já a FDA permite a modelagem de fronteiras de decisão mais complexas ao incorporar funções de suavização, sendo útil para dados não linearmente separáveis.

O Naive Bayes é um classificador probabilístico simples baseado no Teorema de Bayes, assumindo independência entre as variáveis preditoras. Embora essa suposição raramente se sustente na prática, o modelo costuma apresentar bom desempenho em diversas aplicações, especialmente com dados textuais e de alta dimensionalidade.

Os modelos de Máquinas de Vetores de Suporte (SVM) foram utilizados com diferentes kernels: Linear, Polinomial e RBF (Radial Basis Function), cada um com diferentes capacidades de modelar padrões lineares ou não lineares. A SVM Linear busca maximizar a margem entre as classes, sendo mais eficaz quando os dados são linearmente separáveis. A SVM Polinomial adiciona curvatura às fronteiras de decisão, enquanto a SVM RBF é indicada para problemas mais complexos com estruturas não lineares (Morettin, Singer, 2022).

A Regressão Logística Multinomial é uma extensão da regressão logística clássica para múltiplas classes, modelando diretamente as probabilidades de cada classe com base em funções lineares das variáveis preditoras.

A Rede Neural Multilayer Perceptron (MLP) é uma arquitetura de aprendizado profundo composta por múltiplas camadas de neurônios artificiais. Seu poder de generalização e capacidade de capturar relações complexas entre as variáveis a torna uma opção robusta para tarefas de classificação.

A Floresta Aleatória (Random Forest) é um conjunto de árvores de decisão treinadas com amostragem aleatória dos dados e dos atributos, reduzindo o risco de superajuste e aumentando a robustez do modelo.

O algoritmo dos K-Vizinhos Mais Próximos (KNN) classifica novas observações com base na proximidade a exemplos rotulados, sendo sensível à escala dos dados e ao valor de K escolhido.

A Árvore de Decisão constrói um modelo baseado em divisões sequenciais dos dados, facilitando a interpretação, embora esteja mais propensa ao superajuste.

Por fim, a Árvore Boosting (XGBoost) é um método de aprendizado em conjunto baseado em boosting, que combina várias árvores fracas para formar um classificador forte, com alta capacidade preditiva e controle de

regularização, sendo um dos algoritmos mais eficazes em competições de ciência de dados.

Os modelos finais, ajustado com os hiperparâmetros otimizados, foi avaliado no conjunto de teste a partir de diversas métricas, como acurácia, acurácia balanceada, AUC-ROC, F1-score, Kappa, sensibilidade, especificidade, e valores preditivos positivo (VPP) e negativo (VPN). Para a análise da performance dos modelos com base nas métricas utilizadas, foram consideradas as classificações propostas por Landis *et al*., (1977) e Hosmer e Lemeshow (2000), conforme apresentado no Quadro 4.

Quadro 4: Classificação para os valores do Coeficiente Kappa (LANDIS et al., 1977), acurácia, sensibilidade e especificidade (HOSMER E LEMESHOW, 2000).

Variável	Valores	Classificação
Coeficiente Kappa ponderado	0,81 – 1,00	Quase perfeita
	0,61 – 0,80	Bom
	0,41 – 0,60	Moderada*
	0,21 – 0,40	Regular
	0,00 – 0,20	Discreta
	<0,00	Pobre
Acurácia, AUC-ROC, F1-socre, Sensibilidade, Especificidade, VPP e VPN	> 0,90	Excelente
	0,80 – 0,90	Bom
	0,70 – 0,80	Aceitável
	0,60 – 0,70	Regular*
	<0,60	Sem capacidade de discriminação aceitável

Legenda: *Classificação aceitável para a variável.

Fonte: Elaborado pelo próprio autor, 2025.

A sensibilidade indica a capacidade do modelo em identificar corretamente as vozes que pertencem a uma determinada classe, por exemplo, aquelas realmente difíceis de julgar. Altos valores de sensibilidade significam que poucos casos difíceis foram classificados erroneamente como pertencentes a outras categorias. Já a especificidade revela a habilidade do modelo em reconhecer corretamente os casos que não pertencem a uma determinada classe, reduzindo a ocorrência de falsos positivos — por exemplo, ao evitar classificar como “difícil” uma voz que, de fato, não o é.

O valor preditivo positivo (VPP), por sua vez, corresponde à proporção de vozes classificadas como pertencentes a uma categoria (ex.: “difícil”) que realmente pertencem a essa categoria, sendo essencial para a confiabilidade

do modelo em suas decisões positivas. Complementarmente, o valor preditivo negativo (VPN) representa a proporção de vozes classificadas como “não pertencentes” àquela classe que, de fato, não pertencem, sendo relevante para avaliar a confiança nas predições negativas.

O F1-score combina, de forma equilibrada, sensibilidade e VPP, oferecendo uma visão mais robusta do desempenho do modelo, especialmente útil em situações com desequilíbrio entre as classes. Já a acurácia balanceada é calculada como a média entre a sensibilidade e a especificidade, sendo particularmente adequada quando há assimetrias no número de amostras por classe. A acurácia geral, por outro lado, expressa a proporção total de acertos do modelo, sem considerar o equilíbrio entre as classes, podendo superestimar o desempenho quando há predominância de uma classe.

Por fim, o índice Kappa ponderado mede o grau de concordância entre os rótulos previstos pelo modelo e os rótulos reais, descontando os acertos que ocorreriam por acaso. Quando ponderado (por exemplo, com pesos quadráticos), esse índice é ainda mais informativo em tarefas com classes ordenadas, como neste estudo, que considera três níveis crescentes de dificuldade.

4 RESULTADOS

Inicialmente, foi realizada uma etapa de agrupamento dos dados com o objetivo de explorar técnicas de AM não supervisionado para classificar automaticamente as vozes em clusters segundo o grau de dificuldade de JPA de cada parâmetro (GG, GR, GS e GT).

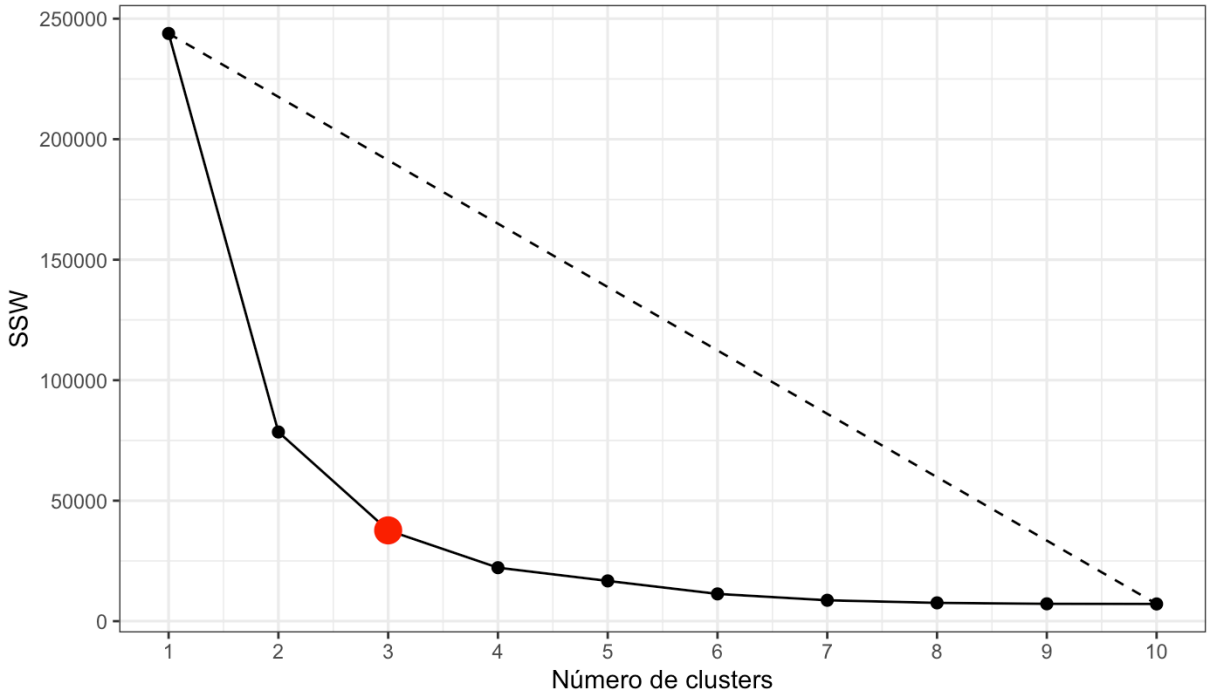
O agrupamento para rotular o nível de dificuldade foi aplicado considerando duas variáveis principais: o CV da EAV fuzzy e a variável de Concordância EAV Fuzzy. Três modelos de agrupamento foram comparados: K-means, Método hierárquico e Fuzzy C-means. Os resultados indicaram que o método K-means obteve o maior valor de Calinski-Harabasz para todos os parâmetros (Quadro 5). Além disso, a análise do número ideal de clusters também foi realizada por meio do método do cotovelo (Gráfico 1). Observa-se uma redução acentuada na soma total dos quadrados intragrupos (SSW) até o ponto correspondente a três clusters, indicado pelo ponto de inflexão marcado em vermelho. A partir desse ponto, o acréscimo de novos agrupamentos resulta em uma diminuição mínima do WSS, sugerindo que a divisão dos dados em três clusters oferece o melhor equilíbrio entre simplicidade do modelo e homogeneidade interna dos grupos.

Quadro 5: Índice interno de avaliação dos modelos de agrupamento para o JPA.

Parâmetros da voz	Método	nº de clusters	Calinski-Harabasz
GG	K-means	3	346,78
	Hierárquico	4	326,16
	Fuzzy C-means	3	345,82
GR	K-means	3	238,88
	Hierárquico	3	227,97
	Fuzzy C-means	3	140,08
GS	K-means	3	294,37
	Hierárquico	3	261,72
	Fuzzy C-means	3	231,72
GT	K-means	3	231,17
	Hierárquico	3	230,45
	Fuzzy C-means	3	183,12

Fonte: Elaborado pelo próprio autor, 2025.

Gráfico 1: Método do cotovelo para identificação do número ideal de clusters do grau de dificuldade do JPA segundo o modelo K-means.



Legenda: SSW - Soma total dos quadrados intragrupos.

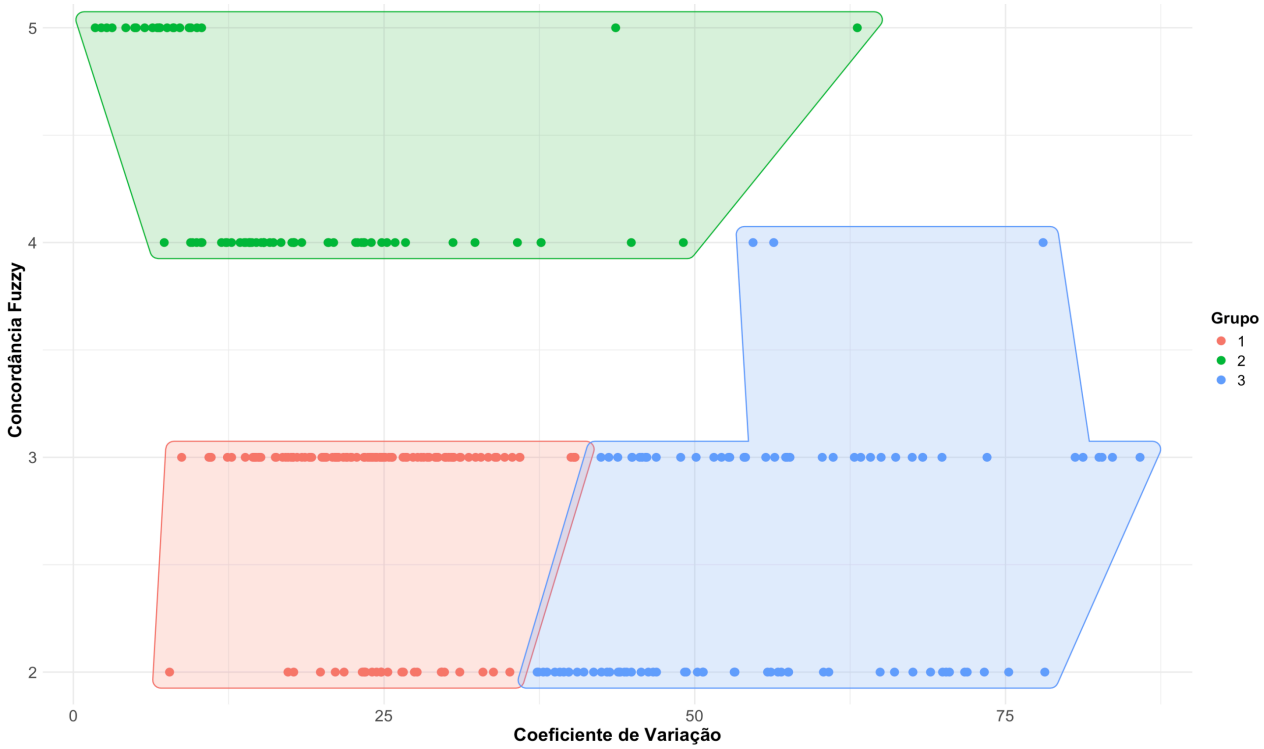
Fonte: Elaborado pelo próprio autor, 2025.

Posteriormente, foram gerados gráficos para a visualização dos clusters formados por cada modelo. Cada gráfico apresenta as amostras de voz distribuídas em função das duas variáveis de agrupamento, com os grupos destacados por cores distintas e contornos poligonais que indicam a forma e a extensão de cada cluster. A análise visual dos agrupamentos permite verificar não apenas a separação entre os grupos, mas também padrões estruturais nas vozes julgadas como mais fáceis ou mais difíceis de avaliar. Também fornece informações sobre a forma, o tamanho e a densidade dos agrupamentos, indicando se são compactos, alongados ou irregulares. Essa visualização auxilia na validação do agrupamento e reforça a coerência do método K-means como alternativa adequada para rotular os dados conforme a dificuldade do JPA.

Observa-se que o modelo K-means gerou três clusters relativamente bem definidos e com separações claras em relação ao CV e concordância dos

juízes, apresentando clusters mais compactos e distribuídos ao longo do eixo do coeficiente de variação (Gráfico 2, 3, 4 e 5).

Gráfico 2: Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GG.



Fonte: Elaborado pelo próprio autor, 2025.

Para o parâmetro GG, os agrupamentos formados foram interpretados em termos do nível de dificuldade do julgamento perceptivo-auditivo (JPA). O grupo 2 foi rotulado como "Fácil", por apresentar baixo coeficiente de variação (CV) e alta concordância fuzzy, indicando menor divergência entre os avaliadores. O grupo 3 foi rotulado como "Difícil", por apresentar alto CV e baixa concordância fuzzy, refletindo maior variabilidade nos julgamentos. Já o grupo 1 foi classificado como "Médio", por apresentar baixo CV, mas ainda assim baixa concordância fuzzy, sugerindo um nível intermediário de dificuldade no JPA. O Quadro 6 apresenta a frequência absoluta e o percentual de vozes atribuídas a cada nível de dificuldade.

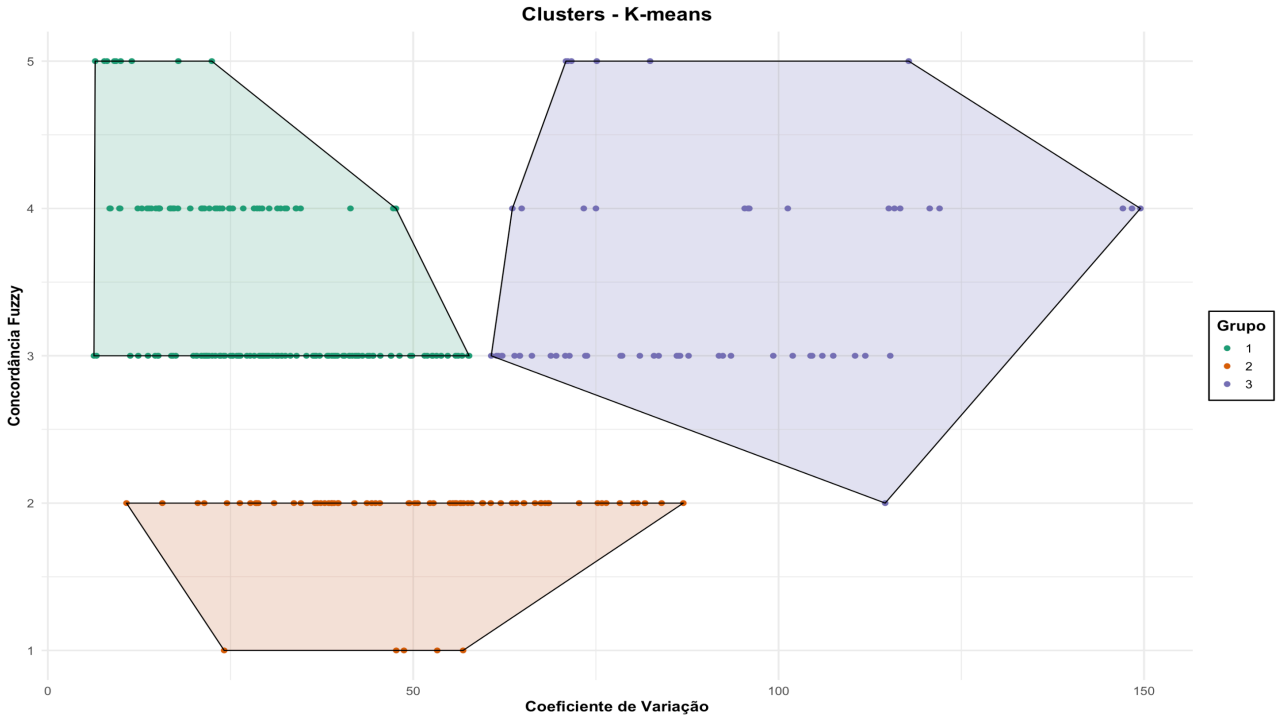
Quadro 6: Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GG.

Cluster	Frequência	Percentual
Fácil	72	24,4%
Médio	128	43,4%
Difícil	95	32,2%
Total	295	100%

Fonte: Elaborado pelo próprio autor, 2025.

Todos os modelos de agrupamento aplicados para os parâmetros de GR, GS e GT resultaram na formação de três clusters distintos, o que indica uma estrutura consistente nos dados quanto ao número de grupos. No entanto, optou-se por adotar o modelo K-means como principal referência, pois, além de apresentar os melhores resultados nos índices de validação interna, como o índice de Calinski-Harabasz, também demonstrou a melhor separação visual entre os grupos nos gráficos de dispersão com contornos (Gráficos 3, 4 e 5).

Gráfico 3: Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GR.



Fonte: Elaborado pelo próprio autor, 2025.

Para o parâmetro GR, o grupo 1 foi rotulado como “Fácil” – baixo CV e alta concordância Fuzzy, o grupo 2 foi rotulado como “Difícil” – alto CV e baixa concordância Fuzzy, e o grupo 3 foi rotulado como “Médio” – alto CV e alto concordância Fuzzy. O Quadro 7 apresenta a frequência absoluta e o percentual de vozes em cada grupo.

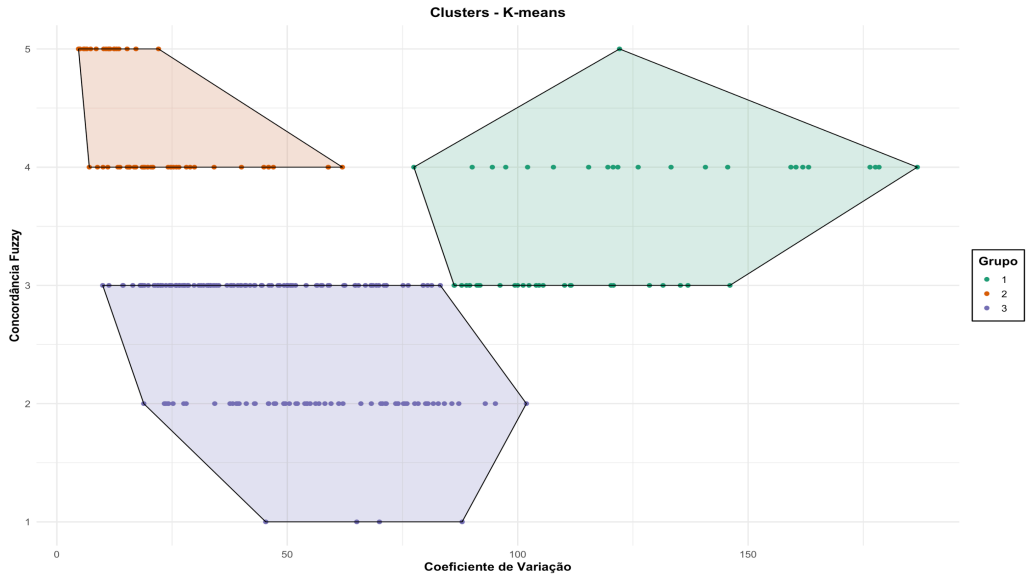
Quadro 7: Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GR.

Cluster	Frequência	Percentual
Fácil	163	55,3%
Médio	59	20%
Difícil	73	24,7%
Total	295	100%

Fonte: Elaborado pelo próprio autor, 2025.

Para o parâmetro GS, o grupo 2 foi rotulado como “Fácil” – baixo CV e alta concordância Fuzzy, o grupo 3 foi rotulado como “Difícil” – alto CV e baixa concordância Fuzzy, e o grupo 1 foi rotulado como “Médio” – alto CV e alto concordância Fuzzy. O Quadro 8 apresenta a frequência absoluta e o percentual de vozes em cada grupo.

Gráfico 4: Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GS.



Fonte: Elaborado pelo próprio autor, 2025.

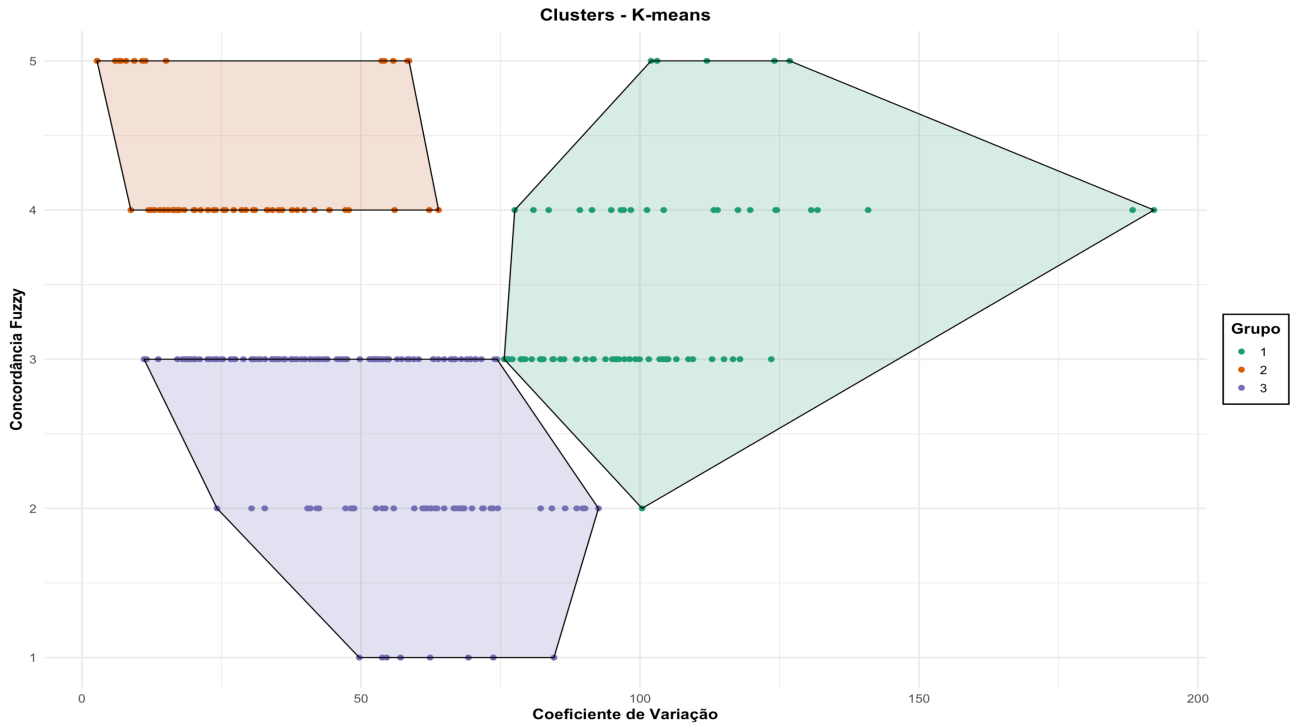
Quadro 8: Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GS.

Cluster	Frequência	Percentual
Fácil	183	62%
Médio	62	21%
Difícil	50	17%
Total	295	100%

Fonte: Elaborado pelo próprio autor, 2025.

Para o parâmetro GT, o grupo 2 foi rotulado como “Fácil” – baixo CV e alta concordância Fuzzy, o grupo 3 foi rotulado como “Difícil” – alto CV e baixa concordância Fuzzy, e o grupo 1 foi rotulado como “Médio” – alto CV e alto concordância Fuzzy. O Quadro 9 apresenta a frequência absoluta e o percentual de vozes em cada grupo.

Gráfico 5: Visualização dos clusters gerados pelo método K-means para o CV do JPA e concordância fuzzy do GT.



Fonte: Elaborado pelo próprio autor, 2025.

Quadro 9: Frequência e percentual de vozes classificadas segundo os clusters do modelo K-means para o CV do JPA e concordância fuzzy do GT.

Cluster	Frequência	Percentual
Fácil	60	20,4%
Médio	75	25,4%
Difícil	160	54,2%
Total	295	100%

Fonte: Elaborado pelo próprio autor, 2025.

Após a etapa de agrupamento, foram ajustados modelos supervisionados de AM com o objetivo de classificar o nível de dificuldade do JPA de cada parâmetro (GG, GR, GS e GT), a partir de suas características acústicas.

A Tabela 1 contêm os desempenhos médios de acurácia para os 14 modelos testados na validação cruzada repetida para o GG. Os melhores desempenhos foram observados nos modelos Discriminante Regularizada, Discriminante Flexível e Naive Bayes, que também apresentaram baixos erros padrão, indicando estabilidade ao longo das repetições (Tabela 1).

Tabela 1: Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GG no conjunto de treinamento.

Modelo	Acurácia Média	Erro Padrão
Discriminante Regularizada (RDA)	0,6409	0,0045
Discriminante Flexível (FDA)	0,6385	0,0042
Naive Bayes	0,6328	0,0046
SVM Linear	0,6311	0,0043
SVM Polinomial	0,6310	0,0043
Regressão Logística Multinomial	0,6301	0,0042
Discriminante Quadrática	0,6239	0,0044
Rede Neural MLP	0,6218	0,0040
Discriminante Linear	0,6217	0,0045
Floresta Aleatória	0,6155	0,0045
K-Vizinhos Mais Próximos	0,6020	0,0039
SVM com Kernel RBF	0,5966	0,0038

Árvore de Decisão	0,5902	0,0049
Árvore Boosting (XGBoost)	0,5873	0,0046

Fonte: Elaborado pelo próprio autor, 2025.

Para o GR, os quatro modelos com acurácia regular foram o Naive Bayes (0,6319), a Análise Discriminante Quadrática (0,6246), a Análise Discriminante Regularizada (0,6185), e Regressão Logística Multinomial (0,6146). Esses modelos também apresentaram baixos erros padrão, indicando estabilidade nos resultados ao longo das repetições (Tabela 2).

Tabela 2: Desempenho dos modelos AM supervisionados para classificação da dificuldade do JPA do GR no conjunto de treinamento.

Modelo	Acurácia Média	Erro Padrão
Naive Bayes	0,6319	0,0052
Discriminante Quadrática (QDA)	0,6246	0,0048
Discriminante Regularizada (RDA)	0,6185	0,0051
Regressão Logística Multinomial	0,6146	0,0051
SVM Linear	0,6058	0,0042
SVM Polinomial	0,6022	0,0047
Gradient Boosting (XGBoost)	0,6018	0,0047
Rede Neural MLP	0,5960	0,0048
Floresta Aleatória (Random Forest)	0,5922	0,0045
SVM com Kernel RBF	0,5889	0,0046
Discriminante Flexível (FDA)	0,5879	0,0044
K-Vizinhos Mais Próximos (KNN)	0,5835	0,0045
Árvore de Decisão	0,5587	0,0044
Discriminante Linear (LDA)	0,5252	0,0051

Fonte: Elaborado pelo próprio autor, 2025.

Para o GS, observou-se que os modelos com acurácia regular foram a Análise Discriminante Regularizada (RDA) (0,6798), seguida pela Rede Neural MLP (0,6759) e pela Regressão Logística Multinomial (0,6756). Os baixos erros padrão, sugerem estabilidade nos resultados ao longo das repetições (Tabela 3).

Tabela 3: Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GS no conjunto de treinamento.

Modelo	Acurácia Média	Erro Padrão
Análise Discriminante Regularizada (RDA)	0,6798	0,0047
Rede Neural MLP	0,6759	0,0055
Regressão Logística Multinomial	0,6756	0,0043
SVM Linear	0,6713	0,0042
Análise Discriminante Quadrática (QDA)	0,6680	0,0041
Análise Discriminante Linear (LDA)	0,6431	0,0045
Análise Discriminante Flexível (FDA)	0,6423	0,0049
SVM Polinomial	0,6411	0,0042
Árvore de Decisão	0,6262	0,0058
Boosting (XGBoost)	0,6248	0,0049
Floresta Aleatória (Random Forest)	0,6177	0,0048
K-Vizinhos Mais Próximos (KNN)	0,6063	0,0036
Naive Bayes	0,6010	0,0042
SVM com Kernel RBF	0,5381	0,0023

Fonte: Elaborado pelo próprio autor, 2025.

Em relação ao GT, observou-se que os modelos com melhor acurácia foram: Análise Discriminante Regularizada (RDA) (0,5939); Naive Bayes (0,5914); e Análise Discriminante Quadrática (QDA) (0,5855) (Tabela 4).

Tabela 4: Desempenho dos modelos de AM supervisionados para classificação da dificuldade do JPA do GT no conjunto de treinamento.

Modelo	Acurácia Média	Erro Padrão
Análise Discriminante Regularizada (RDA)	0,5939	0,0043
Naive Bayes	0,5914	0,0039
Análise Discriminante Quadrática (QDA)	0,5855	0,0044
Regressão Logística Multinomial	0,5809	0,0044
SVM Linear	0,5797	0,0044
SVM Polinomial	0,5739	0,0042
SVM com Kernel RBF	0,5725	0,0043
Floresta Aleatória (Random Forest)	0,5690	0,0033
Rede Neural MLP	0,5676	0,0049

Boosting (XGBoost)	0,5569	0,0032
Árvore de Decisão	0,5557	0,0034
Análise Discriminante Flexível (FDA)	0,5539	0,0041
K-Vizinhos Mais Próximos (KNN)	0,5513	0,0040
Análise Discriminante Linear (LDA)	0,4956	0,0043

Fonte: Elaborado pelo próprio autor, 2025.

A partir das análises conduzidas na etapa de validação cruzada, os modelos supervisionados passaram por uma avaliação final utilizando o conjunto de teste com 30% das vozes (n=89), estratificada pelo nível de dificuldade do JPA para manter o equilíbrio entre os grupos. Essa etapa teve como objetivo verificar a capacidade de generalização dos modelos, ou seja, sua performance ao classificar novas amostras não utilizadas no processo de treinamento.

Foram calculadas métricas de desempenho como acurácia, acurácia balanceada, sensibilidade, especificidade, F1-score, kappa e AUC ROC permitindo uma análise mais completa sobre a robustez dos modelos na predição do grau de dificuldade do JPA das vozes com base em medidas acústicas.

Os modelos com maior acurácia balanceada e área sob a curva ROC demonstraram maior capacidade de generalização. O índice Kappa indicou o grau de concordância real entre as predições dos modelos e os rótulos de dificuldade, ajustando os acertos esperados por acaso.

A Regressão Multinomial foi o modelo que apresentou melhor desempenho global, boa especificidade, aceitável acurácia balanceada, F1-score regular e Kappa moderado, mostrando-se mais eficaz na tarefa de classificar corretamente os diferentes níveis de dificuldade de JPA do GG (Tabela 5).

Tabela 5: Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GG no conjunto de teste.

Modelo	Acurácia	Acurácia Balanceada	F1-score	Especificidade	Kappa	AUC-ROC
Regressão Multinomial	0.6667	0.7227	0.6480	0.8142	0.4647	0.7961
Análise Discriminante	0.6222	0.6899	0.5987	0.7923	0.3949	0.7849

Linear (LDA)						
SVM Linear	0.6000	0.6579	0.5252	0.7755	0.3448	0.7929
SVM Polinomial	0.6000	0.6579	0.5252	0.7755	0.3448	0.7916
Rede Neural (MLP)	0.5889	0.6622	0.5640	0.7722	0.3370	0.7523
Análise Discriminante Flexível (FDA)	0.5778	0.6421	0.5079	0.7640	0.3094	0.7623
SVM RBF	0.5556	0.6134	0.4579	0.7450	0.2537	0.7739
Análise Discriminante Quadrática (QDA)	0.5556	0.6811	0.5579	0.7747	0.3347	0.7404
Análise Discriminante Regularizada (RDA)	0.5444	0.6372	0.5180	0.7541	0.2789	0.7592
Árvore de Decisão	0.5000	0.6107	0.4879	0.7367	0.2201	0.5901
K-Vizinhos Mais Próximos (KNN)	0.4889	0.5882	0.4542	0.7231	0.1813	0.6735
Floresta Aleatória (Random Forest)	0.4778	0.5778	0.4459	0.7138	0.1574	0.7040
Naive Bayes	0.4778	0.5974	0.4565	0.7257	0.1898	0.6886
Árvore de Boosting (XGBoost)	0.4667	0.5850	0.4580	0.7155	0.1612	0.6849

Fonte: Elaborado pelo próprio autor, 2025.

O modelo de Regressão Multinomial obteve uma baixa sensibilidade (0,4091) para o grupo "Fácil", aceitável para o grupo "Médio" (0,7949) e regular para o grupo "Difícil" (0,6897), indicando que o modelo foi mais eficiente em identificar corretamente as vozes com dificuldade intermediária. A especificidade, que representa a capacidade de reconhecer corretamente as vozes que não pertencem a um determinado grupo, foi excelente para os grupos "Fácil" (0,9559) e "Difícil" (0,9180). Os valores preditivos positivos também foram aceitáveis, sendo 0,7500 para "Fácil" e 0,8000 para "Difícil". O F1-score variou entre 0,5294 ("Fácil") e 0,7407 ("Difícil"), refletindo um equilíbrio adequado entre precisão e sensibilidade. A acurácia balanceada foi 0,6825 para "Fácil" e 0,6817 "Médio", e de 0,8038 para "Difícil", sugerindo um bom desempenho geral. O Kappa ponderado foi de 0,4849, o que indica um nível moderado de concordância entre as classificações do modelo e os rótulos

reais atribuídos às amostras (Tabela 6). O Quadro 10 apresenta a matriz de confusão entre o valor real e a predição do modelo de Regressão Multinomial.

Quadro 10: Matriz de confusão do modelo de Regressão Multinomial para os níveis de dificuldade do JPA do GG.

Predição \ Realidade	Fácil	Médio	Difícil
Fácil	9	3	0
Médio	13	31	9
Difícil	0	5	20
Total	22	39	29

Fonte: Elaborado pelo próprio autor, 2025.

Tabela 6: Métricas de desempenho do modelo de Regressão Multinomial baseado na matriz de confusão para os níveis de dificuldade do JPA do GG.

Métrica	Classe: Fácil	Classe: Médio	Classe: Difícil
Sensibilidade	0,4091	0,7949	0,6897
Especificidade	0,9559	0,5686	0,9180
VPP	0,7500	0,5849	0,8000
VPN	0,8333	0,7838	0,8615
F1-score	0,5294	0,6739	0,7407
Acurácia Balanceada	0,6825	0,6817	0,8038
Acurácia Geral	0,6667		
Kappa Ponderado	0,4849		

Legenda: VPP – Valor preditivo positivo; VPN – Valor preditivo negativo.

Fonte: Elaborado pelo próprio autor, 2025.

Para a predição do grau de dificuldade do JPA do GR, a partir do conjunto de teste, a Regressão Multinomial destacou-se com a maior AUC-ROC (0,7961), seguida pelos modelos LDA (0,7849), SVM Linear (0,7929) e SVM Polinomial (0,7916), demonstrando aceitável capacidade de generalização. Esses modelos também apresentaram valores satisfatórios para as outras métricas, sugerindo que são promissores para a tarefa de classificação do nível de dificuldade do GR com base em características acústicas da voz (Tabela 7).

Tabela 7: Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GR no conjunto de teste.

Modelo	Acurácia	Acurácia Balanceada	F1-score	Especificidade	Kappa	AUC-ROC
Regressão Multinomial	0.6667	0.7227	0.6480	0.8142	0.4647	0.7961
Análise Discriminante Linear (LDA)	0.6222	0.6899	0.5987	0.7923	0.3949	0.7849
SVM Linear	0.6000	0.6579	0.5252	0.7755	0.3448	0.7929
SVM Polinomial	0.6000	0.6579	0.5252	0.7755	0.3448	0.7916
Rede Neural (MLP)	0.5889	0.6622	0.5640	0.7722	0.3370	0.7523
Análise Discriminante Flexível (FDA)	0.5778	0.6421	0.5079	0.7640	0.3094	0.7623
SVM RBF	0.5556	0.6134	0.4579	0.7450	0.2537	0.7739
Análise Discriminante Quadrática (QDA)	0.5556	0.6811	0.5579	0.7747	0.3347	0.7404
Análise Discriminante Regularizada (RDA)	0.5444	0.6372	0.5180	0.7541	0.2789	0.7592
Árvore de Decisão	0.5000	0.6107	0.4879	0.7367	0.2201	0.5901
K-Vizinhos Mais Próximos (KNN)	0.4889	0.5882	0.4542	0.7231	0.1813	0.6735
Floresta Aleatória (Random Forest)	0.4778	0.5778	0.4459	0.7138	0.1574	0.7040
Naive Bayes	0.4778	0.5974	0.4565	0.7257	0.1898	0.6886
Árvore de Boosting (XGBoost)	0.4667	0.5850	0.4580	0.7155	0.1612	0.6849

Fonte: Elaborado pelo próprio autor, 2025.

Em relação ao modelo Regressão Multinomial aplicado para a classificação do grau de dificuldade do JPA do GR, a sensibilidade foi boa para o grupo “Fácil” (0,8163), enquanto o grupo “Difícil” apresentou sensibilidade reduzida (0,2727), indicando menor capacidade do modelo em identificar corretamente vozes mais difíceis. A especificidade, por outro lado, foi boa para “Médio” (0,8873) e “Difícil” (0,8656), sugerindo boa habilidade do modelo em reconhecer os casos que não pertencem a essas categorias. Os VPP variaram entre 0,4000 (“Difícil”) e 0,7273 (“Fácil”), e os VPN ficaram acima de 0,7353 em todas os grupos. O F1-score foi mais alto para o grupo “Fácil” (0,7692),

indicando um aceitável equilíbrio entre precisão e sensibilidade nessa categoria. A acurácia balanceada foi baixa para o grupo “Difícil” (0,5692) e aceitável para o grupo “Médio” (0,7492), a acurácia geral foi regular (0,6404) e o Kappa ponderado do modelo foi de 0,2469, representando um nível de concordância regular entre as predições do modelo e os rótulos reais (Tabela 8). O Quadro 11 contém a matriz de confusão entre o valor real e a predição do modelo de Regressão Multinomial para o GR.

Quadro 11: Matriz de confusão do modelo de Regressão Multinomial para os níveis de dificuldade do JPA do GR.

Predição \ Realidade	Fácil	Médio	Difícil
Fácil	40	2	13
Médio	5	11	3
Difícil	4	5	6
Total	49	18	22

Fonte: Elaborado pelo próprio autor, 2025.

Tabela 8: Métricas de desempenho do modelo de Regressão Multinomial baseado na matriz de confusão para os níveis de dificuldade do JPA do GR.

Métrica	Classe: Fácil	Classe: Médio	Classe: Difícil
Sensibilidade	0,8163	0,6111	0,2727
Especificidade	0,6250	0,8873	0,8656
VPP	0,7273	0,5789	0,4000
VPN	0,7353	0,9000	0,7837
F1-score	0,7692	0.5946	0,3243
Acurácia Balanceada	0,7207	0,7492	0,5692
Acurácia Geral	0,6404		
Kappa Ponderado	0,2469		

Legenda: VPP – Valor preditivo positivo; VPN – Valor preditivo negativo.

Fonte: Elaborado pelo próprio autor, 2025.

Na predição da dificuldade do JPA do GS, observou-se que diversos modelos apresentaram desempenhos satisfatórios no conjunto de teste. Os melhores resultados foram obtidos com os modelos MLP (Rede Neural Perceptron Multicamadas), Regressão Multinomial e SVM Linear, com acurácias regulares de 0,6854, 0,6966 e 0,7191, respectivamente. Além da acurácia, esses modelos também apresentaram valores aceitáveis de AUC-

ROC e F1-socre, o que indica equilíbrio entre sensibilidade e precisão na classificação (Tabela 9).

Tabela 9: Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GS no conjunto de teste.

Modelo	Acurácia	Acurácia Balanceada	F1-score	Especificidade	Kappa	AUC-ROC
Rede Neural (MLP)	0.6854	0.7028	0.6324	0.7870	0.4007	0.8358
Regressão Multinomial	0.6966	0.6772	0.6023	0.7711	0.3726	0.8227
SVM Linear	0.7191	0.6716	0.5703	0.7708	0.3850	0.8081
SVM Polinomial	0.7191	0.6525	0.5440	0.7602	0.3573	0.8073
Árvore de Decisão	0.6629	0.6519	0.5637	0.7621	0.3254	0.6781
Análise Discriminante Regularizada (RDA)	0.6404	0.6341	0.5143	0.7382	0.2664	0.7926
Análise Discriminante Flexível (FDA)	0.6854	0.6316	0.5444	0.7459	0.3004	0.8257
Árvore de Boosting (XGBoost)	0.6517	0.6301	0.5411	0.7427	0.2789	0.7140
Floresta Aleatória (Random Forest)	0.6742	0.6240	0.5349	0.7414	0.2827	0.7483
Análise Discriminante Quadrática (QDA)	0.3708	0.6051	0.3839	0.6967	0.1253	0.7523
Análise Discriminante Linear (LDA)	0.6180	0.5866	0.4775	0.7131	0.1844	0.7911
K-Vizinhos Mais Próximos (KNN)	0.6180	0.5784	0.4581	0.7083	0.1691	0.7000
Naive Bayes	0.6180	0.5538	0.3935	0.6982	0.1249	0.7132
SVM RBF	0.6180	0.5000	0.7639	0.6667	0.0000	0.7985

Fonte: Elaborado pelo próprio autor, 2025.

No modelo de classificação MLP para o grau de dificuldade do JPA do GS, observou-se uma acurácia geral regular de 0,6854 e um índice de Kappa ponderado regular de 0,4237, indicando boa concordância entre as predições e os rótulos reais. O grupo “Fácil” obteve índices de desempenho aceitáveis, com sensibilidade de 0,7818, F1-score de 0,7555 e acurácia balanceada de 0,6556. Já o grupo “Difícil” apresentou desempenho equilibrado, com F1-score regular de 0,6429 e especificidade excelente (0,9459). O grupo “Médio” foi a que

apresentou menor sensibilidade (0,4737), apesar de ter uma boa especificidade (0,8857) (Tabela 10). O Quadro 12 mostra a matriz de confusão entre os reais rótulos e a predição da dificuldade do JPA do GS.

Quadro 12: Matriz de confusão do modelo MLP para os níveis de dificuldade do JPA do GS.

Predição \ Realidade	Fácil	Médio	Difícil
Fácil	43	10	6
Médio	8	9	0
Difícil	4	0	9
Total	55	19	15

Fonte: Elaborado pelo próprio autor, 2025.

Tabela 10: Métricas de desempenho do modelo MLP baseado na matriz de confusão para os níveis de dificuldade do JPA do GS.

Métrica	Classe: Fácil	Classe: Médio	Classe: Difícil
Sensibilidade	0,7818	0,4737	0,6000
Especificidade	0,5294	0,8857	0,9459
VPP	0,7288	0,5294	0,6923
VPN	0,6000	0,8611	0,9211
F1-score	0,7555	0,5000	0,6429
Acurácia Balanceada	0,6556	0,6797	0,7730
Acurácia Geral	0.6854		
Kappa Ponderado	0,4237		

Legenda: VPP – Valor preditivo positivo; VPN – Valor preditivo negativo.

Fonte: Elaborado pelo próprio autor, 2025.

Em relação a predição da classificação do grau de dificuldade do JPA do GT o método Random Forest obteve acurácia (0,6517) e acurácia balanceada (0,6541) regulares, demonstrando desempenho superior na distinção dos grupos comparado aos outros modelos. Além disso, apresentou aceitável especificidade (0,7760) e área sob a curva ROC (0,7342), indicando capacidade robusta de discriminação (Tabela 11).

Tabela 11: Desempenho dos modelos de AM supervisionados para classificação da dificuldade de JPA do GT no conjunto de teste.

Modelo	Acurácia	Acurácia Balanceada	F1-score	Especificidade	Kappa	AUC-ROC
Floresta Aleatória (Random Forest)	0.6517	0.6541	0.5442	0.7760	0.3545	0.7342
K-Vizinhos Mais Próximos (KNN)	0.5955	0.6221	0.4767	0.7549	0.2766	0.6828
Naive Bayes	0.5506	0.6142	0.4733	0.7402	0.2342	0.6425
SVM Linear	0.5506	0.6031	0.4728	0.7365	0.2200	0.6626
Árvore de Boosting (XGBoost)	0.6067	0.6021	0.6207	0.7462	0.2537	0.6781
Análise Discriminante Flexível (FDA)	0.5618	0.6020	0.4714	0.7385	0.2233	0.6251
Árvore de Decisão	0.5955	0.5931	0.6148	0.7350	0.2278	0.6108
SVM RBF	0.5393	0.5789	0.4375	0.7219	0.1761	0.6757
Análise Discriminante Regularizada (RDA)	0.4494	0.5736	0.4010	0.7139	0.1380	0.6156
SVM Polinomial	0.4494	0.5508	0.3978	0.6935	0.0903	0.5726
Análise Discriminante Quadrática (QDA)	0.4157	0.5389	0.3831	0.6841	0.0649	0.5784
Regressão Multinomial	0.4045	0.5377	0.3730	0.6874	0.0648	0.5348
Rede Neural (MLP)	0.3933	0.5265	0.3735	0.6761	0.0376	0.5403
Análise Discriminante Linear (LDA)	0.2697	0.4245	0.2313	0.6214	-0.1345	0.3940

Fonte: Elaborado pelo próprio autor, 2025.

O modelo Random Forest para a classificação do grau de dificuldade do JPA do GT apresentou desempenho heterogêneo entre os grupos. O grupo "Difícil" apresentou os melhores resultados, com boa sensibilidade (0,8958), indicando que o modelo conseguiu identificar corretamente a maioria dos casos verdadeiramente difíceis. O F1-score de 0,7692 reforça esse bom desempenho, apesar da baixa especificidade (0,4634), o que sugere que houve classificações incorretas de outros grupos como "difícil" (Tabela 12).

O grupo "Médio" teve desempenho intermediário, com baixa sensibilidade (0,4783) e boa especificidade (0,8788), o que se reflete num F1-

score baixo (0,5238). Já o grupo "Fácil" foi o mais desafiador para o modelo, com baixa sensibilidade (0,2222) — o que indica que muitos casos fáceis foram classificados incorretamente —, apesar do bom valor preditivo positivo (0,8000), sugerindo que, quando o modelo classifica como fácil, geralmente acerta (Tabela 12).

Globalmente, a acurácia balanceada para todas os grupos girou em torno de 0,6041 a 0,6796, e o índice kappa ponderado (0,3883) aponta para uma concordância regular entre as predições do modelo e os julgamentos reais, considerando a gravidade dos erros (Tabela 12). O Quadro 13 representa a matriz de confusão do modelo Random Forest para o GT.

Quadro 13: Matriz de confusão do modelo Random Forest para os níveis de dificuldade do JPA do GT.

Predição \ Realidade	Fácil	Médio	Difícil
Fácil	4	0	1
Médio	4	11	4
Difícil	10	12	43
Total	18	23	48

Fonte: Elaborado pelo próprio autor, 2025.

Tabela 12: Métricas de desempenho do modelo de Random Forest baseado na matriz de confusão para os níveis de dificuldade do JPA do GT.

Métrica	Classe: Fácil	Classe: Médio	Classe: Difícil
Sensibilidade	0,2222	0,4783	0,8958
Especificidade	0,9959	0,8788	0,4634
VPP	0,8000	0,5789	0,6615
VPN	0,8333	0,8286	0,7917
F1-score	0,3478	0,5238	0,7611
Acurácia Balanceada	0,6041	0,6785	0,6796
Acurácia Geral	0,6517		
Kappa Ponderado	0,3883		

Legenda: VPP – Valor preditivo positivo; VPN – Valor preditivo negativo.

Fonte: Elaborado pelo próprio autor, 2025.

5 DISCUSSÃO

A discussão será dividida em três partes, a saber: 1) Desempenho dos modelos de AM para agrupamento; 2) Desempenho dos modelos de AM para classificação e 3) Implicações para o TPA.

5.1 Desempenho dos modelos AM para agrupamento

A análise comparativa entre os modelos de agrupamento — K-means, Método Hierárquico e Fuzzy C-means — revelou que o K-means foi consistentemente superior em termos do índice de Calinski-Harabasz para todos os parâmetros vocais. Esse índice indicou que os agrupamentos gerados pelo K-means apresentaram maior coerência interna e melhor distinção entre os grupos.

Apesar do algoritmo K-means funcionar melhor quando os dados apresentam grupos de forma simétrica e bem definidos, o que nem sempre é o caso em contextos clínicos complexos como o JPA, os melhores valores obtidos pelo índice Calinski-Harabasz indicam que, neste estudo, os dados se ajustaram bem a esse modelo. Isso pode ser explicado pelo fato de que as variáveis utilizadas no agrupamento (coeficiente de variação e concordância fuzzy) conseguiram capturar, de maneira estável, a variabilidade perceptiva e o nível de concordância entre os juízes. Como resultado, foi possível formar agrupamentos mais nítidos entre os diferentes níveis de dificuldade de julgamento das vozes.

O desempenho inferior do Fuzzy C-means pode estar relacionado à sobreposição entre os grupos, característica esperada em dados subjetivos, mas que torna o agrupamento mais difuso e menos útil para fins de classificação posterior. Já o método hierárquico, embora apresente valores próximos aos do K-means, geralmente é mais sensível a outliers e não otimiza diretamente a separação entre grupos como o K-means. Portanto, a escolha do K-means não apenas se justifica pelo índice mais altos, mas também pela sua aplicabilidade prática para gerar rótulos bem definidos, essenciais para a etapa de modelagem supervisionada.

Além dos bons resultados observados neste estudo, a escolha do K-means também encontra respaldo na literatura. O estudo de Almeida (2022),

ao empregar o K-means para classificar a presença e ausência de lesão vocal, demonstrou que medidas acústicas, quando utilizadas isoladamente, resultam em agrupamentos com melhor qualidade do que a EAV ou combinações de múltiplas fontes de informação.

No estudo de Patil e Bhalke, (2004), o classificador híbrido K-means e o k-*nearest neighbor* (KNN) foi aplicado para distinguir entre vozes patológicas e normais. O K-means agrupou os padrões de energia espectral, criando centróides representativos de cada classe. Em seguida, um classificador KNN foi utilizado para comparar novos exemplos com os agrupamentos existentes. Os autores relataram resultados promissores, demonstrando que essa abordagem pode ser aplicada como uma ferramenta de diagnóstico não invasivo. Isso reforça a robustez do algoritmo para identificar padrões estruturados em dados vocais, especialmente quando o objetivo é gerar rótulos consistentes e aplicáveis em modelos supervisionados posteriores.

5.2 Desempenho dos AM para classificação

Na etapa de classificação do grau de dificuldade do JPA, o modelo de Regressão Multinomial obteve o melhor desempenho para a classificação da dificuldade do JPA do GG e do GR. Esse modelo destaca-se por ser um modelo estatístico interpretável e eficiente em cenários de classificação com mais de duas categorias. Sua principal vantagem é capacidade de modelar diretamente a probabilidade de pertencimento a cada grupo com base em variáveis preditoras, sem exigir suposições rígidas sobre a distribuição dos dados (Sainani, 2021).

Além dos aspectos destacados, a Regressão Multinomial tem sido amplamente utilizada em tarefas de classificação, como no reconhecimento de emoções na fala, onde demonstrou eficiência satisfatória mesmo com um modelo relativamente simples (Poovammal *et al*., 2016). Essa abordagem permite modelar diretamente a probabilidade de pertencimento a diversos grupos, o que a torna especialmente adequada para problemas com múltiplas categorias de saída, como no caso do nível de dificuldade do JPA do GG e do GR.

Na classificação dos níveis de dificuldade do JPA do GS, o modelo MLP demonstrou o melhor desempenho. Sua capacidade de lidar com padrões

complexos a partir de dados multivariados justifica sua eficácia na discriminação entre os diferentes níveis de dificuldade perceptiva. Além disso, trata-se de um modelo com ampla aplicabilidade em tarefas de classificação envolvendo dados acústicos, especialmente quando se busca identificar características sutis baseadas em julgamentos humanos (Sabir *et al*., 2017; Leite; Moraes; Lopes, 2022).

O estudo de Liu *et al.,* (2025) reforça essa perspectiva ao demonstrar que o modelo MLP foi capaz de classificar automaticamente a tensão perceptiva em vozes cantadas com alta acurácia, a partir de um conjunto selecionado de características acústicas. A comparação entre diferentes conjuntos de dados e classificadores revelou que o MLP não apenas se destacou frente a outros modelos, mas também apresentou robustez na análise de variações vocais relacionadas ao uso vocal intenso. Esses achados sustentam o uso do MLP como uma abordagem eficaz em contextos clínicos, contribuindo para a automatização de tarefas classificatórias tradicionalmente realizadas por especialistas.

Para a classificação da dificuldade do JPA do GT, o modelo Random Forest apresentou o melhor desempenho. Esse modelo é reconhecido por sua robustez em cenários com variáveis preditoras heterogêneas e relações não lineares, além de lidar bem com dados ruidosos e desequilibrados — características frequentemente presentes em tarefas perceptivo-auditivas (Denisko; Hoffman, 2018; Brieuc; Waters; Drinan; Naish, 2018).

Uma pesquisa recente sobre crianças com TEA falantes do mandarim (Guo *et al*., 2022) utilizou o modelo Random Forest para identificar parâmetros acústicos capazes de diferenciar vozes atípicas, alcançando acurácia de 78,5% e destacando o papel de variáveis como *shimmer* e *jitter* para a classificação. Esses achados reforçam o potencial do Random Forest como uma ferramenta confiável e aplicável à construção de modelos automatizados para análise de voz, inclusive em contextos clínicos e diagnósticos.

Para o parâmetro GG, o modelo de Regressão Multinomial apresentou o melhor desempenho na classificação das vozes pertencentes ao grupo “Difícil”, alcançando os maiores valores de sensibilidade, valor preditivo positivo, valor preditivo negativo, F1-score e acurácia balanceada. Isso indica que as vozes com maior discordância entre os juízes foram corretamente identificadas com

maior frequência. Já para do parâmetro GR, o modelo de Regressão Multinomial apresentou melhor desempenho para o grupo “Fácil”, evidenciado pelos melhores valores de sensibilidade, VPP, F1-score e acurácia balanceada.

Em relação GS, o modelo MLP apresentou desempenho superior na identificação das vozes classificadas como “Fácil”, destacando-se pelas maiores taxas de sensibilidade, F1-score e acurácia balanceada. Esse resultado indica que o modelo é particularmente eficiente na detecção de vozes com menor grau de complexidade perceptiva.

Um possível fator que contribui para o desempenho dos modelos supervisionados do GG, GR e GS é a natureza desses parâmetros vocais, frequentemente associado a altos índices de concordância entre juízes no JPA, o que proporciona uma referência mais estável e reduz a variabilidade entre julgamentos (Brinca et al., 2015; Iawrsson et al., 2018; Yamasaki & Gama, 2019). Além disso, estudos demonstram a correlação entre o JPA e a análise acústica da voz, o que pode ter contribuído para o melhor desempenho dos modelos supervisionados nos parâmetros GG, GR e GS (Omori et al., 1998; Lopes., *et al*., 2019; Lopes *et al*, 2020). Essa relação sugere que, ao menos para esses parâmetros, as características acústicas captadas objetivamente refletem, em certa medida, as percepções subjetivas dos juízes. Dessa forma, a convergência entre os dados acústicos e o JPA pode ter favorecido a aprendizagem dos modelos, facilitando a identificação de padrões consistentes e melhorando sua capacidade de classificação.

Além disso, na etapa de agrupamento, a distribuição desigual das vozes entre os grupos pode ter influenciado os resultados, já que os grupos “Fácil” dos modelos para o GS e GR foi o mais representativo da amostra (183 vozes, 62% e 163 vozes 55,3, respectivamente). Essa predominância numérica pode ter fornecido mais exemplos consistentes para o treinamento do modelo, favorecendo sua capacidade de generalização para essa classe específica.

Por outro lado, a acurácia geral e o valor do kappa ponderado indicam que, apesar do desempenho satisfatório na identificação de vozes com menor complexidade perceptiva, o modelo de regressão Multinomial para o GR apresentou limitações na classificação de vozes com maior grau de dificuldade, sugerindo a necessidade de estratégias adicionais para aprimorar sua capacidade preditiva nesses casos.

O modelo Random Forest, para o GT, apresentou desempenho heterogêneo entre os grupos, com resultados melhores para o grupo “Difícil”, indicando que o modelo conseguiu identificar corretamente a maioria dos casos verdadeiramente difíceis. Em contrapartida, o grupo “Fácil” apresentou o pior desempenho, com baixa sensibilidade, o que revela que muitas vozes com menor dificuldade perceptiva foram classificadas de forma incorreta. Esse resultado pode ser parcialmente explicado pelo fato de que o grupo “Fácil” foi o menor entre os três, com apenas 60 vozes (20,4%), o que pode ter limitado a capacidade do modelo em aprender e distinguir adequadamente suas características durante a etapa de treinamento.

Além disso, a própria natureza do parâmetro GT pode ter contribuído para essa limitação, já que sua definição auditiva ainda representa um desafio para os juízes. Estudos apontam que, ao tentar descrever o grau de tensão vocal, os juízes frequentemente recorrem a interpretações baseadas em aspectos anatômico-fisiológicos ou confundem os conceitos de esforço (sensação do falante) e tensão (atividade muscular), o que pode comprometer a padronização do julgamento (Hunter *et al*, 2020).

De forma geral, o índice kappa ponderado revelou uma concordância apenas regular entre as predições do modelo em relação ao GT e os julgamentos realizados por especialistas, sugerindo que, embora o desempenho do modelo seja promissor, há espaço para melhorias — especialmente na detecção mais precisa de vozes com menor complexidade perceptiva.

Embora os modelos utilizados neste estudo não permitam inferir sobre a importância individual de medidas acústicas específicas na formação dos agrupamentos, eles se mostraram suficientes para garantir que as vozes foram classificadas de forma consistente. A coerência observada entre os grupos gerados e as características perceptivo-auditivas valida a abordagem proposta e atende adequadamente aos objetivos deste trabalho, que buscou desenvolver um método objetivo para estimar o nível de dificuldade de estímulos vocais, sem a pretensão de identificar preditores acústicos individuais.

5.3 Implicações para o TPA

Ao propor uma abordagem baseada em algoritmos de AM para identificar e classificar o grau de dificuldade do JPA, este estudo contribui para a objetivação de um processo tradicionalmente subjetivo e suscetível à variabilidade entre juízes. Isso permite maior padronização e confiabilidade na seleção de estímulos vocais utilizados em treinamentos perceptivos. A possibilidade de utilizar classificadores treinados a partir de parâmetros acústicos reforça a integração entre a análise perceptiva e instrumental, valorizando o uso de evidências objetivas no raciocínio clínico.

Dessa forma, os resultados desta tese abrem caminho para o aprimoramento de ST, como o Ouvindo Vozes, baseados em vozes rotuladas segundo o grau de dificuldade, promovendo um ensino mais eficiente e personalizado para estudantes e profissionais da Fonoaudiologia. Ao incorporar vozes com diferentes níveis de dificuldade, esses simuladores poderão fortalecer habilidades auditivas fundamentais para o exercício clínico qualificado, promovendo maior acurácia na identificação de desvios vocais e, consequentemente, melhorando a qualidade do cuidado oferecido aos pacientes.

6 LIMITAÇÕES E PERSPECTIVAS FUTURAS

Esta pesquisa apresenta contribuições significativas ao propor um modelo para estimar a dificuldade do JPA de vozes a partir de medidas acústicas e técnicas de AM. No entanto, algumas limitações devem ser reconhecidas. A assimetria na distribuição entre os grupos “Fácil”, “Médio” e “Difícil” pode ter influenciado negativamente o desempenho dos modelos de classificação, principalmente na discriminação de classes minoritárias, como por exemplo, o grupo “Fácil” no GT. Essa desproporção pode ter impactado a sensibilidade e a acurácia balanceada, reduzindo a capacidade dos modelos de aprender padrões robustos em todas as classes.

Além disso, os dados utilizados pertencem a um banco específico, com vozes previamente julgadas por um conjunto restrito de juízes. Essa limitação pode comprometer a generalização dos resultados para outras populações ou contextos clínicos distintos, o que impede uma avaliação mais abrangente da robustez dos modelos.

Outra limitação está relacionada ao fato de que os modelos supervisionados não consideraram a importância relativa de cada medida acústica na predição da dificuldade do JPA. Embora o desempenho global tenha sido satisfatório, compreender quais variáveis contribuem mais significativamente para a classificação poderia auxiliar na interpretação clínica e na otimização dos modelos futuros.

Por fim, o uso de três clusters pré-definidos para rotular os dados, embora alinhado aos objetivos práticos da pesquisa, impôs uma estrutura fixa à classificação da dificuldade. Isso pode ter limitado a detecção de nuances intermediárias, especialmente em um fenômeno subjetivo e contínuo como o JPA.

Como perspectivas futuras, destaca-se a importância de ampliar o banco de vozes utilizado, com maior variedade de amostras e juízes, o que permitirá maior generalização dos modelos desenvolvidos. Também se recomenda a aplicação dos modelos a conjuntos de dados externos e independentes, visando validar sua eficácia em contextos distintos.

Outra possibilidade relevante é a incorporação da TRI, com base nos rótulos de dificuldade aqui desenvolvidos, para investigar a severidade e a

capacidade discriminativa das vozes, o que pode beneficiar o desenvolvimento de um ST mais robusto.

Além disso, é desejável que estudos futuros explorem o papel das variáveis acústicas individuais na predição da dificuldade do JPA, avaliando a importância relativa de cada medida na classificação final. Isso não apenas aumentaria a interpretabilidade dos modelos, mas também permitiria identificar quais parâmetros acústicos são mais relevantes nessa tarefa.

Finalmente, integrar essas descobertas à construção de ST interativo e baseado em evidências poderá contribuir significativamente para a formação de novos juízes e para a padronização da prática clínica na Fonoaudiologia.

7 CONCLUSÃO

Modelos não supervisionados de AM permitiram agrupar as vozes em três níveis de dificuldade — fácil, médio e difícil — sendo o modelo K-means selecionado como o mais eficaz, com base no índice de Calinski-Harabasz e na separação visual entre os grupos formados. A partir desses rótulos, modelos supervisionados foram treinados para prever automaticamente o grau de dificuldade do JPA com base em medidas acústicas. Diferentes modelos apresentaram desempenhos superiores a depender do parâmetro vocal analisado: a Regressão Multinomial foi mais eficaz para os parâmetros GG e GR; o modelo MLP para o parâmetro GS; e o Random Forest para o GT. Esses achados reforçam o potencial da combinação entre técnicas não supervisionadas e supervisionadas para classificar o nível de dificuldade do JPA da voz, proporcionando uma abordagem objetiva, reprodutível e potencialmente aplicável em contextos clínicos e educacionais na Fonoaudiologia.

8 PUBLICAÇÕES

Relacionado ao tema anterior da tese, uma revisão de escopo sobre alterações auditivas em indivíduos com disfonia foi desenvolvida. Um artigo com os resultados parciais foi apresentado publicado nos anais ao 32º Congresso Brasileiro de Fonoaudiologia 2024, na categoria de resumos expandidos. Além disso, o artigo com a versão final da revisão, intitulado "*Occurrence of Auditory Impairments in Individuals With Dysphonia: A Scoping Review*", foi publicado no periódico internacional *Journal of Voice*.

Os resultados da presente tese estão sendo submetidos ao 33º Congresso Brasileiro de Fonoaudiologia (2025) para apresentação. Além disso, o artigo contendo os resultados completos encontra-se em fase final de ajustes para submissão ao *Journal of Voice*.

REFERÊNCIAS

ABREU, S. R. *et al*. VoxMore: artefato tecnológico para auxiliar a avaliação acústica da voz no processo ensino-aprendizagem e prática clínica. *CoDAS*, São Paulo, 2023. p. e20220166.

ABREU, S. R. *Índice do grau geral do desvio vocal : desenvolvimento, avaliação e validação de um modelo de suporte à decisão para os falantes do português brasileiro*. 2024. Tese (Doutorado em Modelos de Decisão e Saúde) – Departamento de Estatística, Universidade Federal da Paraíba, João Pessoa, 2024.

ALBON, C. *Machine learning with Python cookbook: practical solutions from preprocessing to deep learning*. Sebastopol: O'Reilly Media, Inc., 2018.

AL-DHIEF, F. T. *et al*. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*, v. 9, p. 77293–77306, 2021. Disponível em: https://doi.org/10.1109/ACCESS.2021.3082565.

ALMEIDA, W. F. *Aplicação do algoritmo k-means para detecção de padrões em dados vocais*. 2022. Trabalho de Conclusão de Curso (Licenciatura em Matemática) – Instituto Federal da Paraíba, Cjazeiras, 2022.

ALVES, J. N. *Influência da experiência do ouvinte e da tarefa de fala na avaliação perceptivo-auditiva da qualidade vocal*. 2019. Dissertação (Mestrado em Linguística) – Universidade Federal da Paraíba, João Pessoa. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/18733.

BISPO, N. O. *et al*. Repetição de estímulos âncoras e natureza das amostras vocais no julgamento perceptivo-auditivo realizado por estudantes de fonoaudiologia. *CoDAS*, v. 34, n. 4, p. e20210064, 2022. Disponível em: https://doi.org/10.1590/2317-1782/20212021064.

BRIEUC, M. S. O.; WATERS, C. D.; DRINAN, D. P.; NAISH, K. A. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Molecular Ecology Resources, [S.l.], v. 18, n. 4, p. 755–766, jul. 2018. DOI: 10.1111/1755-0998.12773.

BRINCA, L. *et al*. The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *Journal of Voice*, v. 29, n. 6, p. 776.e1–776.e14, 2015. DOI: 10.1016/j.jvoice.2015.01.007.

BRINK, H.; RICHARDS, J.; FETHEROLF, M. Real-world machine learning. New York: Simon and Schuster, 2016.

BROCKMANN-BAUSER, M.; DRINNAN, M. J. Routine acoustic voice analysis: time to think again? *Current Opinion in Otolaryngology & Head and Neck Surgery*, v. 19, n. 3, p. 165–170, 2011. Disponível em: http://dx.doi.org/10.1097/moo.0b013e32834575fe.

BUDER, E. H. Acoustic analysis of voice quality: a tabulation of algorithms 1902–1990. In: KENT, M. J.; BALL, R. D. *Voice Quality Measurement*. 1. ed. San Diego: Singular Publishing Group, 2000.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics, v. 3, n. 1, p. 1–27, 1974. DOI: 10.1080/03610927408827101.

CHAN, K. M.; YIU, E. M. The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research*, v. 45, n. 1, p. 111–126, 2002. DOI: 10.1044/1092-4388(2002/009).

CHARILAOU, P.; BATTAT, R. Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, v. 28, p. 605, 2022. https://doi.org/10.3748/wjg.v28.i5.605.

CONTRERAS-RUSTON, F. *et al*. Auditory-perceptual assessment of healthy and disordered voices using the Voice Deviation Scale. *Journal of Voice*, 2021.

COSTA, C. C. *et al*. Avaliação perceptivo-auditiva da voz: comparação de diferentes tarefas de fala na identificação de crianças com e sem lesões laríngeas. *CoDAS*, 2023. p. e20210198.

D'SOUZA, S.; PREMA, K. V.; BALAJI, S. Feature selection and modeling using statistical and machine learning methods. In: *2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER 2020)*. p. 18–22. DOI: 10.1109/DISCOVER50404.2020.9278093.

DEJONCKERE, P. H. *et al*. A basic protocol for functional assessment of voice pathology. *European Archives of Oto-Rhino-Laryngology*, v. 258, p. 77–82, 2001. DOI: https://doi.org/10.1007/s004050000299.

DENISKO, D; HOFFMAN, M. M. Classification and interaction in random forests. Proceedings of the National Academy of Sciences of the United States of America, v. 115, n. 8, p. 1690–1692, 20 fev. 2018. Disponível em: https://doi.org/10.1073/pnas.1800256115.

EADIE, T. L.; BAYLOR, C. R. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice*, v. 20, p. 527–544, 2006.

EADIE, T. L. *et al*. The effect of musical background on judgments of dysphonia. *Journal of Voice*, v. 24, n. 1, p. 93–101, 2010.

FLORENCIO, V. O. *et al*. Differences and reliability of linear and nonlinear acoustic measures as a function of vocal intensity in individuals with voice disorders. *Journal of Voice*, p. 1–3, jun. 2021. https://doi.org/10.1016/j.jvoice.2021.04.011.

FREITAS, R. *Metodologia científica: um guia prático para profissionais da saúde*. 1. ed. Petrolina: [s.n.], 2017.

GHIO, A. *et al*. Perceptual evaluation of dysphonic voices: can a training protocol lead to the development of perceptual categories? *Journal of Voice*, v. 29, n. 3, p. 304–311, 2015. DOI: 10.1016/j.jvoice.2014.07.006.

GUO, C. *et al*. Applying Random Forest classification to diagnose autism using acoustical voice-quality parameters during lexical tone production. *Biomedical Signal Processing and Control*, [S.l.], v. 76, p. 103811, 2022. Disponível em: https://doi.org/10.1016/j.bspc.2022.103811.

HEGDE, S. *et al*. A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, v. 33, p. 947.e11–947.e33, 2019.

HIRANO, M. *Clinical examination of voice*. New York: Springer-Verlag, 1981.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. New York: Wiley, 2000.

HUNTER, E. J. et al. Toward a consensus description of vocal effort, vocal load, vocal loading, and vocal fatigue. *Journal of Speech, Language, and Hearing Research*, v. 63, n. 2, p. 509-532, 2020. Disponível em: http://dx.doi.org/10.1044/2019_JSLHR-19-00057.

IAWRSSON, J. et al. Auditory-perceptual evaluation of dysphonia: a comparison between narrow and broad terminology systems. *Journal of Voice*, v. 32, n. 4, p. 428-436, 2018. Disponível em: http://dx.doi.org/10.1016/j.jvoice.2017.07.006.

JO, T. *Machine learning foundations: supervised, unsupervised, and advanced learning*. Cham: Springer International Publishing, 2021.

KANEKO, H. Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables. *Heliyon*, v. 7, e07356, 2021.

KEMPSTER, G. B. *et al*. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, v. 18, n. 2, p. 124–132, 2009.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977. https://pubmed.ncbi.nlm.nih.gov/843571/.

LATOSZEK, B. B. *et al*. A meta-analysis: acoustic measurement of roughness and breathiness. *Journal of Speech, Language, and Hearing Research*, v. 61, n. 2, p. 298–323, 2018.

LATOSZEK, B. B. *et al*. The Acoustic Breathiness Index (ABI): a multivariate acoustic model for breathiness. *Journal of Voice*, v. 31, n. 4, p. 511.e11–511.e27, 2017.

LEITE, D. R. A.; MORAES, R. M.; LOPES, L. W. Different performances of machine learning models to classify dysphonic and non-dysphonic voices. *Journal of Voice*, 2022. https://doi.org/10.1016/j.jvoice.2022.11.001.

LEITE, D. R. Desenvolvimento de um modelo de classificação da tipologia dos sinais vocais com base no *deep learning*. 2022. Tese (Doutorado em Modelos de Decisão e Saúde) – Departamento de Estatística, Universidade Federal da Paraíba, João Pessoa, 2022.

LIU, Yuanyuan et al. Automatic classification of strain in the singing voice using machine learning. *Journal of Voice*, [S.l.], v. 35, n. 6, p. 1-13, 2025. Disponível em: https://doi.org/10.1016/j.jvoice.2025.03.040.

LIMA-FILHO, L. M. A; LOPES, L. W; SILVA-FILHO, T. M. Integrated Vocal Deviation Index (IVDI): a machine learning model to classifier of the general grade of vocal deviation. Journal of Voice, [S.l.], 2024. Disponível em: https://doi.org/10.1016/j.jvoice.2024.11.002.

LOPES, L. W. *et al*. Vocal characteristics during child development: perceptual-auditory and acoustic data. *Folia Phoniatrica et Logopaedica*, v. 65, p. 143–147, 2013.

LOPES, L. W.; CAVALCANTE, D. P.; COSTA, P. O. Severity of voice disorders: integration of perceptual and acoustic data in dysphonic patients. *CoDAS*, v. 26, n. 5, p. 382–388, 2014.

LOPES, L. W. *et al*. Classificação espectrográfica do sinal vocal: relação com o diagnóstico laríngeo e a análise perceptivo-auditiva. *Audiology - Communication Research*, v. 25, 2020.

LOPES, L. W. *et al*. Classificação espectrográfica do sinal vocal: relação com o diagnóstico laríngeo e a análise perceptivo-auditiva. Audiology & Communication Research, São Paulo, v. 25, e2194, 2020. Disponível em: https://www.scielo.br/j/acr/a/TsswWgM8Q876pQxSSrcFbYp/.

LOPES, L. W. et al. Effectiveness of Recurrence Quantification Measures in Discriminating Subjects With and Without Voice Disorders. Journal Of Voice, [S.L.], v. 34, n. 2, p. 208-220, mar. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.jvoice.2018.09.004.

MALIK, S. *et al*. Data driven approach for eye disease classification with machine learning. *Applied Sciences*, v. 9, n. 14, p. 2789, 2019.

MARYN, Y. *et al*. Acoustic measurement of overall voice quality: a meta-analysis. *The Journal of the Acoustical Society of America*, v. 126, n. 5, p. 2619–2634, 2009.

MARYN, Y. *et al*. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *Journal of Voice*, v. 24, n. 5, p. 540–555, 2010.

MARTINS, P. C.; COUTO, T. E.; GAMA, A. C. C. Avaliação perceptivo-auditiva do grau de desvio vocal: correlação entre escala visual analógica e escala numérica. *CoDAS*, 2015. p. 279–284.

MELLEY, L. E.; SATALOFF, R. T. Beyond the buzzwords: artificial intelligence in laryngology. *Journal of Voice*, v. 36, p. 2–3, 2022.

OATES, J. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica*, v. 61, n. 1, p. 49–56, 2009.

OMORI, K. et al. Influence of size and etiology of glottal gap in glottic incompetence dysphonia. The Laryngoscope, v. 108, n. 4, p. 514-518, 1998.

PAIVA, M. A. A. *et al*. Auditory skills as a predictor of rater reliability in the evaluation of vocal quality. *Journal of Voice*, v. 35, n. 4, p. 559–569, 2021.

PAIVA, M. A. A. *Proposta de um simulador de treinamento para a avaliação perceptivo-auditiva da voz*. 2022. Dissertação (Mestrado em Modelos de Decisão e Saúde) – Departamento de Estatística, Universidade Federal da Paraíba, João Pessoa, 2022.

PAIVA, M. A. A.; MACHADO, L. S.; LOPES, L. W. Proposição de requisitos para o desenvolvimento de um simulador de treinamento para julgamento

perceptivo-auditivo da voz. *Codas*, [S.l.], v. 35, n. 6, p. 1-8, 2023. Disponível em: http://dx.doi.org/10.1590/2317-1782/20232022209pt.

PATIL, H. A.; BHALKE, D. G. K-means nearest neighbor classifier for voice pathology. In: IEEE INDICON 2004 - Annual Conference of the IEEE India Council. Proceedings [...]. Kharagpur: IEEE, 2004. p. 244–247. DOI: 10.1109/INDICO.2004.1497770.

PAZ, K. E. S. *Efetividade do treinamento de habilidades auditivas temporais, associado ao treinamento perceptivo-auditivo convencional no julgamento perceptivo-auditivo da voz*. 2024. Tese (Doutorado em Modelos de Decisão e Saúde) – Departamento de Estatística, Universidade Federal da Paraíba, João Pessoa, 2022.

PAZ, K. E. S. *et al*. Treinamento para análise perceptivo-auditiva da voz: revisão de escopo. *Audiology - Communication Research*, v. 28, p. 1–6, 2023.

PETERSON, E. A. *et al*. Toward validation of the cepstral spectral index of dysphonia (CSID) as an objective treatment outcomes measure. *Journal of Voice*, v. 27, n. 4, p. 401–410, 2013.

POOVAMMAL, E.; VERMA, S.; SHARMA, S.; AGARWAL, V. Emotional analysis using multinomial logistic regression. *Indian Journal of Science and Technology*, [S.l.], v. 9, n. 39, p. 1-10, 2016. Disponível em:https://doi.org/10.17485/ijst/2016/v9i39/102106.

RIBAS, D. *et al*. Automatic voice disorder detection using self-supervised representations. *IEEE Access*, v. 11, p. 14915–14927, 2023.

ROY, N. *et al*. Evidence-based clinical voice assessment: a systematic review. *American Journal of Speech-Language Pathology*, v. 22, p. 212–226, 2013. DOI: https://doi.org/10.1044/1058-0360(2012/12-0014).

SABIR, B. *et al*. Improved algorithm for pathological and normal voices identification. *International Journal of Electrical and Computer Engineering*, v. 7, p. 238–243, 2017.

SAINANI, Kristin L. Multinomial and ordinal logistic regression. PM&R, [S.l.], v. 13, n. 9, p. 1050–1055, 2021. Disponível em: https://doi.org/10.1002/pmrj.12622.

SANCHEZ-PINTO, L. N. *et al*. Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics*, v. 116, p. 10–17, 2018.

SANTOS, P. C. M. D. *et al*. Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. *Journal of Voice*, v. 33, n. 2, p. 220–225, 2019.

SANTOS, P. C. M. D. *et al*. Effect of synthesized voice anchors on auditory-perceptual voice evaluation. *CoDAS*, v. 33, n. 3, 2021.

SHI, B.; IYENGAR, S. S. *Mathematical theories of machine learning: theory and applications*. Cham: Springer, 2020.

SHRIVASTAV, R.; SAPIENZA, C. M.; NANDUR, V. Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, v. 48, p. 323–335, 2005.

SOFRANKO, J. L.; PROSEK, R. A. The effect of experience on classification of voice quality. *Journal of Voice*, v. 26, n. 3, p. 299–303, 2012.

SOFRANKO, J. L.; PROSEK, R. A. The effect of levels and types of experience on judgment of synthesized voice quality. *Journal of Voice*, v. 28, n. 1, p. 24–35, 2014.

TITZE, I. R. *Workshop on acoustic voice analysis: summary statement*. Iowa City: National Center for Voice and Speech, 1995.

VAN STAN, J. H.; MEHTA, D. D.; HILLMAN, R. E. Recent innovations in voice assessment expected to impact the clinical management of voice disorders. *Perspectives of the ASHA Special Interest Groups*, v. 2, p. 4–13, 2017.

WALDEN, P. R.; KHAYUMOV, J. The use of auditory-perceptual training as a research method: a summary. *Journal of Voice*, 2020. https://www.sciencedirect.com/science/article/abs/pii/S0892199720302496.

WERNICK, M. N. *et al*. Machine learning in medical imaging. *IEEE Signal Processing Magazine*, v. 27, n. 4, p. 25–38, 2010.

WIET, G. *et al*. Translating surgical metrics into automated assessments. *Proceedings of Medicine Meets Virtual Reality*, p. 543–548, 2012. https://pubmed.ncbi.nlm.nih.gov/22357055/.