INSTITUTO FEDERAL DA PARAÍBA
CAMPUS CAJAZEIRAS
CURSO DE LICENCIATURA EM MATEMÁTICA

WELLINGTON FERREIRA DE ALMEIDA

APLICAÇÃO DO ALGORITMO K-MEANS PARA DETECÇÃO DE PADRÕES EM DADOS VOCAIS

CAJAZEIRAS
2022

WELLINGTON FERREIRA DE ALMEIDA

APLICAÇÃO DO ALGORITMO K-MEANS PARA DETECÇÃO DE PADRÕES EM DADOS VOCAIS

Trabalho de conclusão de curso apresentado ao **Curso de Licenciatura em Matemática** do **Instituto Federal da Paraíba**, como requisito à obtenção do título de **Licenciado em Matemática**.
Orientador(a): Prof. Me. Alisson de Oliveira Silva.

Cajazeiras

2022

WELLINGTON FERREIRA DE ALMEIDA

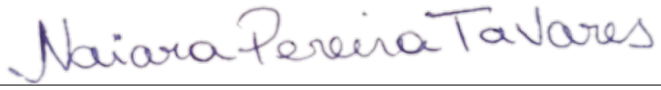
APLICAÇÃO DO ALGORITMO K-MEANS PARA DETECÇÃO DE PADRÕES EM DADOS VOCAIS

Trabalho de conclusão de curso apresentado ao programa de **Curso de Licenciatura em Matemática** do **Instituto Federal da Paraíba**, como requisito à obtenção do título de **Licenciado em Matemática**.

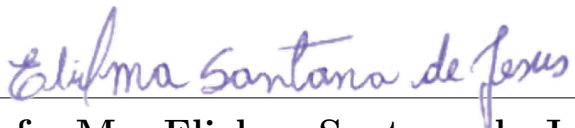
Data de aprovação: 13/05/2022

Banca Examinadora:

Prof. Me. Alisson de Oliveira Silva
Instituto Federal da Paraíba - IFPB



Profa. Esp. Naiara Pereira Tavares
Instituto Federal da Paraíba - IFPB



Profa. Ma. Elielma Santana de Jesus
Universidade Federal Rural de Pernambuco - UFRPE

IFPB / Campus Cajazeiras
Coordenação de Biblioteca
Biblioteca Prof. Ribamar da Silva
Catalogação na fonte: Suellen Conceição Ribeiro CRB-2218

A447a Almeida, Wellington Ferreira de

Aplicação do algoritmo k-means para detecção de padrões em dados vocais / Wellington Ferreira de Almeida. – Cajazeiras/PB: IFPB, 2022.

47f.:il.

Trabalho de Conclusão de Curso (Graduação em Matemática) - Instituto Federal de Educação, Ciência e Tecnologia da Paraíba-IFPB, Campus Cajazeiras. Cajazeiras, 2022.

Orientador(a): Prof. Me. Alisson de Oliveira Silva.

1. Matemática. 2.Medidas Acústicas. 3. K-means. 4. Algoritmo.

I. Almeida, Wellington Ferreira de. II. Titulo.

CDU: 51 A447a

Dedico este trabalho a toda minha familia, mas em especial meus pais: Angelita Ferreira da Silva e José Alves de Almeida e a meu irmão Wilamy Ferreira de Almeida.

AGRADECIMENTOS

A elaboração deste trabalho de conclusão de curso não seria possível sem a colaboração de alguns intermediários que me apoiaram ao longo desta caminhada. Desde já, queria deixar meus agradecimentos a todos que contribuíram de forma direta ou indireta nessa missão, ajudando-me finalizar com sucesso essa etapa complementar tão importante da minha futura carreira profissional acadêmica.

Inicialmente, agradeço a meu orientador Professor Alisson de Oliveira Silva que disponibilizou a proposta do tema, dando-me incentivo durante a pesquisa e mostrando o caminho a ser trilhado para alcançar nosso objetivo. Pra mim foi uma grande satisfação tê-lo como orientador, já que demosntrou uma excelente parceria no desenvolvimento deste trabalho.

Agradeço também a todos os professores que fizeram parte da minha formação acadêmica no decorrer do curso por todo apoio e conhecimentos compartilhados.

Não poderia esquecer de deixar meus agradecimentos aos colegas de sala que sempre foram colaborativos ajudando uns aos outros em ocasiões dificéis. No ensino presencial por exemplo, em semana de avaliações, a gente da turma formava um grupinho de estudo na sala da biblioteca para explorar melhor o conteúdo e saciar as dúvidas um do outro, porém surgiu a pândemia e impossibilitou este hábito. Então, durante a fase de ensino remoto tivemos que se adaptar ao distanciamento e passar a estudar de forma individual. Entretanto, mesmo com essa falta do contato físico com os colegas, não impediu que fosse promovido o dialógo entre nós.

Finalmente, agradeço a toda minha familia que sempre deu maior apoio e insentivo para seguir com meus estudos. É importante evidenciar que esta foi uma fase bastante difícil da minha vida, pois no decorrer desse percurso deparei-me com diversos encargos desafiadores que sugaram de mim um maior esforço, tanto na área acadêmica, como também, na minha vida pessoal. Contudo, através da minha persistência e dedicação, fui capaz de superar todos esses obstáculos presenciados em minha vida e hoje estou grato por toda a experiência que consegui adquirir. Com certeza essa formação favorecerá um futuro melhor para minha personalidade já que agora estou preparado para prestigiar a profissão como educador.

"Ensinar não é transferir conhecimento, mas criar as possibilidades para a sua própria produção ou a sua construção."

Paulo Freire, Pedagogia da Autonomia

RESUMO

A mineração de dados é uma tarefa importante quando se trata de encontrar padrões em conjuntos de dados. Esta técnica conta com o auxílio de algoritmos computacionais de análise estatística e sistema de banco de dados. Um dos principais procedimentos é a clusterização, que depende de funções provenientes de medidas de similaridade para ser executado. Neste trabalho, é apresentado o desempenho do algoritmo de particionamento K-means para separar grupos de indivíduos com lesão e sem lesão na laringe. Essa análise foi feita com base em medidas acústicas, extraídas a partir de sintomas elencados por pacientes, como autopercepção e queixas vocais. Para isso, foi utilizado um banco de dados referente ao projeto de pesquisa intitulado por Acurácia das Medidas de Análise Acústica Linear na Avaliação dos Distúrbios da Voz, desenvolvido no Laboratório Integrado de Estudos da Voz (LIEV), da Universidade Federal da Paraíba (UFPB). Neste estudo, foram incluídos pacientes com idade superior a 18 anos e inferior a 65 anos que apresentassem queixa vocal e diagnóstico laringológico prévio. Para avaliar a performance dos diferentes grupos de medidas foram calculadas algumas medidas externas: F1, índice de Rand e sua versão ajustada, como também, a medida interna, gráfico de silhueta. Os resultados sugerem, de maneira geral, que as medidas acústicas apresentaram desempenho superior, quando comparado as medidas das Escalas Analógicas Visuais (EAV) e as Escalas de Sintomas Vocais (ESV).

Palavras-chave: Medidas acústicas, Clusterização, K-means, Lesão laríngea.

ABSTRACT

Data mining is an essential task for finding patterns in data sets. This technique relies on the help of computer algorithms for statistical analysis and database systems. Clustering methods play a highlighted role in the literature and using similarity measures. In this paper, we present the performance of the K-means partitioning algorithm to separate groups of individuals with and without laryngeal lesions. This analysis was done based on several acoustic measurements extracted from symptoms listed by patients, such as self-perception and vocal complaints. We used a database from the research project entitled Accuracy of Linear Acoustic Analysis Measures in the Evaluation of Voice Disorders, carried out at the Integrated Laboratory for Voice Studies (LIEV) from the Federal University of Paraíba (UFPB). This study included patients aged over 18 years and under 65 years who had a voiced complaint and a previous laryngological diagnosis. We calculated the external measures F1, Rand index, and its adjusted version to assess the performance of the different groups of measures. The results suggest, in general, that the acoustic measures showed superior performance when compared to the EAV and ESV scales.

Keywords: Acoustic measures, Clustering, K-means, Laryngeal injury.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A voz é um dos elementos fundamentais na vida do ser humano, tanto para resolver problemas diários, quanto como ferramenta de comunicação para interagir na sociedade em que se vive. Habitualmente, a voz é um fator também utilizado para demonstar um estado de satisfação, como emoções, tristezas e sensações de felicidade(GUIMARÃES et al., 2010). Para cada situação, o individuo é capaz de expor seus sentimentos através de uma única entonação, como por exemplo, para expressar um sentimento de tristeza o som da voz fica cabisbaixa e com pouca ênfase, já um sentimento de alegria apresenta características mais entusiasmadas na entonação vocal.

Behlau et al. (2018) enfatizam que há uma série de agentes que compõem a estrutura para a construção da voz. De acordo com os autores, tais estruturas são chamadas de articuladores dos sons da fala, que fazem parte do trato vocal e estão localizados nas cavidades de resonância. Dessa forma, os sons produzidos pela concavidade da boca são articulados através do movimento da lingua, lábios, mandíbula e vél palatino, onde os mesmos permitem a entrada de ar no nariz para a produção dos sons nasais. Contudo, para produzir sons consistentes e intelegivéis para o receptor é necessário que os movimentos entre esses componentes sejam bem consolidados. Ainda, partindo desse mesmo ponto de vista, Silva (2014) complementa explicando que a voz é produzida em conjunto com outros órgãos que fazem parte do nosso corpo, especificamente o sistema respiratório, digestivo, nervoso e múscular, mas que envolve também a cartilagem e a estrutura óssea.

A laringe é um dos principais órgãos responsáveis pela produção da voz, visto que, permite a passagem de ar enquanto respiramos e protege os pulmões da entrada de substâncias indesejadas durante a alimentação. Entretanto, para impedir que esse órgão seja danificado durante essas ocasiões, as pregas vocais atuam enquanto o indivíduo está ingerindo tais substâncias tóxicas ou engolindo os alimentos de forma brusca. Nesse momento, as pregas vocais aproximam-se uma das outras, impossibilitando a passagem destes elementos para a laringe por alguns instantes (BEHLAU et al., 2018).

Silva (2014) salienta que com o passar do tempo esses órgãos emitentes da voz passam a sofrer certas influências de desgastes. Behlau et al. (2018) apontam que alguns dos principais motivos para essa defasagem nos órgãos vocais são ocasionadas por hábitos nocivos, como a utilização do fumo, consumo de bebidas alcoólicas e uso de drogas ilícitas. Outros fatores como a poluição, hábitos vocais inadequados, recursos que promovem alergia como o ar-condicionado, má alimentação, prática esportivas em excesso, falta de repouso e uso de medicamentos também comprometem a saúde da voz.

Caso o indivíduo não consiga controlar essas indicações básicas para manter a emissão da voz saudável, consequentemente a insistência em cometer tais hábitos, podem comprometer o relaxamento das pregas vocais, limitando o movimento das articulações e fazendo com que a musculatura laríngea seja atrofiada, culminando um mal funcionamento na qualidade vocal (SILVA, 2014).

Dentre outros, Silva (2018) destaca que há diversos tipos de transtornos que podem afetar o trato vocal tais como: nódulos, cistos, pólipos e calos vocais, que são ocasionados pelo esforço expressivo da voz. Justificando o que foi mencionado por Behlau et al. (2018) anteriormente, Sodré et al. (2016) evidenciam que o uso excessivo do cigarro e consumo de bebidas alcoólicas causam patologias de risco, promovendo irritações na laringe e aumentando a possibilidade de desenvolver câncer de pulmão e laringe.

Nessa perspectiva, é importante que cada um tenha os cuidados necessários e fiquem atentos as possíveis inconsistências apresentadas por seus órgãos vocais. Alguns aspectos presentes numa voz saudável mencionados por Behlau et al. (2018) são identificados quando o som transmitido é limpo, claro e agradável para o ouvinte, sendo emitido sem muito esforço. Além disso, uma voz pode ser considerada saudável quando o indivíduo consegue fazer variações quanto a sua qualidade, frequência, modulação e intensidade, que podem variar de acordo com o ambiente ou espaço de comunicação.

Sodré et al. (2016) ressaltam que avaliar a qualidade vocal consiste em diagnosticar a existência de distúrbios que provocam impedimentos ou mal funcionamento da voz. Pessoas que utilizam a voz constantemente em um tom mais expressivo, como é o caso dos professores, palestrantes e cantores, tendem a sofrer um desgate maior em seus órgãos vocais. É importante que o indivíduo ao perceber algum tipo de inconsistência vocal, ou sentir dificuldades em transmitir o som da fala, busque o quanto antes o auxílio de um especialista para fazer o diagnóstico, pois quanto mais rápida for identificada a lesão, mais simples e eficaz será o tratamento.

A detecção de patologias laríngeas é feita através de exames específicos e solicitados por profissional capacitado. De maneira geral, tratam-se de exames de diagnóstico por imagem e consistem na introdução de uma microcâmera pela cavidade nasal, permitindo uma visualização completa da laringe. Dentre os exames mais utilizados destacam-se a videolaringoscopia e a nasofibrolaringoscopia. Dessa forma, os exames são invasivos e ocasionam desconforto aos pacientes na sua realização (SODRÉ et al., 2016).

Para que seja possivél obter uma visão bem detalhada da laringe através desses equipamentos é necessário que o procedimento seja feito de forma adequada. Nesse sentido, a câmera destes aparelhos devem estar bem possicionadas de modo que a laringe fique

visível e com uma ótima qualidade para que o especialista possa dar um diagnóstico concreto da patologia localizada. No entanto, isso requer a colaboração por parte do paciente para que o médico consiga executá-lo corretamente. Nem sempre esses exames são fáceis de serem realizados, pois na videolaringoscopia, por exemplo, quando é inserido o laringoscópio na boca do paciente são causados efeitos de náuseas e dores constantes, visto que, esse equipamento toca a garganta, que é um órgão de alta sensibilidade. Dessa forma, esses sintomas promovem a inquietação do paciente, dificultando a realização adequada do exame (SODRÉ et al., 2016).

A fim de amenizar tais desconfortos nos pacientes, diversos pesquisadores têm buscado alternativas não invasivas para realizar esse tipo de diagnóstico. Autores como Byeon (2018), Castro e Prado (2002), Neto et al. (2012), Sodré et al. (2016), Takakura et al. (2018) e Teixeira et al. (2011) apresentam propostas baseadas em aprendizado de máquina, que consistem em métodos matemáticos capazes de identificar padrões acústicos, utilizando diferentes tipos de classificadores e medidas de dissimilaridade.

Na maioria destes trabalhos, as informações para a análise são coletadas com base no efeito sonoro transmitido pelo indivíduo, denominada de análise acústica (TEIXEIRA et al., 2011). As medidas acústicas têm sido utilizadas sob diferentes modelagens estatísticas, objetivando a construção de modelos de decisão que auxiliem profissionais no que tange a detecção precisa de patologias laríngeas, e consequentemente no tratamento precoce de tais doenças.

Em Teixeira et al. (2011) por exemplo, foi usado a técnica de análise acústica para quantificar e caracterizar sinais sonoros dos pacientes objetivando-se na identificação de desordens vocais provocadas por patologias na laringe. O autor justifica que a nálise acústica permite de forma não invasiva determinar a qualidade vocal do indivíduo utilizando parâmetros acústicos que compõem o sinal da voz. Visto que, a mesma é capaz de fornecer o formato completo da onda sonora, permitindo com que sejam avaliadas determinadas características, como a frequência do número de vibrações produzidas pelas cordas vocais, medidas de pertubação da frequência definida por Jitter e medidas de pertubação da amplitude definida por Shimmer. Ambas nomenclaturas, são constituídas como principais parâmetros acústicos utilizados na detecção de patologias laríngeas. Além deles, foi empregada nesta pesquisa a medida F_0 que é constituída pela média, mediana, desvio padrão, máximo e mínimo local. Para extração desses parâmetros foram utilizados dois métodos distintos, o Cepstral e o método da Autocorrelação.

O método Cepstral é uma operação matemática que consiste em extrair a Transformada de Fourier do espectro do sinal na forma de logaritmos. Este método permite separar os efeitos da fonte e do meio de transmissão do sinal da voz com frequencias difrentes. Já

o método da autocorrelação é uma equação definida por um sinal estacionário $x(t)$ e uma função de atraso dada por $r_x(\tau)$, a frequência fundamental associada ao sinal da fala $x(t)$ é a F_0. Desta forma, os parâmetros jitter, shimmer e F_0, exprimiram valores normativos para emissão do sinal sonoro entre crianças, homens e mulheres. Diante disso, foi observado que as mulheres apresentaram maior índices de patologias vocais comparado-se aos homens. Entretanto, de acordo com resultados obtidos a partir dos gráficos, verificou-se que os métodos usados mostraram-se eficazes para a análise.

Por outro lado, em um estudo realizado por Byeon (2018), é apresentado um modelo de algoritmo baseado em aprendizado de máquina denominado Support Vector Machine (SVM), onde a utilização dessa ferramenta objetivou-se prever a ocorrência de lesões laríngeas benignas em adultos. Para essa análise, foram utilizadas informações relativas a índices de análise acústica extraídas a partir de dados do ouvido, nariz e garganta dos pacientes. Neste estudo, o SVM foi capaz de encontrar o limite de decisão apropriado dos dados e separá-los em hiperplanos, fornecendo o mapeamento completo dos pacientes que apresentavam característas de lesões benignas na laringe.

Sodré et al. (2016) também desenvolveu um trabalho nesta mesma perspectiva, porém optou em utilizar além do SVM, um outro método embasado em duas configurações de redes neurais para identificar padrões acústicos em pacientes com patologias laríngeas. A primeira configuração, foi gerada a partir de $524,287$ combinações de 19 medidas acústicas para classificar as vozes em normal ou patológicas, e com ela foi possível atingir uma acurácia máxima de $99,5\%$ $(96,99 \pm 2,08\%)$. Já a segunda, classificou 23 tipos de vozes incluindo (vozes normais), mostrando melhor acurácia na identificação de hiperfuncionalidades e vozes normais com $(58,23 \pm 18,98\%)$ e $(52,15 \pm 18,31\%)$, respectivamente.

Na pesquisa desenvolvida por Takakura et al. (2018), é mostrado uma proposta de automatização para o diagnóstico de patologias laríngeas também utilizando técnicas e métodos de Aprendizagem de Máquinas. Neste trabalho os autores utilizaram algoritmos classificados como supervisionados e não-supervisionados para executar a análise. Os métodos supervisionados são aqueles que precisam da supervisão de um agente externo para fornecer as informações de entrada e comparar os resultados finais esperados. Ao contrátrio, os métodos não-supervisionados não precisam da supervisão de agentes externos para comparar os resultados finais. Os métodos computacionais usados nesta pesquisa, foram o Support Vector Machine (SVM), Redes Neurais, OPF (Optimum-Path Forest), (K Nearest Neighbor), Classificadores Bayes, Density Based Spatial Clustering of Application with Noise (DBSCAN) e o K-means com objetivo de comparar a aplicação desses métodos e sua eficácia no contexto da medicina. Sendo assim, a análise se deu por meio de indícios levantados pelas características de cada anomalia, onde foi disponibilizado um banco de dados pela UCI -Machine Learning Repository, fornecendo os tipos de doenças, classes,

atributos e instâncias de cada uma delas. Após a inserção dos dados nesses algoritmos e feito o processamento, alguns deles demonstraram resultados satisfatórios, enquanto outros não foram tão eficazes como era esperado. Sendo assim, os pesquisadores chegaram a conclusão que esses metódos podem servir como uma ferramenta auxiliar no contexto da medicina para realizar uma avaliação prévia de determinadas patologias, evitando que o paciente submeta-se a procedimentos invasivos.

Dessa forma, o uso de métodos de aprendizagem de máquina, supervisionados ou não supervisionados, têm se aprensetando como uma alternativa viável no auxílio aos profissionais da saúde. Em particular, métodos que consistem na identificação de agrupamentos subjacentes aos dados são fundamentais, já que dados rotulados requerem a intervenção de especialistas e tornam-se, dessa forma, cada vez mais escassos.

Neste trabalho, pretendeu-se avaliar o desempenho de diversas medidas acústicas na detecção de grupos de pacientes com lesão e sem lesão na laringe. Para isso, foi aplicado um algoritmo particional de clusterização, denominado K-means que é um dos mais amplamente utilizados na literatura devido a sua simplicidade e eficiência em aplicações práticas, além de sua fácil implementação computacional.

O referente trabalho está estruturado em três capítulos. O primeiro introduz os assuntos abordados e a problemática. O segundo expõe a metodologia adotada para o desenvolvimento da pesquisa, onde foi definido o banco de dados, a técnica de agrupamento K-means e alguns métodos de validação interna e externa, como o gráfico de silhueta a medida F1, Acurácia, índice de rand e rand ajustado. Ainda é especícado sobre os testes de hipóteses que serão utilizados como o de Shapiro-Wilk e Wilcoxon. No terceiro são expostos os resultados obtidos a partir da aplicação do método k-means. E finalmente, são apresentadas as considerações finais a respeito da análise realizada.

1.1 DEFINIÇÃO DO PROBLEMA

Ainda hoje, milhares de pessoas presenciam indícios de patologias vocais, mas lamentavelmente, as medidas de prevenção na maioria das vezes não são adotadas por parte da cultura brasileira. A Academia Brasileira de Laringologia e Voz, estima que 30% dos brasileiros são acometidos por algum tipo de lesão nas pregas vocais. Quando essas enfermidades não são tratadas com antecedência e de maneira adequada, a situação pode piorar, levando o indivíduo a alterações mais rigorosas como é o caso do câncer de laringe, que afeta cerca de 15 mil brasileiros por ano (GUIMARÃES et al., 2010).

Especialistas consideram que sintomas como o cansaço vocal, rouquidão, pigarros, dificuldade para engolir ou respirar, dores na garganta e a perda da voz, são indícios

de lesão na laringe. Dessa forma, quando o diagnóstico é feito com antecedência e logo iniciado o tratamento, o indivíduo tem 90% de chance de ser curado e ficar livre da doença (GUIMARÃES et al., 2010).

Por outro lado, o indivíduo mesmo percebendo que há algum tipo de inconsistência na voz, sente receio de passar por um procedimento médico ou consultar um fonoaudiólogo para avaliar o estado de sua saúde vocal. Um fator que culmina tanto medo nas pessoas são os procedimentos invasivos utilizados para realizar esse tipo de análise, pois podem causar dores ou até sensações desagradáveis no paciente.

De acordo com a Organização Mundial da Saúde (OMS,1993) apud (REGATIERI et al., 2018, pg. 20), os "procedimentos invasivos podem ser descritos como técnicas operativas ou diagnósticas que envolvem o uso de instrumentos que penetram os tecidos ou invadem algum orifício do corpo". Justamente por esse motivo, projeta-se buscar outras alternativas que possam suprir essa necessidade, evitando que o paciente se submeta a certos tipos de exames para iniciar o tratamento.

Desde o século XX a análise acústica da voz tem despertado grande interesse em pesquisadores em avaliar a qualidade vocal através de recursos computacionais. Porém, só por volta dos anos 70 foram utilizados os primeiros softwares de análise acústica, e consequentemente, a aplicação destes recursos promoveram um excelente progresso ao permitir realizar uma avaliação convicta da qualidade vocal, coletando apenas o sinal da voz (FREITAS, 2010).

De acordo com Capucho (2018, p. 102):

> O processamento digital do sinal permite a análise, transformação ou interpretação de sinais através de algoritmos computacionais. Deste modo, as medidas obtidas na análise acústica correspondem a parâmetros físicos definidos (CAPUCHO, 2018, p. 102).

Esses parâmetros mencionados correspondem as propriedades físicas das ondas sonoras da voz que são transmitidas de órgãos internos para um espaço externo.

Deste modo, a análise acústica realizada por meio de ferramentas computacionais possibilita ao indivíduo um diagnóstico de maneira não invasiva para a identificação da patologia vocal. Guimarães (2007 apud CAPUCHO, 2018) ressalta que esse diagnóstico é feito com base em diferentes parâmetros acústicos, tais como: periodicidade, amplitude, duração e composição espectral. Contudo, a confiabilidade das informações coletadas dependem da forma em que o sinal sonoro da voz foi captado, pois alguns parâmetros como a fonte glótica e as cavidades de ressonância do trato vocal apresentam interações

complexas. Assim, esses parâmetros ficam dependendo apenas das forças biomecânicas e aerodinâmicas da laringe e as estruturas supraglóticas. Caso esses componentes apresentem características anatômicas ou fisiológicas fora do padrão, indica que os resultados tiveram um desvio ao que se era esperado, sendo assim considerados como indicadores de patologia vocal (FREITAS, 2010).

Com base em uma revisão bibliográfica, Freitas (2010, p. 46) evidencia alguns dos principais objetivos e vantagens da utilização desses equipamentos para a análise acústica:

1. Oferece uma maior compreensão acústica do output vocal e aproxima formas distintas de avaliação da voz, nomeadamente a análise áudio-perceptual e a acústica ou a laringo-estroboscópica e a acústica;
2. Proporciona de modo expedito e user-friendly dados normativos para realidades vocais distintas culturais, profissionais e/ou patológicas;
3. Propicia informação importante sobre o impacto do sinal vocal no ouvinte (Weismer, 1984 in Murdock, 2005);
4. Oferece a documentação – gráfica e numérica – necessária para descrever a qualidade vocal de um indivíduo, seja ele um utilizador profissional da voz ou um paciente em tratamento, por disfonia, auxiliando e ratificando pareceres judiciais ou outros atestados com carácter legal;
5. Proporciona imagens e gráficos de análises acústicas, com fácil compreensão por parte do paciente/falante em avaliação ou acompanhamento terapêutico, favorecendo um melhor prognóstico associado ao maior envolvimento e consequente motivação para o processo de mudança vocal;
6. Monitoriza a eficácia de um tratamento e permite comparar resultados vocais de diferentes metodologias de intervenção, em fases distintas do processo terapêutico ou cirúrgico/medicamentoso;
7. Acompanha o desenvolvimento de uma voz profissional, e orienta a sua adequação ao longo do tempo, inclusive com a possibilidade de sistemas de feedback-análise acústica em tempo-real;
8. Assume-se como um instrumento de detecção precoce de problemas vocais e laringeos, por exemplo, em campanhas de triagem, pela detecção de níveis de perturbação fonatória acima dos valores de referência de uma população não-disfónica.

Como proposta de intervenção, pretende-se utilizar nesta pesquisa métodos não supervisionados com base em medidas acústicas, para avaliar sua performance na detecção

de padrões em dados de pacientes com patologias laríngeas. Uma vez que os dados foram rotulados previamente por profissionais da área, a performance do método K-means e do uso dessas medidas serão avaliados comparando os agrupamentos formados com os grupos originais. Sendo assim, almeja-se que as medidas acústicas e o método adotado apresentem bons resultados na detecção de padrões, especificamente, na identificação de grupos de pacientes com e sem lesões laríngeas, visto que dados rotulados são escassos e requerem a análise de especialistas.

Nesta perspectiva, espera-se ainda que os resultados deste trabalho possam subsidiar especialistas para o uso de métodos não supervisionados para identificar grupos subjacentes aos dados sem a necessidade de um profissional, além de permitir avaliar a performance das medidas acústicas na separação desses grupos com base em dados já rotulados.

1.2 OBJETIVO GERAL

O objetivo principal deste trabalho, consiste em avaliar o desempenho de diversas medidas acústicas e das Escalas de Sintomas Vocais (ESV) e Escalas Analógias Visuais (EAV) na detecção de padrões em dados de indivíduos com patologias laríngeas.

1.2.1 Objetivos Específicos

Para alcançar o objetivo geral do referente trabalho, será necessário contemplar os seguintes objetivos específicos:

- Caracterizar o perfil de indivíduos com e sem patologias laríngeas com base em medidas acústicas;
- Identificar grupos de indivíduos com e sem patologias laríngeas com base em diferentes grupos de medidas acústicas;
- Avaliar o desempenho do método K-means e dos diferentes grupos de variavéis na identificação de indivíduos com e sem patologias laríngeas.

2 METODOLOGIA

Este capítulo será destinado para expor os aspectos metodológicos fundamentais para o desenvolvimento desta pesquisa, pelo qual, estará dividido em 9 seções. A Seção 2.1, apresenta o banco de dados utilizado para realização da análise. Na Seção 2.2 é descrito sobre o método de agrupamento k-means. A Seção 2.3 mostra o método da silhueta para avaliar particionamentos de dados. Na Seção 2.4 defini-se o conceito sobre a medida acurácia para comprar proximidades entre os resultados obtidos. As Seções 2.5 e 2.6 são definidos métricas e índices de validação externa denominados F1, índice rand e rand ajustado respectivamente. Nas Seções 2.7 e 2.8 são apresentados os testes de hipóteses de Shapiro- Wilk e Wilcoxon nesta ordem. E finalmente na Seção 2.9 será mostrado o suporte computacional `R` utilizado para a implementação do algoritmo e processamento dos dados.

2.1 BANCO DE DADOS

Os dados considerados neste trabalho são referentes ao projeto de pesquisa intitulado por: Acurácia das Medidas de Análise Acústica Linear e não Linear na Avaliação dos Distúrbios da Voz, desenvolvido no Laboratório Integrado de Estudos da Voz (LIEV) da Universidade Federal da Paraíba (UFPB). O estudo foi aprovado pelo comitê de ética do Centro de Ciências da Saúde da UFPB sob parecer de número 508.200.

Foram incluídos no estudo, pacientes com idade superior a 18 anos e inferior a 65 anos, que apresentassem queixa vocal e diagnóstico laringológico prévio. Foram excluídos indivíduos com alterações cognitivas ou neurológicas, que impossibilitassem o preenchimento do questionário utilizado. A amostra, obtida por acessibilidade, foi composta por participantes atendidos no setor de triagem do LIEV.

A coleta dos dados foi conduzida no momento inicial da avaliação do paciente no LIEV-UFPB antes do tratamento vocal, com uma duração média de 60 minutos. Para a coleta das informações os seguintes procedimentos foram realizados:anamnese com dados pessoais e de queixa vocal e o registro da voz. Foram utilizadas a Escala de Sintomas Vocais (ESV) e a Escala Analógica Visual(EAV) para avaliação de sintomas e desvios vocais, respectivamente, sendo a última aplicada por fonoaudiólogo especialista em análise perceptivo-auditiva. No que tange o registro da voz, foram consideradas medidas acústicas obtidas a partir do sinal vocal. A tabela 2.1 descreve os grupos de medidas mencionados com seus respectivos parâmetros.

Tabela 2.1 – Descrição dos parâmetros de cada grupo de medida fornecido pelo LIEV-UFPB que foram utilizados para realizar a análise.

Grupo de Medidas	Variável (Sigla)
Escala Analógica Visual	Grau Geral (EAV - GG)
	Rugosidade (EAV - R)
	Suprosidade (EAV - S)
	Tensão (EAV - T)
Escala de Sintomas Vocais	Domínio Total (ESV - T)
	Domínio de Limitação (ESV - L)
	Domínio Emocional (ESV - E)
	Domínio Físico (ESV - F)
Medidas Acústicas	Frequência Fundamental (F0 - média)
	Frequência fundamental (F0 - desvio padrão)
	Primeiro formante (F1 - média)
	Primeiro formante (F1 - desvio padrão)
	Segundo formante (F2 - média)
	Segundo formante (F2 - desvio padrão)
	Terceiro formante (F3 - média)
	Terceiro formante (F3 - desvio padrão)
	Variabilidade da frequência fundamental a curto prazo (Jitter)
	Variabilidade da amplitude da onda sonora a curto prazo (Shimmer)
	Glottal to Noise Excitation (gne)
	Delay (tau)
	Dimensão de imersão (graus de liberdade) (m)
	Raio de vizinhança (raio)
	Taxa de recorrência (rec)
	Determinismo em um RP (det)
	Comprimento médio das linhas diagonais no gráfico de recorrência (Imed)
	Comprimento máximo das linhas diagonais (Imax)
	Entropia de Shannon (entr)
	laminaridade (lam)
	Medida tempo de permanência (Trapping Time) (tt)
	Comprimento máximo das linhas verticais (vmax)
	Medida tempo de recorrência do tipo 1 (t1)
	Medida tempo de recorrência do tipo 2 (t2)
	Medida da entropia do tempo de recorrência do tipo 1 (Recurrence Probability Density Entropy) (rpde)
	Coeficiente de clusterização (clust)
	Transitividade (trans)
	Medida divergência (div)
	Medida da relação entre determinismo e taxa de recorrência (ratio)

Fonte: Autor, 2022.

2.2 ANÁLISE DE AGRUPAMENTO - K-MEANS

A análise de agrupamento ou clusterização, consiste de uma série de técnicas estatísticas que têm como objetivo obter uma partição dos dados, frequentemente denominados de padrões (indivíduos, objetos, etc.), de forma que padrões dentro de um mesmo grupo sejam os mais similares possíveis, enquanto padrões pertencentes a grupos distintos sejam os mais dissimilares possíveis, segundo algum critério pré-estabelecido (GORDON, 1999; JAIN et al., 1999; XU; WUNSCH, 2005). Essas técnicas têm sido empregas em diversos contextos práticos, tais como em taxonomia, processamento de imagens, mineração de dados, recuperação de informação, dentre outras.

As técnicas de agrupamento podem ser divididas, de forma geral, em métodos hierárquicos e métodos particionais. Os métodos hierárquicos produzem uma resposta representada por uma hierarquia, isto é, uma sequência aninhada de partições do conjunto de padrões de entrada, resultando em uma estrutura de grupos conhecida como dendrograma. Por outro lado, nos métodos particionais o objetivo é obter uma partição única do conjunto de padrões de entrada em um número fixo de grupos, tipicamente através da otimização (geralmente local) de uma função objetivo (FERREIRA, 2013).

Dentre os métodos particionais, o K-means, proposto por MacQueen et al. (1967) é o algoritmo mais utilizado em situações práticas, devido a sua simplificadade de entedimento e facilidade de implementação computacional. Dado o número de clusters K, o algoritmo inicia escolhendo K padrões aleatoriamente como os centróides inciais, também denominados de protótipos. Em seguinda, cada observação é atribuída ao centróide mais próximo, baseada em uma medida de proximidade adequada. Uma vez que os clusters estão formados, seus centróides são atualizados. O algoritmo alterna, iterativamente, entre essas duas etapas, até que não haja mais mudanças nos centróides, ou até que outro critério de convergência seja alcançado. Reddy (2018, p. 89), ressalta que "o agrupamento K-means é um algoritmo guloso que é garantido convergir para um mínimo local [...], normalmente, a condição de convergência é relaxada e uma condição mais fraca pode ser usada."Selim e Ismail (1984) apresenta uma prova detalhada da convergência matemática do K-means.

O algoritmo 1, a seguir, apresenta um resumo sucinto da execução do algoritmo K-means. Na Figura é ilustrada as iterações do método K-means com $K = 3$ clusters. Na primeira iteração, os centróides são escolhidos aleatoriamente. Nas iterações seguintes, os centróides mudam de posição até que haja a convergência. Para calcular a distância de cada observação para o centróide mais próximo é necessário definir uma medida de similaridade apropriada. Dentre as várias propostas na literatura, a distância euclidiana (L_2). Outras escolhas possíveis, destacam-se as distâncias Manhattan (L_1) e a cosseno. Mas em geral, o K-means utiliza a distância Euclidiana, por ser o método mais usual

(REDDY, 2018).

A distância Euclidiana entre dois conjuntos de pontos $X = \{x_1, X_2, ...X_n\}$ e $Y = \{y_1, y_2, ...y_n\}$, n-dimensional, é definida como:

$$d_{euc}(X,Y) = \left[\sum_{j=1}^{d}(x_j - y_j)^2\right]^{\frac{1}{2}} = \left[(X-Y)(X-Y)^T\right]^{\frac{1}{2}},$$

em que x_j e y_j , são os valores do j-ésimo atributo de X e Y , respectivamente. Esta expressãopode ser ainda reescrita da seguinte maneira:

$$d_{euc}(X,Y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2}.$$

Algoritmo 1: Algoritmo K-means

Entrada: Dados, K

Saída: Agrupamento

início

  Selecione K pontos como centróides;

  repita

    Forme K clusters atribuindo cada ponto ao centróide mais próximo;

    Recalcule os centróides de cada cluster;

  até *Convergência do algoritmo*;

fim

Seja $D = \{x_1, \ldots, x_N\}$ um conjunto de dados de tamanho N e $C = \{C_1, C_2, \ldots, C_k \ldots, C_K\}$ o cluster obtido através do K-means. A função objetivo empregada no K-means, denominada como Soma de Quadrados dos Erros (SQE) ou Soma de Quadrados dos Resíduos (SQR) é dada por

$$SQE(C) = \sum_{k=1}^{K}\sum_{x_i \in C_k}\|x_i - c_k\|^2,$$

em que

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|}, \quad k = 1, \ldots, K,$$

é o centróide do k-ésimo cluster.

O método K-means é essencialmente um problema de otimização, cujo objetivo é minimizar a função SQE. É possível mostrar, matematicamente, que a escolha da média

das observações do cluster como centróide otimiza a função objetivo. Para isso, denote C_k como o k-ésimo cluster, $x_i \in C_k$ e c_k a média do k-ésimo cluster. O centróide do cluster C_j, que minimiza a função objetivo pode ser obtido diferenciando SQE em relação a c_j como segue

$$\begin{aligned}\frac{\partial}{\partial c_j} SSE &= \frac{\partial}{\partial c_j} \sum_{k=1}^{K} \sum_{x_i \in C_k} (c_k - x_i)^2 \\ &= \sum_{k=1}^{K} \sum_{x_i \in C_j} \frac{\partial}{\partial c_j} (c_j - x_i)^2 \\ &= 2 \sum_{x_i \in C_j} (c_j - x_i)\end{aligned}$$

Igualando a expressão anterior a zero e isolando c_j, tem-se

$$2 \sum_{x_i \in C_j} (c_j - x_i) = 0 \Rightarrow |C_j| \cdot c_j = \sum_{x_i \in C_j} x_i \Rightarrow c_j = \frac{\sum_{x_i \in C_j} x_i}{|C_j|}, j = 1, \ldots, K.$$

Embora o K-means seja amplamente utilizado na literatura, destaca-se algumas limitações importantes do método tais como a escolha inicial dos centróides e do números de clusters, a presença de observações atípicas, etc. Nesse sentido, variações do K-means têm sido propostas na literatura no sentido de superar essas limitações. Algumas extensões de destaque do K-means são K-medoids, K-medians, Weighted K-means, Kernel K-means e Fuzzy K-means. Tais métodos são obtidos a partir de modificações adequadas na função objetivo. Para uma revisão desses métodos ver Aggarwal e Reddy (2014).

2.3 ÍNDICE DE SILHUETA

O método de silhueta foi desenvolvida por Rousseeuw (1996) utilizada para avaliar particionamentos de dados. Esta técnica consiste em mostrar o grau de semelhança entre um objeto e o cluster a qual pertence, isso significa que, há um índice de concordância entre eles em comparação com os outros clusters. Dessa forma, cada cluster é representado por uma silhueta que varia entre o intervalo de -1 a $+1$. Então, quanto mais alto for o valor da silhueta, melhor ajustado está o objeto ao cluster, enquanto que um baixo valor indica que este objeto foi alocado incorretamente ao cluster.

Desta forma, todo o agrupamento é representado por meio de um único gráfico, gerado a partir da combinação entre as silhuetas. Através dele, é possível estimar a qualidade dos clusteres, ou seja, distinguir quais deles apresentam resultados mais relevantes ou irrelevantes, possibilitando ainda uma visão mais ampla da configuração do conjunto de

dados. Entretanto, a largura média da silhueta fornece uma avaliação prévia e concisa do agrupamento, permitindo com que seja escolhido o número ideal de cluster para realizar a análise.

Rousseeuw (1987) define que para construir as silhuetas, primeiramente é necessário obter a partição dos objetos. Essa partição pode ser feita por meio de alguma técnica de agrupamento, como por exemplo o K-means. Em seguida, calculam-se todas as distâncias entre cada um desses objetos.

Diante disso, para cada objeto i é assosciado um valor $s(i)$, denominado *silhueta*, calculada com base na consistência e separação dos clusters. Deste modo, para definir $s(i)$, é tomado aleatoriamente um objeto i do conjunto de dados e representa-se por A o cluster em que o objeto foi atribuído (ROUSSEEUW, 1987).

A dissimilaridade média ou (largura média) de i em relação a todos os outros objetos de A é dado pela seguinte expressão:

$$a(i) = \frac{1}{N_A - 1} \sum_{j \in A, j \neq i} d(i,j),$$

em que, N_A representa o número de objetos contidos no agrupamento e $d(i,j)$ é a distância ou (dissimilaridade) entre os pontos i e j de N_A. É associado o valor -1 na divisão por conta que não é incluído a distância $d(i,i)$ na soma.

Agora, considera-se um cluster C qualquer, de modo que $C \neq A$ e calcula-se $d(i,C)$, a dissimilaridade média de i para todos os outros objetos pertencentes a C. Então, seleciona-se o menor desses comprimentos e defini-se a dissimilaridade média entre i e os objetos de C como sendo:

$$b(i) = min_{C \neq A} d(i,C).$$

Interpreta-se $b(i)$ como o cluster vizinho de i já que há um melhor ajuste entre eles em termos de proximidade. Assim, considerando que o número k de clusters seja maior que um, obtém-se o índice $s(i)$ combinando $a(i)$ e $b(i)$ como apresentado em (1).

$$s(i) = \begin{cases} 1 - a(i)/b(i) & se \quad a(i) < b(i), \\ 0 & se \quad a(i) = b(i), \\ b(i)/a(i) - 1 & se \quad a(i) > b(i). \end{cases} \tag{1}$$

Alternativamente, pode-se reescrever esta expressão em uma única fórmula:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}.$$

A Figura 2.1 exemplifica as distâncias consideradas para o cálculo de $s(i)$. Com base em Rousseeuw (1987), foram adotados três agrupamentos arbitários A, B e C, e denotado

o i-ésimo objeto como ponto de referência para a análise, interligando-o com os outros objetos do próprio cluster e dos clusters vizinhos, cujas distâncias são reprepresentadas por meio de linhas contínuas a fim de encontrar os valores de $a(i)$ e $b(i)$, respectivamente, sendo que i pertence ao cluster A.

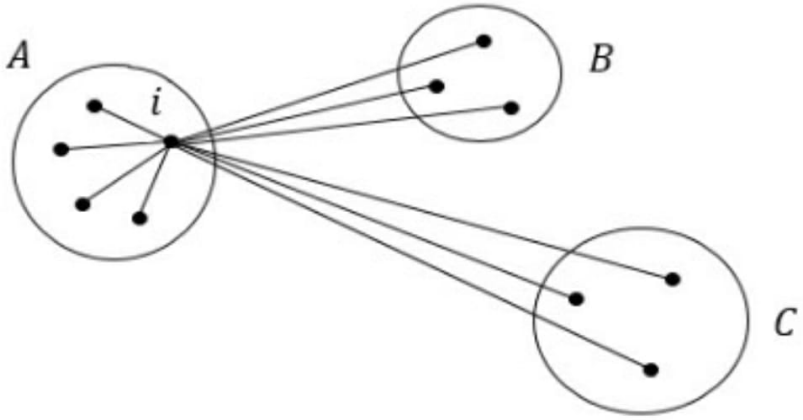


Figura 2.1 – Elementos envolvidos no cálculo de $s(i)$, com as linhas representando a distância entre o i-ésimo ponto e os demais objetos da amostra.

Caso seja atribuído apenas um único objeto ao cluster A, será impossível definir a média $a(i)$. Nesta situação, Rousseeuw (1987) sugere que $s(i)$ seja igualado a zero. Logo, a partir da definição (1), destacada acima, percebemos claramente que a valor da silhueta estará compreendida entre o intervalo $-1 \leq s(i) \leq 1$.

Dessa forma, se o valor de $s(i)$ for mais próximo de -1, significa que o i-ésimo objeto foi mal classificado no agrupamento, e neste caso $b(i) << a(i)$. Em outras palavras, pode-se afirmar que i está mais distante dos objetos de seu próprio grupo do que mesmo os objetos usados para calcular a média em relação a $b(i)$. Em contrapartida, se o valor de $s(i)$ estiver próximo de 1, considera-se que o objeto foi classificado adequadamente ao grupo, de modo que $b(i) >> a(i)$. Finalmente, se o valor de $s(i)$ for aproximadamente zero, indica que $a(i)$ e $b(i)$ são semelhantes, ou seja, o objeto analisado está compreendido num ponto intermediário entre os dois grupos (SOUZA, 2007).

Kaufman e Rousseeuw (1990 apud SOUZA, 2007) recomendam que uma maneira de determinar o número ideal de k cluster adequadamente, é construindo-se todas as possibilidades de agrupamentos com $k = 2, 3, ..., n-1$ grupos, e em seguida, selecionar aquele que apresentar o maior coeficiente de silhueta médio (CSM), que pode ser obtido utilizando-se a seguinte expressão:

$$CSM = \frac{\sum_{i=1}^{n} s(i)}{n}.$$

Após calcular todas as silhuetas, pode-se agora realizar sua construção gráfica.

Como exemplo, é ilustrado uma análise desenvolvida por Souza (2007), que utilizou dados originalmente publicados por Ruspini (1970), composto por 75 observações e duas variáveis. Foi delineado um agrupamento formado por quatro grupos, que foram gerados através do algoritmo de particionamento PAN, implemementado no pacote *cluster* do `R`.

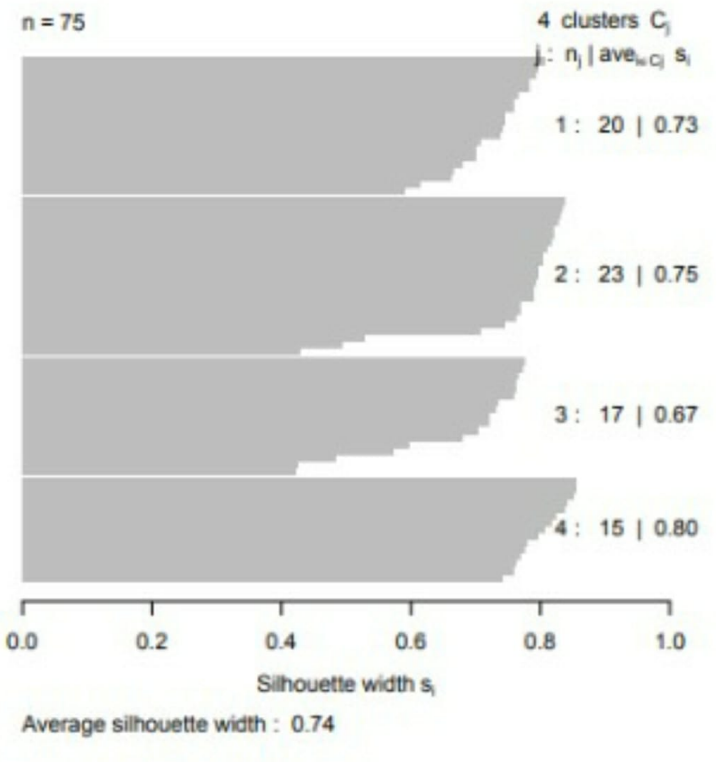


Figura 2.2 – Gráfico de silhueta para os dados de Ruspini (1970), com k=4.

Os dados amostrais são representados por barras de comprimento equivalente ao valor de $s(i)$, sendo descritos em direção ao eixo das abscissas do plano cartesiano e classificados em ordem decrescente para todos os objetos do agrupamento investigado. Dessa forma, as silhuetas mostraram quais objetos estavam bem ou mal ajustados em seus respectivos grupos. Pode-se observar na Figura 2.2, que o menor valor de $s(i)$ está no terceiro grupo que apresenta similaridade médida igual a $0,67$. Isso significa que o grupo não reflete uma boa qualidade de alocação para os objetos, comparativamente aos demais clusters.

De acordo com Rousseeuw (1987):

> As silhuetas oferecem a vantagem de dependerem apenas da partição real dos objetos, e não do algoritmo de agrupamento que foi usado para obtê-lo. Assim, as silhuetas poderiam ser usadas para melhorar os resultados da análise de cluster (por exemplo, movendo um objeto com negativo s(i) para seu vizinho), ou para comparar a saída de

> diferentes algoritmos de agrupamento aplicados aos mesmos dados (ROUSSEEUW, 1987, pg.59).

2.4 ACURÁCIA

A acurácia é uma medida geralmente utilizada para estimar a proximidade de um resultado experimental em relação a o seu valor verdadeiro, ou seja, será obtido um grau de aproximação do valor estimado comparando-se ao valor do parâmetro original. Dessa forma, quanto maior for o valor da acurácia, mais vericído é a validade do resultado.

De acordo com Monico et al. (2009), a acurácia é entendida como o afastamento entre um valor de referência e o valor estimado. A medida que a representa, foi proposta por Gauss, denominada Erro Quadrático Médio (EQM), ou Mean Square Error (MSE) em inglês, que pode ser expressada da seguinte forma:

$$MSE = m^2 = \left\{\left(P - E\left\{(P)\right\}^2\right)\right\} = \sigma_p^2 + b^2 \cong \sum_{i=1}^{n} \frac{\varepsilon_1^2}{n},$$

em que, σ_p^2 simboliza a dispersão das medidas, b é a tendência ou vício estimador, e ε é a diferença entre um valor observado (medido) e o valor tomado como referência (conhecido).

2.5 MEDIDA F1

Como métrica de avaliação, será utilizado neste trabalho a medida F_1, que é um parâmetro de avaliação externa que serve para comparar soluções de referências com a saída do algoritmo de agrupamento. A medida F_1 mais conhecida como $F-measure$ é definida como uma média harmônica, baseada na combinação entre precisão e revocação de dados (MONTALVO et al., 2012).

A *Precisão* e *Revocação*, são definidas como (2) e (3), respectivamente:

$$Precisão = \frac{TP}{FP+TP}, \tag{2}$$

$$Revocação = \frac{TP}{FN+TP}, \tag{3}$$

em que TP e FP correspondem, respectivamente a verdadeiros e falsos positivos, enquanto FN representa os falsos-negativos. Esses termos comparam os resultados do classificador em teste com julgadores confiáveis externos.

De acordo com Chimieski e Fagundes (2013), a expressão que exprime a medida F_1 é obtida através da derivada dos valores da *Precisão* e *Revocação*:

$$F_1 = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação}. \tag{4}$$

Um caso geral da medida F_1 é dada por:

$$F_\beta = \frac{(\beta^2+1)Precisão \times Revocação}{\beta^2 Precisão + Revocação}$$

em que, β é um parâmetro que controla o equilibrio entre a *Precisão* e *Revocação*. O valor deste parâmetro está compreendido entre $0 \leq \beta \leq +\infty$. Quando β assume valor igual a 1, a medida F_1 passa a ser proporcinal à média harmônica da *Precisão* e *Revocação*. Se $\beta > 1$, F_1 torna-se mais orientado à *Revocação*, enquanto que $\beta < 1$, F_1 estará mais orientado para à *Precisão* (SASAKI; FELLOW, 2007).

2.6 ÍNDICE RAND E RAND AJUSTADO

O Índice Rand (IR) e o Índice de Rand Ajustado (IRA), são considerados índices de validação externa, geralmente utilizados para estimar a similaridade entre agrupamentos, isto é, avaliar o grau de semelhança entre partições geradas a partir da mesma base de dados (SOUZA, 2021).

Com o IR é possível fazer comparações entre duas partições distintas sem necessitar que a quantidade de grupos em cada partição sejam iguais. Essa comparação é feita com base na quantidade de pares de pontos concordantes aos dois grupos simultaneamente.

Dado $A = \{\alpha_1, \alpha_2, ..., \alpha_n\}$, um conjunto formado por n elementos, sejam $P = \{C_1, C_2, ...C_k\}$ e $P' = \left\{C'_1, C'_2, ...C'_{k'}\right\}$ duas partições de A em k e k' subgrupos, respectivamente.

Baseando-se em Souza (2021), denota-se a, b, c e d como segue:

$a =$ número de objetos pertencentes aos mesmos subgrupos P e P' simultaneamente;

$b =$ número de objetos que pertencem a subgrupos distintos, mas estão em P e P' simultaneamente;

$c =$ número de objetos que pertence ao mesmo subgrupo de P, e em distintos subgrupos em P';

$d =$ número de objetos que pertencem a distintos subgrupos de P, mas estão no mesmo subgrupo em P'.

Dessa forma, o IR descrito por Rand (1971) é dado por:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{n(n-1)/2},$$

em que $a+b$ representa o total de acordos entre P e P', e $n(n-1)/2$, pode ser interpretado como o número de possibilidades recorrentes sobre o total de pares concordantes entre as duas partições. É possivél deduzir $n(n-1)/2$, como sendo, a probabilidade de um par de pontos escolhidos aleatoriamente serem combinados entre as duas partições. Assim, o IR pode ser calculado como:

$$R = \frac{a+b}{\binom{n}{2}}$$

Verifica-se que o IR exibe valores entre $[0,1]$. Dessa forma, quanto mais próximo de 0 for o valor de R, significa que as partições apresentam certo tipo de divergência entre si. Caso contrário, se R apresenta valores próximos a 1, indica que as partições são parcialmente concordantes, ou totalmente concordantes quando IR for igual a 1.

Por outro lado, o Índice Rand Ajustado assume valores entre $[-1,1]$, onde o valor 1 indica total concordância entre as partições, enquanto que valores iguais ou menores que zero, correspondem a concordância entre partições encontradas ao acaso (MILLIGAN, 1996, apud (FERREIRA, 2006).

Segundo Souza (2021) o IRA é uma versão corrigida do IR. Essa correção foi feita com base num modelo probabilistico envolvendo a permutação dos objetos para gerar novos agrupamentos. Nessa perspectiva, o IRA pode ser escrito da seguinte forma:

$$IRA = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n'_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} - \sum_j \binom{n'_j}{2}\right] - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n'_j}{2}\right]/\binom{n}{2}}$$

em que, n_{ij}, n_i e n_j são obtidos da tabela de contigência $k \times k'[n_{ij}]$, na tabela 2.1, formada pela sobreposição entre P e P', em que cada entrada, n_{ij}, apresentada simboliza o número de objetos em comum a combinação de C_i com C'_j, para todo C_i,$C'_j \in P$ e P', respectivamente, com $n_{ij} = \left|C_i \cap C'_j\right|$.

Tabela 2.2 – Tabela de Contigência.

$P \backslash P'$	C'_1	C'_2	$\ldots$	$C'_{k'}$	somas
C_1	n_{11}	n_{12}	$\ldots$	$n_{1k'}$	n_1
C_2	n_{21}	n_{22}	$\ldots$	$n_{2k'}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_k	n_{k1}	n_{k2}	$\ldots$	$n_{kk'}$	n_k
somas	n'_1	n'_2	$\ldots$	$n'_{k'}$	

2.7 TESTE SHAPIRO-WILK

O teste de Shapiro-Wilk é utilizado para verificar se um determinado conjunto de dados com amostras independentes seguem a distribuíção normal. A suposição de normalidade é um requisito fundamental para a aplicação do teste $t-Student$ e também de outras técnicas estatísticas. O teste de Shapiro-Wilk consiste em verificar as seguintes hipóteses: $\mathcal{H}_0$: os dados seguem a distribuição normal, versus $\mathcal{H}_1$: os dados não seguem a distribuição normal (SHAPIRO; WILK, 1965).

A estatística do teste é obtida através da seguinte expressão:

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

em que y_i é o valor de cada medida independente, $\bar{y}$ é a média de todas as medidas calculadas e as constantes $a_1, a_2, ...a_n$, são obtidas como a solução de

$$(a_1, a_2, ...a_n) = \frac{m^\top V^{-1}}{\left(m^\top V^{-1} V^{-1} m\right)^{\frac{1}{2}}},$$

sendo $m = (m_1, m_2, ..., m_n)^\top$ o vetor dos valores esperados das estatísticas de ordem da amostra e V a matriz de covariância dessas estatísticas (LUCAMBIO, 2008).

Dessa forma, para testar a normalidade dos dados, deve ser calculado, inicialmente, a estatística de W e compará-la com o valor tabelado $W_{n;\alpha}$, acessível em: <http://www.uel.br/projetos/experimental/pages/arquivos/Probabilidades_Shapiro.pdf>. Se o valor de W for menor que o valor tabelado, têm-se evidências para rejeitar a hipótese de normalidade dos dados. Caso contrário, não há evidências para rejeitá-la.

Outra maneira de testar a hipótese de normalidade é calculando o p-valor e compará-lo com o nível de significância α, previamente estabelecido. Se p-valor $< \alpha$, têm-se evidências para rejeitar a hipótese de normalidade dos dados. Ao contrário disto, não há evidências para rejeitá-la.

2.8 TESTE DE WILCOXON

Caso seja rejeitada a suposição de normalidade os dados, uma alternativa ao teste $t-Student$ para amostras independente é o teste não-paramétrico de Wilcoxon, baseado na soma dos postos (ranks) de duas amostras independentes. O teste de Wilcoxon verifica se as medianas das amostras são iguais em situações que a suposição de normalidade não é satisfeita (BARROS; MAZUCHELI, 2005).

Considere $(x_1, x_2, ...x_n)$ uma amostra aleatória proveniente de uma distribuíção simétrica. A estatística de Wilcoxon é descrita da seguinte forma:

$$W = \sum_{i=1}^{n^+} R_i^+ - \frac{n_t(n_t+1)}{4},$$

em que R_i^+ é o posto de $|x_i - \mu_0|$ para $x_i \neq \mu_0$, em que μ_0 representa algum valor especificado de determinado parâmetro de alocação μ, n_t e n_+ são, respectivamente, o número de observações em que $x_i \neq \mu_0$ e $(x_i - \mu_0) > 0$. Em caso de empates, é utilizada a média dos postos, como por exemplo, se $x_1 = x_2$ e o posto de $|x_1 - \mu_0| = 1$ e o posto de $|x_2 - \mu_0| = 2$, então os postos dessas diferenças são respectivamente, $R_1^+ = 1.5$ e $R_2^+ = 1.5$ (BARROS; MAZUCHELI, 2005).

Para $n_t \leq 20$, a distribuição amostral de W é obtida a partir da enumeração de todas as possíveis amostras sob H_0. Por outro lado, a estatística para $n_t > 20$, é dada por:

$$W^* = \frac{W\sqrt{n_t - 1}}{\sqrt{n_t V - W^2}},$$

em que

$$V = \frac{1}{24} n_t(n_t+1)(2n_t+1) - 0.5 \sum_{i=1}^{n^-} t_i(t_i+1)(t_i-1),$$

n representa o número de grupos com empates e t_i é o número de empates no i-ésimo grupo.

2.9 SUPORTE COMPUTACIONAL

Todos os resultados gráficos e numéricos apresentados neste trabalho foram obtidos utilizando o ambiente de programação, análise de dados e gráficos `R`, em sua versão 4.2.0 para sistema operacional Microsoft Windows, que se encontra disponível gratuitamente através do site <http://www.R-project.org>. O `R` foi criado por Ross Ihaka e Robert Gentleman na Universidade de Auckland com o objetivo de produzir um ambiente de programação parecido com `S`, uma linguagem desenvolvida no AT&T Bell Laboratories, cuja versão comercial é o `S-Plus`, tendo as vantagens de ser de livre distribuição e possuir código fonte aberto. Maiores detalhes sobre o `R` podem ser encontrados em Cribari-Neto e Zarkos (1999).

Este trabalho foi digitado utilizando o sistema de tipografia LaTeXdesenvolvido por Leslie Lamport em 1985, que consiste em uma série de macros ou rotinas do sistema TeX(criado por Donald Knuth na Universidade de Stanford) que facilitam o desenvolvimento da edição do texto. Detalhes sobre o sistema de tipografia LaTeXpodem ser encontrados em Lamport (1994) ou através do site <http://www.tex.ac.uk/CTAN/latex>.

3 RESULTADOS

Neste capítulo, são apresentados os resultados da análise exploratória de dados e da aplicação do algoritmo K-means, a fim de avaliar a performance dos diferentes grupos de medidas na separação automática de indivíduos com presença e ausência de lesão na laringe. São aprensentados ainda, resultados dos testes de hipóteses para comparação entre os grupos de pacientes.

3.1 ANÁLISE EXPLORATÓRIA

Para analisar o perfil dos grupos de pacientes com lesão e sem lesão laríngea, foram calculadas a média e desvio padrão para as diferentes medidas analisadas. Os resultados são sumarizados na Tabela 3.1. O objetivo dessa análise é investigar se as variáveis consideradas se distribuem de forma diferenciada em cada um dos grupos. Testes de hipóteses foram realizados a fim de verificar, formalmente, se há diferenças, em média, entre os grupos de pacientes, para cada uma das variáveis. O teste de normalidade de Shapiro-Wilk foi realizado para confirmar a suposição de aderência da distribuição normal aos dados. Isso é necessário para uma escolha adequada do teste a ser utilizado, sendo a normalidade um requisito para aplicação do teste paramétrico t para amostras independentes.

Os testes sugerem que há evidências para rejeitar a normalidade dos dados , ao nível de significância de 5% (p-valores $< 0{,}05$), inviabilizando o uso do teste t. Nesses casos, opta-se por testes livre de distribuição (testes não paramétricos), cujo objetivo é propiciar uma comparação entre os grupos, com base na mediana, sem supor uma distribuição de probabilidade. Dessa forma, foram aplicados testes de Wilcoxon, que é uma alternativa não paramétrica ao teste t. Os resultados são também apresentados na Tabela 3.1.

Para os domínios da escala analógica visual, nota-se que as médias foram superiores para o grupo de pacientes com lesão na laringe, comparativamente aqueles sem lesão. Além disso, os testes de hipóteses indicam que há diferenças entre as medianas dos grupos para essas varíáveis, considerando o nível de significância de 5%. Em relaçao às medidas de autopercepção dos pacientes, apenas para o domínio ESV l foi verificado um valor superior no grupo de pacientes com lesão. Já para os domínios ESV e ESV f, nota-se valores levemente superiores em pacientes sem lesão na laringe. De maneira análoga, foram aplicados testes para avaliar se as diferenças destacadas pelas estatísticas descritivas são significativas estatísticamente.

Como indicado na coluna dos p-valores, apenas para o domínio ESV não foi detectada uma diferença expressiva entre os grupos (p-valor $> 0{,}05$). O mesmo procedimento

foi adotado em relação às medidas acústicas, a fim de identificar aquelas com potencial para distinguir entre os dois grupos de pacientes. De fato, várias medidas, comumentente utilizadas na análise acústica, foram destacadas nos testes de hipóteses quanto a diferença entre as medianas dos grupos analisados. Uma vez que são muitas variáveis, destacam-se as medidas f0 desvio padrão, Jitter, Shimmer, gne, cujos p-valores $< 0{,}001$.

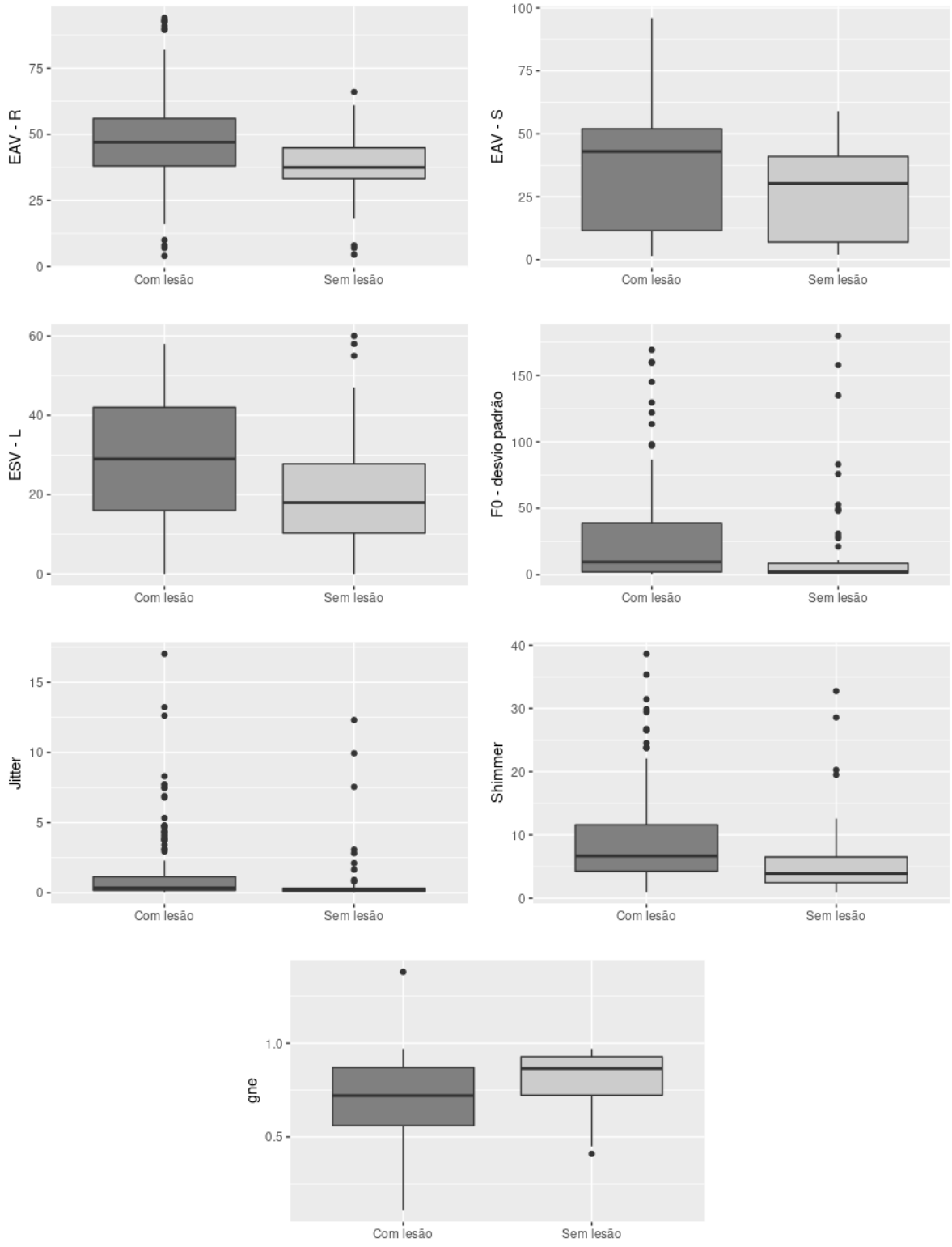
Adicionalmente, foram construídos boxplots para as medidas cujos p-valores $<$ 0,001, a fim de verificar, visualmente, sua distribuição dos grupos de pacientes com lesão e sem lesão. Para as medidas apresentadas nos boxplots, com exceção da medida acústica gne, nota-se um aumento nos valores das medidas para o grupo de pacientes que possuem algum tipo de patologia laríngea. De modo geral, os gráficos também destacaram diversos valores atípicos, ou seja, pacientes com padrões distoantes da distribuição central dos dados. Os resultados apresentados, sugerem que as medidas acústicas e as medidas que compoem as escalas EAV e ESV podem, potencialmente, detectar padrões subjacentes aos dados, mas especificamente agrupar pacientes com presença e ausência de patologias na laringe.

Tabela 3.1 – Estatística descritiva e p-valores do teste de Wilcoxon para as medidas acústicas e das escalas EAV e ESV, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.

Medidas	Presença de Lesão				p-valor
	Com lesão		Sem lesão		
	Média	Desvio Padão	Média	Desvio Padrão	
EAV - GG	53,02	15,50	42,91	11,16	$< 0,001$
EAV - R	48,06	16,41	37,83	11,55	$< 0,001$
EAV - S	37,52	22,81	26,04	18,03	$< 0,001$
EAV - T	33,04	19,53	26,27	17,60	0,0171
ESV - T	49,11	26,41	39,40	24,98	0,0068
ESV - L	29,12	15,68	20,68	14,26	$< 0,001$
ESV - E	9,06	8,93	9,52	13,49	0,4514
ESV - F	12,74	8,82	10,53	6,58	0,0450
F0 - média	185,82	70,06	174,72	50,11	0,7650
F0 - desvio padrão	24,84	33,03	17,29	36,93	$< 0,001$
F1 - média	614,66	163,09	614,22	132,05	0,4774
F1 - desvio padrão	68,97	113,52	65,24	117,28	0,1925
F2 - média	2.050,35	200,96	2.028,47	199,86	0,6404
F2 - desvio padrão	102,12	106,84	87,52	94,66	0,0765
F3 - média	2.869,11	213,27	2.839,10	210,55	0,6054
F3 - desvio padrão	136,19	128,95	120,42	119,05	0,0869
Jitter	1,33	2,40	0,84	2,19	$< 0,001$
Shimmer	8,87	6,87	5,66	5,99	$< 0,001$
gne	0,70	0,21	0,82	0,14	$< 0,001$
tau	15,43	5,63	13,30	4,42	0,0058
m	5,08	1,44	5,19	1,17	0,1120
raio	6,71	2,73	7,51	2,35	0,0103
rec	0,79	0,07	0,82	0,03	0,0385
det	84,69	11,90	84,38	9,83	0,2956
Imed	5,42	1,89	5,10	1,54	0,3067
Imax	205,85	142,42	187,38	78,66	0,9582
entr	1,84	0,37	1,75	0,33	0,0365
lam	77,28	11,26	77,31	12,22	0,5375
tt	3,07	0,46	3,06	0,40	0,5774
vmax	11,71	4,37	12,23	3,75	0,0882
t1	131,37	20,22	118,31	26,51	0,0022
t2	282,96	61,27	260,79	74,77	0,1390
rpde	0,40	0,08	0,42	0,10	0,0114
clust	0,02	0,07	0,01	0,02	0,0010
trans	0,57	0,07	0,55	0,07	0,0100
div	0,01	0,01	0,01	0,00	0,3497
ratio	110,68	22,10	105,35	13,71	0,0188

Fonte: Autor, 2022.

Figura 3.1 – Boxplots das medidas acústicas e das escalas EAV e ESV com p-valores < 0,001 nos testes de Wilcoxon, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.



Fonte: Autor, 2022.

3.2 ALGORITMO K-MEANS

A análise descritiva realizada na seção anterior, permitiu identificar algumas evidências sobre o potencial das medidas das escalas EAV, ESV e medidas acústicas, na caracterização dos grupos de pacientes com lesão e sem lesão laríngea, atendidos pelo LIEV-UFPB. Em particular, os testes de hipóteses sugeriram diferenças significativas entre as medianas dos grupos para várias das medidas analisadas. Embora a análise descritiva permita avaliar o perfil dessas variáveis nos grupos de interesse, são necessários métodos mais adequados, e que permitam, de forma automática, identificar grupos adjacentes aos dados. Nesse sentido, nesta seção serão apresentados os resultados da aplicação do algorítmo K-means, a fim de confirmar a performance dos diferentes grupos de medidas na separação de pacientes com lesão daqueles sem lesão, sem o uso prévio dessa informação.

Para aplicação do algoritmo K-means, foram consideradas duas perspectivas. Na primeira, as medidas acústicas e as medidas das escalas EAV e ESV foram padronizadas. A padronização utilizada foi $z_i = (x_i - \mu)/\sigma$, de modo que todas as variáveis tenham média 0 e variância 1. Esse procedimento é adotado, já que o K-means é influenciado pelas escalas das variáveis. Para cada caso, o algoritmo foi aplicado 100 vezes, e o melhor resultado segundo a função objetivo foi selecionado.

Para comparar o desempenho dos diferentes grupos de medidas, na clusterização de pacientes com e sem lesão laríngea, foram consideradas medidas internas e externas de validação. Especificamente, foram consideradas as medidas externas acurácia, F1, índice de Rand e índice de Rand ajustado e a medida interna coeficiente de silhueta. De maneira geral, quanto maior o valor dessas medidas, melhor a qualidade dos agrupamentos formados. Na Tabela 3.2. Considerando os dados padronizados, nota-se que os melhores desempenhos foram verificados quando todas as medidas são utilizadas na tarefa de agrupamento, ou quando somente as medidas acústicas são utilizadas. Contudo, o ganho em utilizar todas as medidas, comparativamente às medidas acústicas não é expressivo. De fato, os valores estimadados para acurácia, F1, índice de Rand e sua versão ajustada são muito próximas. Por exemplo, considerando a medida F1, o valor estimado para o caso das medidas acústicas foi de 0,7810, enquanto que para o caso em que todas as medidas são utilizados esse valor foi de 0,7842. O mesmo comportamento é observado para as demais medidas.

De maneira análoga, foram calculados os valores das medidas de perfomance para os dados considerados na sua escala original. Contudo, vale destacar que o algoritmo K-means é extremamente sensível as escalas. Para esse cenário, os resultados corroboram com aqueles obtidos para as medidas padronizadas quando consideradas a acurácia e F1. Ainda assim, de modo geral, as medidas padronizadas apresentaram desempenho superior aqueles obtidos na escala original, de modo que a padronização é preferível.

Tabela 3.2 – Medidas de validação externa dos agrupamentos para as medidas acústicas e das escalas EAV e ESV, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.

	Grupo de Medidas	Acurácia	F1	Rand	Rand Ajustado
Padronizadas	EAV	0,5373	0,5755	0,5008	0,0272
	ESV	0,4824	0,6140	0,4987	0,0092
	ACÚSTICAS	0,6745	0,7810	0,5592	0,0679
	TODAS	0,6784	0,7842	0,5620	0,0719
Originais	EAV	0,3686	0,4429	0,5327	0,0512
	ESV	0,4431	0,5749	0,5045	0,0072
	ACÚSTICAS	0,5137	0,6265	0,4984	0,0043
	TODAS	0,5137	0,6265	0,4984	0,0043

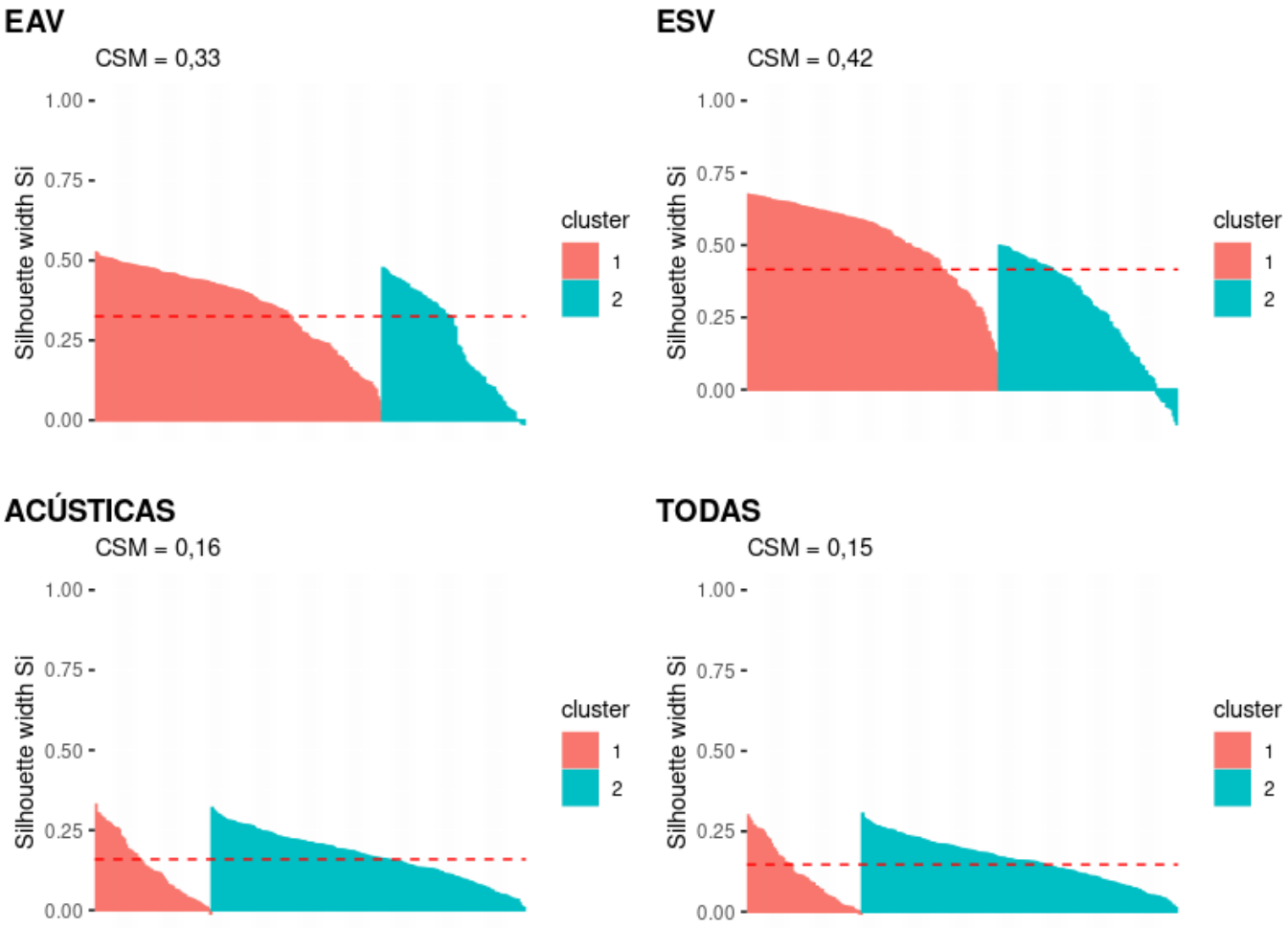
Fonte: Autor, 2022.

Para visualizar os agrupamentos formados, foram construídos gráficos no espaço bidimensional. Para tanto, foi utilizado a técnica de Análise de Componentes Principais (ACP) para reduzir a dimensão de cada grupo de medidas, a fim de permitir sua visualização. Os resultados são aprensentados na Figura 3.2. As diferentes medidas utilizadas resultam em estruturas diferentes de clusterização. Contudo, os gráficos para as medidas acústicas e considerando todas as medidas apresentam estruturas similares. De fato, os resultados das medidas indicam o desempenho similar desses grupos de medidas. Adicionalmente, foram calculados os coeficientes de silhueta para cada conjunto de medidas, além do gráfico de silhueta, a fim de avaliar o desempenho dos agrupamentos (Figura 3.3). De modo geral, o coeficiente de silhueta médio para os diferentes grupos de medidas foi baixo, já que quanto mais próximo de 1, melhor a qualidade dos agrupamentos. Além disso, os melhores resultados, quando analisado o índice interno, foram verificados para as medidas das escalas EAV (CSM = 0,33) e ESV (CSM = 0,42), quando comparados as medidas acústicas (CSM = 0,16) e todas as medidas (CSM = 0,15). Os baixos valores associados ao coeficiente de silhueta é devido a alocação incorreta de observações nos respectivos clusters. Esse fato é observado nos gráficos, em que o coeficiente de silhueta indivivual é negativo.

Figura 3.2 – Gráficos dos agrupamentos K-means para as medidas acústicas e das escalas EAV e ESV padronizadas, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.

Fonte: Autor, 2022.

Figura 3.3 – Gráficos de silhueta para as medidas acústicas e das escalas EAV e ESV padronizadas, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.

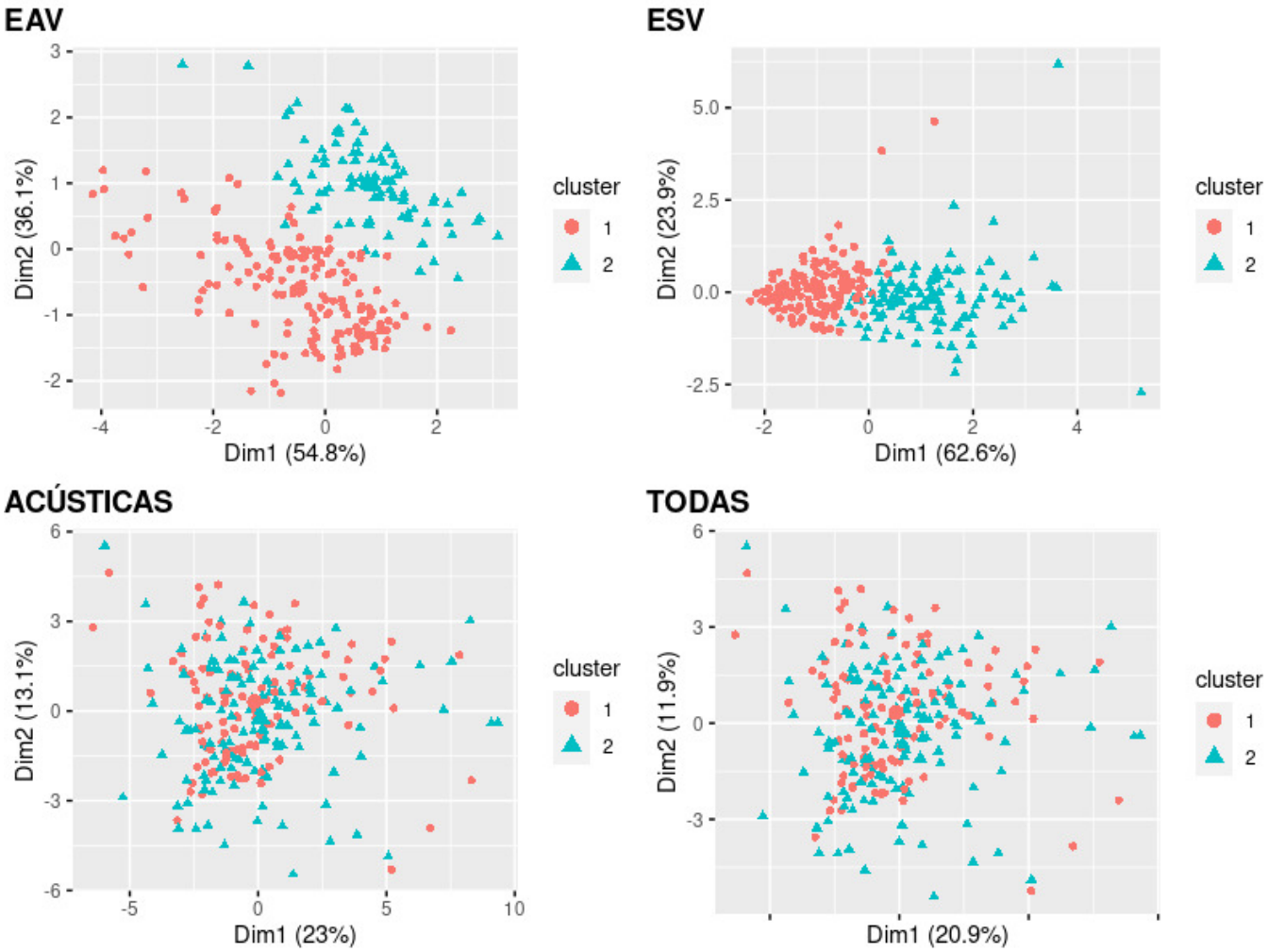


Fonte: Autor, 2022.

No sentido de avaliar também a qualidade interna dos clusters formados através das medidas acústicas, e das escalas EAV e ESV na escala original das variáveis, foram apresentados os gráficos dos agrupamentos formados (Figura 3.4), com base no método ACP, e também os gráficos de silhueta (Figura 3.5). Os gráficos da Figura 3.4 ilustram os agrupamentos para cada um dos grupos de medidas analisados. Os clusters formados com base nas medidas da escala ESV não padronizadas é similar aquele encontrado no caso padronizado. Para os demais casos, observa-se que os agrupamentos são distintos, quando comparados as suas versões padronizadas, principalmente para as medidas acústicas e considerando todas as medidas, cujos agrupamentos formados foram sobrepostos. Com relação ao coeficiente médio de silhueta, as medidas das escalas EAV e ESV apresentaram os maiores coeficientes, 0,40 e 0,46, respectivamente. Já para as medidas acústicas e considerando todas as medidas para a tarefa de agrupamento, tais valores foram inferiores,

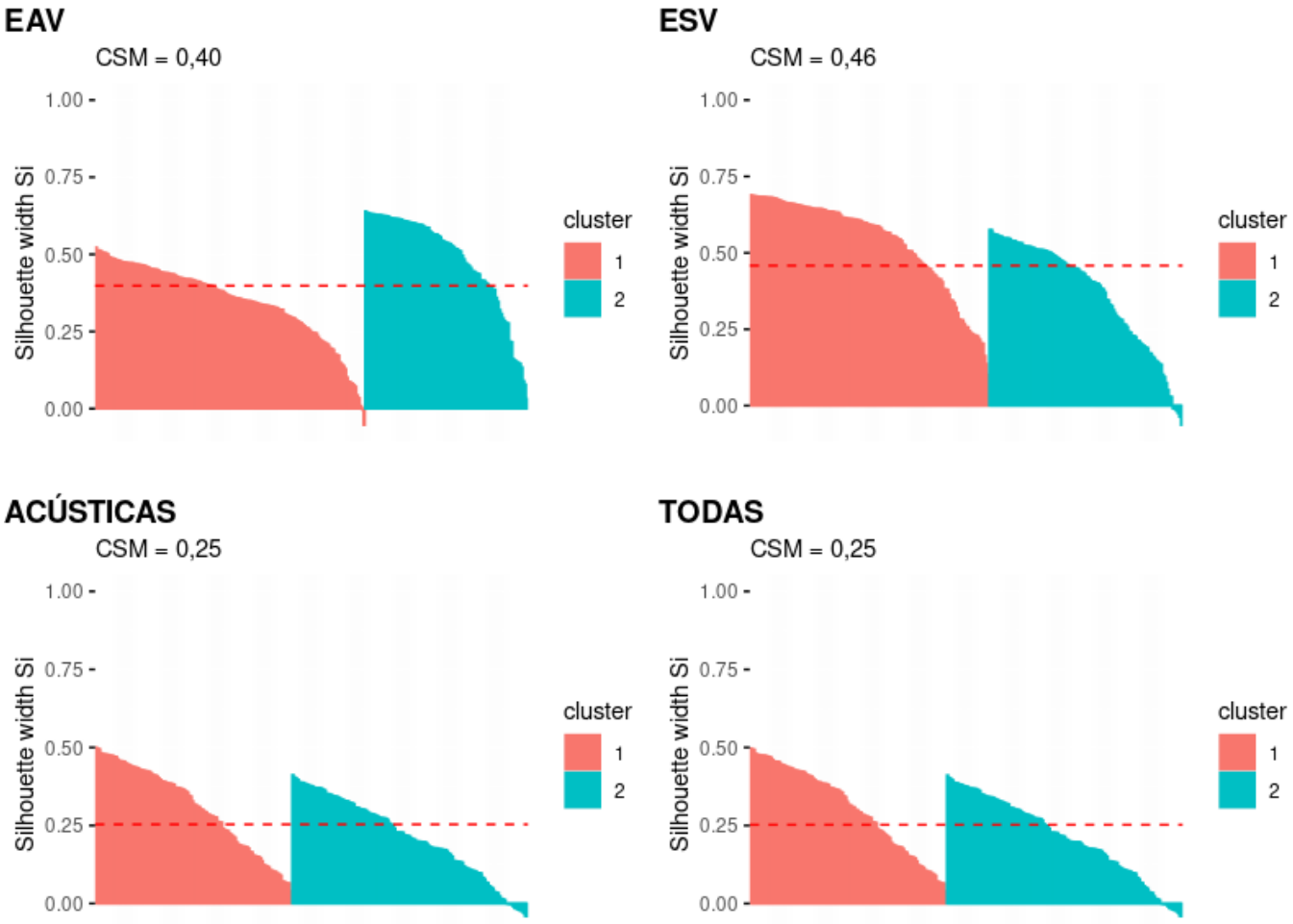
cujas estimativas são 0,25 em ambos os casos. Esses resultados também corroboram com aqueles obtidos para as medidas padronizadas, em que as escalas EAV e ESV apresentaram clusters com coeficientes médios maiores. Embora os coeficientes de silhueta tenham indicado grupos de medidas distintos daqueles observados na análise de medidas externas, essas últimas fornecem resultados mais concisos, já que na sua concepção são utilizadas informações à priori sobre o verdadeiro agrupamento. O uso no presente trabalho foi apenas no sentido de avaliar se as indicações fornecidas por essas duas abordagens resultariam na escolha do mesmo grupo de medidas. Vale ressaltar ainda, que o coeficiente de silhueta é extremamente importante em situações práticas de análise de agrupamento, já que não há nenhuma informação prévia, favorecendo assim, a escolha de agrupamentos mais homogênos. Além disso, configura-se como uma técnica para a seleção do número adequado de clusters, que não foi um problema abordado no trabalho, já que o número de grupos é fixado, estando associado a presença ou ausência de patologias nos pacientes analisados.

Figura 3.4 – Gráficos dos agrupamentos K-means para as medidas acústicas e das escalas EAV e ESV originais, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.



Fonte: Autor, 2022.

Figura 3.5 – Gráficos de silhueta para as medidas acústicas e das escalas EAV e ESV originais, para pacientes com e sem lesão laríngea, atendidos no LIEV-UFPB.



Fonte: Autor, 2022.

De maneira geral, os resultados apresentados permitiram analisar o perfil dos agrupamentos formados pelo uso de diferentes medidas, utilizadas no contexto da análise vocal, mais especificamente para detecção dos grupos de paciente com lesão e sem lesão laríngea. Em particular, os resultados sugerem o grupo de medidas mais adequada para essa tarefa, em contextos em que não há uma classificação de um especialista. De modo geral, os resultados favorecem o uso da medidas acústicas, ou o uso combinados das medidas com aquelas das escalas EAV e ESV.

CONSIDERAÇÕES FINAIS

Neste trabalho objetivou-se analisar a performance de diferentes medidas acústicas na detecção de padrões em dados vocais. Mais especificamente, foram analisados dados provenientes de um projeto de pesquisa desenvolvido pelo Laboratório Integrado de Estudos da Voz da Universidade Federal da Paraíba. Foram considerados vários parâmetros para analisar os desvios vocais como, medidas acústicas, medidas das escalas de sintomas vocais e medidas da escala analógica visual.

Para analisar a performance dos diferentes grupos de medidas consideradas, foram, calculadas inicialmente, algumas medidas descritivas, para os grupos de pacientes com e sem lesão na laringe. Adicionalmente, foram realizados testes de hipóteses, para comparar a mediana das medidas utilizadas entre os grupos. O teste de Wilcoxon foi utilizado com esse propósito, já que o teste de Shapiro-Wilk evidencou a rejeição da hipótese de normalidade dos dados. Os resultados sugerem que várias medidas apresentam potencial para separação dos dados nos grupos de pacientes com e sem lesão na laringe.

O método K-means foi aplicado aos diferentes grupos de medidas para confirmar o poder de separação dessas medidas. Para tanto, foram calculados índices externos e um índice interno, para avaliar a qualidade dos agrupamentos formados, e consequentemente, das medidas utilizadas. De maneira geral, as medidas acústicas apresentaram o melhor desempenho com base nesses índices, quando comparado ao uso das demais medidas das escalas EAV e ESV, bem como do uso combinado dessas medidas, já que o ganho não foi expressivo.

Dessa forma, sugere-se o uso do método K-means associado às medidas acústicas para detecção de padrões em dados vocais, especificamente para agrupamento de pacientes com e sem lesão laríngea, em situações em que não é possível a classificação supervisionada de um especialista. Vale destacar que as análises apresentadas não esgotam todas as possibilidades, de modo que outras combinações entre algoritmos e diferentes medidas utilizadas no contexto da análise vocal, e na detecção de patologias laríngeas, devem ser investigadas.

REFERÊNCIAS

AGGARWAL, C. C.; REDDY, C. K. **Data Clustering: Algorithms and Applications**. 1st. ed. [S.l.]: Chapman & Hall/CRC, 2014. ISBN 1466558210.

BARROS, E. A. C.; MAZUCHELI, J. Um estudo sobre o tamanho e poder dos testes t-student e wilcoxon. **Acta Scientiarum. Technology**, Universidade Estadual de Maringá, v. 27, n. 1, p. 23–32, 2005.

BEHLAU, M.; PONTES, P.; MORETI, F. **Higiene vocal: cuidando da voz**. [S.l.]: Thieme Revinter Publicações LTDA, 2018.

BYEON, H. Model development for predicting the occurrence of benign laryngeal lesions using support vector machine: focusing on south korean adults living in local communities. **Int. J. Adv. Comput. Sci. Appl**, v. 9, p. 222–227, 2018.

CAPUCHO, M. C. P. avaliação multidimensional na voz profissional. 2018.

CASTRO, A. A. M. D.; PRADO, P. P. L. D. Algoritmos para reconhecimento de padrões. **Revista Ciências Exatas**, v. 8, n. 2002, 2002.

CHIMIESKI, B. F.; FAGUNDES, R. D. R. Association and classification data mining algorithms comparison over medical datasets. **Journal of health informatics**, v. 5, n. 2, 2013.

CRIBARI-NETO, F.; ZARKOS, S. G. **R: Yet another econometric programming environment**. [S.l.]: Wiley Online Library, 1999.

FERREIRA, C. Designing neural networks using gene expression programming. In: **Applied Soft Computing Technologies: The Challenge of Complexity**. Springer Berlin Heidelberg, 2006, (Advances in Soft Computing, v. 34). p. 517 – 535. ISBN 978-3-540-31649-7. Disponível em: <http://dx.doi.org/10.1007/3-540-31662-0_40>.

FERREIRA, M. R. P. Agrupamento baseado em kernel com ponderação automática das variáveis via distâncias adaptativas. Universidade Federal de Pernambuco, 2013.

FREITAS, S. Correlação entre a avaliação acústica e perceptual na caracterização de vozes patológicas. **Porto (Portugal): Universidade do Porto**, 2010.

GORDON, A. D. **Classification**. [S.l.]: CRC Press, 1999.

GUIMARÃES, I. A ciência e a arte da voz humana. **Alcoitão, Escola Superior de Saúde de Alcoitão**, 2007.

GUIMARÃES, V. d. C.; VIANA, M. A. d. D. E. S. R.; BARBOSA, M. A.; PAIVA, M. L. d. F.; TAVARES, J. A. G.; CAMARGO, L. A. d. Cuidados vocais: questão de prevenção e saúde. **Ciência & saúde coletiva**, SciELO Public Health, v. 15, p. 2799–2803, 2010.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys (CSUR)**, Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data: An introduction to cluster analysis. a wiley and sons. **New York**, 1990.

LAMPORT, L. **LATEX : a document preparation system : user's guide and reference manual**. Reading, Mass.: Addison-Wesley Pub. Co., 1994. – p. ISSN 0201529831 9780201529838. Disponível em: <http://www.worldcat.org/oclc/29225162>.

LUCAMBIO, F. Diferentes testes para verificar normalidade de uma amostra aleatória. **Statistic Research of Paraná. e**, v. 1, p. 1–12, 2008.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MONICO, J. F. G.; POZ, A. P. D.; GALO, M.; SANTOS, M. C. D.; OLIVEIRA, L. C. D. Acurácia e precisão: revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, Universidade Federal do Paraná, v. 15, n. 3, p. 469–483, 2009.

MONTALVO, S.; FRESNO, V.; MARTÍNEZ, R. Nesm: A named entity based proximity measure for multilingual news clustering. **Procesamiento del lenguaje natural**, Sociedad Española para el Procesamiento del Lenguaje Natural, n. 48, p. 81–88, 2012.

NETO, A. F. G.; JR, S. B.; GUIDO, R. C. Reconhecimento de vogais para identificação de patologias baseado em redes neurais artificiais. 2012.

REDDY, C. K. **Data Clustering: Algorithms and Applications**. [S.l.]: Chapman and Hall/CRC, 2018.

REGATIERI, K. F. et al. Jogo"hospital mirim"como facilitador para o enfrentamento do procedimento invasivo. Universidade Federal de Mato Grosso, 2018.

ROUSSEEUW, P. J. Silhuetas: um auxílio gráfico para a interpretação e validação da análise de cluster. **Jornal de matemática computacional e aplicada**, v. 20, 1987.

SASAKI, Y.; FELLOW, R. The truth of the f-measure, manchester: Mib-school of computer science. **University of Manchester**, p. 25, 2007.

SELIM, S. Z.; ISMAIL, M. A. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 1, p. 81–87, 1984.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

SILVA, M. A. B. da. **Modelos Neuro-Evolucionários de Redes Neurais Spiking Aplicados ao Pré-Diagnóstico de Envelhecimento Vocal**. Tese (Doutorado) — PUC-Rio, 2014.

SILVA, S. S. L. da. Main laryngeal pathologies in teachers. 2018.

SODRÉ, B. R. et al. **Reconhecimento de padrões aplicados à identificação de patologias de laringe**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2016.

SOUZA, D. C. d. **Análise de agrupamentos para dados espaciais: estudo de simulação e aplicação a dados educacionais do Estado do Paraná**. Tese (Doutorado) — Dissertação de Mestrado, Universidade Federal do Paraná, Curitiba, 2021.

SOUZA, E. F. d. **Comparação e escolha de agrupamentos: uma proposta utilizando a entropia**. Tese (Doutorado) — Universidade de São Paulo, 2007.

TAKAKURA, A. M.; PEREIRA, D. R.; SILVA, F. A. da; PAZOTI, M. A.; ALMEIDA, L. L. de; SAPIA, H. M. Uso do aprendizado de máquina no diagnóstico médico de patologias. In: **Colloquium Exactarum. ISSN: 2178-8332**. [S.l.: s.n.], 2018. v. 10, n. 1, p. 78–89.

TEIXEIRA, J. P.; FERREIRA, D.; CARNEIRO, S. M. Análise acústica vocal-determinação do jitter e shimmer para diagnóstico de patalogias da fala. In: INEGI. **6º Congresso Luso-Moçambicano de Engenharia, 3º Congresso de Engenharia de Moçambique**. [S.l.], 2011.

XU, R.; WUNSCH, D. Survey of clustering algorithms. **IEEE Transactions on neural networks**, Ieee, v. 16, n. 3, p. 645–678, 2005.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Cajazeiras
Rua José Antônio da Silva, 300, Jardim Oásis, CEP 58.900-000, Cajazeiras (PB)
CNPJ: 10.783.898/0005-07 - Telefone: (83) 3532-4100

Documento Digitalizado Restrito

Entrega de Trabalho Conclusão de Curso

Assunto:	Entrega de Trabalho Conclusão de Curso
Assinado por:	Wellington Almeida
Tipo do Documento:	Anexo
Situação:	Finalizado
Nível de Acesso:	Restrito
Hipótese Legal:	Informação Pessoal (Art. 31 da Lei no 12.527/2011)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- Wellington Ferreira de Almeida, ALUNO (201712020035) DE LICENCIATURA EM MATEMÁTICA - CAJAZEIRAS, em 23/05/2022 19:26:42.

Este documento foi armazenado no SUAP em 23/05/2022. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 525212
Código de Autenticação: 4a4fd0f2ec