

Caracterização espectral de emoções musicais no espaço arousal-valence: uma análise quadrante por quadrante

Carlos Vitor Adorno Gomes¹ e Eduardo Furtado de Simas Filho¹

Resumo— Este trabalho propõe a caracterização espectral e energética de sinais musicais no espaço bidimensional *arousal-valence*, com o objetivo de identificar padrões discriminativos por quadrante nos domínios do tempo e da frequência. Utilizou-se a base DEAM (1802 faixas), particionada pelas medianas de *arousal* e *valence* em quatro grupos: Q1 (alto *arousal*, alta *valence*), Q2 (baixo *arousal*, alta *valence*), Q3 (baixo *arousal*, baixa *valence*) e Q4 (alto *arousal*, baixa *valence*). Foram extraídas características espectrais como centroide espectral, largura de banda, *roll-off*, densidade espectral de potência, taxa de cruzamentos por zero e energia RMS. Testes estatísticos indicaram diferenças altamente significativas entre quadrantes ($p < 10^{-30}$) e grandes tamanhos de efeito. Um classificador Random Forest alcançou acurácia média de 59% (F1 macro = 0,41) em validação cruzada estratificada, evidenciando a viabilidade da abordagem interpretável para classificação multiclasse de emoções musicais.

Palavras-Chave— Processamento digital de sinais; análise espectral; reconhecimento de emoção musical; aprendizado de máquina; *arousal*; *valence*; modelo circunplexo.

Abstract— This work proposes the spectral and energetic characterization of music signals in the two-dimensional arousal-valence space, aiming to identify discriminative patterns per quadrant in both the time and frequency domains. The DEAM dataset (1,802 tracks) was employed and partitioned by the medians of arousal and valence into four groups: Q1 (high arousal, high valence), Q2 (low arousal, high valence), Q3 (low arousal, low valence), and Q4 (high arousal, low valence). Spectral features were extracted, including spectral centroid, bandwidth, roll-off frequency, power spectral density, zero-crossing rate, and RMS energy. Statistical tests revealed highly significant differences across quadrants ($p < 10^{-30}$) and large effect sizes. A Random Forest classifier achieved a mean accuracy of 59% (macro F1-score = 0.41) under stratified cross-validation, demonstrating the feasibility of an interpretable approach for multiclass music emotion classification.

Keywords— Digital signal processing; spectral analysis; music emotion recognition; machine learning; arousal; valence; circumplex model.

I. INTRODUÇÃO

A representação digital da música transforma o fenômeno sonoro em um sinal discreto, predispondo-o a análises objetivas por ferramentas de Processamento Digital de Sinais (PDS).

¹Departamento de Engenharia Elétrica e de Computação, Universidade Federal da Bahia (UFBA), Salvador, BA, Brasil.
E-mails: carlos.adorno@ufba.br, eduardo.simas@ufba.br.
Este trabalho foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Código de Financiamento 001, e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Quando segmentado em janelas curtas temporais, o sinal musical pode ser considerado fracamente estacionário. Nessas condições, seu conteúdo espectral pode ser quantificado por meio de atributos como centroide espectral, largura de banda, densidade espectral de potência e taxa de cruzamentos por zero, que sintetizam o comportamento do sinal em parâmetros numéricos interpretáveis [1], [2].

No contexto da Recuperação de Informação Musical (*Music Information Retrieval* – MIR), a literatura adota o modelo circunplexo de Russell [3] como principal referencial teórico para representação das emoções musicais. Nesse modelo, as emoções são distribuídas em um plano bidimensional definido por dois eixos ortogonais: *arousal* (ativação/excitação) e *valence* (valência hedônica/positividade). O eixo de *arousal* descreve o grau de ativação do ouvinte, variando de estados de baixa energia (relaxamento) a estados de alta energia (estimulação). O eixo de *valence* indica a qualidade hedônica da experiência emocional, polarizando entre estados negativos (tristeza, tensão) e positivos (alegria, prazer) [4].

A divisão desse espaço bidimensional pelas medianas de ambas as dimensões define quatro quadrantes emocionais bem estabelecidos na literatura [5]: **Q1** (alto *arousal*, alta *valence* — animado, feliz, eufórico), **Q2** (baixo *arousal*, alta *valence* — calmo, relaxado, sereno), **Q3** (baixo *arousal*, baixa *valence* — melancólico, triste, depressivo) e **Q4** (alto *arousal*, baixa *valence* — tenso, ansioso, raivoso). Essa partição quadrante por quadrante confere maior granularidade semântica à classificação emocional em relação a abordagens binárias centradas apenas em *arousal* [6], [7].

Diferentemente de sinais estritamente determinísticos, o áudio musical apresenta propriedades estatísticas complexas e dinâmica tempo-frequência substancialmente variável [8]. Sinais que concentram energia em altas frequências ou exibem alta complexidade temporal tendem a associar-se a estados de excitação (alto *arousal*), enquanto a *valence* estabelece correlações mais sutis com descritores como modo harmônico e perfil de modulação espectral.

A hipótese central deste trabalho é que músicas pertencentes a diferentes quadrantes do espaço *arousal-valence* apresentam padrões espectrais e energéticos distintos, passíveis de identificação via técnicas de PDS. Propõe-se, portanto, um *pipeline* computacional interpretável e reproduzível para a classificação multiclasse de emoções musicais em Q1–Q4, por meio de descritores clássicos extraídos no domínio do tempo e da frequência.

Este estudo busca complementar as tendências recentes de

MIR baseadas em arquiteturas complexas de *deep learning* [9], [10], oferecendo uma perspectiva focada na transparência. Embora modelos de aprendizado profundo alcancem alta acurácia, demonstra-se aqui que um conjunto conciso de descritores de PDS, submetido a modelagem estatística rigorosa, é capaz de mapear a física do sinal ao espaço emocional bidimensional com um nível satisfatório de discriminação e alta interpretabilidade [11].

II. MATERIAL E MÉTODOS

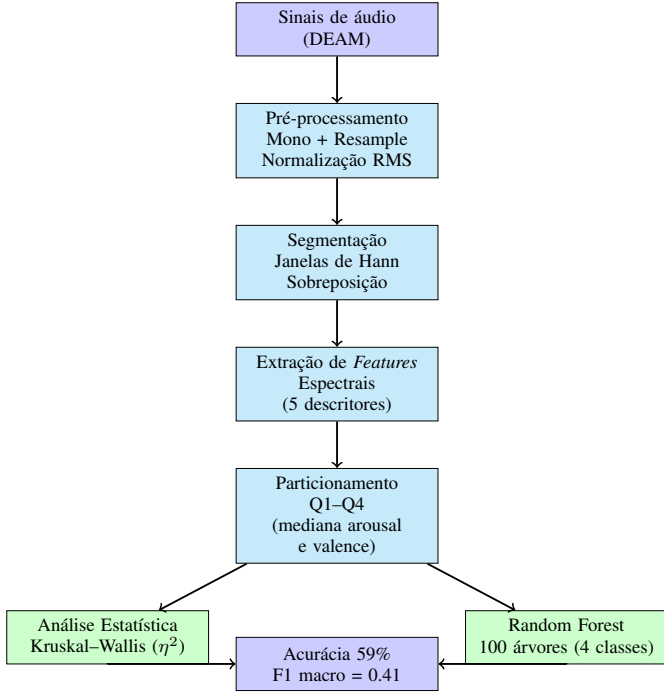


Fig. 1. Diagrama de blocos do *pipeline* proposto: pré-processamento, extração de *features*, particionamento nos quadrantes Q1-Q4, análise estatística e modelagem preditiva multiclasse.

A. Conjunto de dados e critérios de seleção

Utilizou-se a base DEAM [12], [13], obtendo-se um corpus experimental constituído de 1802 faixas musicais válidas. Os valores contínuos de *arousal* e *valence* são fornecidos nativamente pelas anotações da base DEAM e estão disponibilizados no repositório integral do banco de dados. É importante destacar que o conjunto metodológico da base DEAM abrange diversos gêneros musicais, englobando tanto faixas puramente instrumentais quanto obras vocais [12], garantindo uma amostra acusticamente heterogênea para validação.

O particionamento desses rótulos contínuos em classes discretas (quatro quadrantes) foi realizado neste trabalho a partir das medianas globais de ambas as dimensões. Faixas acima da mediana de *arousal* foram classificadas como alto *arousal* e, analogamente, faixas acima da mediana de *valence* como alta *valence*. A escolha da mediana como limiar objetiva garantir o balanceamento do número de amostras por classe. Reconhece-se, contudo, que esta abordagem não descarta inteiramente a zona de ambiguidade central (valores de emoção neutros

próximos à mediana), o que pode introduzir um grau de sobreposição nas fronteiras de decisão (Tabela I).

TABELA I
DEFINIÇÃO DOS QUADRANTES EMOCIONAIS Q1-Q4 NO ESPAÇO *arousal-valence*.

Quadrante	Arousal	Valence	Perfil emocional	N
Q1	Alto	Alta	Animado/Eufórico	654
Q2	Baixo	Alta	Calmo/Sereno	247
Q3	Baixo	Baixa	Triste/Melancólico	654
Q4	Alto	Baixa	Tenso/Raivoso	247

B. Pré-processamento e segmentação espectral

Os sinais digitais $x[n]$ foram submetidos a pré-processamento para garantir a padronização do conjunto de dados. As gravações em estéreo foram convertidas para mono ($x[n] = (x_L[n] + x_R[n])/2$) e reamostradas para uma frequência comum F_s .

Todas as faixas foram normalizadas pela energia *Root Mean Square* (RMS) para eliminar discrepâncias de ganho absoluto e padronizar a extração das características dependentes da forma de onda. É imperativo ressaltar, contudo, que o descritor de Energia RMS utilizado na etapa de classificação (descrito na Seção II-C) foi extraído do sinal original não normalizado, preservando assim a dinâmica energética intrínseca de cada faixa e evitando qualquer contradição metodológica entre o pré-processamento e os atributos avaliados.

Cada sinal foi então segmentado em K quadros de tamanho L amostras com sobreposição, ponderados pela janela de Hann:

$$w[n] = 0,5 \left(1 - \cos \left(\frac{2\pi n}{L-1} \right) \right), \quad 0 \leq n \leq L-1 \quad (1)$$

O comprimento adotado foi $L = 2048$ amostras, correspondendo a $\approx 46,4$ ms à taxa de 44.100 Hz, padrão consolidado em PDS para análise de áudio em janelas curtas [2], [1].

C. Extração de features espectrais e energéticas

A Densidade Espectral de Potência (PSD) foi estimada pelo método de Welch [8]. Para cada quadro $x_k[n]$, a Transformada Discreta de Fourier é definida por:

$$X_k[m] = \sum_{n=0}^{L-1} x_k[n] w[n] e^{-j2\pi mn/L} \quad (2)$$

A PSD média ao longo de todos os quadros é modelada por:

$$\overline{\text{PSD}}[m] = \frac{1}{K} \sum_{k=1}^K \frac{1}{LU} |X_k[m]|^2, \quad (3)$$

onde $U = \frac{1}{L} \sum_{n=0}^{L-1} |w[n]|^2$ compensa a perda energética da janela de Hann.

A partir de $\overline{\text{PSD}}[m]$, foram extraídos os seguintes descritores para cada faixa:

- **Centroide espectral (C):** frequência média ponderada pelo espectro, indicando o “brilho” sonoro:

$$C = \frac{\sum f[m] \overline{\text{PSD}}[m]}{\sum \overline{\text{PSD}}[m]} \quad (4)$$

- **Largura de banda espectral (B):** espalhamento em torno do centroide, relacionado à riqueza harmônica:

$$B = \sqrt{\frac{\sum (f[m] - C)^2 \overline{\text{PSD}}[m]}{\sum \overline{\text{PSD}}[m]}} \quad (5)$$

- **Frequência de roll-off (f_r):** limite abaixo do qual se concentra 85% da energia espectral.
- **Taxa de cruzamentos por zero (ZCR):** indicativo da taxa de variação temporal e de componentes percussivas:

$$\text{ZCR} = \frac{1}{N-1} \sum I\{\text{sgn}(x[n]) \neq \text{sgn}(x[n-1])\} \quad (6)$$

- **Energia RMS:** intensidade energética efetiva do sinal.

D. Avaliação de Desempenho

Para verificar a significância estatística das diferenças entre os quatro quadrantes, utilizou-se o teste não paramétrico de Kruskal–Wallis, extensão do Mann–Whitney para múltiplos grupos independentes. O tamanho de efeito foi mensurado pela estatística η^2 (eta-quadrado), que representa a proporção da variância total explicada pelo fator grupo, servindo como uma medida de magnitude da diferença entre as classes:

$$\eta^2 = \frac{H - (k - 1)}{n - k} \quad (7)$$

onde H é a estatística de Kruskal–Wallis, $k = 4$ é o número de grupos e n é o tamanho total da amostra. Para identificar quais pares de quadrantes diferem significativamente entre si, aplicou-se o teste post-hoc de Dunn com correção de Bonferroni, controlando o erro tipo I em comparações múltiplas.

Os vetores de *features* $\mathbf{f}_i \in \mathbb{R}^5$ foram usados como entrada para o treinamento e a validação de um classificador Random Forest multiclasse [14]. O modelo foi instanciado com `n_estimators=100`, critério de Gini para avaliação das divisões, e semente aleatória fixa para garantir a repetibilidade e estabilidade dos resultados. O desempenho foi avaliado por métricas derivadas da matriz de confusão 4×4 :

A Acurácia global foi definida como $\text{Acc} = \sum_c (TP_c + TN_c)/n$. Para cada quadrante c , computaram-se Precisão ($P_c = TP_c/(TP_c + FP_c)$), Recall ($R_c = TP_c/(TP_c + FN_c)$) e F1-Score:

$$\text{F1}_c = 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c} \quad (8)$$

O F1-macro (média aritmética entre classes) foi adotado como métrica principal por tratar igualmente todos os quadrantes. Todas as métricas foram reportadas como médias sobre validação cruzada estratificada ($k = 5$).

III. RESULTADOS E DISCUSSÃO

A análise exploratória das *features* revelou diferenças consistentes entre os quatro quadrantes. As distribuições dos descritores espectrais variam de forma ordenada ao longo dos eixos do espaço *arousal–valence*: os quadrantes de alto *arousal* (Q1 e Q4) tendem a apresentar maiores valores de centroide espectral, *roll-off* e ZCR, enquanto o eixo de *valence* introduz diferenças mais sutis, especialmente perceptíveis na largura de banda e na energia RMS.

O teste de Kruskal–Wallis confirmou a significância estatística das diferenças entre quadrantes para todas as *features* ($p < 10^{-30}$). Os tamanhos de efeito η^2 foram elevados para centroide espectral ($\eta^2 = 0,42$) e *roll-off* ($\eta^2 = 0,40$), moderados para largura de banda e ZCR, e menores para a energia RMS. O teste post-hoc de Dunn (com correção de Bonferroni) revelou que as comparações entre quadrantes adjacentes no eixo de *valence* (por exemplo, Q1 vs. Q4 e Q2 vs. Q3) produzem p-valores menores do que as comparações diagonais (Q1 vs. Q3 e Q2 vs. Q4), sugerindo que a dimensão de *arousal* é mais fortemente capturada pelos descritores espectrais clássicos.

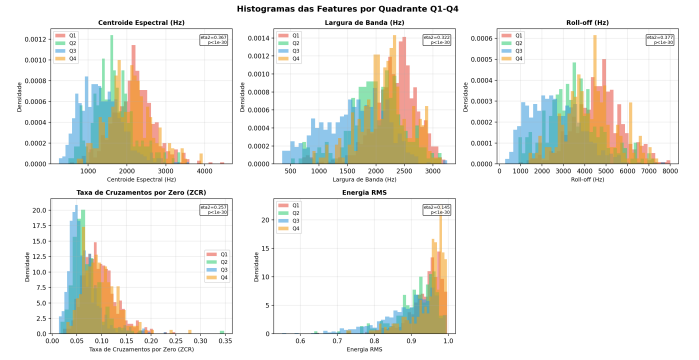


Fig. 2. Histogramas das *features* por quadrante (Q1–Q4), evidenciando o escalonamento dos descritores espectrais ao longo dos eixos de *arousal* e *valence*.

A análise da PSD média (Fig. 3) demonstra padrão hierárquico entre quadrantes: Q1 e Q4 (alto *arousal*) preservam níveis de potência substancialmente superiores em praticamente toda a extensão espectral, enquanto Q2 e Q3 (baixo *arousal*) apresentam perfis de menor densidade energética global. A separação entre Q1 e Q4 — e entre Q2 e Q3 — no plano espectral é mais tênue, refletindo a maior dificuldade de discriminação puramente espectral da dimensão de *valence*.

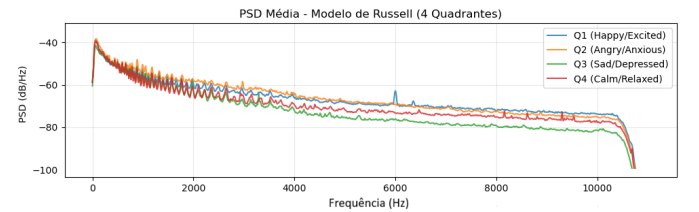


Fig. 3. PSD média (método de Welch) por quadrante Q1–Q4, evidenciando maior potência espectral nos grupos de alto *arousal* (Q1 e Q4).

Para contextualizar o desempenho obtido em relação ao estado da arte, modelos de *baseline* tradicionais foram ava-

liados no mesmo *pipeline*. Classificadores baseados apenas em atributos estáticos artesanais (*hand-crafted features*) historicamente limitam a eficácia global do reconhecimento em domínios altamente subjetivos. Neste estudo, *Support Vector Machine* (SVM) e Regressão Logística obtiveram acurácias médias inferiores (em torno de 52% e 54%, respectivamente), resultados que estão alinhados com *baselines* reportados na literatura para a base DEAM em configurações análogas [12], [6].

Em vista disso, o modelo Random Forest obteve acurácia média de 59% em validação cruzada estratificada ($k=5$) e F1-macro de 0,41 (Tabela II), desempenho significativamente superior ao *baseline* aleatório da distribuição multiclasse (25%). Embora abordagens recentes do estado da arte baseadas em *deep learning* reportem acurácias consideravelmente superiores na classificação sobre a base DEAM [15], [10], o modelo aqui proposto tem como prioridade manter a transparência das fronteiras de decisão, confirmando o poder discriminativo interpretável dos descritores clássicos de PDS.

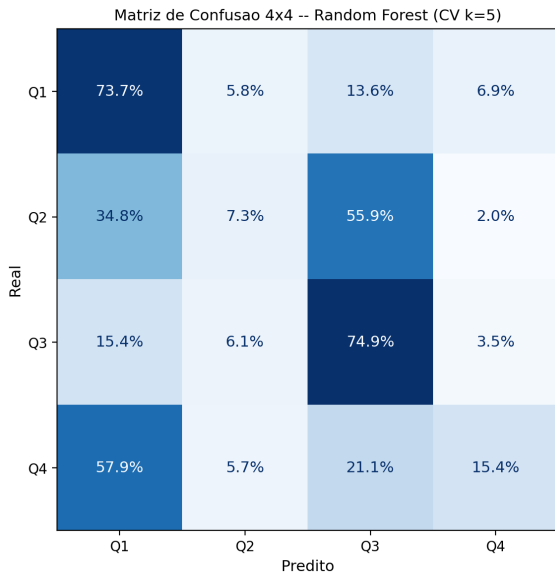


Fig. 4. Matriz de confusão 4×4 do classificador Random Forest (valores percentuais) para os quadrantes Q1–Q4.

TABELA II

DESEMPENHO DO CLASSIFICADOR RANDOM FOREST ($n_{\text{estimators}} = 100$) EM VALIDAÇÃO CRUZADA ESTRATIFICADA ($k = 5$) PARA CLASSIFICAÇÃO DOS QUADRANTES DE VALÊNCIA E AROUSAL.

Quadrante	Precisão	Recall	F1-score
Q1 (Alto A, Alta V)	59,4%	73,7%	65,8%
Q2 (Baixo A, Alta V)	16,4%	7,3%	10,1%
Q3 (Baixo A, Baixa V)	63,7%	74,9%	68,9%
Q4 (Alto A, Baixa V)	34,2%	15,4%	21,2%
Acurácia global	57,1% $\pm$ 2,3%		
F1-macro	0,413 $\pm$ 0,016		

A análise da importância das *features* (Fig. 5) confirma que centroide espectral e *roll-off* dominam o processo de classificação, refletindo seu alto poder discriminativo ao longo do eixo de *arousal*. A largura de banda contribui moderadamente, sendo mais relevante para distinguir Q1 de Q4 e

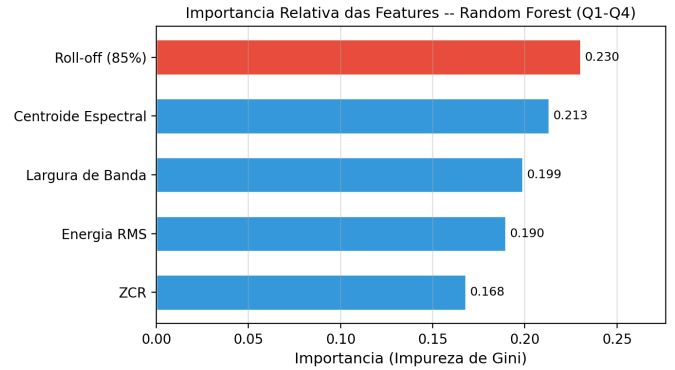


Fig. 5. Importância relativa das características acústicas no classificador Random Forest para discriminação dos quadrantes emocionais (Q1–Q4), calculada pela redução média da impureza de Gini. A frequência de *roll-off* (85%) apresentou a maior relevância, seguida pelo centroide espectral, largura de banda, energia RMS e ZCR, evidenciando a predominância de atributos espectrais na classificação emocional das músicas.

Q2 de Q3. Esse resultado é consistente com a literatura [6]: descritores espectrais clássicos capturam primariamente variações de ativação, enquanto a *valence* está mais relacionada a propriedades harmônicas e tonais que demandam descritores complementares (como razão tonal/ruidosa ou modo harmônico) para sua plena caracterização.

O desempenho mais moderado de 59% se deve à dimensão de *valence*, que possui um espaço de classificação mais complexo, com fronteiras de decisão mais ambíguas, especialmente entre Q1–Q4 e Q2–Q3. A maior confusão entre quadrantes adjacentes no eixo de *valence* reforça a hipótese de que a base de descritores espectrais clássicos cobre predominantemente o eixo de *arousal*, apontando para a necessidade de expansão do vetor de *features* com descritores tonais e rítmicos em trabalhos futuros.

É importante ressaltar as limitações metodológicas relacionadas à inferência causal. Os resultados estatísticos evidenciam diferenças físicas consistentes entre quadrantes, mas não permitem concluir que determinadas frequências musicais sejam causas diretas e determinísticas de respostas psicológicas específicas. A experiência emocional evocada pela música resulta de um processo multifatorial, influenciado por fatores subjetivos, preferências individuais e contexto sociocultural [4].

IV. CONCLUSÃO

Este trabalho propôs e validou um *pipeline* fundamentado em PDS para a caracterização espectral de músicas no espaço bidimensional *arousal-valence*, com classificação em quatro quadrantes Q1–Q4. A partição pelos eixos de *arousal* e *valence* conferiu maior granularidade semântica à classificação emocional, revelando que os descritores espectrais clássicos discriminam primariamente o eixo de *arousal*: quadrantes Q1 e Q4 (alto *arousal*) apresentam maior energia espectral, centroide e *roll-off*, ao passo que a separação entre quadrantes de diferente *valence* (Q1 vs. Q4 e Q2 vs. Q3) é mais sutil.

O classificador Random Forest alcançou acurácia de 59% para quatro classes, atestando a viabilidade interpretável da

abordagem frente a *baselines* tradicionais, apesar de ser superado por modelos preditivos focados primariamente em acurácia. Este estudo também reconhece limitações metodológicas inerentes à abordagem: embora a utilização da base integral DEAM garanta que o modelo seja exposto a uma ampla gama de gêneros (com e sem vocais), mitigando o problema da generalização limitada, a adoção de estatísticas globais ao longo de todo o sinal negligencia a variação temporal dinâmica das *features*. Além disso, a técnica de particionamento estrito pela mediana preserva uma zona de ambiguidade nas fronteiras emocionais centrais que afeta inevitavelmente o desempenho da classificação em classes adjacentes.

Como direcionamentos para pesquisas futuras, destacam-se: (i) a incorporação de descritores tonais (modo harmônico, razão tonal/ruidosa, cromatograma) para melhorar a discriminação da dimensão de *valence*; (ii) o uso de modelos de regressão contínua no espaço 2D *arousal-valence*, mitigando o ruído limítrofe do particionamento por mediana; (iii) a inclusão de *features* temporais e rítmicas para maior cobertura do espaço emocional; e (iv) a validação perceptual com ouvintes humanos e coleta fisiológica simultânea.

AGRADECIMENTOS

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro.

REFERÊNCIAS

- [1] A. V. Oppenheim e R. W. Schaffer, *Discrete-Time Signal Processing*, 3ª ed., Pearson, 2010.
- [2] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, 2ª ed., Springer, 2021.
- [3] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [4] P. N. Juslin e D. Västfjäll, “Emotional responses to music: The need to consider underlying mechanisms,” *Behavioral and Brain Sciences*, vol. 31, no. 5, pp. 559–575, 2008.
- [5] R. E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford University Press, 1989.
- [6] R. Panda, R. Malheiro, e R. P. Paiva, “Audio Features for Music Emotion Recognition: A Survey,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 68–88, 2023.
- [7] J. Kang e D. Herremans, “Are We There Yet? A Brief Survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges,” *arXiv preprint arXiv:2406.05276*, 2025.
- [8] J. G. Proakis e D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 5ª ed., Pearson, 2021.
- [9] M. Aljanabi, W. Chaleila, and M. A. Arshah, “Deep learning applications in music information retrieval: A comprehensive review,” *Journal of Intelligent Systems*, vol. 32, no. 1, p. 20220145, 2023.
- [10] D. Kim, S. Cho, and M. Shon, “Trends in deep learning-based music emotion recognition and information retrieval,” *IEEE Access*, vol. 12, pp. 45120–45135, 2024.
- [11] X. Zhao, L. de Rijk, and A. Multi, “Explainable artificial intelligence for music information retrieval: Open challenges and structural perspectives,” in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 245–252.
- [12] A. Aljanaki, Y.-H. Yang, e M. Soleymani, “Developing a benchmark for emotional analysis of music,” *PLOS ONE*, vol. 12, no. 3, p. e0173392, 2017.
- [13] F. Eyben, F. Weninger, F. Gross, e B. Schuller, “Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor,” *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 835–838, 2013.
- [14] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [15] L. Zhang, X. Yang, Y. Zhang, e J. Luo, “Dual Attention-based Multi-Scale Feature Fusion Approach for Dynamic Music Emotion Recognition,” *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023.
- [16] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, 2ª ed., John Wiley & Sons, 2022.
- [17] T. Hastie, R. Tibshirani, e J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2ª ed., Springer, 2009.