REPRESENTAÇÕES TEMPO-FREQUENCIAIS COM RESOLUÇÃO ADAPTATIVA COM APLICAÇÃO EM ÁUDIO

Gabriel Mendes Gouvêa

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Luiz Wagner Pereira Biscainho
Wallace Alves Martins

Rio de Janeiro
Março de 2016

REPRESENTAÇÕES TEMPO-FREQUENCIAIS COM RESOLUÇÃO ADAPTATIVA COM APLICAÇÃO EM ÁUDIO

Gabriel Mendes Gouvêa

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. Wallace Alves Martins, D.Sc.

Prof. Diego Barreto Haddad, D.Sc.

Prof. Flávio Rainho Ávila, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2016

Gouvêa, Gabriel Mendes

Representações Tempo-frequenciais com Resolução Adaptativa com Aplicação em Áudio/Gabriel Mendes Gouvêa. – Rio de Janeiro: UFRJ/COPPE, 2016.

XVII, 66 p.: il.; 29, 7cm.

Orientadores: Luiz Wagner Pereira Biscainho

Wallace Alves Martins

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2016.

Referências Bibliográficas: p. 63 – 66.

1. esparsidade. 2. resolução adaptativa. 3. tempo-frequência. 4. STFT. 5. CQT. 6. NSGT. I. Biscainho, Luiz Wagner Pereira *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Gostaria de agradecer aos meus orientadores Luiz Wagner e Wallace, que foram excelentes professores ao longo deste curso. Com paciência e dedicação, me permitiram a conclusão desta dissertação.

Agradeço também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela ajuda financeira.

Gostaria de agradecer aos amigos que me deram suporte ao longo desta caminhada, tanto durante as aulas quanto na pesquisa do mestrado.

Agradeço à minha família pelo apoio, carinho e por sempre me guiar, especialmente durante situações difíceis e complicadas da minha vida.

Gostaria de agradecer, por fim, a minha namorada, Bruna, por ter tanta paciência nos momentos em que mais necessitei. Seu carinho e apoio foram fundamentais no desenvolvimento deste trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

REPRESENTAÇÕES TEMPO-FREQUENCIAIS COM RESOLUÇÃO ADAPTATIVA COM APLICAÇÃO EM ÁUDIO

Gabriel Mendes Gouvêa

Março/2016

Orientadores: Luiz Wagner Pereira Biscainho
Wallace Alves Martins

Programa: Engenharia Elétrica

A presente dissertação possui como foco de investigação a comparação de representações tempo-frequência com resolução variável, que favorecem a análise de sinais de áudio, visando à reconstrução perfeita ou quase-perfeita.

A fim de realizarmos esse estudo comparado, partimos da avaliação de técnicas presentes na literatura que possuem a reconstrução perfeita (como a NSGT) e outra dedicada especificamente a sinais musicais (como a CQ-NSGT). Entretanto, verifica-se que a resolução tempo-frequência dessas representações, apesar de variável, não é adaptativa, sendo assim necessária uma análise prévia do sinal para uma resolução adequada ser obtida.

Propomos, a partir disso, aprimorar a reconstrução de uma técnica de representação tempo-frequência com resolução automática já existente na literatura. Através de simulações com sinais de áudio, observamos que a melhoria elaborada por esta dissertação atenua significativamente o problema de reconstrução indicado pelo trabalho original.

Além disso, desenvolvemos um método de análise de resultados que permite um refinamento dos parâmetros que configuram a técnica. Por fim, comparamos com uma representação tempo-frequência básica (STFT), sendo possível concluir que, apesar do alto custo de processamento, aquela se revela mais vantajosa quanto à natureza da representação.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TIME-FREQUENCY REPRESENTATIONS WITH ADAPTATIVE RESOLUTION IN AUDIO APPLICATIONS

Gabriel Mendes Gouvêa

March/2016

Advisors: Luiz Wagner Pereira Biscainho
Wallace Alves Martins

Department: Electrical Engineering

This dissertation has as its investigation focus the comparison of time-frequency representations with variable resolution, which favors the analysis of audio signals, aiming at a perfect or quasi-perfect reconstruction.

In order to accomplish this comparative study, we start from the evalution of techniques in the literature that have the perfect reconstruction proprety (as the NSGT) and other specifically taylored to musical signals (as the CQ-NSGT). However, it is noted that the time-frequency resolution of these representations, although variable, is non-adaptive, requiring a previous analysis of the signal for an adequate resolution to be obtained.

Regarding this issue, we propose to improve the reconstruction of a time-frequency representation technique with automatic resolution from the literature. Through simulations with audio signals, we observe that the improvement developed by this dissertation significantly attenuates the reconstruction problem inherent to the original work.

In addition, we developed a method to assess results that allows one to refine the parameters that configure the technique. Finally, we compare this proposed technique with a basic time-frequency representation (the STFT), and conclude that, despite its high cost of processing, the former is more advantageous as to the nature of the representation.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

$\mathbf{p}_i$	i-ésima projeção de um sinal x sobre uma função $\boldsymbol{\varphi}_i$, p. 13
A	Limite inferior de *frame*, p. 12
B	Limite superior de *frame*, p. 12
B_k	Largura de banda da k-ésima frequência, p. 22
$C_{i,\bar{l},p}$	Melhor conjunto de coeficientes para representar o i-ésimo intervalo de tempo e a p-ésima região frequencial de acordo com a esparsidade, p. 32
$C_{i,l,p}$	Conjunto de coeficientes do i-ésimo intervalo de tempo, da l-ésima análise tempo-frequência e da p-ésima região frequencial, p. 32
H_G	Índice de Gini, p. 33
H_H	Medida de Hoyer, p. 33
H_α	Entropia de Rényi, p. 31
N_k	Duração amostral da k-ésima janela, p. 10
N_w	Duração amostral da janela w, p. 6
N_x	Duração amostral do sinal x, p. 5
N_{Fl}	Comprimento de cada janela na l-ésima análise tempo-frequência, p. 37
Q	Fator de seletividade, p. 9
$W_{m,k}$	Função-base no domínio frequencial com m atrasos temporais e k deslocamentos frequenciais, p. 7
X	Sinal discreto no domínio frequencial, p. 5

Λ_l	l-ésima grade tempo-frequência, p. 29
Ω_c	Frequência digital de corte, p. 37
Ω_k	k-ésima frequência digital, p. 5
Θ	Banco de filtros, p. 30
α	Ordem da entropia de Rényi, p. 31
α_h	Parâmetro da janela de Hann, p. 11
$\mathbf{\Gamma}$	Família de funções que representam um *frame* dual, p. 14
$\mathbf{\Phi}$	Família de funções que representam um *frame*, p. 12
$\mathbf{\Psi}$	Família de funções que representam um *frame*, p. 18
$\boldsymbol{\gamma}_i$	i-ésima função-base na forma vetorial que compõe um *frame* dual, p. 14
$\boldsymbol{\gamma}_{m,k}$	Função-base na forma vetorial que compõe um *frame* dual de Gabor, p. 17
$\boldsymbol{\psi}_k$	k-ésima função-base na forma vetorial que compõe um *frame*, p. 18
$\boldsymbol{\psi}_{m,k}$	Função-base na forma vetorial que compõe um *frame* de Gabor, p. 18
$\boldsymbol{\varphi}_{m,k}$	Função-base na forma vetorial que compõe um *frame* de Gabor no contexto do espaçamento frequencial variável, p. 14
$\boldsymbol{\varphi_i}$	i-ésima função-base na forma vetorial que compõe um *frame*, p. 12
δ	Impulso de Dirac, p. 16
$\mathbb{H}$	Espaço de Hilbert, p. 13
$\mathbf{S}$	Operador *frame*, p. 13
$\mathbf{x}$	Sinal na forma vetorial, p. 13
μ_X	Média de uma variável aleatória X, p. 48
μ_Y	Média de uma variável aleatória Y, p. 48

$\rho_{X,Y}$ Coeficiente de correlação de Pearson entre duas variáveis aleatórias X e Y, p. 48

σ_X Desvio padrão de uma variável aleatória X, p. 49

σ_Y Desvio padrão de uma variável aleatória Y, p. 49

σ_p p-ésima função-peso, p. 31

θ_p p-ésima reposta ao impulso do filtro do banco de filtros Θ, p. 30

$\widehat{\boldsymbol{\psi}_{m,k}}$ DFT da função $\boldsymbol{\psi}_{m,k}$, p. 18

$\widehat{\boldsymbol{\rho}_{m,k}}$ DFT da função $\boldsymbol{\rho}_{m,k}$, p. 20

$\widehat{\mathbf{x}}$ DFT do sinal $\mathbf{x}$, p. 18

$\widehat{\theta_p}$ p-ésima reposta em frequência do filtro do banco de filtros Θ, p. 30

$\widetilde{C_{i,\bar{l},p}}$ Conjunto de coeficientes $C_{i,\bar{l},p}$ após modificações, p. 34

$\widetilde{x_{i,p}}$ Sinal no i-ésimo intervalo de tempo e na p-ésima região frequencial após modificações, p. 34

$\widetilde{x}$ Sinal reconstruído após modificações, p. 34

a Passo de deslocamento temporal, p. 6

a_i i-ésimo deslocamento temporal para segmentar o sinal, p. 25

a_l Passo de deslocamento temporal da l-ésima análise tempo-frequência, p. 29

a_m Passo de deslocamento temporal variante ao longo de m, p. 15

b Passo de deslocamento frequencial, p. 14

b_l Passo de deslocamento frequencial da l-ésima análise tempo-frequência, p. 29

b_m Passo de deslocamento frequencial variante ao longo de m, p. 15

c_i i-ésimo coeficiente de decomposição sobre um *frame*, p. 13

$c_{i,l}$ Coeficiente de decomposição para o i-ésimo intervalo na l-ésima análise tempo-frequência, p. 29

$c_{m,k}$	Coeficiente de decomposição sobre um *frame* de Gabor, p. 15
$d_{m,k}$	Coeficiente de decomposição sobre um *frame* de Gabor, p. 18
$e_{\max}$	Erro de pico, p. 36
e_{rms}	Erro eficaz normalizado, p. 36
f_c	Frequência analógica de corte, p. 37
f_k	k-ésima frequência analógica, p. 5
$f_{\min}$	Menor frequência analisada, p. 9
g_l	Função-janela para a l-ésima análise tempo-frequência, p. 29
h_l	Função-janela para a l-ésima síntese tempo-frequência, p. 29
j	Unidade imaginária, p. 5
k	k-ésima amostra de frequência, p. 5
m	m-ésimo atraso temporal, p. 6
w	Função-janela discreta no domínio temporal, p. 6
w_i	Função-janela do i-ésimo intervalo, p. 25
w_i'	Função-janela w_i modificada, p. 27
$w_{m,k}$	Função-base no domínio temporal com m atrasos temporais e k deslocamentos frequenciais, p. 6
x	Sinal discreto no domínio temporal, p. 5
x_i	Sinal janelado no i-ésimo intervalo, p. 25

Lista de Abreviaturas

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, p. iv
CQ-NSGT	*Constant Q-Transform with non-stationary Gabor Frames*, p. v
CQT	*Constant-Q Transform*, p. 9
DFT	*Discrete Fourier Transform*, p. 4
ERB	*Equivalent Rectangular Bandwidth*, p. 22
FFT	*Fast Fourier Transform*, p. 5
NSGT	*Non-stationary Gabor Transform*, p. v
ODG	*Objetive Difference Grade*, p. 44
PEAQ	*Perceptual Evaluation of Audio Quality*, p. 43
STFT	*Short-Time Fourier Transform*, p. v

Capítulo 1

Introdução

1.1 Motivação

Muitas vezes, dentro da área de processamento de sinais, precisamos fazer uma interpretação do sinal no domínio tempo-frequencial para extração de dados úteis a ele concernentes. Isso ocorre porque esse tipo de representação bidimensional pode fornecer mais detalhes do que as representações tradicionais nos domínios puramente temporais ou frequenciais.

Entretanto, as representações tempo-frequenciais mais comuns vistas na literatura demandam uma análise prévia do sinal para alcançar uma resolução apropriada para o sinal em questão. Além disso, nem sempre é possível chegar através delas a uma reconstrução perfeita, propriedade necessária em muitas aplicações. Portanto, faz-se necessária uma técnica na qual essas características, essenciais para tornar a representação completa, estejam presentes.

À vista disso, propomos, na presente dissertação, aprimorar uma técnica de representação tempo-frequência já existente na literatura, gerando uma nova estratégia com resolução adaptativa e reconstrução quase-perfeita.

1.2 Aplicações

O uso das representações tempo-frequenciais como etapa de análise é recorrente em diversas aplicações, como [2], [3], [4], [5], [6]. Técnicas de separação de fontes, por exemplo, podem valer-se de tais representações para extrair algumas informações dos sinais de mistura a fim de distinguir melhor os sinais provenientes de diferentes fontes [7], [8], [9], [10].

Uma outra aplicação que comumente utiliza essas representações são as técnicas de compressão de áudio. Nesse caso, as características extraídas a partir das representações tempo-frequenciais são necessárias para diminuir a quantidade de

parâmetros que descrevem o sinal de áudio, de acordo com determinados critérios psicoacústicos, viabilizando, assim, sua compressão [11], [12], [13].

1.3 Delimitação

Para sinais de áudio, as representações são particularmente atrativas em várias das aplicações já listadas quando os coeficientes gerados são esparsos [14], isto é, apresentam uma predominância de coeficientes nulos ou aproximadamente nulos. Como esse tipo de sinais possui um comportamento temporal ou espectral rico, as técnicas tempo-frequenciais de resolução fixa, normalmente vistas na literatura, não são as mais adequadas. Portanto, para esses casos o uso das representações tempo-frequenciais com resolução variável ou adaptativa se torna uma necessidade.

Dentre as técnicas utilizadas para representar sinais no domínio tempo-frequência com resolução variável, iremos descrever a *Constant-Q Transform*, a *Non-stationary Gabor Transform* e a *Constant Q-Transform with non-stationary Gabor Frames*. Contudo, nenhuma destas técnicas permite uma resolução variável no tempo e na frequência simultaneamente.

Para suprir esta necessidade e tornar a representação mais esparsa, um método apresentado em [1] foi estudado e implementado neste trabalho. A principal característica desta técnica é a capacidade de selecionar a melhor resolução temporal e frequencial para cada trecho do sinal com base em uma medida de esparsidade.

Como a reconstrução perfeita é essencial em diversas aplicações já mencionadas, avaliamos esta propriedade através de uma série de testes. Os testes consistem em projetar o sinal no domínio tempo-frequência gerando coeficientes; em seguida, reconstruímos o sinal a partir destes coeficientes e o comparamos com o sinal original. Contudo, mesmo variando diversos parâmetros da técnica, o resultado obtido estava muito distante de ter propriedade de reconstrução perfeita.

Com o objetivo de solucionar este problema, sugerimos uma alteração, com base em [15], na etapa de janelamento original e, também, a implementação de duas outras medidas de esparsidade [14]. Realizamos um novo conjunto de testes comparando o efeito destas modificações sobre a reconstrução dos sinais. Através de análises estatísticas dos resultados (cálculo de coeficiente de Pearson, por exemplo), concluímos que atingimos reconstrução quase perfeita devido, principalmente, à alteração na etapa de janelamento.

Por fim, realizamos uma comparação com uma técnica comum na literatura (*Short-Time Fourier Transform*), examinando as duas representações e os indicadores resultantes que avaliam a reconstrução do sinal. Concluímos que a técnica proposta permite uma maior interpretabilidade dos eventos acústicos devido à natureza da representação ao custo de uma baixa perda na reconstrução do sinal.

1.4 Organização

No Capítulo 2, iremos descrever os procedimentos necessários para a implementação de diversos métodos encontrados na literatura que projetam um sinal discreto no domínio tempo-frequencial. Como o objetivo é a representação de sinais de áudio neste domínio, discutiremos as vantagens de cada técnica sobre essa aplicação.

No Capítulo 3, explicaremos as etapas fundamentais para realizar a representação tempo-frequencial com resolução adaptativa desenvolvida em [1]. Apresentaremos também a modificação proposta na etapa de janelamento para aumentar a eficácia do método.

O foco do Capítulo 4 é descrever os testes realizados com o método tempo-frequencial proposto na presente dissertação e analisar os indicadores que avaliam esta técnica.

E por fim, no Capítulo 5, iremos expor as conclusões obtidas deste trabalho, além de sugerir alterações na técnica que potencialmente produzirão melhorias adicionais em sua eficácia.

Capítulo 2

Fundamentação Teórica

Ao analisarmos um sinal unidimensional que varia ao longo do tempo, normalmente consideramos suas características apenas no domínio do tempo ou no domínio da frequência. Contudo, tais caracterizações não permitem descrever informações frequenciais que variam ao longo do tempo. Nesse caso, as representações tempo-frequenciais são ideais para extrair informações de tempo e frequência conjuntamente.

Neste capítulo, dedicado à análise da bibliografia utilizada como base desta pesquisa, descreveremos uma representação tempo-frequência tradicional, a *Short-Time Fourier Transform*, assim como outras representações mais complexas: a *Constant-Q Transform*, a *Non-stationary Gabor Transform* e a *Constant Q-Transform with non-stationary Gabor Frames*.

2.1 *Short-Time Fourier Transform*

Os sinais digitais de áudio[1] são normalmente representados pela variação da amplitude — que desempenha o papel de pressão sonora — ao longo do tempo. Para ilustrar esta análise temporal será utilizado um sinal[2] digital gerado artificialmente apresentado na Figura 2.1.

Nela temos um sinal contendo 3 notas musicais tocadas e intercaladas entre si por 1,5 segundo de silêncio. O sinal tem aproximadamente 12 segundos de duração e uma amplitude de até 0,25 no eixo das ordenadas.

Contudo, não é possível concluir nesta representação qual nota está presente no sinal. Esta informação pode ser obtida através de uma análise frequencial. A *Discrete Fourier Transform* (DFT) é um algoritmo que converte uma representação

[1]Nesta dissertação, serão analisados sinais exclusivamente digitais (tempo amostrado e amplitude quantizada); todas as simulações e experimentos serão feitos com 44100 Hz de taxa de amostragem e 16 bits por amostra.

[2]O sinal foi gerado a partir do registro 011PFNOF adquirido pelo banco de gravações de áudio RWC *Music Database: Musical Instrument Sound Database* [16].

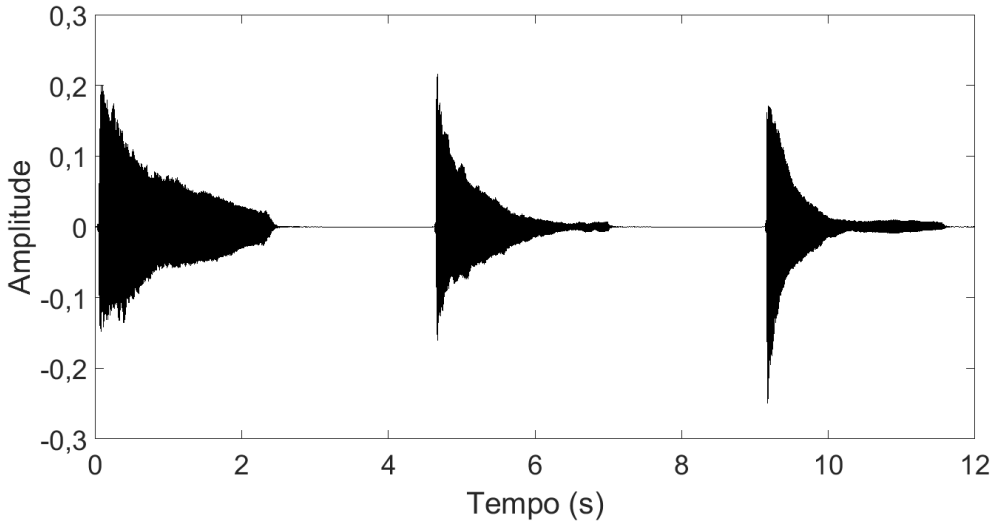

Figura 2.1: Sinal no domínio do tempo.

no tempo para um representação na frequência. A equação

$$X[k] = \frac{1}{N_x} \sum_{n=0}^{N_x-1} x[n] e^{-jn\frac{2k\pi}{N_x}}, 0 \leq k \leq N_x \tag{2.1}$$

explicita o cálculo que realiza essa transformação a partir de um sinal discreto no domínio temporal. O termo $X[k]$ da Equação (2.1) representa o sinal $x[n]$ no domínio da frequência, com o índice k identificando a componente frequencial; N_x é o comprimento do sinal analisado, n é o índice relativo à amostra no tempo e j é a unidade imaginária. O fator $2k\pi/N_x$ que compõe a exponencial complexa é denominado frequência digital (Ω_k). A relação de equivalência entre essa frequência e a frequência analógica f_k é expressa por

$$\Omega_k = \frac{2k\pi}{N_x} \Rightarrow f_k = \frac{kf_s}{N_x}, \tag{2.2}$$

onde f_s é a frequência de amostragem do sinal.

Uma técnica denominada de *zero-padding* pode ser utilizada na DFT para reduzir o custo computacional de um algoritmo rápido para cálculo da DFT, a *Fast Fourier Transform* (FFT). Essa técnica consiste em adicionar zeros ao início ou ao fim do sinal antes de realizar a transformada propriamente dita. No domínio da frequência, este sinal irá apresentar mais amostras frequenciais do que a DFT de um sinal sem *zero padding*.

A Figura 2.2 contém o módulo da representação frequencial obtido pelo algoritmo de DFT do sinal utilizado no exemplo anterior. Apesar de o sinal ser amostrado a uma taxa de 44,1 kHz, limitamos o eixo das abscissas a 3 kHz porque as frequências fundamentais e os principais harmônicos deste sinal estão presentes nessa faixa.

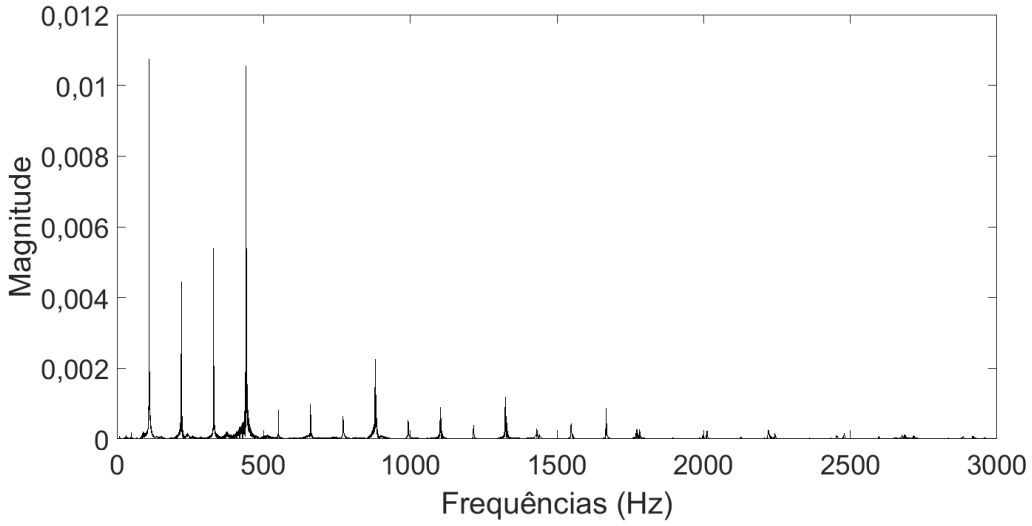

Figura 2.2: Sinal no domínio da frequência.

Neste gráfico, podemos observar vários máximos separados entre si por 110 Hz. E os quatro maiores picos identificam alguns tons relevantes presentes no sinal: 110 Hz (Lá 1 - A2), 220 Hz (Lá 2 - A3), 330 Hz (um harmônico de A2) e 440 Hz (Lá 3 - A4).

Ao contrário da representação na Figura 2.1, o gráfico da Figura 2.2 pode permitir determinar quais foram as notas tocadas. Entretanto, ainda é impossível relacionar qual nota foi tocada em qual instante de tempo. Para alcançar este objetivo, faz-se necessária uma análise que utilize simultaneamente as duas representações: a análise tempo-frequência.

Ao nos dedicarmos ao estudo das representações discretas no domínio tempo-frequência, verifica-se que a *Short-Time Fourier Transform* (STFT) se destaca como a mais utilizada. A decomposição de um sinal através da STFT consiste em projetá-lo sobre um conjunto de funções-base definidas por

$$w_{m,k}[n] = w[n - ma]e^{-j\Omega_k n}, \tag{2.3}$$

cada uma é construída a partir de uma função-janela $w[n]$ de N_w amostras com um deslocamento temporal múltiplo de um passo a em amostras e um deslocamento frequencial de Ω_k em rad/s. Os subíndices m e k estão relacionados, respectivamente, ao número de atrasos temporais e à frequência digital analisada. A frequência digital Ω_k é dada por $2k\pi/N_w$.

Dessa forma, a STFT é uma projeção do sinal sobre cada uma dessa funções-base, a qual gera um coeficiente que representa o sinal nesse ponto destacado. Define-se,

portanto, essa representação, apresentada em [17], como

$$X_{\mathrm{STFT}}[m,k] = \frac{1}{N_w} \sum_{n=0}^{N_x-1} x[n] w_{m,k}[n]. \tag{2.4}$$

Analisando a equação (2.4) juntamente com a equação (2.3), podemos inferir que o sinal primeiramente é subdividido através da função-janela atrasada no tempo e em seguida calcula-se a DFT de tamanho N_w de cada trecho do sinal. Desta forma, é possível considerar que a STFT consiste em uma evolução espectral do sinal ao longo do tempo.

Uma interpretação dual da STFT pode ser obtida aplicando-se o teorema de Parseval [18] na equação (2.4). Assim, temos

$$X_{\mathrm{STFT}}[m,k] = \sum_{k'=0}^{K'-1} X[k'] W_{m,k}[k'], \tag{2.5}$$

onde $X[k']$ é a DFT do sinal $x[n]$, $W_{m,k}[k']$ é a DFT do sinal $w_{m,k}[n]$, k' é o índice da DFT variando entre 0 e $K'-1$ e K' é o tamanho da DFT. Este procedimento descreve a projeção de um sinal representado originalmente no domínio da frequência sobre um conjunto de funções-base no domínio tempo-frequência.

As funções-janela mais usuais são: janela retangular, janela de Hann e janela de Hamming [19]. Apesar de qualquer tipo de janela ter a capacidade de representar o sinal no domínio tempo-frequência, as funções-base construídas são diferentes e, portanto, as projeções geram coeficientes diferentes. Assim sendo, cada tipo de janela define uma representação única do sinal no domínio tempo-frequência.

O sinal da Figura 2.1 é representado no domínio tempo-frequência na Figura 2.3 por meio da sua STFT. O gráfico contém os coeficientes de decomposição do sinal sobre um conjunto de funções-base gerado a partir de uma janela de Hann de comprimento $N_w = 4096$ amostras e de um passo temporal de 2048 amostras. Cada coeficiente da STFT apresenta um valor de intensidade na escala logarítmica de acordo com os níveis de cinza usados na figura. O eixo das abscissas contém os atrasos temporais *ma* em segundos de cada ponto do plano tempo-frequência. O eixo das ordenadas expõe a componente frequencial analisada em Hertz.

As regiões em branco representam os momentos de silêncio do sinal. Ao longo de cada região mais escura, observamos raias horizontais espaçadas igualmente, pois cada região representa uma nota tocada e seus respectivos harmônicos. Desse modo, essa representação permite distinguir cada nota no tempo e, simultaneamente, identificá-la pela frequência.

Dependendo da escolha de parâmetros, a STFT discreta pode não permitir a reconstrução perfeita [20]. Isto pode ocorrer porque algumas janelas possuem uma

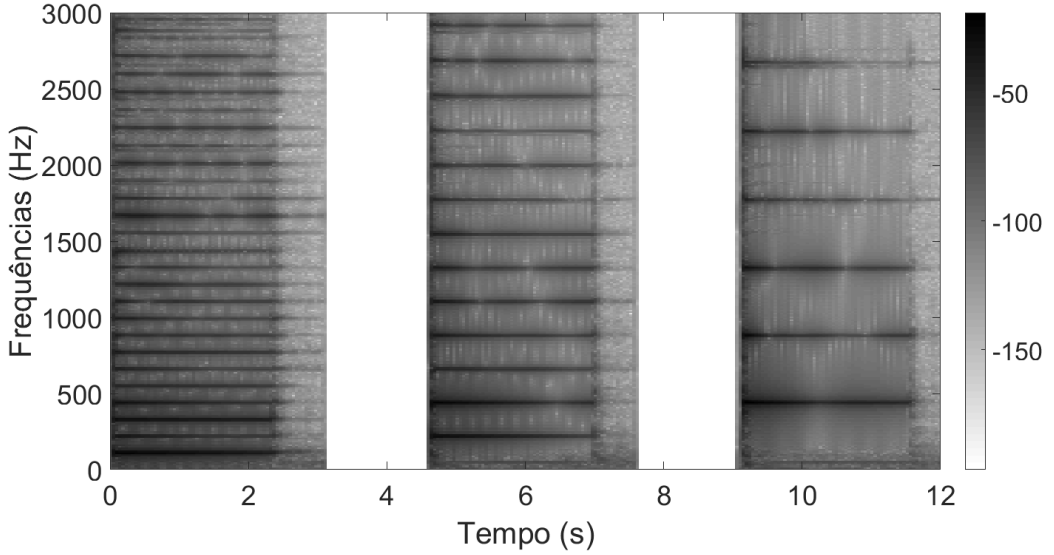

Figura 2.3: Módulo do sinal no domínio do tempo-frequência da STFT.

banda limitada e os deslocamentos frequenciais não possibilitariam a análise de todas as componentes frequenciais do sinal. Portanto, a representação tempo-frequencial não iria conter todas as informações a respeito do sinal em seus coeficientes. Um problema similar ocorre ao analisar os deslocamentos temporais. Nesse caso, o passo temporal pode ser maior do que a largura da janela e haveria informações temporais não representadas.

Nas situações em que essa reconstrução se faz possível, o procedimento para reconstruir o sinal original (adaptado de [20]) é descrito por

$$x[n] = N_w \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} X_{\mathrm{STFT}}[m,k] w^*_{m,k}[n], \tag{2.6}$$

com * sendo a operação de conjugação.

Através da equação (2.4) podemos observar que o cálculo de cada coeficiente da STFT utiliza um grupo de pontos selecionados por uma determinada função-base $w_{m,k}$. O número de pontos presentes nesse grupo é determinado pela largura N_w da função-janela. Se a largura for maior, cada coeficiente irá concentrar um horizonte temporal maior de pontos do sinal e, dessa forma, a STFT terá menos detalhes temporais, ou seja, uma baixa resolução temporal. Analogamente, a adoção de um valor menor de N_w implicaria uma alta resolução temporal.

Uma análise dual a este raciocínio ocorre observando-se que na equação (2.5) cada coeficiente é resultado de um cálculo sobre um grupo de pontos do sinal no domínio da frequência. A quantidade de pontos é determinada pela largura das funções-base $W_{m,k}$. A largura de banda dessas funções-base depende do formato da função-janela $w[n]$ original e é inversamente proporcional à largura N_w da janela. Portanto, se

N_w for menor, a largura de banda será maior, os coeficientes irão concentrar maior pontos do sinal e, dessa forma, a STFT terá menos detalhes frequenciais, isto é, uma baixa resolução frequencial. Analogamente, com N_w maior, a STFT terá uma alta resolução frequencial. Essa relação de proporcionalidade inversa entre a resolução frequencial e temporal é conhecida como o princípio da incerteza [19].

Observamos que as duas resoluções dependem da largura da janela, porém segundo relações opostas: a resolução temporal é inversamente proporcional a N_w e a resolução frequencial é diretamente proporcional a N_w. Conclui-se, portanto, que não é possível obter uma representação de STFT que apresente alta resolução temporal e frequencial simultaneamente.

Em sinais predominantemente musicais[3], os tons produzidos pelos instrumentos são baseados na escala cromática [21]. Esta, por sua vez, define a altura das notas por frequências espaçadas geometricamente. A STFT, porém, representa as componentes frequenciais com espaçamento uniforme. Logo, esse tipo de sinal é retratado com menor resolução frequencial para tons de baixa frequência, enquanto que para tons de alta frequência distingue-se mais facilmente a informação presente no sinal, como ilustrado no exemplo da Figura 2.3.

As representações tempo-frequência apresentadas nas seções seguintes deste trabalho possuem intervalos variáveis entre frequências, característica necessária para atingir uma melhor resolução frequencial na análise desse tipo de sinais.

2.2 *Constant-Q Transform*

A *Constant-Q Transform* (CQT), inicialmente introduzida por Brown em [22], é uma representação tempo-frequencial que, diferentemente da STFT, apresenta um espectro de frequências com fator de seletividade (também conhecido como fator Q) constante. Devido a isto, as componentes frequenciais estão espaçadas geometricamente, ou seja, uma componente f_k é dada por

$$f_k = f_{k-1}\left(\frac{1}{Q}+1\right) = f_{\min}\left(\frac{1}{Q}+1\right)^k, \tag{2.7}$$

onde $f_{\min}$ é a menor componente frequencial analisada pela CQT e Q é o fator de seletividade. Desenvolvendo a equação (2.7) temos

$$f_k - f_{k-1} = \frac{f_{k-1}}{Q}. \tag{2.8}$$

Percebe-se que o espaçamento entre componentes frequenciais ($f_k - f_{k-1}$) desta transformada é dado pela razão entre a frequência anterior e o fator de seletividade

[3]Estamos nos referindo à música ocidental tradicional.

(f_{k-1}/Q). A relação entre o comprimento da janela e esse espaçamento é expressa por

$$\frac{f_s}{N_w} = \frac{f_{k-1}}{Q} \Rightarrow N_k = \frac{f_s}{f_k}Q. \tag{2.9}$$

Assim, conclui-se que o comprimento N_k da janela para analisar cada componente frequencial é dependente de k. Se analisarmos o comportamento do comprimento da janela vemos que ele decresce conforme as frequências aumentam. Isto é justificável pelo fato de uma senoide de baixa frequência necessitar de mais amostras para completar um ciclo que uma senoide de alta frequência.

Da equação (2.2), concluímos que $\Omega_k = 2\pi\frac{f_k}{f_s}$. Portanto, ao utilizar a equação (2.9), temos

$$\Omega_k = \frac{2\pi Q}{N_k}, \tag{2.10}$$

ou seja, a frequência digital (Ω_k), na qual o sinal é projetado, é diretamente proporcional ao fator de seletividade.

Na Tabela 2.1, adaptada do artigo [22], podemos observar os parâmetros analisados entre as duas representações tempo-frequência já mencionadas neste trabalho.

	STFT	CQT
Componentes frequenciais	kf_s/N_w (linear)	$f_{\min}(1/Q+1)^k$ (geométrico)
Espaçamento frequencial	f_s/N_w (constante)	f_{k-1}/Q (variável)
Largura da janela	N_w (constante)	Qf_s/f_k (variável)
Frequência digital	$2k\pi/N_w$	$2Q\pi/N_k$

Tabela 2.1: Tabela de comparação de parâmetros entre STFT e CQT

A partir das equações (2.3) e (2.4), que descrevem a STFT, podemos definir a equação que gera os coeficientes da CQT como

$$X_{\mathrm{CQT}}[m,k] = \frac{1}{N_k}\sum_{n=0}^{N_k-1} x[n]w_{m,k}[n], \tag{2.11}$$

$$w_{m,k}[n] = w[n-ma,k]e^{-j\frac{2Q\pi}{N_k}n}. \tag{2.12}$$

Neste caso, a função de janela $w[n,k]$ está relacionada a k, uma vez que o comprimento da janela N_k depende de k. A janela empregada por [22] é a janela de *Hamming*, dada por

$$w[n,k] = \alpha_h + (1+\alpha_h)\cos\left(\frac{2\pi n}{N_k}\right), \tag{2.13}$$

onde $\alpha_h = 25/46$ e n varia entre 0 a $N_k - 1$. Contudo, não há limitações sobre o tipo de janela a ser utilizada para o cálculo de coeficientes da CQT. Em [23], um algoritmo de análise rápido é descrito também por Brown.

Como dito na Seção 2.1, os sinais de natureza musical apresentam espaçamento entre componentes frequenciais geométrico, assim como a CQT é construída. Esse comportamento qualifica a representação como opção ideal para analisar esse tipo de sinais. Através da equação (2.7) e com $Q = 17$, teremos $f_k \approx f_{k-1}2^{1/12}$, isto é, as componentes frequenciais aproximadamente separadas de um semitom, que é a separação entre as notas na escala cromática. Para uma resolução ainda melhor, utiliza-se $Q = 34$ e portanto tem-se $f_k \approx f_{k-1}2^{1/24}$, ou seja, um espaçamento de um quarto de tom.

A Figura 2.4 contém o exemplo de 3 tons, visto na seção anterior, representado através da CQT. Para gerar essa Figura foi utilizada uma janela de Hann, um Q de 34, e uma frequência mínima de 55 Hz. A barra lateral indica, em níveis de cinza, a magnitude dos coeficientes em dB. Comparando com a Figura 2.3, notamos diferença especialmente na resolução frequencial. A principal vantagem da CQT é destacada ao permitir a identificação nítida dos harmônicos de cada tom.

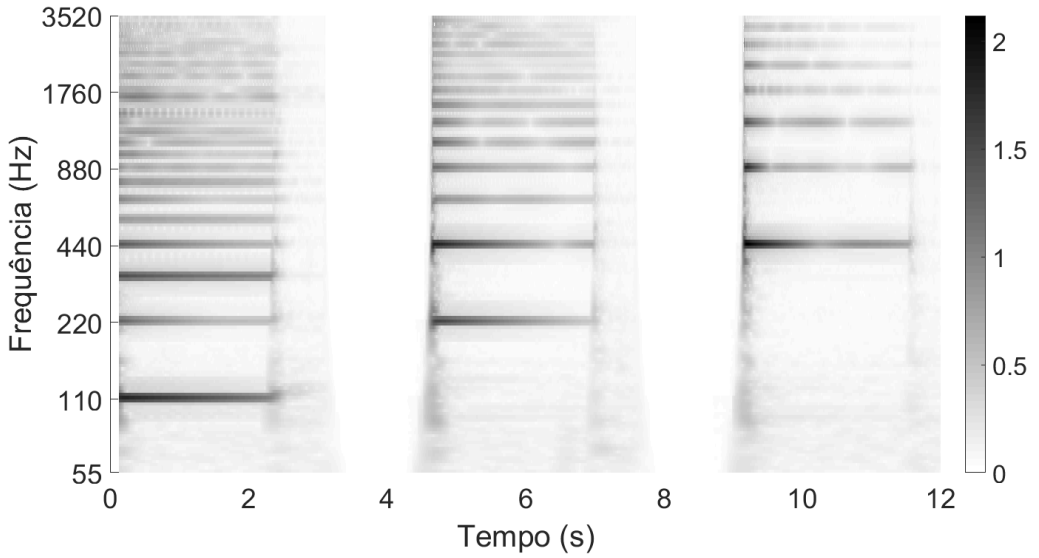

Figura 2.4: Módulo do sinal no domínio do tempo-frequência da CQT.

Apesar de a CQT apresentar um desempenho melhor para representar sinais de música, a técnica, ao contrário da STFT, não possui nenhuma combinação de parâmetros que permita uma reconstrução perfeita do sinal original a partir dos coeficientes. Isto ocorre devido ao fato de que nas construções dos coeficientes de alta frequência as larguras das janelas são menores do que os atrasos temporais. Por esse motivo, nem todas as amostras são analisadas e, portanto, verifica-se que não é possível ressintetizar o sinal a partir dos coeficientes gerados pela CQT proposta

em [22].

Em [24] é introduzida uma nova forma de construir representações tempo-frequenciais com espaçamento não-uniformes (no tempo ou na frequência, mas não simultâneo em ambos os eixos) e ainda permitindo ressíntese do sinal original a partir dos coeficientes gerados: a *Non-stationary Gabor Transform*(NSGT). Assim, a *Constant Q Non-stationary Gabor Transform*(CQ-NSGT) surge em [25] apresentando as mesmas características da CQT como uma adaptação da NSGT. Na Seção 2.3 e na Seção 2.4 detalharemos a NSGT e a CQ-NSGT, respectivamente.

2.3 *Non-stationary Gabor Transform*

Na análise de sinais de áudio vemos constantemente misturas de sinais com características tonais e não-tonais, como, por exemplo, em notas tocadas a partir de um piano. No início das notas o sinal apresenta um ataque que é bem espalhado ao longo do espectro, porém bem localizado no tempo. Ao longo da sustentação das notas, observamos componentes senoidais que se espalham no tempo, mas são bem definidas na frequência.

A solução mais simples para descrição de sinais com tal característica seria usar uma grande resolução temporal e frequencial para obter uma representação mais clara. Porém, haveria uma quantidade desnecessária de informação que elevaria o esforço computacional tanto no processamento desta representação quanto no seu pós-processamento.

Assim sendo, faz-se necessária uma representação tempo-frequencial cujas componentes temporais e frequenciais possam ser distribuídas irregularmente. Dessa forma, podemos utilizar, por exemplo, uma maior resolução temporal para acompanhar elementos bem localizados no tempo e, quando ocorrerem componentes espalhadas no tempo, diminuir a resolução temporal. Com essa configuração do plano tempo-frequência, é possível representar o sinal mais adequadamente sem prejudicar a capacidade prática de processamento.

Para alcançar esse tipo de representação, utilizaremos o *frame*. Na STFT e na CQT, projetamos o sinal em um conjunto de funções-base linearmente independentes que possuem a mesma capacidade geradora de funções-base e respeitam certas condições. O *frame* é um conjunto de funções linearmente dependentes. À vista disso, uma projeção sobre o *frame* geraria coeficientes com informações redundantes. A teoria sobre *frames* que será apresentada de forma resumida a seguir é baseada em [26].

Sejam uma família de funções $\mathbf{\Phi} = \{\boldsymbol{\varphi}_i \mid \sum |\varphi_i[n]|^2 < \infty, i \in \{0, 1, \ldots, I-1\}\}$, onde I é o número de funções na família $\mathbf{\Phi}$, e duas constantes reais A e B tais que

$0 < A \leq B < \infty$. Para que $\mathbf{\Phi}$ seja um *frame*, a condição

$$A\|\mathbf{x}\|^2 \leq \sum_{i=0}^{I-1} |c_i|^2 \leq B\|\mathbf{x}\|^2 \tag{2.14}$$

deve ser satisfeita para todo $\mathbf{x} \in \mathbb{H}$, onde $\mathbf{x}$ é um sinal na forma vetorial de tamanho $N_x \times 1$, $\boldsymbol{\varphi}_i$ é uma função na forma vetorial de tamanho $N_x \times 1$, $|\cdot|$ é a operação de módulo, $\|\cdot\|$ é a operação de norma 2 e $\mathbb{H}$ é o espaço de Hilbert. O termo c_i é denominado coeficiente de decomposição desse *frame* e é dado pelo produto interno descrito em

$$c_i = \sum_{n=0}^{N_x-1} \varphi_i^*[n]x[n] = \langle \mathbf{x}, \boldsymbol{\varphi}_i \rangle. \tag{2.15}$$

As projeções do sinal geradas sobre esse *frame* são dadas por

$$\mathbf{p}_i = c_i \boldsymbol{\varphi}_i. \tag{2.16}$$

Se a família de funções $\mathbf{\Phi}$ formasse uma base ortogonal, o somatório das projeções iria compor o sinal original $\mathbf{x}$. Contudo, na teoria de *frames*, este somatório produz um sinal modificado como

$$\mathbf{x}' = \mathbf{S}\mathbf{x} = \sum_{i=0}^{I-1} c_k \boldsymbol{\varphi}_i, \tag{2.17}$$

onde $\mathbf{S}$ é uma matriz quadrada de dimensões $N_x \times N_x$ e é denominado operador *frame*. Com essa definição do operador podemos escrever que

$$\langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle = \sum_{i=0}^{I-1} |c_i|^2 \tag{2.18}$$

$$A\|\mathbf{x}\|^2 \leq \langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle \leq B\|\mathbf{x}\|^2. \tag{2.19}$$

Se a condição de *frame* for satisfeita, podemos afirmar que o operador *frame* $\mathbf{S}$ é positivo definido, possui inversa $\mathbf{S}^{-1}$ e é hermitiano, ou seja, $\mathbf{S} = \mathbf{S}^*$. Com a propriedade de invertibilidade, é possível reconstruir o sinal a partir desses coeficientes

gerados (c_i)

$$\begin{aligned}\mathbf{x} &= \mathbf{S}^{-1}\sum_{i=0}^{I-1} c_k\boldsymbol{\varphi}_i \\ &= \sum_{i=0}^{I-1} c_i\mathbf{S}^{-1}\boldsymbol{\varphi}_i \\ &= \sum_{i=0}^{I-1} c_i\boldsymbol{\gamma}_i,\end{aligned} \tag{2.20}$$

onde $\boldsymbol{\gamma}_i$ é denominado *frame* dual e forma uma outra família de funções $\boldsymbol{\Gamma}$ que é capaz de sintetizar o sinal $\mathbf{x}$ a partir dos coeficientes c_i.

O *frame* de Gabor é um tipo de família de funções definida no espaço $L^2(\mathbb{R})$ formada através de atrasos temporais e deslocamentos frequenciais de uma função-janela qualquer (similar à construção das funções-base no caso da STFT). Um exemplo desta família $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_{m,k} \mid \sum|\varphi_{m,k}[n]|^2 < \infty, m \in \{0,1,\ldots,M-1\}, K \in \{0,1,\ldots,K-1\}\}$ pode ser construído como

$$\varphi_{m,k}[n] = \varphi[n-ma]e^{j2\pi bkn}, \tag{2.21}$$

onde $\varphi[n]$ é uma função-janela, a é um passo temporal e b é um passo frequencial. Os subíndices m e k, assim como na STFT, são referentes aos atrasos e deslocamentos de frequência, respectivamente. Nesta família de funções de exemplo, foram escolhidos a e b constantes; contudo, para gerar um *frame* de Gabor não é necessário que esses parâmetros sejam constantes. Devemos ressaltar que a função-janela $\boldsymbol{\varphi}$ e os parâmetros a e b devem ser escolhidos de forma que a família de funções $\boldsymbol{\Phi}$ esteja dentro da condição de *frame* descrita em (2.14).

Como veremos a seguir, o uso de *frames* de Gabor para gerar uma representação tempo-frequencial permite grades com espaçamentos variáveis e ainda possibilita a reconstrução perfeita para alguns casos.

2.3.1 Espaçamento temporal variante

Tendo como fundamento o conceito de *frames*, explicaremos agora uma forma de representação tempo-frequencial proposta em [24], na qual o espaçamento entre as componentes temporais é variável ao longo do plano tempo-frequência. A partir de um conjunto de funções $\boldsymbol{\varphi}_m = \varphi[n-ma_m]$, onde $\varphi[n]$ é uma função-janela de comprimento N_x, $m \in \{0,1,...,M-1\}$ e a_m é um passo temporal dependente de m, podemos gerar um *frame* de Gabor $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_{m,k} \mid \sum|\varphi_{m,k}[n]|^2 < \infty, m \in$

$\{0, 1, \ldots, M-1\}, K \in \{0, 1, \ldots, K-1\}\}$ tal que

$$\boldsymbol{\varphi}_{m,k}[n] = \boldsymbol{\varphi}_m[n] e^{-j2\pi b_m k n}, \tag{2.22}$$

onde b_m é o passo frequencial que depende de m. A decomposição em coeficientes de projeção sobre essa base é dada por

$$c_{m,k} = \langle \mathbf{x}, \boldsymbol{\varphi}_{m,k} \rangle, \tag{2.23}$$

onde $\mathbf{x}$ e $\boldsymbol{\varphi}_{m,k}$ são, respectivamente, o sinal e uma função na forma vetorial de tamanho $N_x \times 1$.

Interpretando esse *frame* de Gabor como um plano tempo-frequência, é possível ver que o espaçamento temporal a_m é variante ao longo de m, podendo, portanto, ser irregular. Contudo, o espaçamento frequencial b_m é fixo em relação a k e varia de acordo com o m. Assim, para cada m podemos ter um espaçamento frequencial distinto porém uniforme. Um exemplo da configuração de espaçamento tempo-frequência que esse *frame* de Gabor pode gerar é apresentado na Figura 2.5.

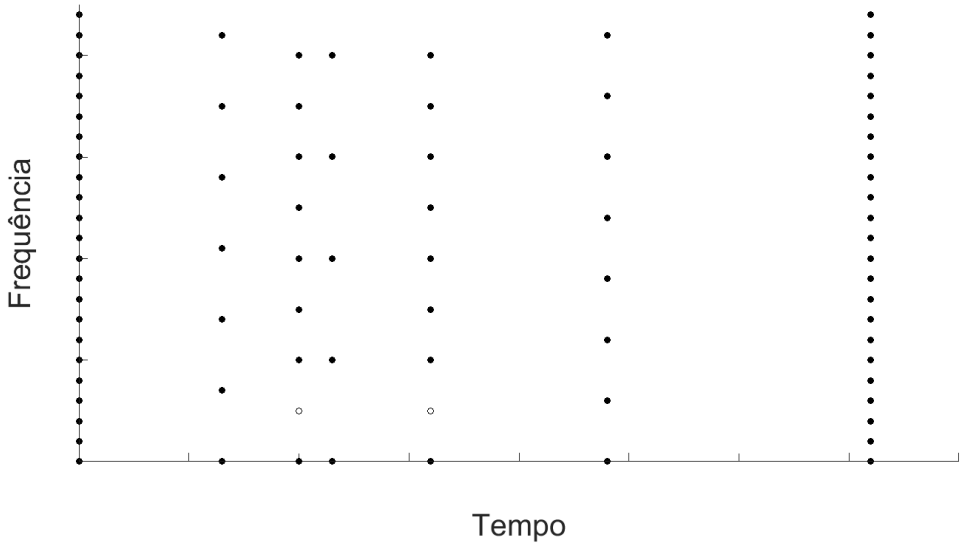

Figura 2.5: Plano tempo-frequência com resolução temporal variante.

Contudo, o *frame* dual que permite a reconstrução perfeita do sinal original ainda necessita ser explicitado. Assim, definiremos o operador *frame* dessa família de funções $\boldsymbol{\Phi}$ como

$$\mathbf{Sx} = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{m,k} \boldsymbol{\varphi}_{m,k}, \tag{2.24}$$

onde $\mathbf{S}$ é uma matriz de dimensão $N_x \times N_x$. Para obter o *frame* dual, precisaremos

de um operador *frame* **S** que seja invertível. Como **Sx** é um novo sinal deformado a partir de $x[n]$, chamá-lo-emos de **y** para facilitar. Desenvolvendo a equação (2.24) temos,

$$\begin{aligned}
\mathbf{Sx} = \mathbf{y} &= \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} c_{m,k}\boldsymbol{\varphi}_{m,k} \\
y[n'] &= \sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\left(\sum_{n=0}^{N_x-1}\varphi^*_{m,k}[n]x[n]\right)\varphi_{m,k}[n'] \\
&= \sum_{n=0}^{N_x-1} x[n]\sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\varphi^*_{m,k}[n]\varphi_{m,k}[n'] \\
&= \sum_{n=0}^{N_x-1} x[n]\sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\varphi^*_{m}[n]e^{j2\pi b_m kn}\varphi_{m}[n']e^{-j2\pi b_m kn'} \\
&= \sum_{n=0}^{N_x-1} x[n]\sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\varphi^*_{m}[n]\varphi_{m}[n']e^{j2\pi b_m k(n-n')} \\
&= \sum_{n=0}^{N_x-1} x[n]\sum_{m=0}^{M-1}\varphi^*_{m}[n]\varphi_{m}[n']\left(\sum_{k=0}^{K-1}e^{j2\pi b_m k(n-n')}\right). \qquad (2.25)
\end{aligned}$$

O somatório de exponenciais complexas (contido dentro dos parênteses) é equivalente ao somatório de impulsos de Dirac , de acordo com o teorema de *Parseval* [18]. Desse modo, podemos reescrever a equação acima como

$$\begin{aligned}
y[n'] &= \sum_{n=0}^{N_x-1} x[n]\sum_{m=0}^{M-1}\varphi^*_{m}[n]\varphi_{m}[n']\left(\frac{1}{b_m}\sum_{s=0}^{S-1}\delta\left[n-n'-\frac{s}{b_m}\right]\right) \\
y[n'] &= \sum_{m=0}^{M-1}\frac{1}{b_m}\varphi_m[n']\sum_{s=0}^{S-1} x\left[n'+\frac{s}{b_m}\right]\varphi^*_m\left[n'+\frac{s}{b_m}\right]. \qquad (2.26)
\end{aligned}$$

De acordo com [24], identifica-se um caso particular para essa expressão que simplificará os cálculos com relação ao segundo somatório. Limita-se o suporte temporal de $\boldsymbol{\varphi}_m$ por

$$\sup(\boldsymbol{\varphi}_m) \leq \frac{1}{b_m}. \qquad (2.27)$$

Dessa forma, sobreposições entre $\varphi_{m_1}\left[n'+s_i/b_{m_1}\right]$ e $\varphi_{m_2}\left[n'+s_j/b_{m_2}\right]$ (para quaisquer s_i e s_j adjacentes) não ocorrerão. Portanto, analisando os elementos

do somatório em s, conclui-se que a equação (2.26) é 0 para $s \neq 0$. Logo, temos

$$\begin{aligned} y[n'] &= \sum_{m=0}^{M-1} \frac{1}{b_m} \varphi_m[n'] \varphi_m^*[n'] x[n'] \\ \mathbf{Sx} &= \sum_{m=0}^{M-1} \frac{1}{b_m} \boldsymbol{\varphi}_m \boldsymbol{\varphi}_m^* \mathbf{x} \\ \mathbf{S} &= \sum_{m=0}^{M-1} \frac{1}{b_m} \boldsymbol{\varphi}_m \boldsymbol{\varphi}_m^*. \end{aligned} \tag{2.28}$$

O operador *frame* $\mathbf{S}$ pode ser visto como uma matriz hermitiana, e portanto apresenta a propriedade de invertibilidade. À vista disso, o *frame* dual $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}_{m,k} \mid \sum |\gamma_{m,k}[n]|^2 < \infty, m \in \{0, 1, \ldots, M-1\}, K \in \{0, 1, \ldots, K-1\}\}$, é dado por

$$\begin{aligned} \mathbf{Sx} &= \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{m,k} \boldsymbol{\varphi}_{m,k} \\ \mathbf{S}^{-1}\mathbf{Sx} &= \mathbf{S}^{-1} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{m,k} \boldsymbol{\varphi}_{m,k} \\ \mathbf{x} &= \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{m,k} \mathbf{S}^{-1} \boldsymbol{\varphi}_{m,k} \\ &= \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} c_{m,k} \boldsymbol{\gamma}_{m,k}, \end{aligned} \tag{2.29}$$

com

$$\begin{aligned} \boldsymbol{\gamma}_{m,k} &= \mathbf{S}^{-1} \boldsymbol{\varphi}_{m,k} \\ &= \frac{\boldsymbol{\varphi}_{m,k}}{\sum_{i=0}^{I-1} \frac{1}{b_i} |\boldsymbol{\varphi}_i|^2}. \end{aligned} \tag{2.30}$$

Conclui-se que a família de funções $\boldsymbol{\Phi}$ é considerada um *frame* de Gabor se impusermos a restrição ao suporte temporal definida na equação (2.27). Ressaltamos que esta condição foi utilizada apenas para a simplificação dos cálculos e, portanto, é possível gerar um *frame* de Gabor com essa configuração de espaçamento, desde que exista um operador *frame* $\mathbf{S}$ não-singular.

O exemplo do sinal de três tons no domínio tempo-frequência com este tipo de espaçamento proporcionado pelo *frame* de Gabor pode ser observado na Figura 2.6. Neste caso, escolhemos um espaçamento temporal de acordo com a envoltória das notas musicais, ou seja, intervalos curtos em oscilações rápidas e intervalos maiores com oscilações lentas. A função-janela escolhida foi a janela de Hann e o

comprimento de cada função Φ é determinada pela equação de suporte (2.27).

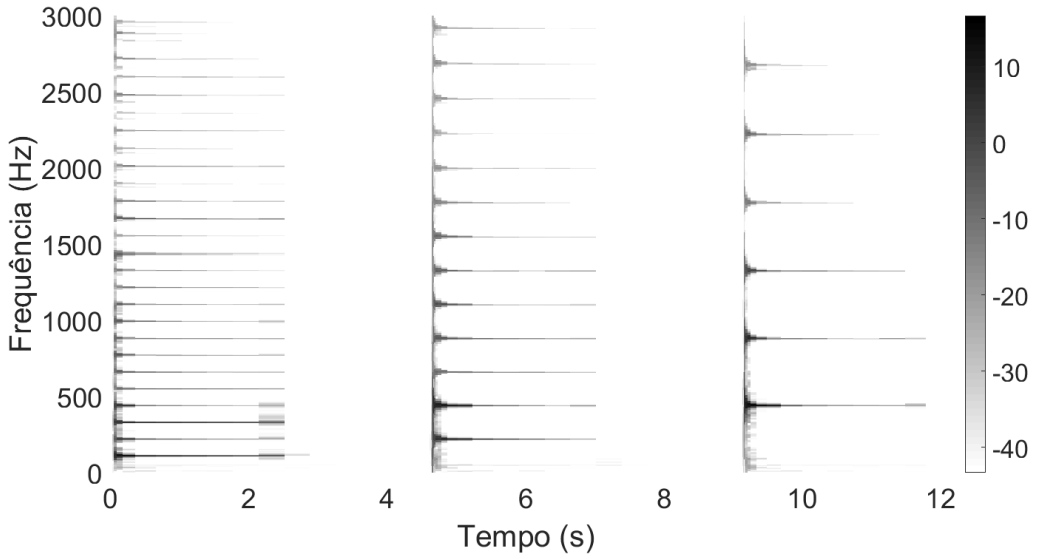

Figura 2.6: Módulo do sinal no domínio do tempo-frequência com espaçamento temporal variante.

2.3.2 Espaçamento espectral variante

Uma construção análoga à vista na Subseção 2.3.1 permite uma resolução frequencial variante ao longo do plano tempo-frequência. Neste caso, utilizaremos um conjunto de funções base $\boldsymbol{\psi}_k = \psi[n]e^{-j2k\pi b_k n}$, onde $\psi[n]$ é uma função-janela de comprimento N_x, $k \in \{0, 1, ..., K-1\}$ e b_k é um passo frequencial dependente de k, para gerar um *frame* de Gabor $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_{m,k} \mid \sum |\psi_{m,k}[n]|^2 < \infty, m \in \{0, 1, \ldots, M-1\}, K \in \{0, 1, \ldots, K-1\}\}$

$$\psi_{m,k}[n] = \psi_k[n - ma_k], \tag{2.31}$$

onde a_k é o passo temporal que depende de k. Calculando a DFT desta expressão e utilizando a propriedade de deslocamento no tempo explicada em [18], temos

$$\widehat{\psi_{m,k}}[s] = \widehat{\psi_k}[s]e^{-j2\pi a_k ms}, \tag{2.32}$$

denotando como $\widehat{\ }$ a operação de DFT.

A decomposição em coeficientes de projeção sobre essa família de funções é dada por

$$d_{m,k} = \langle \widehat{\mathbf{x}}, \widehat{\boldsymbol{\psi}_{m,k}} \rangle, \tag{2.33}$$

onde $\widehat{\mathbf{x}}$ e $\widehat{\boldsymbol{\psi}_{m,k}}$ são, respectivamente, o sinal e uma função na forma vetorial de

tamanho $N_x \times 1$.

De forma análoga ao que foi discutido na Subseção 2.3.1, o plano tempo-frequência gerado por esse *frame* de Gabor possui um espaçamento frequencial b_k variante ao longo de k e, dessa forma, pode ser irregular. Já o espaçamento temporal a_k, por ser dependente de k e fixo em relação a m, pode ser distinto para cada componente frequencial, porém distribuído uniformemente ao longo dessa componente frequencial. Na Figura 2.7, temos um exemplo de plano tempo-frequência que esta especificação de *frames* de Gabor pode construir.

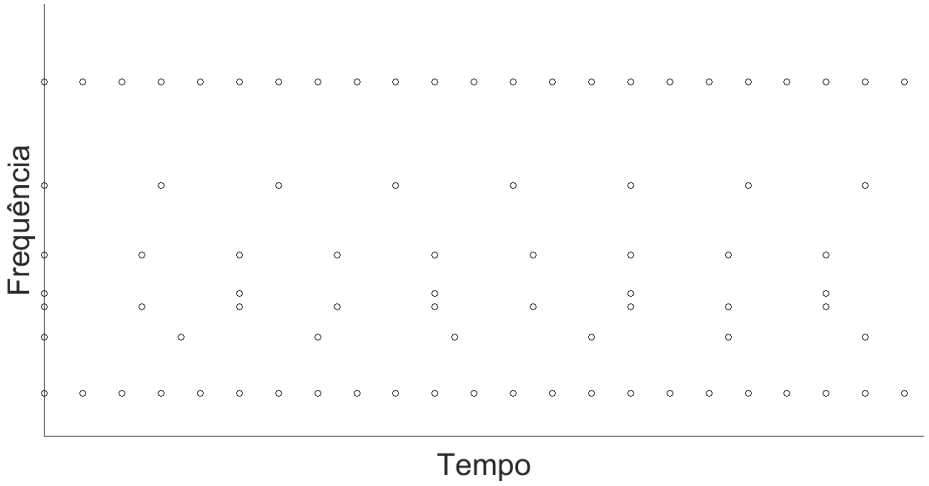

Figura 2.7: Plano tempo-frequência com resolução espectral variante.

Para cumprir com a condição de *frame* e apresentar um *frame* dual que reconstrua o sinal original é necessário realizar a dedução matemática similar à desenvolvida na Subseção 2.3.1[4]. O operador *frame* $\mathbf{S}$ desta família de funções $\mathbf{\Psi}$ é dado por

$$\mathbf{S}\hat{\mathbf{x}} = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{m,k} \widehat{\boldsymbol{\psi}_{m,k}}, \tag{2.34}$$

onde $\mathbf{S}$ é uma matriz $N_x \times N_x$. Novamente realizaremos uma substituição de função

[4]Resumiremos alguns detalhes sobre essa dedução, uma vez que o desenvolvimento é análogo.

para simplificar os cálculos. Assim, desenvolvendo a equação (2.34) temos

$$\begin{aligned}
\mathbf{S}\widehat{\mathbf{x}} = \widehat{\mathbf{y}} &= \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} d_{m,k}\widehat{\boldsymbol{\psi}_{m,k}} \\
\widehat{y}[s'] &= \sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\left(\sum_{s=0}^{S-1}\widehat{\psi_{m,k}}^{*}[s]\widehat{x}[s]\right)\widehat{\psi_{m,k}}[s'] \\
\widehat{y}[s'] &= \sum_{k=0}^{K-1}\frac{1}{a_k}\widehat{\psi_k}[s']\sum_{m'=0}^{M'-1}\widehat{x}\left[s'+\frac{m'}{a_k}\right]\widehat{\psi_k}^{*}\left[s'+\frac{m'}{a_k}\right].
\end{aligned} \tag{2.35}$$

Restringindo o suporte frequencial de $\widehat{\boldsymbol{\psi}_k}$ com

$$\sup(\widehat{\boldsymbol{\psi}_k}) \leq \frac{1}{a_k}, \tag{2.36}$$

não haverá sobreposições entre duas funções $\widehat{\boldsymbol{\psi}_k}$ ao longo do eixo frequencial. Logo, a equação (2.35) é 0 para $m' \neq 0$. Portanto temos

$$\begin{aligned}
\widehat{y}[s'] &= \sum_{k=0}^{K-1}\frac{1}{a_k}\widehat{\psi_k}[s']\widehat{\psi_k}^{*}[s']\widehat{x}[s'] \\
\mathbf{S} &= \sum_{k=0}^{K-1}\frac{1}{a_k}\widehat{\boldsymbol{\psi}_k}\widehat{\boldsymbol{\psi}_k}^{*}.
\end{aligned} \tag{2.37}$$

Sendo o operador *frame* $\mathbf{S}$ uma matriz não singular, podemos obter o *frame* dual $\widehat{\boldsymbol{P}} = \{\widehat{\boldsymbol{\rho}_{m,k}} \mid \sum|\rho_{m,k}[n]|^2 < \infty, m \in \{0,1,\ldots,M-1\}, K \in \{0,1,\ldots,K-1\}\}$ através de

$$\begin{aligned}
\mathbf{S}\widehat{\mathbf{x}} &= \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} d_{m,k}\widehat{\boldsymbol{\psi}_{m,k}} \\
\widehat{\mathbf{x}} &= \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} d_{m,k}\widehat{\boldsymbol{\rho}_{m,k}},
\end{aligned} \tag{2.38}$$

com

$$\widehat{\boldsymbol{\rho}_{m,k}} = \frac{\widehat{\boldsymbol{\psi}_{m,k}}}{\sum_{i=0}^{I-1}\frac{1}{a_i}\left|\widehat{\boldsymbol{\psi}_i}\right|^2}. \tag{2.39}$$

Analogamente à NSGT com espaçamento temporal variante, deduzimos que a restrição do suporte frequencial, vista na equação (2.36), implica que a família de funções $\boldsymbol{\Psi}$ é um *frame* de Gabor.

Um caso particular desta configuração do domínio tempo-frequência será deta-

lhado e exemplificado na Seção 2.4.

□

Através dessas duas técnicas apresentadas, podemos representar um sinal sobre grades tempo-frequenciais cujo espaçamento tempo-frequencial é variante ao longo do tempo ou da frequência. Isso permite uma análise de cada sinal com resoluções temporais ou frequenciais específicas.

Contudo, essas configurações ainda não permitem um plano tempo-frequência com as duas resoluções variando simultaneamente. Uma outra desvantagem dessas técnicas é a necessidade de realizar um pré-processamento do sinal para escolher os espaçamentos adequados. Em resumo, não há um método automático para escolha de resolução em cada trecho do sinal. No Capítulo 3 apresentaremos um novo procedimento que suprime essas necessidades.

2.4 *Constant Q-Transform with non-stationary Gabor Transform*

A CQT, apresentada na Seção 2.2, baseia-se na STFT para criar um plano tempo-frequência com espaçamento espectral geométrico. Contudo, não é possível sintetizar os sinais a partir dos coeficientes gerados pelo procedimento proposto por [22].

A técnica da NSGT de resolução espectral variante, mostrada na Subseção 2.3.2, permite uma escolha arbitrária de espaçamento espectral. Assim sendo, é possível escolher o espaçamento geométrico da CQT e ainda reconstruir o sinal a partir dos coeficientes gerados por este plano tempo-frequência. Essa técnica é denominada de *Constant Q-Transform with Non-stationary Gabor Trasform* (CQ-NSGT) e foi apresentada em [25].

Para obter as especificações necessárias, utilizaremos o *frame* de Gabor apresentado na Subseção 2.3.2 $\widehat{\boldsymbol{\Psi}} = \{\boldsymbol{\psi}_{m,k} \mid \sum |\widehat{\psi}_{m,k}[n]|^2 < \infty, m \in \{0, 1, \ldots, M-1\}, K \in \{0, 1, \ldots, K-1\}\}$, em que

$$\widehat{\psi_{m,k}}[s] = \widehat{\psi_k}[s] e^{-j2\pi a_k m s}. \tag{2.40}$$

O *frame* dual $\widehat{\boldsymbol{P}} = \{\boldsymbol{\rho}_{m,k} \mid \sum |\widehat{\rho}_{m,k}[n]|^2 < \infty, m \in \{0, 1, \ldots, M-1\}, K \in \{0, 1, \ldots, K-1\}\}$ para sintetizar o sinal é dado por

$$\widehat{\mathbf{x}} = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{m,k} \widehat{\boldsymbol{\rho}_{m,k}}. \tag{2.41}$$

Para criar o plano tempo-frequência da CQT, devemos centralizar cada função $\widehat{\boldsymbol{\psi}_k}$ sobre as componentes frequenciais descritas na Seção 2.2

$$f_k = f_{\min}\left(\frac{1}{Q}+1\right)^{k-1}. \tag{2.42}$$

com $k \in \{1, 2, ..., K\}$. Apesar de a CQT não permitir a análise da frequência 0 Hz, a CQ-NSGT pode representar esta componente DC e com isso de permitir a reconstrução perfeita. Assim, definimos $f_0 = 0$ Hz.

Com exceção da componente 0 Hz, as larguras das bandas B_k das funções devem ser escolhidas de forma a manter o fator de seletividade da CQT. Portanto,

$$B_0 = 2f_{\min}, \tag{2.43}$$

$$B_k = f_k/Q, \tag{2.44}$$

com $k \in \{1, 2, ..., K\}$.

Por meio da janela $\widehat{h}[s]$ de Hann padrão, isto é, com centro em 0 e com comprimento unitário, podemos construir as funções

$$\widehat{\psi_k}[s] = \widehat{h}\left(\frac{s\frac{f_s}{N_w} - f_k}{B_k}\right), \tag{2.45}$$

com $k \in \{1, 2, ..., K\}$, N_x sendo o comprimento do sinal e f_s a frequência de amostragem. A janela utilizada para a criação da função $\widehat{\psi_0}$ foi a janela de Tukey, que é mais plana no domínio do tempo.

Na Figura 2.8, podemos ver um exemplo de grade tempo-frequência gerada pela CQ-NSGT.

Podemos observar na Figura 2.9 o sinal de exemplo de três tons representado através da CQ-NSGT com Q de 34 e frequência mínima de 55 Hz. Estes parâmetros escolhidos foram idênticos àqueles usados na representação de CQT na Seção 2.2.

A técnica apresentada se baseia na escala cromática para representar os sinais em um plano tempo-frequência. Uma outra técnica similar utiliza uma escala espectral que simula a resolução frequencial do sistema auditivo humano, a *Equivalent Rectangular Bandwidth* (ERB). O método, denominado *ERBlet*, é explicado em [27] e favorece a interpretação tempo-frequencial dos sinais de acordo com a percepção humana de sons.

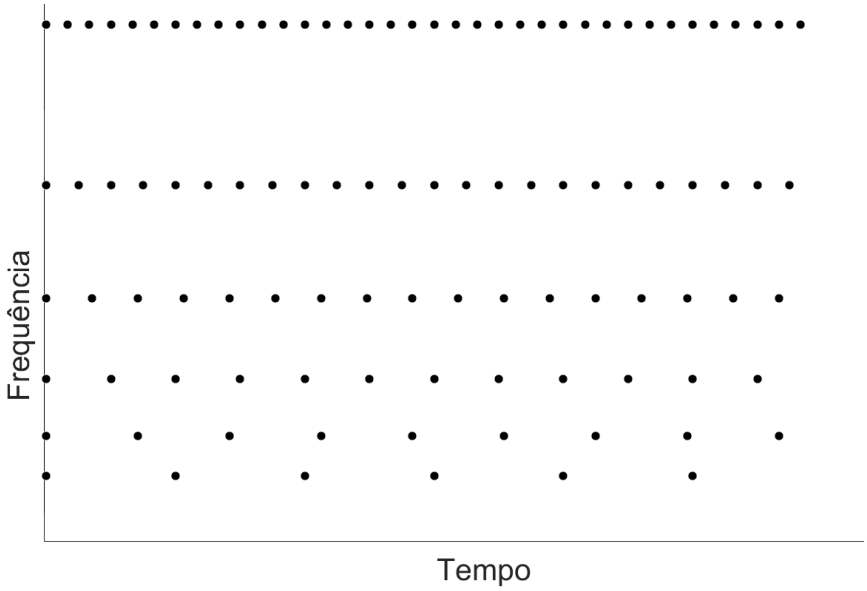

Figura 2.8: Plano tempo-frequência com espaçamento frequencial geométrico.

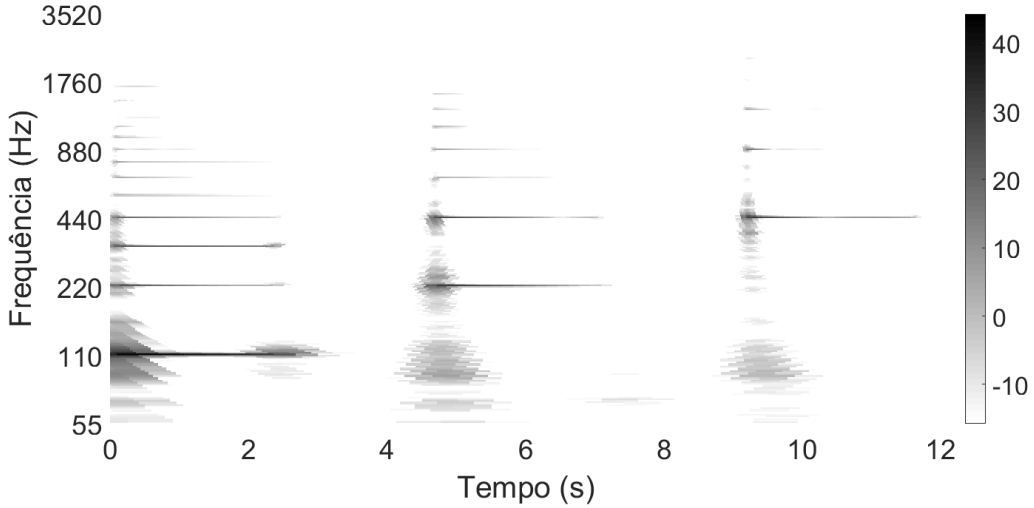

Figura 2.9: Módulo do sinal no domínio do tempo-frequência gerado pela CQ-NSGT.

Capítulo 3

Representação Tempo-Frequência com Resolução Adaptativa

Como visto no Capítulo 2, as representações discretas tempo-frequência constroem uma divisão em grade do plano tempo-frequência que organiza as componentes temporais e espectrais através de uma ponderação de funções centradas em pontos da grade. Na STFT, através das janelas atrasadas no tempo e deslocadas uniformemente na frequência, gera-se uma grade regular. Na CQ-NSGT, a grade é uniformemente distribuída ao longo do eixo temporal e geometricamente distribuída no eixo frequencial. E na NSGT, a grade pode ser irregular ao longo de um dos eixos desde que o outro eixo seja uniforme.

Entretanto, há alguns sinais de música que possuem componentes espectrais que não seguem um comportamento regular, por exemplo quando contêm trechos de comportamentos tonal e percussivo. Nestes casos, a análise tempo-frequencial que se vale das transformadas já mencionadas não será a melhor opção para representar tais sinais.

O presente capítulo trata de uma técnica proposta em [1], que consiste em uma representação tempo-frequência com otimização local da resolução (estrutura da grade subjacente). Esse método se resume em construir a grade no plano tempo-frequência que melhor (segundo um critério de esparsidade) se ajustaria ao sinal analisado em particular através de pequenas regiões complementares entre si e que compõem esse plano. Assim, o sinal é dividido de acordo com cada região e representado múltiplas vezes por meio das técnicas descritas no Capítulo 2. Para cada região, seleciona-se o conjunto de coeficientes que melhor representa o sinal segundo um critério de esparsidade. Dessa forma, o sinal pode ser visualizado em coeficientes distribuídos de forma otimizada e específica para a aplicação pretendida.

Na Figura 3.1, é possível observar uma grade tempo-frequência irregular que pode ser gerada por este método proposto, onde cada símbolo indica uma resolução diferente.

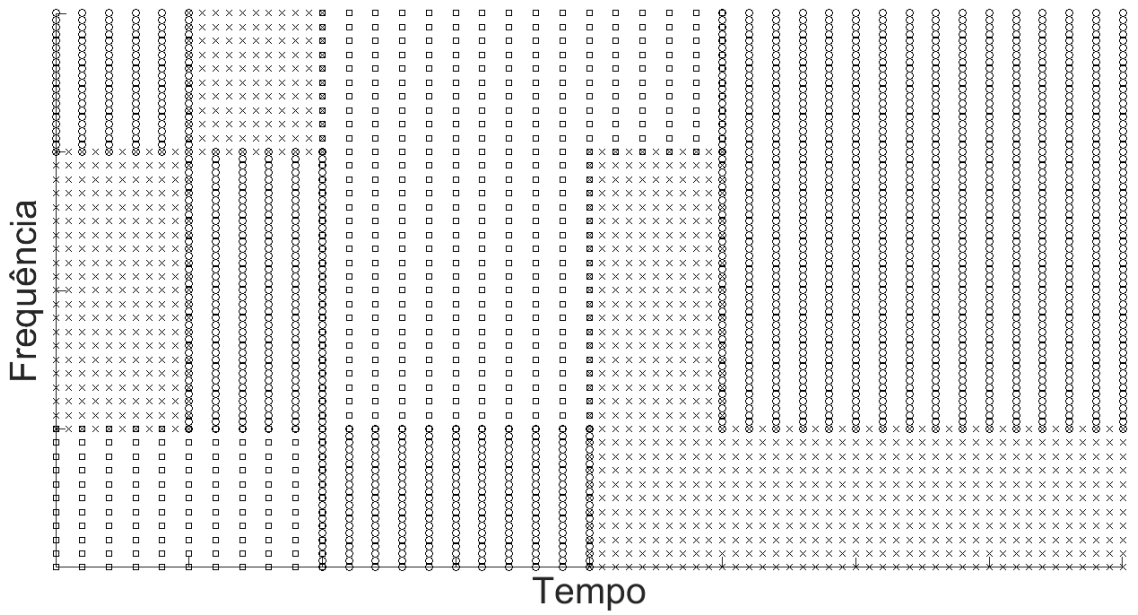

Figura 3.1: Grade tempo-frequência irregular.

Para apresentar esse método, este capítulo se divide em seções que denotam, em ordem cronológica de implementação, os procedimentos a ele inerentes.

3.1 Divisão temporal

Como já mencionamos, essa técnica subdividirá o espaço tempo-frequência em regiões para representar o sinal. Estas regiões, de acordo com o proposto em [1], estão restritas a um formato retangular. Para implementar essa subdivisão, são realizadas subetapas: a divisão temporal e a divisão frequencial. Nesta seção, será especificada apenas a divisão ao longo do eixo temporal realizada através de janelamento. Na Seção 3.3, será detalhado o método usado para implementar a outra subetapa.

O janelamento é um procedimento simples que consiste em multiplicar o sinal original por uma função-janela, extraindo apenas as amostras necessárias dentro de um intervalo de tempo determinado pela janela. Esse processo pode ser descrito matematicamente como

$$x_i[n] = x[n]w_i[n - a_i], \tag{3.1}$$

onde $x[n]$ é o sinal contendo N_x amostras, $w_i[n]$ é a função-janela contendo N_i amostras e o atraso a_i é o deslocamento da janela sobre o intervalo de interesse do sinal a ser analisado. O subíndice i se refere ao trecho temporal em que é feito o janelamento. Uma vez que o sinal original foi janelado, $x_i[n]$ possui, portanto, o comprimento da janela N_i.

Destacamos que os sinais janelados podem apresentar sobreposição de amostras

para melhor representar determinados trechos no domínio tempo-frequência. Devemos ressaltar também que a técnica se compromete a reconstruir o sinal original e, portanto, a soma entre todos os sinais janelados deve, em princípio, recompor o sinal original. Assim temos:

$$x[n] = \sum_{i=0}^{I-1} x_i[n], \quad \forall x[n] \quad \Rightarrow \sum_{i=0}^{I-1} w_i[n - a_i] = 1, \tag{3.2}$$

onde I é a quantidade de divisões que devem ser realizadas no eixo temporal para a criação das regiões.

A restrição expressa em (3.2) pode ser satisfeita para quase todo n, por exemplo, com janelas de Hann e atrasos de 50% da duração da janela ($a_i = N_i/2$), como na Figura 3.2. Porém, podemos observar que as primeiras amostras e as últimas não teriam sobreposição e, logo, não seria possível reconstruir perfeitamente o sinal usando essa divisão.

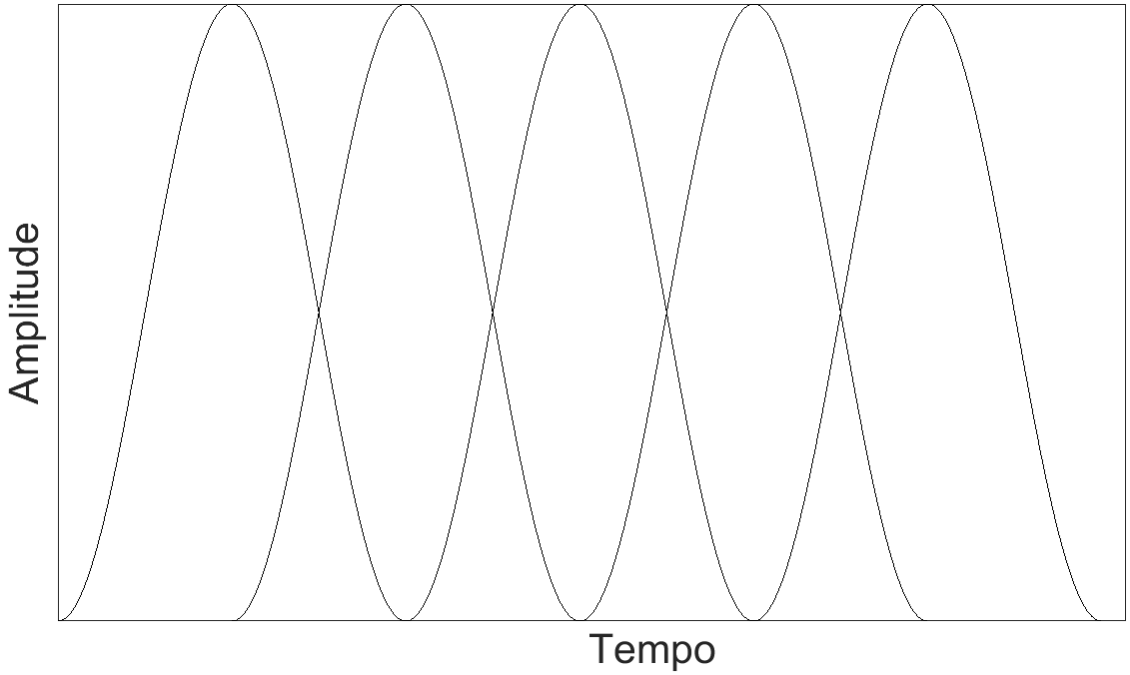

Figura 3.2: Exemplo de sobreposição entre janelas de Hann de comprimento 1024 amostras.

Uma solução que adotamos para resolver este problema nesta dissertação é adicionar $N_i - a_i$ zeros no início e no fim do sinal. Dessa forma, as sobreposições entre as janelas iniciais e as sobreposições entre as janelas finais permitem que a condição da equação (3.2) seja cumprida nas amostras úteis do sinal.

Em [15], explica-se que as funções-janela normalmente utilizadas só satisfazem a restrição (3.2) com uma combinação de parâmetros específica. Para contornar isso, são oferecidas duas soluções: uma modificação das janelas tradicionais e um novo

tipo de janela. Ambas estendem o conjunto de parâmetros para que as funções-janela obedeçam a limitação da equação (3.2).

A adaptação proposta em [15] sobre as janelas é dada por:

$$w_i'[n] = \begin{cases} \frac{\sum_{p=0}^{n} w_i[p]}{\sum_{p=0}^{N_i-a_i} w_i[p]}, & 0 \leq n \leq N_i - a_i - 1 \\ 1, & N_i - a_i \leq n \leq a_i - 1 \\ \frac{\sum_{p=n-a_i+1}^{N_i-a_i} w_i[p]}{\sum_{p=0}^{N_i-a_i} w_i[p]}, & a_i \leq n \leq N_i - 1, \end{cases} \tag{3.3}$$

onde $w_i'[n]$ é a função-janela modificada e $w_i[n]$ é a função-janela original. E o novo tipo de janela, denominada senoidal, é expresso por

$$w_i[n] = \begin{cases} \text{sen}\left[\frac{\pi(n+0{,}5)}{2(N_i-a_i)}\right], & 0 \leq n \leq N_i - a_i - 1 \\ 1, & N_i - a_i \leq n \leq a_i - 1 \\ \text{sen}\left[\frac{\pi(N_i-n-0{,}5)}{2(N_i-a_i)}\right], & a_i \leq n \leq N_i - 1. \end{cases} \tag{3.4}$$

Deve-se notar que, nos dois casos as janelas só podem ser construídas se $a_i \geq N_i/2$, o que impõe que apenas janelas adjacentes se sobreponham. Além disso, apesar de facilitar a reconstrução dos sinais após seu janelamento, a alteração de janela descrita na equação (3.3) modifica o formato original da janela e suas características espectrais.

Na Figura 3.3, temos um exemplo de sobreposição entre janelas retangulares adaptadas de comprimento de 1024 amostras atrasadas (a_i) entre si de 768 amostras.

Nesse ponto, ainda temos uma limitação na construção de janelas no que diz respeito ao atraso a_i. Para resolver este problema, observamos o janelamento com as janelas de Hann sem qualquer modificação: com um atraso de 50% do comprimento da janela, a soma das janelas resulta em $\sum w_i[n - N_i/2] = 1$ (exceto para o início e o fim do sinal); já com um atraso de 25% do comprimento, essa mesma soma é dada por $\sum w_i[n - N_i/4] = 2$; repetindo isso para um atraso de 12,5% do comprimento, obteremos uma soma igual a 4. Em suma, mantendo o mesmo formato de janela e usando um atraso equivalente à metade do original, o somatório apenas duplica o valor final sem que isso provoque uma deformação irreversível no sinal.

Inspirados por este comportamento, desenvolvemos nesta dissertação um algoritmo para computar qualquer percentual janelamento com reconstrução perfeita. Dado um tipo de janela de tamanho N_i que desejamos adaptar e um atraso $a_i < N_i$:

1. Calculamos $a_i' = 2^\beta a_i$, para um $\beta \in \mathbb{N}$ convenientemente escolhido para que a

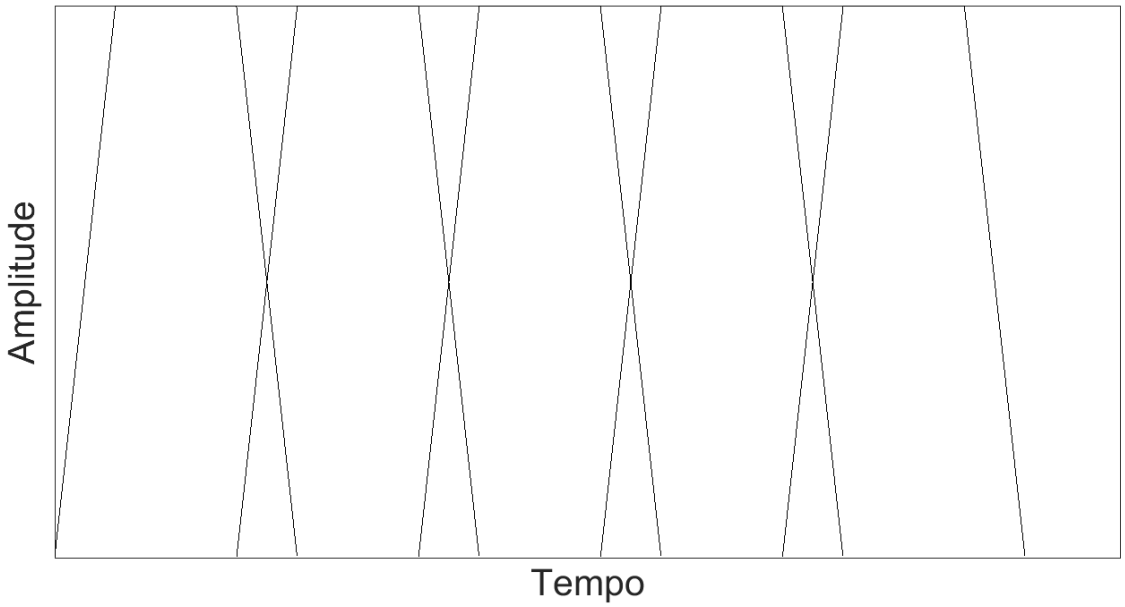

Figura 3.3: Exemplo de sobreposição entre janelas de retangulares adaptadas de comprimento 1024 amostras como estruturada na equação (3.3).

restrição de construção de janelas de [15] seja satisfeita, isto é, $a'_i/N_i \geq 1/2$;

2. Com o tipo da janela, a'_i e N_i já definidos, construímos uma janela a partir da equação descrita na equação (3.4) ou (3.3);

3. Por fim, com esta janela criada, ajustamos a condição de ganho unitário para o janelamento com o atraso a_i original.

De forma equivalente ao caso da janela de Hann, o somatório resulta em um valor constante $2^\beta = a'_i/a_i$. Isso ocorre porque, uma vez a janela já construída, a razão entre os atrasos em forma de potência de 2 permite que cada amostra do sinal seja observada a'_i/a_i vezes pelas janelas e cada observação é complemento de uma outra. Assim, o somatório resulta constante, possibilitando a reconstrução perfeita do sinal.

Através desse algoritmo, o uso de um atraso $a_i \leq N_i$ na etapa de janelamento torna a obtenção do sinal original factível, sendo necessária apenas uma multiplicação na etapa de reconstrução pelo fator a_i/a'_i.

Portanto, a reconstrução do sinal a partir dos sinais janelados pode ser descrita por

$$\widetilde{x}[n] = \frac{a_i}{a'_i} \sum_{i=0}^{I-1} \widetilde{x}_i[n], \tag{3.5}$$

onde $\widetilde{x}[n]$ é o sinal reconstruído e $\widetilde{x}_i[n]$ é o sinal reconstruído referente ao intervalo analisado i.

3.2 Análise tempo-frequência

Após ter realizado o janelamento no sinal, iremos representar cada trecho resultante no domínio tempo-frequência com diferentes resoluções. Essas várias representações servirão como opções para a etapa de escolha de coeficientes definir a melhor representação para cada região.

Cada sinal janelado será representado L vezes no domínio tempo-frequência com resoluções distintas. Esse procedimento consiste em projetar cada sinal janelado em uma função-janela de análise g_l que irá se deslocar ao longo do domínio tempo-frequência através da grade Λ_l, com $l \in \{0, 1, \ldots, L-1\}$. Obteremos, portanto, L conjuntos de coeficientes que representam cada um destes sinais janelados. Através de uma função-janela de síntese h_l, cada grupo de coeficientes possibilita a reconstrução do sinal janelado a que se refere.

As transformadas exploradas no Capítulo 2 podem ser aplicadas nesta etapa para representar cada intervalo do sinal no domínio tempo-frequência. Na STFT, a grade Λ_l é distribuída uniformemente e a função-janela desempenha o papel das funções de análise g_l e de síntese h_l. A grade Λ_l construída pela representação NSGT ou CQ-NSGT permite uma resolução mais irregular em um dos eixos, e as funções de análise g_l e de síntese h_l são dadas pelos *frames* de Gabor e seu dual, respectivamente.

Como já explicitado, um mesmo intervalo de tempo do sinal pode ser representado por L conjuntos de coeficientes no domínio tempo-frequência. Uma vez que o objetivo final é obter a melhor representação (de acordo com o critério de esparsidade) que define cada região, o processo de formação de cada conjunto de coeficientes deve se dar através de grades com resoluções ou até de transformadas diferentes. Assim, com várias opções distintas, a escolha de coeficientes será feita de forma mais eficaz e a técnica obterá melhores resultados para cada região.

Dessa forma, os coeficientes de decomposição são expressos por

$$c_{i,l}[m,k] = \sum_{n=0}^{N_i-1} \left(g_l[n - ma_l]e^{j2\pi b_l kn}\right)^* x_i[n], \tag{3.6}$$

onde m é o índice que localiza o coeficiente no eixo temporal, k é o índice que localiza o coeficiente no eixo frequencial, e a_l e b_l são, respectivamente, os espaçamentos entre componentes temporais e entre componentes frequenciais definidos pela grade Λ_l.

A reconstrução do sinal janelado $\widetilde{x_i}[n]$ é inicialmente dada por

$$\widetilde{x_i}[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \widetilde{c_{i,l}}[m,k] h_l[n], \tag{3.7}$$

onde $\widetilde{c_{i,l}}[m,k]$ são os coeficientes de decomposição após o processamento pretendido, e M e K são as quantidades de pontos da grade Λ_l ao longo dos eixos temporal e frequencial, respectivamente.

3.3 Divisão frequencial

As etapas anteriores foram responsáveis por dividir e limitar o sinal em intervalos temporais e representá-los por uma técnica tempo-frequencial. Portanto, iremos agora concluir a criação das regiões tempo-frequenciais através da divisão frequencial. Em [1], são citadas duas formas de realizar este procedimento: através de banco de filtros ou por funções-peso.

O primeiro método consiste em projetar um banco de filtros $\Theta = \{\theta_p[n] \mid \sum |\theta_p[n]|^2 < \infty, p \in \{0,1,\ldots,P-1\}\}$, onde $\theta_p[n]$ é a resposta ao impulso do p-ésimo filtro. As respostas em frequência desses filtros devem ser compatíveis com os intervalos frequenciais previstos pela divisão em regiões. Assim, ao filtrar os sinais janelados por esse banco, obtemos, basicamente, o sinal original dividido de acordo com as regiões. Essa separação pode ser descrita como

$$\left(x_i \circledast \theta_p\right)[n] = x_{i,p}[n], \tag{3.8}$$

onde $\circledast$ indica a operação de convolução circular entre sinais [28]. Para obter a reconstrução, utilizamos o teorema de Parseval e deduzimos que uma condição similar à do janelamento deve ser satisfeita, a saber:

$$\sum_{p=0}^{P-1} \widehat{\theta_p} = 1. \tag{3.9}$$

Entretanto, de acordo com [1], as técnicas de processamento espectral de sinais frequentemente evitam manipulações no domínio temporal e, portanto, modificações no domínio frequencial ou tempo-frequencial são mais apropriadas. Uma vez que já faremos uso da transformação de sinais no domínio tempo-frequência, podemos usar funções-peso para implementar essas divisões no eixo frequencial.

Ao gerar os coeficientes de decomposição $c_{i,l}[m,k]$, como descrito na Seção 3.2, obtemos essencialmente uma matriz de duas dimensões, onde uma dimensão (referente ao índice m) representa o eixo temporal e a outra dimensão (referente ao índice k) representa o eixo frequencial. Através dessa interpretação, construiremos uma função-peso de mesma dimensão que será multiplicada por esses coeficientes. Os elementos da função-peso referentes ao eixo espectral devem coincidir com a

magnitude dos filtros θ_p. Assim teremos

$$c_{i,l,p}[m,k] = c_{i,l}[m,k]\sigma_p[m,k], \tag{3.10}$$

onde $\sigma_p[m,k]$ é a função-peso puramente real de dimensão $M \times K$.

Com o sinal subdividido de acordo com as regiões, faz-se necessário escolher quais das L representações tempo-frequência geradas em cada uma dessas regiões é a mais adequada para representar o sinal neste trecho.

3.4 Escolha dos coeficientes

O artigo [1] indica que as representações esparsas são as mais úteis em processamento de sinais em diversas aplicações. O fato de ter mais informação contida em poucos coeficientes favorece, por exemplo, a área de compressão de dados. Outras aplicações de representações esparsas em processamento de sinais podem ser vistas em [29]. Portanto, utilizando o critério de esparsidade iremos decidir quais dos coeficientes melhor representam cada região do espaço tempo-frequência analisado.

Um indicador de esparsidade proposto no artigo [1] baseia-se na entropia de Rényi. Esta medida de entropia, apresentada por Rényi em [30], é uma generalização da entropia de Shannon e calcula a quantidade de informação de uma distribuição de probabilidades de uma variável aleatória. A expressão dessa medida para uma variável discreta é dada por

$$H_\alpha\left[p_0, p_1, \cdots, p_{K-1}\right] = \frac{1}{1-\alpha}\log_2 \sum_{k=0}^{K-1} \left(p_k\right)^\alpha, \tag{3.11}$$

onde $p_0, p_1, \cdots, p_{K-1}$ são as probabilidades (assumidas não-nulas) de cada evento, K é o número total de eventos, α é a ordem da entropia de Rényi ($\alpha > 0$ e $\alpha \neq 1$) e H_α é o valor da entropia de Rényi. Analisando a ordem da entropia, podemos obter algumas interpretações sobre este cálculo: com α maior que 1, a entropia tende a desconsiderar as baixas probabilidades; quando α se aproxima de 1, a entropia de Rényi tende à entropia de Shannon.

A abordagem utilizada pelo artigo [1] de quantificação da esparsidade é formulada originalmente por [31] como uma adaptação desta entropia. Assim, temos um cálculo de entropia alterado e específico para a interpretação de esparsidade de espaços

tempo-frequência gerados pela STFT como:

$$H_\alpha\left[C_{i,l,p}\right] = \frac{1}{1-\alpha}\log_2 \sum_{m=0}^{M-1}\sum_{k=0}^{K-1}\left(\frac{\|C_{i,l,p}\|^2[m,k]}{\sum_{m'=0}^{M-1}\sum_{k'=0}^{K-1}\|C_{i,l,p}\|^2[m',k']}\right)^\alpha + \log_2\frac{a_l b_l}{MK}. \quad (3.12)$$

Nesta definição, têm-se $C_{i,l,p}$ como os coeficientes de decomposição de uma região construída pelo janelamento de $w_i[n]$ e a função-peso $f_p[m,k]$ sobre uma grade Λ_l, m como o índice relativo ao intervalo temporal, k como o índice relativo ao componente espectral, M como o número de divisões temporais desta grade Λ_l, K como o número de componentes espectrais desta grade Λ_l e, a_l e b_l como os espaçamentos temporais e frequenciais desta grade Λ_l.

A principal alteração observada é a normalização dos coeficientes de decomposição, uma vez que estes não apresentam as mesmas restrições de uma distribuição de probabilidades. Como em [31] o cálculo é originalmente desenvolvido sobre representações tempo-frequência contínuas, a transformação para o uso em representações discretas requer uma outra parcela ($\log_2(ab/MK)$) referente à discretização do espaço tempo-frequência. Para usar esta medida de esparsidade em espaços tempo-frequência gerados por outras técnicas é necessária uma alteração nesta parcela de discretização coerente com a resolução da grade.

Analisando-se a equação (3.11), é possível concluir que quanto mais esparsa for uma distribuição de probabilidades, menor será sua entropia de Rényi. Analogamente, podemos observar que para a equação (3.12) quanto menor for a entropia de Rényi, menor será o espalhamento (incerteza) dos coeficientes, ou seja, maior será a esparsidade destes coeficientes.

Portanto, a escolha dos conjuntos de coeficientes para cada região será determinada pela representação que possuir o menor valor de entropia de Rényi. Assim, temos o conjunto de coeficientes ótimos expresso por $C_{i,\bar{l},p}$, onde

$$\bar{l} = \mathrm{argmin}_l\left(H_\alpha\left[C_{i,l,p}\right]\right). \quad (3.13)$$

Em [1], é sugerida uma ordem de entropia $\alpha = 0{,}3$, porque este valor é análogo à potência usada no mapeamento dos níveis de audibilidade em fons para o nível de audibilidade subjetiva em sones. Esta equivalência implica uma escolha de coeficientes relacionada com a Psicoacústica e, portanto, favorece uma análise mais próxima da percepção humana.

Apesar de [1] citar apenas a entropia de Rényi para obter um indicador de esparsidade, em [14], são apresentadas diferentes medidas para quantificar esta propriedade. Dentre todas as apresentadas, duas são avaliadas como sendo os melhores

indicadores: medida de Hoyer e índice de Gini.

A medida de Hoyer, como explicado em [14], é dada por

$$H_H = \left(\sqrt{N} - \frac{\sum_n c_n}{\sqrt{\sum_n c_n^2}} \right) \left(\sqrt{N} - 1 \right)^{-1}, \tag{3.14}$$

onde N é o número de coeficientes e c_n é o n-ésimo coeficiente. Adaptando a equação (3.14) para calcular a esparsidade dos coeficientes no domínio tempo-frequência temos

$$H_H \left[C_{i,l,p} \right] = \left(\sqrt{MK} - \frac{\sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \| C_{i,l,p} \|^2 [m,k]}{\sqrt{\sum_{m'=0}^{M-1} \sum_{k'=0}^{K-1} \| C_{i,l,p} \|^4 [m',k']}} \right) \left(\sqrt{MK} - 1 \right)^{-1}. \tag{3.15}$$

Ao analisar o cálculo da equação anterior, podemos observar que quanto maior for o valor na medida de Hoyer, menor será o espalhamento dos coeficientes. Assim, de acordo com essa medida, o conjunto de coeficientes ótimos pode ser obtido por

$$\bar{l} = \text{argmax}_l \left(H_H \left[C_{i,l,p} \right] \right). \tag{3.16}$$

Por fim, o índice de Gini é descrito por

$$H_G = 1 - 2 \sum_n \frac{c_n}{\sum_{n'} c_{n'}} \left(\frac{N - n + 1/2}{N} \right), \tag{3.17}$$

onde os coeficientes devem estar em ordem crescente, ou seja, $c_1 \leq c_2 \leq ... \leq c_N$. Para calcular esse índice em um conjunto de coeficientes bidimensionais, devemos organizá-lo como um vetor e ordená-lo. Assim, o cálculo adaptado para um conjunto de coeficientes bidimensionais é

$$H_G \left[C_{i,l,p} \right] = 1 - 2 \sum_{n=0}^{MK-1} \left(\frac{\| C_{i,l,p} \|^2 [n]}{\sum_{n'=0}^{MK-1} \| C_{i,l,p} \|^2 [n']} \left(\frac{N - n + 1/2}{N} \right) \right), \tag{3.18}$$

onde $C_{i,l,p}[n]$ é um elemento do vetor ordenado criado a partir do conjunto de coeficientes $C_{i,l,p}[m,k]$.

Dessa forma, assim como o caso anterior, ao analisar esta equação podemos observar que a esparsidade do conjunto de coeficientes é diretamente proporcional ao valor do índice de Gini. E, portanto, temos o conjunto de coeficientes ótimos

dado por

$$\bar{l} = \text{argmax}_l \left(H_G \left[C_{i,l,p} \right] \right). \tag{3.19}$$

Através dessa otimização pela esparsidade, o sinal será representado por um conjunto de regiões, onde cada região contém um grupo de coeficientes[5] com a resolução mais adequada.

3.5 Reconstrução do sinal

Ao fim deste algoritmo, encontramos o conjunto de coeficientes $C_{i,\bar{l},p}$ para cada sub-região temporal e para cada sub-região frequencial. Este conjunto, representado sobre a grade tempo-frequência $\Lambda_{\bar{l}}$, possui a distribuição considerada mais adequada de acordo com o critério de esparsidade de entropia de Rényi. Entretanto, na maioria dos casos, esses conjuntos de coeficientes não possuem a mesma dimensão, uma vez que cada região pode possuir uma resolução ótima distinta. Assim, as manipulações do sinal que requerem um processamento neste domínio devem ser realizadas de região em região.

Para obter o sinal no domínio temporal, devemos usar os procedimentos de reconstrução já mencionados nas seções anteriores. Cada conjunto de coeficientes $\widetilde{C_{i,\bar{l},p}}$ [6] deve ser ressintetizado através da função de síntese $h_{\bar{l}}[n]$ descrita na Seção 3.2. Assim, temos

$$\widetilde{x_{i,p}}[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \widetilde{c_{i,\bar{l},p}}[m,k] h_{\bar{l}}[n], \tag{3.20}$$

onde $h_{\bar{l}}[n]$ é a função de síntese referente à resolução ótima e $\widetilde{x_{i,p}}[n]$ é uma das parcelas do sinal reconstruído. Como as funções de janelamento e as funções-peso (a princípio o banco de filtros) satisfizeram as condições dadas pelas equações (3.2) e (3.9), respectivamente, a soma de cada uma dessas parcelas poderá compor o sinal reconstruído. Desse modo,

$$\widetilde{x}[n] = \sum_{i=0}^{I-1} \sum_{p=0}^{P-1} \widetilde{x_{i,p}}[n] = \sum_{i=0}^{I-1} \sum_{p=0}^{P-1} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \widetilde{c_{i,\bar{l},p}}[m,k] h_{\bar{l}}[n]. \tag{3.21}$$

Observamos que na construção dos coeficientes por regiões a etapa de repre-

[5] É importante ressaltar que apesar de o cálculo de esparsidade ocorrer a partir da potência do módulo dos coeficientes, as informações armazenadas sobre cada região são complexas, ou seja, possuem módulo e fase.

[6] O uso de $\widetilde{\ }$ os diferencia dos coeficientes $C_{i,\bar{l},p}$ porque estes geralmente sofrem modificações de acordo com a aplicação.

sentação no domínio tempo-frequência antecedeu a etapa de divisão na frequência. Contudo, na reconstrução do sinal, primeiro sintetizamos o sinal no domínio do tempo e, em seguida, juntamos as parcelas de diferentes zonas frequenciais (e temporais). Esta inversão de operações prejudica, principalmente, a reconstrução de sinais quando há coeficientes adjacentes no eixo frequencial com resoluções diferentes. Como veremos no Capítulo 4, esta ressíntese não é perfeita, mas em alguns casos o erro de reconstrução é imperceptível para algoritmos de avaliação objetiva de áudio como o *Perceptual Evaluation of Audio Quality* [32].

Capítulo 4

Testes e Avaliações Objetivas

Neste capítulo, iremos descrever os testes realizados com a representação tempo-frequencial com resolução otimizada descrita no Capítulo 3. Esses experimentos são essenciais para avaliarmos a eficácia do método de acordo com os parâmetros de configuração da técnica.

A fim de realizarmos essa avaliação, representamos o sinal original x no domínio tempo-frequencial através do método em questão e reconstruímos uma estimativa do sinal $\widetilde{x}$. Comparamos o sinal original e o sinal reconstruído calculando algumas medidas que exprimem o quão diferentes estão os sinais. Mais especificamente, determinam-se o erro de pico, dado por $e_{\max} = ||x - \widetilde{x}||_{\infty}$, e o valor eficaz normalizado, dado por

$$e_{\text{rms}} = \frac{||x - \widetilde{x}||_2}{||x||_2}, \tag{4.1}$$

onde $||\cdot||_R$ é a operação de norma R.

Para analisar os coeficientes através de uma imagem bidimensional como as figuras geradas no Capítulo 2, sugerimos a realização de uma interpolação dos coeficientes. Isso se faz necessário porque os grupos de coeficientes gerados para cada região do espaço tempo-frequencial podem ter resoluções diferentes. Dessa forma, interpolamos linearmente os coeficientes de cada região até que estes grupos possuam dimensões comparáveis.

Os arquivos referentes aos sinais de áudio utilizados para os testes, bem como os resultados obtidos neste capítulo, se encontram no site: `www.smt.ufrj.br/~gabriel.gouvea`.

4.1 Teste de divisão frequencial

Este primeiro teste, proposto em [1], tem como objetivo investigar os efeitos de múltiplos conjuntos de coeficientes, gerados por representações com características

distintas (STFT com comprimento de janelas diferentes), na presença da divisão frequencial. Para isso, criou-se um sinal senoidal de 2 s de duração em um formato dado por

$$x[n] = \text{sen}\left(2\pi\frac{f_0}{f_s}n + T_\Delta f_\Delta \text{sen}\left(\frac{2\pi}{T_\Delta f_s}n\right)\right), \tag{4.2}$$

cuja frequência inicial é $f_0 = 350$ Hz com modulação de $f_\Delta = 220$ Hz, ou seja, a frequência varia entre 130 Hz e 570 Hz, com período de $T_\Delta = 0{,}5$ s. O método de representação utilizado foi baseado na técnica explicada no Capítulo 3. Para simplificar esta análise, não foi realizada a etapa de divisão temporal do sinal descrita pelo método no Capítulo 3.

Representamos diretamente os sinais janelados com a STFT através da janela de Hann tradicional (sem a alteração descrita na Seção 3.1) com comprimento $N_1 = 512$ e $N_2 = 4096$ amostras. Em [1], não foram citadas mais informações sobre o procedimento de STFT para este teste e, portanto, escolheu-se o passo temporal $a_l = N_l/2$ (valor que permite a reconstrução perfeita do sinal, de acordo com [15]) e comprimento da DFT de cada janela $N_{Fl} = 2N_l$.

Como o foco é a divisão frequencial, em [1] foram criadas apenas duas regiões para separar o espaço tempo-frequencial: acima de 350 Hz e abaixo de 350 Hz. Assim, a separação frequencial ocorre exatamente no meio da modulação do sinal de teste. Para uma análise mais ampla, foram realizadas três simulações com conjuntos de funções-peso no formato dado por

$$\sigma_1[m,k] = \begin{cases} 0 & 0 \leq \Omega_k \leq \Omega_{c1}, \\ \frac{\Omega_k - \Omega_{c1}}{\Omega_{c2} - \Omega_{c1}} & \Omega_{c1} < \Omega_k < \Omega_{c2}, \\ 1 & \Omega_{c2} \leq \Omega_k \leq \pi, \end{cases} \tag{4.3}$$

$$\sigma_2[m,k] = 1 - \sigma_1[m,k], \tag{4.4}$$

onde Ω_k é a frequência digital referente à componente frequencial k expressa por $\Omega_k = 2k\pi/N_F$. A função-peso deve ser especificada como simétrica ao longo do eixo frequencial, porque os coeficientes da STFT apresentam essa mesma simetria. Os valores de Ω_{c1} e Ω_{c2} são frequências de corte digitais da função-peso referentes às frequências analógicas f_{c1} e f_{c2}, respectivamente. A relação entre essas duas escalas de frequências, já mencionada, no Capítulo 2 é dada por $\Omega_k = 2\pi f_k/f_s$.

A primeira simulação utiliza um conjunto de frequências de corte $f_{c1} = f_{c2} = 350$ Hz, traduzindo uma separação abrupta no eixo espectral do domínio tempo-frequencial. A segunda simulação foi realizada com $f_{c1} = 200$ Hz e $f_{c2} = 500$ Hz, proporcionando uma transição mais suave e permitindo sobreposição entre σ_1 e σ_2 nesse intervalo. A última simulação foi implementada com frequências de corte

$f_{c1} = 50$ Hz e $f_{c2} = 650$ Hz, configurando uma região de transição que englobaria toda a modulação do sinal de teste. Na Figura 4.1, podemos observar um exemplo de um par de funções-peso complementares.

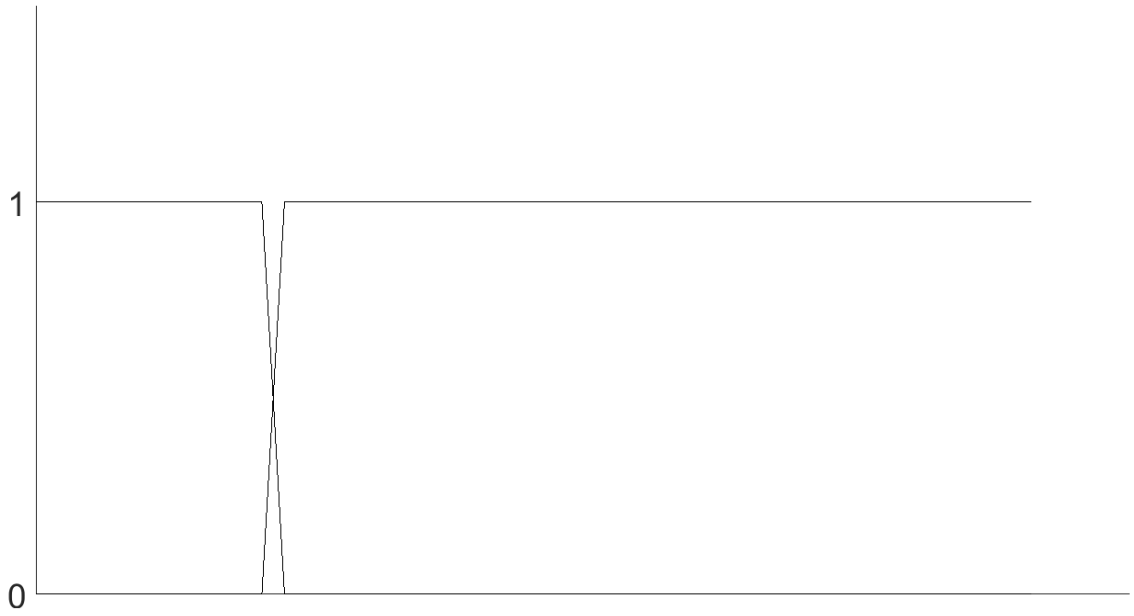

Figura 4.1: Exemplo de um par de funções-peso complementares.

Para cumprir com o objetivo do teste, atribuímos a cada uma das regiões uma representação diferente. Assim, para uma representação da região cujas componentes frequenciais são menores que 350 Hz utilizamos a STFT com janelas de tamanho 4096, e para a região complementar utilizamos a STFT de tamanho 512. Essa escolha é justificada pela necessidade de maior resolução em baixas frequências do que nas altas frequências.

Por fim, com os coeficientes de cada região definidos, reconstruímos o sinal e calculamos o erro que este procedimento provoca.

4.1.1 Análises de resultados

A Tabela 4.1 contém os valores dos erros entre o sinal original e o sinal reconstruído retirados do artigo [1] e da reprodução do trabalho na presente dissertação.

Analisando a tabela, observamos que independentemente das frequências de corte escolhidas para a divisão frequencial, os erros são altos (considerando que o sinal possui uma amplitude unitária). Assim, podemos afirmar que a reconstrução desse processo com esses parâmetros não é perfeita. Além disso, constatamos que quanto maior for o intervalo entre f_{c1} e f_{c2}, menor será a diferença entre o sinal original e o sinal reconstruído.

Parâmetros	Artigo		Dissertação	
	$e_{\max}$	e_{rms}	$e_{\max}$	e_{rms}
$f_{c1} = 350$ Hz $f_{c2} = 350$ Hz	0,5102	0,0967	0,4986	0,1166
$f_{c1} = 200$ Hz $f_{c2} = 500$ Hz	0,1856	0,0725	0,0770	0,0245
$f_{c1} = 50$ Hz $f_{c2} = 650$ Hz	0,0576	0,0262	0,0331	0,0072

Tabela 4.1: Tabela de erros do teste de divisão frequencial retirados do artigo [1] e obtidos nas simulações desta dissertação.

Notamos que há uma discrepância nos valores dos erros do artigo [1] e o teste realizado por esta dissertação, certamente porque não foram informados no artigo todos os parâmetros da STFT: com exceção do e_{rms} do primeiro conjunto de frequências de cortes, os valores de erros do artigo se mostraram maiores do que os valores de erro reproduzidos nesta dissertação (e portanto, podem ser considerados sem polarização).

As Figuras 4.2, 4.3 e 4.4 contêm o módulo (em dB) dos coeficientes da representação tempo-frequencial do sinal gerado para cada uma das simulações em um plano bidimensional. O eixo das abscissas contém os atrasos temporais e localiza cada coeficiente no tempo, enquanto o eixo das ordenadas informa a qual componente espectral cada um se refere. As regiões inferiores do gráfico dizem respeito aos coeficientes gerados pela STFT com $N_2 = 4096$, enquanto as regiões superiores representam os coeficientes produzidos pela STFT com $N_1 = 512$.

Observamos que os coeficientes das regiões inferiores às frequências de corte f_{c1} de cada figura possuem coeficientes menores (regiões mais claras) do que as regiões superiores às frequências de corte f_{c2}. Podemos inferir, portanto, que a escolha de janelas com tamanho $N_2 = 4096$ gerou coeficientes mais esparsos para esta faixa espectral analisada.

4.2 Teste de sinal de áudio

O objetivo deste teste é analisar o comportamento de escolha de coeficientes em um sinal real de áudio através da entropia de Rényi. Consideramos um sinal contendo uma amostra de som de um instrumento de percussão indiano, a tabla, e, em 2,22 s um instrumento de corda indiano, o sitar, é adicionado.

Para representar o sinal no domínio tempo-frequencial, dividimos o sinal através de janelas de tamanho $N_i = 6144$ amostras com passos de $a_i = 1024$ amostras. Em [1], utilizam-se janelas retangulares tradicionais, o que, como já discutido no Capítulo 3, não permite reconstrução perfeita nessas configurações. Como o processo

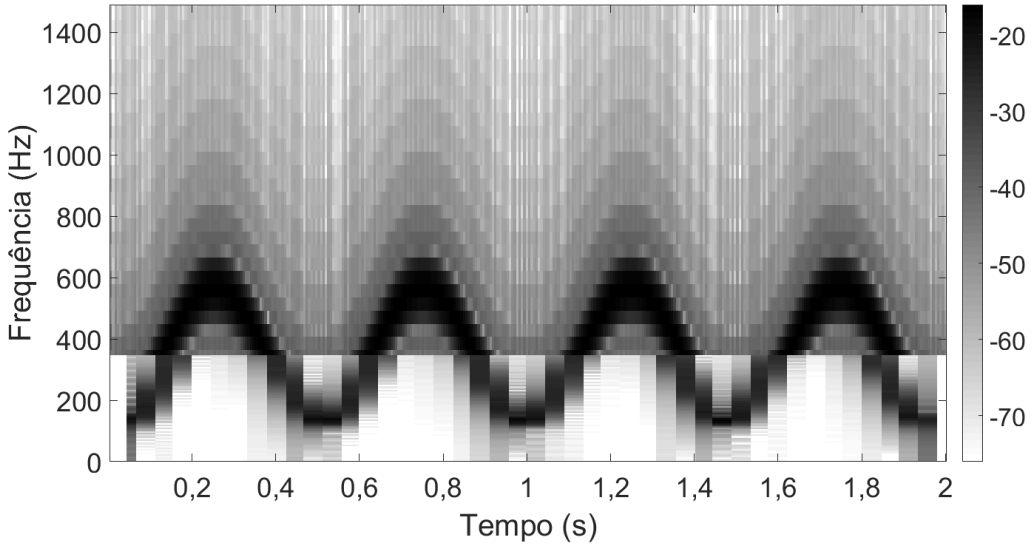

Figura 4.2: Módulo do sinal de senoide modulada no domínio tempo-frequencial de resolução variável com divisão espectral nas frequências de corte $f_{c1} = f_{c2} = 350$ Hz.

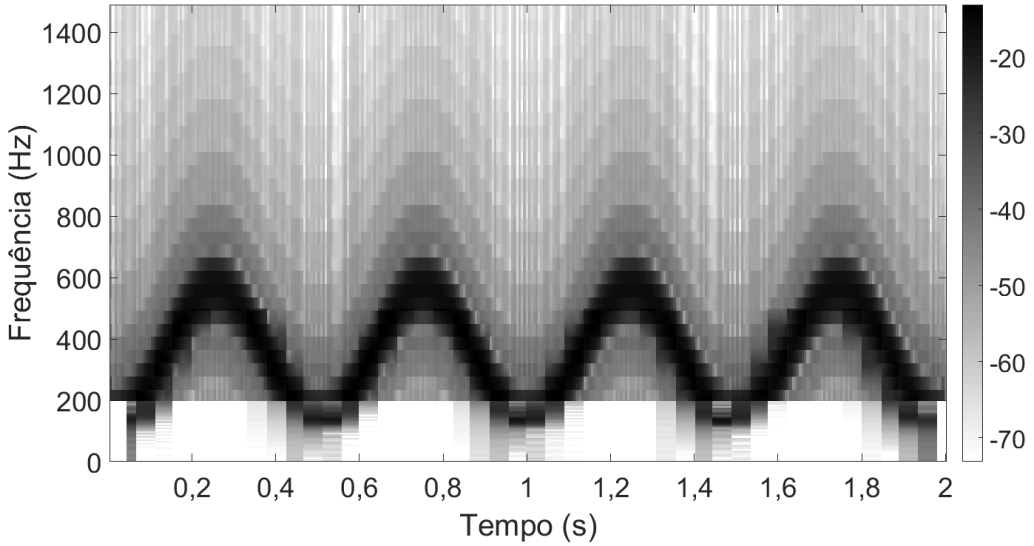

Figura 4.3: Módulo do sinal de senoide modulada no domínio tempo-frequencial de resolução variável com divisão espectral nas frequências de corte $f_{c1} = 200$ Hz e $f_{c2} = 500$ Hz.

de síntese deste teste não foi especificado pelo autor do artigo, escolhemos modificar a janela através do algoritmo descrito no Capítulo 3 e não alteramos os parâmetros da mesma.

Cada região do sinal é então convertida para o domínio tempo-frequencial pela STFT com janelas de Hann. Neste caso, possibilitamos $L = 8$ diferentes comprimentos N_l de janelas para a representação: 1024, 1248, 1522, 1854, 2262, 2756, 3360, 4096. Além disso, o passo temporal é dado por $a_l = 0{,}15N_l$, e o comprimento da DFT de cada janela é $N_{Fl} = 0{,}15N_l$.

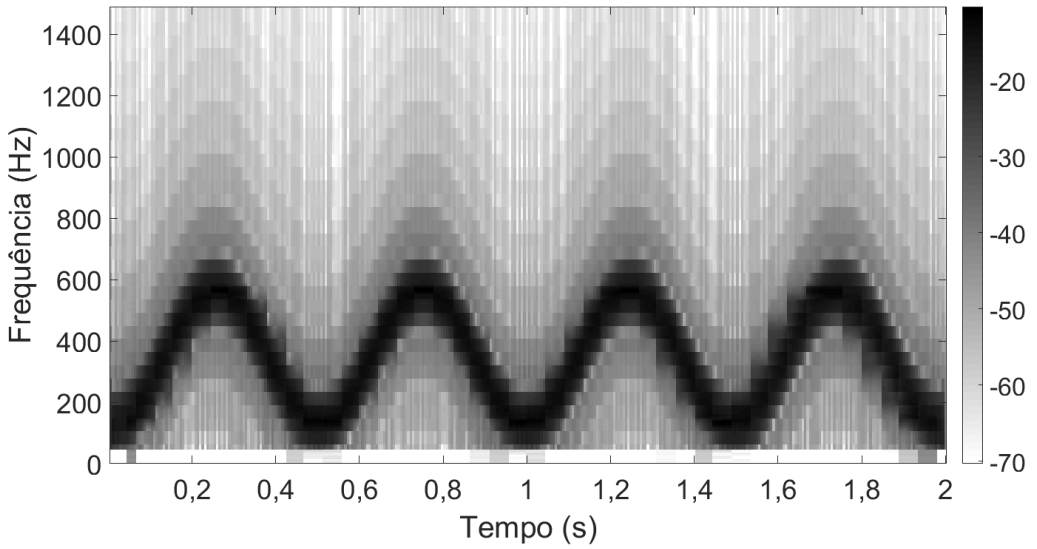

Figura 4.4: Módulo do sinal de senoide modulada no domínio tempo-frequencial de resolução variável com divisão espectral nas frequências de corte $f_{c1} = 50$ Hz e $f_{c2} = 650$ Hz.

Para a divisão frequencial do espaço tempo-frequencial utilizamos o mesmo formato de função-peso descrito na equação (4.3). Da mesma forma, realizamos simulações com diferentes frequências de corte para uma análise mais detalhada: $f_{c1} = f_{c2} = 1$ kHz, $f_{c1} = 750$ Hz e $f_{c2} = 1{,}25$ kHz, e $f_{c1} = 500$ Hz e $f_{c2} = 1{,}5$ kHz. Assim temos duas regiões frequenciais a serem analisadas pela entropia de Rényi para cada trecho do sinal.

De acordo com [1], calcula-se o valor da entropia com $\alpha = 0{,}3$ apenas para a parcela cujas componentes frequenciais são inferiores a 1 kHz, onde os primeiros harmônicos dos instrumentos são predominantes, isto é, ignoramos, para cada trecho do sinal, a parcela superior a 1 kHz. Em seguida, para cada intervalo temporal do sinal, selecionamos a janela cuja entropia foi eleita a menor. Por fim, reconstruímos o sinal a partir dos coeficientes escolhidos e calculamos os erros em relação ao sinal original.

4.2.1 Análise de resultados

Na Tabela 4.2, apresentamos os valores de erro citados no artigo [1] e, para realizarmos uma comparação, exibimos os valores dos erros referentes à reprodução do teste por nós simulado, como explicado no início desta seção.

Podemos observar que os testes realizados pelo artigo [1] possuem um erro muito superior aos erros calculados pelos testes deste trabalho. Como não foi informado o método de reconstrução com o uso da janela retangular tradicional, assumimos que, como o esperado de acordo com [15], esse valor de sobreposição da etapa de janelamento não permite reconstrução perfeita.

Parâmetros	Artigo		Dissertação	
	$e_{\max}$	e_{rms}	$e_{\max}$	e_{rms}
$f_{c1} = 1$ kHz $f_{c2} = 1$ kHz	$4{,}7 \times 10^{-3}$	$4{,}4 \times 10^{-3}$	$8{,}3267 \times 10^{-17}$	$1{,}6221 \times 10^{-16}$
$f_{c1} = 750$ Hz $f_{c2} = 1{,}25$ kHz	$3{,}4 \times 10^{-3}$	$2{,}6 \times 10^{-3}$	$8{,}3267 \times 10^{-17}$	$1{,}5287 \times 10^{-16}$
$f_{c1} = 500$ Hz $f_{c2} = 1{,}5$ kHz	$3{,}7 \times 10^{-3}$	$2{,}2 \times 10^{-3}$	$8{,}3267 \times 10^{-17}$	$1{,}5353 \times 10^{-16}$

Tabela 4.2: Tabela de erros do teste com sinal de áudio retirados do artigo [1] e obtidos nas simulações desta dissertação.

Os pequenos erros obtidos pelos testes deste trabalho podem ser atribuídos aos erros de arredondamento durante o processamento dos coeficientes e da ressíntese do sinal. Portanto, podemos assumir que tal procedimento para esse teste permite a reconstrução perfeita do sinal.

Na Figura 4.5, é apresentado o módulo (em dB) dos coeficientes de simulação com frequências de corte $f_{c1} = f_{c2} = 1$ kHz em formato de imagem. Podemos observar que há uma presença maior de linhas verticais antes de 2,22 s, configurando a existência de sons que são bem definidos no tempo e preenchem o espectro nesse instante. Essas linhas, portanto, indicam a presença do instrumento musical tabla no sinal de teste. Similarmente, notamos uma maior recorrência de linhas horizontais após 2,22 s, configurando sons bem definidos na frequência, como os emitidos pelo sitar.

Na Figura 4.6, exibimos a escolha do comprimento de janelas usadas para representar cada intervalo do sinal no domínio tempo-frequencial. Comparando as duas Figuras 4.5 e 4.6, podemos notar que, na presença exclusiva da tabla, há uma escolha de janelas predominantemente de comprimento $N_l = 1024$. E, após o sitar ser introduzido, a seleção de janelas prioriza comprimentos $N_l = 4096$. Isso já era esperado, uma vez que os sons percussivos da tabla necessitam de uma melhor resolução temporal comparadamente aos sons tonais do sitar, os quais requerem uma resolução espectral superior.

4.3 Teste com parâmetros variados de sinais de áudio

O foco deste teste é avaliar mais amplamente o método proposto através de variações nos parâmetros e simulações sobre um grupo de sinais com características distintas. Este grupo de sinais contém 6 sinais sintéticos, ou seja, criados artificialmente, tais como o sinal modulado do teste descrito na Seção 4.1. Além disso, outros

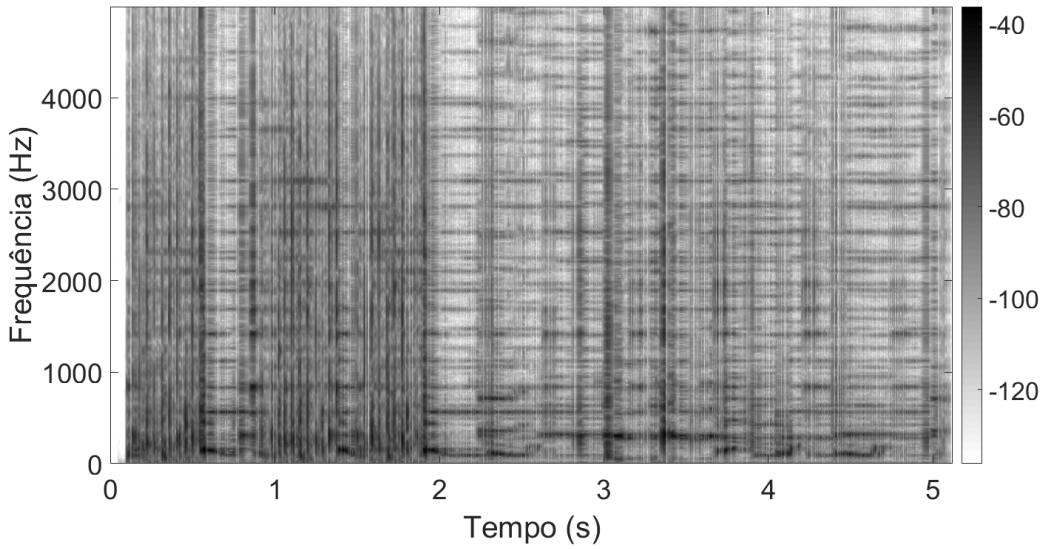

Figura 4.5: Módulo do sinal de teste no domínio do tempo-frequencial de resolução variável com frequências de corte $f_{c1} = f_{c2} = 1$ kHz.

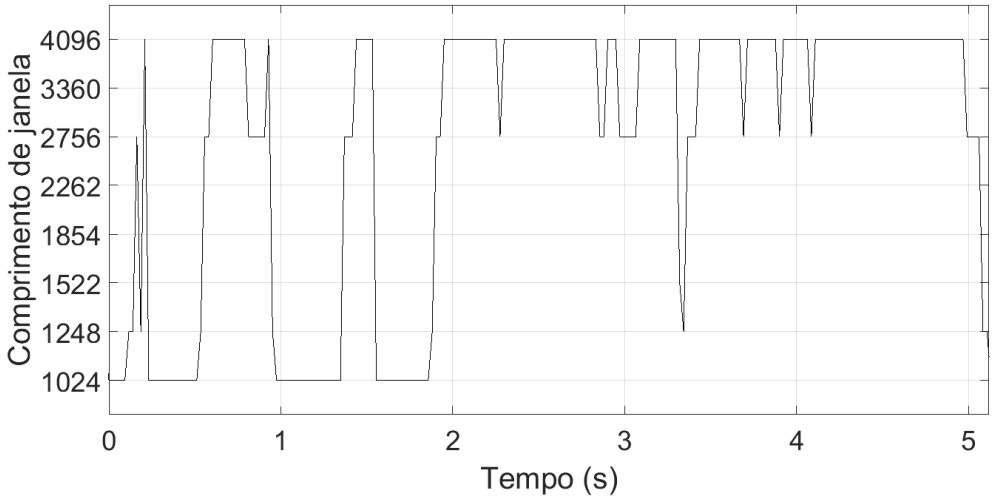

Figura 4.6: Escolha de comprimentos de janelas através do cálculo da entropia de Rényi para o teste com sinal de áudio com frequências de corte $f_{c1} = f_{c2} = 1$ kHz.

11 exemplos de áudio[7] produzidos por instrumentos musicais reais serão utilizados para avaliar este método.

Além dos indicadores de erro já mencionados anteriormente, utilizamos outras medidas para avaliar os resultados da reconstrução através desse método. Uma forma de comparar o sinal reconstruído e o sinal original é através de um algoritmo que avalia a qualidade perceptiva do áudio, como o PEAQ apresentado em [32]. Este algoritmo quantificará as diferenças entre os sinais comparados através de uma nota

[7]Estes sinais integram o banco de sinais utilizados por [1] e se encontram no site `http://recherche.ircam.fr/equipes/analyse-synthese/liuni/TESTS/sounds/`. Acessado em 15 de Março de 2016.

denominada *Objetive Difference Grade* (ODG), cujos valores convencionais variam de 0 (sinais perfeitamente iguais) a -4 (sinais extremamente diferentes)[8].

Um outro indicador que utilizamos foi a contagem de coeficientes não-nulos do sinal no domínio tempo-frequencial (após a atribuição do valor zero a coeficientes desprezivelmente baixos segundo algum critério). Apesar de não estar relacionado com a entropia de Rényi, esse valor é uma forma de avaliar quão esparsa é a representação gerada pelo método (sua eficácia, portanto). Além disso, este indicador avalia a capacidade do método sob análise ser utilizável em compressão de sinais.

Por fim, calculamos também a duração de todo o procedimento (englobando a etapa de análise e a etapa de reconstrução do sinal) para cada configuração de parâmetros e sinal. Através desse indicador quantizamos a eficiência do método proposto.

O teste consiste em representar cada sinal no domínio tempo-frequencial através do método proposto e ressintetizar o sinal. Como realizamos várias simulações com diferentes combinações de parâmetros para cada sinal, detalharemos, a seguir, o procedimento implementado, apontando esses parâmetros variáveis, bem como as opções definidas para eles.

Para a divisão no eixo temporal, segmentamos cada sinal através de um formato de função-janela. Considerando este formato um parâmetro que influencia nos resultados, optamos por variá-lo com três tipos diferentes de janelas: a retangular tradicional, a retangular adaptada e a senoidal adaptada de acordo com a Seção 3.1. O comprimento da janela será fixado em $N_i = 6144$. O passo temporal a_i entre as janelas foi um outro parâmetro que decidimos variar, disponibilizando três opções: 4608, 3072 e 1024.

Em seguida, representamos o sinal no domínio tempo-frequencial, pela STFT com a função-janela senoidal adaptada com diferentes comprimentos N_l de janelas: 1024, 1248, 1522, 1854, 2262, 2756, 3360, 4096. O passo temporal da STFT a_l e o tamanho da DFT N_{Fl} também foram parâmetros que preferimos variar para investigar seus efeitos no resultado final. Trabalhamos com valores de $a_l = 0{,}15N_l$, $0{,}25N_l$ e $0{,}5N_l$, e valores de $N_{Fl} = N_l$, $2N_l$ e $4N_l$.

A divisão frequencial foi realizada sobre os coeficientes do sinal no domínio tempo-frequencial e as funções-peso foram criadas no formato de trapézios (uma generalização das funções desenvolvidas em (4.3)) que dividem o eixo espectral de acordo com as frequências de corte dadas por $f_c = \{21{,}5; 43{,}1; 86{,}1172{,}3; 344{,}5; 689; 1378{,}1; 2756{,}3; 5512{,}5; 11025; 22050\}$ em Hertz. Visto que a frequência de amostragem dos sinais é de 44,1 kHz, escolhe-

[8] A escala prática do PEAQ é ligeiramente diferente, por razões de projeto: sinais absolutamente iguais resultam em valor ligeiramente acima de zero, e a escala satura inferiormente um pouco acima de -4.

mos frequências de corte divisoras deste número em potências de dois. Esta escolha foi inspirada no intervalo de oitavas utilizado na separação das notas da escala cromática.

Cada função possui uma pequena região de sobreposição com cada função adjacentes. O tamanho dessa sobreposição é proporcional à frequência de corte e, portanto, essas funções-peso se comportam como trapézios escalenos ao longo do eixo frequencial. Na Figura 4.7, apresentamos um exemplo desse conjunto de funções-peso. O valor do tamanho de sobreposição foi um parâmetro considerado influente no resultado e, por isso, foi variado entre $0f_c$ (ou sem sobreposição), $0{,}1f_c$ e $0{,}25f_c$.

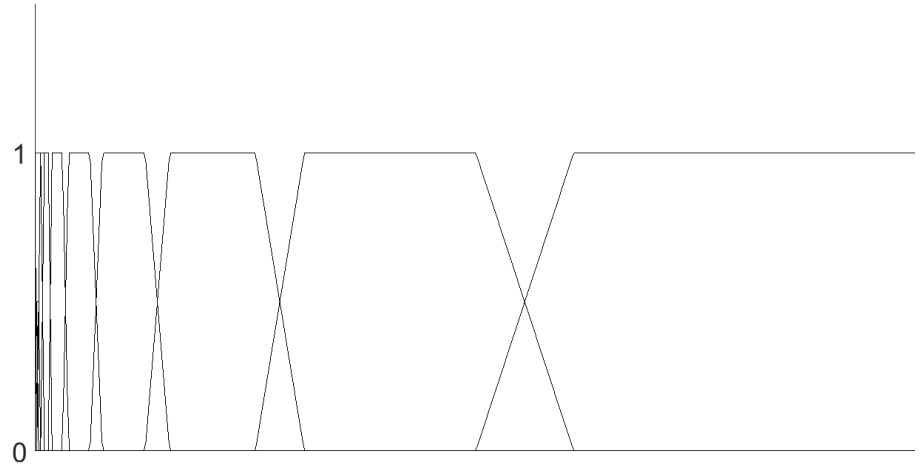

Figura 4.7: Exemplo do conjunto de funções-peso utilizadas nas simulações.

Em seguida, selecionamos os melhores coeficientes de acordo com o critério de esparsidade utilizando entropia com $\alpha = 0{,}3$. Com os melhores coeficientes para cada região, obtemos o sinal representado no domínio tempo-frequencial com a resolução otimizada para os parâmetros em questão.

Uma etapa adicional inserida para analisar o método é a substituição por zeros de coeficientes que estejam abaixo de um limiar. Para isso, descobrimos o valor máximo dentre todos os coeficientes da representação e adotamos como limiar uma fração, denominado *cutoff*, desse valor máximo. Como esta etapa é responsável por eliminar os coeficientes, consideramos o *cutoff* como um parâmetro importante no resultado e, portanto, simulamos com três opções: 90, 50 e 10 dB abaixo do valor máximo. É importante ressaltar que o indicador de eficácia do método, ou seja, a contagem de coeficientes não-nulos é realizada após essa etapa de substituição.

Após a eliminação desses coeficientes, reconstruímos o sinal como descrito no Capítulo 3.

4.3.1 Análise de resultados

O teste foi realizado em 17 sinais, com 6 parâmetros variando, cada um deles em 3 opções diferentes, configurando um total de 12393 simulações. Para avaliar esse teste, criamos um conjunto de resultados ótimos selecionando apenas as simulações cujo ODG seja maior do que -1. Em seguida, contabilizamos a frequência de resultados dentro deste conjunto de cada opção e de cada parâmetro. Nas Figuras 4.8, 4.9, 4.10, 4.11, 4.12 e 4.13, podemos observar os histogramas usados para ilustrar esse conjunto gerado.

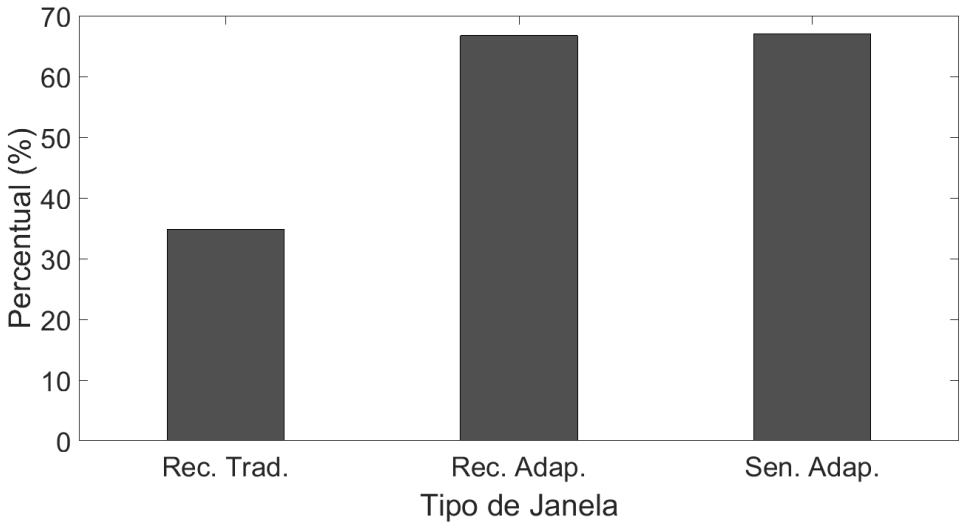

Figura 4.8: Histograma do parâmetro de tipo de janela com resultados acima de -1 de ODG.

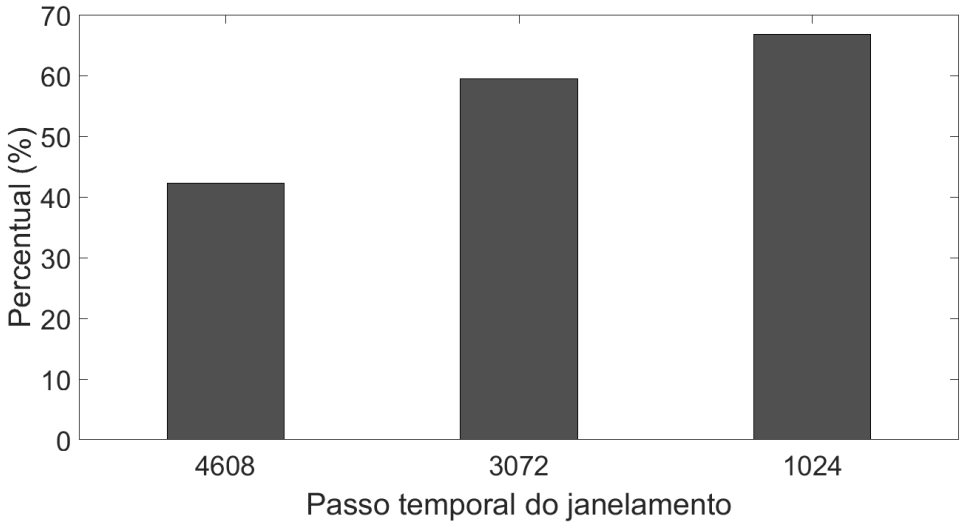

Figura 4.9: Histograma do parâmetro de passo temporal de janelamento com resultados acima de -1 de ODG.

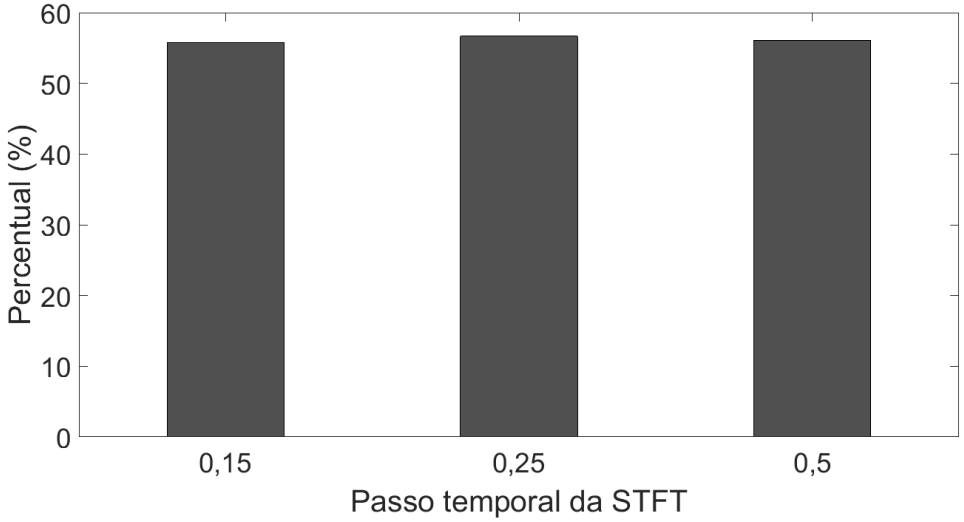

Figura 4.10: Histograma do parâmetro de passo temporal da STFT com resultados acima de -1 de ODG.

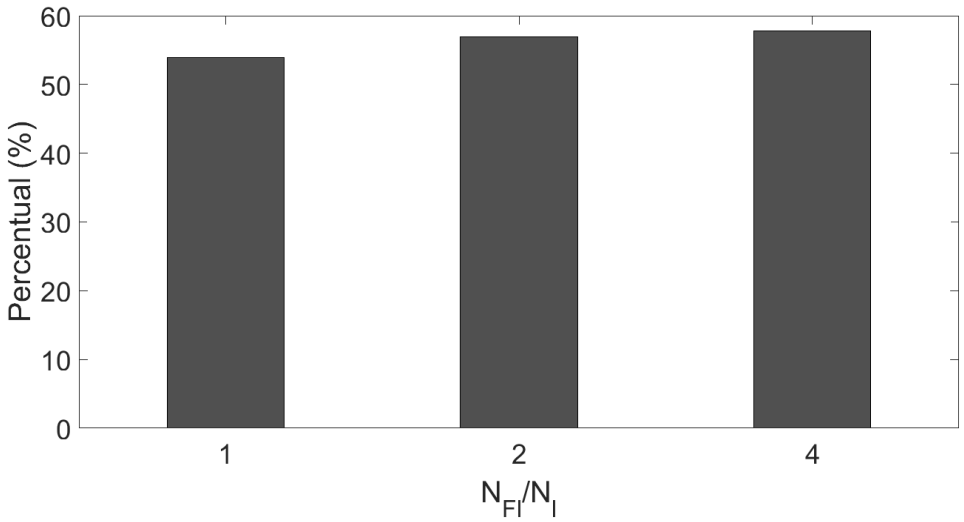

Figura 4.11: Histograma do parâmetro N_{Fl}/N_l com resultados acima de -1 de ODG.

Cada histograma está relacionado a um parâmetro, e as barras indicam a razão entre a frequência da opção contida no conjunto de resultados ótimos e o número total de testes realizados com esta opção. Assim, quanto maior for a altura da barra maior é a presença desta opção dentro do conjunto de resultados ótimos. Após uma análise do histograma da Figura 4.13, observou-se que a opção de 10 dB do parâmetro *cutoff* possui pouca representatividade, com menos de 10% presentes no conjunto de resultados ótimos. Além disso, comparando os valores das barras no histograma da Figura 4.8, notamos que a opção de janela retangular tradicional possui também pouca participação no conjunto ótimo.

Como essas opções prejudicam a avaliação global do desempenho, também irão prejudicar a estatística aferida a partir dos histogramas. Assim, decidimos eliminar

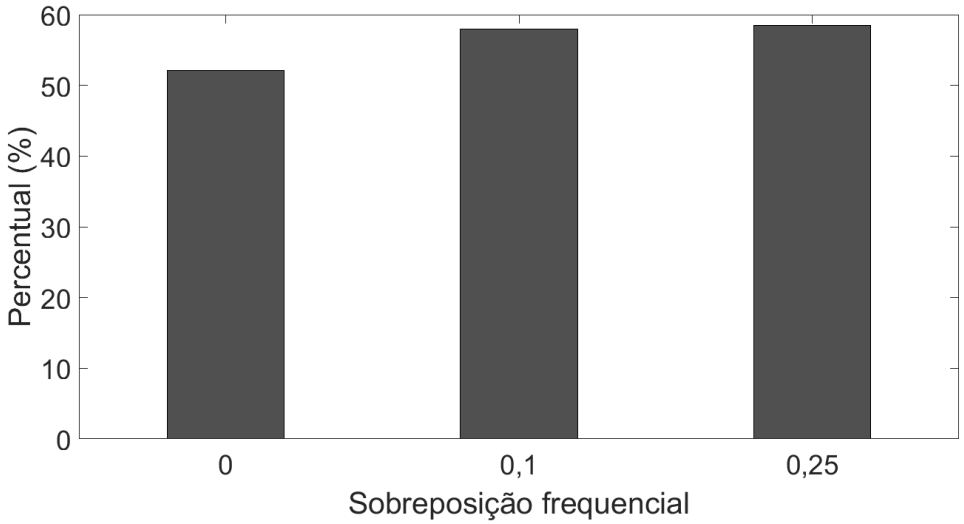

Figura 4.12: Histograma do parâmetro de sobreposicão frequencial com resultados acima de -1 de ODG.

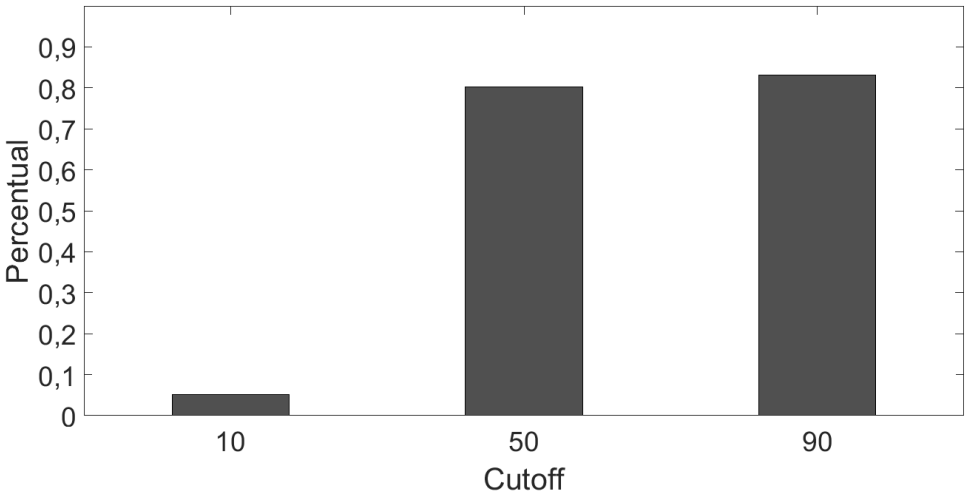

Figura 4.13: Histograma do parâmetro de *cutoff* (em dB) com resultados acima de -1 de ODG.

todos os resultados que utilizavam essas duas opções e, em seguida, realizamos novamente o procedimento de construção desses histogramas. Nas Figuras 4.14, 4.15, 4.16, 4.17, 4.18 e 4.19, podemos observar novos gráficos de barras com o percentual de presença dos parâmetros após esta eliminação.

Um outro modo de analisar esses resultados é através do cálculo do coeficiente de correlação de Pearson [33]. O cálculo deste coeficiente entre duas variáveis aleatórias X e Y é dado por

$$\rho_{X,Y} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{4.5}$$

onde μ_X e μ_Y são, respectivamente, as médias das variáveis aleatórias X e Y, e

Figura 4.14: Histograma do parâmetro de tipo de janela com resultados acima de -1 de ODG após eliminação das opções 10dB e janela retangular tradiconal.

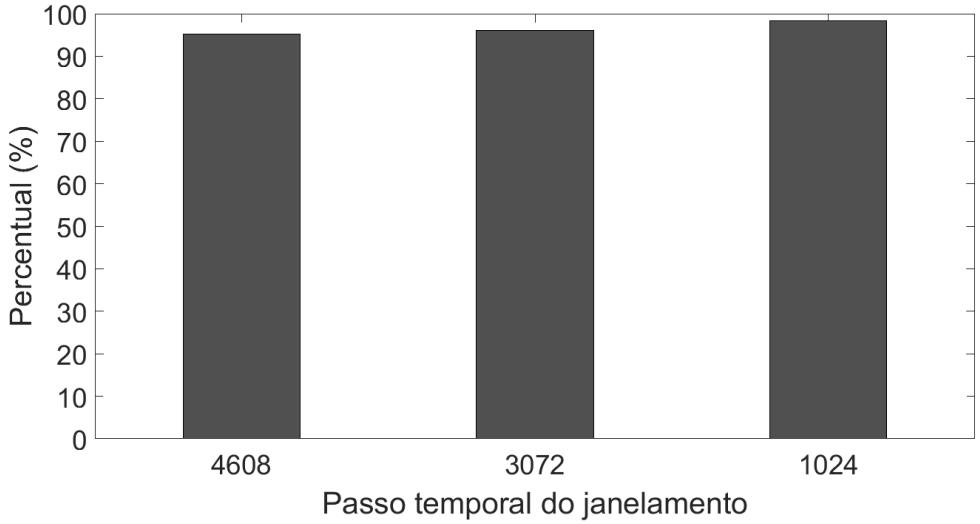

Figura 4.15: Histograma do parâmetro de passo temporal de janelamento com resultados acima de -1 de ODG após eliminação das opções 10dB e janela retangular tradicional.

σ_X e σ_Y são seus respectivos desvios padrões. Como estamos trabalhando com um conjunto de resultados com muitos elementos, este coeficiente pode ser calculado por

$$\rho_{X,Y} = \frac{1}{IJ} \sum_{I} \sum_{J} \frac{(x_i - \bar{X})}{s_X} \frac{(y_i - \bar{Y})}{s_Y}, \tag{4.6}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias amostrais calculadas por

$$\bar{X} = \frac{1}{I} \sum_{I} x_i, \tag{4.7}$$

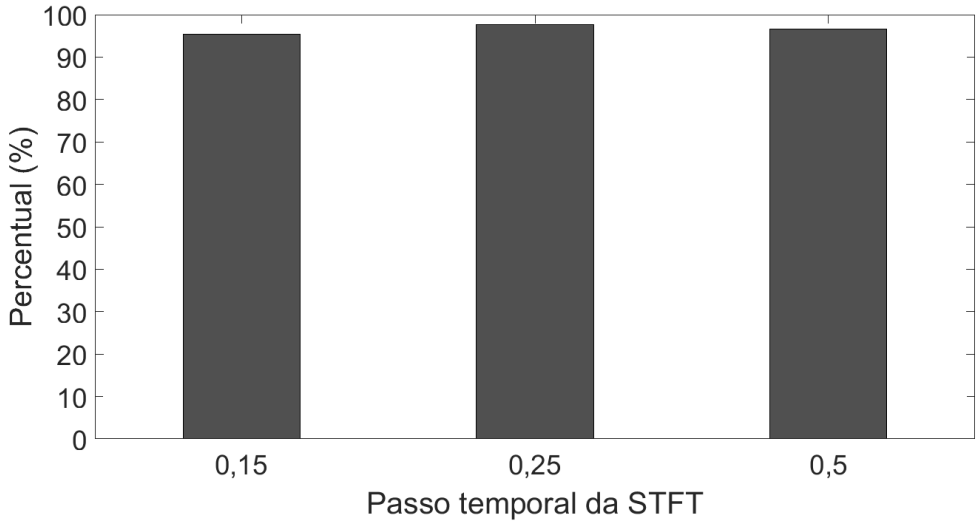

Figura 4.16: Histograma do parâmetro de passo temporal da STFT com resultados acima de -1 de ODG após eliminação das opções 10dB e janela retangular tradicional.

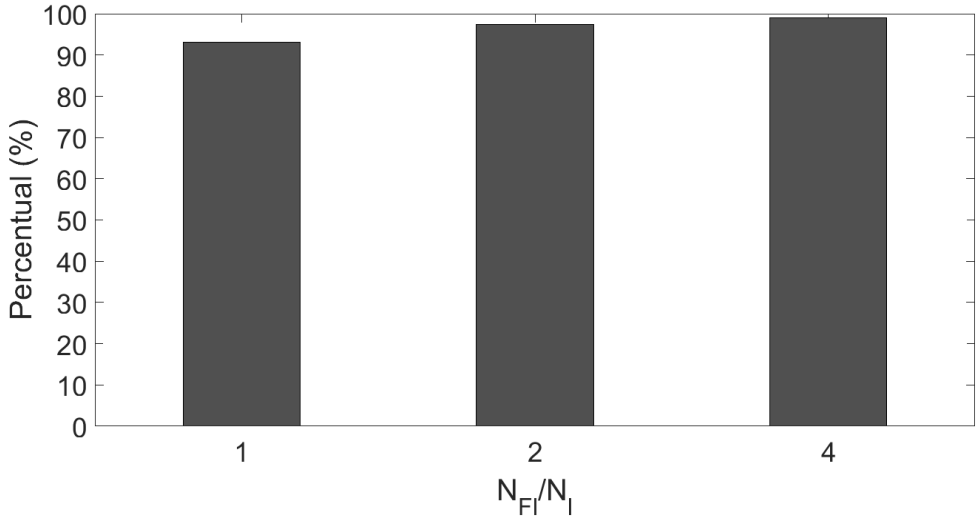

Figura 4.17: Histograma do parâmetro N_{Fl}/N_l com resultados acima de -1 de ODG após eliminação das opções 10dB e janela retangular tradicional.

e s_X e s_Y são os desvios padrões amostrais calculados por

$$s_X = \sqrt{\frac{1}{I}\sum_{I}(x_i - \bar{X})^2}. \tag{4.8}$$

O cálculo desse coeficiente de correlação informa a relação entre dois parâmetros quantificando-o entre -1 e 1. Caso esse valor esteja próximo de 0, isto indica que esses dois parâmetros são poucos correlacionados, e no caso de o módulo desse valor estar próximo de 1 podemos afirmar que esses parâmetros estão muito correlacionados. Com isso, montamos a Tabela 4.3 com os coeficientes de correlação de Pearson

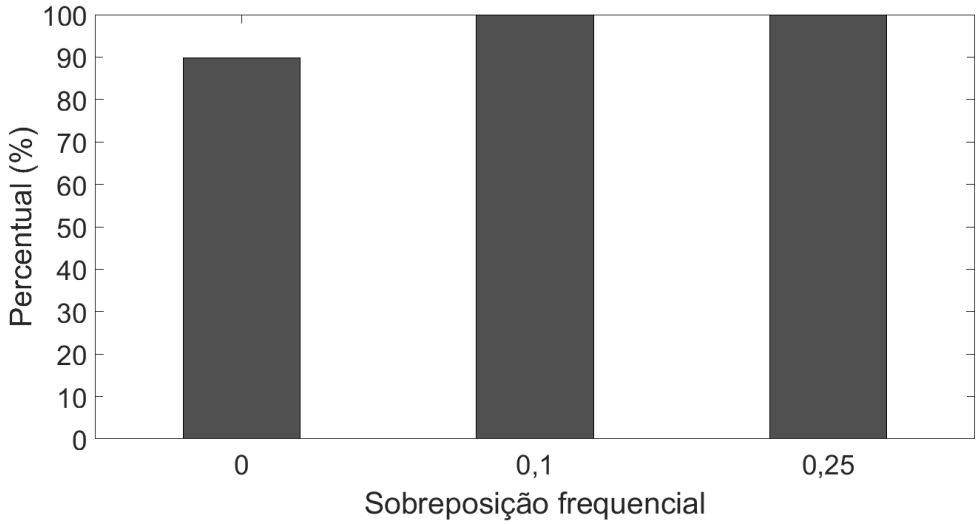

Figura 4.18: Histograma do parâmetro de sobreposição frequencial com resultados acima de -1 de ODG após eliminação das opções 10dB e janela retangular tradicional.

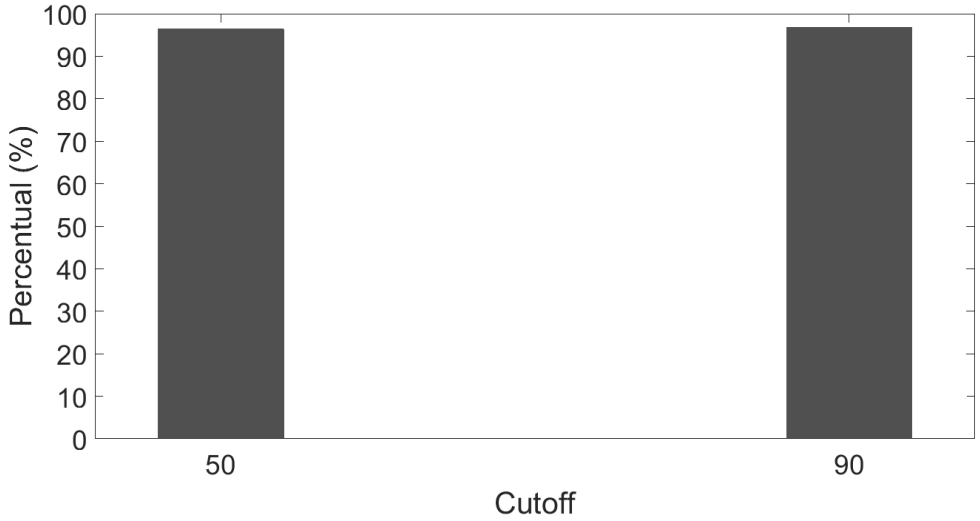

Figura 4.19: Histograma do parâmetro de *cutoff* (em dB) com resultados acima de -1 de ODG após eliminação das opções 10dB e janela retangular tradicional.

entre os parâmetros variáveis[9] e os indicadores de eficácia e eficiência.

De acordo com a Tabela 4.3, os parâmetros tipo de janela, passo temporal de janelamento e passo temporal da STFT são pouco correlacionados com o $e_{\max}$, o e_{rms} e o ODG. Portanto, selecionamos as opções dentre esses parâmetros que menos prejudicassem o tempo de processamento do método: janela senoidal adaptada, $a_i = 4608$ e $a_l = 0{,}5N_l$.

Com relação aos parâmetros de N_{Fl}/N_l e sobreposição frequencial, nada pode se inferir a partir das duas análises discutidas nesta subseção.

[9]Para calcular este coeficiente de correlação com o parâmetro tipo de janela, designamos valores numéricos para as opções dele.

Correlação	$e_{\max}$	e_{rms}	ODG	Coeficientes não nulos	Tempo
Tipo de janelas	0,00472	0,00121	−0,00538	−0,004562	−0,00959
Passo de janelamento	0,03364	0,01952	−0,07139	−0,41022	−0,44700
Passo de STFT	0,01942	−0,00498	0,00803	−0,29264	−0,27501
N_{Fl}/N_l	−0,24087	−0,24412	0,21622	0,34837	0,29958
Sobreposição frequencial	−0,51091	−0,46842	0,43751	0,14290	0,01836
Cutoff	−0,00665	−0,00233	0,07084	0,06118	0,00311

Tabela 4.3: Tabela com coeficientes de correlação de Pearson entre os parâmetros variáveis e os indicadores do teste com parâmetros variáveis.

4.4 Teste com esparsidade de sinais de áudio

A realização do teste anterior contava com a variação de seis parâmetros diferentes. A partir da análise dos resultados deste teste, concluímos que os parâmetros de tipo de janela, passo temporal do janelamento e passo temporal da STFT não influem significativamente na reconstrução dos sinais. Contudo, como o parâmetro *cutoff* está diretamente relacionado com a esparsidade da representação tempo-frequência, uma maior variação desse parâmetro é necessária.

Para investigar melhor esta técnica, realizamos um novo teste fixando três parâmetros (janela senoidal adaptada, $a_i = 4608$ e $a_l = 0{,}5N_l$), mantendo dois parâmetros variantes (N_{Fl}/N_l e a sobreposição frequencial) e refinando o *cutoff* com mais opções (10, 20, 30, 40, 50, 60, 70, 80 e 90 dB) para esta análise. Além disso, o teste dispôs de três medidas diferentes de esparsidade: entropia de Rényi, medida de Hoyer e índice de Gini. Ainda assim, o valor do parâmetro α na entropia de Rényi se manteve igual a 0,3.

Novamente, representamos cada um dos 17 sinais no domínio tempo-frequência através da técnica, eliminamos os coeficientes abaixo do limiar do *cutoff* e reconstruímos o sinal.

4.4.1 Análise de resultados

Com essa variação de parâmetros foram realizadas 243 simulações para cada sinal. Após se obterem esses resultados, foram realizadas as mesmas análises aplicadas aos resultados obtidos na Sub-seção 4.3.1.

A primeira etapa foi analisar os resultados com base no limiar de −1 de ODG, ou

seja, eliminar os resultados com ODG menor que -1. Em seguida, foram construídos os histogramas por parâmetro em que cada barra representa a razão de frequência da opção do parâmetro e o número total de resultados com esta opção de parâmetro. Nas Figuras 4.20, 4.21, 4.22 e 4.23, são mostrados histogramas de cada parâmetro cujos resultados possuem notas acima de -1 ODG.

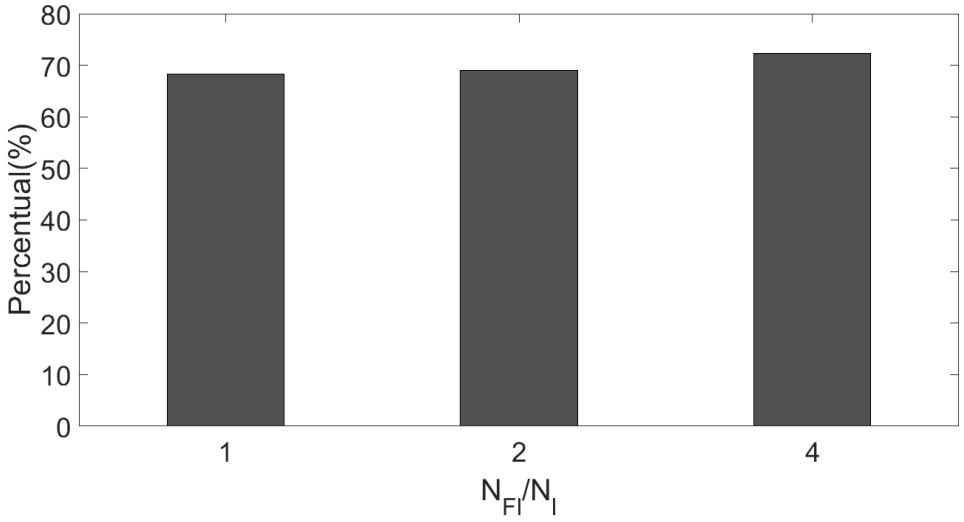

Figura 4.20: Histograma do parâmetro N_{Fl}/N_l com resultados acima de -1 de ODG gerado pelo teste com esparsidade.

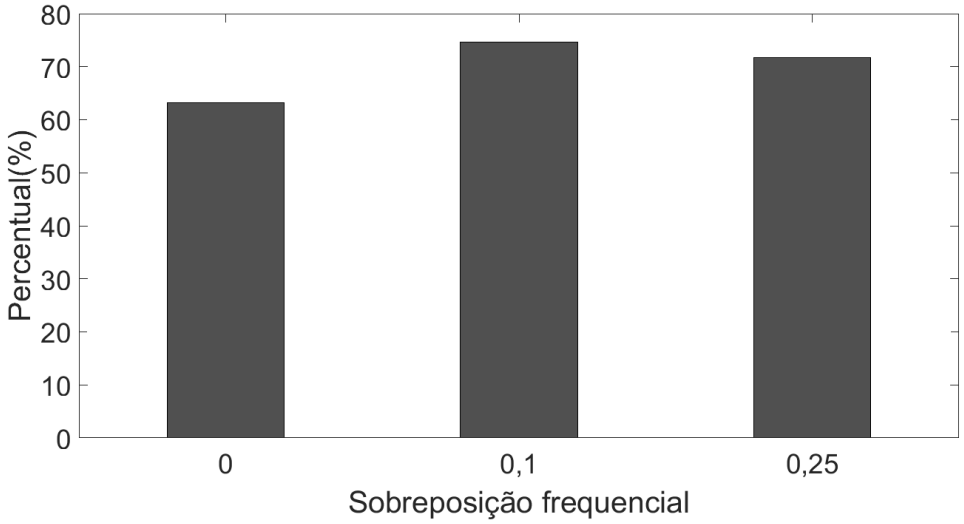

Figura 4.21: Histograma do parâmetro de sobreposição frequencial com resultados acima de -1 de ODG gerado pelo teste com esparsidade.

Apesar de os histogramas dos parâmetros de N_{Fl}/N_l e da técnica de esparsidade não fornecerem nenhuma informação, o *cutoff* indica mais claramente qual o valor limite para o qual os resultados passam a ser considerados ruins, ou seja, com muitos resultados abaixo de -1 de ODG. Os valores de 10, 20 e 30 dB, por possuírem uma

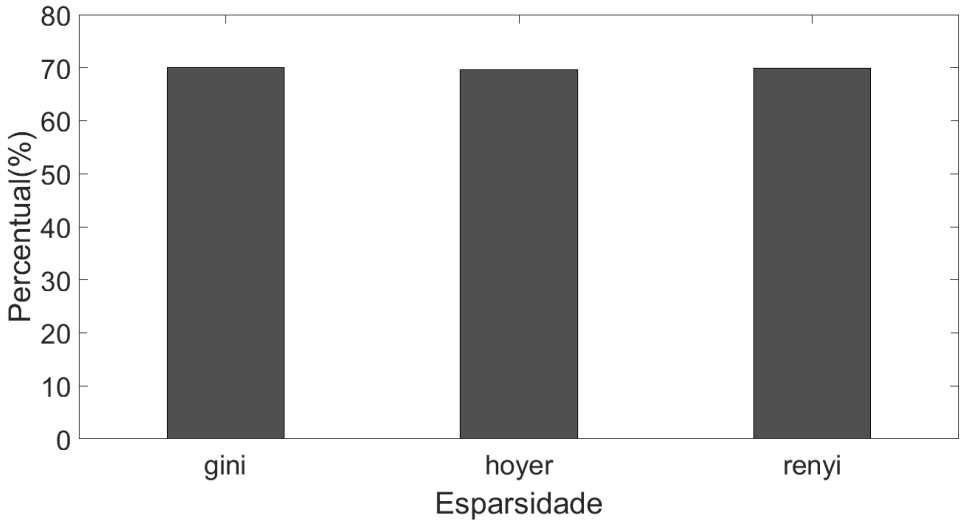

Figura 4.22: Histograma do parâmetro de medida de esparsidade com resultados acima de -1 de ODG gerado pelo teste com esparsidade.

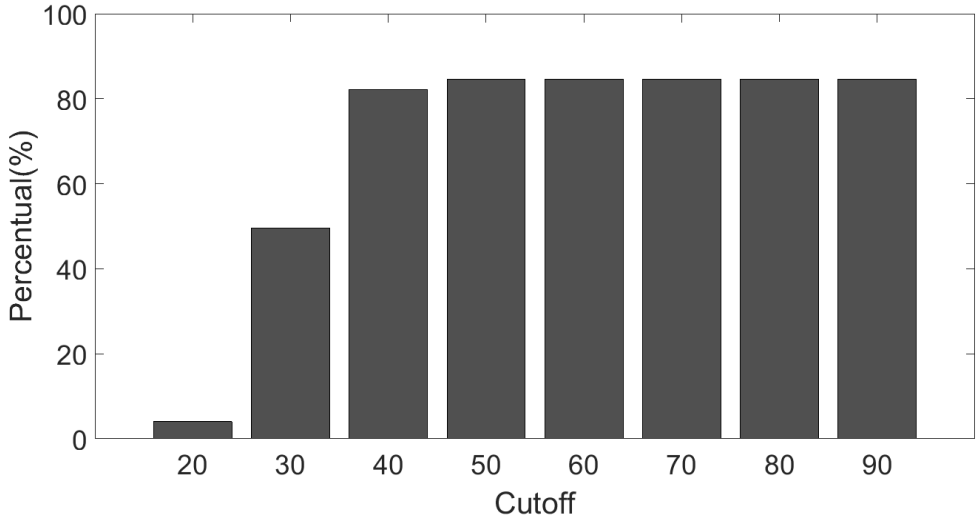

Figura 4.23: Histograma do parâmetro de *cutoff* (em dB) com resultados acima de -1 de ODG gerado pelo teste com esparsidade.

frequência menor do que as outras opções deste parâmetro foram eliminados do conjunto de resultados totais deste teste. Podemos observar também que a presença da opção de sobreposição frequencial 0 neste conjunto é relativamente baixa quando comparada com outras do mesma parâmetro. Portanto, esta é uma opção que também foi eliminada para uma melhor seleção de parâmetros.

Um novo conjunto de histogramas foi gerado, sem estes valores do parâmetro *cutoff* e da sobreposição frequencial, e podem ser vistos nas Figuras 4.24, 4.25, 4.26 e 4.27. Observando estes últimos histogramas, notamos que todas as opções definidas pelos parâmetros são ótimas para o teste proposto. Portanto, concluímos que não é possível extrair mais nenhuma informação destes gráficos.

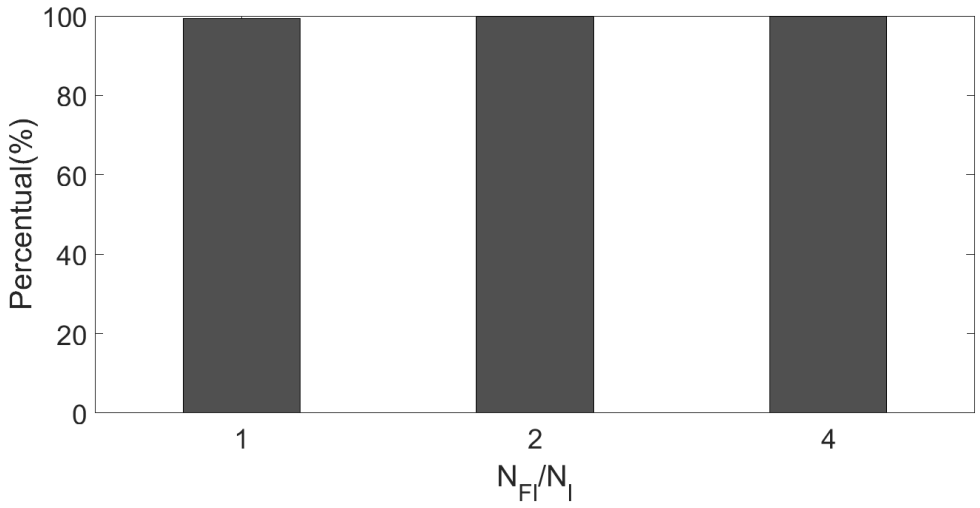

Figura 4.24: Histograma do parâmetro N_{Fl}/N_l com resultados acima de -1 de ODG após eliminação de opções e gerado pelo teste com esparsidade.

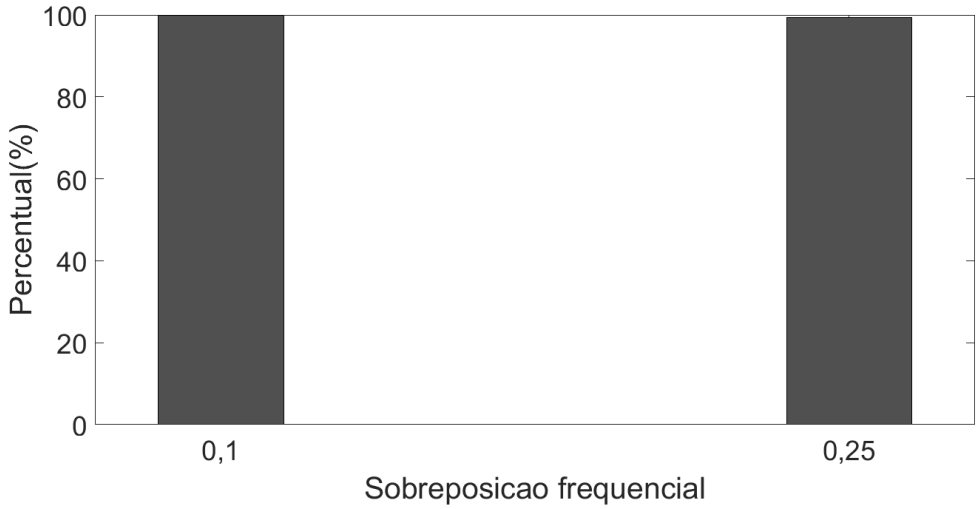

Figura 4.25: Histograma do parâmetro de sobreposição frequencial com resultados acima de -1 de ODG após eliminação de opções e gerado pelo teste com esparsidade.

A segunda etapa de análise consistiu em realizar o cálculo de coeficiente de Pearson com o parâmetros e os indicadores de eficácia e eficiência. A Tabela 4.4 construída com esses coeficientes indica que o parâmetro de esparsidade (isto é, a técnica de medida de esparsidade) é pouco correlacionado com os indicadores.

Em geral, os outros parâmetros apresentam valores inconclusivos sobre a correlação com o ODG. Contudo, o coeficiente de Pearson entre N_{Fl}/N_l e o número de coeficientes não nulos apresenta um valor relativamente alto e, portanto, podemos escolher uma opção menor deste parâmetro.

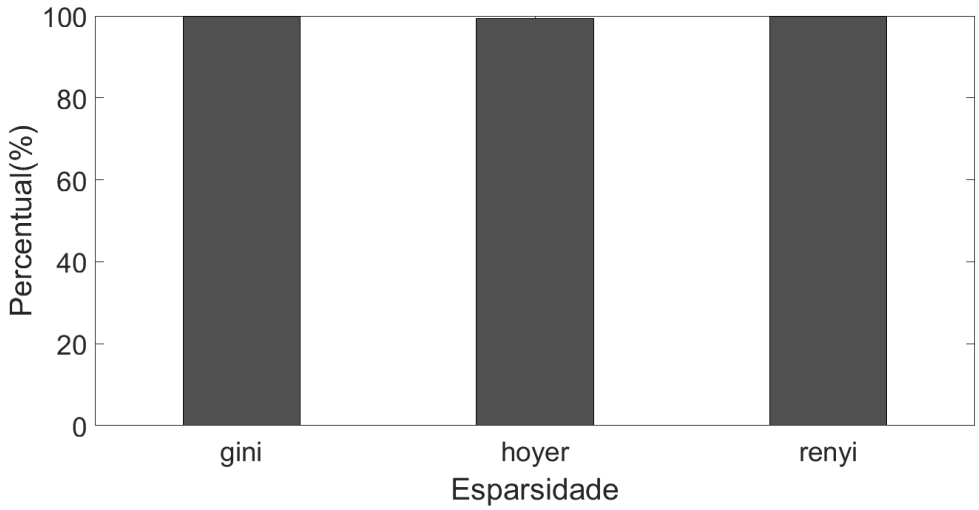

Figura 4.26: Histograma do parâmetro de medida de esparsidade com resultados acima de -1 de ODG após eliminação de opções e gerado pelo teste com esparsidade.

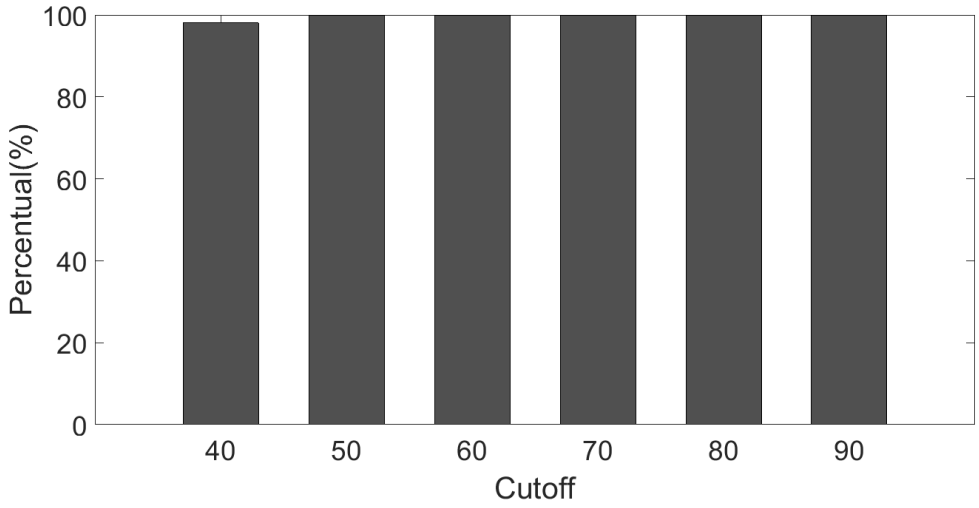

Figura 4.27: Histograma do parâmetro de *cutoff* (em dB) com resultados acima de -1 de ODG após eliminação de opções e gerado pelo teste com esparsidade.

4.5 Comparação com a STFT

Para avaliarmos a eficácia da técnica proposta em comparação com a tradicional STFT, representamos um sinal real de áudio por ambas no domínio tempo-frequência e comparamos os espectrogramas gerados, bem como seus indicadores.

O sinal de teste utilizado para esta comparação foi o exemplo apresentado na Seção 4.2 contendo uma tabla e um sitar. Como cada instrumento presente nesse áudio possui características espectrais distintas (um instrumento percussivo e um instrumento tonal), podemos avaliar essas duas propriedades nas representações.

Na Figura 4.28, podemos observar o sinal no domínio tempo-frequência da STFT. Para essa representação, utilizamos uma função-janela senoidal (proposta por este

Correlação	$e_{\max}$	e_{rms}	ODG	Coeficientes não nulos	Tempo
N_{Fl}/N_l	−0,23648	−0,24707	0,18332	0,57564	0,43130
Sobreposição frequencial	−0,47522	−0,48092	0,39207	0,23691	0,02417
Técnica de esparsidade	−0,00585	−0,03304	−0,00500	−0,01151	−0,00235
Cutoff	−0,04747	−0,001419	0,14779	0,09103	−0,01793

Tabela 4.4: Tabela com coeficientes de correlação de Pearson entre os parâmetros variáveis e os indicadores do teste de esparsidade.

trabalho na Seção 3.1) de comprimento $N_w = 3072$ e com passo temporal de $a = 1536$. A escolha destes parâmetros teve como base os próprios parâmetros da técnica proposta a fim de permitir uma análise competitiva.

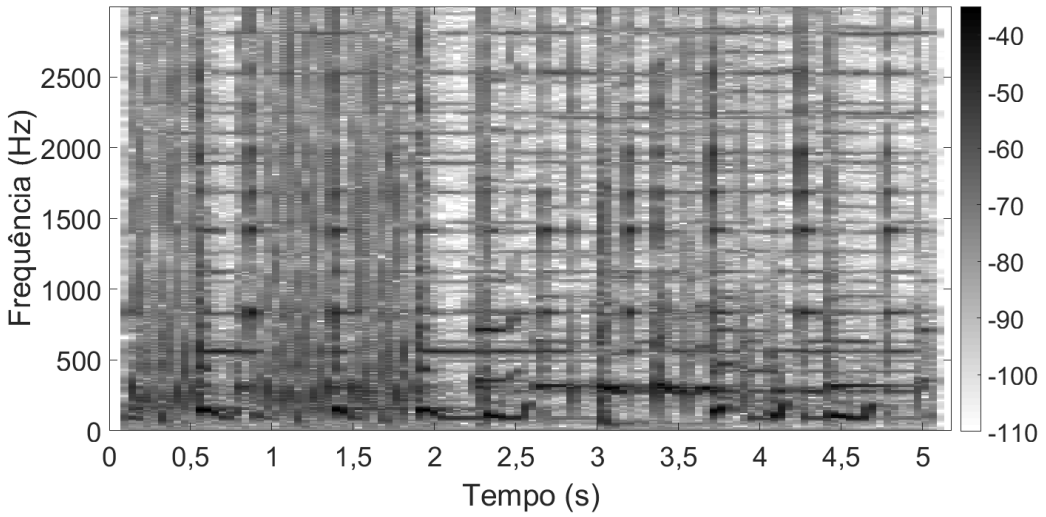

Figura 4.28: Módulo do sinal de teste no domínio tempo-frequencial através da STFT.

Para fins de comparação, na Figura 4.29 construímos a representação tempo-frequência deste sinal através da técnica modificada proposta nesta dissertação[10]. Com base nos testes anteriores, selecionamos os melhores parâmetros que representassem o sinal neste domínio e permitissem uma reconstrução com erro baixo de acordo com a avaliação do PEAQ, ou seja, um alto ODG. Assim, a divisão tempo-frequência foi realizada com janela senoidal de tamanho $N_i = 6144$ em passos de $a_i = 0{,}25N_i$. Geramos representações tempo-frequência de cada intervalo temporal através de uma STFT com uma janela senoidal de comprimentos de $N_l = 1024$, 1248, 1522, 1854, 2262, 2756, 3360, 4096, passo de $a_l = 0{,}5N_l$ e valores de $N_{Fl} = N_l$.

[10]Outros exemplos podem ser ouvidos no site mencionado anteriormente, bem como seus indicadores para uma comparação na reconstrução.

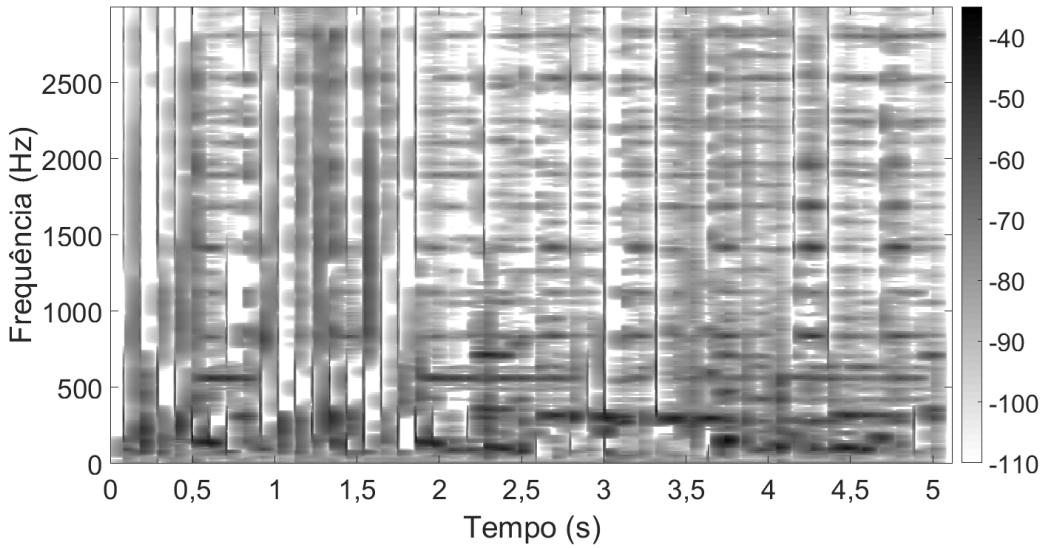

Figura 4.29: Módulo do sinal de teste no domínio do tempo-frequencial através da técnica proposta.

A divisão frequencial implementada refere-se às funções-peso utilizadas nos testes das Seções 4.3 e 4.4 com sobreposição de $0{,}1 f_c$. Para a escolha de coeficientes, utilizamos o indicador de Gini e zeramos coeficientes com 40 dB abaixo do máximo.

Comparando as duas representações, podemos notar que a técnica da STFT não possibilita a distinção explícita dos pulsos temporais (ou seja, sinais percussivos), principalmente na primeira metade do sinal. Em oposição a isto, a técnica proposta, por selecionar adaptativamente a resolução, permite distinguir mais facilmente esse aspecto do sinal. Em relação aos elementos tonais do exemplo, podemos observar que ambos os gráficos representam estas raias frequenciais de forma equivalente nos espectrogramas. Assim sendo, no geral, a interpretação do sinal no domínio tempo-frequência é melhor descrita pela técnica proposta.

Na Tabela 4.5, podemos comparar por outros indicadores os resultados das duas técnicas após a reconstrução do sinal no domínio temporal. Os indicadores de erros e avaliação refletem a reconstrução perfeita da STFT. Contudo, a diferença no ODG entre as duas técnicas não é suficiente para gerar uma discrepância perceptível. Em

Comparação	$e_{\max}$	e_{rms}	ODG	Coeficientes não nulos	Tempo
STFT	0	0	0,21350	446464	0,0743
Técnica proposta	0,0249	0,0815	$-0,1342$	646365	56,099

Tabela 4.5: Comparação de indicadores da reconstrução da STFT com a técnica proposta.

termos de tempo de processamento, ao menos na sua implementação não-otimizada atual, a técnica proposta perde por ordens de grandeza. O indicador de número de coeficientes não nulos aponta que a técnica proposta não é eficaz ao se tratar de compressão imediata de coeficientes. Entretanto, o uso de padrões de comportamento como visto em [12] pode permitir o uso de dicionários para uma compressão mais específica para sinais de áudio, favorecendo o uso desta técnica nessa aplicação.

Portanto, constatamos por meio deste exemplo que os sinais de áudio podem ser representados no domínio tempo-frequência através da técnica proposta, superando a STFT na interpretação do sinal, porém, sendo inferior a esta em aspectos como número de coeficientes e reconstrução perfeita.

Capítulo 5

Conclusões

O principal objetivo desta dissertação foi estudar e sugerir melhorias nas representações tempo-frequenciais na área de processamento de áudio. Para atingir esta meta, descrevemos uma evolução desde a técnica mais comum até a mais complexa, sendo esta foco deste trabalho.

No Capítulo 2, explicamos que a representação temporal possui como vantagem a identificação do momento em que cada som é tocado, porém não permite inferir nenhuma informação sobre as características espectrais do mesmo. Através da DFT, verificamos ser possível converter o sinal deste domínio temporal para um domínio frequencial e, dessa forma, complementamos a representação anterior obtendo dados espectrais sobre o sinal.

Contudo, notamos que não é possível obter essas informações simultaneamente somente com essas representações e, portanto, a projeção de sinais em domínios tempo-frequência se torna uma necessidade. A STFT é apresentada como o método mais comum na literatura e supre essa carência das representações anteriores. Entretanto, esta técnica não é a opção mais adequada para representar sinais de natureza musical.

A CQT é uma representação alternativa cujos coeficientes são gerados pela projeção do sinal em exponenciais complexas espaçadas geometricamente. Esta distribuição frequencial favorece a representação de sinais musicais devido ao espaçamento geométrico das notas musicais da escala de 12 tons em temperamento igual (modernamente utilizada na música ocidental); por conseguinte, é uma técnica mais adequada para eles do que a STFT. Porém, diferentemente de sua antecessora, a CQT não possibilita a reconstrução perfeita do sinal a partir dos coeficientes.

Para solucionar esta dificuldade, a NSGT e sua derivação CQ-NSGT surgiram, permitindo adotar espaçamentos temporais ou frequenciais irregulares, ampliando a dinâmica das representações no domínio tempo-frequência através de *frames* de Gabor. Além disso, como um sucessor direto da CQT, a CQ-NSGT constrói um conjunto de coeficientes similares aos da primeira e ainda permite a reconstrução

perfeita do sinal original.

Ao avaliar estas técnicas estudadas, notamos uma grande deficiência: o fato de a resolução tempo-frequência ter de ser definida antes da análise do sinal. Com vistas a resolver esse problema, este trabalho foi em busca de uma nova técnica na literatura que automaticamente selecionasse a resolução de acordo com o sinal.

Revisada no Capítulo 3, esta técnica consiste em: dividir o sinal em intervalos temporais através de janelamento, projetar cada intervalo em grades tempo-frequência, dividir em blocos frequenciais e, por fim, selecionar para cada intervalo tempo-frequencial a melhor resolução, isto é, a grade que melhor representa esta região.

O método utilizado para escolher automaticamente a melhor resolução por região foi o critério de esparsidade pela entropia de Rényi, uma vez que esta propriedade é recorrente em aplicações para processamento de áudio. Devido a essa divisão em regiões, o plano tempo-frequência construído por esta técnica permite resoluções temporais e frequenciais não-lineares.

Durante o estudo desta técnica, observamos alguns problemas e sugerimos algumas mudanças que poderiam ser investigadas: uma alteração na técnica de janelamento durante a etapa de divisão temporal e o uso de outras medidas de esparsidade além da entropia de Rényi.

Para avaliar o comportamento da técnica original, assim como analisar o efeito dessas mudanças sugeridas, realizamos alguns testes, no Capítulo 4, focando principalmente na reconstrução. Em todos os testes, representamos um ou vários sinais de teste no domínio tempo-frequência através do método e reconstruímos o sinal original. Cada sinal reconstruído foi comparado com o sinal original e foram extraídas informações de erro para serem analisadas.

Inicialmente, já a mudança de janelamento na etapa de divisão temporal permitiu a reconstrução perfeita, que não era possível no método original. Simulamos também os diversos parâmetros em múltiplas combinações em 17 sinais de teste. A partir desta simulação, desenvolvemos uma forma de analisar esses resultados separando os dados dos testes com indicadores mais apropriados, o que foi ilustrado através de histogramas. Ademais, construímos uma tabela de coeficientes de Pearson, relacionando os parâmetros e os indicadores. Essas análises permitiram um estreitamento no conjunto de opções para os parâmetros e, em alguns casos, a escolha da melhor opção.

Em seguida, com alguns dos parâmetros escolhidos, avaliamos o uso de outras técnicas de medida de esparsidade na seleção de coeficientes, e refinamos o parâmetro de *cutoff* (limiar para eliminação de coeficientes) através dos histogramas e da correlação de Pearson. Apesar do estreitamento do *cutoff* ter sido atingido, concluímos que as outras medidas de esparsidade sugeridas não aprimoram a reconstrução dos

sinais.

Por fim, o último teste teve como objetivo comparar no domínio tempo-frequência a técnica proposta (com eliminação de coeficientes) e a STFT. A STFT teve como vantagens a reconstrução perfeita e a complexidade computacional muito menor. Contudo, a técnica proposta, ao custo de pequena redução na qualidade percebida, oferece uma representação bem mais interpretável dos fenômenos acústicos.

5.1 Trabalhos futuros

Este trabalho permite a realização de investigações adicionais para melhorar a eficácia e eficiência desta técnica, entre elas:

- Variação de outros parâmetros que compõem o método, tais como o tamanho do janelamento N_i e a ordem de entropia de Rényi;
- Comparação com outras técnicas de representações tempo-frequenciais no contexto de alguma aplicação específica (compressão de sinais, por exemplo);
- Uso de representações tempo-frequenciais diferentes da STFT como representação-base do método proposto.

Referências Bibliográficas

[1] LIUNI, M., RÖBEL, A., MATUSIAK, E., et al. "Automatic Adaptation of the Time-Frequency Resolution for Sound Analysis and Re-Synthesis", *IEEE Transactions on Audio, Speech and Language Processing*, v. 21, n. 5, pp. 959–970, Maio 2013.

[2] MAJDAK, P., BALAZS, P., KREUZER, W., et al. "A Time-Frequency Method for Increasing the Signal-to-Noise Ratio in System Identification with Exponential Sweeps". In: *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2011)*, pp. 3812–3815, Prague, Czech Republic, Maio 2011.

[3] SIRDEY, A., DERRIEN, O., KRONLAND-MARTINET, R., et al. "Modal Analysis of Impact Sounds with Esprit in Gabor Transforms". In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, pp. 387–392, Paris, France, Setembro 2011.

[4] GRILL, T. "Constructing High-Level Perceptual Audio Descriptors for Textural Sounds". In: *Proceedings of the 9th Sound and Music Computing Conference (SMC 2012)*, pp. 486–493, Copenhagem, Denmark, Julho 2012.

[5] ONCHIS-MOACA, D., GILLICH, G.-R., FRUNZA, R. "Gradually Improving the Readability of the Time-Frequency Spectra for Natural Frequency Identification in Cantilever Beams". In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO 2012)*, pp. 809–813, Bucharest, Romania, Agosto 2012.

[6] TANG, Y., COOKE, M. "Energy Reallocation Strategies for Speech Enhancement in Known Noise Conditions". In: *Proceedings of INTERSPEECH*, pp. 1636–1639, Makuhari, Japan, Setembro 2010.

[7] JAISWAL, R. *Non-Negative Matrix Factorization Based Algorithms to Cluster Frequency Basis Functions for Monaural Sound Source Separation.* Tese de doutorado, School of Electrical Engineering Systems, Dublin Institute of Technology, Dublin, Ireland, Outubro 2013.

[8] RASO, O., BALIK, M., MARTINASEK, Z. "Advantages of Audio Signal Separation to Tonal and Noise Parts for LP modeling". In: *Proceedings of the Workshop of the 12th on Knowledge in Telecommunication Technologies and Optics (KTTO 2012)*, pp. 34–35, Malenovice, Czech Republic, Novembro 2012.

[9] GANSEMAN, J., SCHEUNDERS, P., DIXON, S. "Improving Plca-Based Score-Informed Source Separation with Invertible Constant-Q Transforms". In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO 2012)*, pp. 2634–2638, Bucharest, Romania, Agosto 2012.

[10] BAYRAM, İ., AKYILDIZ, Ö. D. "Primal-Dual Algorithms for Audio Decomposition Using Mixed Norms", *Signal Image and Video Processing*, v. 8, n. 1, pp. 95–110, Dezembro 2013.

[11] MALLAT, S. G., ZHANG, Z. "Matching Pursuits with Time-Frequency Dictionaries", *IEEE Transactions on Signal Processing*, v. 41, n. 12, pp. 3397–3415, Dezembro 1993.

[12] LEWICKI, M. S. *Efficient Coding of Natural Sounds*. Dissertação de mestrado, Fakultät für Mathematik und Naturwissenschaften, Carl-von-Ossietzky-Universität Oldenburg, Augsburg, Deutschland, Julho 2009.

[13] AMBIKAIRAJAH, E., EPPS, J., LIN, L. "Wideband Speech and Audio Coding Using Gammatone Filter Banks". In: *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, v. 2, pp. 773–776, Salt Lake City, USA, Maio 2001.

[14] HURLEY, N., RICKARD, S. "Comparing Measures of Sparsity", *IEEE Transactions on Information Theory*, v. 55, n. 10, pp. 4723–4741, Abril 2009.

[15] BOSI, M., GOLDBERG, R. E. *Introduction to Digital Audio Coding and Standards*. Norwell, USA, Kluwer Academic Publishers, 2002.

[16] GOTO, M., HASHIGUCHI, H., NISHIMURA, T., et al. "RWC Music Database: Music Genre Database and Musical Instrument Sound Database." In: *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, Washington-Baltimore, USA.

[17] OPPENHEIM, A. V., SCHAFER, R. W., BUCK, J. R. *Discrete-Time Signal Processing*. 2ª ed. Upper Saddle River, USA, Prentice-Hall, Inc., 1999.

[18] HAYKIN, S., VAN VEEN, B. *Sinais e Sistemas*. 1ª ed. Porto Alegre, Brasil, Bookman, 2001.

[19] DINIZ, P., DA SILVA, E., NETTO, S. *Processamento Digital de Sinais: Projeto e Análise de Sistemas*. 2ª ed. Porto Alegre, Brasil, Bookman, 2014.

[20] LIM, J. S., OPPENHEIM, A. V. (Eds.). *Advanced Topics in Signal Processing*. 1ª ed. Upper Saddle River, USA, Prentice-Hall, 1987.

[21] LOY, D. *Musimathics: The Mathematical Foundations of Music*, v. 1. 1ª ed. Cambridge, USA, MIT Press, 2006.

[22] BROWN, J. C. "Calculation of a Constant Q Spectral Transform", *Journal of the Acoustical Society of America*, v. 89, n. 1, pp. 425–434, Janeiro 1991.

[23] BROWN, J. C., PUCKETTE, M. S. "An Efficient Algorithm for the Calculation of a Constant Q Transform", *Journal of the Acoustical Society of America*, v. 92, n. 5, pp. 2698–2701, Janeiro 1992.

[24] JAILLET, F., BALAZS, P., DÖRFLER, M. "Nonstationary Gabor Frames". In: *Proceedings of the International Conference on SAMPling Theory and Applications (SAMPTA'09)*, pp. 227–230, Marseille-Luminy, France, Maio 2009.

[25] VELASCO, G. A., HOLIGHAUS, N., DÖRFLER, M., et al. "Constructing an Invertible Constant-Q Transform with Non-Stationary Gabor Frames". In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, pp. 93–99, Paris, France, Setembro 2011.

[26] KOVACEVIC, J., CHEBIRA, A. "Life Beyond Bases: The Advent of Frames (Part I and Part II)", *IEEE Signal Processing Magazine*, v. 24, n. 5, pp. 115–125, Setembro 2007.

[27] NECCIARI, T., BALAZS, P., HOLIGHAUS, N., et al. "The ERBlet Transform: An Auditory-Based Time-Frequency Representation with Perfect Reconstruction". In: *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, Canada.

[28] LATHI, B. P. *Sinais e Sistemas Lineares*. 2ª ed. Porto Alegre, Brasil, Bookman, 2007.

[29] PLUMBLEY, M., BLUMENSATH, T., DAUDET, L., et al. "Sparse Representations in Audio and Music: From Coding to Source Separation", *Proceedings of the IEEE*, v. 98, n. 6, pp. 995–1005, Junho 2010.

[30] RÉNYI, A. "On Measures of Entropy and Information". In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561, Berkeley, USA, 1961.

[31] BARANIUK, R. G., FLANDRIN, P., JANSSEN, A. J. E. M., et al. "Measuring Time-Frequency Information Content Using the Rényi Entropies", *IEEE Transactions on Information Theory*, v. 47, n. 4, pp. 1391–1409, Agosto 1998.

[32] YOU, J., REITER, U., HANNUKSELA, M. M., et al. "Perceptual-based Quality Assessment for Audio-Visual Services: A Survey", *Signal Processing: Image Communication*, v. 25, n. 7, pp. 482–501, Agosto 2010.

[33] PEEBLES, P. *Probability, Random Variables and Random Signal Principles*. 2ª ed. New York, USA, McGraw-Hill, 2002.