TRANSFORMADA DE FOURIER CONTÍNUA E TRANSFORMADA DE GABOR APLICADAS EM REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO DE VOZES

RONEY MENEZES MEIRELLES JÚNIOR[1]; JOÃO INÁCIO MOREIRA BEZERRA[2]; SOFIA PAGLIARINI[2]; MARLON MAURÍCIO HERNANDEZ CELY[2]
JAIRO VALOES DE ALENCAR RAMALHO[2]
[1]Universidade Federal de Pelotas – rjmeirelles999@gmail.com
[2]Universidade Federal de Pelotas – joaoimb97@hotmail.com
[2]Universidade Federal de Pelotas – sofiapagliarini@gmail.com
[2]Universidade Federal de Pelotas – marlon.cely@ufpel.edu.br
[2]Universidade Federal de Pelotas – j.v.a.ramalho@gmail.com

1. INTRODUÇÃO

Grandes áreas na indústria adotam inteligência artificial para aprimorar seus processos de identificação de falhas e classificação de padrões utilizando metodologias cada vez mais precisas e eficientes. Essas técnicas são amplamente aplicadas em diversos setores, como no monitoramento de fluxo bifásico (JIANG et al., 2024), na manutenção preditiva em sistemas mecânicos (SHCERBAKOV; SAI, 2022) e na classificação de sinais de áudios e sons (GUSMÃO et al., 2023).

A classificação de características ou padrões, ainda é feita de forma subjetiva em algumas indústrias, confiando na visão humana por meio de câmeras, observações diretas e experiência prática. No entanto, em setores como o de petróleo e gás, isso pode gerar riscos financeiros e humanos (LI et al., 2023). Já o uso de técnicas automatizadas para esses processos traz benefícios, como maior eficiência na classificação de padrões e prevenção de falhas, garantindo uma operação mais segura e eficiente (PALLA; PANI, 2023).

As diferentes áreas de pesquisa lidam, em geral, com séries temporais de dados que apresentam padrões a serem identificados, sendo uma aplicação correlata a identificação de sinais voz. Neste trabalho, foram utilizadas as Transformadas de Fourier Contínua (CFT) e de Gabor (STFT) para processar sinais de voz de seis pessoas que repetiram a palavra ‘café’ 100 vezes. A classificação desses sinais foi realizada com uma Rede Neural Convolucional (CNN), após a conversão dos dados em imagens geradas das duas transformadas. Os resultados indicam que a Transformada de Gabor, combinada com a CNN, obteve a melhor acurácia, próxima de 100%, enquanto a Transformada de Fourier Contínua atingiu 76%.

2. METODOLOGIA

O estudo seguiu três etapas para seu desenvolvimento, a estrutura é evidenciada a seguir:

A. Detecção de Nível de Ruído

Foi desenvolvido um algoritmo na linguagem de programação Pyhton com a função de detectar e remover amplitudes baixas ou nulas das amostras de áudio de seis pessoas, que repetiram a palavra ‘café’ 100 vezes. Assim, foram gerados 100 áudios por pessoa, totalizando 600 arquivos. Esses áudios foram verificados para garantir que apresentassem baixo nível de ruído.

B. Pré-processamento de Dados

Nesta fase, os dados são ajustados para se adequarem às técnicas utilizadas no treinamento das redes neurais. Foram aplicadas duas transformadas, que geraram várias imagens representativas dos áudios da palavra ‘café’ pronunciada por cada pessoa. Essas imagens servem como entrada para a arquitetura da CNN. Um exemplo dessas imagens dos sinais de áudio pode ser visualizado na Figura 1.

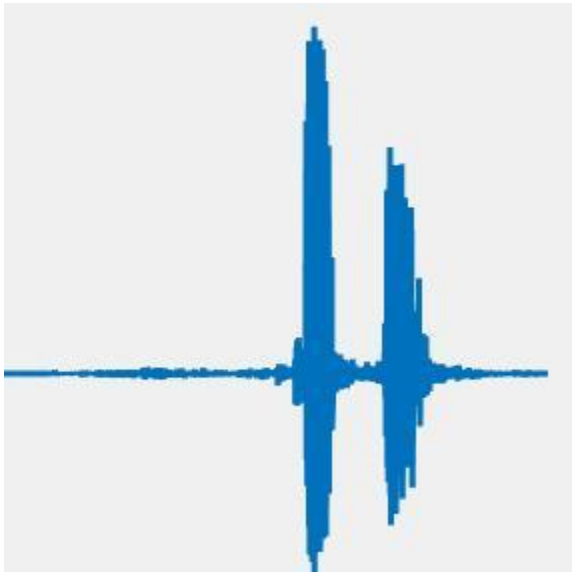

Figura 1. Espectrograma (imagem) obtida ao passar pelo pré-processamento.

C) Aplicação da CNN:

Na Rede Neural Convolucional (CNN), foram fornecidas imagens de 227 × 227 pixels como entrada. A arquitetura da rede consiste em três camadas Convolucionais, cada uma utilizando a função de ativação ReLU (Rectified Linear Unit – Unidade Linear Retificada) e contendo tem 32 filtros, que são dobrados a cada ativação Relu, seguidas por uma camada dropout (desligamento de neurônios), para reduzir ajuste excessive do modelo aos dados de treinamento (overfitting). A classificação é realizada em uma camada densa de seis unidades e ativação softmax (função de ativação que transforma os valores de saída em probabilidade, permitindo a seleção da classe provável). Os dados foram divididos em 75% para treinamento e 25% para testes. Como cada participante forneceu 100 amostras, objetiva-se obter 25 acertos na diagonal principal da matriz de confusão, indicando o desempenho do classificador.

3. RESULTADOS E DISCUSSÃO

Para gerar os resultados deste trabalho com a arquitetura CNN, utilizou-se a métrica de acurácia, que avalia a proporção de classificações corretas em relação ao total de dados.

A Figura 2 apresenta a matriz de confusão da arquitetura CNN para a classificação de fala ao longo de 100, considerando o pré-processamento com a Transformada de Fourier Contínua. A acurácia nos dados de treinamento e validação apresentam um desempenho modesto em torno da 20ª época. A acurácia média de treinamento foi de 45,15%, enquanto a da validação foi de 53,34%. Os valores de perda corroboram essa observação, com uma perda média de treinamento de 1,3380 e uma perda média de validação de 1,2020. Ao final, a acurácia do modelo foi de 76%.

Figura 2. Matriz de Confusão usando o pré-processamento da CFT.

Na Figura 3, a matriz de confusão revela que a CNN classificou a voz das seis pessoas com alta acurácia. As vozes de Eden, Jairo, Letícia, Roney e Sofia foram classificadas com 100% de acerto. A voz de João apresentou uma única classifcação incorreta, sendo equivocadamente atribuída a voz de Eden. No total, foram obtidas 149 classifcações corretas em 150 amostras de teste, resultando em uma acurácia de 99,33%.

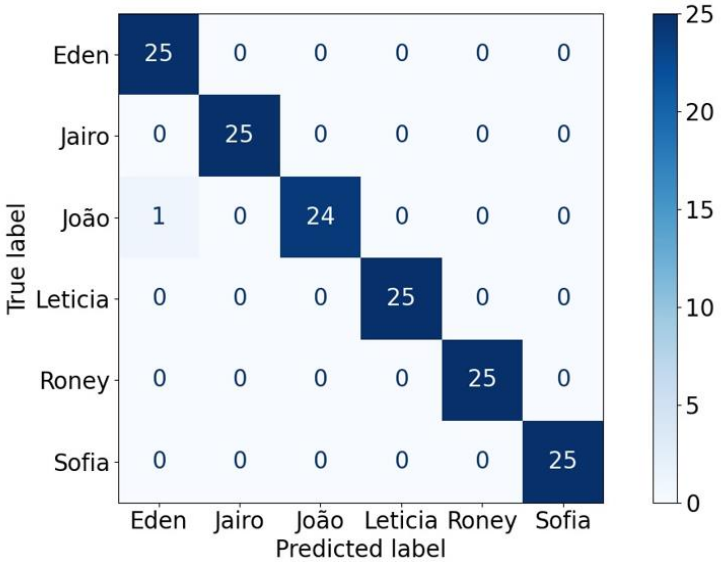

Figura 3. Matriz de confusão usando o pré-processamento da STFT.

Com isso, nota-se nos resultados obtidos na Figura 3 que a Rede Neural Convolucional (CNN) é muito eficiente em classificar padrões utilizando imagens.

4. CONCLUSÕES

Foi desenvolvido um algoritmo para classificar as vozes de seis pessoas utilizando as Transformadas de Fourier Contínua, de Gabor e Redes Neurais Convolucionais (CNN). O estudo demonstrou que a CNN é eficaz na classificação de fala quando os dados são pré-processados com a Transformada de Gabor, alcançando uma acurácia de 99,33%. Em contrapartida, a Transformada de Fourier Contínua obteve uma acurácia de 76%. O algoritmo foi capaz de filtrar os dados com essas transformadas para alimentar a CNN, ressaltando a eficácia da abordagem no processamento de imagens.

5. AGRADECIMENTOS

Agradeço ao João Inácio Moreira Bezerra, Sofia Pagliarini, Letícia Barros Dias Soares, Eden Taylor Dala Barba e Jairo Valoes de Alencar Ramalho pelo apoio nos dados de áudios fornecidos para o desenvolvimento deste trabalho.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Gusmão, R. W. M., Nacif, J. A. M., & Vieira, A. B. (2023). **Estudo de Técnicas de Deep Learning na Classificação de Amostras de Áudio de Instrumentos Musicais**.

Jiang, Y., Liu, Y., Mao, B., Lu, X., Li, Y., & Peng, L. (2024). Um método de estimativa de vazão para fluxo bifásico gás-água baseado em sensores multimodais e modelo híbrido LSTM-CNN. **IEEE Transactions on Instrumentation and Measurement**.

Li, W., Shang, Z., Zhang, J., Gao, M., & Qian, S. (2023). A novel unsupervised anomaly detection method for rotating machinery based on memory augmented temporal convolutional autoencoder. **Engineering Applications of Artificial Intelligence**, 123.

Palla, G. L. P., & Pani, A. K. (2023). Independent component analysis application for fault detection in process industries: Literature review and an application case study for fault detection in multiphase flow systems. **Measurement: Journal of the International Measurement Confederation**, 209.

Shcherbakov, M., & Sai, C. (2022). Uma estrutura híbrida de aprendizagem profunda para manutenção preditiva inteligente de sistemas ciberfísicos. **ACM Transactions on Cyber-Physical Systems (TCPS),** 6 (2), 1–22.