AVALIAÇÃO DE CARACTERÍSTICAS NO RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA VOZ

Leonardo M. Silva (IFPB, Campus João Pessoa), Suzete N. Correia (IFPB, Campus João Pessoa), Silvana C. Costa (IFPB, Campus João Pessoa)

E-mails: marcal.leonardo@academico.ifpb.edu.br, suzete@ifpb.edu.br, silvana@ifpb.edu.br.

Área de conhecimento:(Tabela CNPq): 3.04.06.00-5 Telecomunicações.

Palavras-Chave: reconhecimento de emoções, processamento de sinais de voz, características acústicas, índice de separabilidade.

1 Introdução

O reconhecimento de emoções através da voz tem aplicações práticas significativas em diversos domínios, desde atendimentos automatizados mais empáticos e assistentes pessoais, até triagens clínicas remotas, como detecção de indícios de depressão ou ansiedade (LATIF et al., 2020; ALBESANO; FALCONE; SERVETTI, 2021). Para que esses sistemas funcionem com efetividade, é necessário compreender como as emoções afetam o sinal de fala e quais características acústicas melhor representam essas variações (BAHREINI; SCHERER; MOZGAI, 2021). Esta compreensão é fundamental para o desenvolvimento de sistemas que possam interpretar adequadamente as emoções humanas através da voz.

Embora as abordagens mais modernas utilizem arquiteturas profundas como redes convolucionais e recorrentes, métodos tradicionais que extraem características manuais da voz ainda são amplamente utilizados (ALBESANO; FALCONE; SERVETTI, 2021). Além de exigirem menos dados para treinamento, esses métodos permitem interpretar diretamente quais aspectos do sinal vocal contribuem para a classificação emocional, favorecendo aplicações transparentes e eficientes (ALBESANO; FALCONE; SERVETTI, 2021). A interpretabilidade desses métodos é especialmente importante em aplicações críticas, onde a compreensão do processo de decisão é fundamental (ZHANG; LIU; DU, 2022).

Neste trabalho, é proposta uma investigação sobre 21 métricas acústicas clássicas, avaliando sua capacidade de discriminar emoções na base `RAVDESS`. O foco do estudo é exploratório e visual, buscando compreender como cada métrica se distribui ao longo das emoções por meio de um índice de separabilidade (BAHREINI; SCHERER; MOZGAI, 2021). Esta abordagem metodológica permite identificar quais atributos são mais promissores para tarefas posteriores de classificação, contribuindo para o desenvolvimento de sistemas mais eficientes e interpretáveis (ZHANG; LIU; DU, 2022).

2 Materiais e Métodos

Este estudo utilizou a base `RAVDESS`, composta por 1.440 amostras de áudio falado por 24 atores (12 homens e 12 mulheres), cada um representando oito emoções: neutra (NEU), calma (CAL), feliz (FEL), triste (TRI), raiva (RAI), medo (MED), nojo (NOJ) e surpresa (SUR) (LATIF et al., 2020). As gravações foram realizadas em ambiente controlado com 48 kHz e 16 bits, garantindo uma qualidade adequada para análise acústica (BAHREINI; SCHERER; MOZGAI, 2021). Os áudios foram segmentados em janelas de 20 ms com sobreposição de 50%, aplicando-se janela de *Hamming* para manter a estacionariedade local (BAHREINI; SCHERER; MOZGAI, 2021). Esta técnica de segmentação é amplamente utilizada em processamento de sinais de voz, pois permite uma análise mais precisa das características temporais do sinal (ZHANG; LIU; DU, 2022).

Foram extraídas 21 características agrupadas em quatro domínios: prosódico (*Pitch*, RMSE, ZCR), qualidade vocal (*Jitter*, *Shimmer*, HNR), espectral (MFCC_1–MFCC_{13}, centroides) e articulatório (Formante 1) (BAHREINI; SCHERER; MOZGAI, 2021). Cada métrica foi agregada por média, resultando em um vetor de atributos por amostra. A separabilidade foi quantificada pela razão entre a diferença das médias interclasse e o desvio padrão

médio intraclasse, gerando um índice para cada métrica (BAHREINI; SCHERER; MOZGAI, 2021). Esta metodologia de análise permite avaliar objetivamente a capacidade discriminativa de cada característica (ZHANG; LIU; DU, 2022).

3 Resultados e Discussão

A Tabela 1 apresenta as dez métricas com maior índice de separabilidade. A métrica RMSE se destacou com o maior valor, sendo particularmente eficaz para distinguir emoções de alta ativação como raiva (RAI) e surpresa (SUR) (ALBESANO; FALCONE; SERVETTI, 2021). Esta característica reflete a energia do sinal e está relacionada à intensidade vocal, frequentemente alterada sob estados emocionais intensos (ALBESANO; FALCONE; SERVETTI, 2021). O segundo coeficiente MFCC ($MFCC_2$) apresentou excelente desempenho em capturar nuances espectrais específicas associadas às variações emocionais (BAHREINI; SCHERER; MOZGAI, 2021).

As métricas *Shimmer*, *Pitch* e HNR também demonstraram boa capacidade de separação (BAHREINI; SCHERER; MOZGAI, 2021). *Shimmer* reflete variações de amplitude entre ciclos glóticos sucessivos, frequentemente associadas a instabilidades na voz provocadas por estresse ou tensão emocional (BAHREINI; SCHERER; MOZGAI, 2021). *Pitch*, por sua vez, representa a frequência fundamental da fala, diretamente influenciada por estados emocionais como excitação ou tristeza (BAHREINI; SCHERER; MOZGAI, 2021). HNR mede a proporção entre energia harmônica e ruído, sendo um indicativo de clareza vocal — características que tendem a mudar entre emoções neutras e alteradas (LATIF et al., 2020).

Tabela 1: Ranking das métricas por índice de separabilidade

#	Métrica	Δ Média	Índice Separativo
1	RMSE	0.0138	2.94
2	$MFCC_2$	54.60	1.67
3	Shimmer	0.0334	1.51
4	Pitch	103.07	1.41
5	HNR	3.11	1.40
6	Jitter	0.0079	1.23
7	$MFCC_3$	26.47	1.16
8	$MFCC_4$	21.32	1.07
9	ZCR	0.0343	1.01
10	$MFCC_6$	17.59	0.98

O *boxplot* da RMSE (Figura 1) mostra claramente sua efetividade na distinção entre grupos emocionais. Emoções como RAI e SUR apresentam medianas mais elevadas e maior dispersão, enquanto NEU e CAL permanecem mais concentradas (ALBESANO; FALCONE; SERVETTI, 2021). A distribuição reduzida em classes de menor excitação sugere um perfil acústico mais uniforme, contrastando com as emoções mais intensas (ZHANG; LIU; DU, 2022).

Além da análise geral, foram comparados os resultados entre vozes masculinas e femininas. Observou-se que, embora a tendência de separabilidade se mantenha entre os gêneros, as vozes femininas apresentaram maior variação nos coeficientes MFCC e menor dispersão no *Pitch* (LATIF et al., 2020; ALBESANO; FALCONE; SERVETTI, 2021). Tais diferenças são consistentes com a literatura sobre variações acústicas de gênero (LATIF et al., 2020; ALBESANO; FALCONE; SERVETTI, 2021), e indicam que uma análise sensível ao sexo do locutor pode melhorar o desempenho de classificadores (ZHANG; LIU; DU, 2022). É importante destacar que o uso combinado de métricas, mesmo aquelas com menor índice isolado, pode melhorar a performance geral em tarefas de classificação (ZHANG; LIU; DU, 2022).

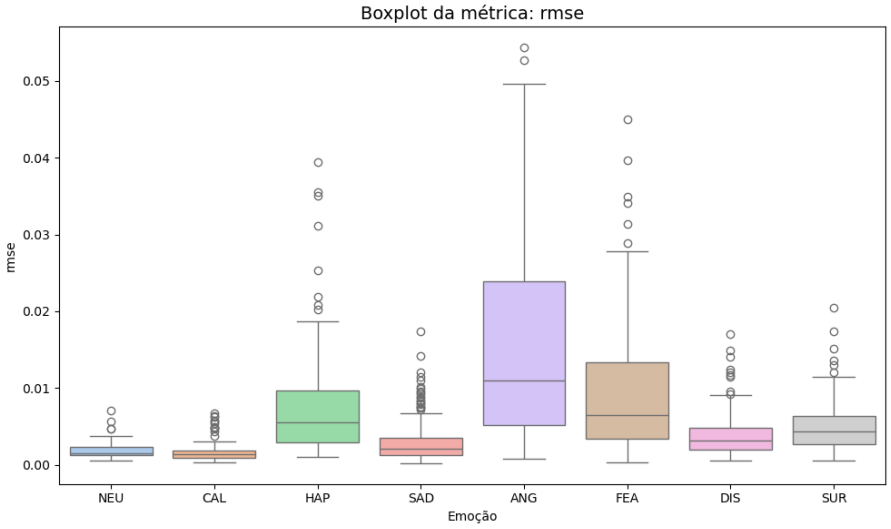

Figura 1: Boxplot da métrica RMSE por emoção.

4 Considerações Finais

A partir da análise exploratória das 21 métricas acústicas aplicadas à base RAVDESS, foi possível identificar as mais relevantes para fins de separação emocional na fala (BAHREINI; SCHERER; MOZGAI, 2021). As métricas RMSE, MFCC_2, *Shimmer*, *Pitch* e HNR se mostraram mais eficazes na distinção entre emoções, tanto do ponto de vista quantitativo quanto visual (ALBESANO; FALCONE; SERVETTI, 2021). Estes resultados reforçam a importância de atributos clássicos e computacionalmente acessíveis para aplicações em reconhecimento de emoções, especialmente em cenários com restrições de infraestrutura ou necessidade de interpretabilidade (ZHANG; LIU; DU, 2022).

Além disso, destaca-se a necessidade de investigar a generalização desses achados em outras bases de dados (ZHANG; LIU; DU, 2022). A literatura indica que o desempenho de métricas pode variar substancialmente entre corpora distintos, devido a fatores como idioma, espontaneidade da fala, qualidade de gravação e variação interindividual (ZHANG; LIU; DU, 2022). Portanto, recomenda-se a aplicação da metodologia em conjuntos como EMO-DB, CREMA-D ou SAVEE, ampliando o escopo experimental e validando a robustez dos atributos selecionados (ZHANG; LIU; DU, 2022).

Agradecimentos

Agradeço ao Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB) pelo apoio institucional e pelas condições proporcionadas, fundamentais para o desenvolvimento deste trabalho.

Referências

ALBESANO, D.; FALCONE, M.; SERVETTI, A. A real-time speech emotion recognition system for edge devices. *Electronics*, v. 10, n. 2, p. 188, 2021.

BAHREINI, K.; SCHERER, S.; MOZGAI, S. Multimodal emotion recognition using facial expressions and vocal prosody. *ACM Transactions on Interactive Intelligent Systems*, v. 11, n. 2, p. 1–24, 2021.

LATIF, S. et al. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, v. 8, p. 48607–48624, 2020.

ZHANG, Y.; LIU, W.; DU, J. Speech emotion recognition using transfer learning and spectral feature fusion. *Neurocomputing*, v. 481, p. 104–115, 2022.