UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA – PPGEEL

RAFAEL KINGESKI

SUBCONJUNTOS GAUSSIANOS EM SISTEMAS DE RECONHECIMENTO DE EMOÇÕES EM SINAIS DE VOZ

JOINVILLE
2025

RAFAEL KINGESKI

SUBCONJUNTOS GAUSSIANOS EM SISTEMAS DE RECONHECIMENTO DE EMOÇÕES EM SINAIS DE VOZ

Tese apresentada ao Programa de Pós–Graduação em Engenharia Elétrica do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Doutor em Engenharia Elétrica.

Orientador: Aleksander S. Paterno
Coorientador: Elisa Henning

JOINVILLE
2025

Ficha catalográfica elaborada pelo programa de geração automática da

Biblioteca Universitária Udesc,

com os dados fornecidos pelo(a) autor(a)

Kingeski, Rafael
Subconjuntos Gaussianos em Sistemas de Reconhecimento de Emoções em Sinais de Voz / Rafael Kingeski. -- 2025.
111 p.

Orientador: Aleksander S. Paterno
Coorientadora: Elisa Henning
Tese (doutorado) -- Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Engenharia Elétrica, Joinville, 2025.

1. voz. 2. emoções. 3. reconhecimento de emoções. 4. PCA. 5. ICA. I. Paterno, Aleksander S.. II. Henning, Elisa. III. Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação em Engenharia Elétrica. IV. Titulo.

RAFAEL KINGESKI

SUBCONJUNTOS GAUSSIANOS EM SISTEMAS DE RECONHECIMENTO DE EMOÇÕES EM SINAIS DE VOZ

Tese apresentada ao Programa de Pós–Graduação em Engenharia Elétrica do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, como requisito parcial para a obtenção do grau de Doutor em Engenharia Elétrica.

Orientador: Aleksander S. Paterno
Coorientador: Elisa Henning

BANCA EXAMINADORA:

Prof. Dr. Aleksander Sade Paterno
Orientador/Presidente

Membros:

Profa. Dra. Anna Alice Figueiredo de Almeida
Universidade Federal da Paraíba - UFPB

Prof. Dr. Miguel Arjona Ramírez
Universidade de São Paulo - USP

Prof. Dr. Paulo Rogério Scalassara
Universidade Tecnológica Federal do Paraná - UTFPR

Prof. Dr. Yuri Kaszubowski Lopes
Universidade do Estado de Santa Catarina - UDESC

Joinville, 17 de dezembro de 2024

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha profunda gratidão aos meus pais, cujo amor e apoio incondicional me acompanharam em todas as etapas da minha vida.

Agradeço sinceramente ao meu orientador, Professor Dr. Aleksander Sade Paterno, por sua orientação valiosíssima e por acreditar em mim ao longo deste percurso. Sua abordagem rigorosa, por vezes desafiadora, foi fundamental para meu crescimento acadêmico e pessoal. As críticas construtivas e o alto padrão que ele exigiu de mim me levaram a alcançar resultados que, de outra forma, eu não teria conseguido. Cada conversa e troca de ideias foram momentos enriquecedores que moldaram a direção da minha pesquisa e ampliaram minha visão sobre o tema.

Um agradecimento especial à minha coorientadora, Professora Dra. Elisa Henning, que se juntou à minha jornada após a metade do doutorado. Sua presença foi um verdadeiro divisor de águas, trazendo não apenas uma nova perspectiva, mas também um suporte essencial em momentos críticos. Estou imensamente grato por sua dedicação e pelas valiosas contribuições que enriqueceram minha pesquisa.

Sou grato também aos meus amigos e familiares, que sempre estiveram ao meu lado, oferecendo suporte e compreensão.

RESUMO

Os sistemas de reconhecimento de emoções a partir da voz tem como objetivo identificar e interpretar as emoções expressas por meio de características acústicas da fala. Essa tecnologia é fundamental para aprimorar a interação entre humanos e máquinas, com aplicações em áreas como saúde, onde pode auxiliar no monitoramento emocional, e em sistemas de assistência inteligente, melhorando a resposta e a adaptação dos sistemas às necessidades dos usuários. Este estudo propõe um método de reconhecimento de emoções em sinais de voz que considera a distribuição dos parâmetros. Ao isolar subconjuntos de parâmetros com distribuição gaussiana, busca-se verificar a influência da distribuição de parâmetros de voz aplicada em técnicas de redução por transformação. Uma metodologia em múltiplas etapas é aplicada, com filtragem inicial baseada em variância, seguida do uso do teste de Kruskal-Wallis para remover dados que não contribuem para a classificação e por fim teste de Anderson-Darling para gerar um subconjunto conforme a distribuição. Parâmetros com distribuição normal são reduzidos pela Análise de Componentes Principais (PCA do inglês *principal component analysis*), enquanto o conjunto todo de parâmetros é reduzido usando a Análise de Componentes Independentes (ICA do inglês *independent component analysis*). A fusão dessas transformações foi utilizada para a classificação com Máquinas de Vetores de Suporte (SVM do inglês support vector machine). O estudo destaca a importância da seleção de parâmetros fundamentada em teorias estatísticas, como uma alternativa simples e eficiente que dispensa o uso de abordagens automatizadas e orientadas por algoritmos iterativos que dependem de modelos. Os resultados demonstram que a segmentação em subconjuntos gaussianos, combinada com PCA e ICA, aprimora a acurácia dos sistemas de reconhecimento de emoções. Para validar o estudo utilizaram-se três bases de dados, EMODB, RAVDESS e SAVEE. Uma melhora na acurácia foi obtida utilizando a fusão de PCA e ICA; além disso, os resultados mostram que dados que possuem distribuição gaussiana têm uma maior contribuição para o modelo.

Palavras-chave: Voz. Emoções. Reconhecimento de Emoções. Seleção de parâmetros. PCA. ICA.

ABSTRACT

Speech emotion recognition systems aim to identify and interpret emotions expressed through the acoustic features of speech. This technology is essential for improving human-machine interaction, with applications in areas such as healthcare, where it can assist in emotional monitoring, and in intelligent assistance systems, enhancing system responsiveness and adaptation to users' needs. This work proposes a speech emotion recognition method that considers the distribution of features. By isolating subsets of features with a Gaussian distribution, the aim is to verify the influence of voice parameter distribution when applying transformation-based dimensionality reduction techniques. A multi-step methodology is employed, starting with variance-based filtering, followed by the Kruskal-Wallis test to remove features that do not contribute to classification, and finally, the Anderson-Darling test to generate a subset according to its distribution. Parameters with a normal distribution are reduced using Principal Component Analysis (PCA), while the entire set of parameters is reduced using Independent Component Analysis (ICA). The fusion of these reduced feature sets was used for classification with Support Vector Machines (SVM). The study highlights the importance of parameter selection based on statistical properties as a simple and efficient alternative to automated approaches guided by iterative, model-dependent algorithms. The results show that segmentation into Gaussian subsets, combined with PCA and ICA, enhances the accuracy of emotion recognition systems. To validate the study, three databases were used: EMODB, RAVDESS, and SAVEE. An improvement in accuracy was achieved using the fusion of PCA and ICA, and the results also show that data with a Gaussian distribution make a greater contribution to the model.

Keywords: Speech. Emotion. Speech Recognition Systems. Features Selection. PCA. ICA.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SRE	Sistema de Reconhecimento de Emoções
PCA	Análise de Componentes Principais (do inglês *Principal Component Analysis*)
ICA	Análise de Componentes Independentes (do inglês *Independent Component Analysis*)
LDA	Análise Discriminante Linear (do inglês *Linear discriminant Analysis*)
t-SNE	Incorporação de Vizinhos Estocásticos Distribuídos em t (do inglês *t-Distributed Stochastic Neighbor Embedding*)
KW	*Kruskal-Wallis*
SFS	Seleção de Característica Sequencial (do inglês *Sequential Feature Selection*)
SFFS	Seleção Sequencial para Frente Flutuante (do inglês *Sequential Forward Feature Selection*)
Relief-F	Algoritmo de seleção (do inglês *Repeated Local Feature Importance with Filtering*)
GWO	Otimização do Lobo Cinzento (do inglês *Grey Wolf Optmizer*)
LPC	Codificação Linear Preditiva (do inglês *Linear Prection Coding*)
MFCC	Coeficientes mel-cepestrais (do inglês *Mel-frequency cepstral coefficients*)
SVM	Máquina de vetor de suporte (do inglês *Support Vector Machine*)
FFT	Transformada rápida de Fourier (do inglês *Fast Fourier Transform*)
PLP	Predição linear perceptual (do inglês *Perceptual Linear Predictive*)
HNR	Razão Harmônico-Ruído (do inglês *Harmonic-to-Noise Ratio*)
DCT	Transformada discreta de cosseno (do inglês *Discrete cossine transform*)
ANOVA	Análise de Varância (do inglês *analysis of variance*)
DNN	Rede Neural Profunda (do inglês *deep neural network*)
KNN	k-Vizinhos Mais Próximos (do inglês *k-nearest neighbors*)
RASTA	Filtro espectral relativo usado para supressão de ruído e realce de sinais de áudio (do inglês *RelAtive SpecTrAl*)
PCM	Modulação por código de pulso (do inglês *pulse code modulation*)

LISTA DE SÍMBOLOS

$x(m)$	Sinal discreto no tempo m
$w(m)$	Janela no tempo discreto m
$h(m)$	Janela ao quadrado utilizada para segmentação do sinal no tempo m
$\mathscr{E}_n$	Energia no tempo discreto n
$s[n]$	Sinal de voz no tempo discreto n
$\tilde{s}[n]$	Sinal previsto de voz no tempo discreto n
$u[n]$	Sinal de excitação no tempo discreto n
G	Parâmetro de ganho
$S(z)$	Sinal de voz no domínio z
$E(z)$	Erro de predição no domínio z
$H(z)$	Função de transferência no domínio z
$P(z)$	Polinômio preditor no domínio z
$A(z)$	Função de transferência do erro pela entrada
a_k	Coeficientes do filtro digital de polos que representam os parâmetros do sistema do trato vocal
α_k	Coeficientes de predição
$E_{\hat{n}}$	Soma total do erro de predição
F_0	Frequência fundamental da voz
$\mathbf{Z_{PCA}}$	Matriz de componentes principais
$\mathbf{W_P}$	Matriz de transformação de PCA
$\mathbf{C_X}$	Matriz de covariância
$\mathbf{J_{PCA}}$	Função objetivo
$\mathbf{S_{ICA}}$	Matriz de componentes independentes
$\mathbf{A}$	matriz de mistura dos sinais independentes
γ	Parâmetro do *kernel* RBF que controla o alcance da influência dos exemplos de treinamento
$\mathbf{U}$	Matriz de autovetores ou matriz de transformação de PCA
$\mathbf{W_I}$	Matriz de transformação de ICA
$J(\mathbf{x})$	Negentropia do vetor de dados $\mathbf{x}$
$g(\cdot)$	função não linear aplicada no algoritmo FastICA

$\mathscr{H}(\mathbf{x})$	Entropia do vetor de dados $\mathbf{x}$
R_j	Soma dos postos atribuídos às observações do grupo j
$R_{\cdot j}$	Média dos postos das observações do grupo j
H	Estatística de Kruskal-Wallis
H_0	Hipótese nula
H_1	Hipótese alternativa
A	Estatística do teste de Anderson–Darling, usada para avaliar a aderência de uma amostra a uma distribuição específica
ζ_i	Dados do teste de Anderson–Darling
$F(\cdot)$	Função de distribuição acumulada
$\mathbf{w}$	Vetor de pesos do SVM
$d(\mathbf{x}, \mathbf{w}, b)$	Função de decisão do SVM, que determina a separação entre as classes
b	Viés da função de decisão do SVM
C	Parâmetro de penalidade de margem do SVM
ξ_i	Variáveis de folga que medem erros de classificação do SVM
K	Função de kernel do SVM
c_0	Parâmetro de ajuste da função de *kernel* polinomial
d	Grau do polinômio na função de *kernel* polinomial
vp_i	Verdadeiros positivos para a i-ésima classe
fp_i	Falsos positivos para a i-ésima classe
fn_i	Falsos negativos para a i-ésima classe
vn_i	Verdadeiros negativos para a i-ésima classe
b	Fator de ponderação do F-*score*
$\bar{x_j}$	Média dos valores da coluna de X
σ	Desvio padrão
$\mathbf{X}$	Matriz de parâmetros acústicos de voz
$\mathbf{Y}$	Matriz coluna de rótulos
$\mathbf{X}_{\mathrm{norm}}$	Matriz de parâmetros normalizados
$\mathbf{X_s}$	Matriz de parâmetros padronizados
$\mathbf{X_n}$	Matriz de parâmetros com distribuição normal
$\mathbf{X_{KW}}$	Matriz de parâmetros filtrados pelo teste de Kruskal-Wallis
α	Significância dos testes estatísticos

SUMÁRIO

1 INTRODUÇÃO

Atualmente, diversos dispositivos inteligentes coletam continuamente dados fisiológicos dos seres humanos, como frequência cardíaca, contagem de passos, períodos de repouso, níveis de oxigenação sanguínea e gasto calórico durante as atividades diárias. Agora, imagine se um desses dispositivos pudesse interpretar suas emoções apenas pela sua voz. Pode-se perguntar: "Qual seria a real necessidade de um dispositivo desse tipo?". A resposta não é única, mas um dispositivo poderia monitorar suas emoções em diferentes ambientes e situações cotidianas. Além disso, esses dados poderiam ser armazenados e consultados da mesma forma que outros dados fisiológicos, permitindo um maior controle sobre a saúde e promovendo o autoconhecimento. A identificação de emoções poderia se tornar um dado valioso para o gerenciamento da saúde mental, exemplificando apenas uma das inúmeras aplicações possíveis de um Sistema de Reconhecimento de Emoções (SRE).

A aplicação de técnicas de reconhecimento de emoções tem se mostrado uma abordagem promissora para a área da saúde, especialmente no auxílio ao diagnóstico e monitoramento de transtornos psicológicos e neurodegenerativos. Estudos recentes demonstram que características da fala e expressões faciais podem ser utilizadas para detectar e diferenciar sintomas como depressão, ansiedade e apatia, contribuindo para um diagnóstico mais preciso e um acompanhamento clínico mais objetivo (Yang et al., 2023). Além disso, a análise linguística de narrativas emocionalmente carregadas tem revelado padrões distintos no discurso de indivíduos com depressão, permitindo a identificação de marcadores linguísticos associados à gravidade dos sintomas (Zhou et al., 2023). Essas abordagens não apenas aprimoram a detecção precoce e a personalização do tratamento, mas também viabilizam soluções de monitoramento remoto, ampliando o acesso a cuidados de saúde mental. Assim, o reconhecimento de emoções é uma ferramenta com potencial relevante para a interseção entre tecnologia e saúde.

Segundo Cowie et al. (2001), a aplicação comercial de maior relevância é em jogos que respondem ao estado emocional dos usuários. Outros estudos também indicam que o reconhecimento de emoções pode ser aplicado em:

a) Sistemas de tutores inteligentes, os quais se fundamentam na teoria da aprendizagem e da cognição para imitar um professor humano (Petrovica; Anohina-Naumeca; Ekenel, 2017);

b) Síntese de voz com emoções, que pode ser utilizada em sistemas de realidade virtual, tornando a experiência do usuário mais natural (Li et al., 2022);

c) Avaliação da qualidade de atendimento (Galanis et al., 2013; Deschamps-Berger; Lamel; Devillers, 2021).

Produzir um sistema capaz de detectar emoções em usuários é um dos principais objetivos do campo de pesquisa chamado computação afetiva (Picard, 1995). Para realizar o reconheci-

mento de emoções em sinais de voz, é necessário dispor de uma base de dados com gravações de vozes emocionalmente etiquetadas.

Uma base de dados destinada ao reconhecimento de emoções deve passar por um processo de validação conduzido por humanos para garantir a confiabilidade das emoções rotuladas nos dados. Estudos anteriores destacam a importância desse procedimento, em Burkhardt et al. (2005), os dados foram validados por avaliadores humanos que ouviram os áudios da base e atribuíram rótulos emocionais com base em sua percepção das emoções transmitidas. No estudo de Haq, Jackson e Edge (2008), a validação dos dados audiovisuais foi realizada por um grupo de 10 avaliadores, assegurando a confiabilidade das emoções rotuladas. O estudo de Livingstone, Russo e Najbauer (2018) utilizou avaliadores para rotular os áudios, verificando a categoria, intensidade e autenticidade das emoções.

Portanto, validar uma base de dados por humanos é um passo metodológico essencial que assegura a qualidade dos rótulos emocionais atribuídos, garantindo que os dados sejam representativos e confiáveis para modelos de aprendizado de máquina.

Um sistema de reconhecimento de emoções pode ser implementado em quatro etapas (Xie; Zhu; Hu, 2023): (1) pré-processamento dos sinais de voz, (2) extração de parâmetros, (3) seleção e fusão de parâmetros e (4) classificação. De forma simplificada, como mostra o diagrama da Figura 1, a base de dados de voz é dividida em grupo de treinamento e grupo de teste. Esse procedimento avalia um classificador reservando uma parte dos dados para verificar a acurácia do modelo posteriormente (Shahin et al., 2023). Outras formas de validação também são empregadas. A validação cruzada, por exemplo, foi utilizada em Xie, Zhu e Hu (2023) e Özseven (2019). Outra abordagem consiste em excluir os dados de um locutor durante o treinamento e utilizá-los exclusivamente para a validação, como descrito em (Xu et al., 2024).

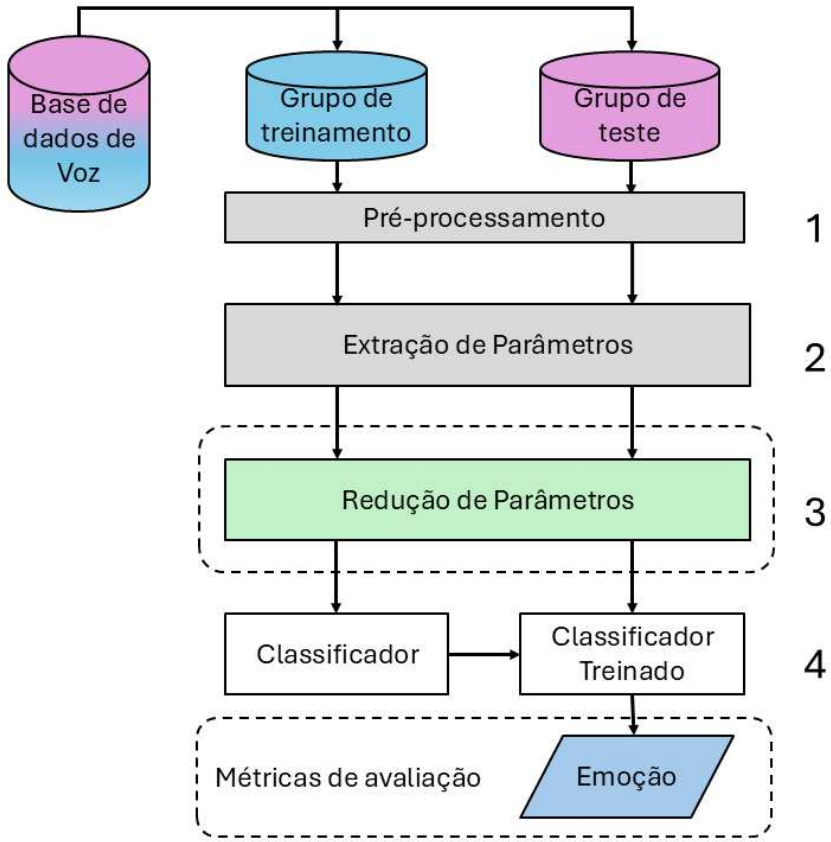
Com técnicas de processamento computacional, muitos SREs foram desenvolvidos utilizando parâmetros acústicos extraídos de ferramentas como:

a) PRAAT (Boersma; Weenink, 2023), um *software* com interface gráfica utilizado para análise e síntese de parâmetros acústicos da voz. Ele oferece ferramentas para manipulação, visualização e extração de parâmetros como *pitch*, formantes, intensidade e duração, sendo amplamente empregado em estudos sobre emoções na voz (Meftah et al., 2020; Wu; Liang, 2011);

b) openSMILE (Eyben; Wöllmer; Schuller, 2010), uma ferramenta para extração de parâmetros acústicos da voz. Permite a extração de características prosódicas, espectrais e de voz, sendo amplamente utilizada no reconhecimento de emoções (Xie; Zhu; Hu, 2023; Song, 2019; Jiang et al., 2017; Özseven, 2019);

c) VOICEBOX (Brookes, 2024), uma *toolbox* para MATLAB destinada ao processamento de sinais de voz. Ela fornece funções para análise e síntese de sinais de fala, incluindo extração de parâmetros de voz, e foi utilizada em estudos de reconhecimento de emoções, como em (Alghifari et al., 2019).

Figura 1 – Diagrama de um Sistema de Reconhecimento de Emoções, destacando 4 passos principais.



Fonte: Elaborado pelo autor (2024).

Essas ferramentas fornecem diversos parâmetros, alguns dos quais podem ser redundantes ou irrelevantes para tarefas de reconhecimento de emoções. A seleção de parâmetros informativos é essencial para melhorar o desempenho dos SRE e otimizar a eficiência computacional (Guyon; Elisseeff, 2003).

Segundo o estudo de Bolón-Canedo, Sánchez-Marono e Alonso-Betanzos (2012), conforme pode ser visto na Tabela 1, as vantagens e desvantagens de três métodos de seleção de parâmetros. Os métodos considerados são: Filtro, Incorporado (Embedded) e Envoltório (Wrapped). Cada método possui características distintas, com vantagens como independência do classificador e custo computacional reduzido, e desvantagens, como a dependência do classificador e risco de sobreajuste. A comparação detalha como cada método interage com o classificador e as implicações de seu uso em termos de desempenho e custo computacional.

A literatura frequentemente discute a redução de parâmetros por diferentes estratégias, como filtragem, métodos incorporados e envoltórios Bolón-Canedo, Sánchez-Marono e Alonso-Betanzos (2012), além de abordagens complementares, como a redução por transformação (Uddin; Hossain, 2021) e o uso de algoritmos meta-heurísticos (Xie; Zhu; Hu, 2023).

A redução por filtragem elimina diretamente os parâmetros considerados menos informativos, com base em critérios como variância ou correlação (Duda; Hart; Stork, 2001). Esse tipo

de abordagem é aplicado para simplificar os modelos sem alterar a representação dos parâmetros remanescentes, mantendo apenas aqueles que mostram maior variabilidade ou relevância para um classificador. A variância, embora não seja um indicador exclusivo de informação e não garanta que um parâmetro contenha informação relevante, é um forte indicativo de sua utilidade. Parâmetros com alta variância são geralmente mais informativos, pois refletem uma maior dispersão dos dados, o que sugere uma maior capacidade de discriminação entre diferentes classes ou comportamentos no modelo (Duda; Hart; Stork, 2001).

Tabela 1 – Vantagens e desvantagens de técnicas de Seleção de Parâmetros.

Método	**Vantagens**	**Desvantagens**
Filtro	• Independente do classificador • Baixo custo computacional • Boa generalização	Não avalia o resultado do classificador
Incorporado (Embedded)	• Interação com o classificador • Custo computacional menor que métodos envoltórios • Identifica dependências dos parâmetros	Depende do classificador
Envoltório (Wrapped)	• Interação com o classificador • Identifica dependências nos parâmetros	• Custo computacional alto • Risco de sobreajuste • Depende do classificador

Fonte: Adpatado de (Bolón-Canedo; Sánchez-Marono; Alonso-Betanzos, 2012).

Para filtragem de parâmetros utilizam-se diversos métodos estatísticos, atribuindo uma classificação ou pontuação a cada parâmetro. Conforme destacado por Bolón-Canedo, Sánchez-Marono e Alonso-Betanzos (2012), a seleção de subconjuntos de parâmetros deve ser realizada antes da construção do modelo de aprendizado, ocorrendo apenas uma vez durante esse processo. Exemplos comuns incluem o uso de variância, ANOVA, utilizada para reconhecimento de emoções em (Sheikhan; Bejani; Gharavian, 2012), e o teste de Kruskal-Wallis, utilizado para reconhecimento de emoções em (Ravi; Taran, 2023), para medir a relevância dos parâmetros. Apesar de serem úteis, essas abordagens não levam a um grupo ótimo de parâmetros, visto que não consideram o resultado do modelo aplicado (Tsamardinos; Aliferis, 2003).

A seleção de parâmetros consiste em identificar os parâmetros mais relevantes utilizando algoritmos heurísticos, que podem ser envoltórios (em inglês *wrapped methods*), como eliminação recursiva (RFE, do inglês *Recursive Feature Elimination*), e incremental (SFS, do inglês *Sequential Feature Selection*), ou métodos de incorporação (em inglês *embedded methods*), como *Lasso* e *Ridge*. Os métodos adaptativos determinam quais características são relevantes para o desempenho do modelo, permitindo a redução da dimensionalidade sem transformar os dados. Alguns desses métodos, aplicados ao reconhecimento de emoções, podem ser vistos na Tabela 2. Embora possam aproximar-se de um conjunto quase ideal de parâmetros, eles estão sujeitos ao teorema do "Não Existe Almoço Grátis", já que dependem de um modelo de classificação (Tsamardinos; Aliferis, 2003). O teorema do "Não Existe Almoço Grátis" afirma que, na ausência de suposições sobre os dados ou de uma afirmação específica, não é possível determinar qual é o melhor algoritmo. Em teoria, isso implicaria a necessidade de testar todos os algoritmos para identificar o mais adequado (Wolpert; Macready, 1997).

Tabela 2 – Estudos que utilizaram algoritmos heurístico para reconhecimento de emoções em sinais de voz.

Referência	Método de Seleção
(Özseven, 2019) (Kuchibhotla; Vankayalapati; Anne, 2016)	SFS
(Kuchibhotla; Vankayalapati; Anne, 2016)	SFFS
(Li et al., 2022)	Relief-F
(Xie; Zhu; Hu, 2023) (Zvarevashe; Olugbara, 2018) (Li et al., 2021)	Florestas Aleatórias
(Kaur; Singh, 2023)	LASSO

Estudos recentes têm explorado o uso de algoritmos meta-heurísticos como uma abordagem para a seleção de parâmetros em reconhecimento de emoções. Esses algoritmos, inspirados em processos naturais ou sistemas biológicos, são, por definição, processos iterativos que orientam uma heurística, combinando conceitos para explorar eficientemente os espaços de busca e encontrar soluções quase-ótimas (Osman; Kelly, 2012). Na Tabela 3, apresentam-se alguns exemplos de estudos que utilizaram algoritmos meta-heurísticos para a seleção de parâmetros, destacando os métodos aplicados.

A redução por transformação altera a representação dos parâmetros, preservando a maior parte das informações em uma nova forma com menos dimensões. Os métodos identificados com base na revisão da literatura são: Análise de Componentes Principais (PCA do inglês *Principal Component Analysis*), Análise Discriminante Linear (LDA do inglês *Linear Discriminant Analysis*), Incorporação Estocástica de Vizinhos com Distribuição t (t-SNE, do inglês *t-Distributed Stochastic Neighbor Embedding*), Análise de Componentes Independentes (ICA do inglês *Independent Component Analysis*).

Tabela 3 – Métodos meta-heurísticos de seleção de parâmetros

Referência	Método de Otimização
(Xie; Zhu; Hu, 2023) (Shahin et al., 2023)	Lobos Cinzentos (*GWO - Grey Wolf Optmizer*)
(Yildirim; Kaya; Kılıç, 2021) (Zhang et al., 2020)	Busca do Cuco (*Cuckoo Search*)
(Dey et al., 2020)	Razão Áurea e Otimização de Equilíbrio

Na Tabela 4 são apresentados exemplos de estudos que aplicaram cada uma dessas técnicas de redução de parâmetros. Entre elas, a Análise de Componentes Principais (PCA) destaca-se como uma das mais amplamente utilizadas. Além disso, alguns estudos combinam diferentes métodos para aproveitar as vantagens de múltiplas abordagens. Em Liu et al. (2018), é feita a combinação o PCA e o LDA para otimizar o desempenho do modelo. O estudo de Palacios et al. (2019a) utilizou as técnicas de PCA e ICA para classificar a voz neutra e a voz sob estresse, empregando duas bases de dados de voz emocional.

Tabela 4 – Estudos que utilizaram métodos de redução por transformação.

Referência	Método de Redução
(Özseven, 2019) (Jain et al., 2022) (Sun et al., 2022) (Quan et al., 2013) (Chen et al., 2012) (Chen; Chellali; Xing, 2018) (Ke et al., 2019)	PCA
(Guo et al., 2022)	ICA
(Dujaili; Ebrahimi-Moghadam, 2023) (Krishnan; Raj; Rajangam, 2021) (Tiwari; Darji, 2022)	LDA
(Liu et al., 2018)	LDA+PCA
(Palacios et al., 2019b)	ICA+PCA

Ao abordar a classificação e os métodos de redução de parâmetros, observa-se que essas técnicas são fundamentais para a melhoria do desempenho em tarefas de aprendizado de máquina. A combinação adequada de métodos de seleção e redução pode otimizar a acurácia, reduzir o custo computacional e melhorar a interpretação dos dados.

1.1 MOTIVAÇÃO

As pesquisas em reconhecimento de emoções em sinais de voz, em geral, buscam alcançar uma maior acurácia na identificação da emoção expressa na fala. Alcançar esse reconhecimento acurado é um dos principais objetivos dos pesquisadores nessa área. No entanto, poucos estudos dedicam atenção à análise prévia dos dados antes de aplicar técnicas avançadas de aprendizado de máquina. É comum a utilização de métodos de seleção de parâmetros baseados em algoritmos e inteligência artificial, como redes neurais. Em muitos casos, o foco está nos resultados obtidos, sem a devida consideração dos custos computacionais e, frequentemente, aplicando-se métodos diversos sem suposições claras sobre as características dos dados.

Neste contexto, e fundamentando-se em teorias amplamente reconhecidas de redução de parâmetros, neste trabalho propõe-se a analisar os parâmetros acústicos da voz em bases de dados populares, utilizando métodos e testes estatísticos. O objetivo central é realizar uma seleção fundamentada nas teorias de transformações de dados para a redução de parâmetros, garantindo uma abordagem embasada na análise dos dados.

Com base no fato de que contêm informações para a classificação, características paralinguísticas, tais como parâmetros acústicos da voz, são comumente extraídas para classificação de emoções (Cowie et al., 2001; Xie; Zhu; Hu, 2023; Özseven, 2019). Um estudo (Cowie et al., 2001) revisou pesquisas anteriores que empregaram parâmetros de voz e avaliaram seu impacto emocional. Além disso, parâmetros acústicos têm sido amplamente utilizados para a detecção de emoções (Scherer; London; Wolf, 1973; Chen et al., 2012; Yildirim; Kaya; Kılıç, 2021; Xie; Zhu; Hu, 2023). No entanto, nem todos os parâmetros extraídos do áudio da fala são úteis para aumentar a acurácia do modelo. Muitos autores optam por usar métodos de redução linear, como PCA, que é o mais utilizado, como visto na Tabela 4.

A Análise de Componentes Principais (PCA) é uma transformação linear amplamente utilizada e classificada como aprendizado de máquina não supervisionado. Frequentemente, é aplicada sem uma seleção prévia de parâmetros. Nos estudos apresentados na Tabela 4, nenhum deles considera a distribuição dos dados ao empregar esta técnica de redução. Por outro lado, a Análise de Componentes Independentes (ICA) também é uma transformação linear, classificada como um método não paramétrico e não supervisionado, sendo recomendada especialmente para dados que não seguem uma distribuição normal (Jolliffe, 2002). Vale destacar que, no caso da ICA, a restrição dos dados não serem normalmente distribuídos aplica-se apenas aos dados de saída (Hyvärinen; Oja, 2010).

Diante desse panorama, é evidente a necessidade de um enfoque mais crítico na análise dos dados utilizados para o reconhecimento de emoções. A integração de métodos de seleção de parâmetros baseados em teorias de transformações de dados pode proporcionar uma abordagem mais eficaz e fundamentada, melhorando a acurácia dos modelos de classificação. Neste trabalho, exploram-se as características paralinguísticas da voz e, ao aplicar métodos estatísticos, busca-se contribuir para o avanço do campo, oferecendo uma base mais consistente para a identificação de

emoções em sinais de voz. Acredita-se que, ao priorizar uma análise cuidadosa dos dados antes da aplicação de técnicas de aprendizado de máquina, seja possível não apenas otimizar os resultados, mas também promover um entendimento mais profundo das dinâmicas emocionais expressas na fala, bem como das bases de dados já existentes e direcionar igualmente o aprendizado de algoritmos de inteligencia computacional.

Durante o desenvolvimento do mestrado, o autor deste trabalho realizou uma pesquisa sobre reconhecimento de emoções em sinais de voz, conforme descrito em Kingeski (2019). Isso motivou a proposta inicial, que envolvia a criação de uma plataforma *online* para coleta de dados de voz, com o objetivo de reunir uma base diversificada com um grande volume de registros vocais. Tal diversificação e volume visavam minimizar erros comuns em reconhecimento de emoções, frequentemente atribuídos à falta de padronização e diversidade nos dados utilizados em outros estudos.

A ideia inicial de construir uma base de dados própria foi motivada pela busca por um conjunto de dados suficientemente representativo, que contribuísse para aumentar a robustez dos modelos de análise emocional.

Em uma revisão sistemática das bases de dados existentes, apresentada no ANEXO A, foi possível identificar uma limitação: muitas das bases de dados mais comumente utilizadas para o reconhecimento de emoções em voz não possuem uma padronização clara e consistente. Essa falta de padronização apresenta desafios consideráveis, principalmente no que diz respeito à comparação entre diferentes idiomas, o que dificulta a aplicação de resultados obtidos em um idioma para outro. Além disso, a ausência de uma estrutura uniforme torna mais difícil a comparação de modelos de classificação entre bases de dados distintas, o que, por sua vez, limita a generalização dos resultados obtidos.

No entanto, durante o avanço da pesquisa, notou-se que a seleção de parâmetros específicos para a representação dos sinais de voz, utilizando outras bases de dados para validar, poderia ter um valor científico. Assim, a pesquisa foi direcionada para o desenvolvimento de métodos de seleção de parâmetros acústicos, e a análise foi realizada utilizando bases de dados de voz previamente disponibilizadas por outros autores.

1.2 IDENTIFICAÇÃO DO PROBLEMA

Embora alguns modelos atuais apresentem alta acurácia, muitos enfrentam um custo computacional elevado e requerem um grande volume de dados de entrada. Enquanto a PCA apresenta um custo computacional que cresce com o número de características devido ao cálculo dos autovalores, sua implementação em grandes conjuntos de dados é geralmente mais eficiente do que os métodos metaheurísticos, que dependem de múltiplas iterações e avaliações da função objetivo. Estudos como o de Özseven (2019) e Mishra, Warule e Deb (2023) destacam que algoritmos como GA (algoritmo genético do inglês *genetic algorithm*) e GWO, embora capazes de identificar subconjuntos ótimos de parâmetros, possuem custo computacional

significativamente maior devido à natureza iterativa e ao uso intensivo de classificadores para validação.

Um problema comum na área é o sobreajuste (em inglês *overfitting*), que torna o modelo ineficaz para novos dados (Duda; Hart; Stork, 2001). O fenômeno do sobreajuste ocorre quando um modelo de aprendizado de máquina se adapta excessivamente ao conjunto de dados de treinamento, resultando em um desempenho ruim em dados não vistos. Isso acontece porque o modelo aprende tanto as partes sistemáticas quanto o ruído dos dados de treinamento, o pode comprometer sua capacidade de generalizar para novos dados. Em resumo, o sobreajuste significa que o modelo se ajusta bem ao conjunto de treinamento, mas não consegue fazer previsões precisas em dados independentes (López; López; Crossa, 2022).

Muitos estudos na área não realizam uma análise profunda dos dados antes de aplicar algoritmos de seleção. A seleção de parâmetros é frequentemente feita sem considerar adequadamente as propriedades dos dados. O uso excessivo de métodos automatizados, como redes neurais e algoritmos heurísticos de alto custo computacional, pode levar a uma compreensão superficial dos dados. Os pesquisadores buscam resultados rápidos sem avaliar os custos e a teoria subjacente às suas escolhas metodológicas. Dos estudos mostrados nas Tabelas 4, 2 e 3, apenas Özseven (2019) faz uma análise estatística para seleção de parâmetros, o autor abordou métricas como média, mediana, desvio padrão e variância, como critério de seleção de parâmetros.

Portanto, a necessidade de uma abordagem que priorize a análise teórica e estatística dos dados se torna evidente. Ao abordar os parâmetros acústicos da voz e aplicar métodos de seleção fundamentados, é possível avançar na precisão do reconhecimento de emoções. Este estudo busca preencher essa lacuna, propondo uma investigação rigorosa e embasada em teorias de transformação de dados para a redução de parâmetros, com o objetivo de contribuir para o desenvolvimento de sistemas mais precisos e confiáveis na identificação de emoções.

1.3 OBJETIVO GERAL

O objetivo deste trabalho é investigar a influência de subconjuntos com distribuição normal de parâmetros vocais no desempenho de reconhecimento das emoções a partir da voz.

1.4 OBJETIVOS ESPECÍFICOS

1. Realizar uma análise detalhada da distribuição dos parâmetros vocais, com o intuito de identificar e separar aqueles que seguem distribuições normais e não normais, facilitando a aplicação de técnicas de redução de parâmetros adequadas a cada grupo.

2. Propor um método para a redução de parâmetros em sistemas de reconhecimento de emoções, destacando o impacto positivo de filtrar os parâmetros por distribuição antes da aplicação das técnicas de redução.

3. Comparar o desempenho de diferentes estratégias de redução de parâmetros, considerando a distribuição dos parâmetros.

1.5 CONTRIBUIÇÕES PARA O TEMA

Este trabalho oferece as seguintes contribuições para a área de reconhecimento de emoções com base em sinais de voz:

1. **Proposição de um novo protocolo de seleção de parâmetros**: O trabalho propõe um protocolo novo que separa os parâmetros vocais com base em sua distribuição, permitindo a aplicação de técnicas de redução mais adequadas para cada grupo, resultando em uma melhora na precisão do reconhecimento de emoções.
2. **Avaliação do impacto da distribuição dos parâmetros**: A pesquisa oferece uma análise detalhada de como a distribuição estatística dos parâmetros vocais (normais ou não normais) influencia o desempenho das técnicas de redução, algo que pode ser relevante para estudos futuros na área.
3. **Aprimoramento na precisão dos sistemas de reconhecimento de emoções**: A aplicação direcionada de técnicas de redução, como a PCA e a ICA, resultou em uma melhoria na precisão dos sistemas de reconhecimento de emoções.

1.6 ESTRUTURA DO TRABALHO

Este trabalho está organizado em cinco capítulos, conforme descrito a seguir:

1. O primeiro capítulo apresenta uma introdução ao tema, detalhando as possíveis aplicações de um sistema de reconhecimento de emoções, sua relevância e estrutura. Além disso, discute-se a motivação que levou à realização deste estudo, seus objetivos e as contribuições esperadas para o avanço na área de pesquisa.
2. O segundo capítulo consiste na fundamentação teórica, onde é realizada uma revisão da literatura sobre a teoria das emoções, os parâmetros acústicos da voz e as técnicas utilizadas para a redução de parâmetros. Também é abordado o método de aprendizado de máquina aplicado ao reconhecimento de emoções utilizados no trabalho.
3. O terceiro capítulo é dedicado à metodologia empregada. Cada etapa do método proposto é detalhada em seções específicas, explicando como as técnicas foram aplicadas para atingir os objetivos deste trabalho. Isso inclui a análise estatística, a seleção e redução de parâmetros, bem como a implementação do algoritmo de aprendizado.
4. O quarto capítulo apresenta os resultados obtidos, seguidos de uma discussão crítica. Os resultados são analisados em profundidade, comparando-os com trabalhos anteriores e destacando as melhorias alcançadas.

5. O quinto e último capítulo traz as conclusões, onde são resumidos os principais achados do estudo, bem como sugestões para trabalhos futuros e possíveis aprimoramentos da metodologia proposta.

1.7 PUBLICAÇÕES

Durante essa pesquisa, realizaram-se as publicações relacionadas ao tema deste trabalho:

- KINGESKI, Rafael; HENNING, Elisa; PATERNO, Aleksander S. **Fusion of PCA and ICA in statistical subset analysis for speech emotion recognition.** Sensors, v. 24, n. 17, 2024. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/24/17/5704. (Kingeski; Henning; Paterno, 2024).
- KINGESKI, Rafael.; SCHUEDA, Larissa A. P.; PATERNO, Aleksander S. **Feature analysis for speech emotion classification.** In: BASTOS-FILHO, Teodiano Freire; CALDEIRA, Eliete Maria de Oliveira; FRIZERA-NETO, Anselmo (Ed.). XXVII Brazilian Congress on Biomedical Engineering. Cham: Springer International Publishing, 2022. p. 2359–2365. ISBN 978-3-030-70601-2. (Kingeski; Schueda; Paterno, 2022).

A primeira publicação Kingeski, Schueda e Paterno (2022) investigou a influência da variabilidade dos parâmetros acústicos da voz no reconhecimento de emoções, analisando seu impacto na classificação emocional. Já a segunda publicação Kingeski, Henning e Paterno (2024) explorou a fusão de Análise de Componentes Principais (PCA) e Análise de Componentes Independentes (ICA) para a seleção estatística de subconjuntos de parâmetros, considerando a distribuição dos parâmetros acústicos de voz.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo está estruturado em três seções. A primeira seção apresenta uma revisão das teorias das emoções, abrangendo desde os conceitos iniciais até os modelos que fundamentam o presente trabalho. Na segunda seção, são discutidos os principais parâmetros acústicos de voz utilizados no reconhecimento de emoções, os quais têm sido amplamente explorados em pesquisas relacionadas ao tema. Por fim, a terceira seção aborda os métodos de redução de parâmetros e o algoritmo de aprendizado de máquina empregado nesta pesquisa, destacando as abordagens adotadas para tratar e analisar os dados.

2.1 TEORIA DAS EMOÇÕES

O estudo das emoções nos seres humanos e nos animais é antigo, com teorias sobre o assunto presentes desde a Grécia Antiga até os dias atuais. O tema tem sido abordado em diversas áreas, como filosofia, psicologia, neurociência, linguística, entre outras.

Charles Darwin foi um dos primeiros a estudar e descrever as emoções como algo biológico, enquanto muitos, até então, atribuíam as emoções ao espírito, uma entidade imaterial. Em seu livro *The Expression of the Emotions in Man and Animals*, publicado pela primeira vez em 1872, Darwin estudou as expressões de seres humanos e animais (Darwin, 1916). Darwin propôs três princípios para explicar as emoções:

a) **princípio dos hábitos úteis associados** — As ações são serviços diretos ou indiretos do estado da mente. Sempre que o mesmo estado emocional é induzido, há uma tendência, através da força do hábito, de que as mesmas ações sejam executadas;

b) **princípio da antítese** — Quando um estado emocional oposto é induzido, existe uma tendência forte e involuntária de que ações e movimentos opostos sejam expressos;

c) **princípio das ações devidas à constituição do Sistema Nervoso** — Independente da vontade e até certo ponto do hábito, quando o sistema sensorial é intensamente estimulado, gera-se uma excessiva força nervosa, que resulta em expressões reconhecíveis. Este efeito pode ser chamado de ação direta do sistema nervoso.

Esses princípios descrevem as expressões e gestos involuntários de animais e seres humanos quando estão sob a influência de emoções e sensações. Darwin foi pioneiro ao comparar as emoções de animais e seres humanos, afirmando que as expressões faciais de emoções são universais. Posteriormente, Paul Ekman confirmou essa teoria (Ekman, 2011).

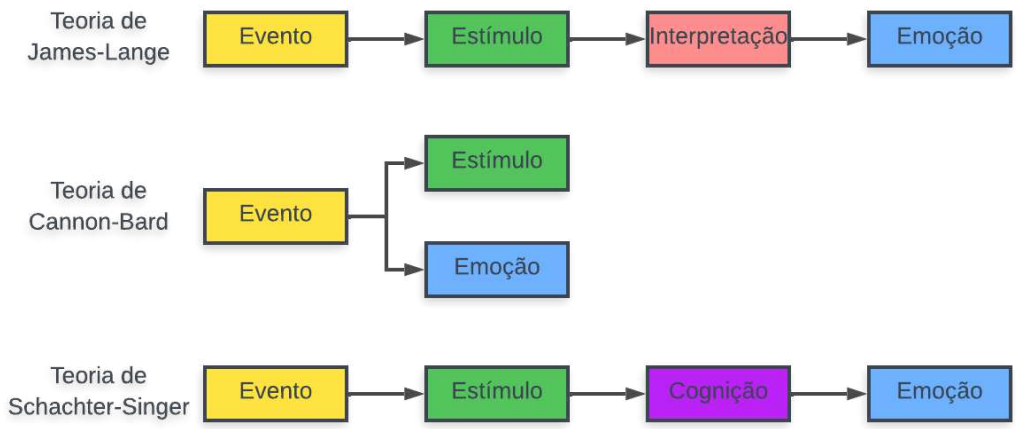
No final do século XIX, os primeiros estudos no campo da psicologia foram realizados por William James e Carl Lange. A teoria que recebeu o nome de Teoria de James-Lange afirma que as emoções ocorrem em uma sequência: estímulo, reação física e emoção (Houwer; Hermans, 2010).

Em 1927, o fisiologista Walter Cannon discordou da Teoria de James-Lange, propondo sua própria teoria. Cannon argumentou que as atividades fisiológicas não poderiam provocar as emoções. Phillip Bard concordou com Cannon, complementando sua teoria, que ficou conhecida como Teoria de Cannon-Bard, e afirmava que as emoções e as reações fisiológicas ocorrem simultaneamente (Cannon, 1987).

Em 1962, os psicólogos americanos Stanley Schachter e Jerome E. Singer desenvolveram uma nova teoria sobre as emoções, na qual propuseram que as emoções possuem tanto um componente cognitivo quanto um componente fisiológico. Essa teoria ficou conhecida como Teoria dos Dois Fatores de Emoção (Schachter; Singer, 1962).

O diagrama na Figura 2 ilustra as três teorias mais comuns, chamadas de teorias clássicas das emoções. Embora compartilhem componentes em comum, elas se diferem em vários aspectos fundamentais.

Figura 2 – Teorias Clássicas das Emoções



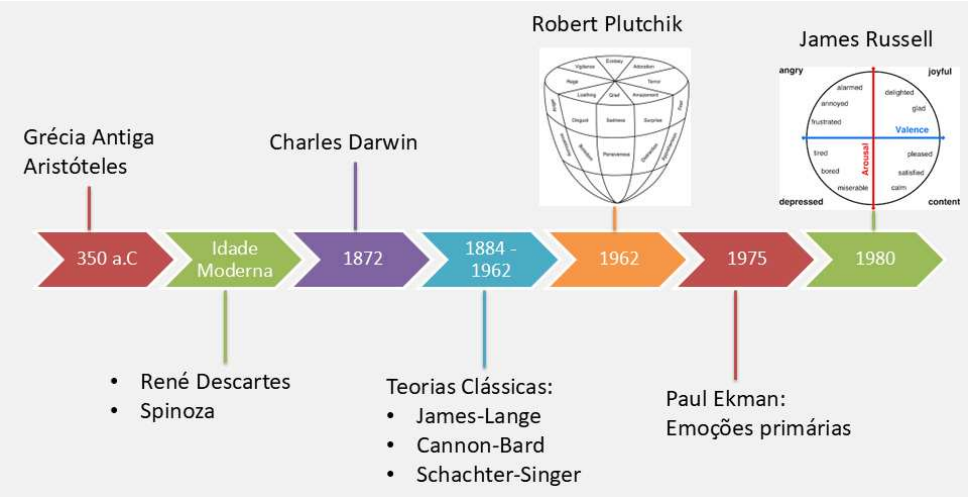
Fonte: Elaborado pelo autor (2022).

Ao longo da história, diversas teorias sobre as emoções foram propostas, refletindo o entendimento e o contexto cultural de cada época. A Figura 3 apresenta uma linha do tempo com algumas das teorias mais significativas, começando na Grécia Antiga com Aristóteles até a teoria das emoções de James Russell.

A concepção aristotélica das emoções, descrita por Aristóteles, estabelece as primeiras bases filosóficas sobre a natureza das emoções humanas. Segundo ele, as emoções são reações tanto do corpo quanto da mente, e devem ser compreendidas em relação à virtude e ao equilíbrio emocional (Yazici, 2015). A partir dessa visão inicial, outras abordagens surgiram ao longo dos séculos, com destaque para a teoria de William James e Carl Lange, que propuseram que as emoções são reações fisiológicas aos estímulos ambientais.

No século XX, a teoria dimensional das emoções, proposta por James Russell, representa um marco importante ao sugerir que as emoções podem ser descritas por duas dimensões principais: valência e ativação (Russell, 1980). Essa teoria oferece uma compreensão mais

Figura 3 – Linha do tempo das teorias das emoções, desde a Grécia Antiga com Aristóteles até a teoria dimensional proposta por James Russell.



Fonte: Elaborado pelo autor (2024), com elementos de (Plutchik, 1962) e (Russell; Fehr, 1994).

abrangente e flexível das emoções, sendo amplamente utilizada nas pesquisas contemporâneas sobre o tema.

Essa evolução das teorias reflete a crescente complexidade no entendimento humano sobre as emoções, desde suas raízes filosóficas até as abordagens científicas modernas, que buscam quantificar e modelar as emoções de maneira mais precisa.

2.1.1 Teoria de Robert Plutchik

Robert Plutchik, psicólogo que, ao analisar pacientes com transtornos mentais, observou dificuldades no reconhecimento de emoções por parte desses indivíduos (Plutchik, 1962). Plutchik formulou seis postulados para descrever formas de analisar e identificar emoções. Para ele, as emoções são semelhantes às cores, existindo emoções primárias e emoções complexas, que resultam de combinações das emoções primárias, assim como as cores secundárias são formadas pela mistura de cores primárias (Plutchik, 1962).

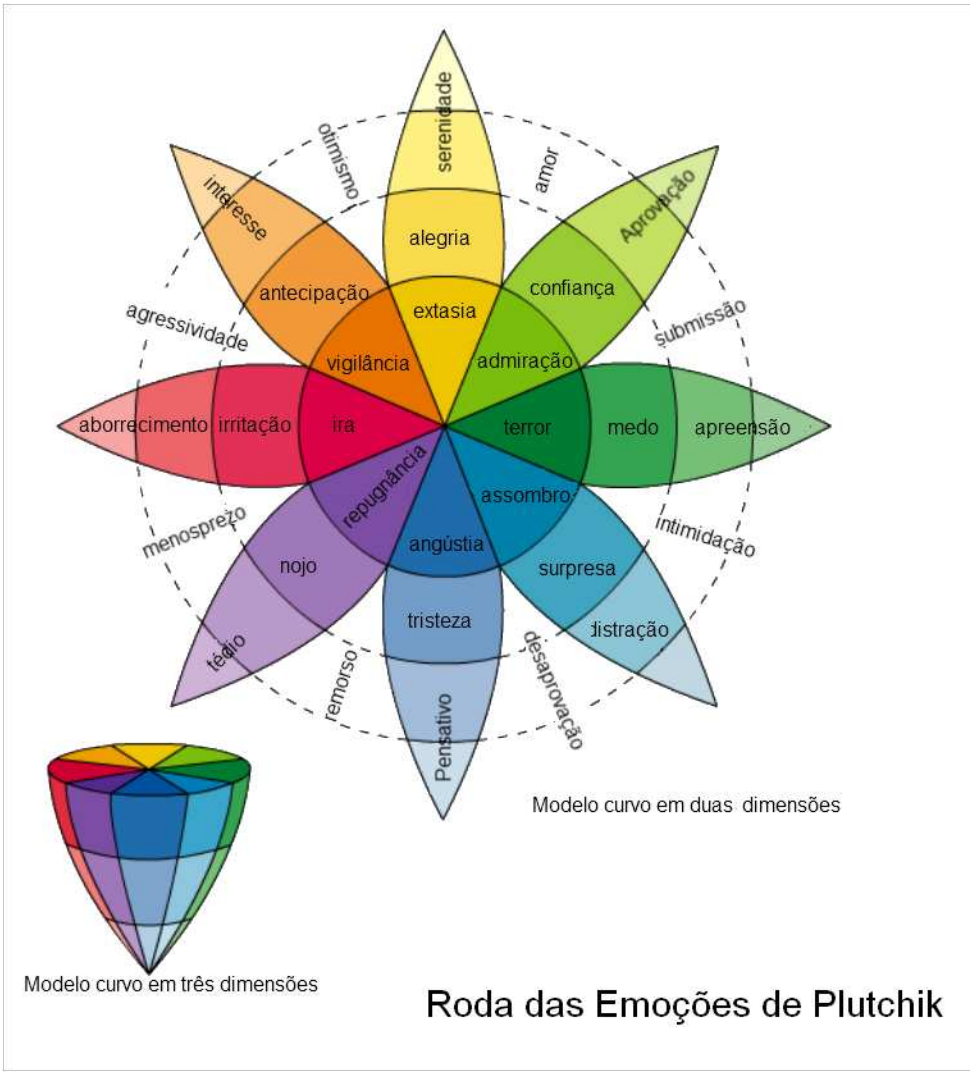
Ele também criou o diagrama das emoções, apresentado na Figura 4, conhecido como a roda das emoções. O diagrama, que se assemelha a uma flor, atribui a cada pétala uma emoção primária em diferentes intensidades. Entre as pétalas, encontram-se as combinações de duas emoções adjacentes. Este foi um dos primeiros estudos a descrever as emoções de forma discreta, e Plutchik não se limitou apenas às emoções primárias, catalogando também outras emoções (Plutchik, 1962).

2.1.2 Emoções primárias segundo Paul Ekman

Um estudo fundamental sobre as emoções humanas foi realizado por Paul Ekman, psicólogo que, além de corroborar as teorias de Darwin, desenvolveu uma abordagem importante para a compreensão das emoções primárias. Ekman descreveu essas emoções com base na

facilidade de expressão e reconhecimento universal, sugerindo que as emoções primárias são: felicidade, medo, raiva, tristeza, surpresa e nojo (Ekman; Friesen, 1975).

Figura 4 – Diagrama de Plutchik das Emoções



Fonte: Página Robert Plutchik Wikipédia[1]

2.1.3 Antônio Damásio: Emoções na Neurociência

Antônio Damásio, médico neurologista e neurocientista português, desempenhou um papel fundamental nos estudos contemporâneos sobre as emoções. Em 1994, publicou o livro intitulado *O Erro de Descartes – Emoção, razão e o cérebro humano*, no qual descreveu a importância das emoções para os seres humanos, com base em estudos realizados com pacientes que apresentavam dificuldades na tomada de decisões e distúrbios emocionais (Damasio, 1996). Damásio argumenta que as emoções são tão importantes para o ser humano que, em algumas situações, desempenham um papel mais relevante do que a razão. Um exemplo claro disso são os

[1] Disponível em: https://pt.wikipedia.org/wiki/Robert_Plutchik, acesso em 18 de dezembro de 2024.

casos em que o medo leva o indivíduo a se afastar do perigo com pouca ou nenhuma intervenção da razão.

Embora o uso habitual da palavra “emoção” inclua o sentimento, Damásio esclarece a diferença entre emoção e sentimento. Segundo o autor:

> [...] as emoções são ações ou movimentos, muitos deles públicos, que ocorrem no rosto, na voz ou em comportamentos específicos. Alguns comportamentos emocionais não são perceptíveis a olho nu, mas podem se tornar “visíveis” com sondas científicas modernas, como a medição dos níveis hormonais no sangue ou dos padrões de ondas eletrofisiológicas. Os sentimentos, por outro lado, são necessariamente invisíveis ao público, assim como todas as outras imagens mentais, escondidas de quem quer que seja, exceto do seu devido proprietário — a propriedade mais privada do organismo no qual ocorrem no cérebro. (Damásio, A., p. 23).

Dessa forma, enquanto as emoções podem ser expressas por meio de sinais observáveis, como expressões faciais, variações na voz e respostas fisiológicas, os sentimentos são intrinsecamente subjetivos e de difícil acesso direto. Isso sugere que a identificação de sentimentos por meio de inteligência artificial ou sinais fisiológicos apresenta desafios significativos, uma vez que sua manifestação não é diretamente acessível. No entanto, as emoções podem ser analisadas a partir de suas expressões externas, sendo a voz uma das formas mais relevantes de comunicação emocional (Cowie et al., 2001).

2.1.4 O modelo de Russell

Diferentemente de todos os outros modelos e diagramas mencionados, Russell propôs uma nova abordagem para a análise das emoções, baseada na identificação de fatores subjacentes. Em sua pesquisa, 34 sujeitos foram convidados a agrupar 28 termos emocionais, e a partir dessa análise, Russell formulou a hipótese de que as emoções possuem fatores ortogonais que podem ser representados em um espaço euclidiano, como ilustrado na Figura 5.

Embora Russell seja frequentemente creditado pela criação desse modelo, a ideia de analisar emoções por meio de fatores não era inteiramente nova. Segundo Nowlis e Nowlis (1956 apud Russell, 1980), Russell já havia utilizado essa abordagem em estudos anteriores sobre autoavaliação de emoções.

O modelo de Russell representa uma contribuição significativa para a análise quantitativa das emoções, proporcionando uma ferramenta valiosa para a compreensão da estrutura e organização das experiências emocionais.

Figura 5 – Modelo Circumplexo de Russell que mapeou os 28 termos emocionais.

Fonte: (Nogueira, 2018)

2.2 VOZ E EMOÇÕES

A principal fonte da fala humana é a energia gerada pelo ar que passa pelas pregas vocais, causando suas vibrações. A vibração produz um espectro de harmônicos que, ao ser seletivamente filtrado pelo trato vocal, resulta em um espectro que varia ao longo do tempo, permitindo a identificação das palavras. A fala carrega dois tipos de informação: a informação linguística, que transmite de forma qualitativa os objetivos do locutor de acordo com as regras do idioma, e as informações paralinguísticas, que também estão relacionadas aos objetivos do locutor, mas não têm relevância direta para o conteúdo semântico das palavras ditas (Cowie et al., 2001).

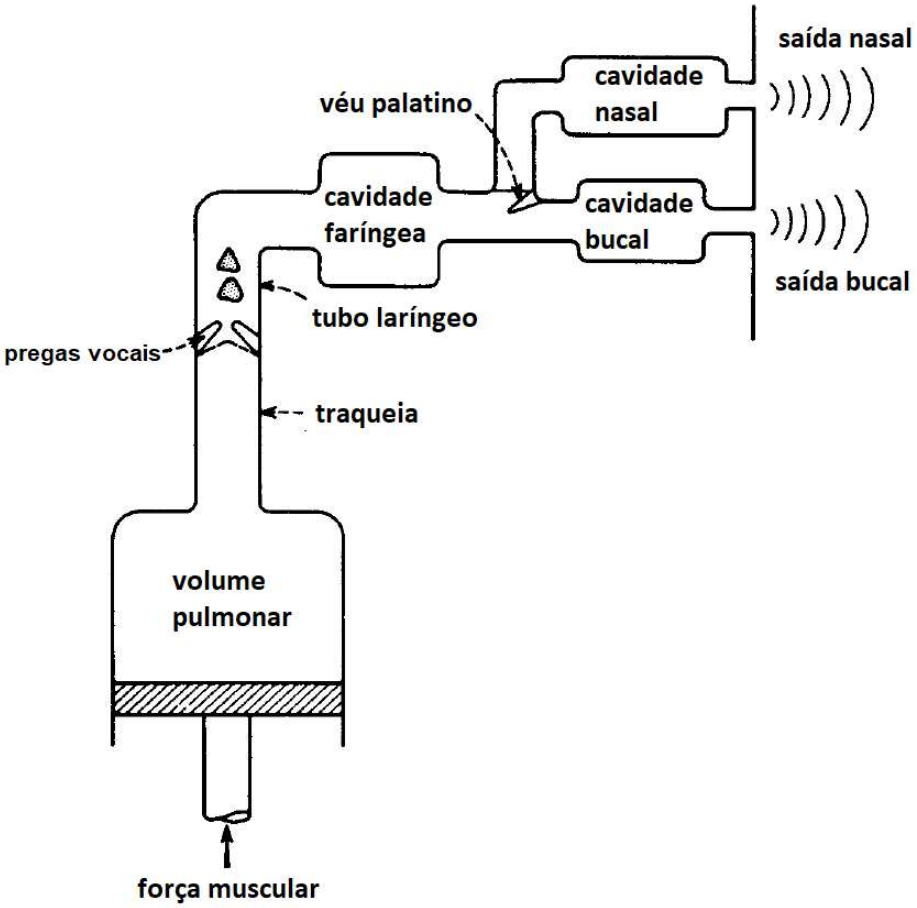
As emoções fazem parte das informações paralinguísticas da fala. Assim, por meio de métodos tradicionais de análise de voz, usados para reconhecer informações linguísticas, é possível também identificar emoções. Desenvolver sistemas que detectem as emoções dos usuários é um dos principais objetivos do campo de pesquisa conhecido como computação afetiva (Picard, 1995).

2.2.1 O Sinal de Voz

Os sons da fala humana são propagados por ondas acústicas, que resultam dos movimentos no aparelho respiratório e mastigatório. Na Figura 6, está ilustrado o aparelho vocal, no qual a força muscular empurra o ar, excitando o mecanismo de voz pela traqueia. Quando as pregas vocais estão tensionadas, o ar causa a vibração, produzindo os sons chamados de sonoros. Quando as pregas vocais estão relaxadas, o fluxo de ar que passa pelo trato vocal torna-se turbulento, resultando na produção dos sons chamados de surdos. Quando ocorre uma

pressão abrupta do ar no trato vocal, são produzidos os sons chamados de oclusivos (Rabiner; Juang, 1993).

Figura 6 – Diagrama de representação do mecanismo fisiológico da produção da fala



Fonte: (Rabiner; Juang, 1993)

Exemplo dos três mecanismos básicos de excitação que podem estar envolvidos durante a fala (Alcaim; Oliveira, 2011):

a) sons Sonoros: Como o som de /u/ em uva, as vibrações nas corda vocais são quase periódicas;

b) sons Fricativos Surdos: como som de s em sala, é criado por uma fonte de ruído contínuo com um espectro amplo e uniforme;

c) sons Oclusivos: como o som de /p/ em pato, consiste de um súbito desprendimento de um excesso de pressão gerado em alguma parte do aparelho vocal.

Os efeitos fisiológicos causados pelas emoções afetam o Sistema Nervoso Autônomo (SNA) e, por consequência, provocam alterações na respiração, salivação e musculatura do peito, da garganta e da cabeça. Essas alterações acabam afetando diretamente as características do trato

vocal (Scherer, 1979). Alguns autores mencionam padrões de alteração nos parâmetros da voz como consequência desses efeitos fisiológicos.

A Tabela 5 apresenta uma descrição das principais alterações em parâmetros de voz segundo estudo de Murray e Arnott (1993), associadas às emoções identificadas. Essas modificações foram analisadas em comparação com a fala de indivíduos em estado neutro, ou seja, sem emoções.

Tabela 5 – Comparação de alterações na voz para cinco emoções diferentes, F_0 – Frequência fundamental da voz.

	Raiva	**Felicidade**	**Tristeza**	**Medo**	**Nojo**
Velocidade da Fala	mais rápida	rápida ou lenta	mais devagar	muito rápido	muito devagar
Média F_0	muito alta	muito alta	mais baixa	altíssima	baixíssima
Variação do F_0	muito alta	muito alta	menor	muito grande	mais larga
Intensidade	alta	alta	baixa	normal	baixa
Qualidade da Voz	voz soprada, voz de peito	voz soprada, estridente	ressonante	voz irregular	resmungada, voz de peito
Mudanças no F_0	abrupta	suave, inflexões para cima	inflexões para baixo	normal	grande, inflexões para baixo
Articulação	tensa	normal	arrastada	precisa	normal

Fonte: (Murray; Arnott, 1993, p. 176, tradução livre)

De acordo com Scherer (1979), ao considerar a análise dos efeitos fisiológicos, as emoções podem ser classificadas em agradáveis e desagradáveis.

Verificação Intrínseca de Agradabilidade: trata-se da forma mais básica de avaliação realizada por todo o organismo. O critério de agradabilidade ou desagradabilidade pode ser esperado tanto de características inatas quanto por associação prévia ou instruída (Scherer, 1986).

Os efeitos das emoções desagradáveis parecem ser consideravelmente mais intensos, incluindo constrição e tensionamento da faringe, encurtamento do trato vocal, o que resulta em maior ressonância na região de alta frequência, estreitamento na largura de banda dos formantes, aumento do primeiro formante, e redução do segundo formante e possivelmente do terceiro, ocasionando o efeito conhecido como “voz estreita”. Por outro lado, os efeitos das emoções agradáveis são mais difíceis de prever. A expansão e o relaxamento da faringe causam a diminuição da frequência do primeiro formante e um amortecimento nas altas frequências; entretanto, uma laringe encurtada pode compensar esse efeito do relaxamento da faringe, equilibrando toda a faixa de frequências e produzindo uma estrutura harmônica clara, o que pode ser descrito como “voz ampla” (Scherer, 1986).

2.2.2 Frequência Fundamental

O *pitch*, também denominado frequência fundamental, é um dos parâmetros mais importantes no processamento de sinais de voz (Rabiner; Schafer, 1980). Durante a fala, as pregas vocais podem apresentar dois estados: o ritmo de abertura para sons vozeados (com vibração) ou a total abertura para sons não vozeados (sem vibração). Quando o som é vozeado, a resposta do trato vocal é um sinal periódico, que consiste em uma resposta a uma série de impulsos. A distância entre os impulsos é a duração, e o inverso dessa distância é a frequência fundamental, denotada por F_0 (Kaminska; Pelikant, 2012). Existem diversas formas de se obter o *pitch*, como a autocorrelação, análise cepstral, coeficientes LPC, entre outras (Rabiner; Schafer, 1980). Para o reconhecimento de emoções por meio dos sinais de voz, obtêm-se parâmetros característicos ou funcionais do *pitch*, como: valor máximo, valor mínimo, média, desvio padrão e variação (Sharma; Neumann; Kim, 2002).

Segundo Huang, Acero e Hon (2001), são típicas as seguintes dificuldades na extração desse parâmetro:

a) erros de sub-harmônicos: Quando o sinal é periódico, com período T, ele também é periódico em $2T$, $3T$, e assim por diante. Espera-se que os valores dessas harmônicas nos algoritmos para extração do *pitch* também sejam elevados;

b) erros de harmônicos: Se o harmônico com energia total do sinal dada por M for dominante, o valor no período T_0/M será maior. Isso pode ocorrer caso o harmônico coincida com a frequência de algum dos formantes, tornando o sinal nesse ponto muito maior, o que pode ocasionar erro na identificação do período fundamental do sinal;

c) ruídos: Quando a relação sinal-ruído é baixa, a extração do *pitch* torna-se quase impossível;

d) laringealização da voz: Nesse caso, a frequência fundamental varia bruscamente. Nessa situação, o *pitch* não é bem definido, e a imposição de suavização de curva pode causar perda de informação. O termo em inglês para a laringealização da voz é *Vocal Fry*;

e) alterações bruscas do *pitch*, como variações de uma oitava acima ou abaixo;

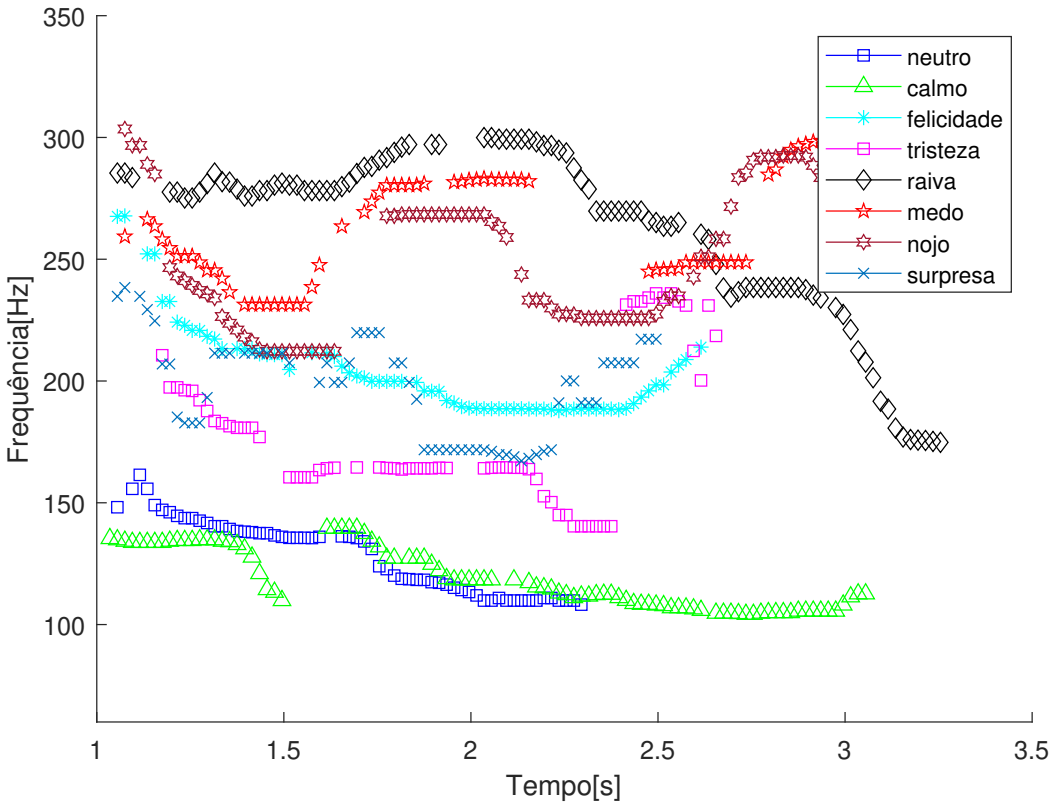
f) voz soprada: Esse fenômeno também prejudica a distinção entre o ruído e a frequência fundamental do sinal;

g) Filtragem de banda estreita, causada por configurações do trato vocal, pode fazer com que o sinal se torne periódico.

Utilizou-se o algoritmo de extração de *pitch* baseado em análise cepstral para calcular o *pitch* em oito diferentes emoções expressas pela mesma frase e pelo mesmo locutor. A Figura 7 apresenta o *pitch* para cada uma das emoções da base de dados RAVDESS (Livingstone; Russo; Najbauer, 2018). As emoções consideradas incluem felicidade, raiva, medo, tristeza, tédio,

surpresa, nojo, calmo e o estado neutro, facilitando a análise comparativa entre as diferentes características emocionais do sinal de voz.

Figura 7 – Frequência fundamental para um sinal de voz da mesma frase e mesmo locutor em oito diferentes emoções, calculado pelo método de análise cepstral, janelamento 30ms com sobreposição de 20ms.



Fonte: Elaborado pelo autor (2024).

2.2.3 Energia de Tempo Curto

Uma função básica de análise para sinais de fala é a energia de tempo curto. Essa função é simples de calcular e útil para estimar propriedades da função de excitação da voz. A energia de tempo curto representa as variações de amplitude de um sinal em função do tempo (Rabiner; Schafer, 1980). A energia de tempo curto é definida na equação (1), onde $x(m)$ é o sinal no tempo discreto m.

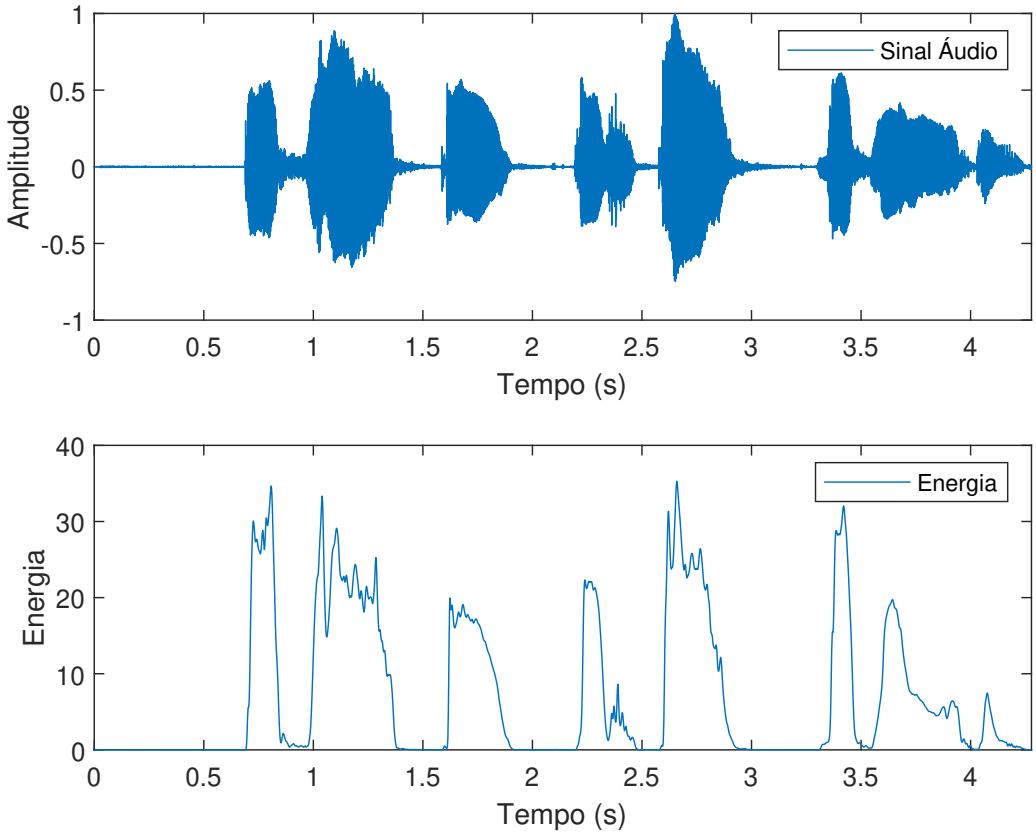
$$\mathscr{E}_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{1}$$

podendo ser reescrita por (2), em que $h(n-m) = w^2(n-m)$ é a janela utilizada para segmentar o sinal.

$$\mathscr{E}_n = \sum_{m=-\infty}^{\infty} x(m)^2 \cdot h(n-m) \tag{2}$$

A análise da energia de curto tempo de um sinal de voz permite observar variações na intensidade do som ao longo do tempo. A Figura 8 apresenta o sinal de voz no domínio do tempo juntamente com sua energia de curto tempo, calculada utilizando uma janela de 20 milissegundos, destacando as flutuações de energia que correspondem às dinâmicas naturais do discurso.

Figura 8 – Representação do sinal de voz no domínio do tempo e sua respectiva energia de curto prazo, evidenciando variações de intensidade ao longo do sinal.



Fonte: Elaborado pelo autor (2024).

2.2.4 Codificação Linear Preditiva - LPC

A *Codificação Linear Preditiva* (LPC), do inglês *Linear Predictive Coding*, é uma das técnicas de análise de fala. Ela é comumente utilizada para estimar os parâmetros do modelo de tempo discreto para a produção da fala (ou seja, altura tonal, formantes, espectros de curto prazo, funções de área do trato vocal) e é amplamente utilizada para representar a fala em transmissões ou armazenamentos de baixa taxa de bits, além de ser aplicada no reconhecimento automático de fala e de locutor. A importância desse método reside tanto na sua capacidade de

fornecer estimativas precisas dos parâmetros de fala quanto na sua relativa rapidez computacional (Rabiner; Schafer, 2007).

Neste modelo, os efeitos do espectro composto são representados por um filtro digital variável no tempo, como ilustrado na Figura 9, cuja função de transferência no estado estacionário é representada pela função racional de polos dada pela equação (3),

$$H(z) = \frac{S(z)}{GU(z)} = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{3}$$

a função $H(z)$ é chamada de função do sistema do trato vocal, embora represente não apenas os efeitos das ressonâncias do trato vocal, mas também os efeitos da radiação nos lábios e, no caso da fala sonora, os efeitos espectrais da forma do pulso glótico. Este sistema é excitado por um trem de impulsos quase periódico para a fala vozeada (visto que a forma do pulso glótico está incluída em $H(z)$ ou por uma sequência de ruído aleatório para a fala não vozeada (Rabiner; Schafer, 2007).

Os parâmetros deste modelo são:

a) classificação vozeada/não vozeada;

b) período do *pitch* (F0) para fala vozeada;

c) parâmetro de ganho G.

Os coeficientes $(a_k, k = 1,2,\ldots,p)$ do filtro digital de polos, representam os parâmetros do sistema de trato vocal. Todos esses parâmetros variam com o tempo. As amostras do sinal de fala $s[n]$ estão relacionadas à excitação $u[n]$ pela equação de diferenças (4),

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n], \tag{4}$$

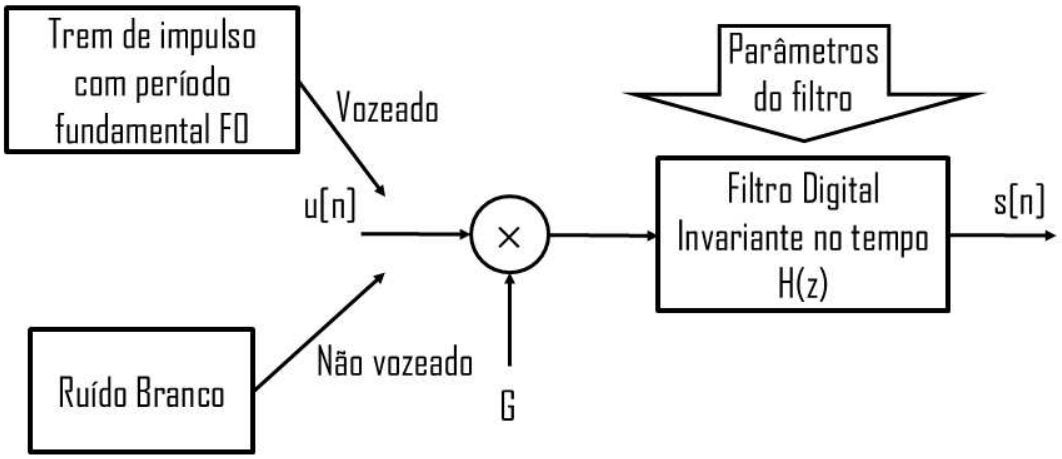
o preditor linear de p-ésimo grau, com coeficientes de predição $(\alpha_k, k = 1,2,\ldots,p)$, é definido como um sistema cuja entrada é $s[n]$ e cuja saída é $\tilde{s}[n]$, dada pela equação (5),

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{5}$$

A função do sistema deste preditor linear de p-ésima ordem é o polinômio na transformada z:

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} = \frac{\tilde{S}(z)}{S(z)}. \tag{6}$$

Figura 9 – Modelo do Trato Vocal Simplificado



Fonte: Elaborado pelo autor (2024). Adaptado de (Rabiner; Schafer, 2007).

O polinômio $P(z)$ é chamado de polinômio preditor. O erro de predição $e[n]$ é definido como a diferença entre $s[n]$ e $\tilde{s}[n]$, ou seja:

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k], \tag{7}$$

da Equação (7), segue-se que a sequência de erro de predição é a saída de um sistema cuja entrada é $s[n]$ e cuja função de sistema é:

$$A(z) = \frac{E(z)}{S(z)} = 1 - P(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}. \tag{8}$$

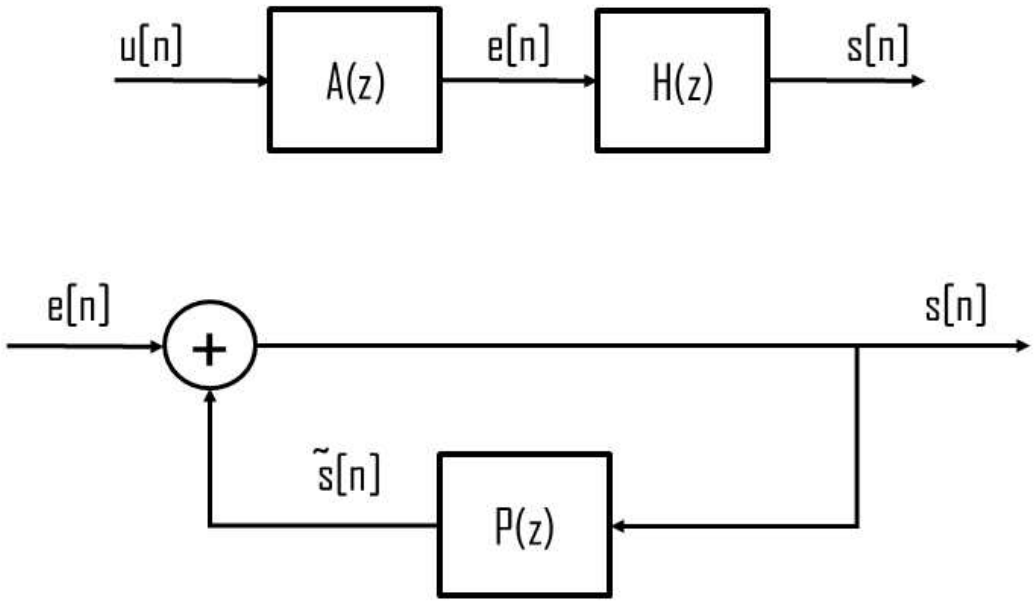
A transformada z dada por $A(z)$, é um polinômio em z^{-1}. Este é frequentemente chamado de polinômio de erro de predição ou, de forma equivalente, de polinômio LPC.

O primeiro diagrama de blocos da Figura 10 ilustra as relações entre os sinais $s[n]$, $e[n]$ e $\tilde{s}[n]$. O sinal de erro $e[n]$ é obtido como saída do filtro $H(z)$, tendo $s[n]$ como entrada. Por definição, $e[n]$ representa a excitação do trato vocal: um trem de impulsos quase periódico para fala sonora e ruído aleatório para fala não sonora. O sinal original $s[n]$ é reconstruído processando o sinal de erro pelo filtro de polos $H(z)$.

No segundo diagrama da Figura 10, tem-se o processo de realimentação para reconstruir $s[n]$ a partir de $e[n]$, ou seja, a combinação da previsão do sinal $\tilde{s}[n]$ com o sinal de erro $e[n]$. O filtro $H(z)$ é uma função racional de p-ésima ordem, descrita por:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - P(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{9}$$

Figura 10 – Modelo do Trato Vocal Simplificado



Fonte: Elaborado pelo autor (2024). Adaptado de (Rabiner; Schafer, 2007).

e $A(z)$ é o polinômio de p-ésima ordem apresentado na equação (8). Este sistema de função de transferência $H(z)$ é frequentemente chamado de modelo do trato vocal ou modelo LPC (Rabiner; Schafer, 2007).

O problema fundamental da análise por predição linear é determinar os coeficientes $(\alpha_k, k = 1,2,\ldots,p)$ diretamente a partir do sinal de fala. Esses coeficientes são usados para estimar as propriedades espectrais do sinal de fala por meio da equação 9. Devido à natureza variável no tempo do sinal de fala, os coeficientes devem ser estimados utilizando uma análise de curto tempo, minimizando o erro médio quadrático de predição em um segmento curto da forma de onda de fala.

A justificativa para essa abordagem pode ser feita de várias maneiras. Primeiro, quando α_k são os coeficientes ideais, o sinal de erro $e[n]$ se aproxima de uma excitação ideal. Para fala sonora, $e[n]$ consiste em um trem de impulsos, enquanto para fala não sonora é um ruído branco. A minimização do erro de predição está de acordo com essa observação. Além disso, quando os coeficientes α_k são constantes e o sistema é excitado por um impulso ou ruído branco estacionário, os coeficientes de predição resultantes são equivalentes aos coeficientes originais do sistema (Rabiner; Schafer, 2007).

A soma total do erro de predição é definida como:

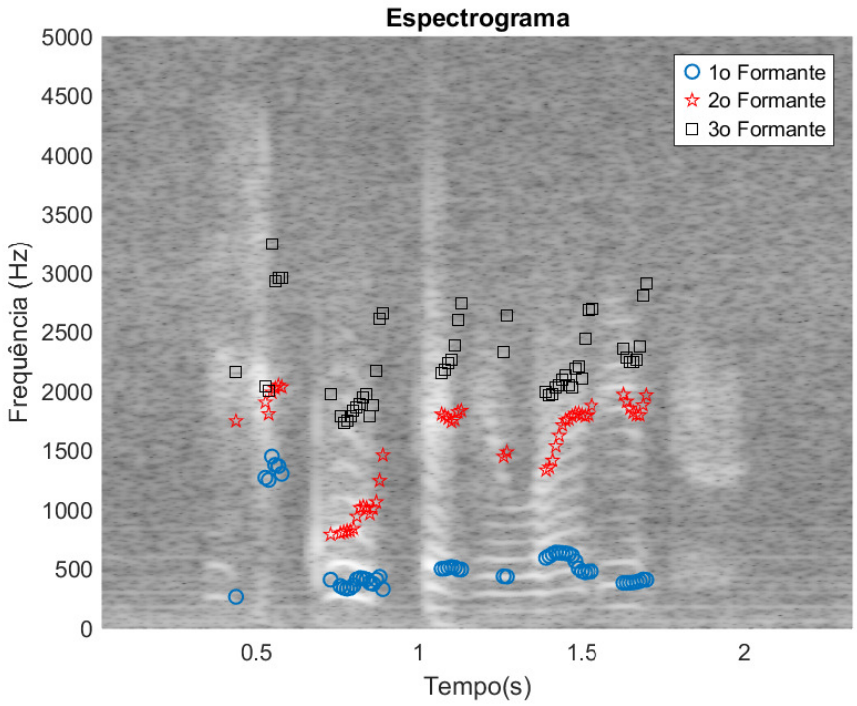
$$E_{\hat{n}} = \sum_m e_{\hat{n}}^2[m] = \sum_m (s_{\hat{n}}[m] - \tilde{s}_{\hat{n}}[m])^2, \tag{10}$$

onde $s_{\hat{n}}[m]$ é um segmento do sinal de fala. A minimização do erro é feita ajustando α_k. O sistema

de análise por predição linear, apesar de complexo em detalhes computacionais, é eficiente e amplamente utilizado devido aos seus resultados consistentes e ao modelo matemático robusto (Rabiner; Schafer, 2007).

Na Figura 11, apresenta-se o espectrograma de um sinal de voz, destacando os três primeiros formantes. Os formantes são parâmetros essenciais que podem ser estimados por meio do filtro preditor LPC. A Figura 12, destaca o espectro de um segmento de voz, calculado por Transformada Rápida de Fourier (FFT), junto à envoltória espectral obtida com os coeficientes LPC, com filtro de ordem 46 utilizando o algoritmo de Levinson-Durbin e janelamento de 60ms com sobreposição de 50ms. Os picos da envoltória são os formantes da voz. Esses gráficos ilustram a análise detalhada da estrutura espectral e a eficácia do uso do LPC na modelagem da fala.

Figura 11 – Gráfico do Espectrograma de um Sinal de Voz com os três primeiros formantes.

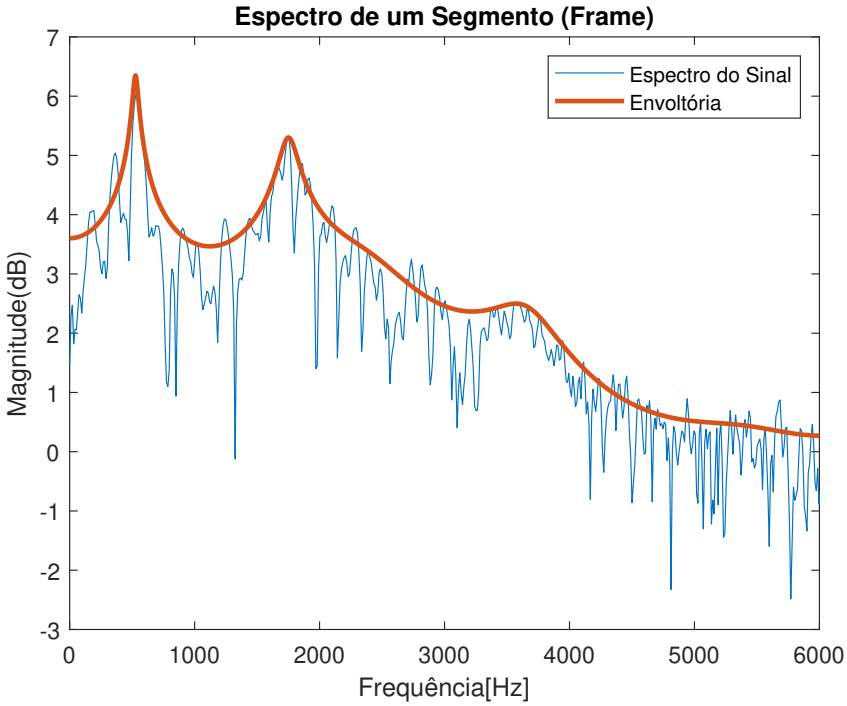


Fonte: Elaborado pelo autor (2024).

2.2.5 Coeficientes Mel-Cepstrais

Os coeficientes mel-cepstrais (*Mel-frequency cepstral coefficients*, MFCC) são baseados no modelo de audição humana. O ouvido humano não possui uma resposta linear à percepção de frequência; a percepção do volume está relacionada à frequência do sinal captado. Os autores Davis e Mermelstein (1980) utilizaram a ideia da escala de audição humana para criar um novo parâmetro de análise de sinais de áudio e voz. A ideia fundamental dos coeficientes mel é realizar uma análise de frequência com base em um banco de filtros, cujo espaçamento de banda crítica e larguras de banda são ajustados para se aproximar do modelo de audição humana (Rabiner;

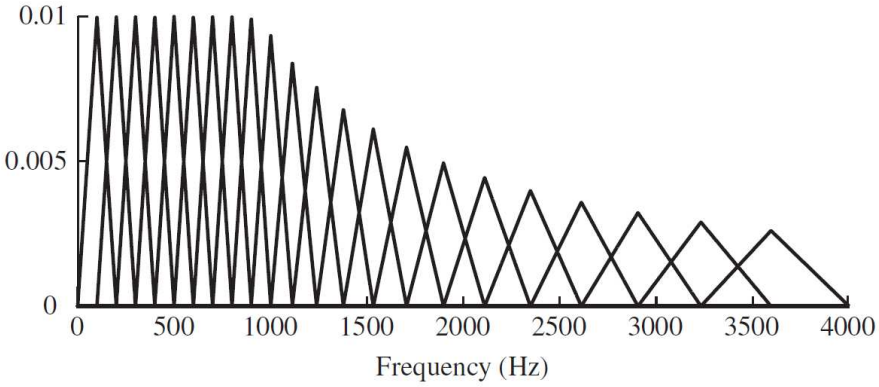
Figura 12 – Formantes de um segmento de voz.



Fonte: Elaborado pelo autor (2024).

Schafer, 2007). A Figura 13 mostra o gráfico de banda dos filtros utilizados na extração dos coeficientes mel-cepstrais.

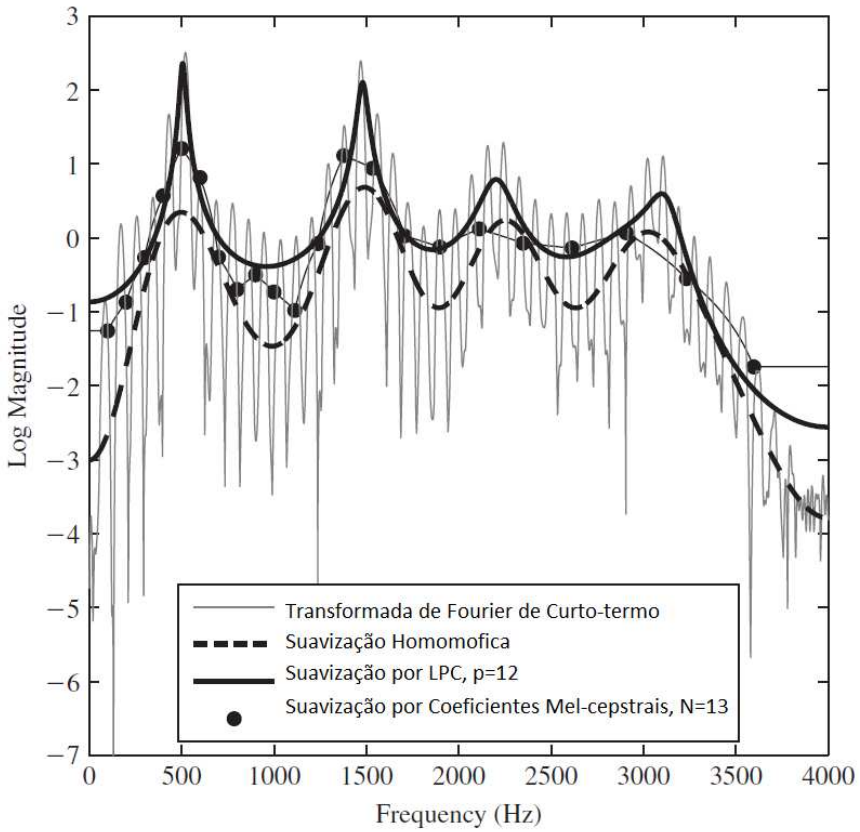
Figura 13 – Funções de ponderação para cálculos de frequência-mel-cepstral



Fonte: (Rabiner; Schafer, 2010)

Para baixas frequências, esses filtros possuem o mesmo ganho, que é gradualmente reduzido conforme a frequência aumenta. Neste caso, utilizam-se 22 filtros. A Figura 14 compara quatro formas de calcular o espectro de um intervalo de sinal de voz (*frame*). No exemplo, o número de coeficientes MFCC utilizados foi de $N_{mfcc} = 13$. A quantidade de filtros é sempre maior do que a de coeficientes (Rabiner; Schafer, 2010).

Figura 14 – Comparação do espectro de um *frame* de um sinal de voz por MFCC



Fonte: (Rabiner; Schafer, 2010)

2.2.6 Considerações Sobre os Parâmetros de Voz

Como discutido nas subseções anteriores, foram destacados e descritos alguns dos principais parâmetros acústicos de voz, com base nas teorias que fundamentam sua importância no reconhecimento de emoções. A escolha desses parâmetros foi motivada por sua relevância em estudos anteriores e por sua pertinência ao contexto da pesquisa presente. Contudo, o *openSMILE* foi utilizado para extrair um conjunto mais amplo de parâmetros, o que permitiu uma análise mais abrangente das características acústicas presentes nos sinais de voz e que está descrita juntamente com a metodologia no Capítulo 3.

2.3 ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais é uma técnica muito utilizada para redução de dimensionalidade, ou seja, para reduzir o número de parâmetros de um conjunto de dados. Essa técnica baseia-se em uma transformação linear dos dados, projetando-os em um subespaço definido pelas direções de maior variância. A ideia principal é encontrar um conjunto de novas variáveis, chamadas componentes principais, que sejam combinações lineares das variáveis originais e expliquem a maior parte da variação presente nos dados (Jolliffe, 2002).

Segundo Jolliffe (2002), pode-se descrever essa transformação por meio da equação (11), onde $\mathbf{X}$ é a matriz que contém os dados originais, composta pelos parâmetros extraídos dos sinais de áudio. Nesse contexto, cada linha da matriz representa um conjunto de valores correspondentes a diferentes parâmetros extraídos de um único sinal de áudio, enquanto cada coluna corresponde aos valores de um mesmo parâmetro calculados para todos os sinais analisados. Essa organização permite que a matriz $\mathbf{X}$ seja usada diretamente para realizar a redução dimensional. A matriz de transformação, $\mathbf{W_P}$, contém os autovetores da matriz de covariância de $\mathbf{X}$, e $\mathbf{Z}_{\text{PCA}}$ representa os dados transformados, ou seja, os dados projetados em uma nova base ortogonal.

$$\mathbf{Z}_{\text{PCA}} = \mathbf{W}_P^T \mathbf{X} \tag{11}$$

Essa transformação é definida com base na análise da variância dos dados. Inicialmente, calcula-se a matriz de covariância $\mathbf{C_X}$ dos dados $\mathbf{X}$:

$$\mathbf{C_X} = \frac{1}{n}\mathbf{X}\mathbf{X}^T = \mathbb{E}\{\mathbf{X}\mathbf{X}^T\} \tag{12}$$

o operador $\mathbb{E}\{\cdot\}$ é a esperança matemática ou valor esperado e representa a média ponderada de todos os possíveis valores que a variável pode assumir (Barbetta; Reis; Bornia, 2024). Definido para uma variável aleatória discreta X, com valores possíveis $x_1, x_2, \ldots, x_n$ e respectivas probabilidades $P(X = x_i)$:

$$\mathbb{E}\{X\} = \sum_i x_i P(X = x_i) \tag{13}$$

Ou seja, o valor esperado é a soma dos valores x_i, cada um ponderado pela sua probabilidade $P(X = x_i)$.

Para que a PCA funcione adequadamente, os dados precisam ser centrados em zero, ou seja, cada variável é ajustada subtraindo-se sua média:

$$\mathbf{x} \leftarrow \mathbf{x} - \bar{\mathbf{x}} \tag{14}$$

A primeira componente principal é definida como a combinação linear dos parâmetros originais que possui a maior variância. Essa combinação é representada como:

$$\mathbf{z}_1 = \mathbf{w_{P1}}^T \mathbf{x} \tag{15}$$

onde $\mathbf{w_{P1}}$ é um vetor de pesos que maximiza a variância de $\mathbf{z}_1$. A restrição $\|\mathbf{w_{P1}}\| = 1$ é imposta para evitar soluções triviais onde a norma do vetor cresce indefinidamente.

O problema de encontrar $\mathbf{w_{P1}}$ que maximiza a variância de $\mathbf{z_1}$ pode ser formulado como um problema de otimização, cuja função objetivo é:

$$J_{\mathrm{PCA}}(\mathbf{w}) = \mathbf{w}^T \mathbf{C_X} \mathbf{w} \tag{16}$$

A solução desse problema é obtida calculado os autovalores e autovetores da matriz $\mathbf{C_X}$, onde $\mathbf{w_{P1}}$ é o autovetor correspondente ao maior autovalor de $\mathbf{C_X}$. De forma geral, para as próximas componentes principais, seguimos o mesmo processo, com a restrição adicional de que cada nova componente seja ortogonal às anteriores. Isso garante que:

$$\mathbb{E}\{z_m z_k\} = 0 \quad (k \neq m) \tag{17}$$

Os autovetores são ortogonais, e seus autovalores indicam a quantidade de variância explicada por cada componente principal. Dessa forma, os dados projetados em cada componente principal podem ser representados como:

$$z_k = \mathbf{e_k}^T \mathbf{x} \tag{18}$$

onde $\mathbf{e_k}$ é o k-ésimo autovetor de $\mathbf{C_X}$.

É importante observar que a PCA não faz nenhuma suposição explícita sobre a distribuição dos dados. No entanto, ela é sensível a *outliers*, pois estes podem distorcer a matriz de covariância e, consequentemente, os autovetores (Candés et al., 2011). Além disso, a PCA é limitada a relações lineares entre as variáveis. Se os dados apresentarem relações não lineares significativas, a projeção nos componentes principais pode não capturar adequadamente a estrutura dos dados, o que é especialmente relevante em tarefas como classificação ou agrupamento (Jolliffe, 2002).

Por fim, o uso da PCA é mais eficiente e confiável quando os dados seguem uma distribuição aproximadamente normal, pois isso minimiza os problemas relacionados a *outliers* e garante que a variância calculada esteja próxima de seu valor real (Jolliffe, 2002).

2.4 ANÁLISE DE COMPONENTES INDEPENDENTES

A Análise de Componentes Independentes (ICA)—uma técnica desenvolvida para separar sinais gerados por fontes independentes e inicialmente proposta para resolver o problema de separação cega de fontes (BSS, do inglês *Blind Source Separation*)—pode separar sinais combinados linearmente. É um método não paramétrico que pode identificar os componentes originais que compõem os sinais observados, mesmo quando as combinações desses componentes são complexas; no entanto, não exige uma separação de características baseada em distribuição (Hyvärinen; Oja, 2010).

Para contextualizar a origem do ICA, vamos analisar a separação cega de fontes, onde temos uma mistura de três sinais de fontes distintas misturadas. Um bom exemplo seriam três vozes de locutores distintos e três microfones dispostos em diferentes posições em um mesmo ambiente. Podemos representar os sinais dos microfones por $x_1(t)$, $x_2(t)$ e $x_3(t)$. Os sinais das vozes são representados por $s_1(t)$, $s_2(t)$ e $s_3(t)$, de modo que,

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t) \\ x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t) \end{cases}$$

ou na forma matricial

$$\mathbf{X} = \mathbf{AS_{ICA}} \tag{19}$$

para encontrar $\mathbf{S_{ICA}}$,

$$\mathbf{S_{ICA}} = \mathbf{W_I X} \tag{20}$$

onde $\mathbf{W_I} = \mathbf{A}^{-1}$.

Busca-se então uma matriz $\mathbf{W_I}$ que separe as variáveis, tornando as componentes de $\mathbf{S}$ maximamente independentes. Diferentemente de PCA, que busca a máxima variância, aqui o objetivo é transformar os dados de entrada em dados com a máxima independência, para ICA, não basta os dados serem não correlacionados (Hyvärinen; Oja, 2010).

Não há como determinar a matriz inversa de $\mathbf{A}$, mas pode-se estimar, baseado no teorema do limite central, valores para $w_{i,j}$ que torna a resposta s_i menos gaussiana o possível. O teorema do limite central diz que a soma de variáveis independentes tende a distribuição gaussiana, logo se buscamos variáveis independentes de sinais não gaussianos, devemos estimar coeficientes da matriz $\mathbf{W_I}$ que resultam em $\mathbf{S}$ com distribuição minimamente gaussiana (Hyvärinen; Oja, 2010).

2.4.1 Algoritmo FastICA

O algoritmo FastICA é um método eficiente para encontrar direções $\mathbf{w_I}$ que maximizam a não gaussianidade de uma projeção $\mathbf{w_I}^T\mathbf{x}$. A não gaussianidade é medida pela aproximação da negentropia $J(\mathbf{w_I}^T\mathbf{x})$, sujeita à restrição de que a variância de $\mathbf{w_I}^T\mathbf{x}$ seja unitária (Hyvärinen; Oja, 2000).

Na equação (21), $J(\mathbf{x})$ representa a negentropia, uma medida da não-gaussianidade de uma variável aleatória $\mathbf{x}$.

$$J(\mathbf{x}) = \mathscr{H}(\mathbf{x_{j_{norm}}}) - \mathscr{H}(\mathbf{x}) \tag{21}$$

O termo $\mathscr{H}(\mathbf{x})$ refere-se à entropia de $\mathbf{x}$, enquanto $\mathscr{H}(\mathbf{x}_{\mathbf{j}_{\text{norm}}})$ é a entropia de uma variável aleatória gaussiana com a mesma variância de $\mathbf{x}$. A negentropia é amplamente utilizada no contexto do ICA para medir a independência estatística entre componentes extraídos.

Passos do algoritmo

1. **Pré-processamento:** Centralizar os dados, subtrair a média de $\mathbf{x}$. Esferizar os dados, transformar os dados de forma que sua matriz de covariância seja a identidade. Isto pode ser feito através da decomposição em valores singulares (SVD).

2. **Inicialização:** Escolher um vetor inicial $\mathbf{w_I}$ tal que $\|\mathbf{w_I}\| = 1$.

3. **Iteração de ponto fixo:** A iteração principal do FastICA é dada por:

$$\mathbf{w_I}^{(t+1)} = \mathbb{E}\{\mathbf{x}g(\mathbf{w_I}^{(t)^T}\mathbf{x})\} - \mathbb{E}\{g'(\mathbf{w_I}^{(t)^T}\mathbf{x})\}\mathbf{w_I}^{(t)} \tag{22}$$

   onde $g(\cdot)$ é uma função não linear, e $g'(\cdot)$ é sua derivada. Esta função é escolhida de acordo com a forma da não-gaussianidade que queremos maximizar.

4. **Normalização:** Após cada iteração, normalizar $\mathbf{w_I}$ para garantir que sua norma seja unitária:

$$\mathbf{w_I} = \frac{\mathbf{w_I}}{\|\mathbf{w_I}\|} \tag{23}$$

5. **Critério de Convergência:** O algoritmo para quando a diferença entre o valor de $\mathbf{w}$ em duas iterações sucessivas for suficientemente pequena, ou seja:

$$\|\mathbf{w_I}^{(t+1)} - \mathbf{w_I}^{(t)}\| < \varepsilon \tag{24}$$

   onde ε é um valor pequeno predefinido.

Resultado do fastICA

No final, o vetor $\mathbf{w_I}$ encontrado representa uma direção tal que a projeção $\mathbf{w_I}^T\mathbf{x}$ maximiza a não-gaussianidade, correspondendo a um dos componentes independentes no modelo ICA. Para estimar mais de uma componente independente, deve-se fazer algumas considerações, visto que o passo a passo anterior estima apenas uma componente independente. Para isso pode-se utilizar um algoritmo similar ao de Gram-Schmidt, onde a matriz final obtida é ortonormal (Hyvärinen; Oja, 2010).

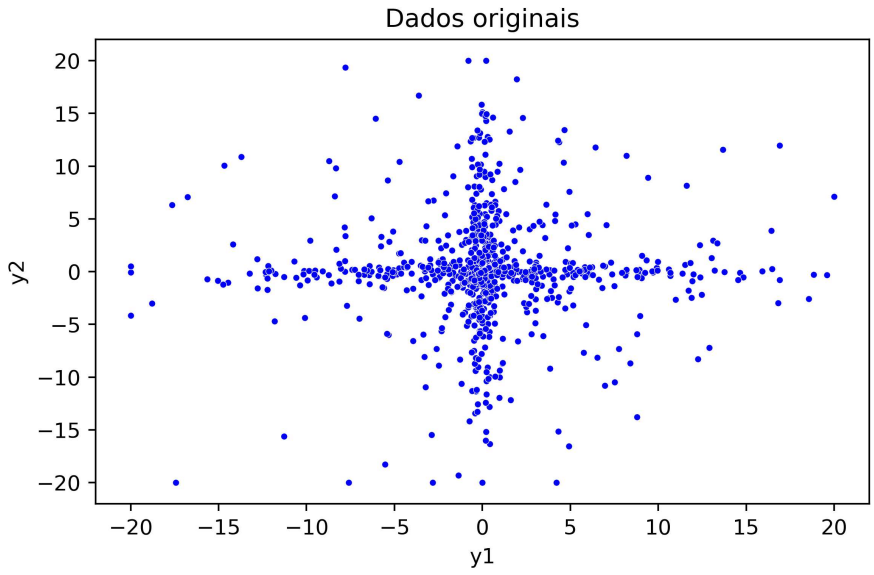
2.4.2 Comparação de PCA e ICA

A Análise de Componentes Principais (PCA) e a Análise de Componentes Independentes (ICA) são técnicas distintas que, embora ambas realizem uma transformação nos dados para gerar um subespaço, possuem objetivos teóricos e metodológicos diferentes. A PCA foca na maximização da variância dos dados, gerando um subespaço que captura a maior parte da variância presente nas amostras. Em contraste, a ICA busca separar os dados de maneira a identificar componentes independentes, o que é especialmente relevante quando se lida com dados não gaussianos.

Para ilustrar essa diferença, considere o exemplo apresentado na Figura 15, onde temos dois dados originais $\mathbf{y}_1$ e $\mathbf{y}_2$. A partir desses dados, é realizada uma mistura que gera um novo dado, conforme mostrado na Figura 16. Enquanto a PCA apenas rotaciona a base dos dados, com o objetivo de alinhar as direções de maior variância, essa transformação não consegue recuperar o sinal original sem mistura. Isso ocorre porque a PCA não leva em consideração a independência dos dados, mas apenas sua variância. Por outro lado, a ICA é capaz de recuperar o sinal original a partir da mistura, já que seu objetivo é separar as fontes independentes.

Note-se que, na Figura 16, os vetores da nova base, representados em preto para PCA e ICA, apresentam direções distintas. Isso evidencia as diferentes abordagens dessas técnicas: enquanto a PCA foca na orientação das componentes de maior variância, a ICA foca na separação dos componentes independentes, sendo, portanto, mais eficaz em situações que envolvem dados não gaussianos e misturados (Jolliffe, 2002; Hyvärinen; Oja, 2010).

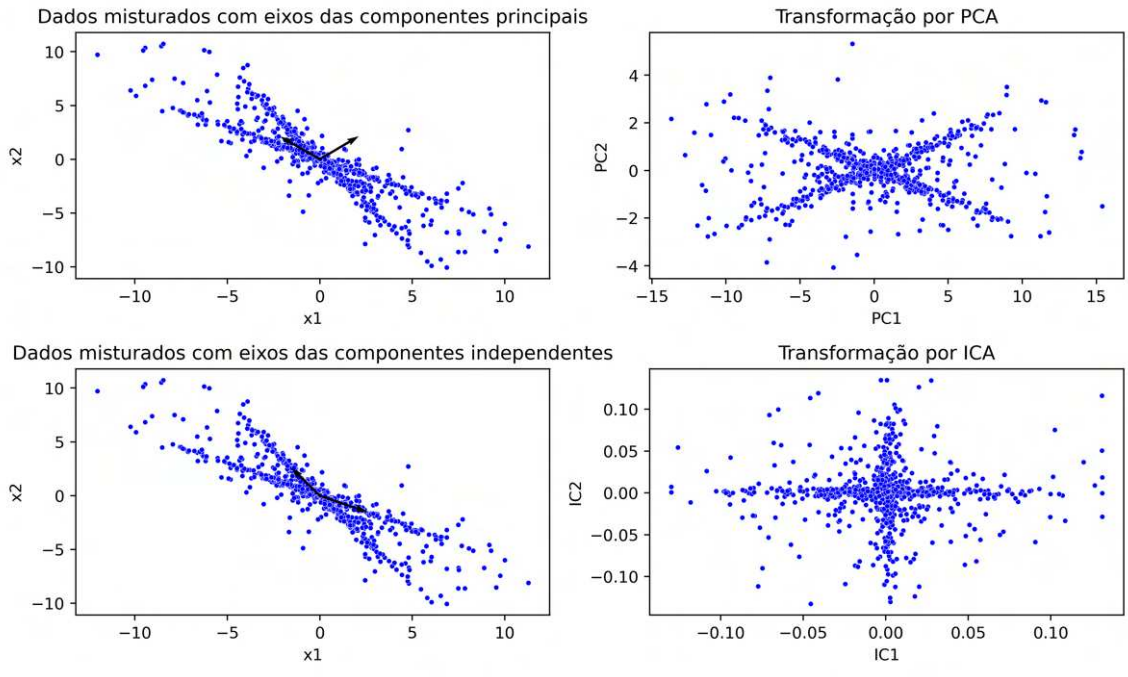
Figura 15 – Dados com distribuição não gaussiana, antes da aplicação dos métodos PCA e ICA.



Fonte: Elaborado pelo autor (2024).

Note-se que, na Figura 16, os vetores da nova base, representados em preto para PCA e ICA, apresentam direções distintas. Isso evidencia as diferentes abordagens dessas técnicas:

Figura 16 – Dados com distribuição não gaussiana após a aplicação dos métodos PCA e ICA.



Fonte: Elaborado pelo autor (2024).

enquanto a PCA foca na orientação das componentes de maior variância, a ICA foca na separação dos componentes independentes, sendo, portanto, mais eficaz em situações que envolvem dados não gaussianos e misturados.

2.5 TESTE DE KRUSKAL–WALLIS

O teste de Kruskal–Wallis é um teste não paramétrico e análogo à ANOVA, é adequado para situações em que a distribuição dos dados é desconhecida. Ele verifica a equivalência estatística dos parâmetros entre diferentes grupos (Hecke, 2012). As emoções podem ser consideradas como classes de resposta, e a distribuição dos parâmetros pode ser analisada em relação a essas classes para verificar se há diferenças significativas entre elas.

Para calcular a estatística de Kruskal–Wallis H, segundo (Hollander; Wolfe; Chicken, 2013) todas as N observações dos k grupos são combinadas e ordenadas do menor para o maior. Seja r_{ij} o posto de $\mathbf{X}_{\mathbf{S}ij}$ nessa classificação conjunta, e defina:

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \quad R_{.j} = \frac{R_j}{n_j}, \quad j = 1, \ldots, k. \tag{25}$$

Por exemplo, R_1 é a soma dos postos atribuídos às observações do tratamento 1, enquanto $R_{.1}$ é o posto médio dessas mesmas observações. A estatística de Kruskal–Wallis H é dada por:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j R_{.j}^2 - 3(N+1), \tag{26}$$

ou equivalentemente,

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1), \tag{27}$$

onde $(N+1)/2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} r_{ij}/N$ é o posto médio atribuído na classificação conjunta.

Para testar a hipótese nula, isto é, todos os grupos tem distribuições iguais:

$$H_0 : [\tau_1 = \cdots = \tau_k], \tag{28}$$

contra a hipótese alternativa geral:

$$H_1 : [\tau_1, \ldots, \tau_k \text{ não são todos iguais}], \tag{29}$$

ao nível de significância α:

- Rejeitar H_0 se $H \geq h_\alpha$; caso contrário, não rejeitar,

onde a constante h_α é escolhida de forma a garantir que a probabilidade de erro do tipo I seja igual a α. Essa constante h_α corresponde ao percentil superior α da distribuição nula de H ($\tau_1 = \cdots = \tau_k$).

Rejeita-se H_0 se $H \geq \chi^2_{k-1,\alpha}$; caso contrário, não se rejeita H_0, (Hollander; Wolfe; Chicken, 2013). Onde $\chi^2_{k-1,\alpha}$ é o ponto do percentil superior α de uma distribuição qui-quadrado com $k-1$ graus de liberdade.

2.6 TESTE DE ANDERSON–DARLING

O teste de Anderson–Darling avalia se uma amostra de dados segue uma distribuição de probabilidade específica. Proposoto em 1954, é baseado na função de distribuição acumulada (Anderson; Darling, 1954).

Matematicamente, a estatística do teste, é definida como:

$$A^2 = -k - \sum_{i=1}^{k} \frac{(2i-1)}{k} \left[\ln F(\zeta_i) + \ln(1 - F(\zeta_{k+1-i}))\right] \tag{30}$$

onde, k é o tamanho da amostra, $F(\zeta_i)$ é a função de distribuição acumulada teórica da distribuição testada e ζ_i representa os dados.

Pode ser utilizado para testar se uma amostra segue ou não uma distribuição normal, sendo utilizado para redução de ruído em imagens Cheng e Chen (2023) e em parâmetros de voz em Pah, Indrawati e Kumar (2022).

2.7 APRENDIZADO DE MÁQUINA

Para o desenvolvimento de um Sistema de Reconhecimento de Emoções (SRE), é comum a utilização de técnicas de aprendizado de máquina, que permitem a programação de computadores para aprenderem a partir de dados (Géron, 2019). O funcionamento de um SRE pode ser descrito de forma estruturada nas seguintes etapas:

1. Seleção ou criação de uma base de dados específica para o problema em questão.
2. Extração de informações relevantes dessa base, como os parâmetros dos sinais de voz.
3. Treinamento de um modelo utilizando m algoritmo de aprendizado de máquina utilizando esses parâmetros, resultando em um modelo de reconhecimento.
4. Validação do modelo treinado para avaliar sua performance.
5. Aplicação do modelo validado em novos dados para a identificação das emoções.

No caso específico de SREs baseados em sinais de voz, o modelo de reconhecimento é geralmente construído a partir de parâmetros extraídos desses sinais. Dessa forma, os sinais de voz não são utilizados diretamente; antes, passam por um processo de extração de parâmetros que serão então aplicados ao modelo. Essa abordagem permite identificar a emoção presente nos sinais de voz de forma eficiente.

A Figura 17 apresenta um diagrama ilustrando o funcionamento de um SRE para o reconhecimento de emoções a partir de sinais de voz. Ressalta-se que a qualidade do modelo de reconhecimento está diretamente relacionada à qualidade da base de dados utilizada. Assim, bases de dados bem construídas e representativas têm maior probabilidade de resultar em modelos mais precisos e robustos. Além disso, a seleção adequada dos parâmetros extraídos dos sinais de voz desempenha um papel essencial no desempenho do sistema, influenciando diretamente a sua capacidade de reconhecimento das emoções.

Os algoritmos de aprendizado de máquina podem ser supervisionados, semisupervisionados, não supervisionados ou de aprendizado por reforço (Géron, 2019). Para um SRE, utilizam-se aprendizados supervisionados, os dados de entrada incluem a solução desejada. Assim, a base de dados deve conter o rótulo da emoção; caso contrário, não será possível criar um modelo eficaz. Alguns algoritmos que já são utilizados para criar modelos de reconhecimento de emoções em sinais de voz incluem:

a) máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) (Abdel-Hamid; Shaker; Emara, 2020; Er, 2020; Özseven, 2019; Ancilin; Milton, 2021; Mishra; Warule; Deb, 2023; Shahin et al., 2023; Vu et al., 2022);

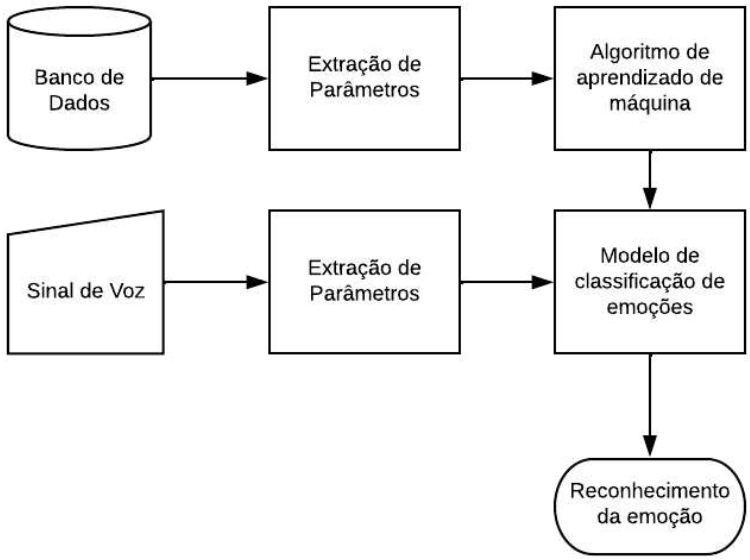
b) k-Vizinhos mais próximos (k-NN, do inglês *k-Nearest Neighbours*) (Aljuhani; Alshutayri; Alahdal, 2021; Verma et al., 2022; Shahin et al., 2023);

c) redes neurais (Chen; Zeng, 2021; Amjad et al., 2022; Aggarwal et al., 2022; Wang et al., 2021);

d) *Ensemble Learning* (Zvarevashe; Olugbara, 2020);

e) modelos de misturas gaussianas (GMMs, do inglês *Gaussian Mixture Models*) (Wu; Liang, 2011).

Figura 17 – Sistema de Reconhecimento de Emoções em Sinais de Voz



Fonte: Elaborado pelo autor (2024).

No contexto de reconhecimento de emoções em sinais de voz, as máquinas de vetores de suporte (SVMs) são frequentemente destacadas como uma abordagem amplamente utilizada. Esse método supervisionado é conhecido por sua eficácia em lidar com problemas de classificação binária e multiclasse, o que pode torná-lo apropriado para tarefas complexas, como a distinção de emoções. Estudos recentes sugerem que a SVM apresenta bons resultados em termos de acurácia, especialmente quando combinada com técnicas de seleção e transformação de características, como PCA e ICA, para otimizar o desempenho do modelo (Özseven, 2019; Palacios et al., 2019a).

2.8 MÁQUINA DE VETORES DE SUPORTE - SVM

Máquina de vetores de suporte SVM do inglês Support Vector Machine é um algoritmo de aprendizado de máquina, proposto por Boser, Guyon e Vapnik (1992). Esta técnica é utilizada para classificação e regressão. Seu objetivo é encontrar um hiperplano que maximiza a margem entre duas ou mais classes, ou seja, a distância mínima entre os pontos mais próximos de classes diferentes. Estes pontos são chamados de vetores de suporte, e determinam a posição e a orientação do hiperplano.

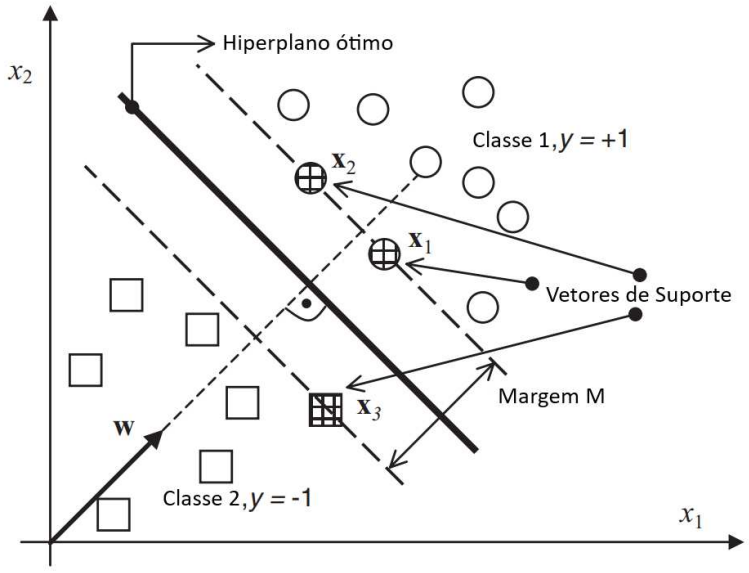
Segundo (Kecman, 2005) o problema de aprendizado para as SVMs é definido da seguinte maneira: existe uma dependência desconhecida e não linear $y = f(x)$ entre um vetor de entrada $\mathbf{x}$ de alta dimensão e uma saída escalar y (ou uma saída vetorial $\mathbf{y}$ no caso de SVMs para múltiplas classes). Não há informações sobre as funções de probabilidade conjuntas subjacentes. Portanto, é necessário realizar um aprendizado livre de suposições sobre a distribuição. A única informação disponível é um conjunto de dados de treinamento $D = \{(x_i, y_i) \in X \times Y\}, i = 1, \ldots, l$, onde l representa o número de pares de dados de treinamento e, consequentemente, o tamanho do conjunto de dados D. Muitas vezes, y_i é denotado como d_i, onde d representa o valor desejado. Assim, as SVMs pertencem às técnicas de aprendizado supervisionado.

Ao utilizar exemplos de treinamento fornecidos, durante a fase de aprendizado, o SVM encontra os parâmetros $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ e b de uma função de decisão $d(\mathbf{x}, \mathbf{w}, b)$, dada pela equação (31),

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^{n} w_i x_i + b, \tag{31}$$

onde $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$, e o escalar b é chamado de viés. Na Figura 18, observa-se o vetor $\mathbf{w}$, que define o hiperplano e a margem determinada por M, onde a saída estabelece a separação entre as classes 1 e 2.

Figura 18 – Representação de um hiperplano definido para separar duas classes utilizando vetores de suporte.



Fonte: (Kecman, 2005)

As máquinas de vetores de suporte (SVMs) são algoritmos supervisionados amplamente utilizados para classificação e regressão. O objetivo principal de uma SVM é encontrar o hiperplano ótimo que separa as classes no espaço de características, maximizando a margem entre os dados de diferentes classes. Para casos onde os dados não são linearmente separáveis, a

SVM incorpora um termo de penalidade que permite erros de classificação, levando à formulação de uma margem suave. Essa formulação é descrita pelo problema de otimização na equação (32):

$$\text{minimizar} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i^k \tag{32}$$

Neste contexto, w representa o vetor de pesos que define o hiperplano, ξ_i são variáveis de folga que medem os erros de classificação, C é o parâmetro de penalidade que controla o equilíbrio entre a margem máxima e a minimização dos erros, e k é o expoente que define a norma da penalidade aplicada.

Quando $k = 1$, a penalidade é baseada na norma $L1$, resultando na SVM L1, que prioriza soluções esparsas, ou seja, com menos vetores de suporte. Já para $k = 2$, a penalidade utiliza a norma $L2$, como na SVM L2, que tende a distribuir os erros de forma mais uniforme. Ambos os modelos são problemas de programação convexa resolvidos por métodos numéricos eficientes (Kecman, 2005).

O parâmetro C é ajustado empiricamente, e seu valor influencia diretamente o desempenho do modelo, podendo ser validado com métricas apropriadas para garantir uma separação eficaz dos dados.

Para problemas cujos dados não podem ser separados linearmente, pode-se utilizar uma função de kernel que mapeia os dados para um espaço de maior dimensão onde a separação linear é possível.

As funções de kernel de base polinomial e de base radial (também conhecidas como RBF) são descritas abaixo:

a) **Funções de base polinomial**

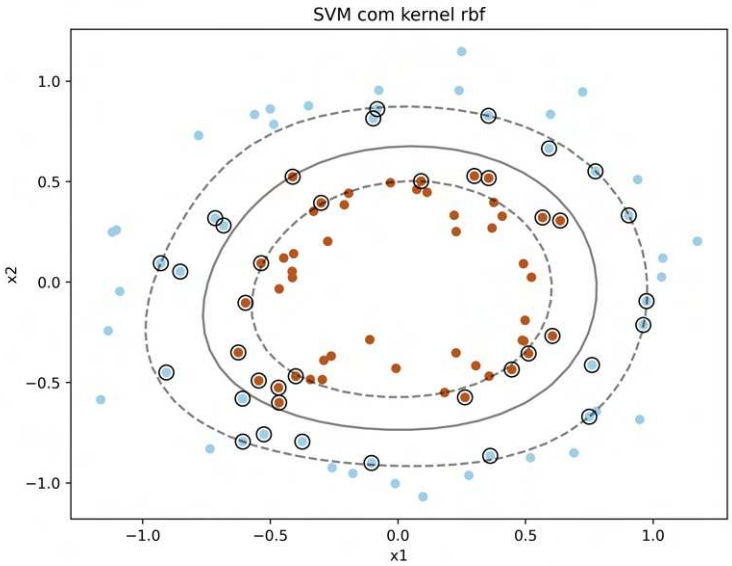
$$K(x_i, x_j) = (\langle x_i, x_j \rangle + c_0)^d \tag{33}$$

b) **Funções de base radial (RBF)**

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{34}$$

O núcleo polinomial, pode ser linear quando $d = 1$. Quadrático se $d = 2$, ou cúbico se $d = 3$. O parâmetro γ define o quão influente os vetores de suporte podem ser no treinamento e seu valor deve ser maior que zero. A representação de um SVM com Kernel de base radial pode ser vista na Figura 19, com a margem tracejada e o hiperplano ilustrado com uma linha contínua, destaca-se os dados que estão na região de margem, como no caso linear são os vetores de suporte para o SVM.

Figura 19 – Representação de um hiperplano definido para separar duas classes utilizando vetores de suporte com Kernel de base radial.



Fonte: Elaborado pelo autor (2024).

2.9 MÉTRICAS PARA AVALIAR O DESEMPENHO DE UM MODELO

Nesta pesquisa, quatro métricas foram utilizadas para avaliar o desempenho dos modelos. A acurácia média por classe de um classificador, ou a acurácia média, dada pela Equação (35), é uma métrica que representa o número de previsões corretas feitas por um modelo. A precisão, dada pela Equação (36), é o acordo médio por classe entre os rótulos de classe dos dados e os do classificador. A revocação, dado pela Equação (37), é a média da eficácia do classificador em identificar os rótulos de classe. O *F-score*, dado pela Equação (38), é a média harmônica entre a precisão e revocação (Sokolova; Lapalme, 2009).

$$\text{Acurácia} = \frac{\sum_{i=1}^{l} \text{vp}_i + \text{vn}_i}{\sum_{i=1}^{l} (\text{vp}_i + \text{fn}_i + \text{fp}_i + \text{vn}_i)}, \tag{35}$$

$$\text{Precisão} = \frac{\sum_{i=1}^{l} \text{vp}_i}{\sum_{i=1}^{l} (\text{vp}_i + \text{fp}_i)}, \tag{36}$$

$$\text{Revocação} = \frac{\sum_{i=1}^{l} \text{vp}_i}{\sum_{i=1}^{l} (\text{vp}_i + \text{fn}_i)}, \tag{37}$$

$$\text{F-score} = \frac{(b^2+1) \times \text{Precisão} \times \text{Revocação}}{b^2 \times \text{Precisão} + \text{Revocação}}, \tag{38}$$

onde vp_i representa os verdadeiros positivos, fp_i representa os falsos positivos, fn_i representa os falsos negativos e vn_i representa os verdadeiros negativos para a *i*-ésima classe. A constante *b* é um fator de ponderação, e neste estudo, *b* foi definido como 1.

Os valores das métricas apresentadas foram analisados para comparar o desempenho dos modelos testados neste estudo. A análise conjunta da acurácia, precisão, *Recall* e F-score permitiu uma avaliação mais robusta, considerando diferentes aspectos do comportamento dos classificadores. Cada métrica contribuiu para identificar as forças e limitações dos modelos, oferecendo uma visão abrangente sobre sua eficácia no reconhecimento de emoções em sinais de voz.

Diante do embasamento teórico apresentado, é possível compreender os principais conceitos e técnicas que fundamentam este estudo, desde as teorias das emoções até os métodos de extração de parâmetros, análise estatística e a revisão sobre o uso de Máquinas de Vetores de Suporte (SVM). Com essas bases estabelecidas, o próximo capítulo apresenta a metodologia adotada, detalhando os procedimentos experimentais, as estratégias de escolha e processamento de dados, bem como a aplicação das técnicas descritas para a análise e modelagem de um sistema de reconhecimento de emoções.

3 METODOLOGIA

O método proposto foi cuidadosamente estruturado em diversas etapas, como ilustrado na Figura 20. Este capítulo detalha cada uma dessas etapas, desde a seleção inicial dos parâmetros acústicos até a aplicação das técnicas de Análise de Componentes Principais (PCA) e Análise de Componentes Independentes (ICA).

Inicialmente, realizou-se uma seleção preliminar dos parâmetros com base em dois critérios principais: variância e a diferença entre classes, avaliada por meio do teste de Kruskal-Wallis. Após essa etapa, foram conduzidos dois procedimentos distintos para avaliar o impacto de filtrar os dados com base na distribuição normal ao utilizar PCA:

a) Os parâmetros com distribuição normal, identificados por meio do teste de Anderson-Darling, foram submetidos à técnica de Análise de Componentes Principais (PCA). Paralelamente, todos os parâmetros selecionados previamente foram submetidos à Análise de Componentes Independentes (ICA) e

b) Todos os parâmetros selecionados foram tratados conjuntamente, sem separação por distribuição, sendo submetidos tanto à PCA quanto à ICA.

A escolha por essas duas abordagens teve como objetivo principal avaliar se o uso dos parâmetros com distribuição normal influencia os resultados obtidos no modelo de reconhecimento de emoções. A comparação foi realizada para determinar se a aplicação seletiva da PCA aos parâmetros com distribuição normal e da ICA aos demais preserva mais informações e melhora o desempenho do modelo em relação ao uso indiscriminado das técnicas para todos os parâmetros.

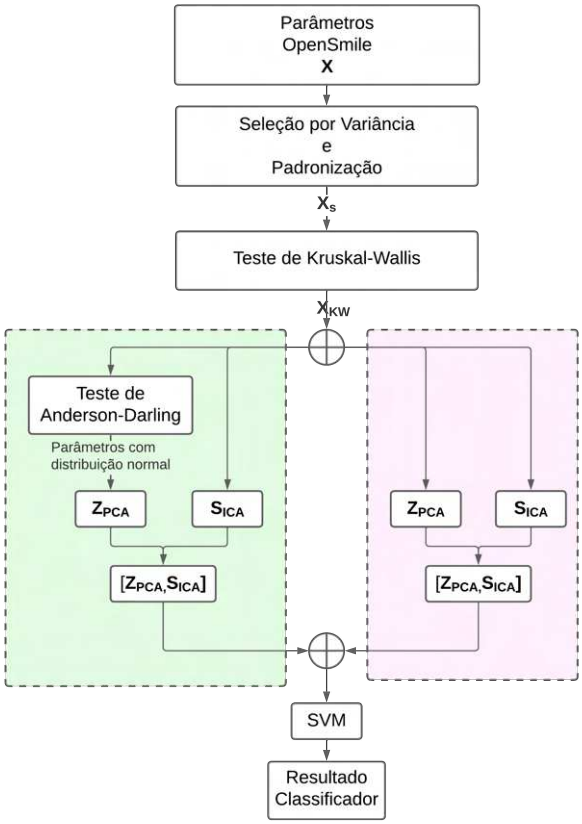
Ao combinar PCA e ICA, aproveitam-se os pontos fortes de ambos os métodos. O PCA é aplicado para reduzir os parâmetros, especialmente ao lidar com conjuntos de dados de alta dimensão (Jolliffe, 2002). Esta etapa garante que a principal variância nos dados seja capturada. Em seguida, o ICA é aplicado aos dados com dimensões reduzidas para extrair componentes independentes que podem revelar estruturas subjacentes adicionais não capturadas apenas pelo PCA.

Essa abordagem, além de garantir a preservação das características mais relevantes dos dados, permitiu uma análise mais aprofundada das diferentes naturezas dos atributos. Nos tópicos subsequentes, cada etapa do método será detalhada, fornecendo uma visão completa de sua aplicação.

Ao combinar PCA e ICA, tanto a variância principal (via PCA) quanto a distribuição independente e não normal (via ICA) nos dados são capturadas, o que é vantajoso ao lidar com conjuntos de dados complexos que contêm parâmetros com distribuições normais e não normais.

O método de seleção de parâmetros em múltiplas etapas proposto tem como objetivo melhorar o processo filtrando os dados menos informativos. Esse processo de filtragem foi projetado para limpar os dados de entrada para PCA e ICA, isso exclui dados que não possuem nenhuma variação, ou seja, tem o mesmo valor para todos os áudios, independente de sua classe.

Figura 20 – Fluxograma do método proposto, iniciando com a filtragem inicial dos parâmetros baseada na variância e no teste de Kruskal-Wallis. Em seguida, dois procedimentos distintos são destacados: um deles, destacado em verde, utiliza o teste de Anderson-Darling para filtrar os parâmetros com base na distribuição antes de aplicá-los em PCA. O outro, destacado em rosa, emprega todos os parâmetros selecionados após a etapa de filtragem por Kruskal-Wallis. Em ambos os casos, a classificação é realizada utilizando o modelo SVM.



Fonte: Elaborado pelo autor (2024).

Além de remover dados com baixa variância que podem aumentar o custo computacional e não conter informações relevantes. Um estudo prévio da influência da variância na classificação de emoções foi feito pelo autor em (Kingeski; Schueda; Paterno, 2022), onde parâmetros com baixa variância mostraram ter uma baixa contribuição para o reconhecimento de emoções.

Além da seleção inicial por baixa variância, fez-se o suo do teste de Kruskal-Wallis, conforme utilizado em (Ravi; Taran, 2023). Todos os scripts foram desenvolvidos em Python 3.12 (Rossum; Jr, 1995) e estão disponíveis no seguinte link https://github.com/rkingeski/pca_ica_speech_emotion, acessado em 1 de setembro de 2024.

3.1 *OPENSMILE*

Um sistema de reconhecimento de emoções pode ser desenvolvido utilizando parâmetros acústicos da voz. Neste trabalho, foi utilizado o *OpenSmile* (Eyben; Wöllmer; Schuller, 2010),

um *toolkit* de código aberto com versão para Python. O *OpenSmile* é um extrator de parâmetros de voz amplamente utilizado para reconhecimento de emoções (Xie; Zhu; Hu, 2023; Özseven, 2019; Song, 2019; Jiang et al., 2017). Optou-se por essa ferramenta devido ao grande número de parâmetros que podem ser extraídos, além de sua facilidade de uso e de possuir grupos de parâmetros específicos para reconhecimento emocional. Os conjuntos disponíveis no *OpenSmile*, criados para esse propósito, estão listados na Tabela 6.

Tabela 6 – Conjuntos de Parâmetros do *OpenSmile* para Reconhecimento de Emoções

Nome do Conjunto	Número de Parâmetros Acústicos	Descrição
Emobase	84	Conjunto de parâmetros básicos para reconhecimento de emoções (Wöllmer et al., 2009).
Emobase Affective	188	Extensão do Emobase com parâmetros afetivos adicionais (Wöllmer et al., 2009).
ComParE	6373	Conjunto de parâmetros utilizado para o desafio de Reconhecimento de Emoções (Schuller et al., 2013).
GeMAPS	62	Conjunto de parâmetros utilizado para o Reconhecimento de Emoções (Eyben; Wöllmer; Schuller, 2010).
eGeMAPS	88	Conjunto de parâmetros utilizado para o Reconhecimento de Emoções (Eyben; Wöllmer; Schuller, 2010).

Devido à sua abrangência, optou-se pelo conjunto de características acústicas *ComParE*. Com 6373 atributos, derivados da combinação de 64 descritores de baixo nível aplicados a descritores funcionais (Eyben; Wöllmer; Schuller, 2010; Weninger et al., 2013), esse conjunto permitiu uma análise detalhada e a aplicação de diversas técnicas de redução de dimensionalidade.

As Tabelas 7 e 8 apresentam, respectivamente, os descritores de baixo nível e as categorias dos descritores funcionais que compõem o conjunto *ComParE* (Eyben; Wöllmer; Schuller, 2010).

A ferramenta *OpenSMILE* foi utilizada para extrair esses descritores acústicos, permitindo a obtenção dos parâmetros relevantes para a análise emocional. A extração foi realizada com a versão *OpenSMILE* 3.0, configurada para gerar os atributos descritos no conjunto *ComParE*. Com isso, os dados processados foram prontamente disponibilizados para as etapas subsequentes de análise e modelagem.

Tabela 7 – Descritores de Baixo Nível do OpenSmile

Grupo de Parâmetros	**Descrição**
Forma de Onda (Temporal)	Cruzamentos por zero, extremos, sinal de DC, energia do sinal, média Quadrática & logarítmica, intensidade Sonora e aproximação de intensidade Sonora
FFT	Espectro de freq. e fase, magnitude (linear, dB, dBA)
ACF	Autocorrelação e cepstrum
Espectro Mel/Bark	Bandas de espectro Mel/Bark (0 até N_{mel})
Espectro de semitons	Baseado em FFT e baseado em filtros
Cepstral	características cepstrais, ex.: MFCC, PLPCC
Pitch	F0 via autocorrelação e métodos SHS
Probabilidade de Voz	qualidade Vocal: HNR, jitter, shimmer
LPC	Coeficientes LPC, coeficientes de reflexão, resíduos, pares espectrais lineares (LSP)
Auditivo	Espectros auditivos e coeficientes PLP
Formantes	frequências centrais e larguras de banda
Energia Espectral	N bandas definidas, pontos de *roll-off* múltiplos, centróide, entropia, fluxo e posição relativa de máx./mín.
Tonal	Características baseadas em CHROMA e CENS

Tabela 8 – Categorias de Parâmetros Funcionais

Categoria	**Descrição**
Extremos	Valores extremos, posições e intervalos
Médias	Média aritmética, quadrática e geométrica
Momentos	Desvio padrão, variância, curtose, assimetria
Percentis	Percentis e intervalos de percentil
Regressão	Coeficientes de aproximação linear e quadrática, erro de regressão e centróide
Picos	Número de picos, distância média entre picos, amplitude média dos picos
Segmentos	Número de segmentos baseados em limiar delta, comprimento médio dos segmentos
Valores de Amostra	Valores do contorno em posições relativas configuráveis
Tempo/Durações	Tempos de níveis alto e baixo, tempos de subida/descida, duração
Inícios (Onsets)	Número de inícios, posição relativa do primeiro/último início/fim
DCT	Coeficientes da Transformação Discreta de Cosseno (DCT)
Cruzamentos por Zero	Taxa de cruzamentos por zero, taxa de cruzamentos médios

3.2 BASES DE DADOS

Neste estudo, foram utilizadas três bases de dados amplamente reconhecidas na área de reconhecimento de emoções, conforme destacado por (Hashem; Arif; Alghamdi, 2023). Essas bases de dados, apesar dos desafios relacionados à padronização e à comparabilidade,

continuam sendo essenciais para o desenvolvimento e a avaliação de novas técnicas de redução de parâmetros. A escolha dessas bases, *Berlin Emotional Database* (EMODB), *Ryerson Audio-Visual Database* (RAVDESS) e *Surrey Audio-Visual Expressed Emotion* (SAVEE), se justifica pela diversidade de emoções representadas, pela qualidade das gravações e pela ampla utilização em pesquisas anteriores (Xie; Zhu; Hu, 2023; Özseven, 2019; Liu et al., 2018; Vu et al., 2022).

Embora existam variações entre as bases de dados em termos de número de falantes, idiomas e condições de gravação, elas oferecem um conjunto de dados consistente para a comparação e avaliação de diferentes métodos. Outra limitação é a impossibilidade de garantir a generalização dos resultados, devido às diferenças entre as bases de dados utilizadas. Apesar desses desafios, o uso dessas bases de dados permite que os resultados obtidos sejam comparados com os de outros estudos, contribuindo para o avanço da área.

3.2.1 *Berlim Emotional Database*

A base de dados de Berlim foi construída por 10 atores, sendo 5 do sexo masculino e 5 do sexo feminino. Cada ator produziu 10 frases no idioma alemão, 5 curtas e 5 mais longas. As emoções expressas pelos atores foram: Raiva, medo, tédio, nojo, felicidade, tristeza e um estado neutro que representa a ausência de emoções. Esta base foi avaliada por 20 pessoas na qual selecionaram-se 535 frases, as expressões vocais com uma taxa de reconhecimento superior a 80% e naturalidade superior a 60%. Destas 302 frases enunciadas por vozes femininas e 233 por vozes masculinas (Burkhardt et al., 2005).

3.2.2 SAVEE

Base de dados audiovisual em inglês, esta base consiste em gravações de 4 atores masculinos em 7 diferentes emoções, totalizando 480 enunciados no idioma inglês britânico. Para validar as gravações foram avaliadas por 10 pessoas. As emoções são separadas categoricamente em: raiva, nojo, medo, felicidade, tristeza, surpresa e neutro (Haq; Jackson; Edge, 2008).

3.2.3 RAVDESS

A base de dados audiovisual RAVDESS é composta por gravações realizadas por 24 atores profissionais, sendo 12 do gênero masculino e 12 do gênero feminino. Cada ator produziu 104 vocalizações únicas no idioma inglês americano, expressando diferentes emoções, como felicidade, tristeza, raiva, medo, surpresa, nojo, calma e neutralidade. A base de dados oferece tanto vídeos quanto áudios; entretanto, para o presente estudo, foi utilizada exclusivamente a parte de áudio do RAVDESS, totalizando 1440 gravações (Livingstone; Russo; Najbauer, 2018).

3.3 SELEÇÃO POR VARIÂNCIA

A variância é uma medida que indica a dispersão de um conjunto de dados em relação à sua média, sendo útil para identificar parâmetros com maior variabilidade ou com ausência de variação significativa (Duda; Hart; Stork, 2001). Essa característica torna a variância uma ferramenta valiosa na seleção de parâmetros, pois permite distinguir aqueles que podem conter informações relevantes (Barbetta; Reis; Bornia, 2024).

No trabalho, foi empregada uma etapa de pré-seleção baseada na variância antes da aplicação do PCA e ICA. Essa abordagem teve como objetivo eliminar parâmetros constantes ou com variâncias muito baixas, garantindo que apenas os parâmetros com maior potencial informativo fossem considerados nas transformações subsequentes.

Dados que possuem variância nula são constantes e não tem informação nenhuma que possa distinguir classes em um conjunto de dados. A exclusão dos parâmetros constantes é fundamental, uma vez que eles não contribuem de forma alguma para o modelo e, portanto, não fornecem informações para a análise (Duda; Hart; Stork, 2001). Os parâmetros com baixa variância podem ter uma contribuição relativamente menor para o modelo, o que pode resultar em uma representação inadequada dos dados. Assim, ao remover os parâmetros com baixa variância, busca-se garantir uma representação mais significativa e informativa dos dados, o que pode resultar em uma melhor performance dos algoritmos de redução de dimensionalidade aqui utilizados, isto é, PCA e ICA.

A variância é um importante indicador de informação, porém não é único, além disso a quantidade de informação não garante que esta será útil para discriminar as classes (Duda; Hart; Stork, 2001).

A variância nos dados foi calculada de acordo com a Equação (39),

$$\mathrm{Var}(\mathbf{X}_{\mathrm{norm}}) = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2. \tag{39}$$

onde $\bar{x}_j$ denota a média dos valores da coluna. Para cada coluna, a variância foi calculada após normalizar a Matriz $\mathbf{X}$ para um intervalo entre 0 e 1. m representa o número de registros de áudio ou o número de linhas na matriz $\mathbf{X}_{\mathrm{norm}}$.

Para a redução de parâmetros e uma análise mais direcionada, as colunas de $\mathbf{X}_{\mathrm{norm}}$ foram selecionadas com base em sua variância, utilizando um valor de corte arbitrário de 1×10^{-5} como a variância mínima aceitável. Esse valor foi escolhido com base na análise preliminar dos dados, buscando equilibrar a retenção de parâmetros. O limite de 1×10^{-5} foi considerado adequado, pois resultou na exclusão de menos de 10% dos parâmetros em todos os conjuntos de dados analisados nesta pesquisa.

Os dados selecionados por variância foram padronizados para uma média de ($\mu = 0$) e desvio padrão de ($\sigma = 1$). A matriz de dados padronizados pode ser representada pelas Equações (40) e (41).

$$x'_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_j}, \tag{40}$$

onde

$$\mathbf{X_s} = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1k} \\ x'_{21} & x'_{22} & \cdots & x'_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x'_{m1} & x'_{m2} & \cdots & x'_{mk} \end{bmatrix}. \tag{41}$$

As dimensões de $\mathbf{X_s}$ são $m \times k$, onde m é o número de registros de áudio de entrada e k é o número de parâmetros selecionadas pela variância. Os valores específicos para cada banco de dados são detalhados na Tabela 10.

3.4 SELEÇÃO DE PARÂMETROS PELO TESTE DE KRUSKAL–WALLIS

No que diz respeito ao teste de Kruskal–Wallis, a hipótese nula foi rejeitada para cada parâmetro que apresentou evidências estatísticas de diferenças de distribuição entre os grupos, com um nível de significância de 1%, garantindo um erro do tipo I de 1%. O valor de $\alpha = 1\%$ foi escolhido para minimizar o número de parâmetros selecionados. Consequentemente, esses parâmetros foram considerados críticos para o modelo, pois a rejeição da hipótese nula sugere que pelo menos um grupo apresenta uma distribuição estatisticamente distinta em relação aos outros.

Embora o teste identifique parâmetros com distribuições estatisticamente diferentes entre os grupos, ele não especifica qual classe exibe a distribuição distinta. Portanto, um novo subconjunto foi definido pela Equação (42), do qual foram excluídas os parâmetros sem diferenças estatísticas.

$$\mathbf{X_{KW}} = \{x_j \mid x_j \in X_s, \text{p-valor}(x_j) \leq \alpha\}. \tag{42}$$

$\mathbf{X_{KW}}$ representa os parâmetros filtrados pelo teste de Kruskal–Wallis. O número de parâmetros para cada banco de dados após o filtro está mostrado na Tabela 12.

A partir dos resultados, gráficos de violino ou do inglês *violin plot*, foram utilizados para ilustrar a distribuição dos parâmetros selecionados.

O *violin plot* combina a representação da densidade de probabilidade com medidas estatísticas, como a mediana e os quartis. Essa visualização facilita a análise comparativa das distribuições dos parâmetros entre as diferentes emoções, destacando padrões relevantes. A associação do gráfico de violino com os resultados do Kruskal-Wallis permitiu observar tanto a dispersão quanto a tendência central dos dados analisados, fornecendo uma base para a interpretação dos parâmetros mais relevantes no contexto do reconhecimento de emoções.

3.5 DISTRIBUIÇÃO DOS PARÂMETROS

A distribuição dos parâmetros é o fundamental para aplicação do método proposto. Os dados foram separados em dois grupos: um com distribuições normais e outro com distribuições não normais. Para realizar essa classificação, utilizou-se o teste de Anderson-Darling, uma técnica estatística que verifica a adequação dos dados a uma distribuição específica (Yazici; Asma, 2007).

O nível de significância estabelecido para o estudo foi de 5%, correspondente a um p-valor crítico de 0,05, utilizado como critério de rejeição da hipótese nula. A hipótese nula do teste de Anderson-Darling assume que os dados da amostra seguem uma distribuição normal, enquanto a hipótese alternativa sugere que os dados não seguem essa distribuição. Esse valor de 5% foi escolhido com base no nível de significância pré-estabelecido para o estudo, visando garantir uma margem de erro controlada nas conclusões. O teste foi aplicado a cada uma das 6373 colunas da matriz de dados $\mathbf{X}$, permitindo uma análise detalhada da normalidade dos dados em cada parâmetro considerado.

3.6 IMPLEMENTAÇÃO DA ANÁLISE DE COMPONENTES PRINCIPAIS

Para reduzir a dimensionalidade dos dados aplicou-se PCA nos dados com distribuição normal. Escolheu-se este subgrupo para que o PCA fosse otimizado, conforme descrito na fundamentação teórica, esta técnica apresenta melhores resultados se aplicada em dados com distribuição normal. O objetivo aqui é analisar a distribuição dos parâmetros, portanto, aplicou-se PCA no conjunto todo selecionado pelo teste de Kruskal-Wallis e em um subconjunto de parâmetros com distribuição normal. Além disso, esperava-se que a combinação de PCA e ICA pudesse melhorar o desempenho do classificador.

Os componentes principais foram calculados conforme a Equação (43).

$$\mathbf{Z_{PCA}} = \mathbf{W_P X_n}, \tag{43}$$

onde $\mathbf{W_P}$ é a matriz de autovetores correspondentes aos autovalores em ordem decrescente da matriz de covariância $\mathbf{C_{X_n}}$.

Usando a PCA, os dados foram transformados em novas variáveis que combinavam linearmente as variáveis originais. Como o objetivo era encontrar a maior variância, podemos assumir um sinal de entrada neste trabalho como X_n com dimensões de $m \times l$, sendo m definido como o número de registros de áudio de entrada, onde l é o número de parâmetros com distribuição normal.

3.7 IMPLEMENTAÇÃO DA ANÁLISE DE COMPONENTES INDEPENDENTES

A análise de componentes independentes (ICA)—uma técnica desenvolvida para separar sinais gerados por fontes independentes e inicialmente proposta para resolver o problema de separação cega de fontes (BSS)—pode separar sinais combinados linearmente. É um método não-paramétrico que pode identificar os componentes originais que compõem os sinais observados, mesmo quando as combinações desses componentes são complexas (Hyvärinen; Oja, 2010).

Neste contexto, não se lida diretamente com sinais, mas sim com parâmetros, que servirão como entrada para o classificador. Aplicou-se a ICA para separar os componentes independentes subjacentes com base na premissa de que esses parâmetros representam combinações de componentes e reduzir o número de parâmetros de entrada, mantendo informações importantes para a classificação. Esta abordagem busca melhorar a capacidade do classificador de reconhecer padrões distintos, assumindo a separação virtual dos parâmetros em componentes independentes.

O modelo ICA pode ser descrito da seguinte maneira: considerando os parâmetros fornecidos como sinais independentes $\mathbf{S_{ICA}} = [s_{1,ica}\, s_{2,ica}\, s_{3,ica} \ldots s_{i,ica}]$, se e somente se eles forem independentes e tiverem uma distribuição não-normal, quando misturados (possivelmente com apenas um componente com distribuição normal incorporado), novos sinais são criados, de forma que $\mathbf{X_{KW}} = [x_{kw1}\, x_{kw2}\, x_{kw3} \ldots x_{kwi}]$, que é uma combinação dos parâmetros $\mathbf{S_{ICA}}$ (Hyvärinen; Oja, 2010). Podemos descrever a combinação em questão com a Equação (44).

$$\mathbf{X_{KW}} = \mathbf{AS_{ICA}}, \tag{44}$$

onde $\mathbf{A}$ é a matriz de mistura dos sinais independentes ou, neste caso, dos parâmetros misturados.

Sob as condições de independência dos sinais, pode-se estimar uma matriz $\mathbf{W}$ que resolve o sistema e recupera os sinais, conforme descrito na Equação (45).

$$\mathbf{S_{ICA}} = \mathbf{W_I X_{KW}}. \tag{45}$$

Neste caso, supondo que tem-se uma fonte desconhecida, representada por $\mathbf{S_{ICA}}$, que, quando misturada, resulta nos parâmetros de voz $\mathbf{X_{KW}}$. Não há uma relação direta nesta mistura; em vez disso, criou-se a hipótese de que há uma combinação de valores em $\mathbf{S_{ICA}}$ que resulta em $\mathbf{X_{KW}}$. Portanto, ao separar os parâmetros, obtemos novos dados, representados por $\mathbf{S_{ICA}}$.

Aplicamos o algoritmo FastICA, que foi implementado na biblioteca Scikit-learn. O algoritmo foi configurado de acordo com as diretrizes teóricas, e os detalhes específicos de sua configuração estão descritos na Tabela 9.

A aplicação da análise de componentes independentes (ICA) neste contexto visa aprimorar a representação dos parâmetros de entrada para a classificação, reduzindo redundâncias e extraindo parâmetros relevantes.

Tabela 9 – Parâmetros para o algoritmo FastICA.

Parâmetros	**Valores**
algorithm	Deflation
whiten	Unit variance
fun	logcosh
fun_args	'alpha':1.0
tol	1×10^{-4}
max_iter	500
w_init	None

Ao separar componentes independentes, espera-se que o modelo obtenha uma estrutura mais informativa dos dados, o que pode contribuir para um melhor desempenho na identificação de padrões.

3.8 MÁQUINA DE VETORES DE SUPORTE

O SVM é frequentemente utilizado em pesquisas sobre reconhecimento de emoção na voz (Chen et al., 2012; Song, 2019; Xie; Zhu; Hu, 2023; Özseven, 2019). Nesta pesquisa, optou-se por utilizar este modelo, já que o foco não era a seleção do algoritmo mais adequado para a geração do modelo.

Para os modelos gerados neste trabalho, foi adotada uma função de base radial para o kernel na classificação por vetores de suporte (SVC). Inicialmente, os parâmetros de ajuste do kernel usados para SVC foram os parâmetros padrão da biblioteca Scikit-learn (Pedregosa et al., 2011). Posteriormente, foram realizados ajustes nos parâmetros do modelo C e γ. O parâmetro γ foi mantido como padrão, conforme representado na Equação (46), onde n representa o número de pontos de dados de entrada e σ_X denota a variância nos dados de entrada. Valores de C de 0.1, 1, 10, 100 e 1000 foram testados, com um valor final de 100 escolhido.

$$\gamma = \frac{1}{(n \cdot \sigma_X)}. \tag{46}$$

Os modelos foram empregados usando conjuntos de dados preexistentes previamente descritos na literatura para testar e validar o método de seleção de parâmetros proposto neste trabalho, conforme mostrado na Seção 3.2. A acurácia foi avaliada usando o método de validação cruzada com $k = 10$, como sugerido por (Lieskovská et al., 2021).

Com a metodologia detalhada, descrevendo a aplicação das técnicas selecionadas, estabelece-se a base para a análise dos fenômenos investigados. A partir dessas etapas, os dados foram processados e os modelos desenvolvidos, possibilitando a avaliação de seus desempenhos e a interpretação dos resultados. No próximo capítulo, são apresentados os resultados obtidos, destacando as principais descobertas e sua relação com os objetivos deste estudo.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da análise dos dados experimentais, com o objetivo de investigar as diferenças nas distribuições dos parâmetros acústicos de voz para o reconhecimento de emoções. Utilizando diferentes métodos de análise estatística e técnicas de redução de dimensionalidade, os resultados são apresentados de forma a evidenciar as características mais relevantes dos dados coletados a partir das bases de dados EMODB, SAVEE e RAVDESS.

Para a análise estatística, foi utilizado o teste de Kruskal-Wallis, que permitiu identificar diferenças significativas entre os grupos de emoções em relação aos parâmetros acústicos extraídos. A seguir, as distribuições dessas variáveis são ilustradas através de gráficos, como os *violin plots*, que permitem visualizar a variação e a concentração dos dados, evidenciando se há ou não distinções entre as classes emocionais.

A organização deste capítulo está estruturada de forma a apresentar primeiramente a análise de cada base de dados individualmente, seguida de uma discussão sobre as implicações dos resultados encontrados. A interpretação dos dados visa proporcionar uma compreensão mais aprofundada sobre a relação entre os parâmetros acústicos da voz e as emoções, com foco na aplicabilidade desses resultados para sistemas de reconhecimento emocional.

4.1 SELEÇÃO DE PARÂMETROS POR VARIÂNCIA E TESTES ESTATÍSTICOS

Inicialmente, 6373 parâmetros foram extraídos usando o openSMILE. Após a filtragem de baixa variância, aplicou-se o teste de Kruskal–Wallis para remover parâmetros sem diferenças estatísticas significativas entre as classes (Hecke, 2012), verificando se pelo menos uma emoção apresentava uma diferença de distribuição em relação às outras. Os parâmetros sem essas diferenças foram descartados, formando um novo conjunto de dados.

Esse conjunto foi então dividido em dois subconjuntos, com o teste de Anderson–Darling separando os parâmetros com e sem distribuição normal. A Tabela 10 apresenta a aplicação dos métodos: a primeira linha mostra os dados selecionados por variância; as linhas seguintes, a divisão por distribuição e por Kruskal–Wallis nos dados selecionados, respectivamente.

Tabela 10 – Parâmetros filtrados por variância e por distribuição.

Tipo de Parâmetro	SAVEE	RAVDESS	Berlin EmoDB
Número de áudios	480	1440	535
Variância	5881	5768	5975
Normal	599	171	475
Não-Normal	5282	5597	5500
Variância e KW	4382	4845	5407
Normal KW	415	133	397
Não-Normal KW	3967	4712	5010

Após a aplicação das etapas de filtragem dos parâmetros, os dados foram organizados em dois subconjuntos de acordo com a distribuição. Os dados da Tabela 10 descrevem os resultados nas três bases de dados: EmodDB, SAVEE e RAVDESS. Separados de acordo com cada etapa e que foram utilizados para comparar os resultados do modelo.

4.1.1 Teste de distribuição Anderson–Darling

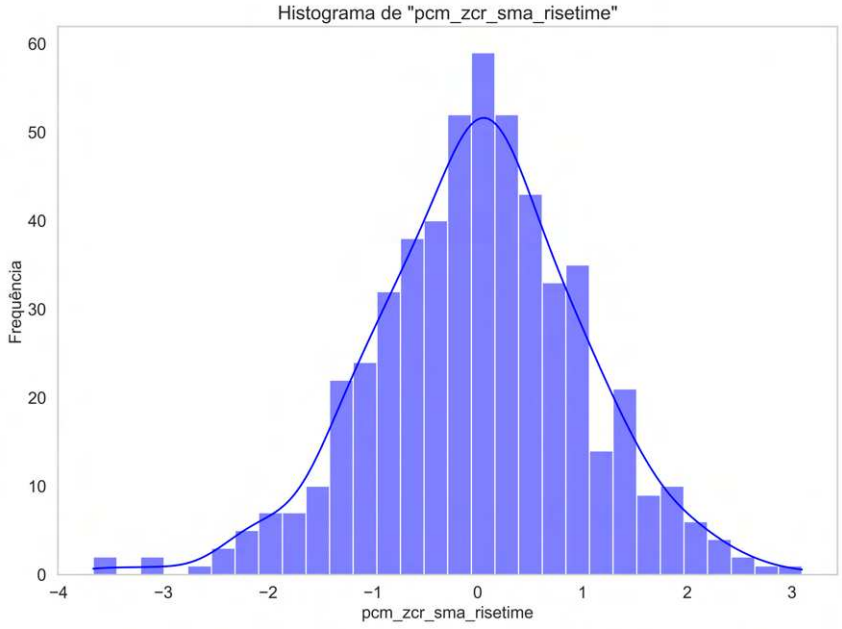
As figuras apresentadas a seguir ilustram exemplos de parâmetros de voz cuja distribuição estatística foi analisada utilizando o teste de Anderson-Darling. Os resultados evidenciam tanto casos em que a hipótese de normalidade é aceita, indicando conformidade com uma distribuição normal, quanto casos em que essa hipótese é rejeitada, sugerindo uma distribuição alternativa.

A Figura 21 apresenta a distribuição da taxa de cruzamento por zero do sinal temporal extraída da base de dados EMODB, identificada como uma distribuição gaussiana de acordo com o teste de Anderson–Darling. Em contrapartida, na Figura 22, observa-se a distribuição do desvio padrão da norma L1 do comprimento do espectro de áudio suavizado com a técnica RASTA, para o qual o teste indicou a ausência de uma distribuição normal. A norma L1, é a soma dos valores absolutos das componentes.

No caso da base de dados SAVEE, a Figura 23 exibe a distribuição do coeficiente LPC de ordem 2 da energia espectral de comprimento auditivo suavizada, classificada como gaussiana pelo mesmo teste. Já o parâmetro representando o 1º percentil da norma L1 do comprimento espectral auditivo, mostrado na Figura 24, não apresentou distribuição normal segundo os resultados do teste de Anderson–Darling.

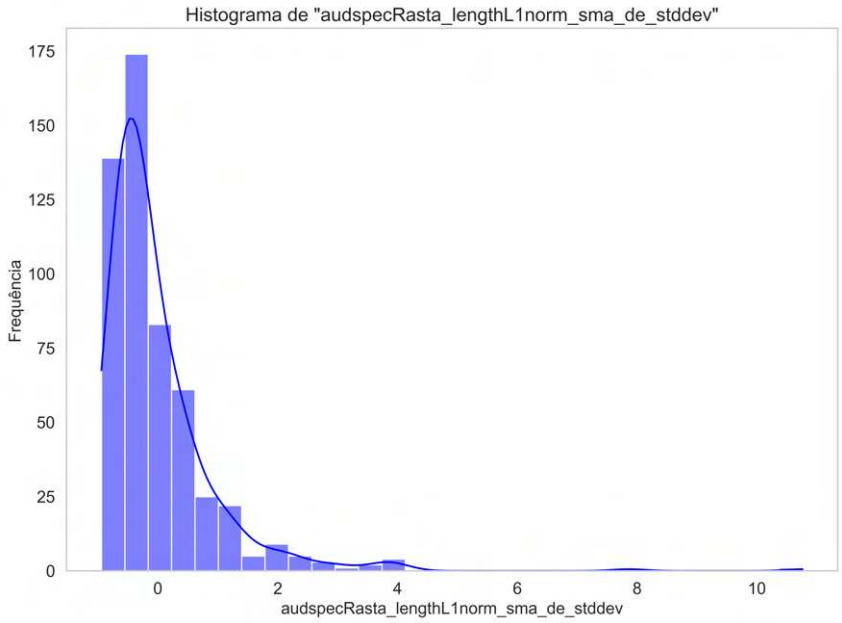
Por fim, na base de dados RAVDESS, a Figura 25 mostra a distribuição do parâmetro norma L1 do comprimento espectral auditivo suavizado com média móvel e derivada de primeira ordem, coeficiente LPC0, considerado com distribuição gaussiana pelo teste. Em contraste, a Figura 26 exibe a distribuição do parâmetro tempo de subida ao nível de 50% da norma L1 do comprimento espectral auditivo suavizado com média móvel e derivada de primeira ordem, para o qual foi identificada a ausência de uma distribuição normal.

Figura 21 – Distribuição do tempo de subida da taxa de cruzamento por zero (ZCR) de um sinal PCM suavizado por média móvel, considerado com distribuição normal de acordo com o teste de Anderson-Darling na base de dados EMODB.



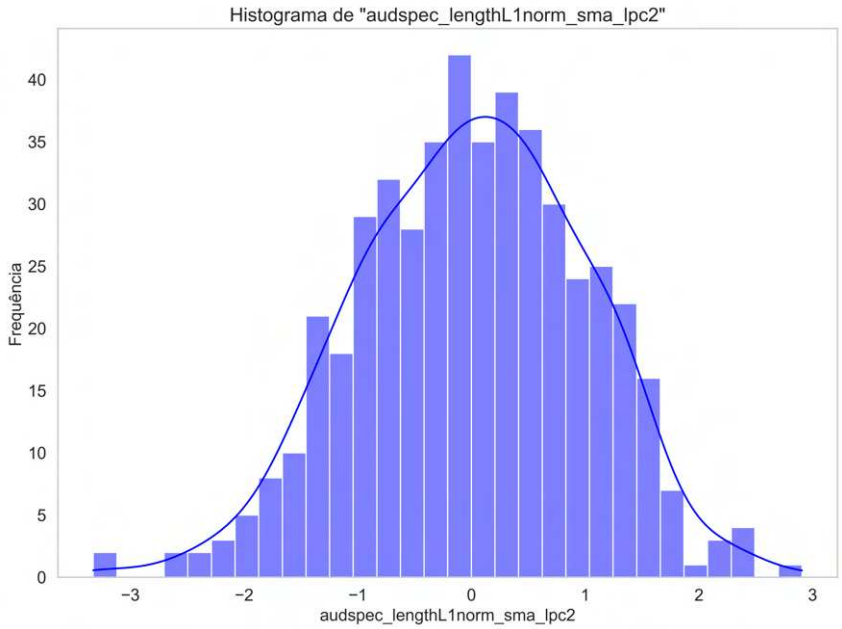
Fonte: Elaborado pelo autor (2024).

Figura 22 – Distribuição do desvio padrão da variação temporal (derivada delta) do comprimento da norma L1 do espectro auditivo processado com RASTA, considerado com distribuição não-normal de acordo com o teste de Anderson-Darling na base de dados EMODB.



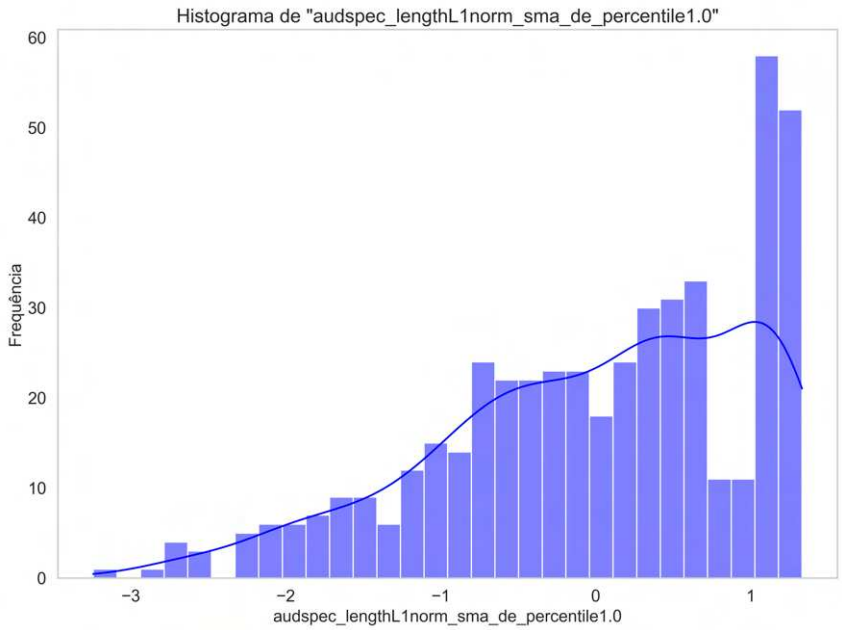
Fonte: Elaborado pelo autor (2024).

Figura 23 – Distribuição do segundo coeficiente de predição linear (LPC) calculado com base no comprimento da norma L1 do espectro auditivo suavizado por média móvel, considerado com distribuição normal de acordo com o teste de Anderson-Darling na base de dados SAVEE.



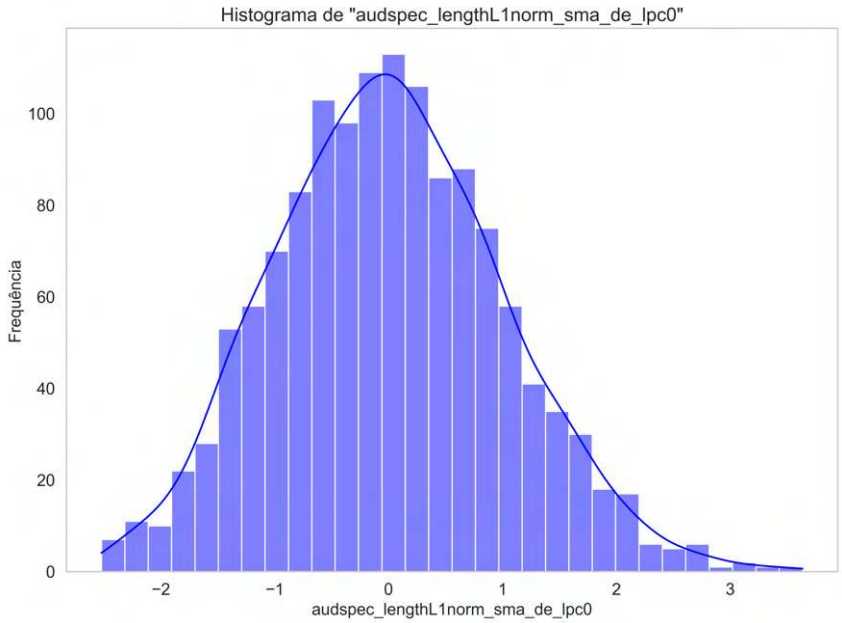
Fonte: Elaborado pelo autor (2024).

Figura 24 – Distribuição do Percentil de 1,0% calculado a partir da derivada temporal (delta) do comprimento da norma L1 do espectro auditivo suavizado por média móvel, um parâmetro considerado com distribuição não-normal de acordo com o teste de Anderson-Darling na base de dados SAVEE.



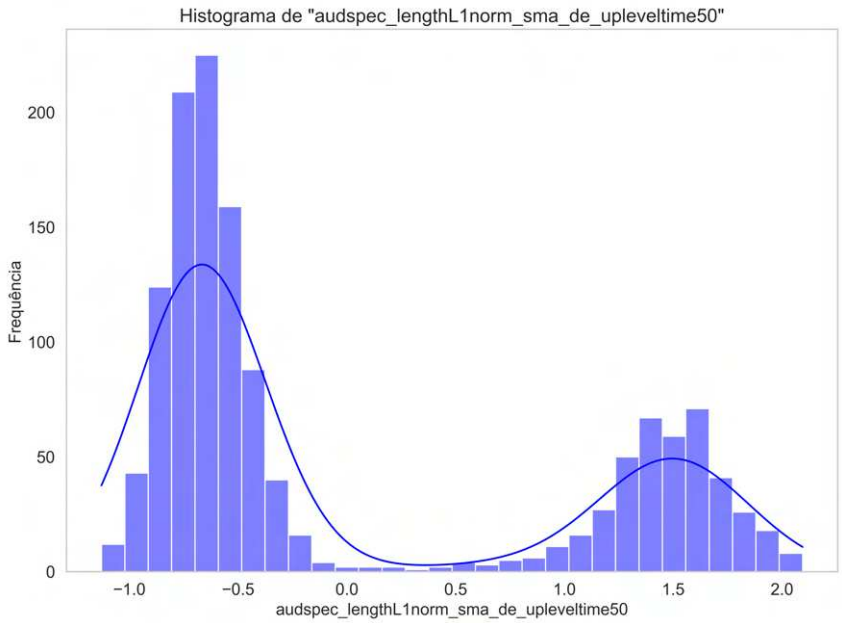
Fonte: Elaborado pelo autor (2024).

Figura 25 – Distribuição da Derivada temporal (delta) do coeficiente de predição linear de ordem 0 (LPC0) calculado a partir do comprimento da norma L1 do espectro auditivo suavizado por média móvel, um parâmetro considerado com distribuição normal de acordo com o teste de Anderson-Darling na base de dados RAVDESS.



Fonte: Elaborado pelo autor (2024).

Figura 26 – Distribuição da derivada temporal do tempo necessário para o comprimento da norma L1 do espectro auditivo suavizado por média móvel atingir 50% do nível máximo, um parâmetro considerado com distribuição não-normal de acordo com o teste de Anderson-Darling na base de dados RAVDESS.



Fonte: Elaborado pelo autor (2024).

4.1.2 Teste de Kruskal-Wallis

As Figuras 27, 28, 29, 30, 31 e 32 apresentam os resultados dos gráficos de violino para as bases de dados EMODB, SAVEE e RAVDESS, nos quais foi utilizado o teste de Kruskal-Wallis para avaliar a diferença nas distribuições dos parâmetros acústicos entre as emoções. Para cada base de dados, são apresentados dois exemplos de parâmetros: um em que o teste estatístico rejeita a hipótese nula, indicando que existe uma diferença significativa na distribuição dos parâmetros, e outro em que a hipótese nula não é rejeitada, sugerindo que não há evidências suficientes para afirmar que há uma diferença na distribuição dos parâmetros.

Na Figura 27, referente à base EMODB, observou-se que os dados rejeitaram a hipótese nula, indicando que há pelo menos um grupo que difere dos demais. Isso sugere que as emoções representadas pelos parâmetros analisados possuem distribuições estatisticamente distintas. Essa diferença pode ser observada pela variação na forma e amplitude dos violinos, refletindo a presença de dados que não se sobrepõem completamente, confirmando a validade do teste de Kruskal-Wallis para esta base.

A Figura 28, também para a base EMODB, onde a hipótese nula não foi rejeitada, indicando que nenhum dos grupos possui diferenças significativas. As distribuições podem ser notadas pelos violinos que não têm uma grande variabilidade entre amplitude, média e mediana, corroborando os achados da análise estatística.

Para a base SAVEE, as Figuras 29 e 30 seguem a mesma linha de raciocínio. Na Figura 29, o teste de Kruskal-Wallis revelou diferenças estatísticas entre os grupos, rejeitando a hipótese nula. A separação observada entre os violinos reflete que as emoções na base SAVEE têm características acústicas distintas, o que pode ser explorado para uma melhor discriminação emocional. A Figura 30 não mostra uma diferença expressiva visualmente, mostrando que para esse parâmetro os grupos de emoções não têm diferença estatística.

Finalmente, as Figuras 31 e 32 apresentam os resultados para a base RAVDESS. Para a Figura 31, o teste de Kruskal-Wallis rejeita a hipótese nula, indicando que as distribuições das emoções analisadas são estatisticamente diferentes. A diferença entre os violinos sugere uma variação significativa entre os grupos, o que é de grande importância para a análise de emoções em sinais de voz. A Figura 32 apresenta uma distribuição parecida entre classes, não é possível observar uma diferenciação nas distribuições, similar aos resultados obtidos nas figuras anteriores de outras bases de dados.

Os resultados do teste de Kruskal-Wallis, apresentados na Tabela 11, mostram a estatística H e os p-valores para cada base de dados utilizada. Para a base EMODB, a estatística H foi de $34,03$ com um p-valor inferior a $0,01$, indicando uma diferença estatisticamente significativa para o parâmetro "Norma L1 do comprimento do espectro de áudio suavizado com o coeficiente LPC de ordem zero". Já para o parâmetro "Posição mínima da variação suavizada da taxa de cruzamento por zero do sinal PCM", obteve-se uma estatística H de $5,13$ e um p-valor de $0,52$, indicando que não há diferença significativa. Na base SAVEE, os parâmetros analisados

mostraram resultados diversos, com a estatística H de $105,7$ e $p < 0,01$ para o desvio padrão do oitavo coeficiente cepstral Mel suavizado, e uma estatística H de $3,09$ com $p = 0,79$ para o parâmetro "LPC ordem 2 da variação temporal suavizada do shimmer local". Na base RAVDESS, o parâmetro "LPC ordem 3 do comprimento do espectro auditivo suavizado pela norma L1" apresentou uma estatística H de $69,95$ com $p < 0,01$, enquanto o parâmetro "Mínimo local do shimmer suavizado" teve uma estatística H de $8,06$ e $p = 0,32$, sem evidência de diferenças significativas.

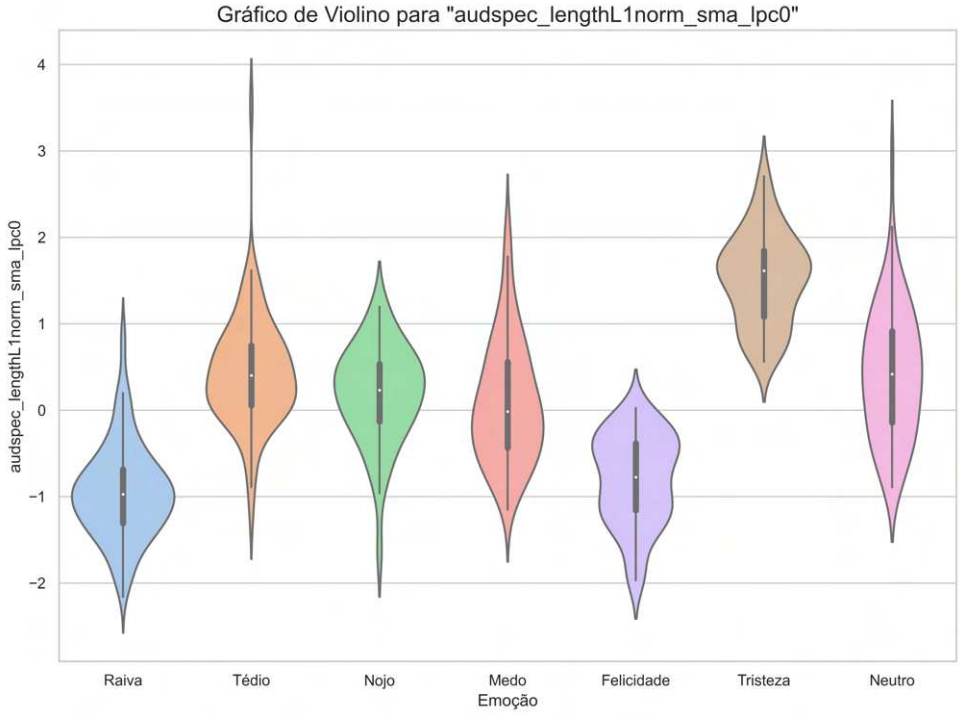
Tabela 11 – Resultados do teste de Kruskal-Wallis para os parâmetros mostrados nos gráficos de violino, valor de significância $\alpha = 0,001$.

Resultados do Teste de Kruskal-Wallis				
Base	**Estatística H - Equação (27)**	**p-valor**	**Nome do parâmetro**	**Figura**
EMODB	34,03	<0,001	Norma L1 do comprimento do espectro de áudio suavizado com o coeficiente LPC de ordem zero	Figura 27
	5,13	0,52	Posição mínima da variação suavizada da taxa de cruzamento por zero do sinal PCM	Figura 28
SAVEE	105,7	<0,001	Desvio padrão do oitavo coeficiente Mel-cepstral suavizado	Figura 29
	3,09	0,79	LPC ordem 2 da variação temporal suavizada do shimmer local	Figura 30
RAVDESS	69,95	<0,001	LPC ordem 3 do comprimento do espectro auditivo suavizado pela norma L1	Figura 31
	8,06	0,32	Mínimo local do shimmer suavizado	Figura 32

Fonte: Elaborado pelo autor (2024).

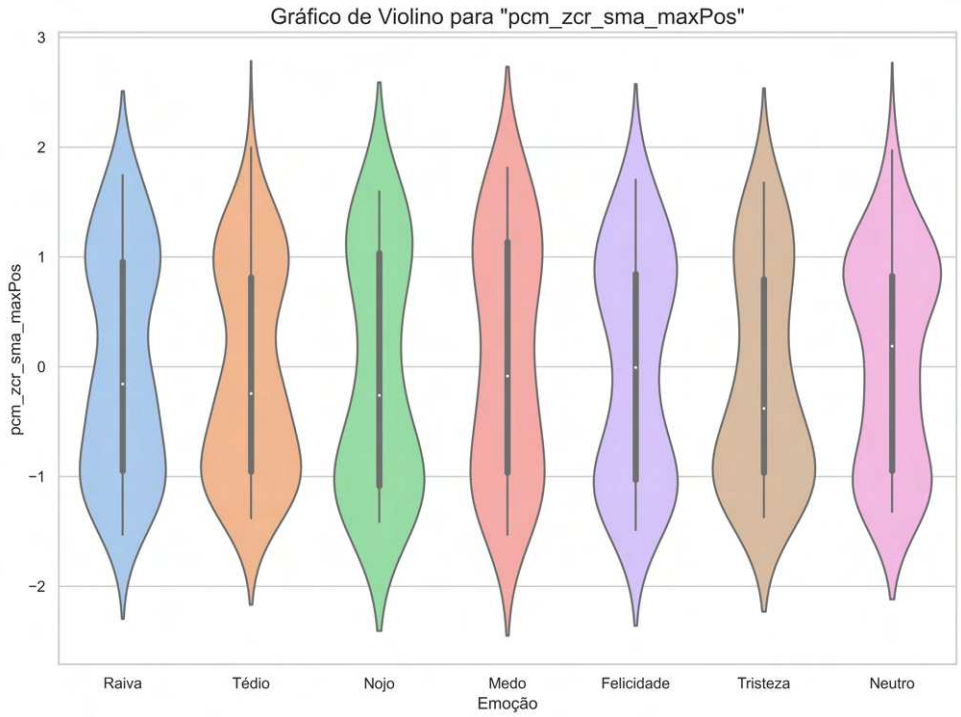
Esses resultados indicam que, para todas as bases analisadas, o teste de Kruskal-Wallis foi eficaz na identificação de diferenças significativas nas distribuições dos parâmetros acústicos das emoções, o que é fundamental para o aprimoramento de modelos de reconhecimento de emoções na fala.

Figura 27 – Distribuição do parâmetro considerado significativamente diferente entre as classes, conforme indicado pelo teste de Kruskal-Wallis na base de dados EMODB.



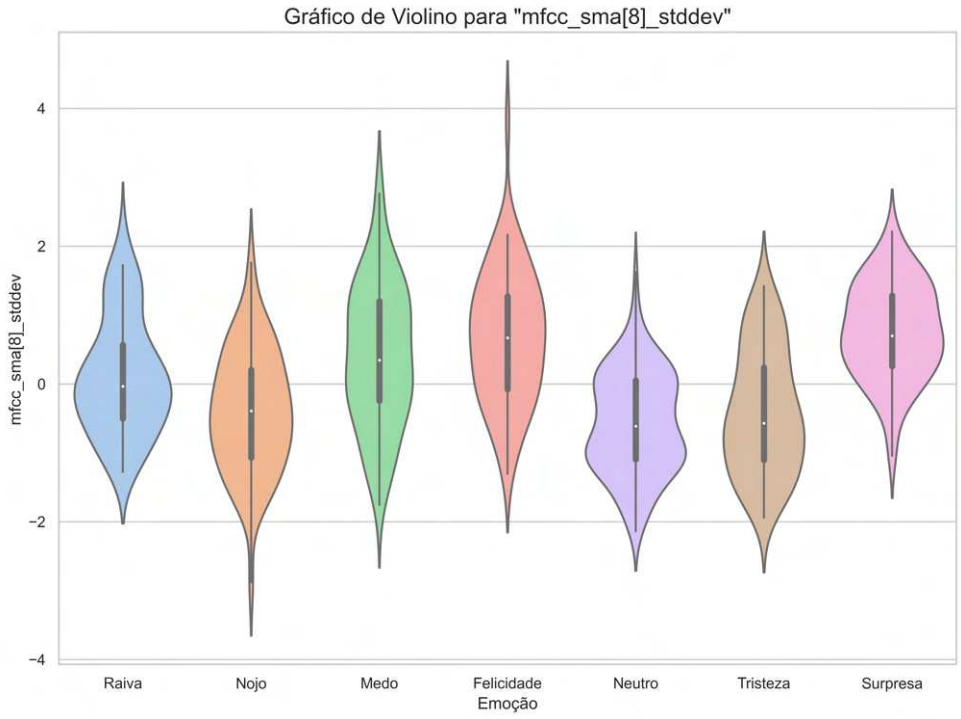
Fonte: Elaborado pelo autor (2024).

Figura 28 – Distribuição do parâmetro considerado não significativamente diferente entre as classes, conforme indicado pelo teste de Kruskal-Wallis na base de dados EMODB.



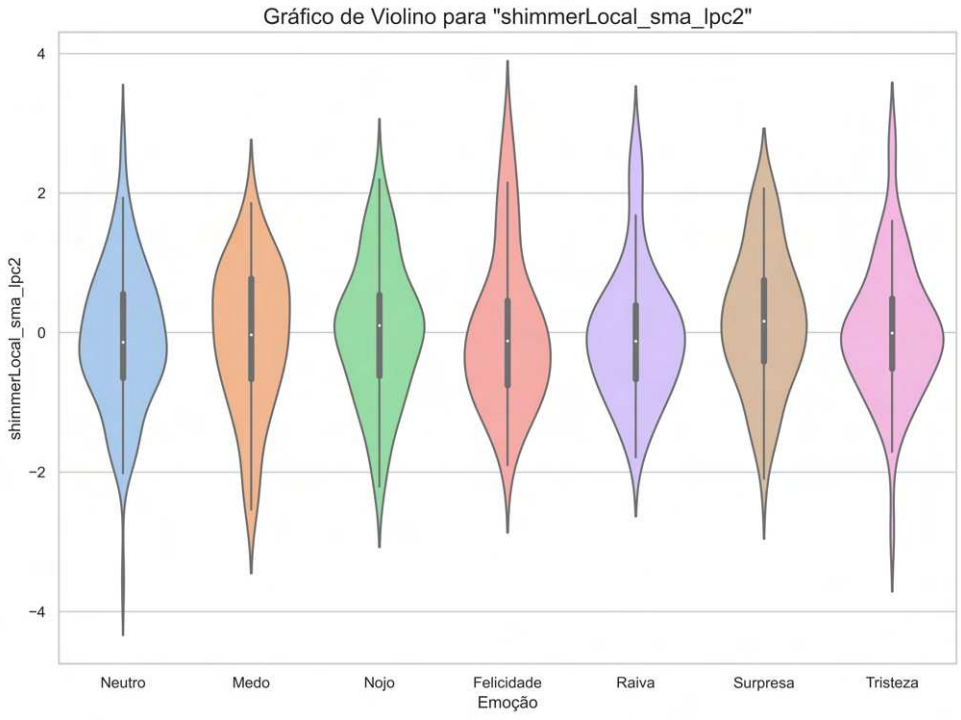
Fonte: Elaborado pelo autor (2024).

Figura 29 – Distribuição do parâmetro considerado significativamente diferente entre as classes, conforme indicado pelo teste de Kruskal-Wallis na base de dados SAVEE.



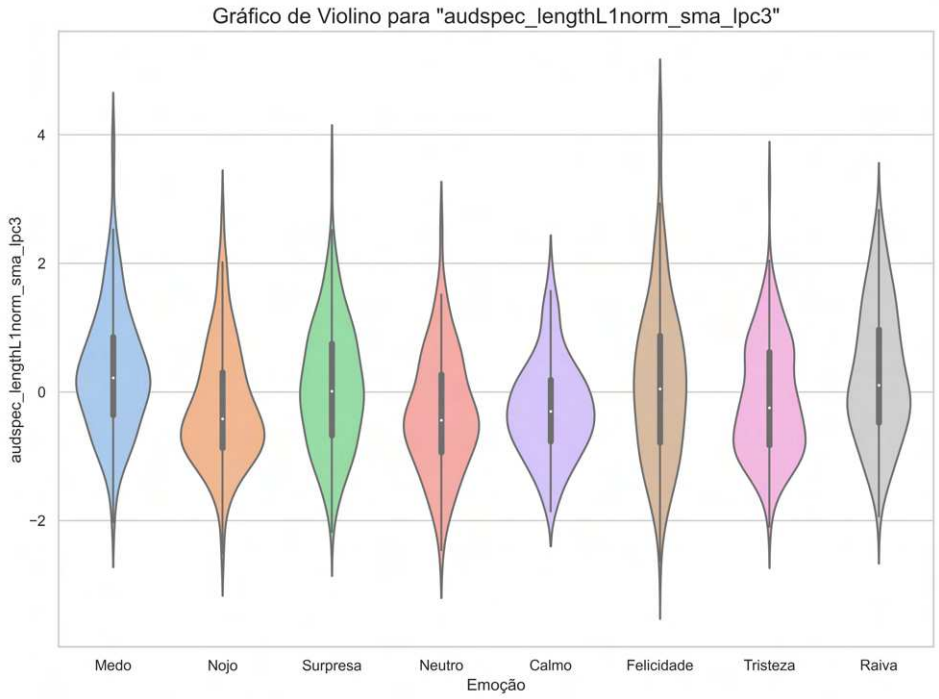
Fonte: Elaborado pelo autor (2024).

Figura 30 – Distribuição do parâmetro considerado não significativamente diferente entre as classes, conforme indicado pelo teste de Kruskal-Wallis na base de dados SAVEE.



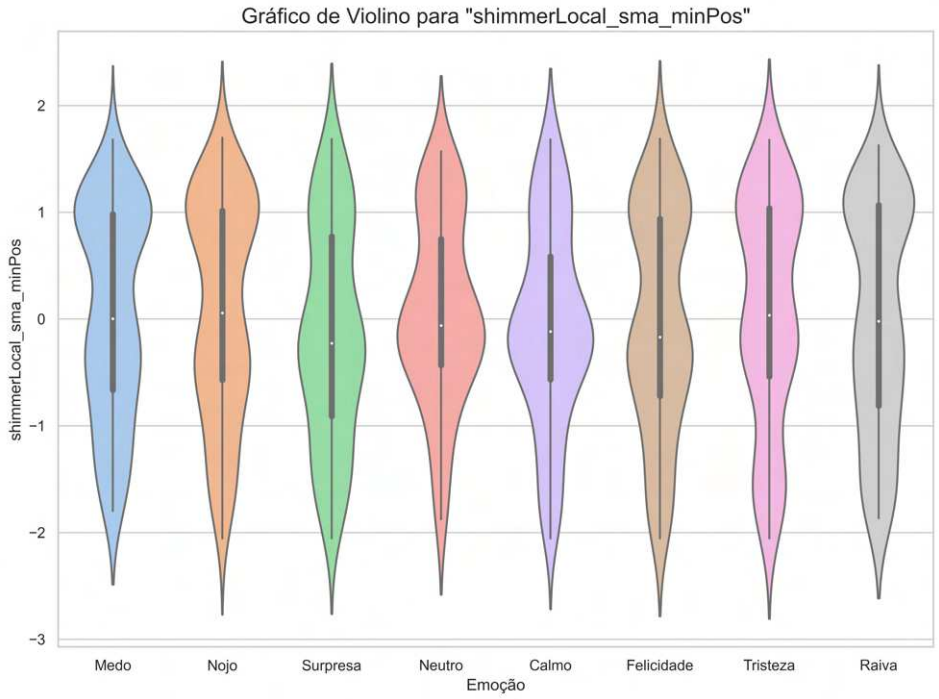
Fonte: Elaborado pelo autor (2024).

Figura 31 – Distribuição do parâmetro considerado significativamente diferente entre as classes, conforme indicado pelo teste de Kruskal-Wallis na base de dados RAVDESS.



Fonte: Elaborado pelo autor (2024).

Figura 32 – Distribuição do parâmetro considerado não significativamente diferente entre as classes, conforme indicado pelo teste de Kruskal-Wallis na base de dados RAVDESS.



Fonte: Elaborado pelo autor (2024).

4.2 SUBCONJUNTO GAUSSIANO EM PCA E ICA

Para comparar os resultados, a acurácia foi testada usando 100 componentes principais e 100 componentes independentes das seguintes duas maneiras: (1) aplicando PCA e ICA a todos os parâmetros, ou seja, 4382, 4845 e 5407 parâmetros para os bancos de dados SAVEE, RAVDESS e Berlin EmoDB, respectivamente, e (2) aplicando ICA a todos os parâmetros e aplicando PCA apenas os parâmetros com distribuição normal, resultando assim em totais de 415, 133 e 397 para os bancos de dados SAVEE, RAVDESS e Berlin EmoDB, respectivamente. O número de parâmetros com distribuições normais e não normais, junto com os parâmetros selecionados pela variância, está apresentado na Tabela 12.

Tabela 12 – Dimensões da matriz para cada banco de dados.

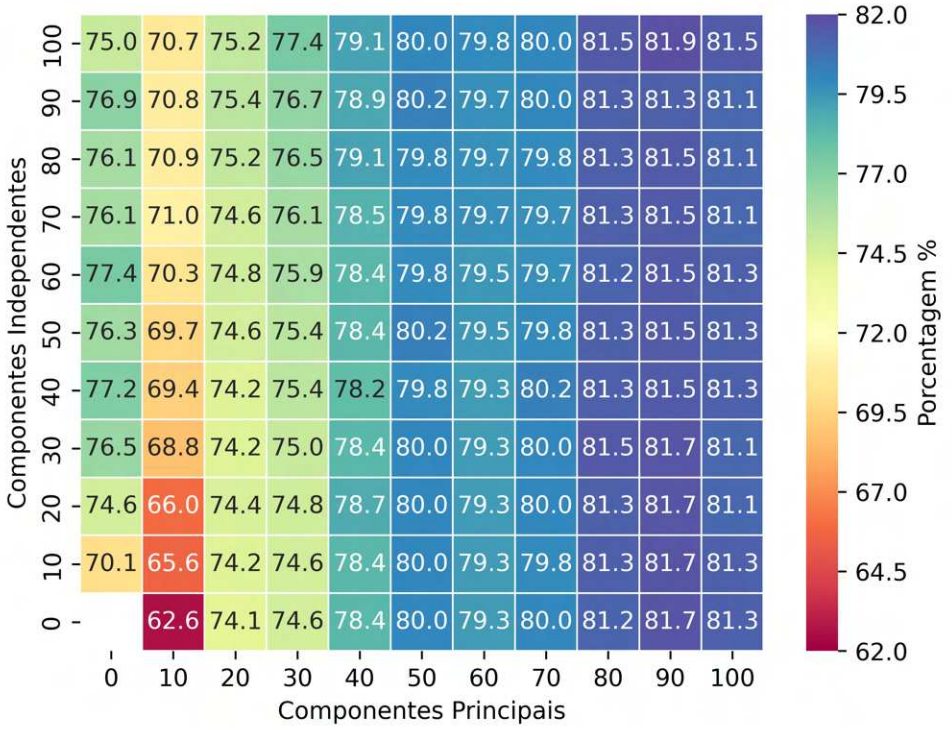
	SAVEE	**RAVDESS**	**Berlin EmoDB**
$\mathbf{X_S}$	480×5881	1440×5768	535×5975
$\mathbf{X_{KW}}$	480×4382	1440×4845	535×5407
$\mathbf{X_n}$	480×415	1440×133	535×397
$\mathbf{X_{nn}}$	480×3967	1440×4712	535×5010

Nas Figuras 33, 34, e 35, a primeira coluna da esquerda para a direita representa as primeiros 100 componentes independentes. Esse procedimento foi realizado assumindo que os parâmetros de entrada poderiam ser decompostos em 10 componentes, 20 componentes, 30 componentes, ..., até 100 componentes. A primeira linha superior dos gráficos nas Figuras 33, 34, e 35 exibe as componentes principais ordenadas pela maior variância. Os outros pontos são concatenações de PCA e ICA.

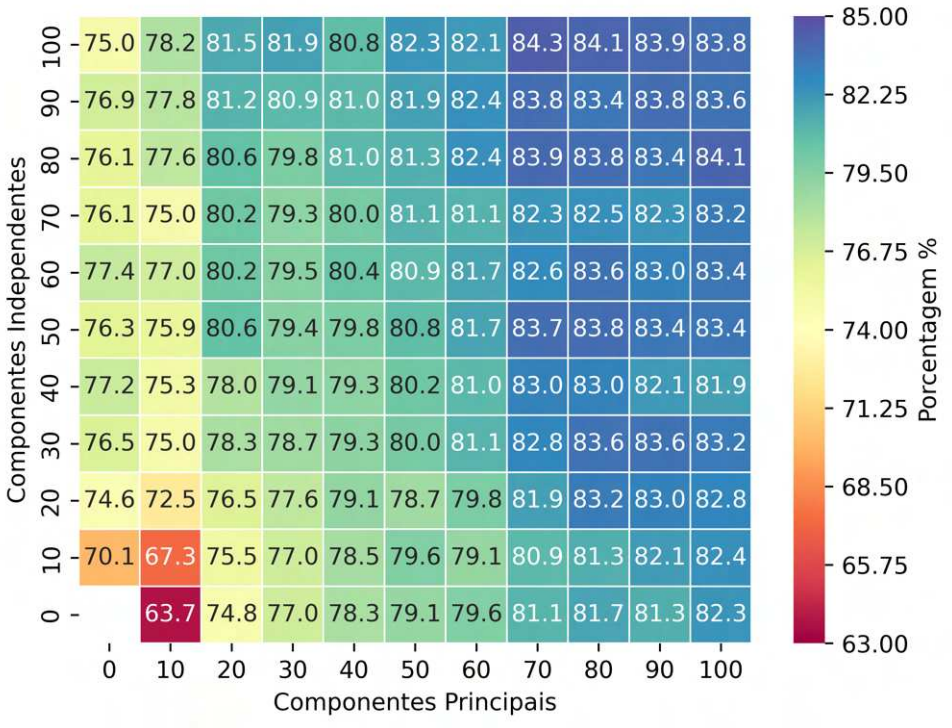
As Figuras 33a, 34a, e 35a representam a acurácia quando os dados não foram segmentados pela distribuição. Nesses casos, as técnicas de análise de componentes principais (PCA) e análise de componentes independentes (ICA) foram aplicadas a todos os dados selecionados via o teste de Kruskal–Wallis, e então os resultados foram combinados conforme o diagrama. Nas Figuras 33b, 34b, e 35b, as componentes principais foram calculados exclusivamente a partir dos parâmetros acústicos que exibiram uma distribuição normal de acordo com o teste de Anderson–Darling.

Nas Figuras 36, 37 e 38, as métricas para cada banco de dados são apresentadas com a distribuição, mediana e valores médios. As métricas utilizadas nessas figuras são descritas pelas Equações (35), (36), (37) e (38), representando acurácia, precisão, recall e F-score. Fundimos PCA e ICA usando o maior valor para as técnicas propostas (ou seja, utilizando o PCA calculado apenas com dados que apresentavam distribuição normal e o ICA calculado em todos os parâmetros). Os valores dos grupos foram separados de acordo com a legenda por distribuição normal e deixados não separados, com o teste de Kruskal–Wallis também aplicado para redução de parâmetros.

Figura 33 – Acurácia da fusão de PCA e ICA para o banco de dados EMODB: (**a**) PCA aplicado em todos os parâmetros e (**b**) PCA aplicado nos parâmetros com distribuição Gaussiana.



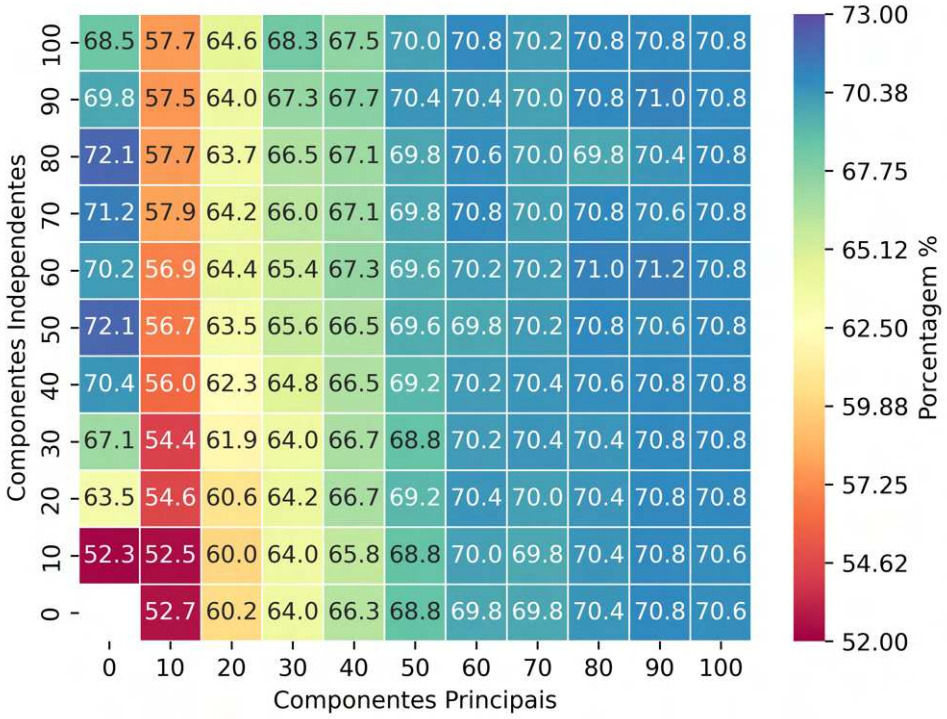
(a)



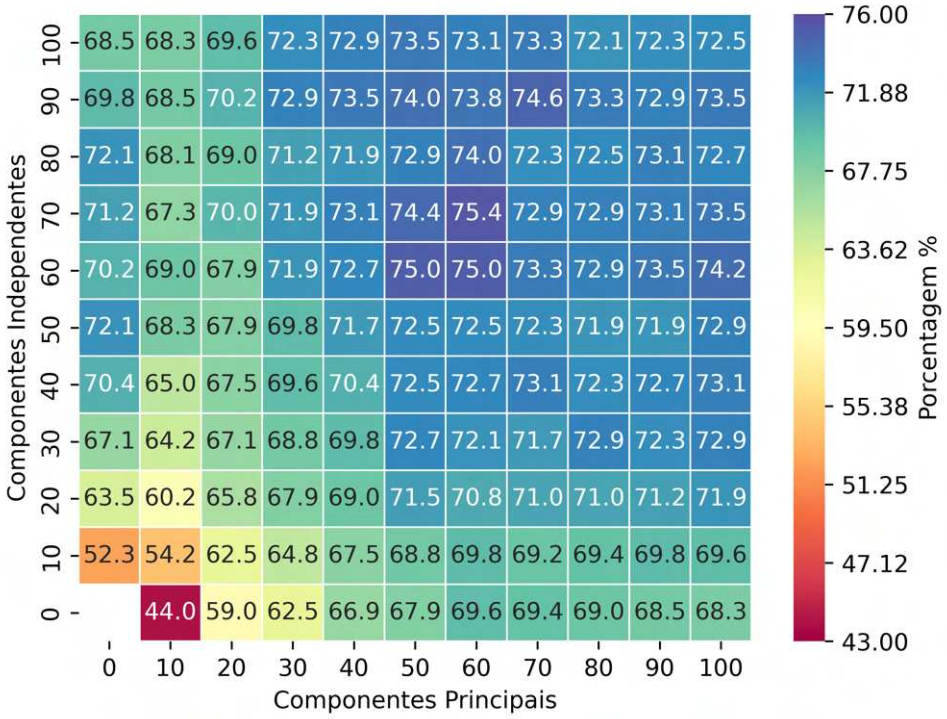
(b)

Fonte: Elaborado pelo autor (2024).

Figura 34 – Acurácia da fusão de PCA e ICA para o banco de dados SAVEE: (**a**) PCA aplicado em todos os parâmetros e (**b**) PCA aplicado nos parâmetros com distribuição Gaussiana.



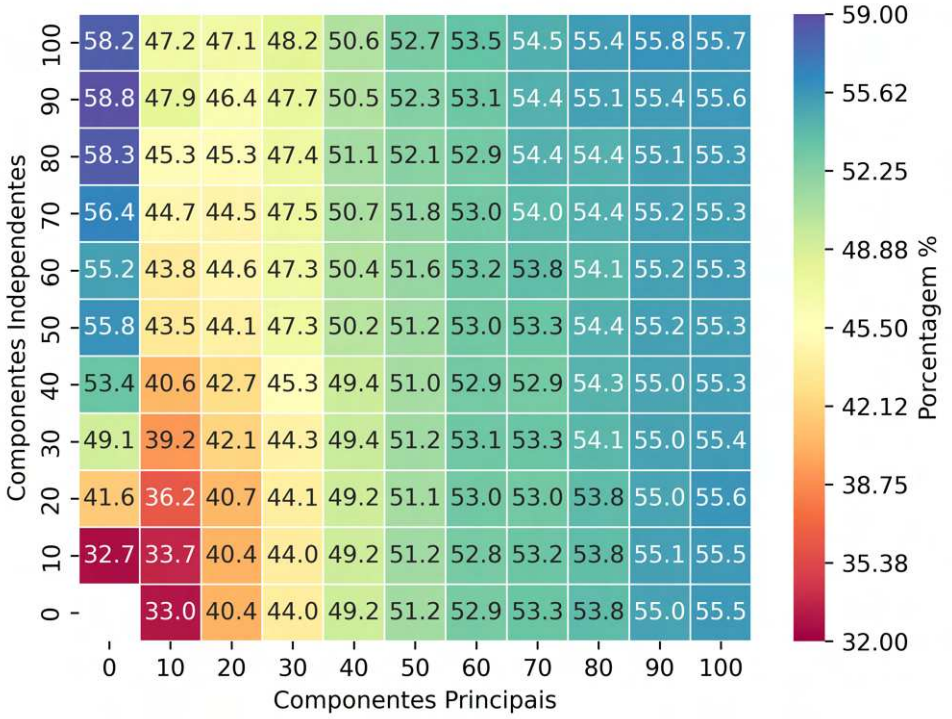
(a)



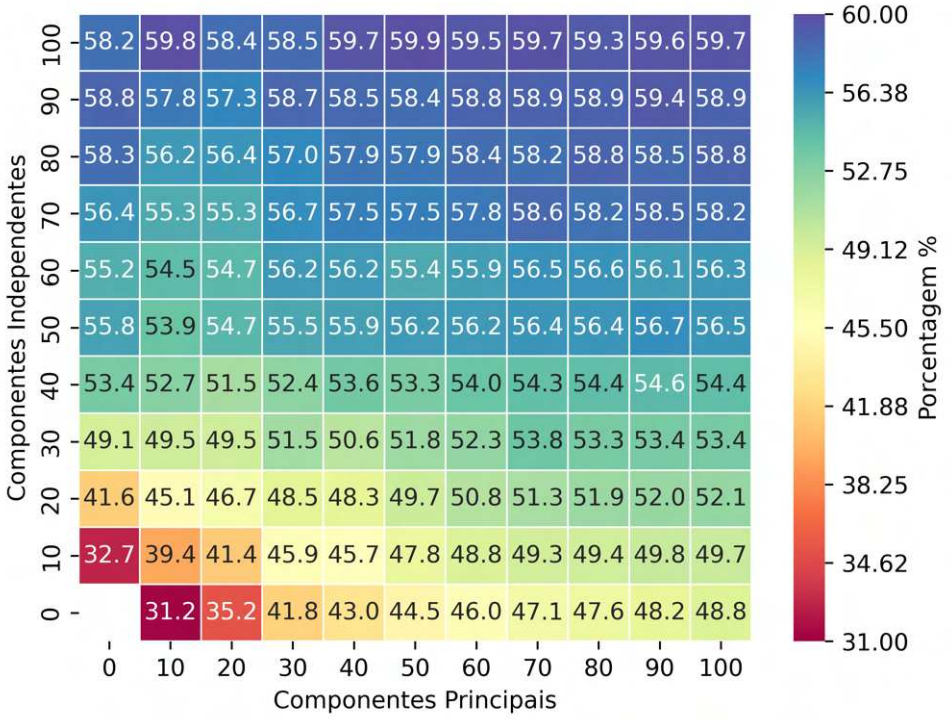
(b)

Fonte: Elaborado pelo autor (2024).

Figura 35 – Acurácia da fusão de PCA e ICA para o banco de dados RAVDESS: (**a**) PCA aplicado em todos os parâmetros e (**b**) PCA aplicado nos os parâmetros com distribuição Gaussiana.



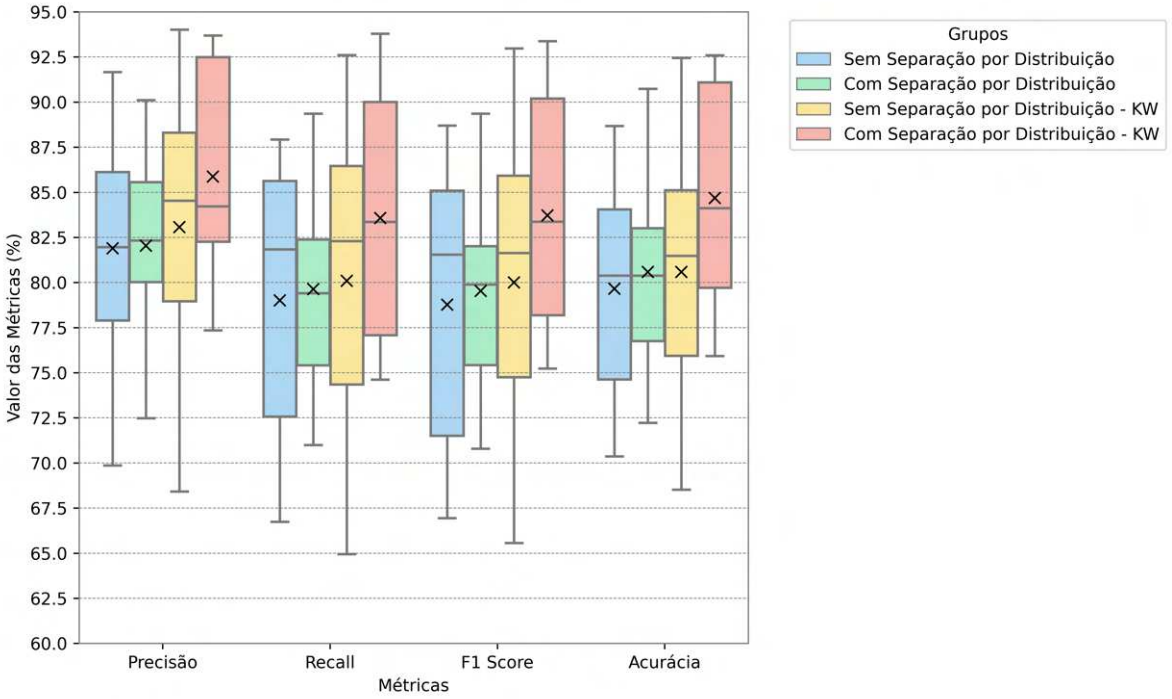
(a)



(b)

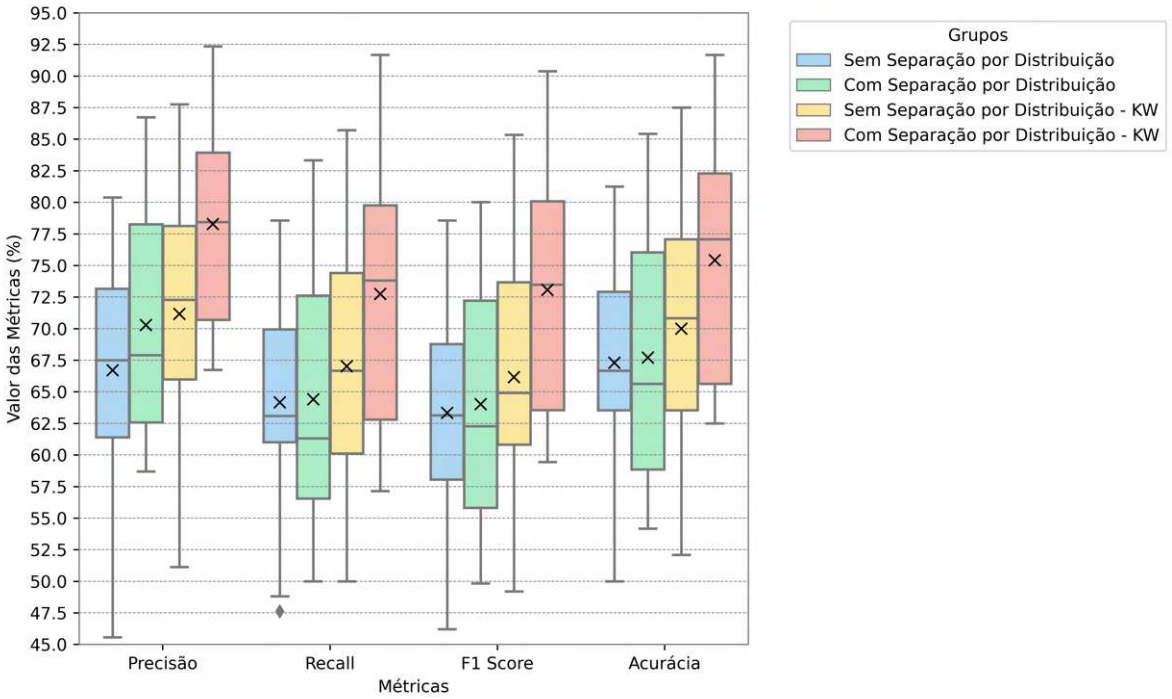
Fonte: Elaborado pelo autor (2024).

Figura 36 – Métricas para a base de dados EmoDB utilizando 100 componentes independentes e 70 componentes principais.



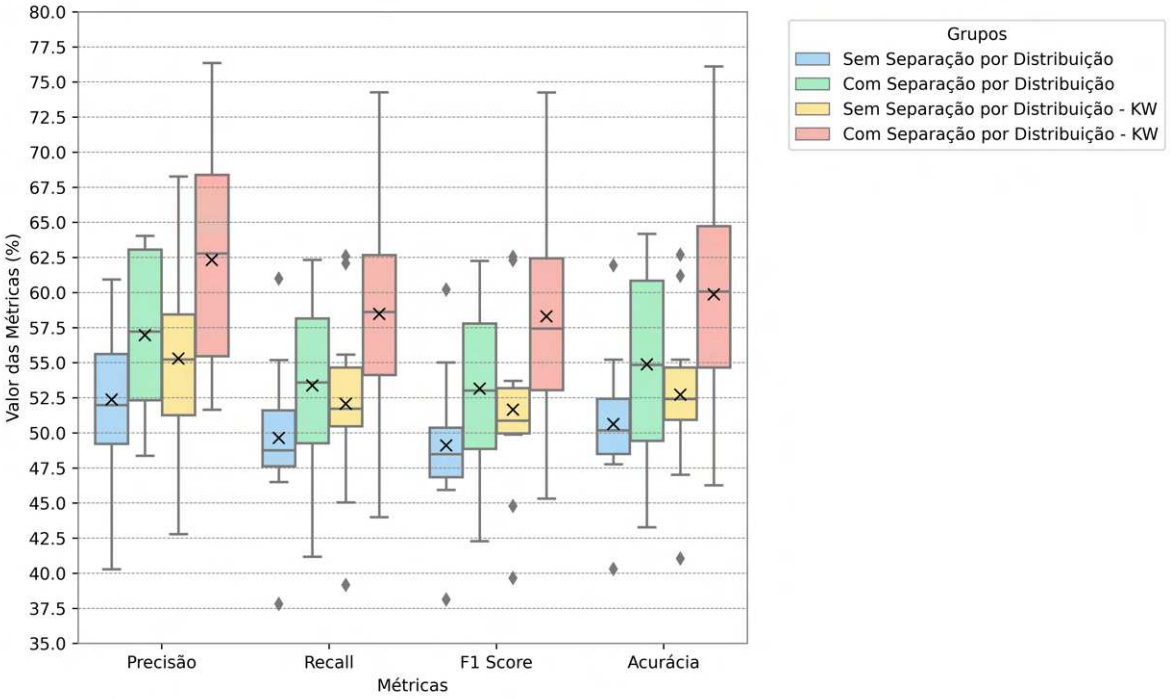
Fonte: Elaborado pelo autor (2024).

Figura 37 – Métricas para a base de dados SAVEE utilizando 60 componentes independentes e 50 componentes principais.



Fonte: Elaborado pelo autor (2024).

Figura 38 – Métricas para a base de dados RAVDESS utilizando 100 componentes independentes e 50 componentes principais.



Fonte: Elaborado pelo autor (2024).

4.2.1 Matrizes de Confusão para Dados Categóricos

As matrizes de confusão geradas para cada uma das três bases de dados (EMODB, SAVEE e RAVDESS), analisando os efeitos do uso de PCA para dados gaussianos em conjunto com ICA para todos os parâmetros. Realizada com duas configurações: uma matriz de confusão sem separar os dados pela distribuição gaussiana e outra com a separação de dados gaussianos antes da aplicação de PCA.

A matriz de confusão é uma ferramenta para avaliar o desempenho de classificadores em problemas de aprendizado supervisionado. Ela organiza as previsões do modelo em uma tabela que relaciona as classificações reais com as preditas, permitindo a visualização detalhada dos acertos e erros. As linhas da matriz representam as classes esperadas ou reais, enquanto as colunas correspondem às classes previstas pelo modelo.

Neste trabalho, as categorias de emoção representam as variáveis de classificação, com objetivo de identificar o quão bem o modelo conseguiu prever as emoções e onde ele falhou, na diagonal principal estão os acertos para cada emoção, quanto mais próximo de 100% maior é a exatidão ou acurácia de um modelo. Valores fora da diagonal principal indicam falsos positivos e falsos negativos.

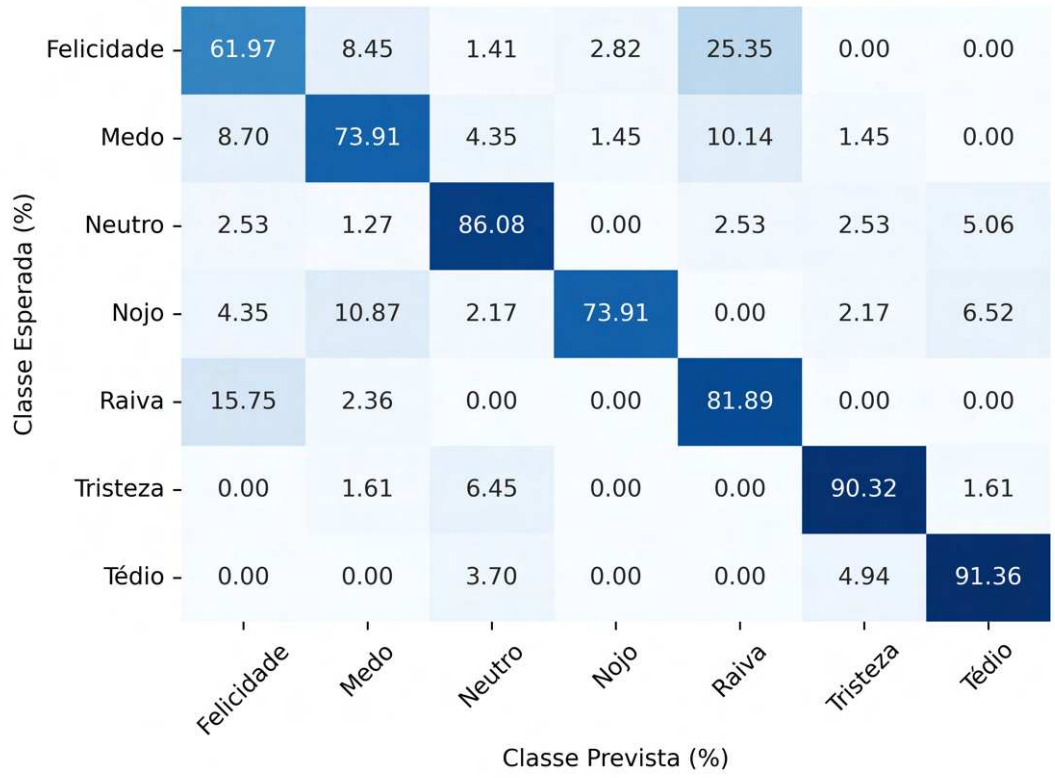
Um exemplo de falso negativo pode ser visto na Figura 39a, onde 25,35% da categoria “felicidade” está classificada como “raiva”, um exemplo de falso positivo seria da categoria “medo” classificada em 10,87% como “nojo”.

Para a base de dados EMODB, ver Figuras 39a e 39b, observou-se que a acurácia da emoção “tristeza” aumentou com o uso de PCA para os dados com distribuição gaussiana, enquanto a acurácia para a emoção “tédio” diminuiu. Isso indica que a separação dos dados com distribuição gaussiana antes da aplicação de PCA e ICA pode ter impactado positivamente algumas emoções, mas de forma adversa para outras, como no caso do tédio. A “raiva” e a “felicidade” tiveram uma maior sobreposição em ambos os métodos. Todas as outras emoções para esta base tiveram um aumento no valor da acurácia.

Na base de dados SAVEE, ver Figuras 40a e 40b, as melhorias mais notáveis ocorreram nas emoções “neutro” e “felicidade”, que apresentaram aumento na acurácia ao usar PCA combinado com ICA nos dados gaussianos. No entanto, observou-se uma piora no reconhecimento da emoção “raiva”, evidenciando que, enquanto alguns estados emocionais se beneficiaram da abordagem proposta, outros não tiveram resultados tão favoráveis.

Por fim, na base de dados RAVDESS, ver Figuras 41a e 41b, a combinação de PCA com ICA e a separação de dados gaussianos mostrou uma melhora significativa para as emoções “raiva” e “calma”, enquanto a emoção “neutro” foi a única a apresentar uma diminuição na acurácia. Essa diferença novamente nos sugere que a aplicação de PCA nos dados gaussianos pode ter ajudado a melhorar o desempenho para algumas emoções específicas, mas pode não ser tão eficaz para outras, como no caso do estado neutro. Outro resultado relevante é a diminuição da sobreposição da “raiva” com a “felicidade” utilizando a filtragem de dados com distribuição normal em PCA.

Figura 39 – Matriz de confusão para dados categóricos: (**a**) EmoDB sem separação de dados por distribuição normal, (**b**) EmoDB com separação de dados por distribuição normal.



(a)

(b)

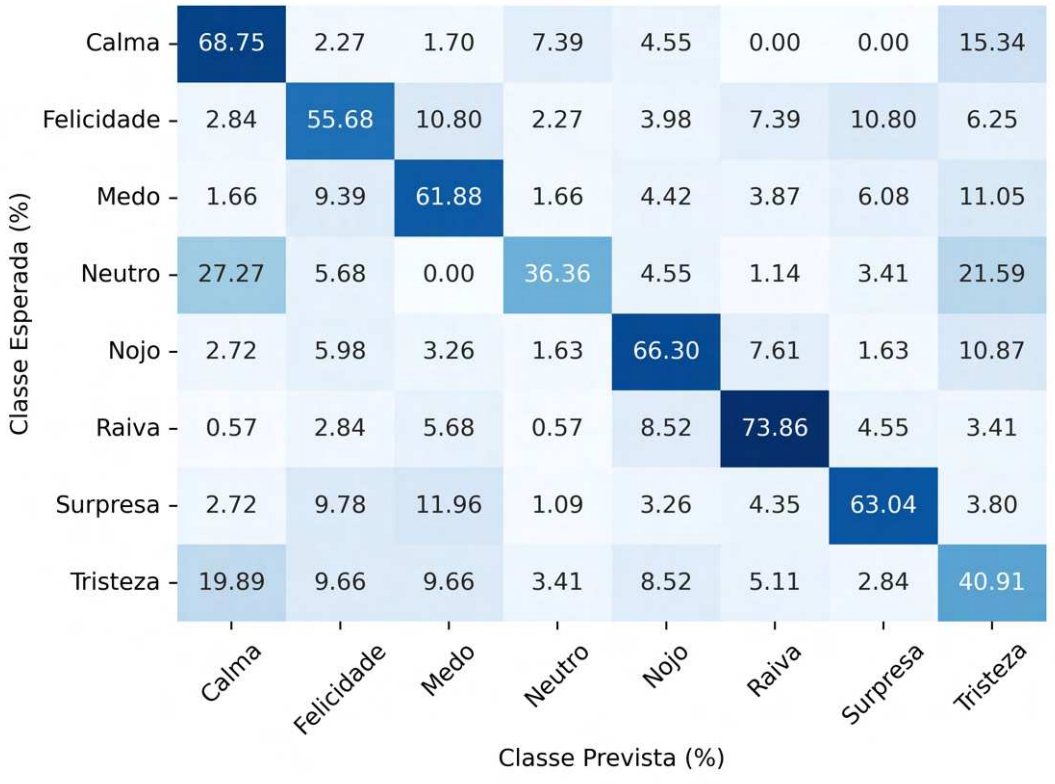
Figura 40 – Matriz de confusão para dados categóricos: (**a**) SAVEE sem separação de dados por distribuição normal, (**b**) SAVEE com separação de dados por distribuição normal.

(a)

(b)

Figura 41 – Matriz de confusão para dados categóricos: (**a**) RAVDESS sem separação de dados por distribuição normal, (**b**) RAVDESS com separação de dados por distribuição normal.

(a)



(b)

4.3 DISCUSSÃO

Esta pesquisa explorou técnicas para seleção e redução de parâmetros com o intuito de analisar o desempenho do método proposto, utilizando a separação dos dados em subconjuntos gaussianos aplicados ao reconhecimento de emoções em sinais de voz.

O número de parâmetros de entrada foi reduzido inicialmente pelo critério de variância e posteriormente pelo teste de Kruskal–Wallis que como mostra a tabela 10, nota-se que há uma redução dos parâmetros modesta, mas suficiente para remover parâmetros irrelevantes e com baixa informação. Essa etapa teve como objetivo otimizar o desempenho dos métodos subsequentes, preservando os parâmetros mais relevantes.

Os dados selecionados foram filtrados para compor um subconjunto de parâmetros com distribuição gaussiana, permitindo a aplicação direcionada do PCA a esses parâmetros, conforme descrito na metodologia. Como alternativa para comparação, também foi testado um segundo procedimento, no qual os dados foram mantidos sem filtragem por distribuição e submetidos tanto ao PCA quanto ao ICA. A Tabela 12 detalha as dimensões das matrizes resultantes, evidenciando a organização e redução do conjunto final de parâmetros antes de aplicar em PCA e ICA.

A filtragem dos parâmetros antes de aplicar PCA e ICA é importante para otimização e redução de custo computacional. Com um número menor de dados irrelevantes, o uso de PCA e ICA torna-se mais eficiente e focado, facilitando a extração de componentes principais significativos e melhorando a interpretação dos resultados. A Tabela 10 resume as dimensões das matrizes de dados resultantes para cada base de dados após a aplicação dessas filtragens, evidenciando a redução e a organização do conjunto final de parâmetros.

PCA foi utilizado para reduzir os parâmetros, capturando a principal variância para o subconjunto de distribuição normal, enquanto ICA foi utilizado para extrair os dados de ambos os conjuntos de dados normais e não normais, assumindo assim uma combinação de parâmetros conforme descrito na metodologia.

Os resultados obtidos com o método proposto, que combina PCA e ICA, sugerem que as componentes principais, isoladamente, podem ter uma influência relevante no desempenho do modelo, como indicado nas Figuras 33a, 34a e 35a.

Isso evidencia um resultado diferente do esperado, uma vez que não se previa que os dados com distribuição gaussiana, tratados isoladamente, conteriam a maior parte das informações necessárias para o reconhecimento de emoções. Inicialmente, supôs-se que a variabilidade capturada pelos dados não normais desempenharia um papel mais expressivo, especialmente considerando que emoções frequentemente manifestam características complexas e não lineares nos sinais de voz.

Outro ponto importante a destacar é o número de parâmetros com distribuição normal, em todas as bases ele é muito menor do que os não normais, intuitivamente poderia supor que mais parâmetros teriam um melhor desempenho. No entanto, a predominância de informações nos parâmetros gaussianos pode ser explicada pela maior estabilidade e regularidade estatística

desses dados, que parecem destacar padrões essenciais para a separação de classes emocionais. Por outro lado, os parâmetros não gaussianos podem conter informações complementares, mas aparentemente são mais suscetíveis a ruídos ou variabilidades que dificultam a extração de características significativas para os classificadores.

Na base de dados RAVDESS, que teve um número muito menor de parâmetros com distribuição gaussiana, estes tiveram uma contribuição muito pequena para o modelo, conforme pode ser visto nas Figuras 35b.

Esse comportamento indica que o uso de PCA, quando direcionado exclusivamente aos parâmetros gaussianos, facilita a captura de componentes principais mais relevantes, maximizando a eficiência do modelo. Entretanto, a análise também destaca a importância de complementar essas informações com ICA aplicado aos dados não gaussianos, visto que as combinações de PCA e ICA continuam a apresentar o melhor desempenho geral. Esse resultado sugere que, embora os dados gaussianos representem a maior parte da informação útil, os dados não gaussianos trazem detalhes que refinam ainda mais a acurácia do modelo.

Os resultados demonstraram que a fusão de componentes principais (PCA) e independentes (ICA) proporcionou uma melhoria geral na acurácia para os três conjuntos de dados analisados, embora com variações de impacto entre eles.

A aplicação de PCA e ICA de forma indiscriminada a todos os parâmetros, sem separação prévia por distribuição, indica que as componentes principais têm maior influência sobre o modelo, como ilustrado nas Figuras 33a, 34a e 35a. A concatenação dos dois grupos de parâmetros ($\mathbf{S_{ICA}}$, $\mathbf{Z_{PCA}}$) sem discriminação por distribuição resultou em pequenas variações na acurácia. No entanto, o conjunto de dados RAVDESS apresentou uma ligeira influência maior das componentes independentes, possivelmente devido ao fato de possuir o menor número de parâmetros com distribuição normal, como observado nas Tabelas 10 e 12.

O modelo proposto alcançou a maior acurácia com o conjunto de dados EmoDB, seguido por SAVEE e RAVDESS. Com EmoDB, a acurácia melhorou de 80,0% para 84,3% ao usar 100 componentes independentes e 70 componentes principais, que representaram a maior acurácia, como mostrado na Figura 33. Ao comparar os valores de acurácia entre as componentes principais e independentes para EmoDB (a primeira linha e coluna nas Figuras 33a e 33b, respectivamente), o uso de apenas 397 parâmetros com distribuição normal (veja a Tabela 10) e 100 componentes principais resultou em uma acurácia melhorada. Isso sugere que componentes com distribuição normal contribuem significativamente para o desempenho do modelo, e aplicar PCA exclusivamente a parâmetros com distribuição normal aumenta a acurácia; no entanto, o melhor desempenho foi alcançado com PCA+ICA.

Ainda na base de dados EMODB, os resultados das métricas visto no gráfico da Figura 36, evidencia uma melhora na precisão, no *Recall* e no *F1-score*, a distribuição destas métricas tem uma dispersão menor para o uso de PCA+ICA utilizando o método proposto, além da média e da mediana terem ficado acima em todos os grupos testados.

Para SAVEE, a redução mais significativa no número de componentes e a melhoria na

acurácia foram observadas, com a acurácia aumentando de 70,0% para 75,4% usando uma combinação de 60 componentes independentes e 50 componentes principais (com distribuição normal). SAVEE teve o maior número de parâmetros com distribuição normal, como mostrado na Tabela 10. Notavelmente, o método proposto levou não apenas a uma maior média e mediana de acurácia, mas também a uma melhoria no quartil superior dentro do mesmo grupo de componentes (com 50 componentes principais e 60 componentes independentes), como é visível na Figura 37.

RAVDESS exibiu o menor desempenho usando o método proposto, provavelmente devido ao fato de conter o menor número de parâmetros com distribuição normal. Além disso, usar ICA isoladamente resultou em uma acurácia melhor do que PCA, como retratado na Figura 35, indicando assim que PCA é menos eficaz para dados não normais.

As métricas na Figura 38 indicam a melhoria do modelo, onde a acurácia aumentou de 52,6% para 59,9% para 100 componentes independentes e 50 componentes principais. Porém, nesta base de dados as métricas acabaram tendo *outliers* e uma dispersão bem maior.

Em todas as bases analisadas, a precisão apresentou médias superiores e menor dispersão em relação ao *recall* utilizando o método proposto (PCA+ICA com subconjunto gaussiano), o que sugere que os modelos cometem menos erros de classificação de falsos positivos ao identificar as emoções. Isso indica que os modelos são mais criteriosos ao reconhecer uma emoção específica, mas podem não captar todos os áudios associados a essa emoção, resultando em menor sensibilidade para identificar corretamente todas as emoções presentes.

Para contextualizar, fez-se a comparação dos resultados desta pesquisa com aqueles de outros estudos, é importante destacar que alguns desses estudos não utilizaram os mesmos 6373 parâmetros acústicos da biblioteca *openSMILE* empregados nesta pesquisa. Além disso, os métodos de validação adotados diferem entre os estudos. Enquanto utilizamos validação cruzada com 10 divisões (*10-fold*), outros trabalhos empregaram métodos como divisão de dados em 80–20 ou *leave-one-speaker-out*, conforme mostrado nas Tabelas 13 e 14, o que pode ter influenciado os valores de acurácia.

A Análise das matrizes de confusão indicou que o método proposto melhorou o reconhecimento de algumas emoções específicas em cada conjunto de dados, embora tenha apresentado desafios para outras. Em EmoDB, houve uma melhora significativa para a emoção "tristeza", mas uma leve redução para "tédio". Em SAVEE, as emoções "neutro"e "felicidade"foram beneficiadas, enquanto "raiva"apresentou desempenho reduzido. Para RAVDESS, "raiva"e "calma"mostraram melhorias, mas "neutro"apresentou queda, destacando a natureza heterogênea dos impactos das técnicas de redução de dimensionalidade sobre diferentes emoções.

Na base de dados SAVEE, a combinação de PCA e ICA mostrou-se benéfica para as emoções "neutro"e "felicidade", enquanto a emoção "raiva"sofreu uma diminuição na acurácia, destacando a necessidade de uma análise mais detalhada sobre como diferentes emoções reagem a esse processo de transformação.

Por outro lado, na base de dados RAVDESS, a abordagem de PCA com ICA teve um

Tabela 13 – Comparação com o método proposto baseado no desempenho de reconhecimento. VC - Validação cruzada.

Base de Dados	Método	Classificador	Autor	Proporção de Divisão	Acurácia (%)
EmoDB	264 PCA	SVM	(Özseven, 2019)	VC 10-fold	81.71
	PCA+ICA	**SVM**	**autor(2024)**	**VC 10-fold**	**84.30**
SAVEE	235 PCA	SVM	(Özseven, 2019)	VC 10-fold	72.39
	PCA+LDA	SVM	(Liu et al., 2018)	80-20	72.23
	PCA+ICA	**SVM**	**autor(2024)**	**VC 10-fold**	**75.40**
RAVDESS	PCA	SVM	(Vu et al., 2022)	VC 10-fold	42.96
	PCA+ICA	**SVM**	**autor(2024)**	**VC 10-fold**	**59.90**

impacto positivo nas emoções "raiva"e "calma", porém causou uma piora na emoção "neutro", sendo a única emoção com desempenho reduzido. Novamente, um resultado que sugere que a aplicação de técnicas de redução de dimensionalidade, como PCA e ICA, pode não ser igualmente eficaz para todas as emoções e que o impacto dessas transformações pode variar de acordo com a natureza dos dados e das emoções envolvidas.

As evidências levantadas nesta pesquisa reforçam a ideia de que, embora a redução de dimensionalidade com PCA e ICA tenha o potencial de melhorar o desempenho em tarefas de reconhecimento de emoções, o efeito dessa abordagem não é homogêneo entre diferentes estados emocionais.

A comparação de resultados de acurácia entre estudos não é precisa devido à falta de padronização. Estudos variam no número de emoções consideradas para uma mesma base de dados, nos parâmetros extraídos e nos processos de pré-processamento aplicados aos áudios. Essas diferenças impedem uma comparação precisa e impedem afirmar com certeza se um método é superior ao outro.

O objetivo principal deste trabalho não foi superar a acurácia alcançada pelo estado da arte, mas sim propor uma abordagem nova que considera a distribuição das características ao aplicar métodos de redução de características, como a análise de componentes principais (PCA) e a análise de componentes independentes (ICA).

Os resultados desta pesquisa demonstram que o método de seleção de características em múltiplas etapas proposto alcança uma acurácia superior em comparação com métodos similares; ver Tabela 13. Garantiu-se que apenas as características mais relevantes e informativas fossem

[1] MFMC -
[2] GA - Algoritmo genético, do inglês *genetic algorithm*
[3] LLDs e VGG -
[4] GWO - Lobos Cinzentos - do inglês *Grey Wolf Optimizer*

Tabela 14 – Comparação do desempenho de reconhecimento do método proposto sem aplicar técnicas de transformação de redução de parâmetros de outros autores.

Base de Dados	**Método**	**Classificador**	**Autor**	**Proporção de Divisão**	**Acurácia (%)**
EmoDB	86 parâmetros	SVM	(Özseven, 2019)	VC 10-fold	84.07
	MFMC [1]	SVM	(Ancilin; Milton, 2021)	VC 10-fold	81.50
	GA [2]	SVM	(Mishra; Warule; Deb, 2023)	VC 5-fold	85.60
	Espectrograma	Rede GRU	(Xu et al., 2024)	leave-one-speaker-out (LOSO)	88.93
	PCA+ICA	**SVM**	**autor(2024)**	**VC 10-fold**	**84.30**
SAVEE	129 parâmetros	SVM	(Özseven, 2019)	VC 10-fold	77.92
	LLDs+VGGs [3]	DNN	(Wang et al., 2021)	NA	66.20
	MFMC	SVM	(Ancilin; Milton, 2021)	VC 10-fold	75.63
	GWO [4]	KNN	(Shahin et al., 2023)	90-10	83.54
	PCA+ICA	**SVM**	**autor(2024)**	**VC 10-fold**	**75.40**
RAVDESS	MFMC	SVM	(Ancilin; Milton, 2021)	VC 10-fold	64.31
	2D+VGG-16	DNN	(Aggarwal et al., 2022)	80-20	81.94
	182 parâmetros	SVM	(Shahin et al., 2023)	90-10	49.65
	GWO	KNN	(Shahin et al., 2023)	90-10	80.48
	PCA+ICA	**SVM**	**autor(2024)**	**VC 10-fold**	**59.90**

utilizadas, filtrando dados não informativos antes de aplicar PCA e ICA, o que levou a melhorias no desempenho do modelo.

No entanto, os resultados também enfatizam a necessidade de abordagens adaptativas que considerem as características específicas de cada base de dados, como o número de emoções, a distribuição dos parâmetros e as peculiaridades de cada estado emocional. Além disso, é possível que os parâmetros acústicos sofram influência em etapas anteriores ao processamento propriamente dito, como durante a gravação, os algoritmos utilizados para extração de parâmetros e pela qualidade do áudio capturado.

5 CONCLUSÃO

Neste trabalho, foi proposto um método de seleção e redução de parâmetros baseado em subconjuntos, levando em consideração a distribuição dos parâmetros acústicos da voz. O método, que combina PCA para os parâmetros com distribuição normal e ICA para o conjunto completo de dados, demonstrou eficácia em melhorar a acurácia do modelo para a detecção de emoções, ao mesmo tempo em que reduziu o número de parâmetros necessários para a classificação. A validação foi realizada em três bases de dados distintas (*Berlin EmoDB*, *SAVEE* e *RAVDESS*), e os resultados confirmaram que a aplicação de PCA e ICA, de maneira adaptada às distribuições dos parâmetros, é um caminho para otimização de modelos de reconhecimento de emoções.

A acurácia para o banco de dados SAVEE foi melhorada com o método proposto, com um aumento de 5% na acurácia para o mesmo número de componentes. Por outro lado, o método proposto teve um desempenho fraco quando foi aplicado ao banco de dados RAVDESS, que continha o menor número de parâmetros que seguiam uma distribuição normal. Na base dados RAVDESS ao comparar a acurácia para 90 componentes independentes com a de 100 componentes independentes + 50 componentes principais mostrou um aumento de apenas 0,1%, enfatizando a relação entre a eficácia de PCA e o número de parâmetros com distribuição normal.

Além disso, as matrizes de confusão revelaram padrões específicos de confundir certas emoções, indicando áreas onde o modelo poderia ser aprimorado. A comparação com a literatura relacionada demonstra que o método proposto é competitivo, alcançando uma acurácia similar.

Os resultados indicaram que a aplicação de PCA exclusivamente a parâmetros com distribuição normal foi eficaz para melhorar a acurácia do modelo, especialmente nas bases de dados EmoDB e RAVDESS. A incorporação de ICA, aplicada a todo o conjunto de dados, resultou em uma melhora adicional no desempenho. O modelo não apresentou um desempenho significativamente superior no banco de dados RAVDESS, indicando que pode não ser tão eficaz devido ao menor número de parâmetros com distribuição normal. As matrizes de confusão também revelaram que, embora o método tenha aumentado a acurácia geral, algumas emoções específicas apresentaram maior dificuldade de reconhecimento, sugerindo que a abordagem pode ser mais eficaz para determinados estados emocionais.

Ao comparar com a literatura existente, o método proposto mostrou-se competitivo, alcançando uma acurácia similar ou superior à de outros métodos, com a vantagem de utilizar um número menor de parâmetros e, portanto, com menor custo computacional. Esse resultado sublinha a importância da análise da distribuição dos dados antes da aplicação de técnicas de redução dimensional, como PCA e ICA, para melhorar o desempenho dos modelos.

Trabalhos futuros podem explorar alternativas ao teste de Kruskal-Wallis para a seleção de parâmetros, com o objetivo de identificar métodos que possam ser mais eficazes em diferentes contextos de dados. Além disso, a substituição do SVM por outros métodos, aprendizado profundo, como redes neurais convolucionais ou redes neurais recorrentes, pode ser investigada,

uma vez que esses algoritmos têm se mostrado promissores em tarefas de reconhecimento de emoções, especialmente com grandes volumes de dados. Outras abordagens para a redução de dimensionalidade, como t-SNE ou *isomap*, também podem ser avaliadas para verificar seu impacto no desempenho do modelo.

A análise dos parâmetros acústicos sob diferentes condições de gravação, qualidade do áudio, como possível fonte de variabilidade nos resultados, também seria uma linha interessante de investigação, para entender melhor como esses fatores afetam o desempenho do reconhecimento de emoções.

Ademais, é relevante investigar o uso de algoritmos alternativos para a extração de parâmetros além do openSMILE, que é amplamente utilizado na literatura, mas pode não capturar todas as nuances presentes nos dados de áudio.

Embora neste trabalho o método tenha sido testado nos bancos de dados EmoDB, SAVEE e RAVDESS, seria interessante avaliar o desempenho desses modelos em cenários do mundo real, com gravações de campo e em ambientes mais dinâmicos. Essa avaliação poderia fornecer uma referência sobre a robustez do modelo em situações de gravação menos controladas.

Além disso, uma possível abordagem futura seria realizar uma comparação que não foi explorada neste estudo, criando um modelo em uma base de dados e testando-o em outras, com o objetivo de verificar se o método proposto reduz o erro causado por diferenças entre idiomas e características individuais de cada base de dados.

REFERÊNCIAS

ABDEL-HAMID, Lamiaa; SHAKER, Nabil H.; EMARA, Ingy. Analysis of linguistic and prosodic features of bilingual arabic–english speakers for speech emotion recognition. **IEEE Access**, v. 8, p. 72957–72970, 2020. Citado na página 51.

AGGARWAL, Apeksha et al. Two-way feature extraction for speech emotion recognition using deep learning. **Sensors**, v. 22, n. 6, 2022. ISSN 1424-8220. Citado 2 vezes nas páginas 52 e 91.

ALCAIM, Abraham; OLIVEIRA, Carlos Alexandre dos Santos. **Fundamentos do Processamento de Sinais de Voz e Imagem**. 1. ed. Rio de Janeiro: Editora Puc-Rio, 2011. ISBN 9788571932678. Citado na página 33.

ALGHIFARI, Muhammad et al. On the use of voice activity detection in speech emotion recognition. **Bulletin of Electrical Engineering and Informatics**, v. 8, 12 2019. Citado na página 17.

ALIMURADOV, Alan K et al. Development of Natural Emotional Speech Database for Training Automatic Recognition Systems of Stressful Emotions in Human-Robot Interaction. In: **2020 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR)**. Innopolis, Russia: DCNAIR, 2020. p. 11–16. Citado na página 110.

ALJUHANI, Reem Hamed; ALSHUTAYRI, Areej; ALAHDAL, Shahd. Arabic speech emotion recognition from saudi dialect corpus. **IEEE Access**, v. 9, p. 127081–127085, 2021. Citado na página 51.

AMJAD, Ammar et al. Recognizing semi-natural and spontaneous speech emotions using deep neural networks. **IEEE Access**, v. 10, p. 37149–37163, 2022. Citado na página 52.

ANAGNOSTOPOULOS, Christos-Nikolaos; ILIOU, Theodoros; GIANNOUKOS, Ioannis. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. **Artificial Intelligence Review**, v. 43, fev. 2012. Citado 3 vezes nas páginas 104, 105 e 108.

ANCILIN, J.; MILTON, A. Improved speech emotion recognition with mel frequency magnitude coefficient. **Applied Acoustics**, v. 179, p. 108046, 2021. ISSN 0003-682X. Citado 2 vezes nas páginas 51 e 91.

ANDERSON, T. W.; DARLING, D. A. A test of goodness of fit. **Journal of the American Statistical Association**, American Statistical Association, v. 49, p. 765–769, 1954. ISSN 0162-1459,1537-274X. Citado na página 50.

BARBETTA, Paulo A.; REIS, Maria M.; BORNIA, André C. **Estatística para Cursos de Engenharia, Computação e Ciência de Dados**. Rio de Janeiro, Brasil: LTC, 2024. Citado 2 vezes nas páginas 44 e 62.

BIOLCHINI, Jorge et al. **Systematic Review in Software Engineering: Relevance and Utility**. Rio de Janeiro, Brasil, 2005. Citado na página 104.

BOERSMA, Paul; WEENINK, David. **PRAAT: Doing Phonetics by Computer**. 2023. https://www.fon.hum.uva.nl/praat/. Accessed: 30 August 2024. Citado na página 17.

BOLÓN-CANEDO, Verónica; SÁNCHEZ-MARONO, Noelia; ALONSO-BETANZOS, Amparo. A review of feature selection methods on synthetic data. **Knowledge and Information Systems**, v. 34, n. 3, p. 483–519, mar. 2012. Citado 2 vezes nas páginas 18 e 19.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the Fifth Annual Workshop on Computational Learning Theory**. New York, NY, USA: ACM Press, 1992. p. 144–152. Citado na página 52.

BROOKES, Mike. **VOICEBOX: Speech Processing Toolbox for MATLAB**. 2024. Disponível em: https://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Acesso em: 30 ago. 2024. Citado na página 17.

BURKHARDT, Felix et al. A database of german emotional speech. In: **Proceedings of Interspeech**. Lisboa: Interspeech, 2005. p. 1517–1520. Citado 3 vezes nas páginas 17, 61 e 104.

CANDÉS, Emmanuel J. et al. Robust principal component analysis. **J. ACM**, Association for Computing Machinery, New York, NY, USA, v. 58, n. 3, jun. 2011. ISSN 0004-5411. Citado na página 45.

CANNON, Walter B. The james-lange theory of emotions: A critical examination and an alternative theory. **The American Journal of Psychology**, v. 100, p. 567–586, 1987. Citado na página 28.

CHEN, Chuang; CHELLALI, Ryad; XING, Yin. Speech emotion recognition based on kernel principal component analysis and optimized support vector machine. In: **2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)**. Harbin, China: IEEE, 2018. p. 751–755. Citado na página 21.

CHEN, Guanghui; ZENG, Xiaoping. Multi-modal emotion recognition by fusing correlation features of speech-visual. **IEEE Signal Processing Letters**, v. 28, p. 533–537, 2021. Citado na página 52.

CHEN, Lijiang et al. Speech emotion recognition: Features and classification models. **Digital Signal Processing**, v. 22, n. 6, p. 1154–1160, Dec. 2012. ISSN 1051-2004. Citado 3 vezes nas páginas 21, 22 e 66.

CHENG, Libo; CHEN, Pengyu. Remote sensing image denoising based on gaussian curvature and shearlet transform. **IEEE Access**, v. 11, p. 97716–97725, 2023. Citado na página 50.

COWIE, Roderick et al. Emotion recognition in human-computer interaction. **IEEE Signal Processing Magazine**, Institute of Electrical and Electronics Engineers Inc., v. 18, n. 1, p. 32–80, Jan. 2001. ISSN 1053-5888. Citado 3 vezes nas páginas 22, 31 e 32.

DAMASIO, Antonio R. **O Erro de Descartes**. São Paulo: Companhia das Letras, 1996. Citado na página 30.

DARCY, Isabelle; FONTAINE, Nathalie M.G. The Hoosier Vocal Emotions Corpus: A validated set of North American English pseudo-words for evaluating emotion processing. **Behavior Research Methods**, Springer, v. 52, n. 2, p. 901–917, Apr. 2020. Citado 2 vezes nas páginas 108 e 110.

DARWIN, Charles. **The expression of the emotions in man and animals**. New York: D. Appleton and Co., 1916. 406 p. Citado na página 27.

DAVIS, Steven B.; MERMELSTEIN, Paul. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 28, n. 4, p. 357–366, 1980. Citado na página 41.

DESCHAMPS-BERGER, Théo; LAMEL, Lori; DEVILLERS, Laurence. End-to-end speech emotion recognition: Challenges of real-life emergency call centers data recordings. In: **2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)**. Nara, Japan: IEEE, 2021. p. 1–8. Citado na página 16.

DEY, Arijit et al. A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition. **IEEE Access**, v. 8, p. 200953–200970, 2020. Citado na página 21.

DOUGLAS-COWIE, Ellen et al. Emotional speech: Towards a new generation of databases. **Speech Communication**, v. 40, p. 33–60, 04 2003. Citado na página 105.

DOUGLAS-COWIE, Ellen; COWIE, Roderick; SCHRÖDER, M. A new emotion database: Considerations, sources and scope. In: **Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research**. Belfast: Textflow, 2000. p. 39–44. Citado na página 104.

DUDA, Richard O.; HART, Peter E.; STORK, David G. **Pattern Classification**. 2. ed. New York: Wiley, 2001. ISBN 978-0-471-05669-0. Citado 4 vezes nas páginas 18, 19, 24 e 62.

DUJAILI, Mohammed Jawad Al; EBRAHIMI-MOGHADAM, Abbas. Automatic speech emotion recognition based on hybrid features with ann, lda and knn classifiers. **Multimedia Tools and Applications**, Apr. 2023. Citado na página 21.

EKMAN, Paul. **A Linguagem das Emoções**. Rio de Janeiro: Editora Leya, 2011. Citado na página 27.

EKMAN, Paul; FRIESEN, Wallace V. **Unmasking the face: A guide to recognizing emotions from facial clues**. Oxford: Prentice-Hall, 1975. Citado na página 30.

El Ayadi, Moataz; KAMEL, Mohamed S.; KARRAY, Fakhri. Survey on speech emotion recognition: Features, classification schemes, and databases. **Pattern Recognition**, v. 44, n. 3, p. 572–587, 2011. ISSN 0031-3203. Citado na página 105.

ER, Mehmet Bilal. A novel approach for classification of speech emotions based on deep and acoustic features. **IEEE Access**, v. 8, p. 221640–221653, 2020. Citado na página 51.

EYBEN, Florian; WÖLLMER, Martin; SCHULLER, Björn. opensmile–the munich versatile and fast open-source audio feature extractor. **ACM Multimedia**, 2010. Citado 3 vezes nas páginas 17, 58 e 59.

GALANIS, Dimitrios et al. Classification of emotional speech units in call centre interactions. In: **2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)**. Budapest, Hungary: IEEE, 2013. p. 403–406. Citado na página 16.

GÉRON, Aurélien. **Mãos A Obra: Aprendizado de Máquina com scikit–learn E tensorflow**. São Paulo: Alta Books, 2019. Citado na página 51.

GUO, Yi et al. A novel speech emotion recognition method based on feature construction and ensemble learning. **PLOS ONE**, Public Library of Science, v. 17, n. 8, p. 1–17, 08 2022. Citado na página 21.

GUYON, Isabelle; ELISSEEFF, André. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157–1182, 2003. Citado na página 18.

HAQ, Sanaul; JACKSON, Philip J. B.; EDGE, James D. Audio-visual feature selection and reduction for emotion classification. **The proceedings of international conference on auditory-visual speech processing**, p. 185–190, 2008. Citado 2 vezes nas páginas 17 e 61.

HASHEM, Ahlam; ARIF, Muhammad; ALGHAMDI, Manal. Speech emotion recognition approaches: A systematic review. **Speech Communication**, p. 102974, 2023. ISSN 0167-6393. Citado na página 60.

HECKE, T. Van. Power study of anova versus kruskal-wallis test. **Journal of Statistics and Management Systems**, Taylor & Francis, v. 15, n. 2-3, p. 241–247, 2012. Citado 2 vezes nas páginas 49 e 67.

HOLLANDER, Myles; WOLFE, Douglas A; CHICKEN, Eric. **Nonparametric Statistical Methods**. Hoboken, NJ: John Wiley Sons, 2013. ISBN 9781118553299. Citado 2 vezes nas páginas 49 e 50.

HOUWER, Jan de; HERMANS, Dirk. **Cognition and Emotion: Reviews of Current Research and Theories**. Hove; New York: Psychology Press, 2010. Citado na página 27.

HUANG, Xuedong; ACERO, Alex; HON, Hsiao-Wuen. **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**. 1st. ed. Upper Saddle River, NJ, USA,: Prentice Hall PTR, 2001. Citado na página 35.

HYVÄRINEN, Aapo; OJA, Erkki. Independent component analysis: algorithms and applications. **Neural Networks**, v. 13, n. 4, p. 411–430, 2000. ISSN 0893-6080. Citado na página 46.

HYVäRINEN, Aapo; OJA, Erkki. **Independent Component Analysis: Algorithms and Applications**. Hoboken, NJ, USA: John Wiley & Sons, 2010. ISBN: 978-0470746669. Citado 6 vezes nas páginas 22, 45, 46, 47, 48 e 65.

JAIN, Kabir et al. Investigation using mlp-svm-pca classifiers on speech emotion recognition. In: **2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)**. Uttar Pradesh, India: IEEE, 2022. p. 1–6. Citado na página 21.

JIANG, Haihua et al. Investigation of different speech types and emotions for detecting depression using different classifiers. **Speech Communication**, v. 90, p. 39–46, 2017. ISSN 0167-6393. Citado 2 vezes nas páginas 17 e 59.

JOLLIFFE, Ian T. **Principal Component Analysis**. New York: Springer, 2002. Citado 5 vezes nas páginas 22, 43, 45, 48 e 57.

KAMINSKA, Dorota; PELIKANT, Adam. Recognition of human emotion from a speech signal based on plutchik's model. **International Journal of Electronics and Telecommunications**, Polish Academy of Sciences Committee of Electronics and Telecommunications, vol. 58, n. No 2, 2012. Citado 2 vezes nas páginas 35 e 104.

KASURIYA, Sawit et al. Developing a Thai Emotional Speech Corpus from Lakorn (EMOLA). **Lang. Resour. Eval.**, Springer-Verlag, Berlin, Heidelberg, v. 53, n. 1, p. 17–55, 2019. ISSN 1574-020X. Citado 2 vezes nas páginas 104 e 110.

KAUR, Kamaldeep; SINGH, Parminder. Comparison of various feature selection algorithms in speech emotion recognition. **AIUB Journal of Science and Engineering (AJSE)**, v. 22, n. 2, p. 125–131, Aug. 2023. Citado na página 20.

KE, Xianxin et al. Speech emotion recognition based on pca and chmm. In: **2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)**. Chengdu, China: IEEE, 2019. p. 667–671. Citado na página 21.

KECMAN, Vojislav. **Support Vector Machines – An Introduction**. Berlin, Germany: Springer, 2005. v. 177. 605–605 p. ISBN 978-3-540-24388-5. Citado 2 vezes nas páginas 53 e 54.

KIM, Jangwon et al. Vocal tract shaping of emotional speech. **Computer Speech Language**, v. 64, nov 2020. Citado na página 110.

KINGESKI, Rafael. **Desenvolvimento de um Banco de Dados de voz com Emoções em Idioma Português Brasileiro**. Dissertação (Mestrado) — Universidade do Estado de Santa Catarina, Joinville, 2019. Citado na página 23.

KINGESKI, Rafael; HENNING, Elisa; PATERNO, Aleksander S. Fusion of pca and ica in statistical subset analysis for speech emotion recognition. **Sensors**, v. 24, n. 17, 2024. ISSN 1424-8220. Citado na página 26.

KINGESKI, Rafael.; SCHUEDA, Larissa A. P.; PATERNO, Aleksander S. Feature analysis for speech emotion classification. In: BASTOS-FILHO, Teodiano Freire; CALDEIRA, Eliete Maria de Oliveira; FRIZERA-NETO, Anselmo (Ed.). **XXVII Brazilian Congress on Biomedical Engineering**. Cham: Springer International Publishing, 2022. p. 2359–2365. ISBN 978-3-030-70601-2. Citado 2 vezes nas páginas 26 e 58.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, Aug. 2004. Citado na página 104.

KOSSAIFI, Jean et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. **CoRR**, abs/1901.0, 2019. Citado 2 vezes nas páginas 109 e 110.

KRISHNAN, Palani Thanaraj; RAJ, Alex Noel Joseph; RAJANGAM, Vijayarajan. Emotion classification from speech signal based on empirical mode decomposition and non-linear features. **Complex & Intelligent Systems**, v. 7, feb 2021. Citado na página 21.

KUCHIBHOTLA, Swarna; VANKAYALAPATI, Hima Deepthi; ANNE, Koteswara Rao. An optimal two stage feature selection for speech emotion recognition using acoustic features. **International Journal of Speech Technology**, v. 19, n. 4, p. 657–667, aug 2016. Citado na página 20.

LASSALLE, Amandine et al. The eu-emotion voice database. **Behavior Research Methods**, Springer New York LLC, v. 51, n. 2, p. 493–506, apr 2019. Citado na página 110.

LI, Dongdong et al. Exploiting the potentialities of features for speech emotion recognition. **Information Sciences**, v. 548, p. 328–343, 2021. ISSN 0020-0255. Citado na página 20.

LI, Tao et al. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 30, p. 1448–1460, 2022. Citado 2 vezes nas páginas 16 e 20.

LIESKOVSKÁ, Eva et al. A review on speech emotion recognition using deep learning and attention mechanism. **Electronics**, v. 10, n. 10, 2021. ISSN 2079-9292. Citado na página 66.

LIU, Zhen-Tao et al. Speech emotion recognition based on an improved brain emotion learning model. **Neurocomputing**, v. 309, p. 145–156, 2018. ISSN 0925-2312. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925231218305344. Citado 3 vezes nas páginas 21, 61 e 90.

LIVINGSTONE, Steven R.; RUSSO, Frank A.; NAJBAUER, Joseph. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. **PLoS ONE**, Public Library of Science, v. 13, 2018. ISSN 1932-6203. Citado 3 vezes nas páginas 17, 35 e 61.

LÓPEZ, Osval A. Montesinos; LÓPEZ, Abelardo Montesinos; CROSSA, Jose. Overfitting, model tuning, and evaluation of prediction performance. **Multivariate Statistical Machine Learning Methods for Genomic Prediction**, p. 109–139, 2022. Citado na página 24.

MARTIN, O. et al. The enterface'05 audio-visual emotion database. In: **Proceedings of the 22nd International Conference on Data Engineering Workshops**. USA: IEEE Computer Society, 2006. (ICDEW '06). Citado na página 104.

MEFTAH, Ali Hamid et al. Speaker identification in different emotional states in arabic and english. **IEEE Access**, v. 8, p. 60070–60083, 2020. Citado na página 17.

MISHRA, Siba Prasad; WARULE, Pankaj; DEB, Suman. Chirplet transform based time frequency analysis of speech signal for automated speech emotion recognition. **Speech Communication**, v. 155, p. 102986, 2023. ISSN 0167-6393. Citado 3 vezes nas páginas 23, 51 e 91.

MURRAY, Iain R.; ARNOTT, John L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. **The Journal of the Acoustical Society of America**, v. 93, n. 2, p. 1097–1108, 1993. Citado na página 34.

NEZAMI, Omid Mohamad; LOU, Paria Jamshid; KARAMI, Mansoureh. Shemo: A large-scale validated database for persian speech emotion detection. **Language Resources and Evaluation**, Berlin, Heidelberg, v. 53, n. 1, p. 1–16, 2019. ISSN 1574-020X. Citado na página 110.

NOGUEIRA, Kennyo. Estudo de respostas emocionais às cores no contexto de cartazes de cinema. **Design e Tecnologia**, v. 8, p. 1, 06 2018. Citado na página 32.

NOWLIS, Vincent; NOWLIS, Helen H. The description and analysis of mood. **Annals of the New York Academy of Sciences**, v. 65, n. 4, p. 345–355, 1956. Citado na página 31.

OSMAN, Ibrahim H; KELLY, James P. **Meta-Heuristics**. New York: Springer Science & Business Media, 2012. ISBN 9781461313618. Citado na página 20.

PAH, Nemuel D.; INDRAWATI, Veronica; KUMAR, Dinesh K. Voice features of sustained phoneme as covid-19 biomarker. **IEEE Journal of Translational Engineering in Health and Medicine**, v. 10, p. 1–9, 2022. Citado na página 50.

PALACIOS, Daniel et al. An ica-based method for stress classification from voice samples. **Neural Computing and Applications**, v. 32, n. 24, p. 17887–17897, Oct 2019. Citado 2 vezes nas páginas 21 e 52.

PALACIOS, Daniel et al. An ica-based method for stress classification from voice samples. **Neural Computing and Applications**, Springer Science, v. 32, n. 24, p. 17887–17897, Oct 2019. Citado na página 21.

PAN, Yixiong; SHEN, Peipei; SHEN, Liping. Speech emotion recognition using support vector machine. **International Journal of Smart Home**, v. 6, n. 2, April 2012. Citado na página 104.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 66.

PETROVICA, Sintija; ANOHINA-NAUMECA, Alla; EKENEL, Hazım Kemal. Emotion recognition in affective tutoring systems: Collection of ground-truth data. **Procedia Computer Science**, v. 104, p. 437–444, 2017. Citado na página 16.

PICARD, Rosalind Wright. **Affective Computing**. Cambridge, MA: Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995. Citado 2 vezes nas páginas 16 e 32.

PLUTCHIK, Robert. **The Emotions: Facts and Theories, and a New Model**. New York, NY: Random House, 1962. (Random House Studies in Psychology). Citado na página 29.

QUAN, Changqin et al. Reduce the dimensions of emotional features by principal component analysis for speech emotion recognition. In: **Proceedings of the 2013 IEEE/SICE International Symposium on System Integration**. Kobe, Japan: IEEE, 2013. p. 222–226. Citado na página 21.

RABINER, L.R.; JUANG, B.H. **Fundamentals of Speech Recognition**. Englewood Cliffs, NJ, USA: Pearson Education, 1993. (Pearson Education Signal Processing Series). ISBN 9788129701381. Citado na página 33.

RABINER, L.R.; SCHAFER, R.W. **Digital Processing of Speech Signals**. Englewood Cliffs, NJ, USA: Prentice-Hall, 1980. (Prentice-Hall Signal Processing Series: Advanced Monographs). Citado 3 vezes nas páginas 35, 36 e 104.

RABINER, Lawrence; SCHAFER, Ronald. **Theory and Applications of Digital Speech Processing**. 1st. ed. USA: Prentice Hall Press, 2010. ISBN 0136034284. Citado 2 vezes nas páginas 42 e 43.

RABINER, Lawrence R.; SCHAFER, Ronald W. **Introduction to Digital Speech Processing**. Hanover, MA, USA: Now Publishers Inc., 2007. Citado 5 vezes nas páginas 38, 39, 40, 41 e 42.

RAMBABU, Banothu et al. Iiit-h temd semi-natural emotional speech database from professional actors and non-actors. In: CALZOLARI, Nicoletta et al. (Ed.). **Proceedings of The 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 1538–1545. Citado na página 110.

RAVI; TARAN, Sachin. A nonlinear feature extraction approach for speech emotion recognition using vmd and tkeo. **Applied Acoustics**, v. 214, p. 109667, 2023. ISSN 0003-682X. Citado 2 vezes nas páginas 19 e 58.

ROSSUM, Guido Van; JR, Fred L Drake. **Python reference manual**. Amsterdam: Centrum voor Wiskunde en Informatica, 1995. Citado na página 58.

RUSSELL, James. A circumplex model of affect. **Journal of Personality and Social Psychology**, v. 39, p. 1161–1178, 12 1980. Citado 2 vezes nas páginas 28 e 31.

RUSSELL, J A; FEHR, B. Fuzzy concepts in a fuzzy hierarchy: Varieties of anger. **Journal of Personality and Social Psychology**, v. 67, n. 2, p. 186–205, 1994. Citado na página 29.

SATO, Ryota et al. Creation and analysis of emotional speech database for multiple emotions recognition. In: **Proceedings of the 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2020**. Yangon, Myanmar: IEEE, 2020. p. 33–37. Citado na página 110.

SCHACHTER, Stanley; SINGER, Jerome E. Cognitive, social, and physiological determinants of emotional state. **Psychological Review**, v. 69, p. 379–399, 1962. Citado na página 28.

SCHERER, Klaus R. Nonlinguistic vocal indicators of emotion and psychopathology. **Emotions in Personality and Psychopathology**, p. 493–529, 1979. Citado na página 34.

SCHERER, Klaus R. Vocal affect expression: A review and a model for future research. **Psychological Bulletin**, v. 99, p. 143–165, 1986. Citado na página 34.

SCHERER, Klaus R; LONDON, Harvey; WOLF, Jared J. The voice of confidence: Paralinguistic cues and audience evaluation. **Journal of Research in Personality**, v. 7, n. 1, p. 31–44, 1973. ISSN 0092-6566. Citado na página 22.

SCHULLER, Björn et al. Recognising realistic emotions and affect in speech: State of the art and challenges. **IEEE Transactions on Affective Computing**, v. 3, n. 1, p. 18–31, 2013. Citado na página 59.

SHAHIN, Ismail et al. An efficient feature selection method for arabic and english speech emotion recognition using grey wolf optimizer. **Applied Acoustics**, v. 205, p. 109279, 2023. ISSN 0003-682X. Citado 4 vezes nas páginas 17, 21, 51 e 91.

SHARMA, R.; NEUMANN, U.; KIM, C. Emotion recognition in spontaneous emotional utterances from movie sequences. **Proceedings of the WSEAS International Conference on Electronics, Control & Signal Processing**, 2002. Citado na página 35.

SHEIKHAN, Mansour; BEJANI, Mahdi; GHARAVIAN, Davood. Modular neural-svm scheme for speech emotion recognition using anova feature selection method. **Neural Computing and Applications**, v. 23, n. 1, p. 215–227, Jan 2012. Citado na página 19.

SNEDDON, I et al. The belfast induced natural emotion database. **IEEE Transactions on Affective Computing**, Institute of Electrical and Electronics Engineers, v. 3, p. 32–41, 2012. Citado na página 104.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, p. 427–437, 07 2009. Citado na página 55.

SONG, Peng. Transfer linear subspace learning for cross-corpus speech emotion recognition. **IEEE Transactions on Affective Computing**, v. 10, n. 2, p. 265–275, 2019. Citado 3 vezes nas páginas 17, 59 e 66.

SULTANA, Sadia et al. Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla. **PLoS ONE**, Public Library of Science, v. 16, n. 4 April, apr 2021. Citado na página 110.

SUN, Linhui et al. Multi-classification speech emotion recognition based on two-stage bottleneck features selection and mcjd algorithm. **Signal, Image and Video Processing**, Springer Science and Business Media LLC, v. 16, n. 5, p. 1253–1261, Jan 2022. Citado na página 21.

TIWARI, Pradeep; DARJI, A. D. A novel s-lda features for automatic emotion recognition from speech using 1-d cnn. **International Journal of Mathematical, Engineering and Management Sciences**, v. 7, n. 1, p. 49–67, 2022. Citado na página 21.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin F. Towards principled feature selection: Relevancy, filters and wrappers. In: BISHOP, Christopher M.; FREY, Brendan J. (Ed.). **Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics**. Key West, FL, USA: PMLR, 2003. (Proceedings of Machine Learning Research, R4), p. 300–307. Citado 2 vezes nas páginas 19 e 20.

UDDIN, Md. Al Mamun Md. Palash; HOSSAIN, Md. Ali. Pca-based feature reduction for hyperspectral remote sensing image classification. **IETE Technical Review**, Taylor & Francis, v. 38, n. 4, p. 377–396, 2021. Citado na página 18.

VERMA, Aman et al. An acoustic analysis of speech for emotion recognition using deep learning. In: **2022 1st International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing (PCEMS)**. Prayagraj, India: IEEE, 2022. p. 68–73. Citado na página 51.

VIDAL, Andrea et al. Msp-face corpus: A natural audiovisual emotional database. In: **Proceedings of the 2020 International Conference on Multimodal Interaction**. New York, NY, USA: Association for Computing Machinery, 2020. (ICMI '20), p. 397–405. Citado 2 vezes nas páginas 108 e 110.

VU, Linh et al. Improved speech emotion recognition based on music-related audio features. In: **2022 30th European Signal Processing Conference (EUSIPCO)**. Belgrade, Serbia: EURASIP, 2022. p. 120–124. Citado 3 vezes nas páginas 51, 61 e 90.

WANG, Chunyi et al. Speech emotion recognition based on multi-feature and multi-lingual fusion. **Multimedia Tools and Applications**, v. 81, n. 4, p. 4897–4907, Aug. 2021. Citado 2 vezes nas páginas 52 e 91.

WENINGER, Felix et al. On the acoustics of emotion in audio: What speech, music, and sound have in common. Frontiers Media SA, v. 4, n. 292, May 2013. Citado na página 59.

WÖLLMER, Florian et al. Speech emotion recognition using long short-term memory recurrent neural networks. In: **Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)**. Taipei, Taiwan: IEEE, 2009. p. 1229–1232. Citado na página 59.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **IEEE Transactions on Evolutionary Computation**, v. 1, n. 1, p. 67–82, 1997. Citado na página 20.

WU, Chung-Hsien; LIANG, Wei-Bin. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. **IEEE Transactions on Affective Computing**, v. 2, n. 1, p. 10–21, 2011. Citado 2 vezes nas páginas 17 e 52.

XIE, Jie; ZHU, Mingying; HU, Kai. Fusion-based speech emotion classification using two-stage feature selection. **Speech Communication**, v. 152, p. 102955, 2023. ISSN 0167-6393. Citado 8 vezes nas páginas 17, 18, 20, 21, 22, 59, 61 e 66.

XU, Chunsheng et al. A new network structure for speech emotion recognition research. **Sensors**, v. 24, n. 5, 2024. ISSN 1424-8220. Citado 2 vezes nas páginas 17 e 91.

YANG, Chaoqing et al. Emotion-dependent language featuring depression. **Journal of Behavior Therapy and Experimental Psychiatry**, v. 81, p. 101883, 2023. ISSN 0005-7916. Citado na página 16.

YAZICI, Asli. Aristotle's theory of emotion. **Turkish Studies**, 01 2015. Citado na página 28.

YAZICI, Berna; ASMA, Senay. A comparison of various tests of normality. **Journal of Statistical Computation and Simulation**, v. 77, p. 175–183, 02 2007. Citado na página 64.

YILDIRIM, Serdar; KAYA, Yasin; KıLıç, Fatih. A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. **Applied Acoustics**, v. 173, p. 107721, 2021. ISSN 0003-682X. Citado 2 vezes nas páginas 21 e 22.

ZHANG, Xiaoye et al. A method of building phoneme-level chinese audio-visual emotional database. In: **2020 International Conference on Culture-oriented Science Technology (ICCST)**. Beijing, China: IEEE, 2020. p. 504–509. Citado na página 21.

ZHANG, Zixing et al. Cooperative learning and its application to emotion recognition from speech. **IEEE/ACM Transactions on Audio Speech and Language Processing**, v. 23, n. 1, p. 115–126, 2015. ISSN 23299290. Citado 2 vezes nas páginas 105 e 110.

ZHOU, Ying et al. Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: A cross-sectional observational study. **International Journal of Nursing Studies**, v. 146, p. 104562, 2023. ISSN 0020-7489. Citado na página 16.

ZVAREVASHE, Kudakwashe; OLUGBARA, Oludayo. Ensemble learning of hybrid acoustic features for speech emotion recognition. **Algorithms**, v. 13, n. 3, 2020. ISSN 1999-4893. Citado na página 52.

ZVAREVASHE, Kudakwashe; OLUGBARA, Oludayo O. Gender voice recognition using random forest recursive feature elimination with gradient boosting machines. In: **2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)**. London, UK: IEEE, 2018. p. 1–6. Citado na página 20.

ÖZSEVEN, Turgut. A novel feature selection method for speech emotion recognition. **Applied Acoustics**, v. 146, p. 320–326, 2019. ISSN 0003-682X. Citado 13 vezes nas páginas 17, 20, 21, 22, 23, 24, 51, 52, 59, 61, 66, 90 e 91.

ANEXO A – REVISÃO SISTEMÁTICA DE LITERATURA

Neste trabalho foi proposto uma forma de pesquisa sistemática bibliográfica baseada em algumas literaturas sobre métodos de revisão bibliográfica sistemática (Biolchini et al., 2005; Kitchenham, 2004). Após uma pesquisa exploratória prévia, onde foi buscado alguns estudos que fizessem um levantamento bibliográfico de bases de dados de voz contendo emoções, não foi identificado nenhum estudo que reunisse as informações necessárias para responder a questão que foi levantada durante a leitura desses trabalhos. Definiu-se então um protocolo para a revisão sistemática com a finalidade de encontrar lacunas em estudos similares.

Reconhecimento de emoções por voz é um tema de pesquisa que vem sendo estudado e aprimorado nos últimos anos, principalmente em relação as técnicas de processamento de sinais e aprendizado de máquina. É comum utilizar bases de dados que já foram feitas em outros estudos e que são bem conhecidas no meio científico. Contudo não existe um padrão nestes dados, nem na forma de coleta, nem na quantidade dos dados que deve ser utilizado para o reconhecimento de uma forma prática. O objetivo dessa revisão é identificar e analisar estudos que tenham construído bases de dados e quais foram os métodos e técnicas utilizados para esta coleta.

Em geral, os estudos de reconhecimento de emoções por voz são feitos em três etapas (Kaminska; Pelikant, 2012):

- Coleta de dados de voz: Nesta parte o pesquisador pode optar em utilizar uma base de dados pronta ou criar sua própria base de dados. Os bancos de dados de voz utilizados para identificar emoções podem ser construídos com emoções expressas por atores (Burkhardt et al., 2005; Martin et al., 2006), por coleta de áudios em vídeos e filmes (Kasuriya et al., 2019), induzindo com vídeos ou atividades (Sneddon et al., 2012) ou coletando emoções em situações cotidianas (Douglas-Cowie; Cowie; Schröder, 2000);

- Extração de parâmetros dos sinais de voz: Nesta parte é feita a extração de parâmetros dos sinais. Utiliza-se as mesmas técnicas de processamento de sinais utilizados em reconhecimento de fala e reconhecimento de locutor (Pan; Shen; Shen, 2012; Rabiner; Schafer, 1980);

- Aprendizado de máquina: Os parâmetros extraídos na segunda fase são aplicados em um aprendizado de máquina supervisionado (Pan; Shen; Shen, 2012; Anagnostopoulos; Iliou; Giannoukos, 2012).

Pode-se abordar questões nas três etapas descritas, porém em estudos similares foram encontrados lacunas maiores no primeiro item. A coleta de dados não possui um método em comum nos estudos, tornando difícil a comparação de um estudo com o outro, as pesquisas de estudos secundários em geral tentam descrever as três etapas em estudos primários, focando muito mais no desempenho que cada estudo obteve na classificação das emoções e não na

construção de um *corpus linguístico* (Anagnostopoulos; Iliou; Giannoukos, 2012; El Ayadi; Kamel; Karray, 2011).

Para que se possa fazer uma afirmação de que existe essa lacuna e que se possa propor uma pesquisa que contribuirá diretamente com estudos similares, faz-se necessário uma revisão bibliográfica sistemática.

A.1 TRABALHOS RELACIONADOS

Em (Anagnostopoulos; Iliou; Giannoukos, 2012), foi feita uma busca sobre trabalhos relacionados entre os anos de 2000 e 2011. Os autores concluíram que os estudos nessa área não possuem um banco de dados e sim conjunto de dados em pequena escala, também cita que não estão amplamente disponíveis. Para os autores é muito difícil a construção de um banco de dados utilizando fala natural e que estudos desse tipo são escassos e os que possuem maior número de dados são os obtidos através de programas de TV, rádio ou *call centers*. Este trabalho descreve as três etapas de um reconhecimento de emoções, descrevendo-as e dispondo informações sobre os resultados do desempenho dos classificadores de emoções obtidos nos estudos.

Em (Douglas-Cowie et al., 2003) o autor descreve detalhadamente sobre as bases de dados utilizadas para este fim e faz uma análise estatística de alguns dados. Porém trata-se trabalho que possui quase 20 anos. O autor afirma que existe uma variabilidade enorme no que diz respeito a coleta desses dados, e que alguns estudos sugerem que estas bases de dados devam ter mais de 10 horas para que se possa utilizar em uma aplicação cotidiana.

Outro estudo similar, de (El Ayadi; Kamel; Karray, 2011), afirma que a maioria das bases de dados encontradas nas literaturas possuem uma taxa de reconhecimento baixa até para humanos, o que torna difícil utilizar em sistemas computacionais para classificação.

O desafio principal consiste em criar um corpus linguístico contendo emoções, tão grande quanto os utilizados em reconhecimento de fala, permitindo assim uma boa generalização (Zhang et al., 2015).

A.1.1 Formulação da Pergunta

Questão Principal: Analisar as bases de dados já construídas para o reconhecimento de emoções.

Questões de Qualidade e Amplitude:

1. Questões a serem respondidas:

    a) Quais são os métodos e técnicas para coleta e avaliação das bases de dados de voz contendo emoções?

    b) Qual o tamanho e qualidade dessas bases?

2. Intervenção: métodos e técnicas de contrução das bases de dados.

3. Controle: Artigos previamente coletados por uma pesquisa exploratória não sistemática.
4. População: estudos com coleta de dados de sinais de voz contendo emoções.
5. Resultados: Visão de como são feitas as coletas e quais são os métodos e técnicas utilizados.
6. Aplicação: Para todos pesquisadores da área de estudos que envolvem reconhecimento de emoções por voz.

A.1.2 Seleção das fontes

1. Bases de dados a serem escolhidas para busca: IEEE, ACM, Springer, dblp;
2. Como foram selecionadas as bases: pelas informações dos artigos encontrados na pesquisa exploratória;
3. Idioma de busca: Inglês;
4. Palavras-Chave: emotion recognition, acoustics, sentiment analysis, database, emotional, speech, speech emotion recognition, emotional speech databases, emotion analysis;
5. Critério de Seleção: Trabalhos revisados de revistas e conferências;

A.2 SELEÇÃO DE ESTUDOS

A.2.1 Definição do Estudo

1. Critério de Inclusão:
    - Artigos que descrevem métodos e técnicas para coleta de dados de voz com emoções;
    - Artigos que contém as informações importantes sobre como foi feita essa coleta de forma clara;
2. Critério de Exclusão:
    - Trabalhos que não descrevem o método de construção da sua base de dados;
    - Bases de dados criadas por síntese de voz;
    - Artigos que descrevem resultados de reconhecimento de uma base de dados pronta;
    - Artigo que não contém de forma clara a construção de sua base de dados;
    - Artigos que descrevem a mesma base de dados (repetidos);

A.2.2 Definição da Estratégia de Seleção de Dados

Para cada base de dados serão utilizadas as palavras chaves selecionadas, testando diferentes strings, as quais não precisam necessariamente conter as palavras chaves definidas na seção de seleção das fontes, podendo ser descartadas ou incluídas caso se encontre outras palavras úteis.

Após ser feita a seleção dos artigos, o autor fez uma leitura destacando quais foram os métodos de coleta de dados e se houve algum teste estatístico para validação dos dados.

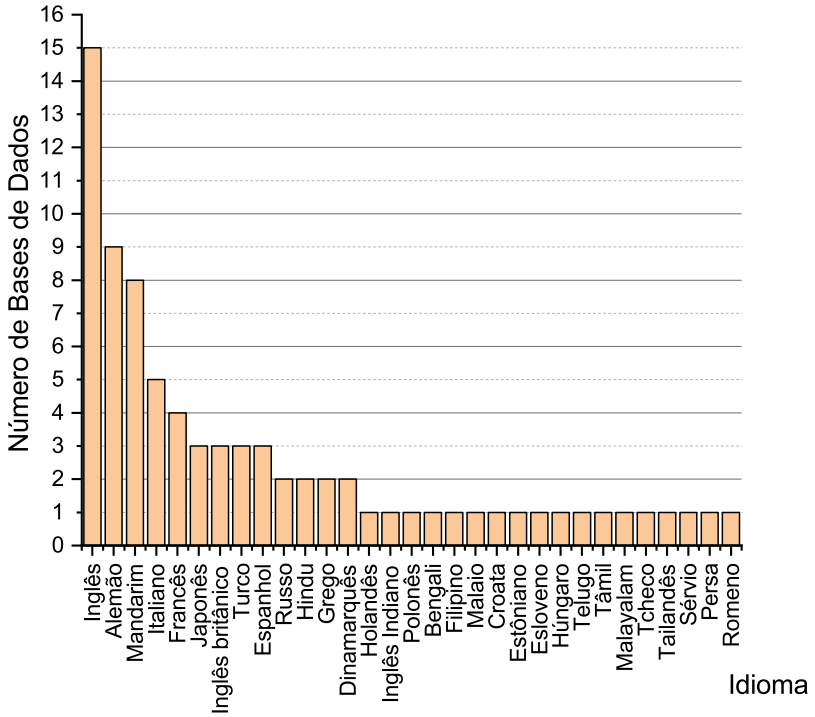
Informações que devem ser extraídas do método de coleta de dados:

- Idioma;
- Se as emoções foram atuadas, espontâneas ou induzidas;
- número de participantes;
- Houve algum questionário?
- Utilizou-se de alguma técnica para avaliar o estado emocional do participante?
- O total de dados (horas e/ou número de frases);
- As emoções, caso tenha sido catalogado de forma categórica;
- O gênero dos participantes;
- A faixa etária dos participantes;
- Se foi gravado em estúdio com equipamentos profissionais;
- Se foram divulgados para livre acesso estes dados;
- Análise e tratamento dos dados após a coleta;
- Qual o objetivo da criação dessa base de dados;

Tabela 15 – Tabela com as emoções mais utilizadas para construção das bases de dados

Emoção	Número de Bases
raiva	43
tristeza	40
felicidade	39
nojo	22
medo	25
surpresa	27

Figura 42 – Gráfico de número de bases de dados por idioma



Fonte: Elaborado pelo autor (2022).

A.3 REVISÃO BIBLIOGRÁFICA DE BASES DE DADOS AUDIOVISUAIS COM EMOÇÕES

Em (Anagnostopoulos; Iliou; Giannoukos, 2012), foi feita uma busca sobre trabalhos relacionados entre os anos de 2000 e 2011. Os autores concluíram que os estudos nessa área não possuem um banco de dados e sim conjunto de dados em pequena escala, também cita que não estão amplamente disponíveis. Para os autores é muito difícil a construção de um banco de dados utilizando fala natural e que estudos desse tipo são escassos e os que possuem maior número de dados são os obtidos através de programas de TV, rádio ou *callcenters*.

Foram coletados 389 artigos. Destes 109 foram selecionados após ter sido feita a leitura do seu resumo, o restante foi descartado por não se encaixarem no critério de seleção. Destes foi feita a leitura e coleta dos dados conforme o descrito no método de coleta de dados. Durante a leitura foram descartados por critérios de exclusão 41 artigos, restando portanto 68 artigos.

Escolheu-se os 13 artigos mais recentes para dispor em uma tabela, com partes de informações coletadas na revisão bibliográfica sistemática, nas Tabelas 16 e 17. Observou-se que estes estudos tem forma de coletas pouco úteis, como por exemplo o estudo do autor [2] (Darcy; Fontaine, 2020), onde além de conter apenas 2 locutores a base é construída por atores o que já foi criticado por outros autores como (Vidal et al., 2020). Outra informação interessante sobre estas bases é que o idioma mais utilizado é o inglês, aparecendo em 4 estudos. Um estudo que teve destaque foi a do autor [6], trata-se de uma base de dados audiovisual de múltiplos idiomas,

não atuada e que utilizou a teoria de fatores para classificar as emoções. Esta base de dados fez a sua coleta utilizando um *site*, onde duas pessoas se comunicavam de qualquer lugar que estivessem. Inicialmente respondiam um questionário de informações pessoais e posteriormente foram induzidas emoções com anúncios na tela. Após ler o anúncio os participantes conversavam sobre o que leram e respondiam um novo questionário (Kossaifi et al., 2019).

Tabela 16 – Informações dos 13 trabalhos mais recentes sobre criação de base de dados de voz com emoções

Autores [1]	Ano	Idioma	Emoções foram atuadas, espontâneas ou induzida	Núm. Locutores	Avaliou o estado emocional	Número de frases/audios	Emoções [2]
[1]	2020	Russo	Espontâneas	293 frases	Pelo contexto	293	Md, Nj, Ra, Su, Tr, Fe
[2]	2020	Inglês	Atuadas	2	sim, 96 ouvintes	1763 frases	Fe, Tr, Md, Ra, Nj
[3]	2020	Alemão	Induzidas	60	Respostas da escala SAM		In, So, No, Fa, Su, Fr
[4]	2019	Tailandês	Atuadas (filmes e séries)	51	sim, por 6 ouvintes	8987 frases	Ne, Fe, Tr, Ra, Ou
[5]	2020	Espanhol	Induzidas	174	sim, por 7 ouvintes	34,88 horas, 19793 frases	Ne, Su, Fe, Tr, Md, Ra, Nj, Ou
[6]	2019	Chinês, Inglês, Alemão, Grego, Húngaro, Sérvio	Induzidas	398	sim, avaliadores utilizaram ferramenta online	2000 minutos	três fatores: Va, Ag, GN

[1] [1](Alimuradov et al., 2020), [2](Darcy; Fontaine, 2020), [3](Kasuriya et al., 2019), [4](Kasuriya et al., 2019), [5](Kim et al., 2020), [6](Kossaifi et al., 2019), [7](Lassalle et al., 2019), [8](Nezami; Lou; Karami, 2019), [9](Rambabu et al., 2020), [10](Sato et al., 2020), [11](Sultana et al., 2021), [12](Vidal et al., 2020), [13](Zhang et al., 2015)

[2] Ne: Neutro, Ra: Raiva, Fe: Felicidade, Nj: Nojo, Su: Surpresa, Tr: Tristeza, Md: Medo, Ou: Outras, Va: Valência, Ag: Agitação, GN: Gosto/Não-gosto, Ve: Vergonha, Td: Tédio, De: Desapontado, En: Enjoado, An: Animado, Fu: Frustração, Ma: Magoa, In: Interesse, Ci: Ciumes, Br: Brincando, Ge: Gentil, Or: Orgulho, So: Sorrateiro, Ho: Hostilidade, Pr: Preocupação, Sr: Sarcasmo, Rx: Relaxado

Tabela 17 – continuação da Tabela 16 - Informações dos 13 trabalhos mais recentes sobre criação de base de dados de voz com emoções

Autores [1]	Ano	Idioma	Emoções foram atuadas, espontâneas ou induzida	Núm. Locutores	Avaliou o estado emocional	Número de frases/audios	Emoções [2]
[7]	2019	Inglês Britânico, Suíço e Hebraico	Atuadas	54	sim	2159 frases	Md, Ra, Ve, Te, De, En, Na, Fu, Fe, Ma, In, CI, Ge, Br, Or, So, Su, Ho, Pr
[8]	2019	Persa	Induzidas	87	sim, 12 ouvintes	3 horas e 25 min	Ra, Md, Fe, Tr, Su, Ne
[9]	2020	Hindu	atuadas (parte por atores e parte por não atores	38	sim, por 18 ouvintes	5317 frases	Ra Fe,Tr, Ne, Su, Md, Nj, Sr, Fu, Rx, Pr, Tm, Ag
[10]	2020	Japonês	atuadas		sim, pelos participantes que fizeram a coleta da base	2025 frases	An, Al, Co, Md, Su, Tr, Nj, Ra
[11]	2021	Bengali	atuadas	20	não	7000 frases,7 horas	Ra, Nj, Md, Fe, Ne, Tr, Su
[12]	2020	Inglês	espontâneas vídeos da internet	297	sim por ouvintes com um questionário com escala SAM e emoções primárias	9,370 frases, 24 hrs e 41 min	Ra, Tr, Fe, Su, Md, Nj, Des, Ne, Ou
[13]	2020	Mandarim	induzidas por diálogo	35	sim por 6 ouvintes	2480 vídeos	Ra, Nj, Md, Fe, Su, Tr, Ne