Universidade Estadual do Paraná
Campus Apucarana

JOÃO PEDRO NAVES BENEDITO

SÍNTESE DE IMAGENS BASEADAS EM DADOS EXTRAÍDOS DE ÁUDIO

APUCARANA–PR

2025

JOÃO PEDRO NAVES BENEDITO

SÍNTESE DE IMAGENS BASEADAS EM DADOS EXTRAÍDOS DE ÁUDIO

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Paraná para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. José Luis Seixas Junior

APUCARANA–PR

2025

Ficha catalográfica elaborada pelo Sistema de Bibliotecas da UNESPAR e
Núcleo de Tecnologia de Informação da UNESPAR, com Créditos para o ICMC/USP
e dados fornecidos pelo(a) autor(a).

Naves Benedito, João Pedro
Síntese de Imagens Baseadas em Dados Extraídos de Áudio / João Pedro Naves Benedito. -- Apucarana-PR,2025.
46 f.

Orientador: José Luis Seixas Junior.
Trabalho de Conclusão de Curso, Ciência da Computação - Universidade Estadual do Paraná, 2025.

1. Relação Áudio Imagem. 2. Conversão de Dados. 3. Extração de Características. I - Seixas Junior, José Luis (orient). II - Título.

João Pedro Naves Benedito
Síntese de Imagens Baseadas em Dados Extraídos de Áudio/ João Pedro Naves Benedito. – Apucarana–PR, 2025-
46 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. José Luis Seixas Junior

– Universidade Estadual do Paraná, 2025.

1. Extração. 2. Frequência. 3. Síntese. I. José Luis Seixas Junior. II. Universidade Estadual do Paraná. III. Faculdade de Ciência da Computação. IV. Síntese de Imagens Baseadas em Dados Extraídos de Áudio

CDU 02:141:005.7

JOÃO PEDRO NAVES BENEDITO

SÍNTESE DE IMAGENS BASEADAS EM DADOS EXTRAÍDOS DE ÁUDIO

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Paraná para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Prof. Dr. José Luis Seixas Junior
Universidade Estadual do Paraná
Orientador

Prof. Dr. Lailla Milainny Siqueira Bine
Universidade Estadual do Paraná

Prof. Dr. Paulo Roberto de Oliveira
Universidade Estadual do Paraná

Apucarana–PR, 17 de dezembro de 2025

Este trabalho é dedicado a minha família e amigos.

AGRADECIMENTOS

Agradeço ao meu orientador José Luis Seixas Júnior.

"Be something more than what you see in the mirror, let the world be your mirror."
Dwayne Michael "Lil Wayne" Carter Jr(Cherry Bomb the Documentary).

BENEDITO, J. P. N.. **Síntese de Imagens Baseadas em Dados Extraídos de Áudio**. 46 p. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual do Paraná, Apucarana–PR, 2025.

RESUMO

A tentativa da representação visual da música não é algo novo, a muitos anos diversos autores já abordaram essa temática sob diferentes perspectivas, em sua maioria com um caráter mais artístico e subjetivo. Contudo, com o avanço da visão computacional, dos estudos da computação gráfica e dos processamentos de sinais digitais, tornou-se possível explorar essa temática de forma mais objetiva. Este trabalho propõe uma abordagem computacional dessa representação visual do som, convertendo dados unidimensionais em dados bidimensionais correspondentes. Utilizando técnicas de extração de características, e as de síntese de imagens, os dados de um arquivo de áudio podem ser convertidos em dados bidimensionais de uma imagem, criando uma representação visual que corresponde diretamente ao comportamento das variações de frequências, amplitude e energia do som em tempo real. Esse estudo visa possibilitar uma experiência sonora e visual diretamente ligada e precisa, de arte generativa.

Palavras-chave: Extração. Frequência. Síntese.

BENEDITO, J. P. N.. **Image Synthesis from Audio-Extracted Data**. 46 p. Final Project (Bachelor of Science in Computer Science) – State University of Paraná, Apucarana–PR, 2025.

ABSTRACT

The attempt to visually represent music is not a new concept, for many years, several authors have explored this theme from different perspectives, generally with a more artistic and subjective character. However, with the advancement of computer vision, computer graphics, and digital signal processing, it has become possible to investigate this subject in a more objective and systematic way. This work proposes a computational approach to the visual representation of sound by converting one-dimensional data into corresponding two-dimensional data. Using feature extraction and image synthesis techniques, the data from an audio file can be transformed into two-dimensional image data, creating a visual representation that directly corresponds to the behavior of frequency, amplitude, and energy variations of the sound in real time. This study aims to enable a directly linked and precise auditory and visual experience through generative art.

Keywords: Extraction. Frequency. Synthesis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

FFT	*Fast Fourier Transform*
MFCC	*Mel Frequency Cepstral Coefficients*
RGB	*red, green, blue*
CMYK	*cyan, magenta, yellow, key*
MB	*mega bytes*
STFT	*Short Time Fourier Transform*
DTF	Transformada Discreta de Fourier
CQT	*Constant-Q Transform*
CENS	*Chroma Energy Normalized*

SUMÁRIO

1 INTRODUÇÃO

A associação entre imagem e som não é algo recente; há séculos cientistas e artistas exploram a relação entre esses domínios em estudos, experimentos e performances, especialmente na intersecção entre artes visuais e música. No campo da expressão artística musical, a vinculação entre cor e som aparece de modo recorrente, articulando-se tanto em teorias quanto em práticas criativas [1, 2].

A similaridade com as ondas mecânicas do som neste aspecto é imediata, e Newton, tendo observado sete cores na decomposição da luz (em referência direta às sete notas da escala diatônica), foi o primeiro a colocar comparativamente o som e a cor lado a lado, presumindo que cada cor corresponderia a uma nota. Assim, produziu dois discos: um contendo as sete cores do espectro visível, que ao ser girado resulta na percepção do branco; e outro no qual cores são associadas às notas musicais [3, 4].

A conversão de dados constitui um processo central em sistemas computacionais modernos, permitindo que informações originalmente codificadas em um formato, estrutura ou domínio sejam transformadas em outra representação adequada a um objetivo específico. Essa transcodificação possibilita operações como análise, visualização, compressão ou síntese, e é amplamente utilizada em aplicações científicas, artísticas e tecnológicas [5].

Representar dados em domínios alternativos permite explorar propriedades estruturais não visíveis em sua forma original, além de facilitar interpretações computacionais que dependem de organização, escala ou granularidade diferenciadas. Assim, a conversão de dados não se limita a uma operação técnica, mas constitui uma etapa conceitual essencial em fluxos de processamento contemporâneos [6].

A transposição de dados para dimensões superiores viabiliza novas classes de análise e visualização, permitindo combinar informação temporal ou sequencial com representação espacial. É justamente esse princípio que fundamenta transformações nas quais a dimensão adicional atua como um eixo interpretativo capaz de revelar propriedades antes implícitas na estrutura original dos dados[7, 8].

O maior obstáculo para performances que combinam arte visual e música é a criação em tempo real, pois não é viável que o performer produza simultaneamente ambas as expressões. Por isso, a maior parte das obras que articulam música e imagem permanece ligada a escolhas estéticas, subjetivas ou simbólicas, sem uma correspondência direta e calculável entre dados acústicos e elementos visuais [9, 10].

A relação proposta neste trabalho entre dados acústicos e representações visuais também se insere no campo das discussões sobre sinestesia, especialmente no que se refere às formas de correspondência intermodal estudadas na neurociência e na psicologia perceptual. A sinestesia genuína, conforme descrevem Cytowic e Eagleman [11], caracteriza-se pela automaticidade e consistência de associações intersensoriais que ocorrem sem esforço consciente, distinguindo-se de qualquer mecanismo computacional ou simbólico. De modo semelhante, Hubbard e Ramachandran [12] afirmam que a experiência sinestésica não envolve tradução voluntária entre modalidades, mas ativação cruzada entre áreas cerebrais adjacentes, indicando bases neurocognitivas específicas para o fenômeno.

Distinto desse processo biológico, o presente trabalho realiza um mapeamento sistemático entre parâmetros sonoros e atributos visuais, alinhando-se às estratégias formais de visualização discutidas por Bertin [13], toda representação gráfica depende da escolha adequada de variáveis visuais capazes de expressar variações de dados. Assim, embora não busque reproduzir a experiência subjetiva da sinestesia, a metodologia adotada permite refletir sobre aproximações conceituais entre percepção multimodal e modelos computacionais de tradução sensorial. Tal abordagem também dialoga com discussões recentes

sobre visualização musical e correlações entre som e cor, como argumenta Gartrell [14].

O presente trabalho propõe transcrever dados unidimensionais, como arquivos de áudio, em dados bidimensionais, como imagens. A conversão ocorre por meio da extração de características espectrais e temporais, representando diretamente variações de frequência, amplitude e energia presentes no sinal sonoro. Esse processo considera igualmente o fator temporal, produzindo visualizações em tempo real dos trechos analisados.

A conversão é realizada mediante diferentes técnicas de extração e processamento de sinais, cujas saídas se complementam na construção de uma composição visual única. Dessa forma, o método integra múltiplas abordagens analíticas para gerar representações diretamente derivadas do comportamento acústico do áudio.

O objetivo deste trabalho é extrair, por meio de técnicas selecionadas, as características necessárias para a síntese de imagens em tempo real tomadas a partir de um arquivo de áudio, produzindo sequências visuais que funcionem como representações da estrutura sonora. A ênfase das análises recai sobre atributos espectrais, dos quais derivam imagens sucessivas capazes de traduzir, sem simbolismo subjetivo, as dinâmicas presentes no sinal.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica aborda as definições técnicas dos tipos de arquivos utilizados no trabalho e das técnicas de extração utilizadas nos dados obtidos.

Amostras de áudio bruto formam um sinal unidimensional em série temporal, que é fundamentalmente diferente de imagens bidimensionais. Sinais de áudio são comumente transformados em representações tempo-frequência bidimensionais para processamento, mas os dois eixos, tempo e frequência, não são homogêneos como os eixos horizontal e vertical de uma imagem. Imagens são capturas instantâneas de um alvo e geralmente analisadas como um todo ou em partes, com poucas restrições de ordem; no entanto, sinais de áudio precisam ser estudados sequencialmente, em ordem cronológica. Essas propriedades deram origem a soluções específicas para áudio[15].

Em um arquivo de áudio, podemos observar dois tipos de características presentes, a **Frequência** e **Amplitude**. Frequência é a velocidade da vibração, o que define a altura, no quesito musical. Ela só é útil ou significativa para sons musicais, nos quais há uma forma de onda fortemente regular. A frequência é medida como o número de ciclos de onda que ocorrem em um segundo. A unidade de medida da frequência é o Hertz.

Amplitude é o tamanho da vibração, e isso determina o quão alto é o volume do som. Já vimos que vibrações maiores produzem sons mais altos. A amplitude é importante ao equilibrar e controlar o volume dos sons, como no controle de volume de um reprodutor de CD. O termo nota é utilizado na música de forma ampla, podendo referir-se tanto ao símbolo musical quanto ao som dotado de altura percebida. Cada nota apresenta atributos que determinam sua duração relativa e altura (*pitch*), relacionando-se à frequência fundamental do som produzido[16].

O conceito de *pitch* corresponde a uma propriedade perceptiva que permite ordenar os sons em uma escala de frequência, de modo que notas com frequências fundamentais em razão de potências de dois (metade, dobro etc.) são percebidas como semelhantes — fenômeno que dá origem à noção de classe de altura (*pitch class*) e oitava.

A discretização do contínuo de alturas leva à definição de escala musical, entendida como um conjunto finito de alturas (notas musicais) representativas distribuídas dentro de uma oitava. Müller [17] destaca ainda que diferentes culturas e períodos históricos propuseram divisões distintas desse espaço sonoro, não havendo uma escala universalmente válida: a adequação de uma escala depende do tipo de música, do instrumento e do contexto cultural no qual é empregada.

Diferente do áudio a imagem possui duas dimensões, a renderização de uma cena 3D produz um *array* de pixels (contração de *picture element*) chamado *raster*, que será exibido em algum tipo de matriz 2D, como um monitor, ou gravado em um arquivo de imagem. Os pixels são modelos computacionais da cor natural, geralmente processados em valores ponto flutuante durante o processo de renderização, podendo ser descritos como $p \in \mathbb{R}, 0 \leq p \leq 1$, sendo 0 ausência de cor e 1 a intensidade total da cor. Cada pixel é descrito como um *array* de três valores usados para representar as três frequências luminosas percebidas pelas células cones presentes no olho humano:

Vermelho, verde e azul (*red, green, blue* - RGB), respectivamente. A variação de intensidade em cada canal RGB é responsável pela percepção das demais cores, tais como amarelo (1, 1, 0) e cinza (0.5, 0.5, 0.5) [18].

É possível existir ainda um canal adicional *alpha* para dar suporte a imagens com transparência, geralmente gravadas em arquivos com extensão `.png`. Existem modelos de cores que se baseiam em outros parâmetros para a formulação da cor, como o modelo CMYK, que ao contrário do RGB (que é um modelo aditivo de cores para dispositivos que emitem luz), trata-se de um modelo subtrativo de cores baseado na absorção da luz, utilizado para impressão de imagens em papel [18].

Uma imagem digital possui dois atributos que influenciam sua qualidade, a resolução espacial, que é a dimensão $m \times n$ da matriz 2D da imagem, determinando assim a sua quantidade total de pixels, e a quantização de cor, que é a quantidade N de bits utilizados por cada canal de cor para representar as variações de cores possíveis.

Ou seja, em um *display* de 24 bits (8 bits para cada canal RGB, gerando $2^8 = 256$ valores distintos por canal) é possível representar $2^{24} = 16.777.216$ de cores. O tamanho de arquivo sem compressão que uma imagem RGB de 24 bits ocupa é dado pela $largura \times altura \times N$, logo, uma imagem em resolução *full hd* ocupa $1920 \times 1080 \times 24 = 49.766.400 bits$, ou aproximadamente $6.22 MB$ (*mega bytes*)[18]. Para extração de características foram selecionadas técnicas com mais material e estudos sobre elas, também as mais utilizadas, por serem mais reconhecíveis e com objetivos mais específicos e claros, tornando-as mais reconhecíveis.

Podemos ver na Tabela 1, a escala feita por Newton e outros demais cientistas e artistas, referentes as notas musicais e suas respectivas transcrições em cores, relacionando simbolicamente cores do espectro visível com notas musicais da escala ocidental, conceito esse que também foi abordado por outros autores, alguns também presentes na tabela, estudos esses utilizados para sinestesia e visualização musical [14].

A partir da análise apresentada por Gartrell [14], compreende-se que essas tentativas de correlacionar sons e cores não se limitam a um simples exercício especulativo, mas refletem uma epistemologia própria da era clássica, na qual diferentes domínios sensoriais eram organizados segundo princípios de proporcionalidade e ordem. O autor discute como, nesse período, a representação adquiria um papel central na construção do conhecimento, permitindo que fenômenos heterogêneos fossem interpretados dentro de um mesmo regime simbólico.

Assim, as escalas cromáticas e sonoras eram percebidas como estruturas paralelas, ambas derivadas de leis naturais consideradas universais. Esse enquadramento contribuiu para consolidar a ideia de que a percepção podia ser sistematizada matematicamente, fundamentando práticas que associavam harmonia musical, ordenação visual e, posteriormente, experimentações sinestésicas.

2.1 Transformada de Fourier de Curto Prazo

O som pode ser representado matematicamente como uma função contínua que descreve a variação da pressão do ar em relação ao tempo. Considerando um sinal analógico, tanto o tempo quanto a amplitude são grandezas reais e contínuas, de modo que o sinal pode ser modelado como uma função $f : \mathbb{R} \to \mathbb{R}$, onde a cada instante de tempo $t \in \mathbb{R}$ corresponde um valor de amplitude $f(t) \in \mathbb{R}$. O gráfico dessa função, que relaciona amplitude e tempo, constitui a forma de onda ($waveform$) do som [19].

Do ponto de vista matemático, uma função define uma relação entre um conjunto de entradas e um conjunto de saídas, de modo que cada elemento de entrada se associa exatamente a um elemento de saída. Assim, é necessário distinguir entre a própria função f e o valor resultante $f(t)$ obtido para um dado argumento t. Enquanto a matemática trata f de maneira abstrata — sem considerar o significado físico do argumento —, a engenharia frequentemente utiliza a notação $f(t)$ para enfatizar a dependência temporal do sinal. Essa abordagem permite representar formalmente a estrutura fundamental de um som

Name	Year	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
Isaac Newton	1704												
Louis Bertrand Castel	1734												
George Field	1816												
D. D. Jamson	1844												
Theodor Seeman	1881												
A. Wallace Rimington	1893												
Bainbridge Bishop	1893												
H. von Helmholtz	1910												
Alexander Scribin	1911												
August Aeppli	1940												
J.Belmont	1944												
Steve Zieverink	2004												

Tabela 1 – Relação de cor e notas musicais por diversos autores

como uma função contínua de tempo e amplitude [17].

A Transformada de Fourier de Curto Prazo (STFT) (ou *short-term Fourier transform*) é uma ferramenta de uso geral poderosa para o processamento de sinais de áudio. Ela define uma classe particularmente útil de distribuições tempo-frequência, que especificam a amplitude complexa em função do tempo e da frequência para qualquer sinal. Nos preocupamos principalmente com o ajuste dos parâmetros da STFT para aproximar a análise tempo-frequência realizada pelo ouvido humano para fins de exibição espectral e medir parâmetros de modelos em um espectro de curto prazo.

A definição matemática do **STFT** para sinais genéricos analógicos é [20]

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-iw(n-mR)} \tag{2.1}$$

onde

$x(n)$ = sinal de entrada no instante n

$w(n)$ = função janela de comprimento M

$X_m(\omega)$ = Transformada de frequencia dos dados janelados centrados no instante mR

R = tamanho do salto (*hop-size*), em amostras.

Estes sinais não são utilizados por qualquer equipamento digital, esses equipamentos digitais necessitam de uma conversão. A fórmula dada para o sinal digital é [21]

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)\left[cos(\omega) - i.sen(\omega)\right] \tag{2.2}$$

Tanto a Equação 2.1 quanto a Equação 2.2 definem as transformadas de Fourier, para sinais

analógicos e digitais, respectivamente. Essa transformada contém números complexos que não existem no mundo real, para este caso, a fórmula é

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)\,[cos(\omega)] \tag{2.3}$$

A Equação 2.3 é a Equação 2.2, porém ela não utiliza números complexos, pois em equipamentos digitais não existe forma de computar números complexos.

2.2 Mel Frequency Cepstral Coefficients

Uma das técnicas de extração utiliza *Mel Frequency Cepstral Coefficients*(MFCC) que pode ser empregados em aplicações músicais. O MFCC leva em conta a percepção não linear do som pelo ouvido humano. O que torna o uso de MFCC interessante é o fato de sua aplicação reduzir um espectro de 1024 pontos para cerca de 15 a 40 pontos que podem ser utilizados para verificar a similaridade ou distinção de sons [22]. O método foi originalmente proposto por Davis [23] para reconhecimento de fala, e posteriormente adaptado para aplicações musicais. Conforme Logan [24], os coeficientes MFCC capturam de forma compacta as propriedades espectrais relevantes para o timbre musical, permitindo a modelagem de similaridade entre canções com boa eficiência. O processamento MFCC é realizado inicialmente por um processo de janela, logo após o janelamento de sinal, é processado o DTF (Transformada Discreta de Fourier), visto na equação 2.4.

A amplitude da DTF é filtrada por janelas triangulares na escala Mel e então, aplicada o logaritmo. A Transformada Discreta de Cosseno é aplicada e os Coeficientes Mel-Cepstrais são as amplitudes resultantes.

A escala Mel é uma escala psicoacústica que explora a relação de percepção da frequência fundamental entre dois tons, criada a partir do estudo da dinâmica do sistema auditivo humano. A unidade de medida Mel (em referência a melodia) refere-se a frequência subjetiva de tons puros percebida pelo ouvido humano [22].

$$X[k] \;=\; \sum_{n=0}^{N-1} x[n]\, e^{-j\,\frac{2\pi}{N}\,kn}, \qquad k = 0, \ldots, N-1 \tag{2.4}$$

O cálculo dos *Mel Frequency Cepstral Coefficients* (MFCCs) é realizado por meio de uma sequência de transformações matemáticas que visam aproximar a percepção auditiva humana da frequência. Inicialmente, o sinal de áudio discreto $x(n)$ é dividido em janelas temporais curtas, sobre as quais se aplica a *Transformada Discreta de Fourier* (DFT) ponderada por uma janela de Hamming $w(n)$, conforme

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n)e^{-j2\pi kn/N} \tag{2.5}$$

onde N é o número de amostras por janela e $f(k) = kf_s/N$ representa a frequência associada ao índice k. A função de janela Hamming é definida como

$$w(n) = 0.54 - 0.46\cos\left(\frac{\pi n}{N}\right) \tag{2.6}$$

utilizada para reduzir descontinuidades entre segmentos adjacentes do sinal.

O espectro de magnitude $|X(k)|$ é então mapeado para a escala Mel por meio de um banco de filtros triangulares $H(k, m)$, que realiza uma compressão logarítmica da escala de frequências de modo

a refletir a resposta perceptiva do sistema auditivo humano. Cada filtro é definido em função de suas frequências centrais $f_c(m)$, calculadas através da conversão entre Hertz e escala Mel:

$$\phi = 2595 \log_{10} \left(1 + \frac{f}{700}\right), \tag{2.7}$$

e sua inversa,

$$f = 700 \left(10^{\phi/2595} - 1\right). \tag{2.8}$$

As energias filtradas são então comprimidas logaritmicamente:

$$X'(m) = \ln \left(\sum_{k=0}^{N-1} |X(k)| \cdot H(k,m)\right), \tag{2.9}$$

onde M é o número de filtros Mel.

Por fim, aplica-se a *Transformada Discreta do Cosseno* (DCT) para obter os coeficientes cepstrais,

$$c(l) = \sum_{m=1}^{M} X'(m) \cos\left[\frac{l\pi}{M}\left(m - \frac{1}{2}\right)\right], \tag{2.10}$$

os quais representam as variações espectrais de forma compacta e aproximadamente descorrelacionada. Esses coeficientes descrevem o contorno espectral do sinal e são amplamente utilizados na análise de timbre e reconhecimento de padrões acústicos. Conforme discutido por Sigurdsson [25], apenas os primeiros 15 coeficientes tendem a capturar as informações perceptualmente mais relevantes, sendo também os mais robustos a variações de codificação e compressão, como o formato MP3.

Estudos posteriores mostraram que o conjunto de coeficientes MFCC apresenta alta consistência entre diferentes implementações, especialmente nos primeiros 15 coeficientes, os quais concentram as informações perceptualmente mais relevantes do sinal de áudio. Essa característica garante boa robustez em aplicações de reconhecimento e classificação musical, mesmo quando os sinais são submetidos a compressão perceptual, como no formato MP3, desde que em taxas superiores a 128 kbit/s [25]. O método, originalmente proposto para reconhecimento de fala [23], foi posteriormente adaptado para modelagem musical e recuperação de similaridade sonora, mostrando eficácia na representação de timbre e estrutura espectral [24].

2.3 Espectograma

Com a escala Mel podemos obter um Mel-Espectograma. Um espectrograma é uma representação visual de um sinal de áudio que foi submetido a uma Transformada de Fourier de curto prazo.

A partir de um espectrograma, é possível obter informações sobre a variação da amplitude de cada frequência presente no sinal ao longo do tempo. Um mel-espectrograma é simplesmente um espectrograma no qual as frequências foram convertidas para a escala mel; uma escala na qual as distâncias entre as frequências são proporcionais à percepção do cérebro humano sobre as diferenças entre as frequências [26].

2.4 Chroma

A percepção humana de altura musical apresenta uma natureza periódica, na qual dois sons separados por uma oitava são percebidos como semelhantes em "cor" ou função harmônica. Essa relação permite decompor o conceito de altura em dois componentes principais: a altura tonal (*tone height*) e a *croma*. A altura tonal indica o número da oitava, enquanto a croma representa a identidade da nota dentro do conjunto de classes de altura definido por $\{C, C\#, D, D\#, E, F, F\#, G, G\#, A, A\#, B\}$.

Enumerando as cromas de forma discreta, pode-se associar esse conjunto ao intervalo numérico $[0 : 11]$, onde 0 corresponde à nota C, 1 a $C\#$, e assim sucessivamente até 11, que representa B. Define-se então uma *classe de altura* (*pitch class*) como o conjunto de todas as notas que compartilham a mesma croma, ou seja, todas as alturas que diferem entre si por um número inteiro de oitavas. Por exemplo, a classe de altura correspondente à croma C é composta pelo conjunto $\{\ldots, C_0, C_1, C_2, C_3, \ldots\}$, reunindo todas as notas C separadas por intervalos de oitava, o que define o *chroma* como uma classe de altura do som [17].

Na música, o termo característica de chroma ou *chromagram* está intimamente relacionado às doze diferentes classes de altura. As características baseadas em chroma, também chamadas de perfis de classe de altura (*pitch class profiles*), são ferramentas poderosas para analisar músicas cujas alturas podem ser categoricamente organizadas (geralmente em doze categorias) e cuja afinação se aproxima da escala temperada igual [16].

Uma das principais propriedades das características de chroma é que elas capturam aspectos harmônicos e melódicos da música, sendo ao mesmo tempo robustas a variações de timbre e instrumentação [16].

As características de chroma têm como objetivo representar o conteúdo harmônico (por exemplo: tonalidades, acordes) de uma janela de tempo curto do áudio. O vetor de características é extraído a partir do espectro de magnitude utilizando transformadas como a STFT, CQT, CENS, entre outras [16].

3 MÉTODO DE PESQUISA

O método de pesquisa elaborado para este trabalho segue o fluxo descrito na Figura 1, primeiro o arquivo de áudio é aberto, é separado os dados importantes para a extração, sendo descartado o cabeçalho do arquivo. Feita a primeira visualização mais simples do arquivo, com o objetivo de uma primeira representação dos dados. Assim, os dados sendo calculados pelos extratores de características e por fim, a saída da FFT é utilizada como entrada da MFCC e do Chroma, aplicando a convolução das técnicas. Os métodos e técnicas escolhidos, foram métodos que são focados no pré-processamento de áudio, extração de características do arquivo de áudio, e técnicas para a síntese de imagens baseadas nas características extraídas.

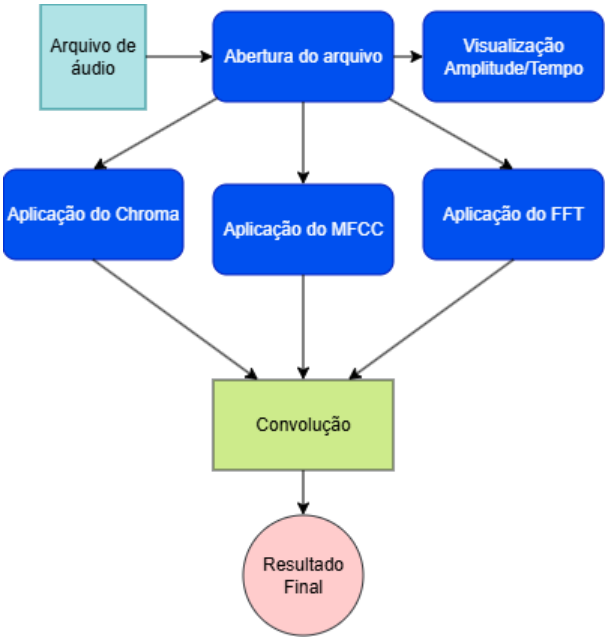

Figura 1 – Diagrama do fluxo das operações com o arquivo de áudio.

Foi escolhido o arquivo de áudio que seria utilizado durante os primeiros testes, primeiramente foram feitos testes mais simples para visualização gráfica do exemplo escolhido. A primeira conversão de áudio para imagem que foi executada foi a de amplitude por tempo, resultando na Figura 2, evidenciando a forma de onda do áudio escolhido, diferente dos seguintes experimentos, esta conversão não consiste em nenhuma forma de manipulação ou extração de características ou dados do áudio. Mais simples do que as utilizadas para o resultado final, como FFT, MFCC ou Chroma.

Utilizando-se da biblioteca do OpenCV, foram feitas as primeiras extrações de características do áudio utilizado as técnicas de extração, com o áudio de exemplo e convertidas graficamente. Com auxílio da biblioteca, os dados obtidos pela técnica da FFT aplicada sobre o áudio, foram utilizados como coordenadas em um plano de duas dimensões e convertidas em uma linha contínua que se altera de acordo com a alteração dos valores das frequências convertidas pela Equação 2.3, na Figura 3 o primeiro resultado e na Figura 4 o segundo resultado. Na Figura 4 obteve-se uma melhor visualização do resultado da FFT, separada por cor onde azul representa as frequências baixas, verde as médias e vermelho as altas.

Foram geradas dois tipos de imagens diferentes, utilizando a mesma técnica de extração do MFCC, a Figura 5 com estilo de intensidade da cor de preto até a cor branca na imagem. Quanto

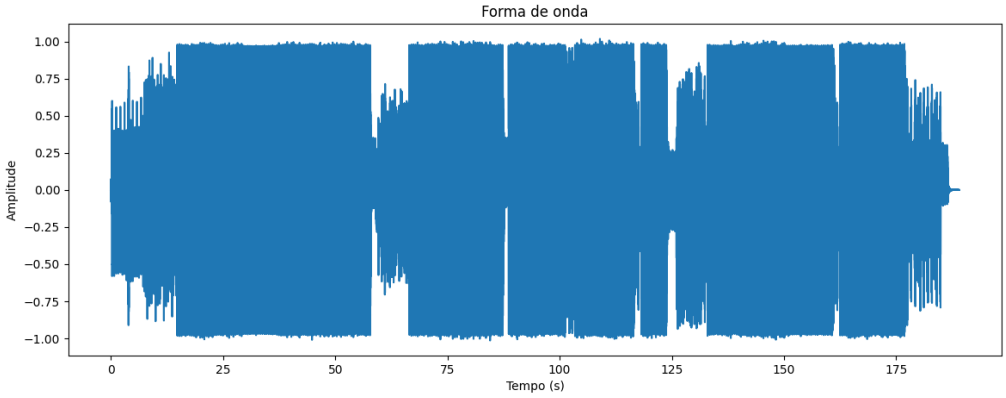

Figura 2 – Visualização gráfica de amplitude por tempo.

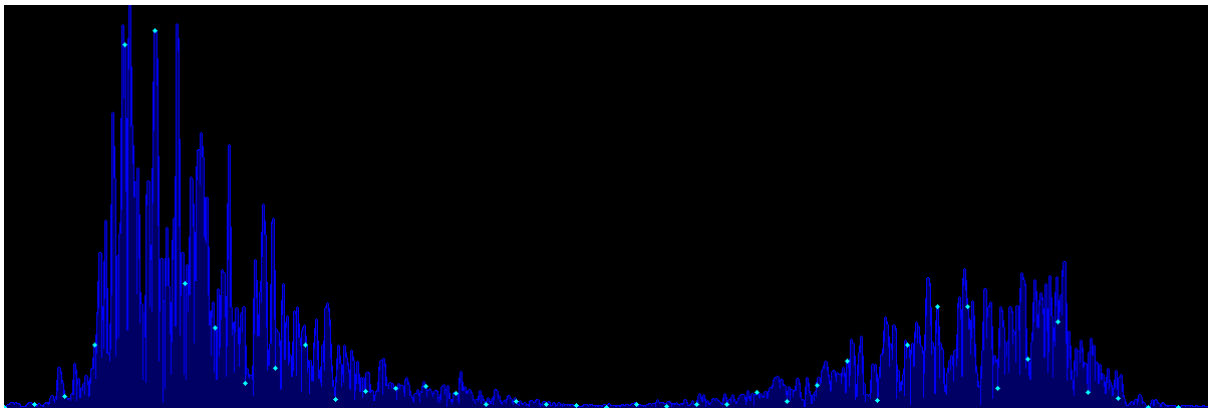

Figura 3 – *Frame* do primeiro resultado da extração da *Fast Fourrier Transform*

maior a frequência do canal, mais perto do branco e quanto mais perto do preto menos frequência. A segunda síntese do MFCC na Figura 6 foi estruturada como um mapa de calor, os valores de frequência são representadas na cor azul claro, verde e vermelho, separados em mais canais formando um gráfico que avança em função do tempo.

A coloração do Chroma foi baseada na tabela de D.D. Jameson e Alexander Scribin representada na Tabela 1, foi montada uma relação do tipo Hash Map entre chave e valor, com a chave sendo a nota musical e o valor sendo o equivalente RGB da cor. A tabela de cores foi montada como uma mescla das duas existentes para exibir uma maior diferença entre as notas mais discrepantes e uma maior proximidade entre as notas semelhantes, que pode ser vista na Tabela 2. Assim como as outras técnicas, o Chroma foi aplicado em função do tempo do arquivo de áudio, gerando uma tabela com a escala de cores onde quanto maior a clareza da frequência equivalente a nota musical maior a nitidez da cor escolhida para representação, ou seja, menos presença da cor preta na cor, Figura 7.

Tabela 2 – Escala de Cores por Notas Musicais

C	C#	D	D#	E	F	F#	G	G#	A	A#	B

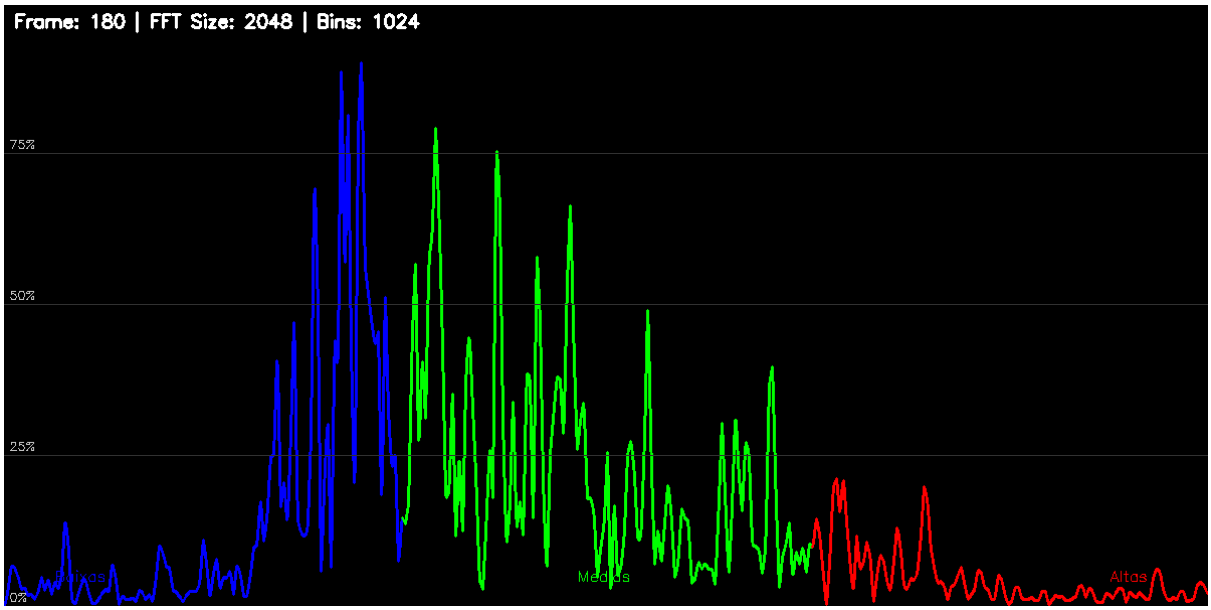

Figura 4 – *Frame* do segundo resultado da extração da *Fast Fourrier Transform*

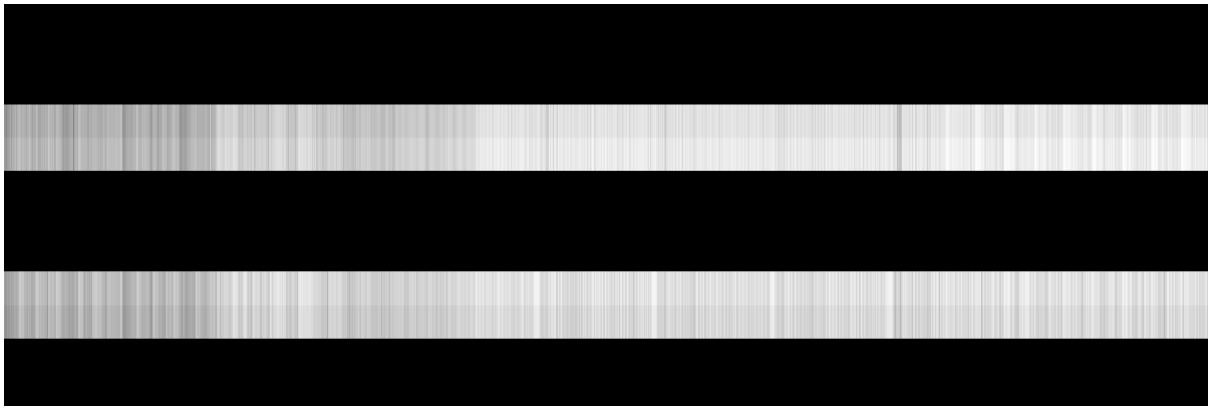

Figura 5 – Primeiro resultado da aplicação da técnica *Mel Frequency Cepstral Coefficients*

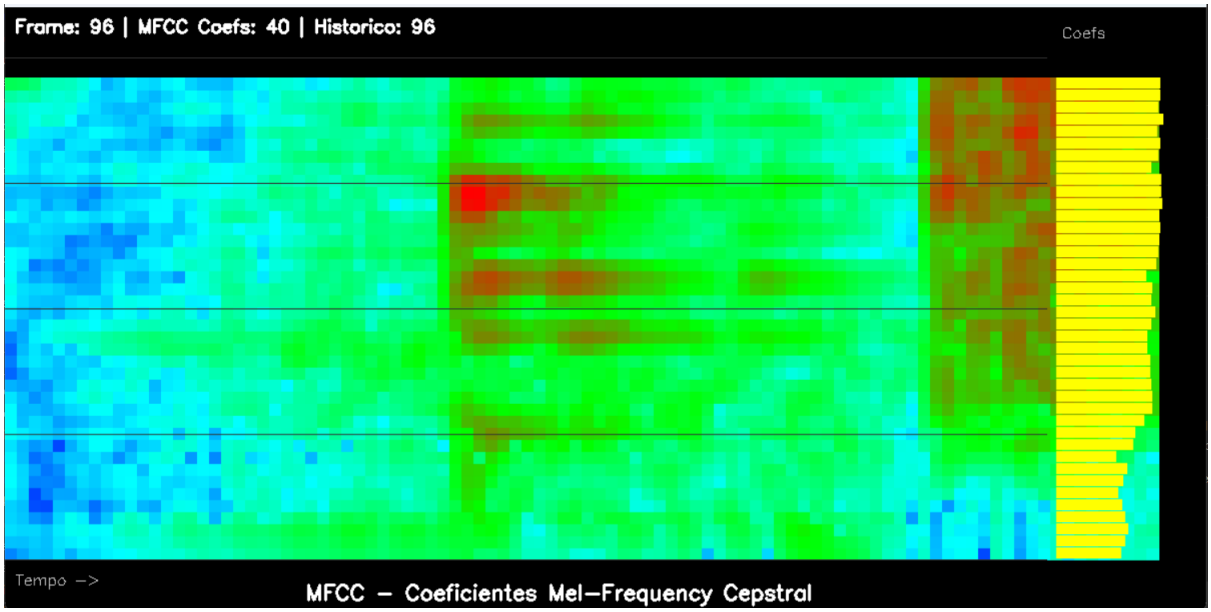

Figura 6 – Segundo resultado da aplicação da técnica *Mel Frequency Cepstral Coefficients*

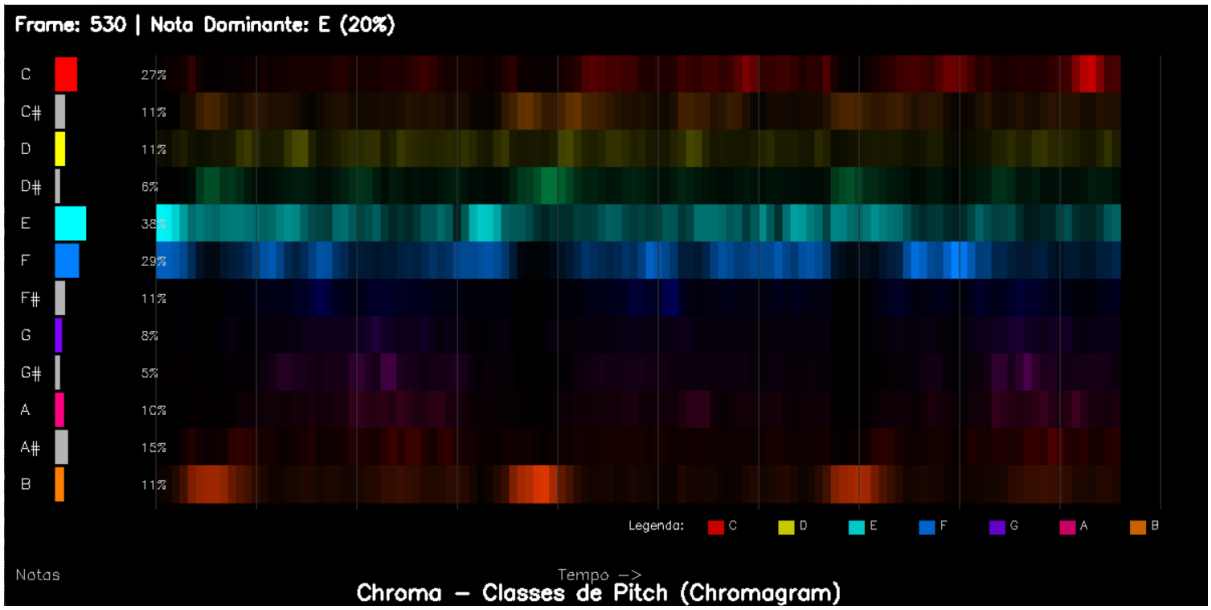

Figura 7 – *Frame* do resultado da aplicação do Chroma

4 EXPERIMENTOS

Os experimentos foram realizados com o objetivo de evidenciar as diferenças entre conteúdos distintos de arquivos de áudio e observar como as técnicas escolhidas se comportam em contextos variados. Três arquivos foram selecionados: uma gravação isolada de baixo elétrico; uma gravação de guitarra elétrica com forte distorção; e trechos vocais, tanto solo quanto coral. A análise desses áudios permite compreender como cada extração reage a timbres, frequências e contextos musicais específicos.

As extrações realizadas no áudio de baixo concentram-se em maior ocorrência em frequências mais baixas, como esperado para um instrumento grave. Já o áudio de guitarra apresenta conteúdo mais médio/baixo, porém com maior quantidade de harmônicos devido ao efeito de distorção. No caso dos vocais, a expectativa é de um chroma mais limpo no vocal solo e maior densidade de notas no coral, refletindo a sobreposição de melodias e a natureza tímbrica da voz humana.

Na Figura 8 observamos o comportamento do Chroma aplicado ao áudio do baixo. A representação mostra notas bem definidas e com espaçamentos claros, uma vez que apenas um instrumento está presente e sem sobreposição de frequências.

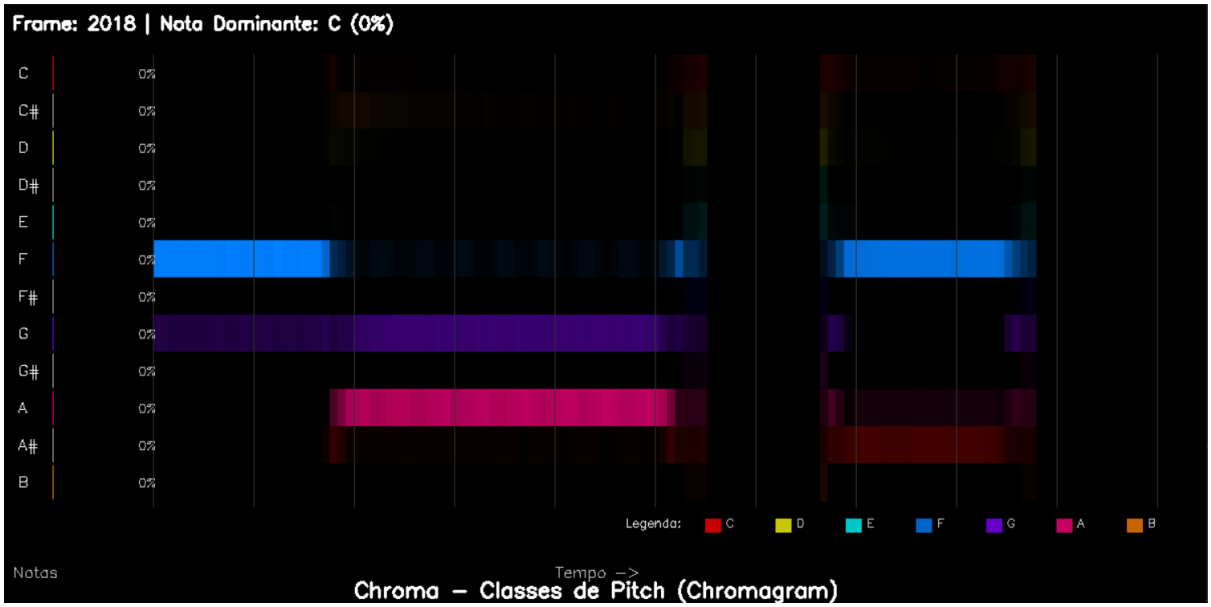

Figura 8 – *Frame* do Chroma aplicado em uma linha de baixo

A seguir, a Figura 9 apresenta o resultado da FFT para o mesmo áudio. Como esperado, a transformada evidencia a predominância de frequências graves, sem a presença de outras componentes significativas, diferentemente das Figuras 3 e 4, onde há maior distribuição devido à presença de múltiplos instrumentos e sintetizadores.

Além disso, o MFCC do áudio do baixo, representado na Figura 10, demonstra baixa variância entre os coeficientes. Isso ocorre devido ao caráter repetitivo das notas e ao espectro limitado do instrumento, contrastando com o comportamento mais variado observado em áudios completos de músicas.

No caso do áudio de guitarra elétrica distorcida, observa-se um comportamento mais complexo. A Figura 11 mostra uma distribuição de frequências mais espalhada, consequência direta da distorção e

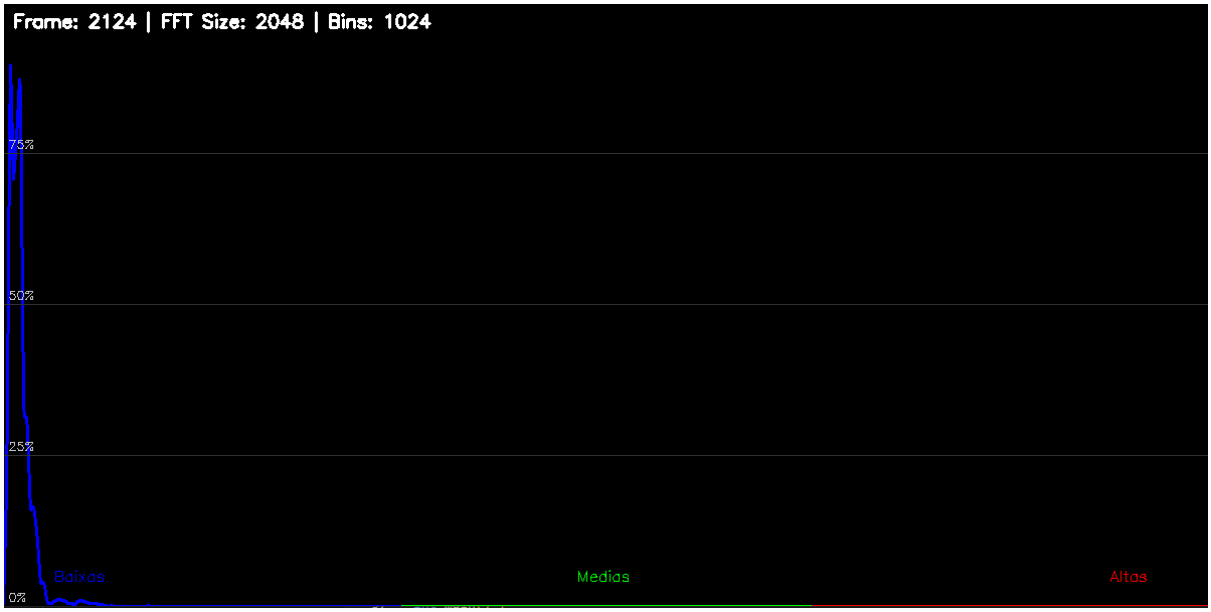

Figura 9 – *Frame* do *Fast Fourier Transform* aplicado em um áudio de baixo elétrico

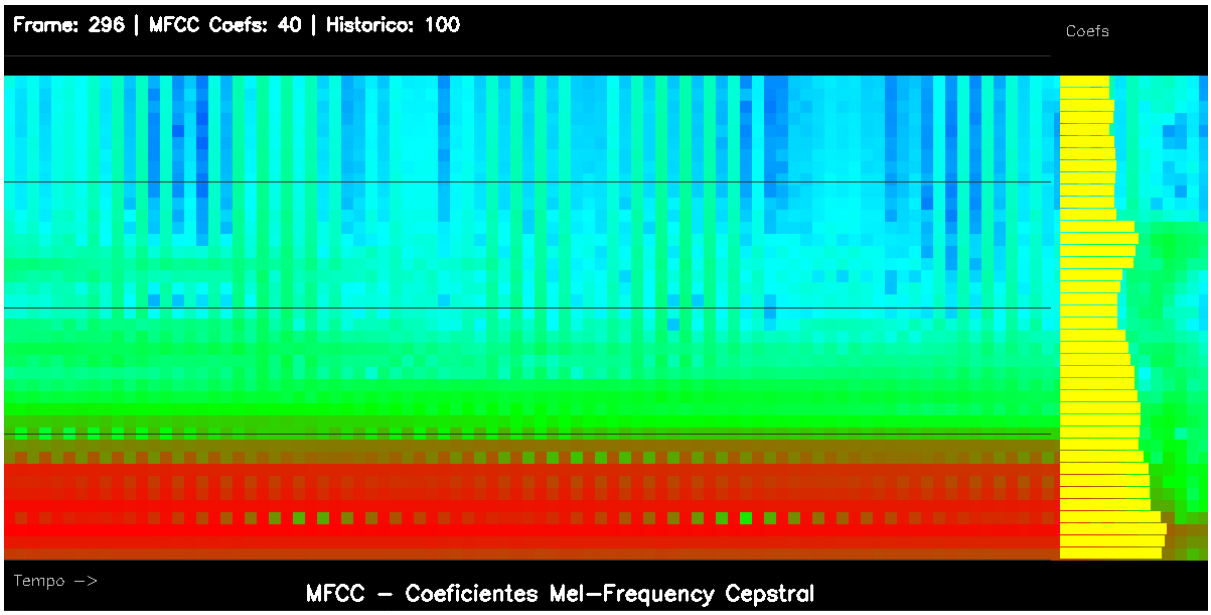

Figura 10 – *Frame* do *Mel Frequency Cepstral Coefficients* aplicado em um áudio de baixo elétrico

das notas mais agudas tocadas em relação ao baixo.

Essa maior instabilidade espectral também aparece na extração do MFCC, como pode ser visto na Figura 12. As mudanças abruptas nas cores indicam transições rápidas e variações repentinas nas características acústicas, típicas do efeito de distorção.

Para os dados vocais, foram analisados dois trechos: um vocal solo e um coral. Na Figura 13, o Chroma aplicado ao vocal solo apresenta notas mais claras e bem definidas, ainda que com variações provenientes de reverberação e da própria natureza da voz humana.

Já a Figura 14 demonstra como a presença de múltiplas vozes no coral resulta em uma distribuição mais difusa no Chroma, revelando diferentes notas simultâneas e menor intensidade de cada componente individual.

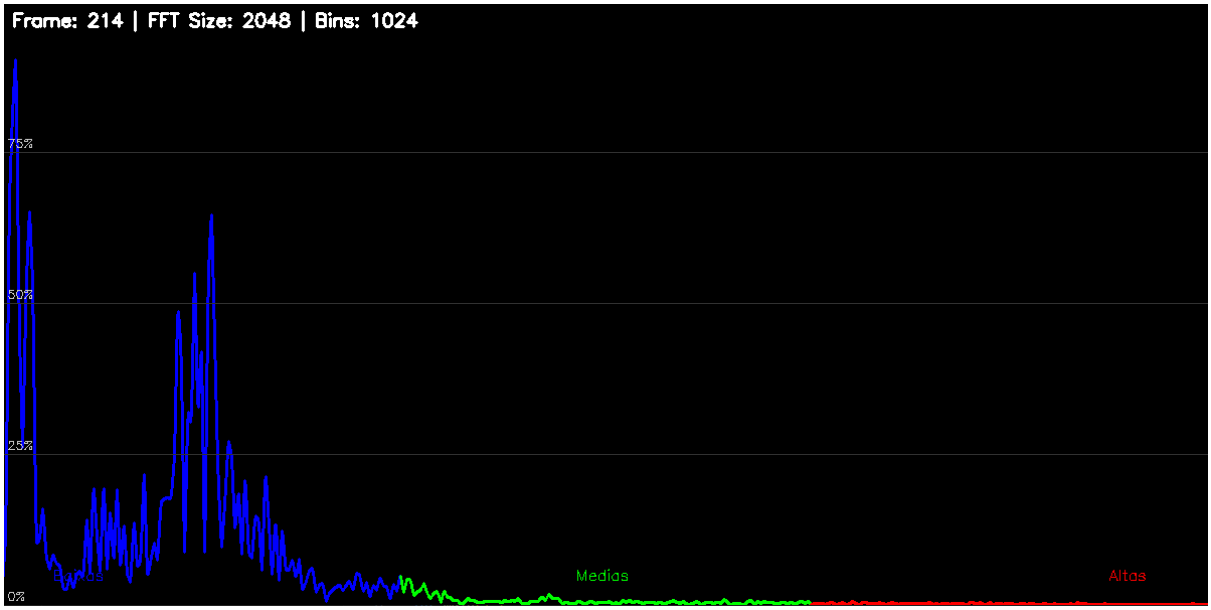

Figura 11 – *Frame* do *Fast Fourier Transform* aplicado em um áudio de guitarra elétrica

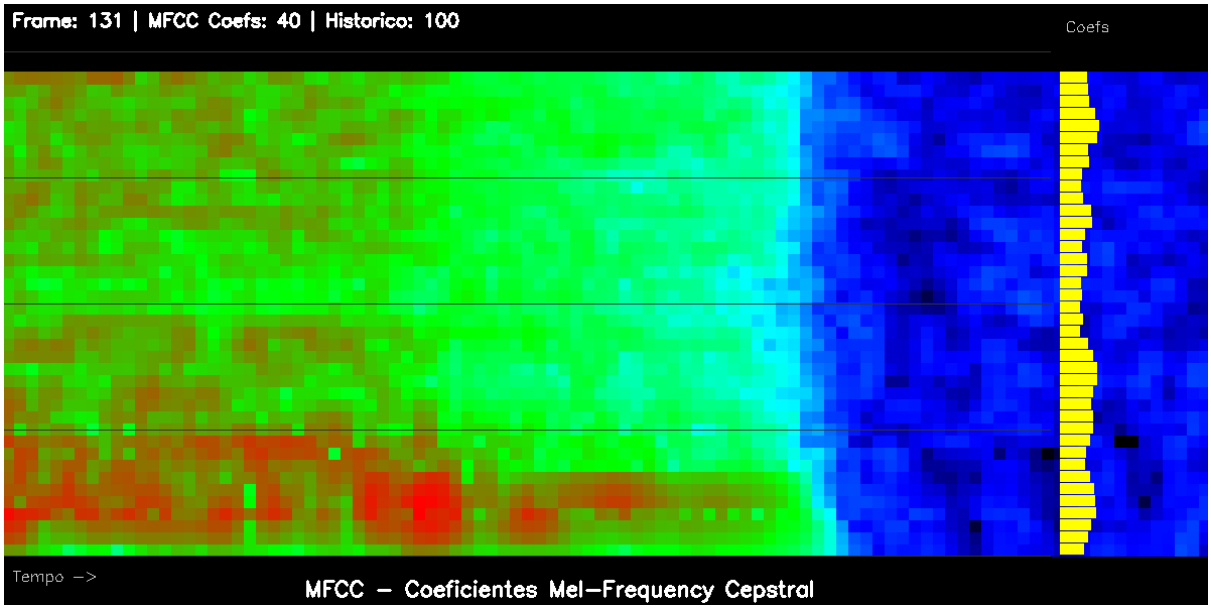

Figura 12 – *Frame* do *Mel Frequency Cepstral Coefficients* aplicado em um áudio de guitarra elétrica

Comparando o FFT dos dois trechos vocais, observa-se na Figura 15 uma predominância nas frequências médias, típica de um vocal isolado executando uma única melodia.

Em contraste, a Figura 16 revela uma maior distribuição espectral no caso do coral, já que diferentes vozes ocupam regiões distintas de frequência.

Essa diferença torna-se ainda mais evidente no MFCC. A Figura 17 apresenta um padrão concentrado, indicando que a energia do vocal solo se mantém em uma mesma faixa de coeficientes.

Já a Figura 18 evidencia regiões distintas nos coeficientes, cada uma correspondente a uma das vozes presentes no coral, permitindo identificar visualmente suas diferenças espectrais.

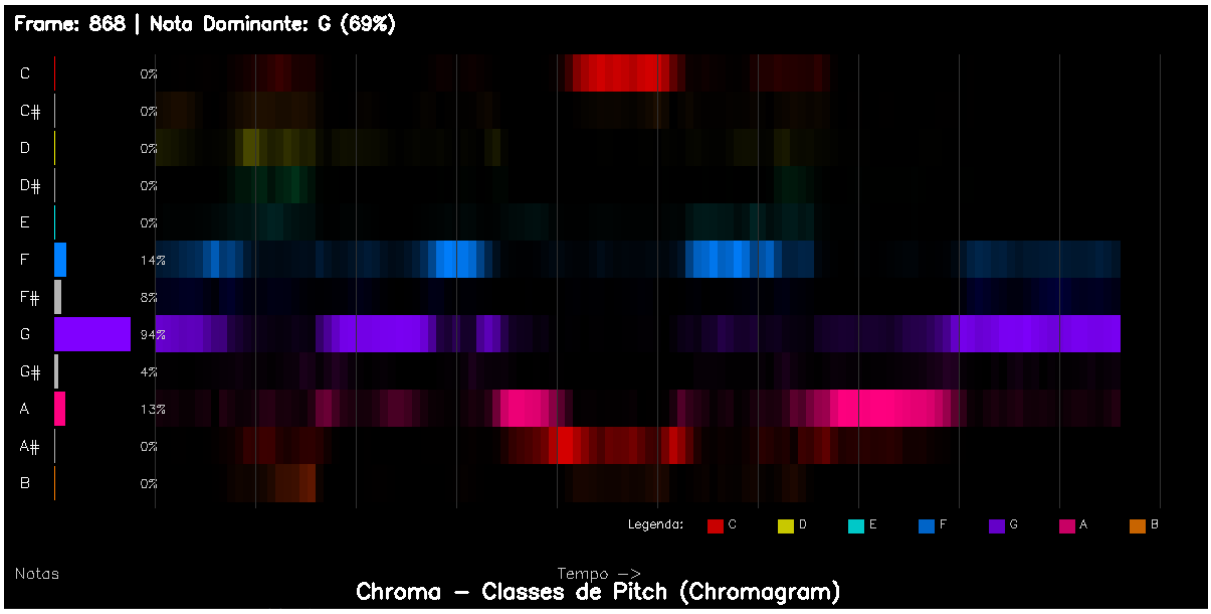

Figura 13 – *Frame* do Chroma aplicado em um áudio de vocal único

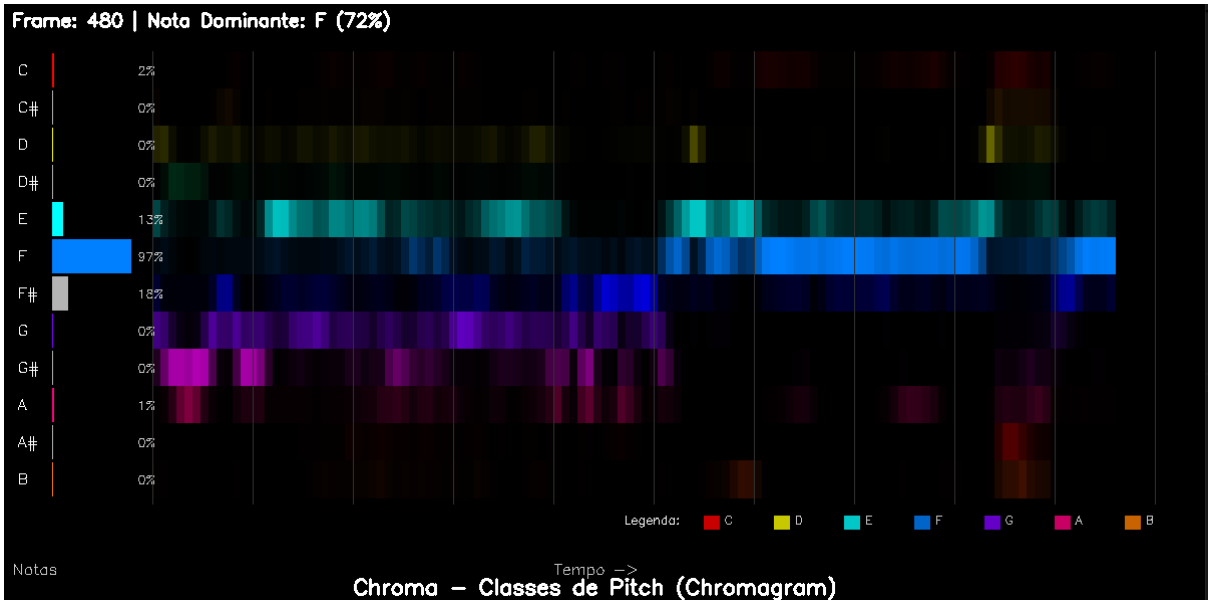

Figura 14 – *Frame* do Chroma aplicado em um áudio de coral

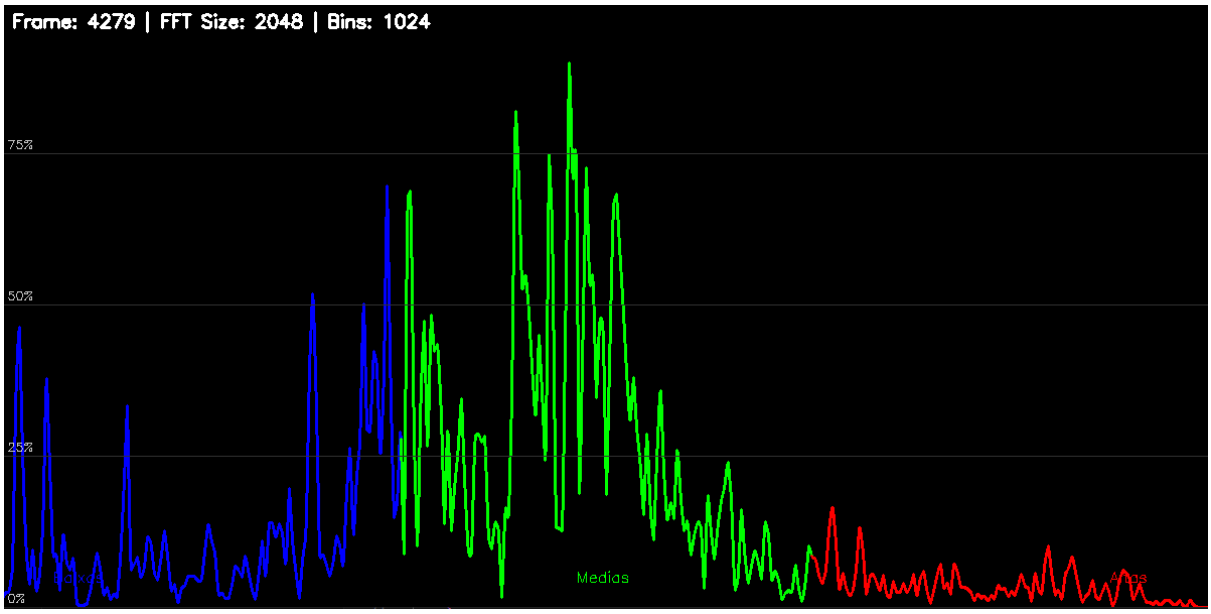

Figura 15 – *Frame* do *Fast Fourier Transform* aplicado em um áudio de vocal único

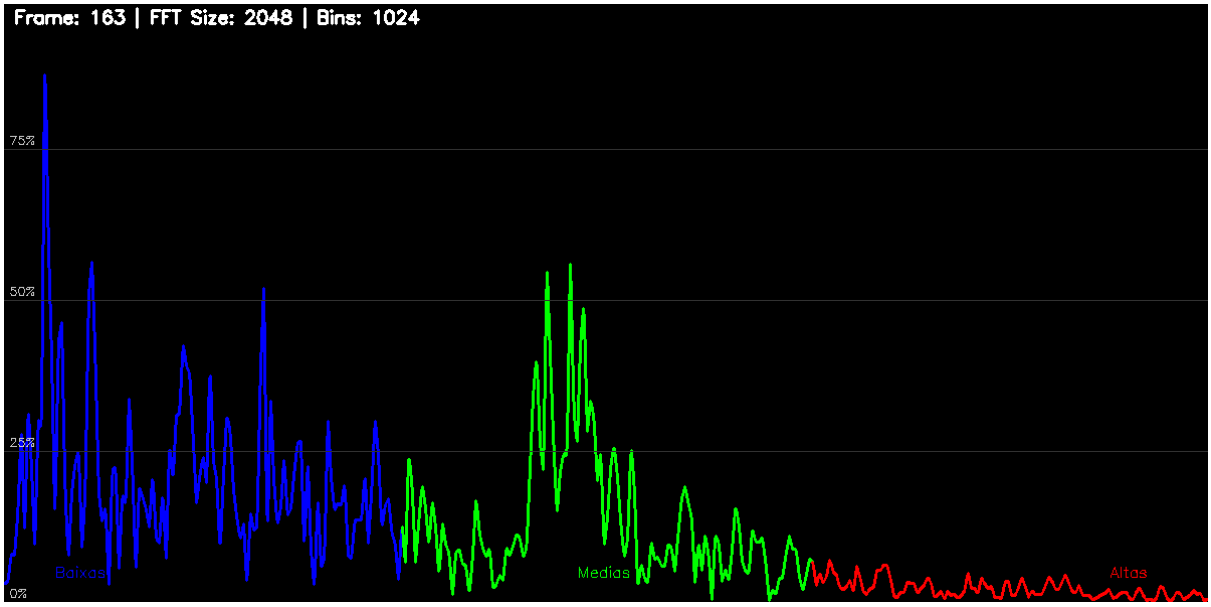

Figura 16 – *Frame* do *Fast Fourier Transform* aplicado em um áudio de coral

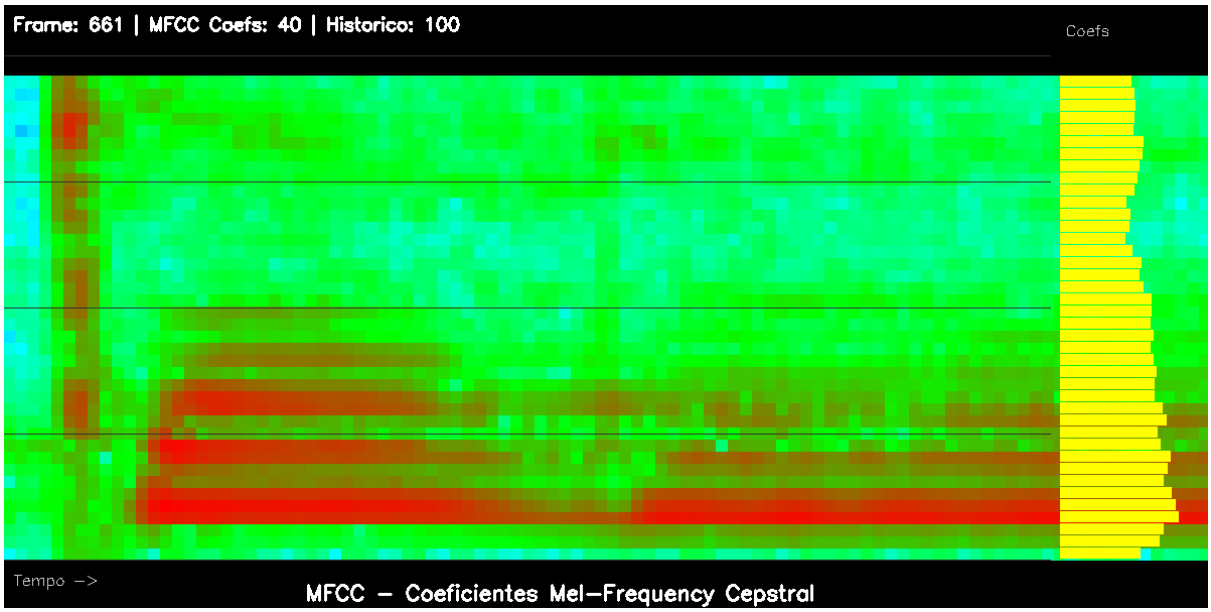

Figura 17 – *Frame* do *Mel Frequency Cepstral Coefficients* aplicado em um áudio de vocal único

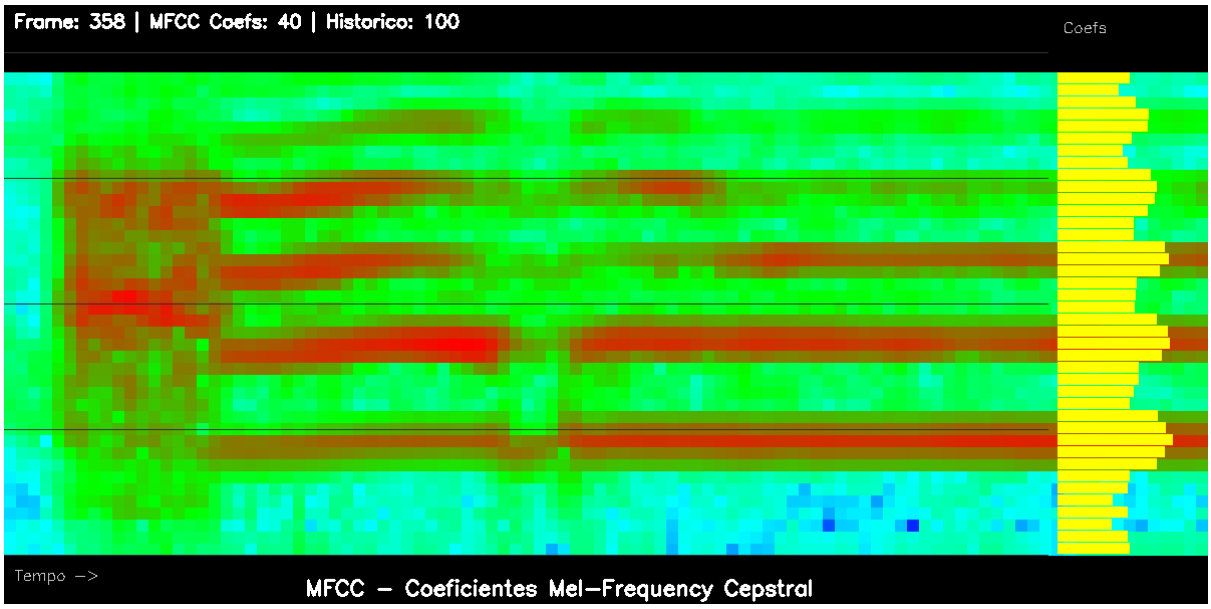

Figura 18 – *Frame* do *Mel Frequency Cepstral Coefficients* aplicado em um áudio de coral

5 RESULTADOS

Os experimentos aplicados nesse trabalho foram desenvolvidos com a finalidade de gerar resultados visuais gerados continuamente, a partir das informações obtidas pela extração de características específicas de um arquivo de áudio. A representação visual que foi obtida a partir das características extraídas do áudio, formam uma sequência de quadros que representam e expõem a particularidade de cada técnica utilizada para extração.

No resultado, é possível observar as características de cada uma das extrações utilizadas nos dados extraídos do arquivo de áudio, a FFT nas linhas geradas, a distorção causada pelo MFCC nessas linhas e a coloração vinda da tabela dos valores das frequências existentes no Chroma. No canto superior esquerdo é evidenciado a qual cor da linha da tabela cromática está sendo colorida a linha desenhada na tela pelo OpenCV, qual a frequência que está sendo identificada pela FFT no momento atual. Os efeitos causados pelas diferentes frequências da música, alteram e distorcem as linhas que aparecem na tela pelo FFT que cria o aspecto de ondas e MFCC que causa textura por cima do resultado da FFT.

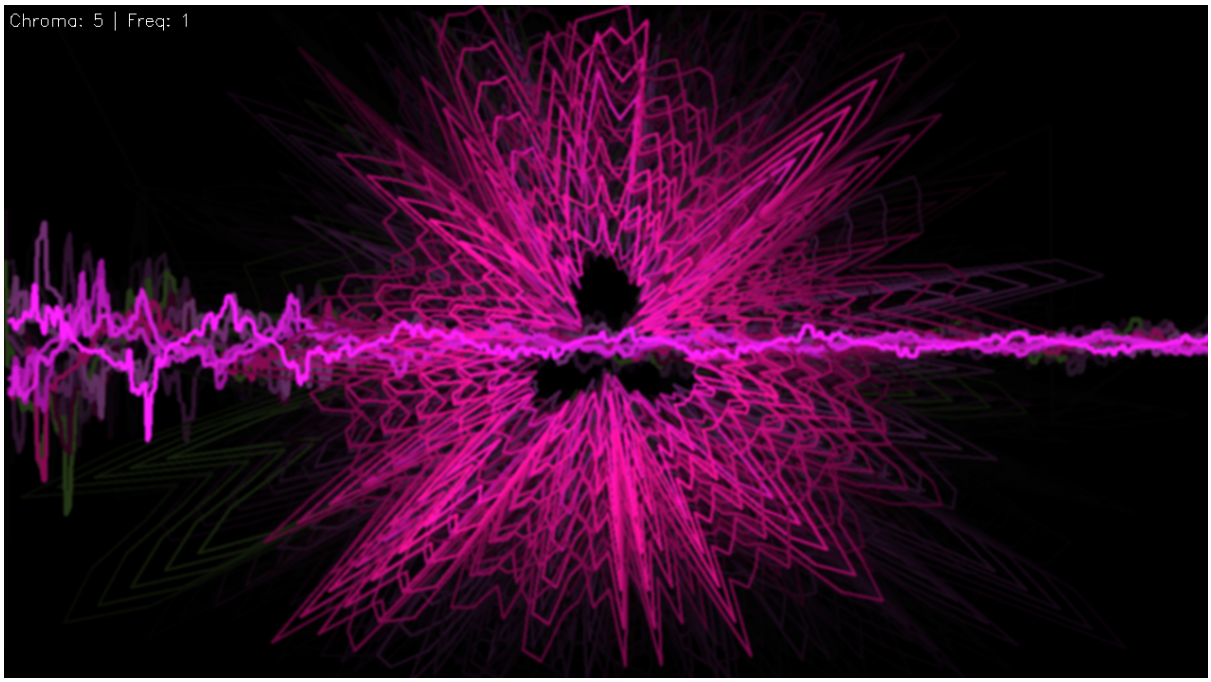

Figura 19 – *Frame* do Resultado Final da Convolução.

Aplicado um filtro de compressão logarítmica, o resultado é menos espelhado e as formas que são geradas apartir do MFCC ficam mais claras do que quando a distorção causada não é controlada. A diferença entre as duas figuras, Figura 19 e 20, é o controle do nível da distorção e textura causada pelos valores do FFT e MFCC.

No resultado final da experiência é possível de identificar as distorções causadas pelo MFCC nas linhas resultantes da FFT, que são introduzidas na tela de fundo preto e coloridas pela frequência que é identificada pelo Chroma.

A Figura 19 apresenta menos espaçamento das informações do FFT na tela, o MFCC esta mais estabilizado, e é perceptível a formação dos três círculos ao fundo.Na 20 as formações estão espalhadas por conta da distorção mais alta, justamente pelo valor sem normalização usado nela. Os círculos ao

fundo se espalham mais formando as linhas que se vê ao fundo, as linhas do FFT ficam menos lineares e compactadas ao centro.

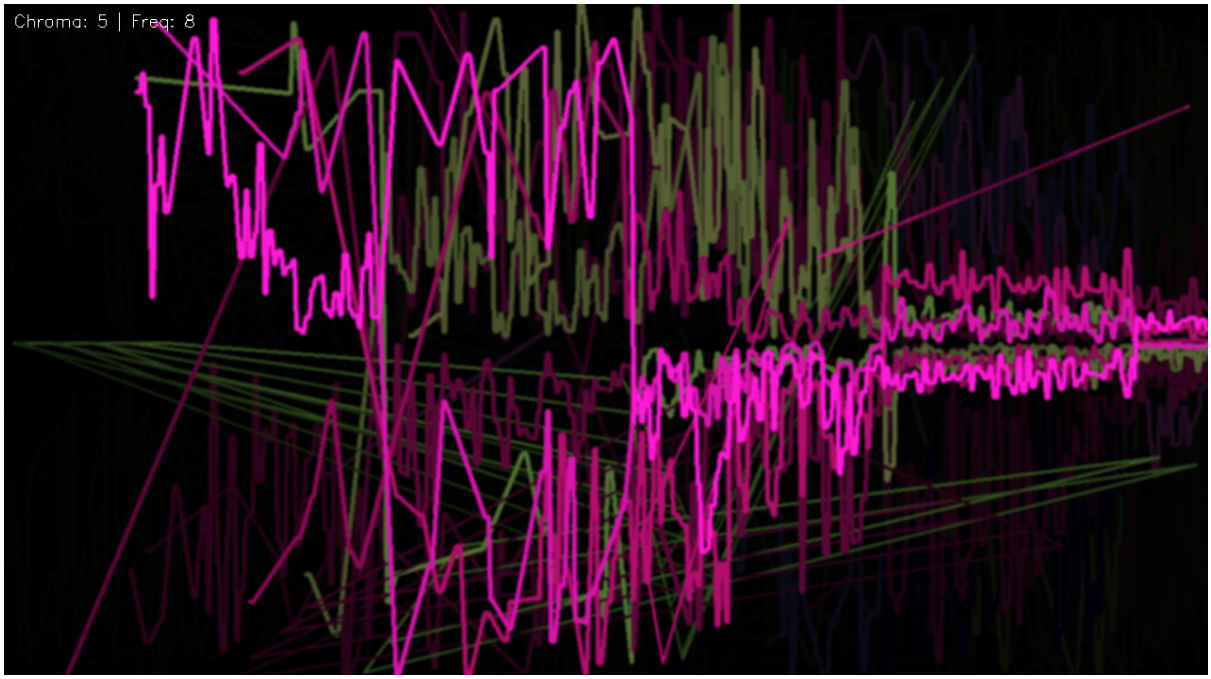

Figura 20 – Outro *frame* do Resultado Final da Convolução.

Utilizando uma convolução das técnicas de extração escolhidas, formam-se *frames* com visuais únicos e característicos. Essa representação visual pode ser aplicada em concertos, shows ou qualquer tipo de apresentação musical por seu processamento paralelo ao áudio, sendo possível as imagens serem processadas em tempo real, poderiam ser utilizadas por DJs, que utilizam imagens para conduzir suas performances, geralmente imagens e vídeos abstratos, a correlação de imagem e som seria explorada de maneira mais fiel, em vista que as imagens seriam geradas a partir do próprio áudio.

6 CONCLUSÃO

O trabalho apresenta uma conversão dos dados unidimensionais em representações bidimensionais, utilizando as técnicas de extração de características FFT, MFCC e Chroma. Tendo experimentos de como diferentes descritores do sinal se comportam quando sintetizados graficamente, focado na transposição direta de parâmetros mensuráveis para elementos visuais. A organização temporal do sinal e sua decomposição espectral permitem gerar estruturas visuais que representam cada característica extraída.

Os experimentos demonstraram que cada técnica de extração produz padrões distintos quando convertidos em imagem, reforçando que frequências graves, médias ou agudas, bem como variações tímbricas e harmônicas, influenciam diretamente o resultado visual. A FFT gerou linhas mais estáveis e contínuas, enquanto o MFCC introduziu texturas perceptíveis e dependentes da densidade espectral e o Chroma as notas predominantes e variações tonais ao longo do tempo. A combinação dessas técnicas em uma única síntese permitiu observar como seus efeitos se complementam, criando composições visuais coerentes com o comportamento dos dados.

Os resultados alcançados indicam que é viável produzir representações visuais derivadas diretamente das características extraídas de um sinal unidimensional, possibilitando aplicações em contextos artísticos, educativos ou performáticos. Ele oferece uma abordagem para se visualizar propriedades espectrais e temporais de forma simultânea.

Este mesmo trabalho se modificado poderia também ser utilizado com sinais de entrada já que as operações são feitas em janela temporal, gerando o que está sendo identificado sequencialmente. O que pode ser interessante para performances ao vivo também.

Uma utilidade a ser mencionada também, é a de $software$s reprodutores de arquivos de áudio, este tipo de representação visual sendo integrado a um sistema do tipo, causaria uma experiência multi sensorial do arquivo de áudio a ser reproduzido pelo $software$ no usuário.

REFERÊNCIAS

[1] CAMPEN, C. van. *The Hidden Sense: Synesthesia in Art and Science*. Cambridge, MA: MIT Press, 2007.

[2] WARD, J. *Synesthesia*. Cambridge, MA: MIT Press, 2013.

[3] GONÇALVES, J. dos S.; OLIVEIRA, A. M. de. Educação musical interativa: propostas interdisciplinares para as tecnologias educacionais. *Revista Digital do LAV*, v. 9, n. 1, p. 103–127, 2016.

[4] SPENCE, C.; STEFANO, N. D. Coloured hearing, colour music, colour organs, and the search for perceptually meaningful correspondences between colour and sound. *i-Perception*, v. 13, n. 3, p. 1–42, 2022.

[5] KUTZ, J. N. *Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data*. [S.l.]: Oxford University Press, 2013.

[6] GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 4. ed. [S.l.]: Pearson, 2018.

[7] MUNZNER, T. *Visualization Analysis and Design*. [S.l.]: CRC Press, 2014.

[8] KEIM, D. A. et al. *Mastering the Information Age: Solving Problems with Visual Analytics*. [S.l.]: Eurographics Association, 2011.

[9] WHITELAW, M. *Metacreation: Art and Artificial Life*. Cambridge, MA: MIT Press, 2004.

[10] MACHOVER, T. *Hyperinstruments: A Progress Report*. Cambridge, MA, 1992.

[11] CYTOWIC, R. E.; EAGLEMAN, D. M. *Wednesday Is Indigo Blue: Discovering the Brain of Synesthesia*. [S.l.]: MIT Press, 2009.

[12] HUBBARD, E. M.; RAMACHANDRAN, V. S. Neurocognitive mechanisms of synesthesia. *Neuron*, v. 48, n. 3, p. 509–520, 2005.

[13] BERTIN, J. *Semiology of Graphics: Diagrams, Networks, Maps*. [S.l.]: University of Wisconsin Press, 1983.

[14] GARTRELL, M. *Cores e Sons Em Port-Royal e Newton: Uma Análise Foucaultiana Da Representação e Sinestesia Clássicas*. Tese (Tese de Doutorado) — Instituição não especificada, 2024.

[15] PURWINS, H. et al. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, v. 13, n. 2, p. 206–219, 2019.

[16] SHAH, A. et al. Chroma feature extraction. In: . [S.l.: s.n.], 2019.

[17] MÜLLER, M. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Cham: Springer, 2015. 11–12 p. ISBN 978-3-319-21945-5.

[18] PAISANTE, D. B. S. *Programação de shaders em Computação Gráfica*. UFMA, 2022. Disponível em: <http://hdl.handle.net/123456789/5813>.

[19] SALLES, L. M. M. *Pressão sonora e processamento de sinais*. Dissertação (Dissertação de Mestrado) — Universidade Estadual Paulista (UNESP), Ilha Solteira, 2017.

[20] THE Short-Time FourierTransform | Spectral Audio Signal Processing. Dsprelated.com, 2019. Disponível em: <https://www.dsprelated.com/freebooks/sasp/Short_Time_Fourier_Transform.html>.

[21] VáRKONYI, D. T.; SEIXAS, J. L.; HORVáTH, T. Dynamic noise filtering for multi-class classification of beehive audio data. *Expert Systems with Applications*, v. 213, p. 118850, 2023. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417422018681>.

[22] BRENT, W. *Physical and perceptual aspects of percussive timbre*. [S.l.]: University of California, San Diego, 2010.

[23] DAVIS, S. B.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, p. 357–366, 1980.

[24] LOGAN, B. et al. Mel frequency cepstral coefficients for music modeling. In: PLYMOUTH, MA. *Ismir*. [S.l.], 2000. v. 270, n. 1, p. 11.

[25] SIGURDSSON, S.; PETERSEN, K. B.; LEHN-SCHIØLER, T. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In: *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*. Victoria, Canada: University of Victoria, 2006. p. 286–289.

[26] HEBBAR, D.; JAGTAP, V. *A Comparison of Audio Preprocessing Techniques and Deep Learning Algorithms for Raga Recognition*. 2022. Disponível em: <https://arxiv.org/abs/2212.05335>.