Síntese de Imagens Baseadas em Dados Extraídos de Áudio

João Pedro Naves Benedito[1], **José Luis Seixas Junior**[1]

[1]Ciência da Computação
Universidade Estadual do Paraná (UNESPAR)
Apucarana – Paraná

Resumo. *Este resumo expandido apresenta um estudo sobre a síntese de imagens derivadas de dados extraídos de áudio, empregando técnicas de processamento de sinais e métodos de visualização computacional. O trabalho propõe a conversão de um sinal unidimensional — o áudio — para representações visuais bidimensionais construídas a partir de extratores como FFT, MFCC e Chroma. Os resultados demonstram que diferentes descritores espectrais produzem padrões visuais particulares, sugerindo potencial para aplicações artísticas, performáticas e analíticas.*

1. Introdução

A relação entre música e imagem possui um longo histórico teórico e experimental, explorado por artistas, matemáticos e pesquisadores da percepção. Com o avanço das técnicas computacionais, tornou-se possível tratar o áudio não apenas como um sinal sonoro, mas como um conjunto estruturado de dados capaz de ser reinterpretado visualmente[1, 2].

A expansão dimensional, isto é, a conversão de um sinal unidimensional para uma representação bidimensional, permite revelar informações que não são diretamente perceptíveis na forma original do som. Este trabalho investiga como diferentes extratores de características podem produzir imagens coerentes com as propriedades temporais e espectrais do áudio, explorando a síntese visual a partir de FFT, MFCC e Chroma[3, 4].

2. Fundamentação Teórica

A análise de áudio envolve etapas de transformação que possibilitam observar frequência, energia e padrões temporais. Inicialmente, utiliza-se a Transformada Rápida de Fourier (FFT), que decompõe o sinal em suas frequências constituintes, permitindo identificar regiões de maior intensidade espectral. A Figura 1 apresenta um exemplo dessa análise aplicada a um trecho sonoro simples.

Além da FFT, emprega-se o MFCC, que utiliza escalas perceptuais para aproximar o processamento humano do timbre. Os coeficientes resultantes fornecem uma representação compacta e útil de características tímbricas. Em contraste, o Chroma organiza frequências de acordo com classes de altura musical, oferecendo uma visão diretamente relacionada às notas presentes no material sonoro[5, 6, 7].

Essas técnicas complementares são essenciais para produzir imagens que refletem estrutura harmônica, timbre e distribuição espectral.

3. Experimentos

Para avaliar a qualidade visual e a coerência dos padrões gerados, foram utilizados áudios de três categorias: baixo elétrico, guitarra distorcida e vocais. Cada extrator produziu uma representação distinta, e sua combinação permitiu explorar diferentes dimensões do som.

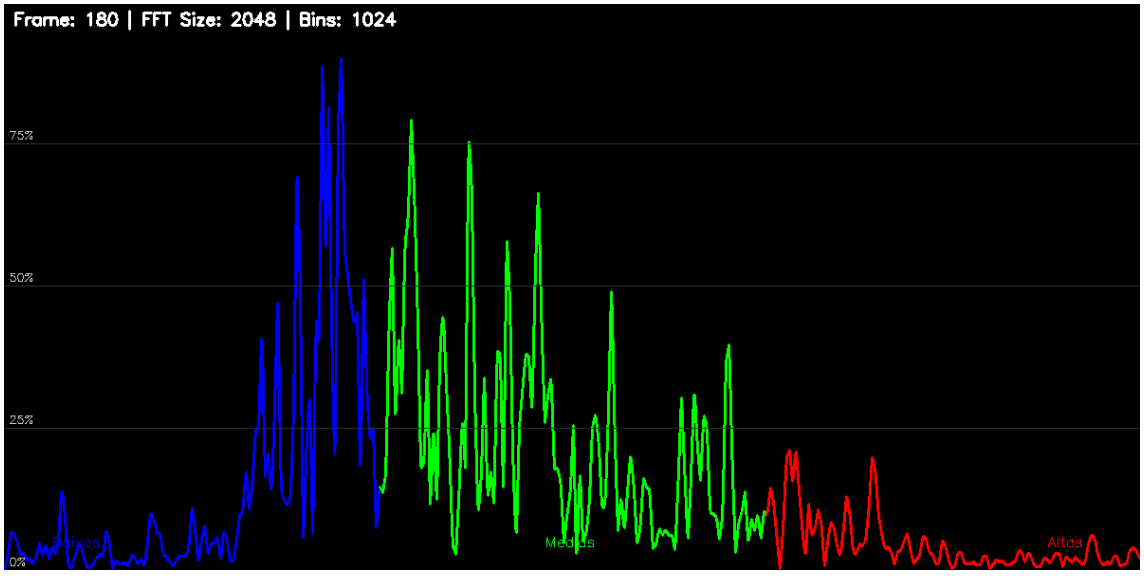

Figura 1. Representação em frequência obtida pela FFT.

A análise pelo MFCC, por exemplo, evidenciou diferenças marcantes entre instrumentos, como mostrado na Figura 2. Nesse caso, o baixo apresentou coeficientes mais estáveis e consistentes, refletindo a predominância de frequências graves e menor variação harmônica.

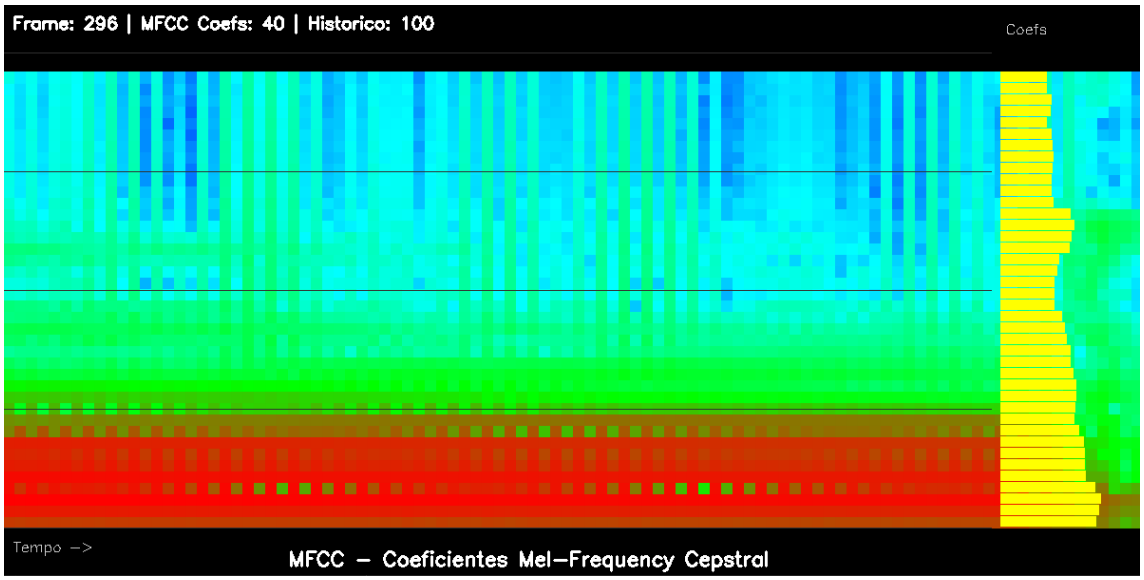

Figura 2. MFCC aplicado a um trecho de baixo elétrico.

Para sinais vocais, o Chroma se mostrou mais adequado, pois realça variações melódicas e mudanças rápidas de altura. Isso é especialmente evidente em trechos com maior expressividade melódica ou com presença de vibrato.

4. Resultados

A síntese visual final utilizou uma junção de FFT (para estrutura), MFCC (para textura) e Chroma (para coloração baseada nas notas). Em vez de empregar todas as figuras in-

termediárias, apresentamos apenas um dos quadros finais, ilustrando a fusão entre os três tipos de informação.

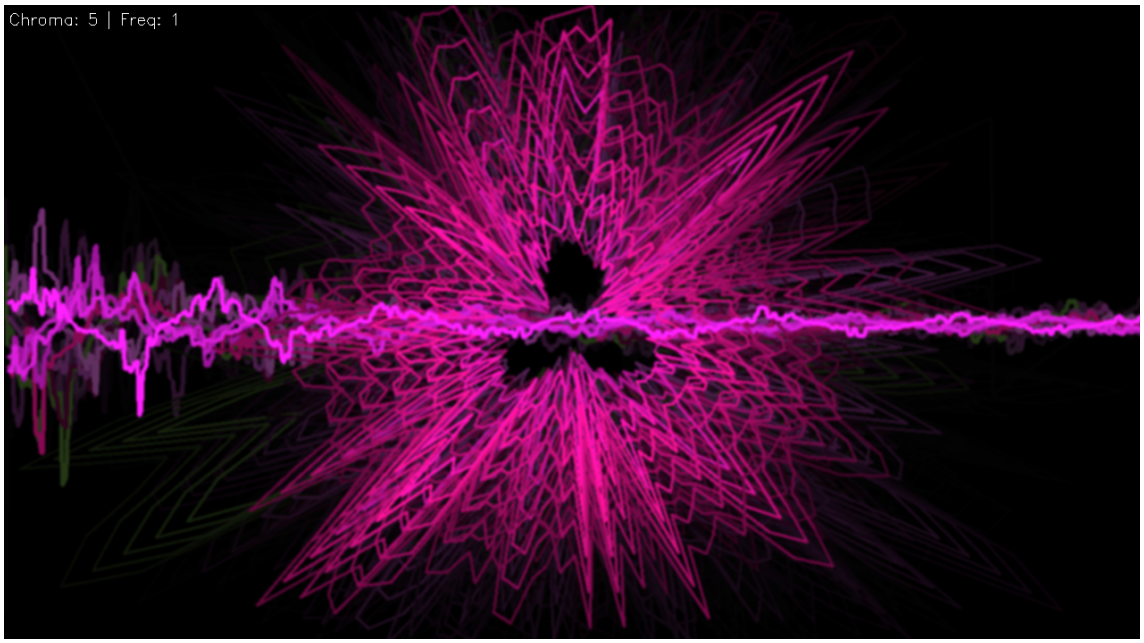

Figura 3. Frame final da síntese de imagem baseada em FFT + MFCC + Chroma.

A combinação das técnicas demonstrou que cada extrator contribui de maneira específica para o resultado visual: a FFT define contornos e distribuições espectrais, o MFCC adiciona densidade e irregularidades tímbricas e o Chroma introduz padrões relacionados à harmonia. Assim, a imagem resultante representa uma síntese coerente entre a estrutura e a expressividade musical.

5. Conclusão

Este trabalho demonstrou a viabilidade da construção de imagens a partir de dados extraídos de áudio utilizando três técnicas complementares. A abordagem evidenciou que sinais sonoros carregam informações estruturais ricas, capazes de serem reinterpretadas visualmente de maneira consistente. Como trabalhos futuros, pretende-se investigar outras formas de composição visual, integrar processamento em tempo real e explorar aplicações artísticas interativas.

Referências

[1] Michael Gartrell. *Cores e Sons Em Port-Royal e Newton: Uma Análise Foucaultiana Da Representação e Sinestesia Clássicas*. Tese de doutorado, Instituição não especificada, 2024.

[2] Jamie Ward. *Synesthesia*. MIT Press, Cambridge, MA, 2013.

[3] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4 edition, 2018.

[4] J. Nathan Kutz. *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*. Oxford University Press, 2013.

[5] William Brent. *Physical and perceptual aspects of percussive timbre*. University of California, San Diego, 2010.

[6] Ayush Shah, Manasi Kattel, Araju Nepal, and D. Shrestha. Chroma feature extraction. 01 2019.

[7] Devayani Hebbar and Vandana Jagtap. A comparison of audio preprocessing techniques and deep learning algorithms for raga recognition, 2022. URL `https://arxiv.org/abs/2212.05335`.