Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Classificação de Áudio Musical a Partir dos Coeficientes da Transformada Wavelet Utilizando Redes Neurais Convolucionais

Milton dos Santos[1]
ITA/IEAv, São Paulo, Brasil
Camilo Rodrigues Neto[2]
Grupo de Modelagem de Sistemas Complexos, EACH/USP, São Paulo, Brasil.

Resumo. A identificação do estilo musical a qual pertence uma música é uma tarefa relativamente simples para um humano, mesmo com pouco treinamento musical. Entretanto, é uma tarefa difícil a ser realizada de forma automatizada. Neste trabalho, utilizamos a transformada Wavelet, que representa uma música em suas componentes de frequência em função do tempo, gerando uma imagem denominada espectrograma. A partir dos espectrogramas, a Rede Neural Convolucional foi treinada com o objetivo de classificar os sinais de áudio de acordo com os seus estilos musicais. Foi utilizada apenas metade de cada sinal de áudio para gerar os espectrogramas, resultando em um total de 6.075 músicas para treinamento e 2.025 para teste provenientes de 10 estilos musicais – Blues, Clássico, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae e Rock. Os dados para treinamento e também para previsão foram escolhidos aleatoriamente em cada estilo musical para que fossem executadas 1000 épocas de treinamento, a medição foi repetida 10 vezes para cada estilo musical, deste modo garantindo o processo estocástico. A acurácia de treinamento obteve o melhor resultado com 90% das imagens (8100) de aproximadamente 82%. Os estilos Reggae, Jazz, Hiphop, Country, Classical e Blues obtiveram os seguintes melhores valores médios de previsões certas respectivamente: 94%, 88%, 82%, 90%, 91% e 83%.

Palavras-chave. Processamento de Sinais, MIR, Transformada Wavelet, Coeficientes Wavelet, Rede Neural Convolucional

1 Introdução

As Redes Neurais Convolucionais (do inglês, "Convolutional Neural Networks (CNN)") têm se destacado na Visão Computacional por sua eficácia em reconhecimento, restauração e geração de imagens. Além de imagens fotográficas, sinais temporais, como áudio, podem ser convertidos em representações visuais para análise. Entretanto, o treinamento dessas redes enfrenta desafios relacionados à quantidade e qualidade das imagens, afetando a generalização e a capacidade de extração de características relevantes [3].

Este trabalho investiga a aplicação de CNNs na classificação de áudio, utilizando imagens geradas a partir dos coeficientes da transformada Wavelet. A contribuição principal é o uso de aprendizagem profunda com representações Wavelet, visando aprimorar a previsão do estilo musical.

[1]miltonsantos@ita.br
[2]camiloneto@usp.br

2 Transformada Wavelet

As transformadas Wavelet, assim como as de Fourier, realizam uma projeção linear do sinal em uma função base para extrair informações relevantes [5]. No entanto, ao contrário das funções cossenoidais de Fourier, as Wavelets possuem domínio compacto, permitindo localizar e analisar características do sinal no tempo e na frequência simultaneamente.

O espectrograma obtido pela transformada de Fourier não é localizado, enquanto a transformada de Fourier janelada melhora essa localização. Já as transformadas Wavelet oferecem localização e multiescala, permitindo uma análise mais refinada. As funções Wavelet derivam de uma Wavelet Mãe ψ de média zero, dilatada pelos parâmetros de escala s e posição u [16]:

$$\Psi_{s,u}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right), \quad u \in \mathbb{R}, s > 0 \tag{1}$$

A transformada Wavelet projeta um sinal $x(t)$ dentro de uma janela de tempo chamada "Wavelet", ela é transformada para permitir um redimensionamento da escala s onde a mudança da localização da frequência acontece e a Wavelet pode ser deslocada no tempo para qualquer localização da translação u que é o termo da direita, Ψ é o tipo de Wavelet escolhida fazendo u variar no tempo t e a escala s variar na frequência ω [4]:

$$\mathcal{W}_\psi(s,u) = \int_{-\infty}^{\infty} f(t)\psi_{s,u}(t)dt \tag{2}$$

A transformada Wavelet Haar foi desenvolvida em 1910 e é a Wavelet mais simples, sendo usada para decomposição de sinais por meio de componentes aproximados e detalhados [8]. Já a Wavelet Morse Generalizada é indicada para sinais modulados, sendo definida no domínio da frequência [15]. A característica fundamental da transformada Wavelet é a manipulação da escala da frequência com a translação no tempo com a vantagem do tamanho de janela não ser constante, isto faz com que a manipulação dos parâmetros de escala "s" e o de translação "u" exerçam um papel importante no emprego dessa técnica. A figura 1 ilustra como esses parâmetros se deslocam de forma mais geral.

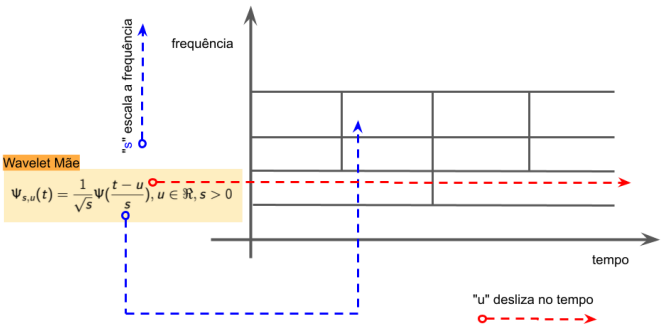

Figura 1: Parâmetro **s** se desloca pela dimensão da frequência, assim como o parâmetro **u** se desloca por meio da dimensão do tempo. Elaborado pelo autor, 2022.

O parâmetro de escala "s" de uma Wavelet tem o significado de poder dilatar ou contrair a Wavelet Mãe, de tal modo que, quanto menor o valor de escala "s", mais contraída será a Wavelet e, considerando esta característica, a escala está relacionada com a frequência do sinal que diminui

DOI: 10.5540/03.2026.012.01.0260

a escala "s" fazendo com que a Wavelet seja comprimida e os detalhes da transformada mudem rapidamente em uma alta frequência ω. Essa capacidade de ajuste dinâmico permite melhor representação de sinais em diferentes escalas de frequência [21]. A flexibilidade da transformada Wavelet permite a análise de sinais variáveis no tempo e na frequência, sendo amplamente aplicada em processamento de sinais e aprendizado de máquina [2, 16].

Uma alta escala "s" implica em uma Wavelet dilatada onde os detalhes da transformada mudam lentamente em uma baixa frequência ω [16], conforme ilustra a figura 1 com a forma na qual os domínios variam de formas diferentes. A transformada Wavelet mais simples, conhecida como Wavelet *Haar*, foi desenvolvida em 1910 [8]. A sua definição, com um exemplo sintético, apresenta o resultado de três Wavelets de Haar com $\Psi_{1,0}$, $\Psi_{1/2,0}$ e $\Psi_{1/2,1/2}$ aplicadas em um sinal:

$$\Psi(t)_{Haar} = \begin{cases} 1, & 0 \leq t < \frac{1}{2} \\ -1, & \frac{1}{2} \leq t < 1 \\ 0, & de\, outra\, forma. \end{cases}$$

Esta transformada Wavelet possui a sua versão Discreta onde, dada uma série $T = t_1, ..., t_n$, a Wavelet Haar Discreta transforma a saída em duas séries: as *aproximadas* A_i e os *detalhadas* D_i, onde $1 \leq i \leq \frac{n}{2}$ [14]:

$$A_i = \frac{t_{2i-1} + t_{2i}}{\sqrt{2}} \tag{3}$$

$$D_i = \frac{t_{2i-1} - t_{2i}}{\sqrt{2}} \tag{4}$$

A Wavelet Morse Generalizada é um tipo de Wavelet utilizada em sinais modulados, definida no domínio da frequência com pico de $\left(\frac{P}{\gamma^2}\right)^{\frac{1}{\gamma}}$:

$$\psi_{P,\gamma}(\omega) = U(\omega)\alpha_{P,\gamma}\omega^{\frac{P^2}{\gamma}}e^{-\omega^\gamma} \tag{5}$$

Uω é a função de etapa de Heaviside. O termo $\alpha_{\beta,\gamma}$ é a constante de normalização da Wavelet. β e γ são os parâmetros respectivamente responsáveis pelo decaimento e pela simetria da função Wavelet, mas estão representados pela variável P^2 que é o produto $\beta\gamma$, sendo a largura de banda pelo tempo [15].

3 Convolutional Neural Network

A Rede Neural Convolucional (CNN) foi desenvolvida para problemas de classificação de imagens, demonstrando grande sucesso nessa tarefa. A imagem de entrada é convertida em uma matriz de pixels, passando por camadas que identificam padrões e calculam a probabilidade de classificação [13].

3.1 Convolução

A operação de convolução é denotada por $\star$ entre duas funções, onde $x(t)$ é a entrada e $w(a)$ é um filtro. A saída é definida como [7]:

$$s(t) = (x \star w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{6}$$

3.2 Filtro de Imagem com Camadas Convolucionais

Em camadas convolucionais, os filtros aplicados extraem características da imagem. O mapa de características é dado por:

$$\mathbf{H}_i = g_i act[\mathbf{K}_i \star \mathbf{X} + \mathbf{B}_i] \tag{7}$$

3.3 Agrupamento (*Pooling*)

O *pooling* reduz a dimensionalidade, mantendo as informações relevantes. A operação de *max-pooling* seleciona os valores máximos dentro de uma região, propagando apenas esses valores na retropropagação [7]:

$$\frac{\partial L}{\partial h_{i,r_j,c_k}} = \begin{cases} 0, & r_j \neq r_j^*, c_k \neq r_j^*, \\ \frac{\partial L}{\partial p_{ijk}}, & r_j = r_j^*, c_k = r_j^*. \end{cases} \tag{8}$$

Essas técnicas tornam as CNNs eficientes para reconhecimento de padrões em imagens [22].

3.4 Camadas Densas

Após as etapas convolucionais e de pooling, os mapas de características são dimensionados e passados para camadas densas, cuja função é classificar. A combinação dessas camadas permite que a rede aprenda as características, bordas e texturas até chegar a formas mais complexas [12].

3.5 Processo de Treinamento

A imagem que é inserida na CNN é propagada pela rede, passando por convoluções, ativações e pooling, até gerar uma predição na saída para tal são necessários algumas iterações.

- **Cálculo da Função de Perda**: predição de saída é comparada ao rótulo real por meio de uma função de custo, como a Cross-Entropy Loss para classificação [7].
- **Retropropagação**: a rede calcula os gradientes dos pesos em relação à perda utilizando o algoritmo de Backpropagation, combinado à regra da cadeia [17].
- **Atualização dos Pesos**: os pesos são atualizados por meio de algoritmos de otimização, como o Stochastic Gradient Descent (SGD), o Adam ou variantes [11].
- **Regularização**: técnicas como Dropout [18] e Batch Normalization [9] são frequentemente aplicadas para prevenir overfitting e acelerar o treinamento.

4 Dados Experimentais

A base de dados utilizada é uma coletânea de áudios disponível *online*[3], inicialmente manipulada por Zanetakis *et al.* para estudos de Recuperação de Informação Musical [20] e amplamente utilizada em pesquisas acadêmicas [6]. Contém 1000 arquivos de áudio distribuídos igualmente entre 10 estilos musicais: *Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae e Rock*. Cada diretório contém 100 músicas, algumas com versões repetidas, como a "Rhapsody in Blue"de George Gershwin, presente nos arquivos 44 e 48, totalizando 13 ocorrências semelhantes [19].

[3]https://github.com/marsyas/marsyas

4.1 Organização de Dados

A organização dos dados seguiu três etapas: extração dos coeficientes aproximados da transformada Wavelet, geração de imagens (espectrogramas) e organização dos dados para treinamento e teste da Rede Neural Convolucional. O método utilizado para garantir a proporcionalidade dos dados para o treinamento foi o de manter a aleatoriedade na escolha dos arquivos para treinamento e também para a previsão, uma vez que os dados utilizados no presente trabalho apresentam balanceamento entre as classes [10], em que cada estilo musical possui a mesma quantidade de músicas.

5 Coeficientes e Espectrogramas

5.1 Geração dos Coeficientes

Os arquivos de áudio foram convertidos de *au* para *wav* para compatibilidade com o código desenvolvido, mantendo as propriedades originais. Os coeficientes aproximados foram extraídos da série temporal de cada áudio utilizando a Wavelet Haar [1]. Esta escolha permitiu capturar a maior parte do sinal com coeficientes aproximados.

Dada a limitação da base de dados, foram extraídos os coeficientes aproximados de ordens 1 (cA1), 2 (cA2) e 3 (cA3), resultando em 3.000 novas séries temporais.

5.2 Geração dos Espectrogramas e Treinamento da CNN

Cada série temporal foi dividida em três partes iguais para a geração de imagens espectrogramas. O algoritmo seguiu as etapas: 1) acessar a série temporal do coeficiente aproximado cA_n; 2) determinar o comprimento da série; 3) dividir a série em três segmentos; 4) aplicar a transformada Wavelet Contínua (CWT); 5) repetir o processo para 3.000 séries, resultando em 9.000 novas imagens para treinar a Rede Neural Convolucional. O método de escolha dos dados para a fase de teste foi feito de forma aleatória, garantindo a estocasticidade, em vez de escolher sempre as 10 primeiras imagens sequenciais de cada estilo para serem testadas. O treinamento da Rede Neural Convolucional teve 75% das figuras utilizadas para treinamento e 25% para validação para cada grupo (estilo musical).

5.3 Principais resultados

Os resultados abaixo apresentam os valores de previsão de cada estilo musical de acordo com a quantidade de imagens utilizadas no treinamento. Para cada conjunto de imagens as redes foram treinadas 10 vezes com 1.000 épocas cada uma com uma taxa de aprendizagem de 0,01. Agrupadas do seguinte modo: 30% (2.700 imagens), 70% (6.300 imagens) e 90% (8.100 imagens). Os valores de acerto foram obtidos por meio da escolha aleatória de 90 novas amostras, previstas 10 vezes, para chegar aos valores médios das tabelas abaixo.

Tabela 1: Utilizados 30% dos dados. Estilo Country obteve média de acerto nas previsões acima de 80%.

30%	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Média	0,67	0,56	**0,82**	0,45	0,57	0,79	0,44	0,63	0,76	0,36
Desvio padrão	0,08	0,23	**0,06**	0,13	0,11	0,07	0,15	0,10	0,06	0,16

Tabela 2: Utilizados 70% dos dados. A média de acertos ficou acima de 80% para a maioria.

70%	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Média	**0,82**	**0,84**	**0,89**	0,67	**0,83**	**0,83**	0,48	0,67	**0,91**	0,71
Desvio padrão	**0,09**	**0,18**	**0,04**	0,12	**0,10**	**0,08**	0,11	0,10	**0,05**	0,12

Tabela 3: Utilizados 90% dos dados. Foram alcançadas algumas médias de acertos maiores iguais a 90%.

90%	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Média	**0,83**	**0,91**	**0,90**	0,70	**0,82**	**0,88**	0,57	0,61	**0,94**	0,66
Desvio padrão	**0,11**	**0,09**	**0,05**	0,08	**0,07**	**0,06**	0,12	0,12	**0,06**	0,09

6 Considerações Finais

Este trabalho abordou a classificação de sinais de áudio em estilos musicais utilizando coeficientes aproximados da transformada Wavelet Discreta. O método proposto gerou novas séries temporais com apenas metade do sinal (15 s), aumentando o conjunto de imagens utilizadas no treinamento da Rede Neural Convolucional, o percentual de acerto de alguns estilos foi maior igual a 90%. Para trabalhos futuros, sugere-se treinar a mesma rede com diferentes tipos de Wavelets, porém com a mesma quantidade de coeficientes aproximados a fim de identificar a Wavelet para obter diferentes resultados.

Agradecimentos

Agradecimento ao meu orientador à época, Prof. Dr. Camilo Rodrigues Neto, e ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 à realização deste trabalho.

Referências

[1] P. A. Addison. **The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine, and finance**. 1st ed. Taylor & Francis, 2002. ISBN: 978-0750306928.

[2] S. L. Brunton e J. N. Kutz. **Data-driven science and engineering: Machine learning, dynamical systems, and control**. Cambridge University Press, 2022. ISBN: 978-1108422093.

[3] S. Dodge e L. Karam. "Understanding how image quality affects deep neural networks". Em: **Eighth International Conference on Quality of Multimedia Experience (QoMEX)** (2016), pp. 1–6. DOI: 10.1109/QoMEX.2016.7498955.

[4] M. O. Domingues, O. Mendes, M. K. Kaibara, V. E. Menconi e E. Bernardes. "Explorando a transformada wavelet contínua". Em: **Revista Brasileira de Ensino de Física** 38 (2016). DOI: 10.1590/1806-9126-RBEF-2016-0019.

[5] F. Germain. **The Wavelet Transform Applications in Music Information Retrieval**. 1a ed. CRC Press, 2009, pp. 3–4. ISBN: 9780750306928.

[6] A. Ghildiyal, K. Singh e S. Sharma. "Music genre classification using machine learning". Em: **4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE** (2020), pp. 1368–1372.

[7] I. Goodfellow, Y. Bengio e A. Courville. **Deep Learning**. MIT Press, 2016. ISBN: 9780262035613.

[8] A. Haar. "Zur theorie der orthogonalen funktionensysteme". Em: **Mathematische Annalen, 69(3):331–371** (1910).

[9] S. Ioffe e C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". Em: abs/1502.03167 (2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

[10] G. James, D. Witten, T. Hastie e R. Tibshirani. **An Introduction to Statistical Learning with Applications in Python**. Springer Texts in Statistics. Springer, 2013. ISBN: 978-3-031-38747-0.

[11] D. P. Kingma e J. Ba. **Adam: A Method for Stochastic Optimization**. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[12] Y. LeCun. "Generalization and network design strategies". Em: **Connectionism in perspective** 19.143-155 (1989), p. 18.

[13] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard e L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". Em: **Neural Computation** 4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541.

[14] D. Li, T. F. Bissyande, J. Klein e Y. L. Traon. "Time series classification with discrete wavelet transformed data: Insights from an empirical study". Em: **The 28th International Conference on Software Engineering and Knowledge Engineering (SEKE 2016)**. 2016. DOI: 10.1142/S0218194016400088.

[15] J. Lilly e S. Olhede. "Generalized Morse Wavelets as a Superfamily of Analytic Wavelets". Em: **IEEE Transactions on Signal Processing** 60 (2012), pp. 2661–2670. DOI: 10.1109/TSP.2012.2210890.

[16] S. Mallat. **A Wavelet Tour of Signal Processing: The Sparse Way**. Academic Press, 2008. ISBN: 9780123743701.

[17] D. E. Rumelhart, G. E. Hinton e R. J. Williams. "Learning representations by back-propagating errors". Em: **Nature** 323(6088) (1986), pp. 533–536. DOI: 10.1038/323533a0.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever e R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". Em: **Journal of Machine Learning Research** 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[19] B. L. Sturm. "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use". Em: **CoRR** abs/1306.1461 (2013). DOI: 10.1080/09298215.2014.894533.

[20] G. Tzanetakis e P. Cook. "Musical genre classification of audio signals". Em: **IEEE Transactions on Audio and Speech Processing** 10.5 (2002). DOI: 10.1109/TSA.2002.800560.

[21] C. Weihs, D. Jannach, I. Vatolkina e G. Rudolph. **Music Data Analysis: Foundations and Applications**. Chapman e Hall/CRC, 2019. ISBN: 9780367872816.

[22] I. H. Witten, E. Frank, M. A. Hall e C. J. Pal. **Data Mining: Practical Machine Learning Tools and Techniques**. 4th ed. Morgan Kaufmann, 2016. ISBN: 9780128042915.

