



Leonardo Marçal da Silva

Avaliação de Características no Reconhecimento de Emoções Através da Voz

**João Pessoa – PB
Abril de 2026**



Leonardo Marçal da Silva

Avaliação de Características no Reconhecimento de Emoções Através da Voz

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Engenharia Elétrica do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, como requisito necessário à obtenção do grau de Mestre em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Telecomunicações

Linha de Pesquisa: Processamento de Sinais

Suzete Élide Nóbrega Correia, Prof.^a Dra.

Orientação

Silvana Luciene do Nascimento Cunha Costa, Prof.^a Dra.

Coorientação

João Pessoa – PB, Abril de 2026

© Leonardo Marçal da Silva — marcal.leonardo@academico.ifpb.edu.br

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha do IFPB, *Campus* João Pessoa

S586a Silva, Leonardo Marçal da.

Avaliação de características no reconhecimento de emoções
através da voz / Leonardo Marçal da Silva. – 2026.

60 f. : il.

Dissertação (Mestrado – Engenharia Elétrica) – Instituto
Federal de Educação da Paraíba / Programa de Pós-Graduação
em Engenharia Elétrica (PPGEE), 2026.

Orientação : Prof^a Dra. Suzete Élide Nóbrega Correia.

Coorientação : Prof^a Dra. Silvana Luciene do N. C. Costa.

1.Reconhecimento de emoções. 2. Processamento de sinais
de voz. 3.Características acústicas. 4. Máquinas de vetor de
suporte. 5. Coeficiente de Fischer. I. Título.

CDU 004.934:159.942(043)

Bibliotecária responsável: Lucrecia Camilo de Lima – CRB 15/132

LEONARDO MARÇAL DA SILVA

AVALIAÇÃO DE CARACTERÍSTICAS NO RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA VOZ

DISSERTAÇÃO submetida à Coordenação do Curso Superior de Engenharia Elétrica, do Instituto Federal da Paraíba (IFPB), linha de pesquisa: “Processamento de Sinais”, em cumprimento aos requisitos institucionais para a obtenção do Título de **MESTRE EM ENGENHARIA ELÉTRICA**.

Aprovado em 10 de abril de 2026.

Membros da Banca Examinadora:

Profa. Dra. Suzete Élide Nóbrega Correia

Instituto Federal da Paraíba (IFPB)
Orientadora

Profa. Dra. Silvana Luciene do Nascimento Cunha Costa

Instituto Federal da Paraíba (IFPB)
Coorientadora

Prof. Dr. Carlos Danilo Miranda Regis

Instituto Federal da Paraíba (IFPB)
Examinador interno

Prof. Dr. Thyago Leite de Vasconcelos Lima

Instituto Federal da Paraíba (IFPB)
Examinador interno

Documento assinado eletronicamente por:

- **Suzete Elida Nobrega Correia**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 13/04/2026 13:45:00.
- **Silvana Luciene do Nascimento Cunha Costa**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 13/04/2026 13:53:36.
- **Carlos Danilo Miranda Regis**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 13/04/2026 14:22:46.
- **Thyago Leite de Vasconcelos Lima**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 15/06/2026 07:35:38.

Este documento foi emitido pelo SUAP em 13/04/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifpb.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código 865067
Verificador: 5df6eaa5be
Código de Autenticação:



AGRADECIMENTOS

A Deus, pela força, saúde e sabedoria concedidas em todos os momentos desta caminhada.

À minha família, que sempre acreditou em mim e esteve ao meu lado com amor, apoio e compreensão, especialmente nos períodos mais desafiadores deste percurso.

Aos meus amigos, pela motivação constante, pelas conversas que aliviaram a rotina e por celebrarem comigo cada pequena conquista.

Expresso minha profunda gratidão à minha orientadora, Prof.^a Dra. Suzete Élide Nóbrega Correia, pela orientação atenta, paciência, dedicação e rigor científico que contribuíram decisivamente para a construção deste trabalho. Agradeço também à minha coorientadora, Prof.^a Dra. Silvana Luciene do Nascimento Cunha Costa, pelas contribuições valiosas, incentivo e apoio ao longo do desenvolvimento desta pesquisa.

A todos os professores e servidores do Programa de Pós-Graduação em Engenharia Elétrica do IFPB, que contribuíram com conhecimento, estrutura e oportunidades de crescimento acadêmico.

Por fim, agradeço a todos que, direta ou indiretamente, fizeram parte desta jornada e contribuíram para que este trabalho se tornasse possível.

RESUMO

O reconhecimento automático de emoções na fala (SER) tem ganhado destaque em diversas áreas, incluindo saúde, interação humano-máquina e sistemas inteligentes. Este trabalho investiga a contribuição individual e conjunta de características acústicas para a classificação de emoções, utilizando a base de dados RAVDESS e um classificador baseado em Máquinas de Vetores de Suporte (SVM). Foram extraídas 21 características acústicas (medidas prosódicas, espectrais e de qualidade vocal) de amostras vocais classificadas em oito estados emocionais: neutro, calmo, feliz, triste, bravo, assustado, enojado e surpreso. O protocolo experimental empregou Validação Cruzada Estratificada de 5 partições e avaliou o desempenho em diferentes níveis de representação: *frame-level*, *utterance-level* e um subconjunto otimizado de variáveis (*Top-10*) ranqueado pelo método *Fisher Score*. A análise revelou que os atributos prosódicos e de qualidade vocal, com destaque para a energia (RMSE), o *shimmer* e a frequência fundamental (F_0), são os marcadores mais consistentes na detecção das emoções. O classificador operando com o conjunto reduzido (*Top-10*) obteve o melhor desempenho do estudo, alcançando uma acurácia de 44,03%. Embora limitado pela confusão entre emoções de mesma ativação (valência), o resultado demonstra que a seleção de atributos acústicos permite a construção de modelos preditivos mais enxutos.

Palavras-chave: Reconhecimento de emoções; processamento de sinais de voz; características acústicas; Máquinas de Vetores de Suporte; Coeficiente de Fisher.

ABSTRACT

Automatic speech emotion recognition (SER) has gained prominence in various fields, including healthcare, human-machine interaction, and intelligent systems. This study investigates the individual and combined contributions of acoustic features to emotion classification, using the RAVDESS dataset and a Support Vector Machine (SVM)-based classifier. Twenty-one acoustic features (prosodic, spectral, and vocal quality measures) were extracted from speech samples classified into eight emotional states: neutral, calm, happy, sad, angry, scared, disgusted, and surprised. The experimental protocol employed 5-fold stratified cross-validation and evaluated performance at different levels of representation: *frame-level*, *utterance-level*, and an optimized subset of variables (*Top-10*) ranked by the *Fisher Score* method. The analysis revealed that prosodic and vocal quality attributes, particularly energy (RMSE), *shimmer*, and fundamental frequency (F_0), are the most consistent markers for emotion detection. The classifier operating with the reduced set (*Top-10*) achieved the best performance in the study, reaching an accuracy of 44.03%. Although limited by confusion between emotions of the same valence, the result demonstrates that the selection of acoustic features allows for the construction of leaner predictive models.

Keywords: Emotion recognition; speech signal processing; acoustic features; Support Vector Machines; Fisher Score.

LISTA DE ILUSTRAÇÕES

Figura 1 – Fluxo metodológico adotado para o reconhecimento de emoções na fala. . .	35
Figura 2 – Distribuição da frequência fundamental (F_0) por emoção.	42
Figura 3 – Distribuição da energia (RMSE) por emoção.	43
Figura 4 – Distribuição da taxa de cruzamento por zero (TCZ) por emoção.	43
Figura 5 – Distribuição do centroide espectral por emoção.	44
Figura 6 – Distribuição dos coeficientes MFCC ₁ a MFCC ₆ por emoção.	45
Figura 7 – Distribuição dos coeficientes MFCC ₇ a MFCC ₁₃ por emoção.	46
Figura 8 – Distribuição do Jitter (%) por emoção.	47
Figura 9 – Distribuição do Shimmer (APQ5) por emoção.	48
Figura 10 – Distribuição da razão harmônico-ruído (HNR) por emoção.	48
Figura 11 – Distribuição da proeminência do pico cepstral suavizada (CPPS) por emoção. . .	49
Figura 12 – Matriz de confusão, cenário <i>frame-level</i>	52
Figura 13 – Matriz de confusão, cenário <i>utterance-level</i>	53
Figura 14 – Matriz de confusão, cenário <i>Top-10</i>	53

LISTA DE TABELAS

Tabela 1 – Resumo detalhado dos estudos sobre reconhecimento de emoções na fala . .	14
Tabela 2 – Distribuição das amostras de fala na base RAVDESS utilizada neste trabalho.	20
Tabela 3 – Métricas médias (5-fold) dos três cenários SVM.	51

SUMÁRIO

1	INTRODUÇÃO	8
1.1	Objetivos	17
1.1.1	Objetivo Geral	17
1.1.2	Objetivos Específicos	17
2	FUNDAMENTAÇÃO TEÓRICA	18
2.1	Reconhecimento de Emoções na Fala	18
2.2	Bases de Dados para Reconhecimento Emocional na Fala	19
2.3	Características Acústicas da Voz	20
2.3.1	Características Prosódicas	21
2.3.2	Características de Qualidade Vocal	24
2.3.3	Características Espectrais	29
2.4	Máquinas de Vetores de Suporte (SVM)	32
3	METODOLOGIA	35
3.1	Pré-processamento	35
3.2	Extração das Características Acústicas	36
3.3	Seleção de Características	37
3.4	Classificação com SVM	38
3.5	Avaliação de Desempenho	39
4	RESULTADOS E DISCUSSÃO	42
4.1	Características Prosódicas	42
4.2	Características Espectrais e Cepstrais	44
4.3	Características de Qualidade Vocal	47
4.4	Seleção de características: o conjunto top-10 (<i>Fisher Score</i>)	49
4.5	Desempenho do Classificador SVM	51
4.5.1	Resumo das Métricas Globais	51
4.5.2	Matrizes de Confusão	52
4.6	Discussão Geral	54
5	CONSIDERAÇÕES FINAIS	56
	REFERÊNCIAS	58

1. INTRODUÇÃO

O conceito de emoção abrange diversas áreas de estudo, incluindo filosofia, biologia e, principalmente, psicologia. Existem várias teorias sobre como as emoções se manifestam em um ser humano, sendo uma das mais conhecidas a teoria de James-Lange. Essa teoria sugere que, ao receber um estímulo do ambiente, o ser humano experimenta alterações fisiológicas, como palpitações, sudorese e falta de ar, as quais geram uma emoção. Em contraste, a teoria de Cannon-Bard propõe que as reações fisiológicas e a experiência emocional ocorrem simultaneamente. As emoções estão integradas a respostas adaptativas do sistema nervoso autônomo e somático, resultando em modificações no funcionamento do sistema de produção de fala. Estudos indicam que, para cada emoção, há alterações na respiração, fonação e articulação, o que provoca mudanças no padrão do sinal acústico (JOHNSTONE; SCHERER, 2000).

O reconhecimento automático das emoções humanas através da fala pode ser utilizado em diversas áreas, como no gerenciamento de estresse e ansiedade, na adaptação de intervenções terapêuticas e no monitoramento da saúde emocional. Essa tecnologia também auxilia na detecção de sinais de solidão, depressão e outros distúrbios mentais (AKÇAY; OĞUZ, 2020; WANI et al., 2021; LOPE; GRAÑA, 2023; DHUHEIR et al., 2021).

Para que o reconhecimento emocional seja realizado por algoritmos computacionais, são necessários parâmetros acústicos específicos da voz humana. Akçay e Oğuz (2020) propõem quatro grupos de medidas acústicas para caracterizar as emoções:

- **Medidas associadas à prosódia:** Relacionadas à comunicação não verbal, como ritmo e entonação. São representadas por indicadores acústicos como frequência fundamental, energia e duração dos momentos com e sem voz.
- **Medidas espectrais:** Baseadas no espectro de frequências dos sinais de voz, como coeficientes mel-cepstrais e coeficientes cepstrais de predição linear.
- **Medidas de qualidade vocal:** Fornecem informações sobre as propriedades físicas do trato vocal, como *jitter*, *shimmer* e razão Harmônico-Ruído (HNR, do inglês *Harmonic-to-Noise Ratio*).
- **Medidas baseadas no operador de energia de Teager:** Representam as não linearidades do sistema de produção vocal.

Chowdhury, Ramanna e Kotecha (2025) desenvolveram um modelo de *ensemble* leve e eficiente para reconhecimento de emoções na fala, combinando arquiteturas de redes neurais convolucionais tradicionais (CNN) com variantes bidirecionais integradas a redes recorrentes do tipo LSTM (do inglês *Long Short-Term Memory*) (CNN_{BiLSTM}). O foco principal do estudo

foi alcançar elevada acurácia com custo computacional reduzido, tornando o modelo adequado para aplicações embarcadas ou em tempo real. Para isso, os autores utilizaram cinco bases de dados públicas amplamente empregadas na literatura, incluindo a RAVDESS, permitindo uma análise comparativa em diferentes contextos e variações de idiomas, atores e protocolos de gravação. A extração de características foi realizada manualmente a partir dos sinais de voz, com ênfase em atributos espectrais e prosódicos como coeficientes MFCC, *Chroma* STFT (transformada de Fourier de curto tempo, do inglês *Short-Time Fourier Transform*), *Root Mean Square Energy* (RMSE) e taxa de cruzamento por zero (ZCR), que foram utilizados como entrada para as redes neurais após normalização e formatação temporal. O modelo de ensemble foi construído com múltiplos caminhos de processamento convolucional e recorrente, cujas saídas foram fundidas por meio de camadas densas, possibilitando o aproveitamento complementar de diferentes representações do sinal. Os resultados demonstraram que o ensemble superou de forma consistente os modelos individuais, tanto os baseados unicamente em CNN quanto os híbridos isolados, apresentando desempenho superior nas métricas de acurácia, AUC-ROC, AUC-PRC e *F1-score* em todos os conjuntos testados, inclusive na base RAVDESS. Na RAVDESS, o modelo atingiu estabilidade mesmo diante da variação inter-locutor e da diferença de intensidade emocional, evidenciando sua capacidade de generalização. O estudo destacou ainda que a incorporação simultânea de múltiplos descritores acústicos permitiu capturar nuances emocionais mais complexas, mesmo em sinais curtos ou com artefatos de gravação. Como conclusão, os autores reforçaram que modelos enxutos baseados em ensemble, quando bem estruturados e alimentados com atributos discriminativos, podem alcançar desempenho comparável ao de arquiteturas mais profundas, sem comprometer a eficiência computacional.

Colunga-Rodriguez et al. (2025) apresentaram um estudo focado na construção de um conjunto de dados secundário derivado da base RAVDESS, com o objetivo de disponibilizar características acústicas prontas para uso em tarefas de reconhecimento de emoções por meio de técnicas de aprendizado de máquina. O principal enfoque do trabalho esteve na definição de um fluxo de pré-processamento acústico rigoroso e reprodutível, englobando etapas como remoção de silêncio, normalização de amplitude, padronização da taxa de amostragem e segmentação temporal das amostras. Com base nesse fluxo, os autores extraíram um conjunto de atributos acústicos clássicos, entre eles coeficientes MFCC, *pitch*, energia, *jitter*, *shimmer* e taxa de cruzamento por zero (ZCR), cobrindo diferentes domínios do sinal, como espectral, prosódico e de qualidade vocal. Embora o artigo não tenha implementado ou avaliado diretamente classificadores, o foco foi a análise descritiva e comparativa da distribuição dessas características por emoção, fornecendo uma base sólida para futuras investigações. A estrutura do conjunto gerado foi organizada em formato tabular, com cada linha representando uma amostra vocal e suas respectivas características agregadas estatisticamente (como média, desvio padrão e curtose), permitindo sua aplicação direta em modelos supervisionados. O repositório resultante foi disponibilizado publicamente em acesso aberto, com documentação completa, facilitando a

reutilização por parte da comunidade científica interessada em tarefas de classificação emocional. Os autores ressaltaram que uma cadeia de processamento bem definida, desde a limpeza do sinal até a formatação dos dados, tem impacto direto na acurácia e estabilidade de qualquer modelo posterior, especialmente quando se trabalha com bases estruturadas como a RAVDESS. Assim, o estudo contribui significativamente ao fornecer uma base de dados pronta para uso e metodologicamente transparente, promovendo reprodutibilidade e padronização em estudos futuros de reconhecimento emocional.

Dar e Delhibabu (2024) realizaram uma revisão abrangente da literatura sobre reconhecimento automático de emoções na fala, com foco nos três pilares fundamentais dos sistemas de SER: bases de dados, características acústicas e classificadores. O estudo analisou um conjunto de trabalhos recentes, abrangendo bases de dados atuadas e naturais amplamente utilizadas, como EMO-DB, IEMOCAP, RAVDESS, SAVEE, TESS e MSP-Podcast, contemplando tanto a classificação de emoções discretas quanto a predição de dimensões emocionais contínuas, como *arousal* e *valence*. No que se refere às características acústicas, os autores identificaram que descritores espectrais, em especial os coeficientes cepstrais em frequência Mel (MFCC), aliados a parâmetros prosódicos como frequência fundamental (F_0), energia e duração, permanecem como os atributos mais recorrentes e eficazes na literatura. A revisão também destacou o papel complementar de características de qualidade vocal, como *jitter*, *shimmer* e entropia espectral, especialmente em cenários com maior variabilidade interlocutor e ruído ambiental. Em relação aos classificadores, o estudo evidenciou uma transição gradual de métodos clássicos, como SVM, GMM e *Random Forest*, para abordagens baseadas em aprendizado profundo, incluindo CNNs, LSTMs e arquiteturas híbridas, que apresentam desempenho superior quando grandes volumes de dados estão disponíveis. Contudo, os autores ressaltaram que modelos tradicionais continuam competitivos em contextos com bases limitadas. Como conclusão, a revisão enfatizou que a combinação criteriosa de características acústicas, aliada à seleção adequada de modelos e, quando possível, à multimodalidade, constitui o caminho mais promissor para o avanço de sistemas robustos e generalizáveis de reconhecimento de emoções na fala.

Mashhadi e Osei-Bonsu (2023) realizaram uma investigação sistemática sobre reconhecimento de emoções na fala utilizando a base RAVDESS, comparando dois paradigmas distintos de classificação: o algoritmo tradicional *Random Forest* e uma rede neural convolucional unidimensional (Conv1D). Para alimentar os modelos, os autores extraíram um conjunto diversificado de características acústicas, incluindo coeficientes MFCC, cromograma, contraste espectral, taxa de cruzamento por zero (ZCR), espectrograma em escala Mel e representação Tonnetz, cobrindo tanto aspectos espectrais quanto harmônicos do sinal de voz. Após a normalização dos dados, foi aplicada uma etapa de seleção de características baseada na importância das variáveis estimada pelo próprio *Random Forest*, com o objetivo de reduzir a dimensionalidade e otimizar o desempenho preditivo. A avaliação foi conduzida por meio de validação cruzada, utilizando métricas como acurácia, precisão, *recall* e *F1-score*. Os resultados revelaram que, em contextos

com conjuntos de dados limitados — como é o caso da RAVDESS —, o *Random Forest* obteve desempenho superior, com acurácia média de 69%, destacando-se especialmente nas emoções “medo” (com 72% de precisão) e “calma” (com 84% de *recall*). Os autores argumentaram que modelos de aprendizado profundo como Conv1D, embora promissores em grandes volumes de dados, tendem a apresentar *overfitting* ou baixa generalização quando aplicados a corpora de menor escala, exigindo ajustes finos e maior custo computacional. Já o *Random Forest* demonstrou maior robustez, interpretabilidade e capacidade de generalização, sendo eficaz mesmo sem necessidade de ajuste extensivo de hiperparâmetros. O estudo concluiu que, em aplicações práticas de reconhecimento emocional, especialmente em ambientes computacionalmente restritos ou com dados moderados, métodos baseados em árvores de decisão continuam oferecendo desempenho competitivo, sobretudo quando aliados a uma engenharia de atributos bem estruturada.

Leem et al. (2022) investigam a robustez de diferentes características acústicas para reconhecimento automático de emoções na fala em ambientes com ruído, partindo da hipótese de que nem todos os descritores acústicos contribuem igualmente para o desempenho de sistemas SER sob condições adversas. O estudo focou nos *Low-Level Descriptors* (LLDs) comuns em conjuntos de características acústicas e avaliou individualmente cada um deles, treinando modelos de reconhecimento em fala limpa e testando-os em versões contaminadas por ruído. A proposta central consistiu em selecionar subconjuntos de características mais robustas ao ruído, por meio de critérios de desempenho absoluto e de estabilidade relativa entre condições limpa e ruidosa, em vez de empregar o conjunto completo de descritores sem discriminação. Os experimentos, realizados com versões limpa e com diferentes níveis de ruído da base de dados *MSP-Podcast*, demonstraram que um modelo treinado com apenas características robustas apresentava desempenho superior em condições ruidosas, sem necessidade de técnicas de compensação adicionais, como aumento de dados ou módulos de melhoria de sinal. Os resultados evidenciaram melhorias significativas nas tarefas de predição de atributos emocionais (*arousal*, *valence* e *dominance*) quando comparados ao uso de todos os LLDs, reforçando que a seleção criteriosa de características acústicas resilientes a ruído pode aumentar a eficácia de sistemas SER em aplicações do mundo real.

Yan et al. (2020) apresentaram um sistema interativo de reconhecimento automático de emoções na fala baseado no classificador *Random Forest*, com aplicação prática implementada em um programa da plataforma *WeChat*. O estudo utilizou a base de dados *Berlin Speech Emotion Database* (EMO-DB), composta por sete classes emocionais, a partir da qual foram extraídas características acústicas por meio da ferramenta *openSMILE*. Foram consideradas 16 características acústicas, incluindo coeficientes cepstrais em frequência Mel (MFCC), energia RMS, taxa de cruzamento por zero, probabilidade de vozeamento e frequência fundamental, às quais se aplicaram 12 funções estatísticas, resultando em um vetor final de 384 atributos por amostra. Os autores compararam diferentes classificadores, como regressão logística, SVM,

XGBoost e *Random Forest*, verificando que este último apresentou o melhor desempenho, alcançando acurácia média de até 89%. Os resultados indicaram maior robustez do *Random Forest* frente a ruído, além de melhor separabilidade entre classes emocionais. O trabalho demonstrou a viabilidade de sistemas de SER em tempo quase real e reforçou a eficácia do uso combinado de descritores acústicos e métodos de aprendizado de máquina para aplicações em interação homem-máquina.

Akçay e Oğuz (2020) realizaram uma revisão sistemática sobre os principais elementos que compõem sistemas de reconhecimento de emoções na fala, com foco especial nas características acústicas extraídas do sinal vocal. Os autores categorizaram essas características em três grupos principais: (i) características prosódicas, que incluem frequência fundamental (F_0), duração da fala, intensidade e ritmo, fortemente associadas à entonação e ao padrão melódico do discurso; (ii) características espectrais, que englobam os coeficientes cepstrais em frequência Mel (MFCC), *Linear Predictive Cepstral Coefficients* (LPCC) e os filtros de banda crítica, utilizados para representar a estrutura de frequência da fala; e (iii) características de qualidade vocal, como *jitter*, *shimmer* e razão harmônico-ruído (HNR), que fornecem informações sobre estabilidade vibratória e regularidade do sinal glótico. A revisão também abordou o impacto de fatores como ruído, variação inter-locutor e tipo de corpus na robustez dos sistemas SER, além de discutir técnicas de normalização e seleção de atributos. Um ponto central da análise foi a recomendação da combinação entre diferentes tipos de características — especialmente prosódicas e espectrais — para capturar as nuances emocionais com maior fidelidade. Os autores enfatizaram ainda que a escolha das características deve considerar a natureza da aplicação (*real-time* ou *offline*) e a qualidade do ambiente de coleta. Ao sintetizar resultados de dezenas de estudos experimentais, Akçay e Oğuz forneceram um referencial teórico sólido para a construção de fluxos de processamento de reconhecimento emocional, apontando que abordagens híbridas tendem a oferecer melhor desempenho em cenários reais e baseando-se em classificadores como SVM, *Random Forest* e redes neurais rasas.

Tarnowski et al. (2017) analisaram o desempenho de diferentes técnicas de reconhecimento automático de emoções na fala a partir da extração de características acústicas do sinal de voz. O estudo concentrou-se na utilização de descritores prosódicos e espectrais, incluindo coeficientes cepstrais em frequência Mel (MFCC), frequência fundamental, energia e parâmetros temporais, extraídos em janelas de curto prazo. As características obtidas foram organizadas em vetores estatísticos para representar o comportamento global do sinal emocional. Para a etapa de classificação, foram avaliados algoritmos de aprendizado de máquina supervisionado, com ênfase em máquinas de vetores de suporte (SVM) e classificadores baseados em distância, aplicados a bases de dados de fala emocional atuadamente utilizadas na literatura. Os resultados experimentais indicaram que os MFCCs apresentaram elevado poder discriminativo para a identificação de estados emocionais, sobretudo quando combinados com características prosódicas. O estudo concluiu que a integração de múltiplos descritores acústicos contribui

significativamente para o aumento da acurácia do sistema, reforçando a relevância da seleção adequada de características na construção de sistemas eficientes de *Speech Emotion Recognition* (SER).

Schuller, Steidl e Batliner (2009) apresentaram um sistema de reconhecimento de emoções na fala desenvolvido no contexto do *Interspeech Emotion Challenge*, uma iniciativa internacional voltada à padronização de tarefas, métricas e protocolos de avaliação em sistemas de SER. O estudo focou na construção de uma base robusta de características acústicas, abrangendo medidas espectrais, prosódicas e articulatórias, incluindo coeficientes cepstrais em frequência Mel (MFCC), formantes (F_1 a F_3), energia média, frequência fundamental (F_0), entropia espectral, taxa de cruzamento por zero e medidas de dinâmica temporal como deltas e delta-deltas. O total de características ultrapassava mil atributos por amostra, o que exigiu a aplicação de técnicas avançadas de seleção e redução de dimensionalidade, como ganho de informação, análise de componentes principais (PCA) e correlação cruzada. Os classificadores utilizados foram baseados em máquinas de vetores de suporte (SVM) com diferentes funções kernel (linear e radial), com validação cruzada estratificada para garantir a generalização dos resultados. Os experimentos foram realizados em diferentes corpora, incluindo o corpus de treino disponibilizado pela competição e bases externas para verificação cruzada. Os resultados mostraram que o uso de conjuntos de características altamente diversificados, aliado a estratégias cuidadosas de seleção, aumentou significativamente a acurácia da classificação, reduzindo a sobreposição entre classes emocionais semelhantes. Além disso, os autores observaram que a inclusão de atributos derivados da dinâmica temporal (como a variação ao longo do tempo de MFCCs e energia) foi determinante para diferenciar emoções com padrões acústicos similares, como tristeza e neutralidade. O estudo reforçou a premissa de que sistemas SER eficazes devem ir além da extração isolada de parâmetros e considerar o comportamento combinado e contextualizado dos atributos.

Busso et al. (2008) introduziram a base de dados IEMOCAP (*Interactive Emotional Dyadic Motion Capture*), uma das corpora mais amplamente utilizados no desenvolvimento e validação de sistemas de reconhecimento de emoções na fala (SER). Essa base foi concebida com o objetivo de fornecer gravações de interações reais e semiestruturadas entre atores profissionais, contendo manifestações emocionais tanto atuadas quanto espontâneas. O corpus é multimodal e inclui não apenas o áudio, mas também dados de vídeo, movimento corporal e transcrições textuais, embora o foco deste estudo tenha sido o canal acústico. Em termos de extração de atributos, os autores aplicaram uma ampla gama de características, entre elas a frequência fundamental (*pitch*), energia do sinal, coeficientes MFCC, duração média das sílabas e medidas de variação temporal. A diversidade e riqueza das expressões emocionais presentes no IEMOCAP permitiram a avaliação detalhada da capacidade discriminativa de cada característica. O estudo testou diferentes algoritmos de aprendizado de máquina, incluindo SVM, GMM (do inglês, *Gaussian Mixture Model*) e classificadores baseados em árvores, para avaliar a eficácia dessas

medidas na distinção entre emoções como raiva, alegria, tristeza, frustração e neutralidade. Os resultados indicaram que características como *pitch* e energia foram particularmente úteis na diferenciação de emoções de alta excitação, como raiva e alegria, enquanto os MFCCs desempenharam um papel mais relevante na identificação de emoções mais sutis, como tristeza e desânimo. Além disso, os autores destacaram a importância da combinação de atributos estáticos e dinâmicos — incluindo derivadas temporais — para capturar a natureza transiente das emoções na fala. A contribuição fundamental deste estudo foi não apenas a análise empírica das características, mas também a disponibilização pública da base IEMOCAP, que desde então se tornou um padrão de referência na área de SER.

A Tabela 1, sintetiza de forma detalhada os estudos analisados neste capítulo, incluindo as bases de dados utilizadas, as emoções avaliadas, as medidas acústicas extraídas e os resultados quantitativos reportados pelos autores. Essa consolidação permite comparar metodologias e desempenhos de diferentes abordagens no reconhecimento de emoções na fala, fornecendo um panorama objetivo do estado da arte.

Tabela 1 – Resumo detalhado dos estudos sobre reconhecimento de emoções na fala

Autor (Ano)	Base de Da- dos	Emoções Avaliadas	Medidas Acústicas Uti- lizadas	Resultados Reporta- dos
Chowdhury et al. (2025)	RAVDESS + 4 bases	Alegria, Raiva, Tris- teza, Medo, Calma, Neutro	MFCC, Chroma STFT, RMSE, ZCR	Ensemble CNN+BiLSTM: ACC=92.4%, F1=0.91, AUC-ROC=0.95; ro- bustez a variação de intensidade.
Colunga- Rodriguez et al. (2025)	RAVDESS derivada	Alegria, Raiva, Tris- teza, Medo, Calma, Surpresa, Neutro	MFCC, pitch, jitter, shimmer, energia, ZCR	Fornecer dataset proces- sado; sem avaliação de modelo (ACC não infor- mado).

Continua na próxima página

Tabela 1 – Continuação

Autor (Ano)	Base de Da- dos	Emoções Avaliadas	Medidas Acústicas Uti- lizadas	Resultados	Reporta- dos
Mohmad Dar & Delhibabu (2024)	Revisão sistemática (múltiplas bases: EMO-DB, IEMOCAP, SAVEE, RAVDESS, TESS, MSP-Podcast, entre outras)	Alegria, Raiva, Tristeza, Medo, Surpresa, Nojo, Neutro	MFCC, F_0 , energia, duração, entropia espectral, jitter, shimmer, características linguísticas e multimodais	Revisão aponta que abordagens híbridas e profundas (CNN, LSTM) superam métodos clássicos; seleção de características aumenta robustez.	
Mashhadi & Osei-Bonsu (2023)	RAVDESS	Alegria, Raiva, Tristeza, Medo, Calma, Nojo, Surpresa, Neutro	MFCC, cromograma, ZCR, Tonnetz	Random Forest: ACC=69%, CNN-Conv1D: ACC=63%; RF superior em dados limitados.	
Leem et al. (2022)	MSP-Podcast (limpo + ruído)	Arousal, Valence, Dominance	LLDs acústicos (MFCC, energia, F_0 , jitter, shimmer, entropia espectral)	Seleção de características robustas melhorou desempenho em ruído; subconjuntos superaram conjuntos completos sem aumento de dados.	
Yan et al. (2020)	EMO-DB	Alegria, Raiva, Tristeza, Medo, Nojo, Tédio, Neutro	MFCC, RMS, ZCR, probabilidade de vozeamento, F_0 (384 atributos estatísticos)	Random Forest: ACC até 89%; superou SVM, XGBoost e regressão logística; maior robustez a ruído.	

Continua na próxima página

Tabela 1 – Continuação

Autor (Ano)	Base de Da- dos	Emoções Avaliadas	Medidas Acústicas Uti- lizadas	Resultados	Reporta- dos
Akçay & Oğuz (2019)	Revisão sistemática (múltiplas bases)	Alegria, Raiva, Tristeza, Medo, Sur- presa, Nojo, Neutro	F_0 , duração, intensidade, MFCC, LPCC, jitter, shimmer, HNR	Revisão ACC variando entre 70–95%, destacando combinações prosódi- cas+espectrais como mais eficazes.	reporta
Tarnowski et al. (2017)	Corpus atuado (fala emocional)	Alegria, Raiva, Tris- teza, Medo, Neutro	MFCC, F_0 , energia, pa- râmetros temporais	MFCC maior poder discrimi- nativo; combinação prosódica + espectral elevou a acurácia global.	
Schuller et al. (2009)	Interspeech Emotion Challenge	Alegria, Raiva, Tris- teza, Tédio, Neutro	MFCC, energia, entro- pia, formantes, deltas	SVM-RBF: ACC=81%, F1=0.84; PCA reduziu dimensionalidade man- tendo desempenho.	
Busso et al. (2008)	IEMOCAP	Raiva, Ale- gria, Tris- teza, Frus- tração, Neu- tro	F_0 , energia, MFCC, du- ração silábica, deltas	SVM: ACC=71%, F1=0.74; Pitch e energia mais discriminativos para emoções de alta excitação.	

Diante desse cenário, a compreensão aprofundada das características acústicas que realmente contribuem para a separação entre classes emocionais torna-se essencial. Avaliar de forma quantitativa a capacidade discriminativa de cada atributo permite não apenas buscar melhorias na acurácia dos modelos, mas, sobretudo, fornecer transparência sobre como as emoções estão sendo inferidas. Essa abordagem é especialmente relevante em aplicações sensíveis, como diagnósticos clínicos, onde decisões automáticas precisam ser justificáveis. Além disso, a escolha criteriosa das características a serem extraídas pode reduzir significativamente o custo computacional, favorecer a interpretabilidade e melhorar a consistência da análise com dados limitados.

A presente dissertação alinha-se a esse objetivo, propondo uma investigação centrada na contribuição estatística de cada característica para o reconhecimento emocional, utilizando uma base de dados padronizada (RAVDESS) e um método de classificação cujos resultados

podem ser analisados e compreendidos, permitindo identificar como cada atributo influencia o processo de decisão do modelo. Ressalta-se que a investigação se limita ao cenário de fala atuada em ambiente controlado, o que favorece a extração precisa de atributos, embora restrinja a generalização imediata para contextos de fala espontânea e ruidosa. Dessa forma, busca-se construir uma metodologia de SER transparente, interpretável e cientificamente fundamentada.

É importante situar que o estado da arte no Reconhecimento de Emoções na Fala (SER) é atualmente dominado por arquiteturas de aprendizado profundo (Deep Learning), as quais alcançam níveis de acurácia elevados, frequentemente superiores a 80% ou 90%. Entretanto, tais modelos operam majoritariamente como "caixas-pretas", o que dificulta a compreensão detalhada de quais parâmetros acústicos específicos fundamentam a decisão final do classificador. Em contrapartida, esta dissertação opta pela utilização de modelos clássicos, especificamente as Máquinas de Vetores de Suporte (SVM), e pela extração de características acústicas manuais. Essa escolha metodológica prioriza a interpretabilidade, permitindo investigar como atributos físicos e fisiológicos contribuem individualmente para a discriminação emocional. Assim, embora os índices de acurácia resultantes desta abordagem sejam mais modestos frente aos sistemas de redes neurais profundas, o foco do estudo reside em fornecer uma análise cientificamente transparente e explicável sobre a assinatura acústica dos estados afetivos.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Analisar características acústicas extraídas de sinais de voz — como medidas prosódicas, espectrais e de qualidade vocal — para a detecção de emoções humanas, utilizando de classificação baseada em SVM.

1.1.2 Objetivos Específicos

- Extrair características acústicas da fala, contemplando parâmetros de prosódia, qualidade vocal e descritores espectrais relevantes para a diferenciação emocional.
- Aplicar o classificador baseado em Máquinas de Vetores de Suporte (SVM) para identificar as emoções humanas a partir das características extraídas.
- Avaliar o desempenho do modelo de classificação, utilizando métricas quantitativas que permitam verificar a eficiência e a consistência dos resultados obtidos.
- Interpretar os resultados, destacando as características acústicas com maior contribuição para a distinção entre as emoções analisadas.

2. FUNDAMENTAÇÃO TEÓRICA

A fala é um dos meios mais naturais e expressivos de comunicação humana, permitindo não apenas a transmissão de informações linguísticas, mas também a manifestação de estados emocionais. A forma como uma pessoa fala, sua entonação, ritmo, intensidade e timbre, pode revelar emoções, intenções e até mesmo condições fisiológicas e psicológicas. Nesse contexto, compreender e reproduzir computacionalmente essa dimensão afetiva da fala tem se tornado um dos principais desafios na área de processamento de sinais e reconhecimento de padrões. Neste capítulo serão apresentados os fundamentos teóricos do reconhecimento de emoções na fala, abordando conceitos, desafios e aplicações da área. Também são descritas a base de dados utilizada no estudo e as principais características acústicas extraídas do sinal de voz. Por fim, são apresentados os princípios das SVM, utilizadas como método de classificação neste trabalho.

2.1 RECONHECIMENTO DE EMOÇÕES NA FALA

O reconhecimento de emoções a partir da fala é uma subárea do processamento de sinais de voz que visa identificar o estado emocional de um indivíduo por meio da análise das propriedades acústicas do discurso. Essa tarefa envolve a identificação de emoções básicas, como alegria, tristeza, raiva, medo, nojo, surpresa e neutralidade, a partir de variações no padrão vocal. A capacidade de interpretar tais sinais emocionais de forma automática tem impulsionado o desenvolvimento de tecnologias, com aplicações em assistentes virtuais, monitoramento clínico remoto, veículos inteligentes, centrais de atendimento automatizado e interfaces acessíveis (KHALIL et al., 2019).

Nos últimos anos, o interesse acadêmico pelo reconhecimento de emoções na fala cresceu significativamente, impulsionado pelo amadurecimento de algoritmos de aprendizado de máquina, pelo aumento da disponibilidade de bases de dados rotuladas com qualidade e pelo aprofundamento da compreensão sobre como os estados emocionais afetam a produção vocal (WANG et al., 2024). O sinal de voz não carrega apenas o conteúdo semântico das palavras, mas também elementos paralinguísticos, tais como ritmo, entonação, intensidade e harmonia, que refletem o estado emocional do locutor. Essas nuances são captadas por meio de alterações em atributos como frequência fundamental, tempo de fala, qualidade vocal e distribuição espectral, permitindo que emoções sejam inferidas de forma não intrusiva (NASSIF et al., 2019).

O processo de reconhecimento emocional automatizado geralmente compreende três etapas centrais: (i) o pré-processamento do sinal, que pode incluir remoção de silêncio, normalização, segmentação e aplicação de janelas temporais; (ii) a extração de características acústicas que representem de forma quantitativa os aspectos prosódicos, espectrais e de qualidade vocal da fala; e (iii) a classificação supervisionada com o uso de algoritmos como máquinas de vetores de suporte, florestas aleatórias, redes neurais profundas, ou modelos baseados em

mistura de gaussianas, entre outros (KHALIL et al., 2019). Cada uma dessas etapas influencia diretamente a acurácia, robustez e interpretabilidade do sistema, e sua escolha deve considerar fatores como variabilidade entre locutores, idioma, condições de gravação e tamanho da base de dados.

Apesar dos avanços recentes, o SER ainda enfrenta desafios. Um dos principais obstáculos é a elevada variabilidade entre locutores, que pode mascarar os padrões emocionais no sinal. Além disso, muitas emoções possuem fronteiras difusas, sendo difíceis de distinguir até mesmo por humanos. O problema é agravado pelo uso recorrente de bases com atuações emocionais exageradas, que não refletem fielmente os padrões espontâneos observados em situações naturais (ILIOU; ANAGNOSTOPOULOS, 2010). A inconsistência nos critérios de rotulação entre diferentes bases, seja no número de emoções, intensidade ou valência, também dificulta a generalização dos modelos, tornando-os sensíveis ao contexto de treinamento.

2.2 BASES DE DADOS PARA RECONHECIMENTO EMOCIONAL NA FALA

A base de dados utilizada em um sistema de reconhecimento de emoções na fala exerce influência direta sobre sua capacidade de generalização, confiabilidade estatística e validade experimental. Embora diversas bases tenham sido desenvolvidas ao longo dos últimos anos, muitas apresentam limitações quanto à padronização, à variabilidade das expressões ou ao controle de fatores extrínsecos, como ruído, sotaque e inconsistência de enunciados.

Entre as bases comumente mais utilizadas destacam-se a EMO-DB, composta por dez locutores do alemão em gravações atuadas de sete emoções básicas, e a IEMOCAP, um corpus multimodal amplamente citado por sua riqueza em termos de diversidade de dados, incluindo gravações de áudio, vídeo, transcrições textuais e anotações emocionais detalhadas. Essa variedade de modalidades e informações contextuais torna a base valiosa para pesquisas em reconhecimento de emoções. Entretanto, ela também é criticada pela variabilidade entre os estilos de atuação e pela falta de controle sobre os enunciados falados. Tais bases têm contribuído para o avanço da área, mas impõem dificuldades metodológicas, especialmente quando se pretende investigar a separabilidade estatística de características acústicas com precisão e replicabilidade (BURKHARDT et al., 2005; BUSSO et al., 2008).

Neste contexto, a presente dissertação adota a base RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) (LIVINGSTONE; RUSSO, 2018), que se destaca na literatura contemporânea por reunir, de forma equilibrada, rigor experimental, diversidade de emoções e alta qualidade técnica de gravação. A RAVDESS foi desenvolvida com o objetivo de fornecer um conjunto de dados padronizado e controlado para pesquisas em análise emocional, contendo amostras de fala e canto emocional de 24 atores profissionais (12 homens e 12 mulheres), todos nativos do inglês norte-americano. O diferencial da RAVDESS está em sua estrutura cuidadosamente equilibrada. Cada ator grava as mesmas duas frases neutras em oito

emoções distintas (neutro, calmo, feliz, triste, bravo, assustado, enojado e surpreso), em duas intensidades (normal e forte), e com repetições. Esse grau de controle garante uniformidade entre classes emocionais e neutraliza o impacto de fatores semânticos ou fonéticos nos resultados, tornando a base ideal para comparações estatísticas rigorosas entre diferentes características acústicas. Além disso, as gravações foram realizadas em ambiente acusticamente isolado, com taxa de amostragem de 48 kHz, resultando em arquivos livres de ruídos de fundo e distorções, assegurando maior fidelidade na extração de medidas acústicas sensíveis, como *jitter*, *shimmer* e *Cepstral Peak Prominence Smoothed* (CPPS).

Outro aspecto é a completa documentação da base e sua disponibilidade em acesso aberto, o que reforça seu valor científico ao permitir replicabilidade e comparabilidade com outros estudos. A Tabela 2 resume a estrutura da RAVDESS, considerando apenas os trechos de fala (ou seja, excluindo as amostras de canto).

Tabela 2 – Distribuição das amostras de fala na base RAVDESS utilizada neste trabalho.

Emoção	Intensidade	Nº de locutores	Total de amostras
Neutro	Normal	24	96
Calma	Normal/Forte	24	192
Felicidade	Normal/Forte	24	192
Tristeza	Normal/Forte	24	192
Raiva	Normal/Forte	24	192
Medo	Normal/Forte	24	192
Nojo	Normal/Forte	24	192
Surpresa	Normal/Forte	24	192
Total	,	,	1.440

2.3 CARACTERÍSTICAS ACÚSTICAS DA VOZ

A voz humana constitui uma das manifestações mais complexas da comunicação, integrando elementos fisiológicos, articulatórios e cognitivos. Para além do conteúdo semântico, o sinal vocal carrega informações sutis sobre o estado emocional do falante, as quais se manifestam por meio de alterações em propriedades acústicas como frequência, intensidade, estabilidade e distribuição espectral (KHALIL et al., 2019). Tais propriedades, quantificadas por meio de descritores acústicos, servem como base para a análise computacional das emoções expressas na fala (WANG et al., 2024).

No contexto do reconhecimento de emoções, a escolha e compreensão dessas características acústicas são fundamentais para a construção de sistemas robustos (CHOWDHURY; RAMANNA; KOTECHA, 2025). Cada característica representa um aspecto distinto da produção vocal, seja relacionado à vibração das pregas vocais, ao controle articulatório, à harmonia do sinal glótico ou à estrutura espectral resultante. Por esse motivo, o desempenho de sistemas de classificação emocional está intimamente vinculado à qualidade e à representatividade do conjunto de descritores utilizados (ILIOU; ANAGNOSTOPOULOS, 2010).

Esta seção apresenta e descreve as 21 características acústicas extraídas dos áudios da base RAVDESS. As características são agrupadas em três categorias principais: (i) características prosódicas, que refletem propriedades de entonação, intensidade e ritmo; (ii) características de qualidade vocal, relacionadas à estabilidade da fonação e ao esforço vocal; e (iii) características espectrais, que descrevem a distribuição de energia em diferentes faixas de frequência, capturando aspectos relacionados ao timbre e à articulação (NASSIF et al., 2019). Para cada característica, serão apresentados sua definição conceitual, sua formulação matemática e sua relevância no contexto da discriminação emocional.

A abordagem adotada neste trabalho enfatiza a análise individual de cada descritor, com o objetivo de avaliar seu poder discriminativo e sua contribuição isolada para a tarefa de reconhecimento de emoções. Essa estratégia visa não apenas melhorar o desempenho dos classificadores, mas também aprofundar a compreensão sobre os mecanismos acústicos subjacentes à expressão emocional vocal (MASHHADI; OSEI-BONSU, 2023).

2.3.1 Características Prosódicas

As características prosódicas refletem propriedades globais da fala relacionadas à entonação, à intensidade e ao ritmo do discurso. Elas são moduladas por mecanismos fisiológicos do trato vocal, com destaque para o sistema subglótico, responsável pela pressão pulmonar que impulsiona a fonação; o controle laríngeo, que regula a frequência de vibração das pregas vocais (frequência fundamental, ou F_0) e os padrões articulatórios que definem a temporalidade e fluência da fala. Juntos, esses mecanismos conferem à voz humana sua expressividade natural, tornando a prosódia um dos principais canais paralinguísticos de transmissão emocional (COSTA et al., 2008).

A relevância das características prosódicas no reconhecimento de emoções decorre do fato de que estados afetivos distintos induzem alterações fisiológicas perceptíveis no padrão vocal. Emoções de alta excitação, como raiva, alegria ou medo, tendem a elevar a frequência fundamental e a intensidade da voz, resultando em fala mais tensa, rápida e com maior projeção energética. Por outro lado, emoções de baixa excitação, como tristeza, resignação ou desânimo, manifestam-se por meio de reduções na entonação, menor variação melódica e intensidade reduzida, refletindo um padrão prosódico mais plano e monótono (WANG et al., 2024; KHALIL et al., 2019).

Além disso, a prosódia desempenha um papel fundamental na segmentação perceptiva do discurso, auxiliando ouvintes a identificar unidades prosódicas, pausas, ênfases e variações de intenção comunicativa. Em contextos emocionais, essas variações carregam significados que transcendem o conteúdo semântico da fala, atuando como marcadores acústicos implícitos do estado psicológico do falante. Em aplicações computacionais, a modelagem eficiente da prosódia permite que sistemas de reconhecimento emocional captem nuances afetivas de forma

mais natural e intuitiva (COSTA et al., 2008).

Nesta dissertação, foram selecionadas três características prosódicas para análise: a frequência fundamental (F_0), que representa a entonação ou “altura” da voz; a raiz quadrática média (RMSE), que expressa a intensidade acústica do sinal; e a Taxa de Cruzamento por Zero (TCZ), que quantifica a rapidez com que o sinal cruza o eixo temporal, refletindo aspectos da aspereza e da presença de componentes de alta frequência na fala. Todas essas métricas são extraídas em janelas curtas de 25 ms e posteriormente agregadas estatisticamente por áudio. A escolha por esse conjunto de características se justifica por sua consistência frente à variabilidade interindividual e por sua ampla aceitação na literatura como marcadores confiáveis de excitação e dinamismo emocional na fala.

A proeminência dessas características deve-se ao fato de que a excitação fisiológica manifesta-se de forma primária e direta nas propriedades mecânicas da laringe e na pressão subglótica. Consequentemente, a RMSE e a F_0 atuam como os marcadores acústicos mais robustos para diferenciar estados de alta ativação (como raiva) de estados de baixa ativação (como tristeza), independentemente de variações tímbricas mais complexas.

Frequência Fundamental (F_0)

A frequência fundamental representa a taxa de vibração das pregas vocais e é determinada pelas características fisiológicas do aparelho fonador, incluindo tensão muscular, comprimento das pregas e pressão subglótica. Tratando-se de um dos principais indicadores prosódicos, a F_0 reflete diretamente aspectos emocionais, como nível de excitação e valência afetiva (WANG et al., 2024; KHALIL et al., 2019). Emoções de alta excitação, como alegria, surpresa e raiva, tipicamente elevam o valor médio da F_0 , enquanto estados emocionais de baixa ativação, como tristeza e tédio, reduzem sua média e variabilidade.

Para estimar a F_0 em sinais de voz, uma abordagem comum é o método de autocorrelação. Dividindo-se o sinal $x(n)$ em janelas curtas (tipicamente 25 ms), calcula-se:

$$R(\tau) = \sum_{n=0}^{N-\tau-1} x(n) \cdot x(n + \tau) \quad , \quad (1)$$

em que τ é o atraso. O primeiro pico significativo de $R(\tau)$, ignorando o valor máximo na origem ($\tau = 0$), corresponde ao período fundamental da vibração glótica, denotado por T_0 . A frequência fundamental F_0 é então calculada como o inverso desse período:

$$F_0 = \frac{1}{T_0} \quad (2)$$

A F_0 possui poder discriminativo em tarefas de SER, algoritmos baseados nela alcançam alta acurácia em diversos trabalhos recentes. Contudo, ela é sensível à variações individuais, como

sexo biológico (homens têm F_0 média mais baixa que mulheres), idade, sotaque e estratégias de fala (canto versus fala natural) (MASHHADI; OSEI-BONSU, 2023).

Energia (*Root Mean Square Energy* – RMSE)

A energia da fala expressa a intensidade com que o som é emitido, refletindo não apenas o grau de ativação do sistema fonatório, mas também o estado emocional do locutor. Emoções de alta excitação, como raiva, surpresa e alegria, geralmente resultam em maior projeção vocal, traduzindo-se em níveis elevados de energia sonora. Por outro lado, estados emocionais de baixa excitação, como tristeza ou tédio, tendem a se manifestar por intensidades vocais reduzidas, evidenciando um padrão de fala mais contido, monótono e introspectivo (WANG et al., 2024; KHALIL et al., 2019).

A métrica de energia utilizada neste trabalho é a *Root Mean Square Energy* (RMSE), amplamente usada por sua simplicidade e eficácia na quantificação da potência acústica do sinal. A equação é dada por:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \quad , \quad (3)$$

em que x_i representa a amplitude da i -ésima amostra de áudio dentro de uma janela de N pontos (tipicamente 25 ms com sobreposição de 50%).

A adoção da RMSE se justifica por sua baixa complexidade computacional, boa tolerância a ruídos de fundo e alta interpretabilidade. Contudo, seu valor absoluto pode ser sensível a fatores como o ganho do microfone e a distância entre locutor e receptor, introduzindo variabilidade não relacionada à emoção (MASHHADI; OSEI-BONSU, 2023).

Do ponto de vista fisiológico, emoções intensas estão associadas ao aumento da pressão subglótica, à maior ativação dos músculos respiratórios e laríngeos e, consequentemente, ao incremento da intensidade vocal. Valores elevados de RMSE indicam esse esforço vocal intensificado, servindo como reflexo direto do grau de excitação emocional (KHALIL et al., 2019). Ainda assim, sua interpretação exige contexto e padronização adequados, especialmente em cenários aplicados ou gravações não controladas.

Em suma, a RMSE configura-se como uma das métricas prosódicas mais relevantes para tarefas de reconhecimento de emoções na fala. Quando tratada estatisticamente, oferece uma visão clara e objetiva da intensidade emocional expressa por meio da voz, complementando outras dimensões prosódicas, espectrais e de qualidade vocal.

Taxa de Cruzamento por Zero (TCZ)

A *Taxa de Cruzamento por Zero (TCZ)*, também conhecida como *Zero-Crossing Rate (ZCR)*, é uma métrica simples, porém eficaz, que representa a frequência com que o sinal de fala cruza o eixo zero em um intervalo de tempo determinado. Em termos físicos, ela indica quantas vezes a polaridade do sinal muda entre amostras consecutivas, sendo particularmente sensível a componentes de alta frequência e à presença de ruído (WANG et al., 2024).

Sua formulação é dada por:

$$TCZ = \frac{1}{2N} \sum_{n=1}^N |\text{sgn}(x[n]) - \text{sgn}(x[n-1])|, \quad (4)$$

em que $\text{sgn}(x[n])$ é a função sinal, que retorna +1 para valores positivos e -1 para negativos. A TCZ é normalmente calculada em janelas curtas (ex. 25 ms) e expressa como uma média por segmento de fala.

Altos valores de TCZ estão associados a sinais ruidosos, agudos ou com alta concentração de energia em frequências elevadas, como ocorre em consoantes fricativas ou fala excitada. Já sinais com baixas taxas de cruzamento tendem a ser mais graves, suaves e sustentados, como em falas calmas, melancólicas ou articuladas com pouca tensão muscular (KHALIL et al., 2019). Essa sensibilidade à estrutura temporal torna a TCZ útil para distinguir emoções com perfis espectrais distintos, como raiva (alta TCZ) e tristeza (baixa TCZ).

Além disso, a TCZ é computacionalmente leve, sendo utilizada em cenários de baixo custo computacional, como sistemas embarcados ou de resposta em tempo real. Sua principal limitação está na baixa robustez frente a ruídos externos, exigindo sinais com boa relação sinal-ruído, o que reforça a adequação do uso da base RAVDESS, cujas gravações são limpas e controladas.

Por sua capacidade de capturar alterações abruptas na forma de onda e sua relação indireta com a frequência dominante, a TCZ atua como uma característica complementar às medidas espectrais e de qualidade vocal. Sua inclusão enriquece a representação acústica de emoções e contribui para o desempenho de classificadores em sistemas SER baseados em análise multivariada.

2.3.2 Características de Qualidade Vocal

As características de qualidade vocal representam atributos do sinal de fala que emergem da biomecânica fina da fonação, refletindo diretamente a estabilidade e regularidade das vibrações das pregas vocais durante a produção sonora. Essas características não apenas capturam propriedades acústicas globais, mas também revelam variações sutis e rápidas nos padrões vibratórios da laringe, como instabilidades no período ou na amplitude da onda glótica,

que muitas vezes não são perceptíveis de forma consciente, mas são indicadores sensíveis do estado fisiológico e emocional do falante (KHALIL et al., 2019; CHOWDHURY; RAMANNA; KOTTECHA, 2025).

Diferentemente das características prosódicas (como F_0 e RMSE), que estão ligadas à entoação e intensidade do discurso, os atributos de qualidade vocal fornecem uma visão mais detalhada sobre o controle motor laríngeo e a eficiência da vibração das pregas vocais. Essas medidas são particularmente úteis na detecção de emoções que envolvem tensões musculares elevadas, alterações respiratórias ou controle emocional comprometido, como raiva, medo, ansiedade ou tristeza profunda (WANG et al., 2024).

Do ponto de vista fisiológico, estados emocionais intensos ativam o sistema nervoso autônomo, alterando o tônus muscular laríngeo e a coordenação entre respiração e fonação. Essas alterações se manifestam acusticamente por meio de tremores, irregularidades na frequência ou na amplitude do sinal, ou aumento do ruído glotal, sendo captadas por medidas como *jitter*, *shimmer*, Razão Harmônico-Ruído (HNR) e CPPS (NASSIF et al., 2019).

Nas subseções seguintes, são descritas quatro métricas centrais de qualidade vocal: *jitter*, *shimmer*, HNR e CPPS, incluindo suas definições matemáticas, fundamentação fisiológica e relevância para a análise emocional.

Jitter

O *jitter* é uma medida que quantifica a irregularidade temporal da vibração das pregas vocais durante a fonação. Especificamente, ele expressa a variação ciclo a ciclo do período glótico, ou seja, o tempo entre sucessivos ciclos de fechamento e abertura das pregas vocais, que corresponde à frequência fundamental (F_0). Essa métrica é particularmente sensível a instabilidades no controle laríngeo, muitas vezes induzidas por fatores emocionais ou neuromusculares (COSTA et al., 2008).

Durante estados emocionais como medo, ansiedade, tensão ou raiva reprimida, o sistema nervoso autônomo pode interferir na precisão do fechamento glótico, gerando flutuações involuntárias no tempo de cada ciclo. Isso se manifesta acusticamente como pequenas variações na F_0 entre ciclos consecutivos, variações estas que são capturadas pelas métricas de *jitter*. Já em estados como tristeza ou neutralidade, a estabilidade laríngea tende a ser maior, produzindo menor *jitter* e uma vibração mais regular das pregas vocais (NASSIF et al., 2019).

Duas formas comuns de estimativa do *jitter* são o valor absoluto e o relativo. O primeiro mede a diferença média entre períodos consecutivos:

$$\text{Jitter}_{\text{abs}} = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_{i+1} - T_i| \quad (5)$$

Já o *jitter* relativo normaliza essa diferença pela média dos períodos, sendo mais adequado para comparações entre locutores com diferentes faixas de F_0 :

$$\text{Jitter}_{\text{rel}} = \frac{\sum_{i=1}^{N-1} |T_{i+1} - T_i|}{(N-1)\bar{T}} \times 100 \quad , \quad (6)$$

em que T_i representa o período do i -ésimo ciclo glótico e $\bar{T}$ a média dos períodos. A unidade é geralmente expressa em percentual.

Por ser fortemente dependente da regularidade da vibração laríngea, o *jitter* é mais confiável quando extraído de sinais com qualidade acústica elevada e sem ruídos de fundo, como é o caso da base RAVDESS, que proporciona condições ideais para extração precisa. Quando combinado com medidas complementares como *shimmer* e HNR, o *jitter* contribui para capturar a dimensão fisiológica da emoção, revelando aspectos internos da produção vocal que não são evidentes em parâmetros mais globais, como F_0 ou RMSE. Assim, sua inclusão neste estudo se justifica pela capacidade de representar microvariações associadas a emoções intensas, sendo um parâmetro acústico de alto valor discriminativo na análise da qualidade vocal emocional.

Shimmer

O *shimmer* é uma medida acústica que representa a instabilidade da amplitude do sinal de fala entre ciclos sucessivos da vibração glótica. Em termos fisiológicos, ele reflete variações rápidas e involuntárias na força com que o ar é expelido pelas pregas vocais a cada ciclo de fonação. Assim como o *jitter* está relacionado à frequência fundamental, o *shimmer* atua como um análogo para a dimensão da intensidade vocal (COSTA et al., 2008).

Tais variações na amplitude geralmente ocorrem por flutuações no controle neuromuscular da pressão subglótica ou por tensões irregulares nas pregas vocais, características comuns em estados emocionais intensos como raiva, angústia, medo e euforia. Vozes com altos níveis de *shimmer* tendem a soar trêmulas, irregulares ou instáveis, o que afeta diretamente a percepção da emoção transmitida (COSTA et al., 2008).

A formulação clássica do *shimmer* relativo, amplamente utilizada por seu caráter normalizado, é dada por:

$$\text{Shimmer} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left| \frac{A_{i+1} - A_i}{A_i} \right| \times 100 \quad , \quad (7)$$

em que A_i representa a amplitude do i -ésimo ciclo glótico, medida em janelas de tempo curtas. A média dessas diferenças é expressa como um percentual, indicando o grau de flutuação da intensidade entre ciclos consecutivos.

Valores mais elevados de *shimmer* são observados em locuções com esforço vocal excessivo ou instabilidade emocional, enquanto vozes controladas e neutras apresentam baixos índices. Estudos demonstram que o *shimmer* é particularmente eficaz para distinguir emoções de alta excitação de emoções neutras ou de baixa intensidade, complementando métricas como energia média (RMSE) e HNR na caracterização da qualidade vocal (COSTA et al., 2008).

Na prática, o *shimmer* é sensível a ruídos de fundo, à qualidade do microfone e à distância entre locutor e receptor, fatores que podem introduzir artefatos na análise. Por isso, sua extração exige gravações de alta fidelidade e ambiente controlado, condições atendidas pela base RAVDESS, que oferece gravações com excelente relação sinal-ruído e equilíbrio entre locutores (COSTA et al., 2008).

Em modelos de reconhecimento de emoções, o *shimmer* tem sido utilizado como uma das principais variáveis associadas ao estado fisiológico do locutor, funcionando como um indicador acústico da tensão laríngea ou instabilidade respiratória. Quando empregado em conjunto com *jitter*, HNR e CPPS, permite traçar um panorama completo da qualidade vocal sob diferentes estados emocionais (COSTA et al., 2008).

Razão harmônico-ruído (HNR)

A Razão Harmônico-Ruído é uma medida acústica que avalia a proporção entre os componentes periódicos e os componentes aperiódicos do sinal de fala. Em termos práticos, ela quantifica o quão “limpo” ou “ruidoso” é um sinal vocal, fornecendo uma estimativa direta da regularidade da vibração glótica e da presença de ruído glotal (COSTA et al., 2008).

O HNR é amplamente utilizado em análises clínicas da voz, sendo considerado um dos principais indicadores de qualidade vocal. Em contextos emocionais, sua aplicação tem ganhado destaque por sua capacidade de distinguir entre vozes controladas, típicas de emoções neutras ou calmas, e vozes instáveis ou ásperas, frequentemente associadas a estados emocionais como raiva, medo ou angústia (WANG et al., 2024). Emoções de alta excitação, por exemplo, tendem a provocar aumento da pressão subglótica e instabilidade na coordenação respiratória-laríngea, elevando o componente aperiódico da fala e reduzindo o valor de HNR.

A formulação geral do HNR é expressa por:

$$\text{HNR} = 10 \log_{10} \left(\frac{P_{\text{harm}}}{P_{\text{noise}}} \right) , \quad (8)$$

em que P_{harm} representa a potência dos componentes harmônicos do sinal (associados à vibração periódica das pregas vocais), e P_{noise} representa a potência dos componentes não harmônicos, frequentemente relacionados ao ruído de fricção, escape de ar ou tremores vocais.

Altos valores de HNR indicam que o sinal é dominado por componentes periódicos, o que é típico de uma fonação estável, regular e limpa, como ocorre em estados emocionais

neutros, calmos ou de fala preparada. Em contraste, baixos valores de HNR refletem um aumento na proporção de ruído, indicando irregularidades vocais compatíveis com esforço fonatório excessivo, tensão muscular ou instabilidade respiratória (MASHHADI; OSEI-BONSU, 2023).

Do ponto de vista computacional, o HNR pode ser extraído a partir da função de autocorrelação do sinal ou por meio de análise espectral com técnicas como a transformada de Fourier. A estimativa depende de janelas curtas de tempo e é sensível à qualidade de gravação, motivo pelo qual bases como a RAVDESS, com alto padrão acústico, são ideais para sua utilização.

Estudos mostram que a combinação de HNR com medidas como *jitter* e *shimmer* permite uma caracterização rica da qualidade vocal, contribuindo significativamente para a performance de sistemas de reconhecimento de emoções (COSTA et al., 2008). Assim, o HNR figura como um componente essencial na representação acústica de estados emocionais, especialmente quando se busca compreender a estabilidade fisiológica da produção vocal.

Contudo, métricas de qualidade vocal como o HNR são extremamente sensíveis a ruídos de extração e a presença de valores discrepantes (*outliers*). Em processos de normalização global (como o Min-Max), a ocorrência de *outliers* extremos pode comprimir a variância dos demais dados, resultando em um fenômeno visual de 'achatamento' nas distribuições estatísticas, o que exige cautela na interpretação visual desses diagramas de caixa.

Proeminência do Pico Cepstral Suavizada (CPPS)

A Proeminência do Pico Cepstral Suavizada (CPPS, do inglês *Cepstral Peak Prominence-Smoothed*) é uma medida acústica que expressa a clareza e a estabilidade da sonoridade da fala. Trata-se de um descritor derivado da análise cepstral, técnica que permite quantificar a regularidade harmônica de um sinal com base na estrutura de repetições periódicas presentes no espectro de frequência.

O cálculo do CPPS baseia-se em três etapas principais: (i) obtenção do espectro de potência do sinal por meio da transformada rápida de Fourier (FFT, do inglês *Fast Fourier Transform*); (ii) aplicação do logaritmo sobre esse espectro para evidenciar os componentes harmônicos; e (iii) cálculo do cepstro, por meio da transformada inversa do log-espectro. O resultado é um domínio alternativo, denominado quefrência, no qual picos proeminentes indicam padrões de periodicidade estáveis, correspondentes ao período fundamental da voz.

O CPPS mede o quanto esse pico cepstral se destaca em relação a uma curva de base suavizada, servindo como um índice de periodicidade vocal. Sua formulação é expressa por:

$$\text{CPPS} = 10 \log_{10} \left(\frac{A_{\text{pico}}}{A_{\text{suav}}} \right) , \quad (9)$$

em que A_{pico} representa a amplitude do pico cepstral localizado na quefrência correspondente à frequência fundamental estimada, e A_{suav} é o valor da envoltória suavizada nesse mesmo ponto,

calculada geralmente por regressão linear local. O valor resultante é expresso em decibéis (dB) e representa o grau de proeminência harmônica do sinal.

Altos valores de CPPS indicam uma estrutura harmônica clara e estável, característica de vozes bem projetadas, regulares e com pouca perturbação glótica. Por outro lado, baixos valores refletem vozes com instabilidade, aspereza, soprosidade ou esforço fonatório excessivo, como ocorre em condições patológicas ou em contextos emocionais de alta carga afetiva.

No reconhecimento de emoções, o CPPS tem se mostrado uma métrica útil para a detecção de emoções como raiva, medo ou frustração, nas quais a emissão vocal pode tornar-se irregular, com perda de periodicidade e aumento do ruído glotal. Por outro lado, vozes neutras, calmas ou com controle emocional tendem a preservar uma estrutura harmônica bem definida, resultando em valores mais elevados de CPPS (WANG et al., 2024).

2.3.3 Características Espectrais

As características espectrais representam a distribuição da energia do sinal de fala ao longo das frequências, permitindo descrever como o conteúdo acústico se organiza no domínio espectral. Esse tipo de análise é fundamental no reconhecimento de emoções, pois diferentes estados emocionais tendem a alterar a concentração de energia, o formato do espectro e a intensidade relativa entre bandas de frequência. Entre as medidas mais utilizadas destacam-se o centroide espectral, a largura de banda, a inclinação espectral, os coeficientes cepstrais na escala Mel (MFCC) e parâmetros derivados da Transformada Rápida de Fourier. Essas características oferecem uma representação compacta e perceptualmente relevante da estrutura espectral, sendo amplamente empregadas em tarefas de classificação de fala, incluindo o reconhecimento de emoções.

Coefficientes Cepstrais na Escala Mel (MFCCs)

Os *Mel-Frequency Cepstral Coefficients* (MFCCs) são amplamente utilizados na análise de sinais de fala devido à sua capacidade de representar a estrutura espectral da voz de maneira compacta e perceptualmente significativa. Inspirados na forma como o ouvido humano percebe diferentes bandas de frequência, os MFCCs aplicam uma escala mel para priorizar as regiões espectrais mais relevantes à audição humana, com maior resolução em baixas frequências e menor em altas (WANG et al., 2024; CHOWDHURY; RAMANNA; KOTECHA, 2025).

O processo de extração dos MFCCs envolve várias etapas sucessivas:

- Segmentação do sinal em janelas curtas (tipicamente 25 ms), com sobreposição de 50% e aplicação de uma janela de Hamming para suavizar descontinuidades;
- Transformada de Fourier em cada segmento, para obter o espectro de magnitude;

- Mapeamento do espectro linear para o domínio mel, por meio de um banco de filtros triangulares sobrepostos;
- Logaritmização das energias de cada filtro;
- Transformada Discreta do Cosseno (DCT, do inglês *Discrete Cosine Transform*) para obter os coeficientes cepstrais.

A sequência de fórmulas envolvidas pode ser sintetizada como:

$$S[k] = |\mathcal{F}\{x[n] \cdot w[n]\}|^2, \quad (10)$$

$$E[m] = \sum_k S[k] \cdot H_m[k], \quad (11)$$

$$\text{MFCC}[n] = \sum_{m=1}^M \log(E[m]) \cdot \cos\left(\frac{\pi n(m - 0.5)}{M}\right), \quad (12)$$

em que: $x[n]$ é o sinal de entrada; $w[n]$ é a janela de Hamming; $S[k]$ é o espectro de potência; $H_m[k]$ são os filtros triangulares na escala Mel; $E[m]$ são as energias nos filtros; M é o número de filtros (geralmente entre 20 e 40); n é o índice do coeficiente (tipicamente os primeiros 12 a 13 são usados).

Os MFCCs capturam informações sobre o timbre e a estrutura articulatória da fala, sendo menos sensíveis à entonação e intensidade. Essa característica os torna especialmente úteis para tarefas que requerem discriminação entre diferentes estilos vocais, articulatórios ou emocionais. Na literatura, os MFCCs são sistematicamente apontados como uma das características mais eficazes em tarefas de reconhecimento emocional (KHALIL et al., 2019).

Os MFCCs são uma representação do espectro de potência de curto termo de um sinal de áudio, baseados na escala não linear de frequências Mel. Amplamente utilizados no reconhecimento e na análise de fala, os MFCCs descrevem a forma geral da envoltória espectral do sinal, imitando o comportamento do sistema auditivo humano para capturar as informações tímbricas e harmônicas.

O processo de extração desses coeficientes pode ser resumido nas seguintes etapas: primeiro, obtém-se o espectro de magnitude do sinal por meio da Transformada de Fourier; em seguida, mapeia-se esse espectro para a escala Mel utilizando um banco de filtros triangulares sobrepostos; calcula-se o logaritmo da energia na saída de cada filtro e, por fim, aplica-se a transformada discreta de cosseno (DCT) ou transformada inversa. Essa técnica descreve a variação de energia ao longo de diferentes bandas de frequência, permitindo separar as características do trato vocal e descartar detalhes irrelevantes para certas análises, como o *pitch*

e o ruído da glote. Além disso, os MFCCs podem ser combinados com outros atributos, como os coeficientes delta e delta-delta, que representam as suas taxas de mudança ao longo do tempo (COSTA et al., 2008).

É justamente por focar na envoltória espectral e isolar o trato vocal que, em contextos emocionais, os MFCCs são capazes de refletir alterações sutis na articulação e na configuração física da fala. Emoções como raiva ou medo podem gerar vocalizações mais tensas, com maior concentração espectral em regiões médias-altas, enquanto tristeza ou desânimo tendem a produzir espectros mais suaves e graves. Essas variações são capturadas de forma implícita pelos coeficientes que, quando usados em conjunto com modelos de aprendizado de máquina, revelam alto poder discriminativo (MASHHADI; OSEI-BONSU, 2023).

Centroide Espectral

O centroide espectral é uma medida que representa o centro de massa do espectro de frequências de um sinal de áudio, ponderado pela magnitude de cada componente espectral. É comumente interpretado como um indicador de brilho ou acuidade sonora: quanto mais alto o centroide, maior a concentração de energia em frequências agudas; quanto mais baixo, maior o predomínio de componentes graves (WANG et al., 2024).

Essa métrica é amplamente utilizada em tarefas de classificação de timbre, análise musical e reconhecimento de emoções na fala, pois responde de forma sensível a alterações no padrão articulatório, na tensão vocal e na excitação emocional do locutor. Em estados emocionais como raiva, surpresa ou medo, a fala tende a se tornar mais tensa e aguda, deslocando o centroide para frequências mais elevadas. Por outro lado, em emoções como tristeza, desânimo ou tédio, o centroide se desloca para regiões mais graves do espectro (DAR; DELHIBABU, 2024).

A formulação clássica do centroide espectral C é dada por:

$$C = \frac{\sum_{k=0}^{N-1} f_k \cdot |X[k]|}{\sum_{k=0}^{N-1} |X[k]|}, \quad (13)$$

em que: $X[k]$ é o valor do espectro de magnitude na frequência f_k ; N é o número total de pontos da transformada rápida de Fourier (FFT).

A equação mostra que o centroide é a média ponderada das frequências presentes no espectro, com pesos proporcionais à energia de cada componente. O cálculo é normalmente realizado em janelas curtas (ex. 25 ms), e os valores são posteriormente agregados por estatísticas como média e desvio padrão para caracterizar o áudio completo.

O centroide espectral é uma métrica de baixo custo computacional e com forte poder discriminativo, sendo especialmente útil para distinguir emoções de valência semelhante mas com diferentes níveis de excitação (como alegria vs. calma, ou raiva vs. tédio). Além disso,

ele complementa outras características espectrais como MFCCs, fornecendo uma medida mais intuitiva da distribuição de energia no eixo de frequência (DAR; DELHIBABU, 2024).

2.4 MÁQUINAS DE VETORES DE SUPORTE (SVM)

As Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*) constituem um dos classificadores supervisionados mais utilizados em reconhecimento de emoções na fala, devido à sua robustez estatística, capacidade de generalização e desempenho consistente mesmo em bases com número limitado de amostras. De forma geral, a SVM busca um hiperplano ótimo que melhor separe as classes no espaço de características, maximizando a distância entre ele e os pontos mais próximos, os vetores de suporte (DAR; DELHIBABU, 2024). Essa estratégia garante um modelo mais estável e menos sensível a ruídos ou variações de amostragem.

A formulação clássica da SVM (YAN et al., 2020) para o caso binário considera um hiperplano definido por:

$$\mathbf{w}^T \mathbf{x} + b = 0, \quad (14)$$

em que $\mathbf{w}$ representa o vetor normal ao hiperplano, b o termo de viés e $\mathbf{x}$ uma amostra no espaço de atributos. A ideia central é que as amostras sejam projetadas de modo que satisfaçam as condições de margem. Para isso, a maximização da margem é reformulada como a minimização da norma do vetor de pesos,

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2, \quad (15)$$

sujeita às restrições de separabilidade:

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, N, \quad (16)$$

em que $y_i \in \{-1, +1\}$ são os rótulos das amostras. Essa formulação pressupõe dados linearmente separáveis. Em tarefas reais, como o reconhecimento emocional, isso raramente ocorre. Para lidar com sobreposições entre classes, introduzem-se variáveis de relaxamento ξ_i , que permitem violações controladas:

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad (17)$$

resultando na forma de otimização mais geral:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i, \quad (18)$$

onde o parâmetro C controla o balanço entre maximizar a margem e penalizar erros de classificação.

Função Dual e Vetores de Suporte

A solução prática da SVM é obtida por meio da formulação dual, derivada via multiplicadores de Lagrange. Essa forma revela explicitamente a dependência do modelo apenas das amostras mais relevantes, os vetores de suporte. A função dual é dada por:

$$\max_{\alpha} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \quad (19)$$

sujeita às restrições:

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C. \quad (20)$$

Os coeficientes α_i não nulos identificam precisamente os vetores de suporte, que determinam o hiperplano final. Assim, a função de decisão assume a forma:

$$f(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^N \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right). \quad (21)$$

Essa formulação evidencia que a decisão depende apenas de produtos internos entre amostras, o que motiva a utilização de funções *kernel* em cenários não lineares.

Kernels e Separação Não Linear

Quando as classes apresentam fronteiras complexas, como ocorre com padrões emocionais na fala —, o truque do kernel substitui o produto interno no espaço original por uma função de similaridade:

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \rightarrow K(\mathbf{x}_i, \mathbf{x}_j), \quad (22)$$

permitindo projetar implicitamente os dados para um espaço de maior dimensão sem custo computacional explícito. Entre as opções existentes, o *kernel* RBF é o mais usado em SER por sua capacidade de modelar fronteiras altamente não lineares. Ele é definido por:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp \left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right), \quad (23)$$

onde γ controla a complexidade da superfície de decisão: valores maiores produzem fronteiras mais flexíveis, e valores menores geram limites mais suaves.

SVM Multiclasse

A SVM é originalmente binária. Para lidar com múltiplas emoções, utilizam-se estratégias de decomposição, como:

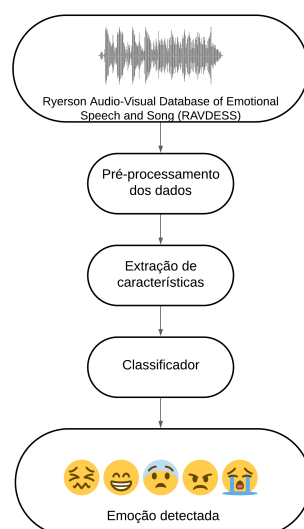
- **One-vs-One (OvO)**, treina-se um classificador para cada par de classes;
- **One-vs-Rest (OvR)**, compara-se cada classe contra todas as demais.

Neste trabalho adota-se a estratégia OvO, que é amplamente relatada na literatura como adequada para emoções, pois divide o problema em subespaços mais simples e melhora a sensibilidade a fronteiras específicas (YAN et al., 2020).

3. METODOLOGIA

Este capítulo descreve de forma detalhada o conjunto de procedimentos adotados para o desenvolvimento do sistema de reconhecimento de emoções na fala. São apresentadas as etapas que compõem o processo metodológico (Figura 1), desde a seleção e preparação da base de dados até a extração e seleção das características acústicas, seguidas pela configuração e avaliação do classificador SVM. Cada fase foi estruturada de modo a garantir a reprodutibilidade dos experimentos e a confiabilidade dos resultados obtidos.

Figura 1 – Fluxo metodológico adotado para o reconhecimento de emoções na fala.



Fonte: Elaborado pelo autor.

3.1 PRÉ-PROCESSAMENTO

A etapa de pré-processamento teve como finalidade preparar os sinais de voz para a extração das características acústicas, assegurando que todos os arquivos apresentassem condições ideais de análise e comparabilidade. Essa fase é essencial para garantir a integridade dos dados, reduzir ruídos indesejados e manter a coerência estatística entre as amostras.

Inicialmente, todos os arquivos de áudio provenientes da base RAVDESS foram convertidos para formato monofônico e ajustados à mesma taxa de amostragem de 48 kHz, preservando a resolução de 16 bits. Essa padronização elimina discrepâncias técnicas entre os sinais, evitando distorções durante o cálculo de medidas espectrais e temporais.

Em seguida, cada gravação foi segmentada em quadros de 25 milissegundos, com sobreposição de 50% entre quadros consecutivos, aplicando-se uma janela de *Hamming*. Essa configuração é comumente utilizada na literatura por garantir que cada segmento de áudio mantenha relativa estacionariedade. A sobreposição parcial permite suavizar as transições entre

quadros e evitar perdas de informação nas bordas dos segmentos, preservando as variações temporais sutis que caracterizam as emoções humanas.

Após a segmentação, foi aplicada uma normalização de amplitude. Para tal, em cada sinal, individualmente, foram detectados os pontos de máximo e mínimo, ajustando a amplitude da forma de onda para a escala $[0, 1]$. Esse procedimento assegura que discrepâncias no ganho de gravação sejam atenuadas, sem, contudo, eliminar a variabilidade na dinâmica da intensidade vocal, a qual permanece preservada na relação energética entre os quadros do sinal. Durante o processamento de extração, valores ausentes (NaN), tipicamente resultantes de quadros sem vozeamento ou instabilidades em algoritmos de frequência fundamental, foram identificados e tratados. Diferentemente de abordagens de preenchimento local, optou-se pela substituição dos valores ausentes pela média da respectiva característica acústica em todo o conjunto de dados. Essa escolha metodológica buscou garantir a integridade das matrizes de atributos para o classificador SVM, minimizando o impacto de falhas pontuais de amostragem na convergência do modelo, especialmente em medidas sensíveis como *jitter* e HNR. Observou-se que a incidência de valores ausentes concentrou-se majoritariamente em medidas de qualidade vocal e frequência fundamental durante trechos de silêncio ou baixa sonoridade, sendo devidamente estabilizada pelo método de substituição adotado.

Por fim, os sinais segmentados e normalizados foram organizados em estruturas tabulares, prontas para a etapa de extração das características acústicas. Essa preparação garante que o modelo opere sobre dados consistentes, com variações atribuíveis exclusivamente aos padrões emocionais presentes na voz, e não a artefatos de gravação ou inconsistências de processamento.

3.2 EXTRAÇÃO DAS CARACTERÍSTICAS ACÚSTICAS

A etapa de extração das características acústicas foi desenvolvida em linguagem Python, empregando-se principalmente as bibliotecas *Parselmouth*, *NumPy* e *SciPy*. O processo foi automatizado por meio de *scripts* específicos, como *prepare_dataset.py* e *prepare_global_metadata.py*, responsáveis por percorrer todas as amostras da base RAVDESS, realizar a segmentação temporal e registrar as medidas extraídas em estruturas tabulares no formato `.csv`. Esse fluxo garante rastreabilidade e reprodutibilidade em todas as etapas do experimento.

O conjunto inicial de parâmetros acústicos totalizou 21 características, organizadas em três grupos principais de acordo com seu papel na produção e percepção da fala:

- **Características prosódicas:** Frequência Fundamental (F_0), RMSE e Taxa de Cruzamento por Zero (TCZ);
- **Características de qualidade vocal:** *Jitter*, *Shimmer*, Razão Harmônico-Ruído (HNR) e Proeminência do Pico Cepstral Suavizada (CPPS).

- **Características espectrais e cepstrais:** Os 13 primeiros coeficientes cepstrais em frequência Mel (MFCC₁ – MFCC₁₃) e Centroide Espectral.

Cada uma dessas características foi extraída individualmente para todos os quadros de 25ms com sobreposição de 50%, resultando em vetores numéricos que descrevem o comportamento acústico e temporal de cada amostra de voz. Especificamente em relação aos atributos cepstrais, foram utilizados os 13 primeiros coeficientes MFCC, extraídos por janela de 25ms e posteriormente agregados estatisticamente para cada arquivo de áudio. Essa abordagem permite capturar de forma sintética a assinatura espectral de cada emoção, contribuindo de maneira significativa para a análise de separabilidade estatística e para a tarefa de classificação emocional supervisionada.

A escolha dessas 21 características foi fundamentada em sua ampla adoção e relevância comprovada na literatura sobre SER. Estudos clássicos e recentes demonstram que essas métricas capturam de forma abrangente os principais aspectos acústicos modulados pela emoção, englobando desde variações prosódicas globais até flutuações microtemporais da fonte glótica (SCHULLER; STEIDL; BATLINER, 2009; AKÇAY; OĞUZ, 2020; YAN et al., 2020; MASHHADI; OSEI-BONSU, 2023).

Em síntese, a seleção das 21 características busca equilibrar abrangência descritiva e eficiência computacional. Ao integrar parâmetros prosódicos, espectrais e de qualidade vocal, o conjunto captura múltiplas dimensões do sinal de voz, fonatória, articulatória e perceptual.

3.3 SELEÇÃO DE CARACTERÍSTICAS

Após a extração completa das 21 características acústicas, foi conduzida uma etapa de seleção de atributos com o objetivo de identificar as variáveis mais relevantes para a distinção entre as diferentes emoções. Essa fase é fundamental em sistemas de SER, pois reduz a dimensionalidade do conjunto de dados, elimina redundâncias e aumenta a interpretabilidade dos resultados, preservando apenas as informações com maior poder discriminativo (AKÇAY; OĞUZ, 2020).

O método adotado para essa seleção foi o *Fisher Score*, implementado no script *fisher_top10.py*. Essa métrica baseia-se na relação entre a variância interclasses e a variância intraclasse, avaliando a capacidade de cada característica individual em separar as categorias emocionais. Em termos matemáticos, o escore de *Fisher* para uma característica x_i é calculado conforme a Equação 24:

$$F_i = \frac{\sum_{c=1}^C n_c (\mu_{c,i} - \mu_i)^2}{\sum_{c=1}^C n_c \sigma_{c,i}^2} \quad (24)$$

em que C representa o número de classes (emoções), n_c é o número de amostras da classe c , $\mu_{c,i}$

e $\sigma_{c,i}^2$ são, respectivamente, a média e a variância da característica x_i dentro da classe c , e μ_i é a média global da característica em todo o conjunto de dados. Assim, quanto maior o valor de F_i , maior a separabilidade entre as classes para aquela característica.

A partir da aplicação do método, as dez características com maiores valores de *Fisher* foram selecionadas, compondo o subconjunto denominado *Top-10*. Ressalta-se que a aplicação do *Fisher Score* foi realizada de forma global sobre o *corpus* em nível de áudio (*utterance-level*), funcionando como uma etapa de análise exploratória e redução de dimensionalidade prévia à fase de classificação. Essa abordagem permitiu identificar a assinatura acústica predominante da base RAVDESS e compreender quais grupos de atributos possuem maior variância interclasses. Embora a seleção global possa introduzir viés de otimismo nos resultados, optou-se por essa estratégia para garantir que o classificador SVM operasse sobre um espaço de características fisicamente interpretável e estatisticamente consistente para todas as partições do experimento.

A utilização do *Fisher Score* apresenta vantagens significativas, como simplicidade computacional, robustez em relação a *outliers* e independência de parâmetros. Além disso, o método não requer ajustes de modelo nem supõe distribuições específicas dos dados, o que o torna particularmente adequado para o tamanho da base de dados utilizada.

A adoção do conjunto *Top-10* teve como objetivo principal equilibrar desempenho e interpretabilidade. A redução dimensional facilita a análise das contribuições individuais de cada atributo, reduz o tempo de treinamento e minimiza a possibilidade de sobreajuste (*overfitting*). Essa estratégia é amplamente relatada na literatura recente de reconhecimento emocional, especialmente em fluxos que priorizam classificadores clássicos, como o SVM. Dessa forma, o processo consolidou um conjunto otimizado de parâmetros, representando as dimensões mais expressivas do sinal vocal para o treinamento do modelo final (WANI et al., 2021; LOPE; GRAÑA, 2023).

3.4 CLASSIFICAÇÃO COM SVM

A etapa de classificação das emoções foi conduzida utilizando exclusivamente o algoritmo SVM, devido à sua comprovada robustez estatística, estabilidade e ampla adoção em tarefas de reconhecimento de emoções na fala.

Os experimentos foram realizados em dois níveis de granularidade: *frame-level* e *utterance-level*. No primeiro, cada quadro de 25ms é tratado como uma amostra independente; no segundo, todos os quadros de um mesmo áudio são agregados em uma única amostra. Embora o cenário *frame-level* forneça maior detalhamento temporal, ele produz uma quantidade de amostras adicionais, aumentando drasticamente o custo computacional. Diante disso, a seleção de características pelo *Fisher Score* foi aplicada apenas no nível *utterance-level*, que apresenta maior viabilidade prática e melhor aproximação da percepção humana. Acrescenta-se que, a partir dessa etapa, também foi feita a classificação com o subconjunto reduzido Top-10,

permitindo comparar seu desempenho com o conjunto completo de atributos.

Para o treinamento do modelo, utilizou-se a estratégia *one-vs-one* (OvO) com o *kernel* radial de base gaussiana (RBF). Os hiperparâmetros foram definidos empiricamente, a partir de testes preliminares de avaliação de desempenho, fixando-se os valores em $C = 10.0$ e $\gamma = \text{scale}$, conforme configurado nos *scripts* de execução. A escolha do valor de C visou equilibrar a imposição de uma margem de separação larga com a penalização de classificações incorretas no conjunto de treino, fator essencial dada a sobreposição natural das classes emocionais acústicas. O ajuste de γ garantiu que a influência de cada amostra fosse proporcional à variância dos atributos. Adicionalmente, adotou-se uma tolerância de convergência de 1×10^{-3} como critério de parada do otimizador numérico.

A avaliação do modelo foi conduzida por meio de Validação Cruzada Estratificada de 5 partições (*5-fold Stratified Cross-Validation*), superando abordagens simples de divisão (como *hold-out* 80/20) ao garantir que todas as amostras fossem testadas e que a proporção das emoções fosse mantida constante em cada *fold*. Todo o processamento foi implementado na linguagem Python, utilizando a biblioteca *scikit-learn*, e executado em uma estação de trabalho Acer Aspire A515-52G, equipada com processador Intel Core i5 (1.60 GHz), 16 GB de memória RAM e sistema operacional Windows 11.

Por uma questão de honestidade metodológica e transparência experimental, ressalta-se que o particionamento estratificado da validação cruzada foi baseado exclusivamente nos rótulos emocionais (*speaker-dependent*). Consequentemente, amostras de fala de um mesmo ator podem figurar tanto no conjunto de treinamento quanto no conjunto de teste. Da mesma forma, no cenário *frame-level*, os quadros oriundos de um mesmo arquivo de áudio foram distribuídos aleatoriamente entre os *folds*. Essa decisão reconhece que a dependência amostral eleva artificialmente as métricas preditivas devido à adaptação do modelo à assinatura acústica do locutor.

3.5 AVALIAÇÃO DE DESEMPENHO

A avaliação do desempenho do classificador foi conduzida a partir de métricas amplamente utilizadas na literatura de aprendizado de máquina e de reconhecimento de emoções na fala, permitindo uma análise quantitativa rigorosa e comparável a estudos correlatos. As métricas selecionadas oferecem uma visão abrangente do comportamento do modelo, contemplando tanto o desempenho global quanto o equilíbrio entre as predições corretas e incorretas.

As medidas adotadas foram as seguintes (YAN et al., 2020):

- **Acurácia (ACC):** representa a proporção total de amostras corretamente classificadas, conforme a Equação 25. É uma métrica global que fornece uma visão geral do desempenho

do modelo.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

- **Precisão (P):** quantifica a proporção de previsões positivas que foram de fato corretas, indicando a confiabilidade do modelo ao identificar uma emoção específica.

$$P = \frac{TP}{TP + FP} \quad (26)$$

- **Revocação (Recall - R):** também denominada sensibilidade, expressa a proporção de instâncias positivas corretamente reconhecidas pelo modelo.

$$R = \frac{TP}{TP + FN} \quad (27)$$

- **F1-score:** corresponde à média harmônica entre precisão e revocação, sendo particularmente útil em cenários de desequilíbrio entre classes ou dificuldade de separação estrita.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (28)$$

Essas métricas foram inicialmente calculadas de forma individual para cada uma das classes emocionais. Para consolidar o resultado em cada partição (*fold*) da validação cruzada, obteve-se a média aritmética simples dos valores das oito classes, garantindo que todas as emoções tivessem o mesmo peso na avaliação final, independentemente de eventuais pequenas variações no número de amostras. Por fim, os valores reportados neste estudo correspondem à média dessas estimativas ao longo das 5 partições, o que assegura uma avaliação de desempenho estável e representativa da capacidade de generalização do modelo.

Todos os resultados gerados durante a execução dos experimentos foram automaticamente armazenados em arquivos .csv, acompanhados de gráficos de desempenho, matrizes de confusão e relatórios textuais consolidados. Essa estrutura de saída padronizada possibilitou não apenas a análise quantitativa, mas também a avaliação visual do comportamento do modelo, permitindo identificar padrões de confusão entre classes e o impacto da redução de características sobre a separabilidade emocional.

A comparação entre os três cenários experimentais, *frame-level*, *utterance-level* e *Top-10*, permitiu observar a influência direta da granularidade temporal e da seleção de atributos no desempenho global. Enquanto o modelo *frame-level* fornece uma análise detalhada das variações instantâneas do sinal, o modelo *utterance-level* captura informações agregadas mais próximas da percepção humana. Já o cenário *Top-10* evidencia o efeito da seleção de características no equilíbrio entre simplicidade e eficiência, demonstrando que é possível mitigar o ruído de atributos menos informativos e alcançar o melhor desempenho preditivo do estudo com menor custo computacional.

Por fim, todas as métricas e figuras de desempenho foram documentadas de forma automatizada para posterior análise comparativa, assegurando a reprodutibilidade dos resultados e a transparência metodológica do processo experimental.

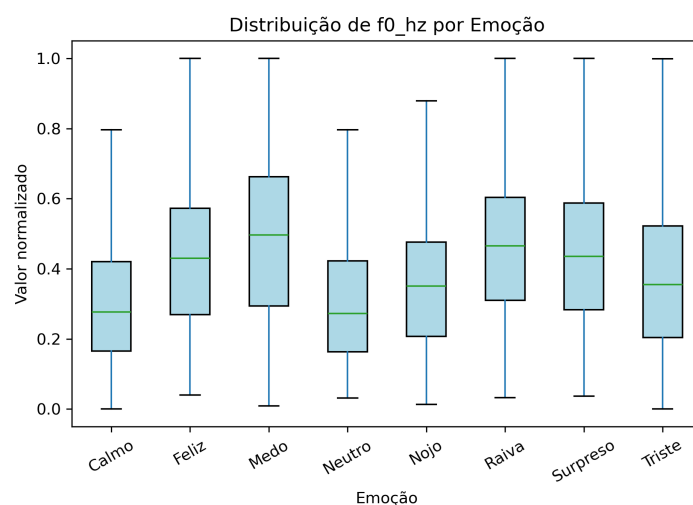
4. RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos a partir da análise das 21 características acústicas extraídas da base RAVDESS. As distribuições estatísticas são exibidas por meio de diagramas de caixa (*boxplots*), permitindo observar o comportamento das variáveis entre as diferentes classes emocionais. As análises são organizadas em três grupos principais: características prosódicas, espectrais/cepstrais e de qualidade vocal. Em seguida, são apresentados os resultados de classificação obtidos por meio do classificador *Support Vector Machine* (SVM) nos cenários *frame-level*, *utterance-level* e com o subconjunto reduzido de dez características selecionadas pelo Fisher Score.

4.1 CARACTERÍSTICAS PROSÓDICAS

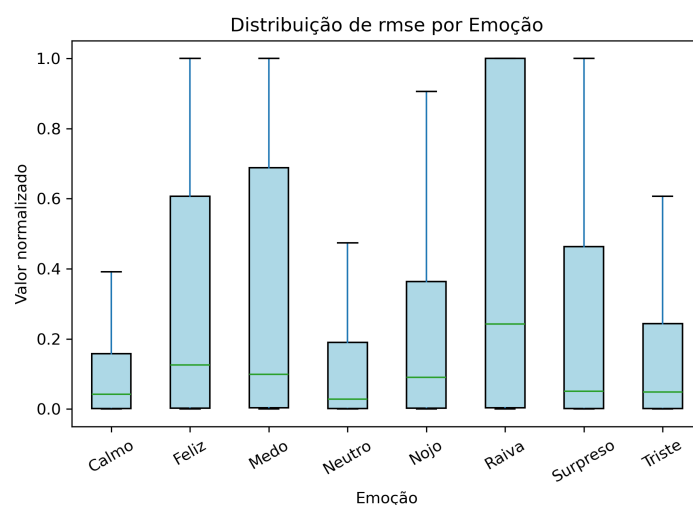
As características prosódicas descrevem o contorno melódico, o ritmo e a intensidade do discurso, sendo fortemente associadas ao grau de excitação emocional do locutor. As principais variáveis analisadas foram a frequência fundamental (F_0), a energia (RMSE) e a taxa de cruzamento por zero (TCZ).

Figura 2 – Distribuição da frequência fundamental (F_0) por emoção.



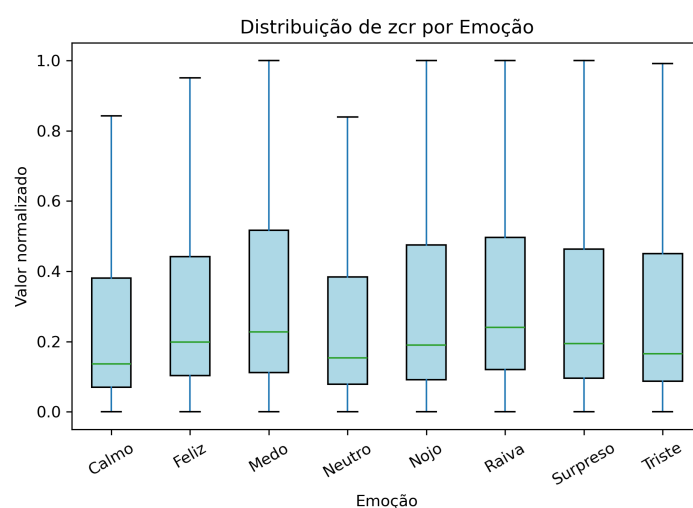
Fonte: Elaborado pelo autor.

A Figura 2 evidencia o comportamento da frequência fundamental. Emoções de alta excitação, como o medo, a surpresa e a raiva, apresentam as maiores medianas de F_0 , o que reflete uma maior tensão laríngea e grande variabilidade melódica na expressão dessas emoções. Por outro lado, emoções como calma, neutralidade e tristeza exibem valores medianos reduzidos e limites superiores mais baixos, um comportamento acústico coerente com estados de baixa energia e menor esforço fonatório.

Figura 3 – Distribuição da energia (RMSE) por emoção.

Fonte: Elaborado pelo autor.

A análise da energia média (RMSE), ilustrada na Figura 3, reforça a separação por níveis de ativação. Emoções altamente ativas, notadamente a raiva e a felicidade, concentram distribuições muito mais dispersas em direção a valores altos, com seus quartis superiores alcançando o topo da escala normalizada. Em contrapartida, as emoções calma, neutro e tristeza mantêm a maior parte de sua distribuição espremida na parte inferior do gráfico, evidenciando menor intensidade sonora e projeção vocal restrita.

Figura 4 – Distribuição da taxa de cruzamento por zero (TCZ) por emoção.

Fonte: Elaborado pelo autor.

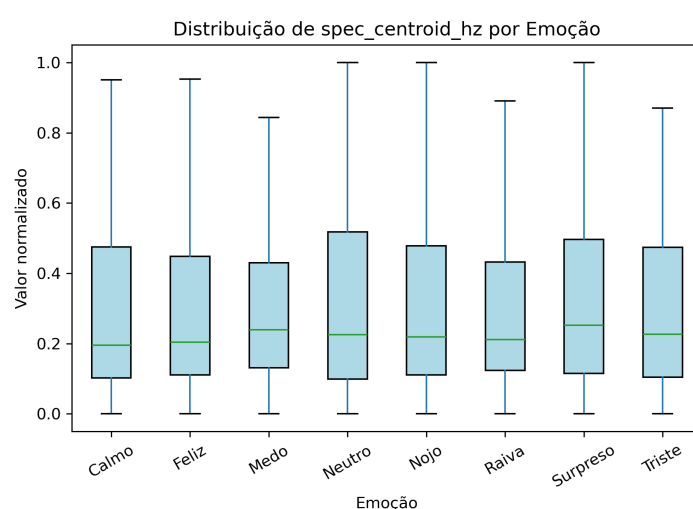
A Figura 4 demonstra que a taxa de cruzamento por zero (TCZ) apresenta valores medianos e dispersões mais altas para as emoções raiva e medo, indicando uma maior instabilidade na

forma de onda e forte presença de componentes de alta frequência (típicos de falas mais ásperas ou fricativas). A tristeza e o estado neutro, por sua vez, exibem menor variabilidade temporal e medianas mais baixas.

4.2 CARACTERÍSTICAS ESPECTRAIS E CEPSTRAIS

As medidas espectrais e cepstrais descrevem a forma e a distribuição de energia ao longo das frequências do sinal de fala. Entre essas características, destacam-se o centroide espectral e os coeficientes cepstrais em frequência Mel (MFCCs).

Figura 5 – Distribuição do centroide espectral por emoção.



Fonte: Elaborado pelo autor.

A Figura 5 revela o comportamento do centroide espectral. Visualmente, nota-se que emoções como surpresa, medo e o estado neutro tendem a apresentar medianas mais elevadas, além de grande dispersão, o que sugere uma concentração de energia voltada para frequências mais agudas nessas amostras específicas da base RAVDESS. Em contrapartida, emoções como calma e, surpreendentemente, a raiva, exibem medianas ligeiramente inferiores neste atributo. Esse comportamento evidencia que o brilho espectral pode variar fortemente dependendo do estilo de atuação do falante.

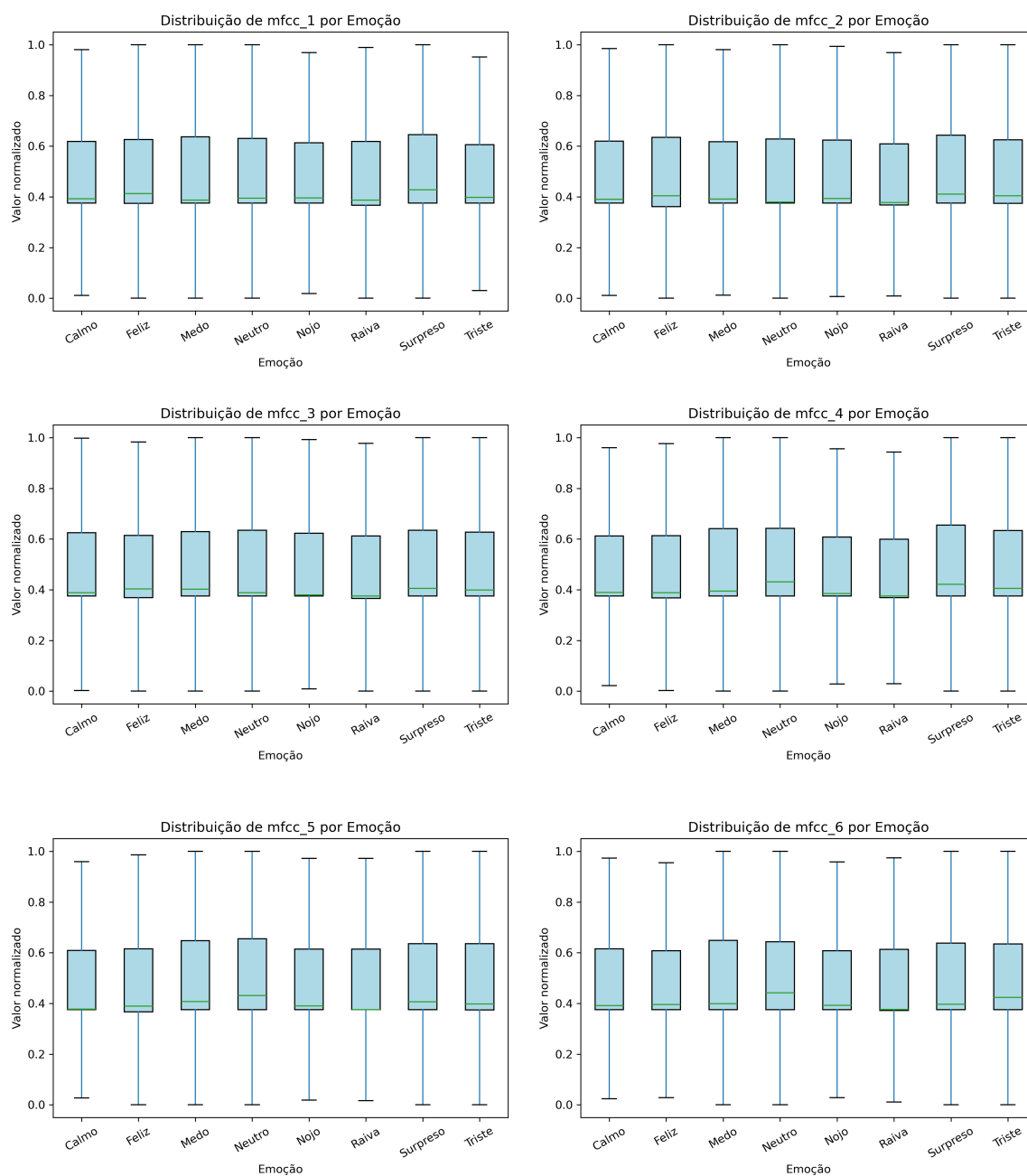
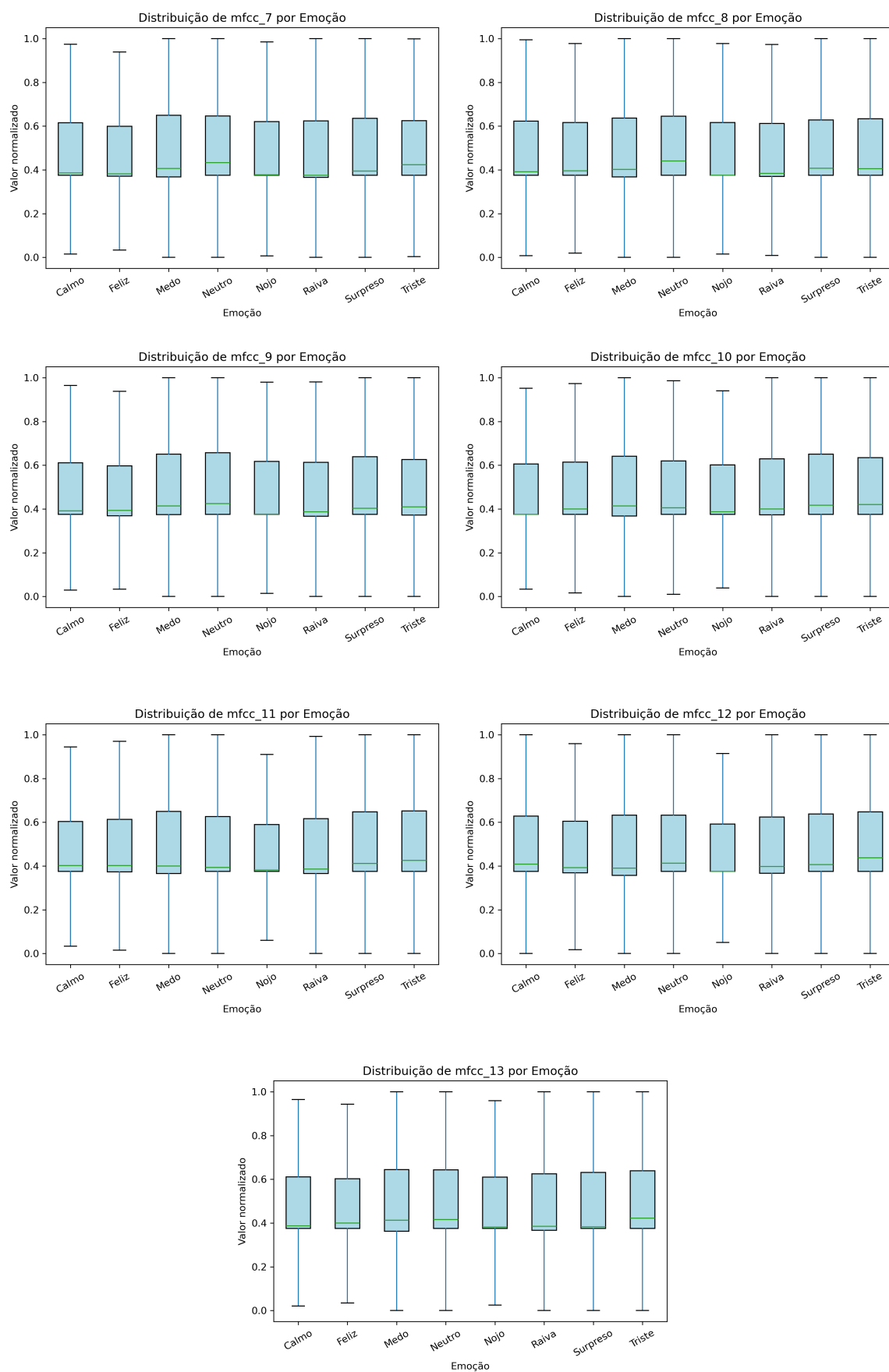
Figura 6 – Distribuição dos coeficientes MFCC₁ a MFCC₆ por emoção.**Fonte: Elaborado pelo autor.**

Figura 7 – Distribuição dos coeficientes MFCC₇ a MFCC₁₃ por emoção.**Fonte: Elaborado pelo autor.**

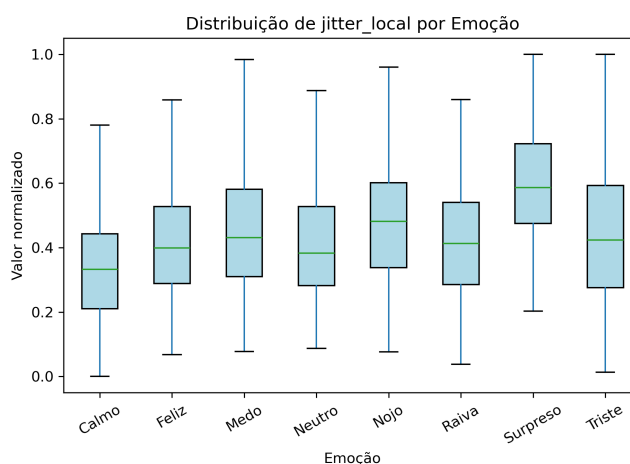
As Figuras 6 e 7 apresentam a distribuição dos 13 coeficientes MFCC. Observa-se que os coeficientes de ordem mais baixa (MFCC1 a MFCC4), responsáveis por capturar a estrutura espectral mais global (a envoltória do espectro), exibem distinções perceptíveis nas medianas entre certas classes. A surpresa, por exemplo, destaca-se com a maior mediana no MFCC1, enquanto a calma e a raiva mantêm valores mais baixos.

Os coeficientes intermediários e superiores (MFCC7 ao 13) mostram medianas muito niveladas, revelando variações mais sutis. O aspecto mais importante evidenciado por esses gráficos é a alta sobreposição (overlap) das caixas entre todas as emoções. Essa sobreposição visual justifica por que a análise isolada de um único MFCC é insuficiente, sendo imprescindível a combinação desses coeficientes em um espaço multidimensional (como feito pelo classificador SVM) para encontrar os hiperplanos capazes de separar características de timbre e articulação.

4.3 CARACTERÍSTICAS DE QUALIDADE VOCAL

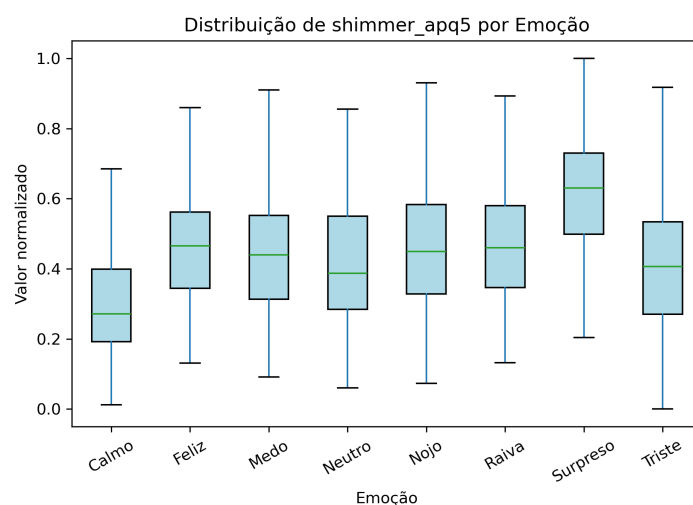
As características de qualidade vocal refletem irregularidades na periodicidade e na amplitude da vibração glótica, abrangendo medidas como *jitter*, *shimmer*, Razão Harmônico-Ruído (HNR) e Proeminência do Pico Cepstral Suavizada (CPPS).

Figura 8 – Distribuição do Jitter (%) por emoção.



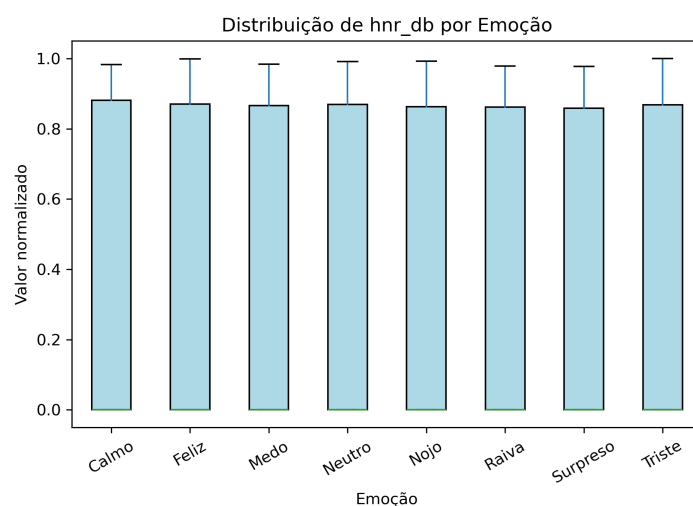
Fonte: Elaborado pelo autor.

A Figura 8 ilustra a variação da frequência glótica (*jitter*). O grande destaque visual é a emoção surpresa, que exibe, de forma isolada, a maior mediana e a maior dispersão de dados. Além dela, o medo e a raiva também apresentam altos níveis de *jitter*, o que caracteriza instabilidade glótica típica de vozes tensas ou tremidas. Em contraste, estados mais contidos, como a calma e o neutro, apresentam as menores medianas, refletindo um controle fonatório superior e maior estabilidade vibratória.

Figura 9 – Distribuição do Shimmer (APQ5) por emoção.

Fonte: Elaborado pelo autor.

O comportamento do *shimmer* (variação de amplitude), ilustrado na Figura 9, segue uma tendência semelhante à do *jitter*. Novamente, a surpresa se destaca com a maior variação de intensidade. Contudo, emoções como felicidade e raiva também apresentam medianas consideravelmente elevadas, sugerindo que a euforia e a agressividade induzem fortes oscilações de amplitude ciclo a ciclo durante a fala. Calma e neutro, mais uma vez, confirmam sua estabilidade acústica exibindo as caixas mais compactas e as menores medianas.

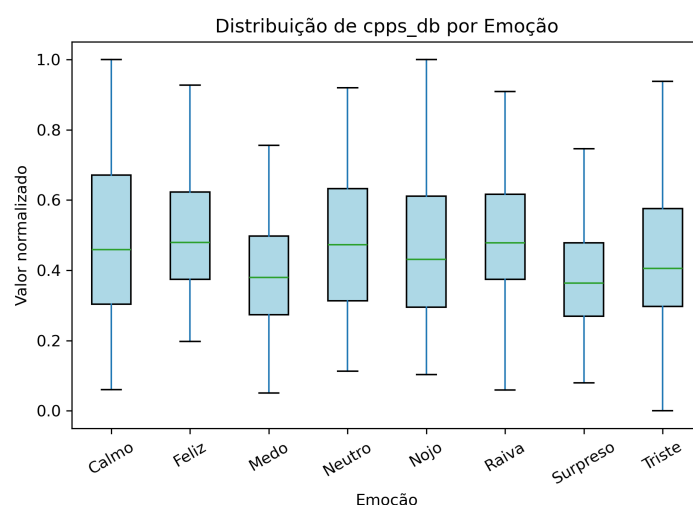
Figura 10 – Distribuição da razão harmônico-ruído (HNR) por emoção.

Fonte: Elaborado pelo autor.

A distribuição da Razão Harmônico-Ruído (HNR), apresentada na Figura 10, demanda uma leitura crítica. Visualmente, todas as classes emocionais apresentam distribuições idênticas,

com a mediana colada em zero e os limites da caixa (terceiro quartil) estendendo-se por quase toda a escala. Esse achatamento gráfico evidencia que a normalização para a escala $[0, 1]$ (também conhecida como Min-Max), aplicada aos valores de todos os atributos para fins de padronização visual nos gráficos e alimentação do classificador, foi fortemente influenciada por valores discrepantes (*outliers* extremos) no caso específico do HNR. Consequentemente, isso comprimiu a variância da maior parte dos dados para próximo de zero, ocultando as reais diferenças entre as classes na representação visual.

Figura 11 – Distribuição da proeminência do pico cepstral suavizada (CPPS) por emoção.



Fonte: Elaborado pelo autor.

Por fim, a proeminência do pico cepstral (CPPS), exposta na Figura 11, mede a regularidade harmônica. Observa-se que emoções de alta projeção vocal, como felicidade e raiva, além dos estados neutro e calmo, apresentam as medianas mais altas, o que caracteriza vozes mais sonoras e com periodicidade bem definida (pouco ruído de fricção). Por outro lado, o medo, a surpresa e o nojo exibem os menores valores medianos, indicando perda de regularidade harmônica, aumento do escape de ar ou ruído glotal exacerbado durante a atuação dessas emoções.

Esses resultados reforçam a relevância das medidas de qualidade vocal na caracterização emocional, evidenciando como diferentes estados afetivos modulam a estabilidade fonatória, o controle da laringe e a clareza harmônica do sinal de voz.

4.4 SELEÇÃO DE CARACTERÍSTICAS: O CONJUNTO TOP-10 (*FISHER SCORE*)

A análise individual (seções 4.1 a 4.3) evidenciou que as 21 características acústicas extraídas apresentam diferentes graus de sobreposição e variabilidade entre as classes emocionais. Para otimizar o desempenho preditivo e mitigar o risco de *overfitting* causado por redundâncias,

aplicou-se o método *Fisher Score* em nível de áudio. Conforme detalhado na metodologia, essa técnica ranqueia os atributos com base na razão entre a variância interclasses e a variância intraclasse, priorizando as métricas que melhor separam as emoções e que apresentam maior coesão interna.

A partir do cálculo do *Fisher Score* para o vetor completo de características, selecionou-se o subconjunto com as dez maiores pontuações, denominado Top-10. As métricas que compõem este cenário otimizado, ordenadas por sua importância discriminativa, foram:

1. Energia (RMSE) – *Fisher Score*: 0.508
2. Shimmer (APQ5) – *Fisher Score*: 0.210
3. Frequência Fundamental (F_0) – *Fisher Score*: 0.164
4. Taxa de Cruzamento por Zero (TCZ) – *Fisher Score*: 0.149
5. Razão Harmônico-Ruído (HNR) – *Fisher Score*: 0.137
6. Jitter (Local) – *Fisher Score*: 0.120
7. Proeminência do Pico Cepstral Suavizada (CPPS) – *Fisher Score*: 0.063
8. Centroide Espectral – *Fisher Score*: 0.023
9. $MFCC_4$ – *Fisher Score*: 0.008
10. $MFCC_{12}$ – *Fisher Score*: 0.007

O ranqueamento estabelecido pelo *Fisher Score* revela padrões sobre o comportamento acústico das emoções na base RAVDESS. Observa-se a absoluta predominância das características prosódicas e de qualidade vocal, que ocuparam as sete primeiras posições do ranking. A métrica de energia (RMSE) obteve uma pontuação superior a 0.50, mais que o dobro da segunda colocada. Este resultado comprova matematicamente a observação visual do *boxplot* (Figura 3), consolidando a intensidade vocal como o marcador acústico isolado mais forte para separar classes emocionais em termos de ativação (como diferenciar a raiva da tristeza). Em seguida, métricas de instabilidade glótica e tensão, como *Shimmer* e F_0 , garantiram o segundo e terceiro lugares do ranking, respectivamente, confirmando a importância das alterações fisiológicas involuntárias na fala emocional.

Um achado particularmente interessante ocorre com a Razão Harmônico-Ruído (HNR). Embora a análise visual do seu *boxplot* (Figura 10) sugerisse uma sobreposição quase total entre as classes devido ao achatamento provocado por normalização, o *Fisher Score* é uma métrica baseada em razões de variância, sendo imune a essas distorções de escala. O algoritmo

identificou boa separabilidade estatística interna no HNR, conferindo-lhe a quinta posição geral. Isso ressalta a importância de aliar a inspeção visual a métodos de seleção estatística. Dessa forma, o HNR serve como um caso de estudo sobre as limitações da normalização linear em atributos com distribuições de cauda longa, reforçando que a separabilidade estatística capturada pelo *Fisher Score* (onde o HNR ocupa a 5ª posição) é mais fidedigna à realidade dos dados do que a representação visual normalizada.

Por outro lado, o grupo de características espectrais (especialmente os MFCCs) apresentou baixo poder discriminativo isolado. Apenas o $MFCC_4$ e o $MFCC_{12}$ entraram no corte do Top-10, ambos com pontuações baixas (inferiores a 0.01). Este fenômeno corrobora a alta sobreposição de caixas observada nas Figuras 6 e 7.

A escolha por manter as dez características mais bem pontuadas (Top-10), em detrimento de um corte mais restritivo (como as cinco melhores, por exemplo), justifica-se pelo interesse de preservar a representatividade das três dimensões acústicas analisadas. Um corte excessivo concentraria o modelo preditivo unicamente em parâmetros prosódicos e de intensidade laríngea, excluindo por completo atributos espectrais complementares, como o Centroide Espectral e os coeficientes MFCC, além de métricas de instabilidade fina, como o Jitter. Considerando a complexidade da separação entre oito classes emocionais distintas, o subconjunto de dez variáveis configurou-se como um ponto ótimo empírico. Ao isolar essas dez características, o sistema reduz drasticamente a dimensionalidade do problema descartando os coeficientes espectrais ruidosos ou sobrepostos, garantindo que o classificador SVM opere apenas sobre o núcleo de informações acústicas mais expressivas, com informações multidimensionais suficientes para estabelecer as fronteiras de decisão, sem incorrer no custo computacional ou no ruído gerado pelas 21 variáveis originais.

4.5 DESEMPENHO DO CLASSIFICADOR SVM

Após a seleção criteriosa das características via *Fisher Score*, avaliou-se a capacidade discriminativa do classificador SVM. Foram investigados três cenários experimentais, permitindo comparar o desempenho do modelo em diferentes níveis de representação temporal (*frame-level* e *utterance-level*) e de dimensionalidade (conjunto completo versus Top-10).

4.5.1 Resumo das Métricas Globais

Tabela 3 – Métricas médias (5-fold) dos três cenários SVM.

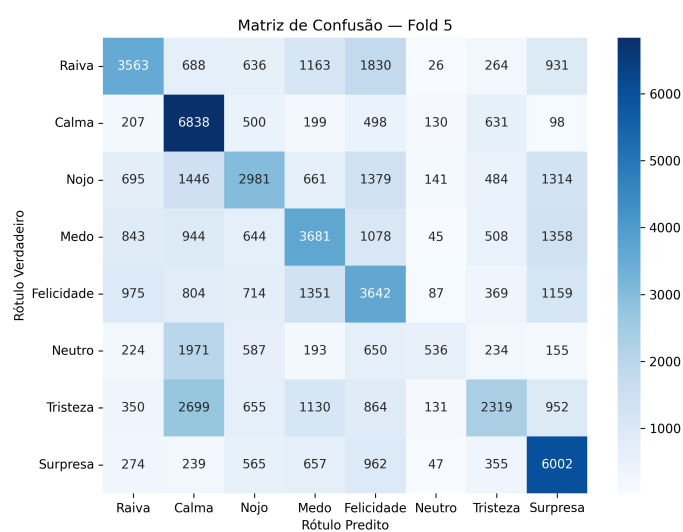
Cenário	Acurácia	Precisão	Sensibilidade	F1-score
<i>Frame-Level</i>	0.4344	0.4378	0.4146	0.4021
<i>Utterance-Level</i>	0.4125	0.3994	0.4038	0.3976
Top-10 (Fisher)	0.4403	0.4284	0.4327	0.4262

A Tabela 3 consolida as métricas de desempenho para os três cenários avaliados. O cenário *frame-level*, que processa cada quadro de 25 ms de forma independente, alcançou uma acurácia média de 43,44%. Embora essa granularidade capture variações rápidas do sinal de voz, o excesso de flutuações locais em quadros específicos prejudica a estabilidade da predição global. O cenário *utterance-level*, que agrega as 21 características de todo o áudio por meio de estatísticas descritivas, obteve 41,25% de acurácia. O melhor desempenho, no entanto, foi alcançado pelo cenário Top-10 (*Fisher*). Utilizando apenas as dez características com maior poder discriminativo estabelecidas na Seção 4.4 (lideradas por RMSE, Shimmer e F_0), o modelo alcançou o melhor desempenho global, com acurácia de 44,03% e F1-score de 0,4262. Esse resultado comprova que a redução de dimensionalidade eliminou atributos ruidosos ou redundantes (como a maioria dos MFCCs isolados) que estavam sobrecarregando o hiperplano do SVM. Considerando que o problema envolve 8 classes emocionais (onde o acerto totalmente aleatório seria de apenas 12,5%), as taxas em torno de 44% obtidas por um classificador clássico demonstram a consistência e a validade das características acústicas selecionadas dentro de um cenário de classificação clássica.

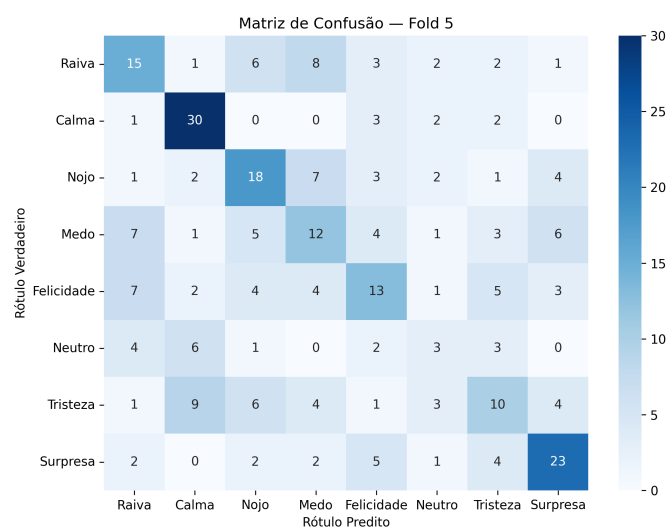
4.5.2 Matrizes de Confusão

As Figuras 12, 13 e 14 apresentam as matrizes de confusão referentes aos três cenários avaliados. Essas representações permitem visualizar o padrão de acertos e erros do classificador, destacando quais emoções tendem a ser corretamente identificadas e quais apresentam maior sobreposição acústica. Observam-se, em especial, confusões recorrentes entre emoções de ativação semelhante, evidenciando limitações inerentes à separabilidade entre certas classes.

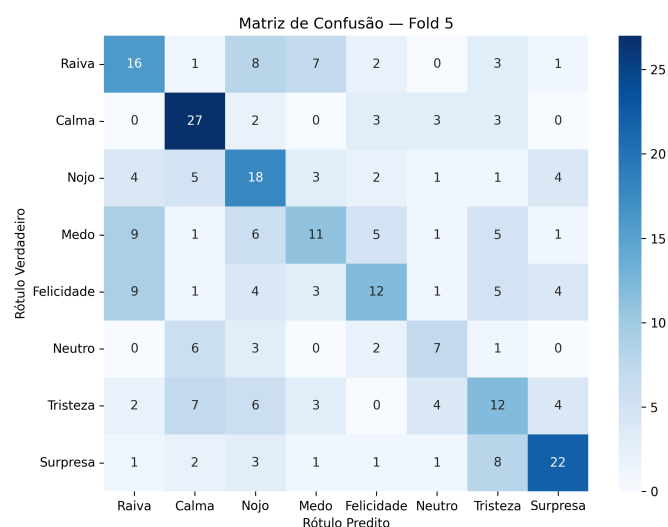
Figura 12 – Matriz de confusão, cenário *frame-level*.



Fonte: Elaborado pelo autor.

Figura 13 – Matriz de confusão, cenário *utterance-level*.

Fonte: Elaborado pelo autor.

Figura 14 – Matriz de confusão, cenário *Top-10*.

Fonte: Elaborado pelo autor.

As Figuras 12, 13 e 14 apresentam as matrizes de confusão referentes aos três cenários, permitindo visualizar o padrão de acertos e erros do classificador. A análise cruzada dessas matrizes revela comportamentos sobre a separabilidade emocional:

- **Classes de Maior Acerto:** em todos os cenários, as emoções calma e surpresa despontam como as mais bem classificadas. Como visto na análise do *Fisher Score* e nos *boxplots*, a surpresa possui uma assinatura acústica muito isolada em termos de instabilidade glótica (*jitter* e *shimmer*), facilitando a criação de margens de decisão pelo SVM. A calma, por sua extrema estabilidade e baixa energia, também é facilmente isolada.

- **Confusões por Sobreposição de Ativação:** o modelo apresenta dificuldades em separar emoções que compartilham o mesmo nível de energia (alta ativação). A emoção felicidade, por exemplo, é frequentemente classificada de forma incorreta como raiva ou medo (padrão nitidamente visível na Figura 12, com altos índices de falsos preditos nestas classes). Isso ocorre porque as três emoções compartilham valores altos nas métricas que dominaram o modelo preditivo, como F_0 e RMSE.
- **Confusões de Baixa Ativação:** de forma análoga, a emoção tristeza apresenta considerável confusão com as emoções calma ou neutro, visto que todas exibem distribuição acústica espremida na base dos gráficos de energia e prosódia, dificultando a distinção fina pelo classificador.

Essas matrizes evidenciam que os atributos extraídos conseguem separar grupos com base na ativação (alta vs. baixa energia), mas apresentam limitações na separabilidade entre emoções de mesma intensidade.

4.6 DISCUSSÃO GERAL

A análise conjunta das características acústicas e dos experimentos de classificação revelam observações não apenas sobre o potencial da análise acústica, mas também sobre as limitações inerentes aos métodos clássicos de reconhecimento de emoções na fala. Do ponto de vista da extração e seleção de atributos, o estudo demonstrou que as emoções moduladas na base RAVDESS alteram drasticamente a estabilidade fonatória e a energia vocal. O ranqueamento estabelecido pelo *Fisher Score* apontou que a intensidade vocal (RMSE) e os marcadores de instabilidade glótica (*Shimmer* e F_0) apresentam o poder discriminativo para detecção das emoções, sobrepondo-se aos marcadores de timbre (MFCCs), que apresentaram forte sobreposição entre as classes.

Ao cruzar essas observações com os resultados do classificador SVM, é necessário realizar uma análise sobre as métricas de desempenho. O modelo otimizado (cenário Top-10) atingiu o melhor resultado global da pesquisa, com uma acurácia de 44,03%, o cenário reduzido com a eliminação de atributos ruidosos ou redundantes elevou o desempenho e diminuiu o custo computacional em relação aos cenários com o conjunto completo de dados.

Contudo, é imperativo reconhecer que uma acurácia em torno de 44% é modesta quando comparada ao estado da arte em SER, que frequentemente ultrapassa a marca dos 80% ou 90% utilizando redes neurais profundas (*Deep Learning*) e arquiteturas multimodais. Essa limitação de desempenho, contudo, não invalida os achados, mas expõe os limites metodológicos da abordagem clássica adotada:

- *O Desafio da Valência vs. Ativação:* as matrizes de confusão evidenciam que o SVM

consegue agrupar as emoções perfeitamente pelo seu nível de ativação (energia), mas falha em distinguir a sua valência (positividade ou negatividade). O modelo confunde sistematicamente felicidade com raiva ou medo, pois todas compartilham valores altos de RMSE e F_0 .

- *Ausência de Dinâmica Temporal:* características agregadas em nível de áudio resumem as médias e variâncias estatísticas, perdendo a evolução temporal da emoção (como o contorno melódico de uma frase). A baixa relevância dos MFCCs no modelo estático reforça que, para capturar o timbre emocional com precisão, é necessário utilizar derivadas temporais (coeficientes delta e delta-delta) que acompanhem a articulação ao longo do tempo.
- *Interpretabilidade versus Acurácia:* a escolha pelo SVM e por atributos acústicos clássicos priorizou a interpretabilidade do modelo em detrimento da acurácia absoluta.
- *O Impacto da Normalização Individual na Amplitude:* a decisão metodológica de aplicar a normalização de amplitude individualmente para cada sinal (ajustando a forma de onda de cada áudio para a escala de 0 a 1) atenuou as discrepâncias de ganho de gravação, mas, como efeito colateral, suprimiu a percepção de intensidade sonora absoluta entre as diferentes amostras. Em um contexto em que emoções de alta e baixa excitação se diferenciam fortemente pela projeção vocal, essa padronização faz com que sinais intrinsecamente contidos (como o da tristeza) atinjam os mesmos picos de amplitude de sinais originados por esforço vocal severo (como a raiva). Essa perda da intensidade relativa global entre os arquivos limitou a capacidade do modelo de usar a magnitude absoluta como fronteira de decisão, atuando como um forte fator limitante que contribuiu para o teto de acurácia em 44,03%.

Ressalta-se que a análise estatística aqui apresentada é de natureza descritiva e exploratória. A ausência de testes inferenciais de hipótese (como ANOVA ou testes t) limita a generalização das afirmações de separabilidade para além do corpus RAVDESS. Contudo, a convergência entre os rankings estatísticos e os padrões de erro das matrizes de confusão oferece evidências que uma acurácia de 44% não é uma falha do classificador, mas um reflexo da insuficiência de descritores acústicos estáticos para resolver o dilema da valência emocional.

Em síntese, os resultados evidenciam que parâmetros acústicos clássicos fornecem uma base sólida e explicável para compreender as respostas fisiológicas ligadas à emoção. O classificador SVM cumpriu o seu papel em mapear o espaço de características para identificar a excitação da voz, mas também evidenciou que, para transpor a barreira da confusão entre emoções de mesma ativação, trabalhos futuros precisarão inevitavelmente integrar sequenciamento temporal (como redes recorrentes) ou combinar a análise acústica com informações textuais ou visuais.

5. CONSIDERAÇÕES FINAIS

O presente trabalho investigou a eficácia e a contribuição individual de diferentes grupos de características acústicas na tarefa de reconhecimento de emoções na fala, utilizando a base de dados RAVDESS e o classificador supervisionado SVM, para isso, o desempenho foi avaliado em três cenários experimentais distintos: representação temporal em nível de quadro (*frame-level*), agregação global (*utterance-level*) e um subconjunto otimizado de características (Top-10), selecionado por meio do método *Fisher Score*.

A análise das características revelou que as emoções moduladas alteram a estabilidade fonatória e o esforço vocal. O ranqueamento estabelecido pelo *Fisher Score* demonstrou o domínio absoluto das métricas prosódicas e de qualidade vocal, com a energia (RMSE), o *shimmer* e a frequência fundamental (F_0) assumindo o topo da relevância discriminativa. Em contrapartida, os descritores espectrais, notadamente os coeficientes MFCC isolados, apresentaram forte sobreposição entre as classes e baixo poder de separação estática. Em termos práticos, conclui-se que atributos de intensidade e instabilidade fonatória funcionam como indicadores robustos de ativação emocional, enquanto descritores puramente espectrais e cepstrais (como os MFCCs) não se mostram eficazes de forma isolada para a separação das oito classes, exigindo alta dimensionalidade ou combinação com dinâmicas temporais para contribuir significativamente com o modelo.

O cenário *frame-level* alcançou uma acurácia média de 43,44%, enquanto o cenário *utterance-level* obteve 41,25%, entretanto, foi o cenário reduzido Top-10, que atingiu o melhor desempenho global com 44,03% de acurácia, eliminando mais da metade das variáveis originais e reduzindo o custo computacional.

Ainda assim, é imprescindível um olhar crítico sobre os limites dessa abordagem. Uma acurácia em torno de 44% evidencia o teto de aprendizado de classificadores clássicos e estáticos diante das nuances afetivas da fala humana. As matrizes de confusão comprovaram que o sistema desenvolvido consegue agrupar as emoções pelo seu nível de ativação (energia), mas esbarra em limitações expressivas ao tentar diferenciar a valência dessas emoções (positividade ou negatividade). Vozes expressando alegria extrema e raiva extrema compartilham graus tão elevados de intensidade e tensão laríngea que se tornam virtualmente indistinguíveis para um algoritmo linear apoiado em descritores agregados. Optou-se, portanto, por um modelo que sacrifica o estado da arte em acurácia, tipicamente alcançado por arquiteturas de aprendizado profundo, em favor da interpretabilidade matemática e fisiológica do modelo preditivo.

Diante dessas constatações, os achados deste estudo pavimentam direcionamentos para trabalhos futuros:

- *Incorporação de Dinâmica Temporal*: a agregação de características (*utterance-level*)

elimina a evolução temporal da emoção. Trabalhos futuros devem incorporar derivadas temporais (coeficientes delta e delta-delta) ou adotar arquiteturas recorrentes, com o objetivo de capturar o contorno dinâmico da prosódia ao longo da frase.

- *Abordagens Multimodais*: Para romper a barreira de confusão entre emoções de mesma ativação, a análise acústica pode ser fundida a características linguísticas (conteúdo textual extraído por reconhecimento de fala) ou visuais, complementando a inferência da valência emocional.
- *Avaliação em Bases Naturais*: a expansão dos testes para corpora de fala espontânea, em vez de atuações controladas, permitirá avaliar a robustez das características selecionadas diante de ruídos e ambiguidades naturais do discurso humano.
- *Modelos de Ensemble*: A investigação de classificadores baseados em múltiplas abordagens (como *Random Forest* associadas a SVMs) pode elevar a capacidade preditiva em bases de dados limitadas, mantendo o controle sobre a importância das variáveis.

Em resumo, a presente dissertação demonstrou que é possível extrair informações justificáveis a partir de apenas 10 atributos acústicos, evidenciando o comportamento fisiológico subjacente à fala emocional. O trabalho contribui para o avanço das pesquisas em reconhecimento de emoções, fornecendo diretrizes sólidas sobre quais características priorizar e reforçando a viabilidade de sistemas enxutos, transparentes e cientificamente embasados.

REFERÊNCIAS

- AKÇAY, Mehmet Berkehan; OĞUZ, Kaya. **Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers**. [S.l.]: Elsevier B.V., 2020. 56-76 p.
- BURKHARDT, Felix; PAESCHKE, Astrid; ROLFES, M.; SENDLMEIER, Walter; WEISS, Benjamin. A database of german emotional speech. **9th European Conference on Speech Communication and Technology**, v. 5, p. 1517–1520, 09 2005.
- BUSSO, Carlos; BULUT, Murtaza; LEE, Chi-Chun; KAZEMZADEH, Abe; MOWER, Emily; KIM, Samuel; CHANG, Jeannette N; LEE, Sungbok; NARAYANAN, Shrikanth S. Iemocap: interactive emotional dyadic motion capture database. **Language Resources and Evaluation**, v. 42, p. 335–359, 2008. ISSN 1574-0218. Disponível em: <<https://doi.org/10.1007/s10579-008-9076-6>>.
- CHOWDHURY, Jaher Hassan; RAMANNA, Sheela; KOTTECHA, Ketan. Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. **Scientific Reports**, v. 15, p. 11824, 2025. ISSN 2045-2322. Disponível em: <<https://doi.org/10.1038/s41598-025-95734-z>>.
- COLUNGA-RODRIGUEZ, Alvaro A.; MARTÍNEZ-REBOLLAR, Alicia; ESTRADA-ESQUIVEL, Hugo; CLEMENTE, Eddie; PLIEGO-MARTÍNEZ, Odette A. Developing a dataset of audio features to classify emotions in speech. **Computation**, v. 13, n. 2, 2025. ISSN 2079-3197. Disponível em: <<https://www.mdpi.com/2079-3197/13/2/39>>.
- COSTA, Silvana Luciene do Nascimento Cunha et al. Análise acústica, baseada no modelo linear de produção da fala, para discriminação de vozes patológicas. Universidade Federal de Campina Grande, 2008.
- DAR, G. H. Mohmad; DELHIBABU, Radhakrishnan. Speech databases, speech features, and classifiers in speech emotion recognition: A review. **IEEE Access**, v. 12, p. 151122–151152, 2024.
- DHUHEIR, Marwan; ALBASEER, Abdullatif; BACCOUR, Emna; ERBAD, Aiman; ABDALLAH, Mohamed; HAMDI, Mounir. **Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey**. 2021. Disponível em: <<https://arxiv.org/abs/2107.05989>>.
- ILIOU, Theodoros; ANAGNOSTOPOULOS, Christos-Nikolaos. Classification on speech emotion recognition: A comparative study. **International Journal on Advances in Life Sciences**, IARIA, v. 2, n. 1–2, p. 18–28, 2010. ISSN 1942-2660.
- JOHNSTONE, Tom; SCHERER, Klaus R. Vocal communication of emotion. 2000.
- KHALIL, R A; JONES, E; BABAR, M I; JAN, T; ZAFAR, M H; ALHUSSAIN, T. Speech emotion recognition using deep learning techniques: A review. **IEEE Access**, v. 7, p. 117327–117345, 2019. ISSN 2169-3536.
- LEEM, S G.; FULFORD, D; ONNELA, J P.; GARD, D; BUSO, C. Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments. In: **ICASSP**

2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2022. p. 6447–6451. ISBN 2379-190X.

LIVINGSTONE, Steven R; RUSSO, Frank A. 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, p. e0196391–, 5 2018. Disponível em: <<https://doi.org/10.1371/journal.pone.0196391>>.

LOPE, Javier de; GRAÑA, Manuel. An ongoing review of speech emotion recognition. **Neurocomputing**, Elsevier B.V., v. 528, p. 1–11, 4 2023. ISSN 18728286.

MASHHADI, Mohammad Mahdi Rezapour; OSEI-BONSU, Kofi. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. **PLOS ONE**, Public Library of Science, v. 18, p. e0291500–, 11 2023. Disponível em: <<https://doi.org/10.1371/journal.pone.0291500>>.

NASSIF, A B; SHAHIN, I; ATTILI, I; AZZEH, M; SHAALAN, K. Speech recognition using deep neural networks: A systematic review. **IEEE Access**, v. 7, p. 19143–19165, 2019. ISSN 2169-3536.

SCHULLER, Björn; STEIDL, Stefan; BATLINER, Anton. **The Interspeech 2009 Emotion Challenge**. [S.l.: s.n.], 2009. 312-315 p.

TARNOWSKI, Paweł; KOŁODZIEJ, Marcin; MAJKOWSKI, Andrzej; RAK, Remigiusz J. Emotion recognition using facial expressions. **Procedia Computer Science**, v. 108, p. 1175–1184, 2017. ISSN 1877-0509. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1877050917305264>>.

WANG, Jiaxin; YIN, Hao; ZHOU, Yiding; XI, Wei. Advancements and challenges in speech emotion recognition: a comprehensive review. In: **Proc.SPIE**. [s.n.], 2024. v. 13077, p. 130770D. Disponível em: <<https://doi.org/10.1117/12.3027122>>.

WANI, T M; GUNAWAN, T S; QADRI, S A A; KARTIWI, M; AMBIKAIRAJAH, E. A comprehensive review of speech emotion recognition systems. **IEEE Access**, v. 9, p. 47795–47814, 2021. ISSN 2169-3536.

YAN, S; YE, L; HAN, S; HAN, T; LI, Y; ALASAARELA, E. Speech interactive emotion recognition system based on random forest. In: **2020 International Wireless Communications and Mobile Computing (IWCMC)**. [S.l.: s.n.], 2020. p. 1458–1462. ISBN 2376-6506.