CONTRIBUIÇÕES A MÉTODOS PARA REPRESENTAÇÃO TEMPO-FREQUENCIAL DE SINAIS DE MÚSICA

Isabela Ferrão Apolinário

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.

Orientador: Luiz Wagner Pereira Biscainho

Rio de Janeiro
Setembro de 2015

CONTRIBUIÇÕES A MÉTODOS PARA REPRESENTAÇÃO TEMPO-FREQUENCIAL DE SINAIS DE MÚSICA

Isabela Ferrão Apolinário

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. Eduardo Antônio Barros da Silva, Ph.D.

Prof. Paulo Antonio Andrade Esquef, D.Sc. (Tech)

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2015

Apolinário, Isabela Ferrão

Contribuições a Métodos para Representação Tempo-Frequencial de Sinais de Música/Isabela Ferrão Apolinário. – Rio de Janeiro: UFRJ/COPPE, 2015.

XI, 76 p.: il.; 29, 7cm.

Orientador: Luiz Wagner Pereira Biscainho

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2015.

Referências Bibliográficas: p. 73 – 76.

1. Sinais de áudio. 2. Análise Tempo-Frequência. 3. Transformada *Fan Chirp*. 4. Esparsidade Estruturada. I. Biscainho, Luiz Wagner Pereira. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Gostaria de agradecer, em primeiro lugar, ao meu orientador e amigo, Luiz Wagner, por ter acreditado em mim e no meu trabalho desde o princípio. Seu reconhecimento fez muita diferença no decorrer desta etapa.

Agradeço aos professores Martín Rocamora e Pablo Cancela da Udelar, Uruguai, pela oportunidade de trabalhar em conjunto.

Agradeço aos meus amigos Allan, Leonardo, Lucas Arrabal, Lucas Simões e Maurício por estarem presentes ao longo desses dois anos de mestrado e alguns mais de graduação. Vocês tornaram o caminho mais fácil e divertido. Agradeço, também, aos meus colegas do SMT. Cada um contribuiu de alguma forma para que eu chegasse até aqui.

Agradeço aos meus pais, pelo suporte e carinho.

Agradeço, por fim, ao meu amigo Victor por ter tido a paciência de me entender e estar ao meu lado na reta final. Sua ajuda foi imprescindível.

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

CONTRIBUIÇÕES A MÉTODOS PARA REPRESENTAÇÃO TEMPO-FREQUENCIAL DE SINAIS DE MÚSICA

Isabela Ferrão Apolinário

Setembro/2015

Orientador: Luiz Wagner Pereira Biscainho

Programa: Engenharia Elétrica

O foco deste trabalho é o estudo de métodos para representação tempo-frequencial de sinais de áudio. Geralmente, estes podem ser modelados por componentes de natureza tonal, transitória ou ruidosa, cada uma das quais apresentando um padrão no plano tempo-frequencial. Tais padrões podem ser explorados para gerar representações mais informativas. Em particular, as parcelas determinísticas do sinal tendem a apresentar alguma estrutura, que pode favorecer representações esparsas.

Dentre as principais ferramentas de representação de sinais já desenvolvidas, estão as transformadas de Fourier de Tempo Curto ou STFT, de Q Constante ou CQT e do *Fan Chirp* ou FChT. A teoria por trás de cada uma delas é apresentada aqui.

A FChT favorece a representação de sinais harmônicos, e considera que a fundamental do sinal sob análise varia linearmente no tempo, o que permite atingir maior esparsidade na representação de sinais não-estacionários. Aqui, o método proposto é estendido para abarcar variações não-lineares, resultando em representações esparsas para sinais de áudio com variações frequenciais rápidas. Além disso, um novo método para a estimação da taxa de variação da fundamental é utilizado como alternativa. Bons resultados foram obtidos para sinais sintéticos com ruído de fundo.

O trabalho conclui abordando a esparsidade estruturada, utilizando dicionários compostos a partir da STFT (como na proposta original) e da CQT (como nova proposta). A principal aplicação em vista é a redução de ruído em sinais de áudio. Uma bateria de experimentos foi elaborada, e os resultados para cada dicionário foram comparados quantitativamente. Para os parâmetros de entrada selecionados, não foi observada melhoria ao substituirmos a STFT pela CQT.

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

CONTRIBUTIONS FOR METHODS OF TIME-FREQUENCY REPRESENTATION OF MUSIC SIGNALS

Isabela Ferrão Apolinário

September/2015

Advisor: Luiz Wagner Pereira Biscainho

Department: Electrical Engineering

This work focuses on the study of methods for time-frequency representation of audio signals. Usually, audio signals can be modeled by components of tonal, transitory or noisy natures, each of which presents a pattern in the time-frequency plane. Such patterns can be explored to create more informative representations. In particular, the deterministic portions of the signal tend to present some structure, which may yield sparse representations.

Among the main tools developed for signal representation are the Short-Time Fourier Transform or STFT, the Constant-Q Transform or CQT, and the Fan-Chirp Transform or FChT. The theory behind each one of them is presented here.

The FChT provides a good representation for harmonic signals, and considers their fundamental to vary linearly in time, thus allowing for sparser representations of non-stationary signals. Here, the proposed method is extended to embrace non-linear frequency variations, resulting in sparser representations for audio signals with rapid frequency variations. Moreover, a new method for the estimation of the fundamental chirp rate is used as an alternative. Good results were obtained for synthetic signals with background noise.

This work concludes by approaching the structured sparsity, using dictionaries consisting of the STFT (as in the original proposal) and the CQT (as a new proposal). Here, the main application considered is audio signal denoising. An extended set of experiments was elaborated, and the results for each dictionary were quantitatively compared. For the chosen input parameters, no improvement was observed by replacing the STFT by the CQT.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Tema

A análise de sinais é o estudo e caracterização das suas propriedades básicas. Sua teoria foi desenvolvida, historicamente, na época do descobrimento dos sinais fundamentais na natureza, tais como o campo elétrico, a onda sonora e correntes elétricas [1]. Um sinal é geralmente uma função de várias variáveis. Neste trabalho, vamos considerar sinais variantes no tempo. Podemos pensar, portanto, em representar sinais neste domínio.

Além do tempo, a representação mais importante é a no domínio da frequência. No caso específico de um sinal de áudio, por exemplo, sabemos que este é composto em grande parte por sons (como notas musicais) com *pitch* bem definido e relacionados, portanto, a informações frequenciais. Para representar o sinal neste domínio, podemos calcular a sua Transformada de Fourier [2].

Até agora, falamos dos domínios "tempo" e "frequência" de forma isolada. Cada representação carrega informações importantes do sinal, porém, de forma pouco localizada. Sabemos, por exemplo, quais frequências estão presentes em um sinal, mas não quando (em quais instantes de tempo) elas ocorreram. Representações tempo-frequenciais tem como principal finalidade sinalizar a evolução temporal das componentes frequenciais presentes.

É necessária, portanto, uma abordagem para tratar o problema de análise tempo-frequencial. Muitos métodos já foram propostos para analisar sinais não estacionários, como a clássica Transformada de Fourier de Tempo Curto ou STFT (do inglês, *Short-Time Fourier Transform*) e outras menos utilizadas, como a Transformada de Q Constante ou CQT (do inglês, *Constant-Q Transform*) [3], a Transformada de Q Limitado ou BQT (do inglês, *Bounded-Q Transform*) [4], a distribuição de Wigner e a Transformada Discreta de Wavelet, entre outras [1], [5], [6].

Quando nos referimos às representações tempo-frequenciais de sinais de áudio,

geralmente as associamos a uma imagem com ocorrências frequentes de padrões na forma de linhas localizadas no tempo ao longo da frequência, sinalizando informações impulsivas, e na forma de linhas localizadas na frequência ao longo do tempo, sinalizando informações tonais. Assim sendo, considerando que o tempo é representado no eixo horizontal e a frequência, no eixo vertical, esperamos, por exemplo, visualizar um gráfico composto por diversas linhas horizontais associado a uma nota de violino, onde cada linha representaria uma parcial harmônica (incluindo a fundamental) da nota tocada.

Diversas aplicações podem se beneficiar da análise tempo-frequencial de sinais de áudio. Podemos pensar, por exemplo, em separação de fontes sonoras [7], detecção da melodia principal de uma música [8], transcrição automática [9], síntese de música [10], alteração do timbre de instrumentos [11], redução de ruído de fundo [12], entre outros.

1.2 Motivação

Aqui, o conceito de esparsidade é utilizado como sinônimo de concentração. Quando nos referimos a representações esparsas, queremos dizer, na verdade, representações bem concentradas em torno das regiões do plano tempo-frequencial correspondentes aos sons tonais e percussivos emitidos pelas fontes sonoras. Procuramos representações esparsas pois estas levam a interpretações que guardam relação com os processos de geração do som, ou seja, com suas fontes.

Como foi dito, existem, hoje em dia, diversos métodos para analisar o conteúdo tempo-frequencial de sinais em geral. Uma das ferramentas mais utilizada em processamento de áudio é a STFT. Isso porque ela pode ser rapidamente calculada (através do algoritmo FFT) e apresenta resultados bastante intuitivos de como o conteúdo frequencial (as notas da música) evolui com o tempo. Ela consiste em separar o sinal em blocos temporais, geralmente com alguma sobreposição, e analisar, em seguida, cada trecho por meio de uma transformada de Fourier [1]. Neste processo, consideramos que o conteúdo frequencial do sinal dentro de cada bloco é invariante.

A STFT, no entanto, apresenta certas limitações. Ao dividir o sinal em blocos, estamos, essencialmente, multiplicando o sinal por uma janela limitada temporalmente. Como resultado, introduzimos informação extra e não desejável na representação tempo-frequencial resultante, já que, ao janelarmos o sinal, o seu conteúdo frequencial original é convoluído pela Transformada de Fourier da janela [2]. Tal operação provoca, por si só, um espalhamento da informação. Além disso, o comprimento dessa janela está diretamente relacionado com o Princípio da Incerteza que, por sua vez, está relacionado com a resolução tempo-frequencial da transformada

como um todo [1].

Uma alternativa para amenizar os efeitos causados pelo janelamento é utilizar, no lugar da STFT, a CQT. Esta analisa diferentes *bins* de frequência com janelas temporais de diferentes tamanhos [3]. Assim sendo, temos uma melhor resolução frequencial para baixa frequências e uma melhor resolução temporal para altas frequências.

Outro fator relevante da STFT é que ela considera sinais localmente estacionários, ou seja, sinais cujas componentes frequenciais não variam dentro dos intervalos de tempo especificados pelo comprimento da janela escolhido. No caso de música, o conteúdo frequencial do sinal varia a todo instante com as notas emitidas e até mesmo com a forma na qual o músico toca o instrumento. Assim, uma análise que leve em consideração essa variação de frequência seria capaz de ilustrar melhor a realidade do sinal.

A Transformada *Fan Chirp* ou FChT (do inglês, *Fan Chirp Transform*) permite que componentes relacionadas harmonicamente possuam uma determinada variação frequencial ao longo do tempo [8]. Assim sendo, a FChT é uma boa ferramenta para lidar com sinais de áudio, pois geralmente consideramos que as componentes com altura definida destes sinais são decompostas em uma frequência fundamental e suas parciais harmônicas. Em [8], a FChT é aplicada diretamente em sinais de música por meio da sua versão de tempo curto, a STFChT (do inglês, *Short-Time Fan Chirp Transform*). Ela modela a frequência fundamental de trechos do sinal como uma reta com taxa de inclinação α.

Por definição, podemos calcular a FChT de um dado sinal para diferentes valores de α. Estamos interessados, no entanto, em encontrar o parâmetro α^* que resulte na representação cuja resolução tempo-frequencial seja ótima, ou seja, a mais esparsa possível. Este valor corresponde, teoricamente, à taxa de inclinação verdadeira da fundamental. Para estimar α^*, realizamos uma busca exaustiva, onde uma das etapas é a acumulação de parciais harmônicas para um conjunto pré-determinado de frequências candidatas a fundamental [8].

Ainda no contexto da FChT, podemos, agora, querer analisar sinais com um alto grau de não-estacionariedade. Dessa forma, diminuímos o tamanho da janela de análise para que a suposição de uma fundamental com variação linear ainda seja válida. Essa medida, porém, aumenta o grau de espalhamento frequencial, de acordo com o Princípio da Incerteza. Uma outra proposta seria acrescentar mais um grau de liberdade na estimação da fundamental do sinal. Isso permitiria melhorar a resolução da transformação de tais sinais.

A operação de busca exaustiva citada acima sofre influências do timbre dos instrumentos e de possíveis parciais inarmônicas. Como alternativa, utilizamos um método que leva em consideração apenas a localização frequencial das parciais neste

processo e, além disso, utiliza um modelo de inarmonicidade para modelar tais parciais inarmônicas.

Por fim, estudamos representações a partir de dicionários redundantes. Estes dicionários são compostos, essencialmente, de um conjunto de vetores que constituem uma base acrescidos de um outro conjunto de vetores formados a partir de combinações lineares desses. Queremos, com isso, obter um maior grau de esparsidade na representação em troca de uma quantidade maior de informação, ou seja, a redundância acrescentada [13]. Além disso, um dos principais objetivos aqui é, como mencionado, levar em conta a estrutura interna do sinal de áudio, traduzida como persistência temporal, no caso de sinais tonais, e persistência frequencial, no caso de sinais percussivos. A técnica abordada em [12] utiliza a STFT como dicionário com o objetivo de eliminar o ruído de fundo presente em sinais de música. Esta transformada, no entanto, é limitada pelo Princípio da Incerteza. Resolvemos, portanto, explorar o algoritmo proposto utilizando, agora, uma versão adaptada da CQT [14] como dicionário, no lugar da STFT.

1.3 Objetivos

O principal objetivo dessa dissertação é analisar diferentes técnicas para representar de forma fiel sinais de música em termos do seu conteúdo tempo-frequencial. Selecionamos as ferramentas abordadas na seção anterior para estudá-las mais a fundo e, mais especificamente, entender suas propriedades de forma a aplicá-las em sinais de música com características de interesse. Queremos, com isso, ser capazes de implementar as melhorias propostas e analisar os resultados obtidos.

De forma geral, a principal meta deste trabalho é formar uma base de conhecimento sólida que permita a continuação dos estudos nesta linha de pesquisa e, além disso, propor e realizar experimentos que abram caminho para outras melhorias nas técnicas estudadas.

1.4 Organização deste Trabalho

Este trabalho prossegue com um capítulo introdutório, o Capítulo 2, que trata dos principais conceitos abordados nessa dissertação, tais como as transformadas STFT e CQT, a decomposição em dicionários redundantes e técnicas de recuperação esparsa de sinais de áudio. Além disso, ele é responsável por introduzir termos técnicos e a nomenclatura a serem utilizados ao longo deste trabalho.

Em seguida, o Capítulo 3 trata da FChT. Apresentamos sua definição, forma de implementação, melhorias propostas e implementadas e resultados.

O Capítulo 4 fala da Esparsidade Estruturada, abordando conceitos como normas mistas e limiarização suave. Falamos, além disso, da integração da técnica utilizada para a redução de ruído em sinais de áudio com a CQT e apresentamos os resultados obtidos.

Por fim, o Capítulo 5 apresenta conclusões e fala de trabalhos futuros.

Capítulo 2

Elementos da Análise Tempo-Frequência

2.1 Considerações Iniciais

Um sinal discreto de L amostras $x(n)$, com $n = 0, 1, \ldots, L-1$, pode ser interpretado como um vetor $\mathbf{x}$ no espaço $\mathbb{R}^L$. Podemos, por isso, selecionar uma base em $\mathbb{R}^L$ para representá-lo de forma a evidenciar certas propriedades. No caso de um sinal de música, por exemplo, queremos ser capazes de identificar desde a presença de notas musicais cuja altura se mantém por tempo relativamente longo até sons percussivos de curta duração com finalidade rítmica.

As notas musicais em um sinal de áudio são um exemplo cujo conteúdo mais relevante está concentrado em frequência, enquanto que os sons impulsivos, em tempo. Queremos, no entanto, obter também informações adicionais, como tempos de ínicio e término das notas, presença de vibratos e tremolos, características transitórias e timbre. Assim sendo, o desejável é uma representação que ilustre a evolução frequencial ao longo do tempo [1]. Entretanto, tempo e frequência são grandezas indissociáveis, o que agrega a essa desejada descrição uma dificuldade intrínseca.

Quando falamos em representação tempo-frequência nos referimos ao mapeamento do sinal de entrada $x(n)$ em um plano bidimensional, no qual um dos eixos representa o tempo e o outro, a frequência. Essa operação pode, também, ser interpretada como a decomposição do vetor $\mathbf{x}$ em uma família de vetores $\{\phi_{k,j}(n)\}_{k,j}$[2]. O conjunto de átomos, como são chamados, forma um dicionário $\mathbf{\Phi}$. Esse modelo será direta ou indiretamente utilizado neste trabalho e será explicado na próxima seção.

[1] Podemos, também, pensar no caso análogo, que é a evolução temporal ao longo da frequência.

[2] Aqui, utilizamos a notação $\{\phi_{k,j}(n)\}_{k,j}$ para nos referir ao conjunto de todos vetores $\boldsymbol{\phi}_{k,j}$, onde $\boldsymbol{\phi}_{k,j} = [\phi(0)_{k,j} \;\; \phi(1)_{k,j} \;\; \ldots \;\; \phi(L-1)_{k,j}]^{\mathrm{T}}$ é um vetor localizado no tempo de índice k e na frequência de índice j, para todos os valores de k e j considerados.

Os vetores $\{\phi_{k,j}(n)\}_{k,j}$ não necessariamente constituem uma base, ainda que consigam gerar todo o espaço de interesse. O dicionário pode ser redundante e proporcionar infinitas formas de decompor $\mathbf{x}$. E nesse caso, é necessário algum critério que determine qual destas é a mais apropriada. Um critério com o potencial de analisar o sinal em partes relacionadas às fontes que o geraram é a esparsidade da representação, que maximiza a contribuição sinalizada por cada coeficiente. Além disso, a esparsidade é facilitadora para várias aplicações, tais como codificação e redução de ruído.

A ideia deste capítulo é apresentar as principais técnicas utilizadas e relacionadas ao trabalho. Além disso, pretendemos introduzir a nomenclatura que será utilizada nos capítulos posteriores. Começamos abordando as decomposições atômicas, onde tratamos dos conceitos de *frames*, Transformada de Fourier de Tempo Curto (STFT) e Transformada de Q constante. A seguir, abordamos uma técnica conhecida como Recuperação Esparsa, que tem como principal objetivo obter representações maximamente esparsas. Esta servirá como base para o Capítulo 4.

2.2 Decomposições Atômicas

2.2.1 *Frames*

Quando expandimos o vetor $\mathbf{x}$ utilizando o dicionário $\mathbf{\Phi}$, estamos, na verdade, reescrevendo $\mathbf{x}$ como uma soma ponderada dos vetores componentes de $\mathbf{\Phi}$, $\{\phi_{k,j}(n)\}_{k,j}$. Os coeficientes de ponderação, $\{c_{k,j}\}_{k,j}$, também chamados de coeficientes de expansão, representam o sinal no domínio da transformada. Se o dicionário for redundante, haverá infinitas possibilidades para a escolha de $\{c_{k,j}\}_{k,j}$. Assim sendo, coeficientes esparsos destacam as principais características do sinal de entrada $x(n)$, já que estes indicam onde (em quais vetores do dicionário) está concentrada a informação do sinal.

Um exemplo de expansão em um dicionário é a Transformada Discreta de Fourier ou DFT (do inglês, *Discrete Fourier Transform*) $X(j)$ de um sinal $x(n)$ com L amostras de duração. Nesse caso, temos uma família de vetores na forma $\{\mathrm{e}^{\mathrm{i}2\pi jn/L}\}_j$ e escrevemos

$$x(n) = \frac{1}{L}\sum_{j=0}^{L-1} X(j)\mathrm{e}^{\mathrm{i}2\pi jn/L}, \tag{2.1}$$

para $n = 0, 1, \ldots, L-1$, onde $X(j)$ são os coeficientes complexos de ponderação. Se tivermos, por exemplo, uma entrada senoidal da forma $x(n) = (-1)^n = \mathrm{e}^{\mathrm{i}\pi n}$, teremos, no domínio da transformada, toda a energia concentrada em um só elemento: $X(j) = L$, para $j = L/2$ e $X(j) = 0$, para os demais valores de j (considerando que L é par). Dessa forma, temos uma representação bem mais esparsa na frequência

que no tempo.

Podemos escrever a operação de DFT na sua forma matricial. O vetor $\mathbf{x}$ é escrito como $\mathbf{x} = [x(0)\ x(1)\ \ldots\ x(L-1)]^{\mathrm{T}}$, os coeficientes de expansão são dados por $\mathbf{c} = \frac{1}{L}[X(0)\ X(1)\ \ldots\ X(L-1)]^{\mathrm{T}}$ e os vetores $\{\phi_j(n)\}_j$ são organizados matricialmente como as colunas da matriz que representa o dicionário $\mathbf{\Phi}$, ou seja,

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_0 & \boldsymbol{\phi}_1 & \ldots & \boldsymbol{\phi}_{L-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & \mathrm{e}^{\mathrm{i}2\pi/L} & \ldots & \mathrm{e}^{\mathrm{i}2\pi(L-1)/L} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{e}^{\mathrm{i}2\pi(L-1)/L} & \ldots & \mathrm{e}^{\mathrm{i}2\pi(L-1)^2/L} \end{bmatrix} \tag{2.2}$$

Assim, a Eq. 2.1 dada por

$$x(n) = \sum_j c_j \phi_j(n) \tag{2.3}$$

pode ser reescrita em sua forma matricial como

$$\mathbf{x} = \mathbf{\Phi}\mathbf{c}. \tag{2.4}$$

Neste caso, os vetores $\{\phi_j(n)\}_j$ formam uma base ortogonal em $\mathbb{C}^L$, pois $\langle \boldsymbol{\phi}_{j_1}, \boldsymbol{\phi}_{j_2} \rangle = 0$, para $j_1 \neq j_2$ e $\langle \boldsymbol{\phi}_{j_1}, \boldsymbol{\phi}_{j_2} \rangle \neq 0$, para $j_1 = j_2$. Assim sendo, podemos calcular os coeficientes de expansão $\{c_j\}_j$ por meio de produtos escalares entre o sinal $x(n)$ e os vetores $\{\phi_j(n)\}_j$ [6]. Quando os vetores da base não são ortogonais entre si, é possível recorrer a uma base dual, formada por um conjunto vetores $\{\tilde{\phi}_j(n)\}_j$. Estes, por sua vez, são ortogonais aos elementos de $\mathbf{\Phi}$. A Eq. (2.3) se torna

$$x(n) = \sum_j \langle x(n), \tilde{\phi}_j(n) \rangle \phi_j(n). \tag{2.5}$$

O dicionário $\mathbf{\Phi}$ pode, ainda, ser completo ou sobrecompleto (também chamado de redundante). Quando completo, não há dependência entre entre os vetores $\{\phi_{k,j}(n)\}_{k,j}$ e, consequentemente, só uma representação possível (abordada acima). Quando redundante, há mais vetores que o necessário para formar uma base, ou seja, estes não são mais linearmente independentes entre si. A proposta da representação por *frames* é justamente esta: torná-la redundante e, com isso, aumentar a flexibilidade na escolha da representação para o sinal analisado [15].

Por definição, a família de vetores $\mathbf{\Phi} = \{\phi_\gamma(n)\}_{\gamma \in \Gamma}$ em um espaço de Hilbert $\mathbb{H}$ é denominada um *frame* se existirem constantes A e B, com $0 < A \leq B < \infty$, tal que para todo $\mathbf{x}$ em $\mathbb{H}$ [15],

$$A\|\mathbf{x}\|^2 \leq \sum_{\gamma \in \Gamma} |\langle \boldsymbol{\phi}_\gamma, \mathbf{x} \rangle|^2 \leq B\|\mathbf{x}\|^2, \tag{2.6}$$

onde A e B são chamados de limites do *frame*. Esse "enquadramento" [3] está relacionado à estabilidade da representação. Podemos pensar em uma analogia ao teorema de Parseval para o caso da decomposição em uma base ortogonal. Se dividirmos a Eq. (2.6) por $\|\mathbf{x}\|^2$, percebemos que A e B determinam os limites inferior e superior da energia normalizada dos produtos internos [6].

Se $A = B$, o conjunto de vetores é denominado *frame* apertado. Nesse caso, os coeficientes de expansão podem ser obtidos por meio de produtos internos entre os vetores $\{\phi_\gamma(n)\}_{\gamma\in\Gamma}$ e o sinal $x(n)$. A expansão de $x(n)$ nos vetores da base é expressa por

$$x(n) = A^{-1}\sum_{\gamma\in\Gamma}\langle\phi_\gamma(n), x(n)\rangle\phi_\gamma(n). \tag{2.7}$$

onde o termo A^{-1} é uma medida da redundância da representação [6].

Ainda neste contexto, se o conjunto de expansão não é um *frame* apertado, ou seja $A \neq B$, recorremos novamente a um *frame* dual (composto por vetores $\tilde{\mathbf{\Phi}} = \{\tilde{\phi}_\gamma(n)\}_\gamma$). Agora, por conta da redundância, há infinitas formas de determinar $\tilde{\mathbf{\Phi}}$. Uma delas, chamada de *frame* dual canônico, é dada por $\tilde{\mathbf{\Phi}}^{\mathrm{H}} = \mathbf{\Phi}^{\mathrm{H}}\mathbf{S}^{-1}$, onde $\mathbf{A}^{\mathrm{H}}$ é o hermitiano da matriz $\mathbf{A}$, ou seja, $\mathbf{A}^{\mathrm{H}} = (\mathbf{A}^{\mathrm{T}})^*$ e $\mathbf{S} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{H}}$ é chamado de operador de *frame* [15]. Esta escolha leva à solução chamada de MOF (do inglês, *Method of Frames*), dada por [13]

$$\mathbf{c} = \mathbf{\Phi}^{\mathrm{H}}\mathbf{S}^{-1}\mathbf{x}. \tag{2.8}$$

No caso das bases ortogonais, vemos que a matriz $\mathbf{\Phi}$ composta pelos vetores de expansão é quadrada. Como observamos na expansão de $x(n)$ por meio da DFT, $\mathbf{\Phi}$ pertence ao espaço $\mathbb{C}^{L\times L}$, o vetor de coeficientes $\mathbf{c}$ ao espaço $\mathbb{C}^L$ e, como já mencionado anteriormente, o vetor $\mathbf{x}$, ao espaço $\mathbb{R}^L$. Quando tratamos de *frames*, o dicionário passa a ser uma matriz retangular cujas colunas não são mais linearmente independentes entre si, ou seja, há um acréscimo no número de linhas que representa a redundância da representação.

Na DFT, notamos que os átomos dependem somente de um índice j, relacionado à frequência. Agora, elaboramos o caso bidimensional, no qual consideramos que a representação depende tanto da frequência quanto do tempo e é, portanto, apropriada para sinais não estacionários. Podemos escrever a decomposição do sinal $x(n)$ nos átomos de tempo e frequência $\{\phi_{k,j}(n)\}_{k,j}$ da seguinte forma:

$$x(n) = \sum_k\sum_j c_{k,j}\phi_{k,j}(n), \tag{2.9}$$

para $n = 0, 1, \ldots, L-1$.

[3] Daí o nome *frame*, que, em inglês, quer dizer quadro.

Escrevemos, agora, a Eq. (2.9) na sua forma matricial. Montamos uma matriz $\mathbf{\Phi}$ que contém, em suas colunas, os átomos $\{\phi_{k,j}(n)\}_{k,j}$. As colunas de $\mathbf{\Phi}$ serão da forma

$$\text{coluna}_{k,j}[\mathbf{\Phi}] = [\phi_{k,j}(0)\ \phi_{k,j}(1)\ \ldots\ \phi_{k,j}(L-1)]^{\mathrm{T}} \tag{2.10}$$

e as linhas,

$$\text{linha}_i[\mathbf{\Phi}] = [\phi_{0,0}(i)\ \phi_{0,1}(i)\ \ldots\ \phi_{0,J-1}(i)\ \ldots\ \phi_{K-1,0}(i)\ \phi_{K-1,1}(i)\ \ldots\ \phi_{K-1,J-1}(i)], \tag{2.11}$$

ou seja, mantemos o índice k fixo e incrementamos j de 0 até $J-1$; em seguida, fazemos o mesmo para $k+1$ e assim por diante, até $k = K-1$. Temos, como resultado, uma matriz de dimensão $L \times P$, com $P = JK$. O vetor de coeficientes $\mathbf{c}$ será, portanto, dado por

$$\mathbf{c} = [c_{0,0}\ c_{0,1}\ \ldots\ c_{0,J-1}\ \ldots\ c_{K-1,0}\ \ldots c_{K-1,J-1}]^{\mathrm{T}}. \tag{2.12}$$

A Eq. (2.9) se torna, então,

$$\mathbf{x} = \mathbf{\Phi c}. \tag{2.13}$$

Essa nomeclatura matricial será útil no Cap. 4, quando trataremos dos métodos de otimização para recuperar um sinal original $x(n)$ a partir de uma versão corrompida por ruído $y(n)$.

2.2.2 A Transformada de Fourier de Tempo Curto

No caso da Transformada Discreta de Fourier, o sinal $x(n)$ é projetado em L amostras contíguas de L senoides complexas harmônicas. Isso faz com que informações temporais, tais como início das notas tocadas, variações rápidas de *pitch* e presença de intrumentos percussivos, sejam diluídas na representação. No caso da Transformada de Fourier de Tempo Curto, as senoides puras são substituídas por átomos localizados em tempo e em frequência [5].

Os átomos $\{\phi_{k,j}(n)\}_{k,j}$ são construídos através de translações no tempo e na frequência de uma dada função janela $h(n)$ com suporte compacto [12]. Assim, temos $\phi_{k,j}(n) = M_{bj}T_{ka}h(n)$, onde T_x é o operador de deslocamento no tempo e M_ω, o de deslocamento na frequência, ou seja, $T_x h(n) = h(n-x)$ e $M_\omega h(n) = h(n)\mathrm{e}^{\mathrm{i}2\pi n\omega/L}$, resultando em $\phi_{k,j}(n) = h(n-ka)\mathrm{e}^{\mathrm{i}2\pi nbj/L}$. Os parâmetros a e b, com $a, b \in \mathbb{N}$, estão relacionados à amostragem no tempo e na frequência, respectivamente, e $j = 0, \ldots J-1$, $k = 0, \ldots, K-1$, com $Ka = Jb = L$.

Podemos pensar que a janela $h(n)$ isola trechos consecutivos de amostras do sinal $x(n)$ para, em seguida, aplicarmos uma DFT de L amostras em cada segmento. O salto entres os inícios de cada trecho é determinado pelo parâmetro a: quanto maior

for a, maior será este intervalo. Uma decimação também é realizada no domínio da frequência: somente alguns *bins* da DFT de L amostras são retidos. O salto em frequência está relacionado ao parâmetro b: quanto maior for b, menor será o número de pontos armazenado. Se considerarmos que a janela $h(n)$ tem N amostras, o parâmetro b passa a ser determinado por $b = L/N$ [16].

Seguindo a teoria desenvolvida até aqui, a definição da Transformada de Fourier de Tempo Curto (ou STFT) é obtida por meio da decomposição do sinal $x(n)$ nos átomos $\{\phi_{k,j}(n)\}_{k,j}$ do dicionário, ou seja,

$$X(k,j) = c_{k,j} \triangleq \sum_{n=0}^{L-1} x(n)h(n-ka)\mathrm{e}^{-\mathrm{i}2\pi nbj/L}. \tag{2.14}$$

Para reconstruir o sinal $x(n)$ a partir dos seus coeficientes $\{c_{k,j}\}_{k,j}$, consideramos que estes formam um *frame* apertado e, assim, podemos utilizar a Eq. (2.7) para a síntese [16].

A ideia por trás da STFT é considerar que o sinal de áudio é estacionário[4] dentro do intervalo de tempo determinado por $h(n)$ e, portanto, a análise por DFT de cada trecho é representativa [8]. Assim sendo, o número de amostras N da janela exerce uma grande influência na representação tempo-frequencial resultante. Além disso, quando janelamos um trecho de sinal ao multiplicá-lo, no domínio do tempo, por uma versão deslocada de $h(n)$, estamos, no domínio da frequência, convoluindo os seus espectros [2]. Dessa forma, a informação frequencial do sinal que obtemos por meio da DFT é misturada com a da janela. É importante, portanto, entender estes efeitos de forma que, manipulando os parâmetros da transformada (número de amostras N, tipo de janela $h(n)$ e parâmetro a), sejamos capazes de amenizá-los ou levá-los em conta na interpretação do resultado.

Se o valor de N utilizado é relativamente pequeno, trechos mais curtos de sinal serão analisados por vez. Por um lado, informações temporais serão bastante precisas. Por outro, informações frequenciais (como altura das notas) irão se espalhar na representação. Da mesma forma, para um valor de N elevado, informações frequenciais serão mais bem representadas que as temporais. Dizemos, então, que para aumentar a resolução frequencial[5], devemos aumentar o número de amostras da janela de análise. Este compromisso é formalmente delimitado pelo Princípio da Incerteza [1].

Como foi dito, as resoluções temporal e frequencial da STFT dependem somente da janela $h(n)$ utilizada e, portanto, não variam de acordo com sua posição—ponto (k,j)—no plano da representação. Como explicado no Cap. 1, as parcelas "tonais"

[4]Um sinal é considerado estacionário se seu conteúdo frequencial não varia com o tempo.

[5]A resolução se opõe ao grau de espalhamento da representação: uma boa resolução frequencial implica linhas bem definidas ao longo do eixo temporal.

de sinais de música são geralmente modeladas por uma fundamental e suas parciais harmônicas superiores. Assumindo que a fundamental varia com o tempo, essa variação se torna mais acentuada para as parciais superiores e, consequentemente, o intervalo de tempo no qual podemos considerá-las estacionários diminui. O ideal é adequar a duração de $h(n)$ à duração do evento que queremos analisar, ou seja, utilizar um número de amostras N menor para altas frequências (resultando em maior resolução temporal) e maior para baixas (resultando em maior resolução frequencial) [2].

A STFT é amplamente utilizada na análise de sinais de música [11], apesar das limitações expostas. A ideia neste trabalho é justamente estudar representações alternativas para a análise destes sinais que procurem reduzir o efeito de tais limitações. A Transformada *Fan Chirp* (ou FChT), que será explicada no Cap. 3, leva em consideração que a fundamental pode variar dentro do trecho de sinal analisado e, assim, busca evitar perda de resolução nas parciais superiores. Já a Transformada de Q Constante, que será explicada a seguir, permite que *bins* de frequências diferentes possuam resoluções temporais diferentes, o que busca equilibrar a esparsidade da representação como um todo.

2.2.3 A Transformada de Q Constante

As frequências da escala de temperamento igual de 12 tons representam as alturas das notas musicais presentes na música ocidental. Por serem geometricamente espaçadas, não são, muitas vezes, corretamente mapeadas nas componentes frequenciais utilizadas na DFT, que apresentam um espaçamento linear [3]. A proposta da Transformada de Q Constante é fixar a razão entre frequência central e resolução e, assim, permitir que componentes distintos passem a ter resoluções diferentes.

Na DFT, cada *bin* j corresponde a uma frequência analógica $f_j = \frac{j}{N} f_\mathrm{s}$, sendo N o número de pontos da janela de análise (considerando o contexto da STFT) e f_s, a taxa de amostragem do sinal $x(n)$. Aqui, definimos a resolução do *bin* j como sendo o espaçamento entre as frequências correspondentes adjacentes, ou seja,

$$\begin{aligned} \Delta_j^{\mathrm{DFT}} &= f_{j+1} - f_j \\ &= \frac{j+1}{N} f_\mathrm{s} - \frac{j}{N} f_\mathrm{s} \\ &= \frac{f_\mathrm{s}}{N}. \end{aligned} \tag{2.15}$$

Percebemos que esta é constante, o que significa que, para um dado valor de N e uma nota com frequência f, a resolução da DFT irá corresponder a $\frac{f_\mathrm{s}}{fN} 100\%$ da frequência escolhida. Para frequências baixas (f pequeno), esse valor pode ser maior que a separação frequencial de 6% entre duas notas adjacentes da escala temperada

e levar, portanto, à perda de informação [3].

A CQT utiliza frequências centrais geometricamente espaçadas da forma

$$f_j = f_{\min} 2^{\frac{j}{B}}, \tag{2.16}$$

onde B é o número de *bins* por oitava, $f_{\min}$, a menor frequência considerada e f_j varia de $f_{\min}$ até uma frequência $f_{\max}$ abaixo da de Nyquist.

A resolução para a j-ésima componente frequencial passa a ser, portanto,

$$\begin{aligned} \Delta_j^{\mathrm{CQT}} &= f_{j+1} - f_j \\ &= f_{\min} 2^{\frac{j}{B}} (2^{\frac{1}{B}} - 1) \\ &= f_j (2^{\frac{1}{B}} - 1). \end{aligned} \tag{2.17}$$

Assim, a razão entre a frequência f_j e a sua resolução Δ_j^{CQT} é dada por

$$Q \triangleq \frac{f_j}{\Delta_j^{\mathrm{CQT}}} = (2^{\frac{1}{B}} - 1)^{-1}, \tag{2.18}$$

que é constante e, daí, o nome "Transformada de Q Constante". Com as escolhas apropriadas de $f_{\min}$ e B, frequências centrais f_j podem vir a corresponder a alturas de notas musicais em uma escala de temperamento igual.

A expressão para o cálculo da CQT pode ser obtida a partir da definição da STFT. Como vimos, a resolução dos *bins* da DFT é inversamente proporcional ao tamanho N da janela. Para que o valor de Q definido acima seja constante, é necessário, portanto, que N varie inversamente com a frequência central f_j, resultando em

$$N_j = \left\lceil \frac{f_{\mathrm{s}}}{f_j} Q \right\rceil. \tag{2.19}$$

Além disso, a frequência digital da j-ésima componente passa, agora, a ser dada por $2\pi Q/N_j$. A partir da Eq. (2.14), chegamos à definição da CQT dada por [3]:

$$X(k,j) = c_{k,j} \triangleq \frac{1}{N_j} \sum_{n=0}^{L-1} x(n) h_j(n - ka) \mathrm{e}^{-\mathrm{i}2\pi n Q/N_j}, \tag{2.20}$$

onde já realizamos a substituição do parâmetro b por $b = N/L$, $h_j(n)$ é a janela de análise com N_j amostras e $j = 0, 1, \ldots, \bar{J}$, com

$$\bar{J} = \left\lceil B \log_2 \left(\frac{f_{\max}}{f_{\min}} \right) \right\rceil. \tag{2.21}$$

Aqui, os átomos de tempo-frequência são dados por $\phi_{k,j}(n) = (1/N_j) h_j(n - ka) \mathrm{e}^{\mathrm{i}2\pi n Q/N_j}$. A divisão por N_j é uma normalização da janela $h_j(n)$, já que seu

número de amostras varia com o *bin* j.

De acordo com a definição dada pela Eq. (2.20), podemos interpretar que, assim como na STFT, o sinal é primeiramente dividido em trechos para, em seguida, aplicarmos a transformação. O salto entre trechos é, novamente, dado pelo parâmetro a. Consideramos, aqui, que cada trecho possui N_0 amostras, que é o número de amostras corresponde à analise do *bin* $j = 0$. Para cada *bin* de frequência há, no entanto, um tamanho de janela diferente. Como consequência, para as oitavas superiores, pode haver uma lacuna na análise se o valor de a não for suficientemente pequeno.

Em [14], o salto a entre os trechos de sinal é variável com o *bin* j de frequência. Dessa forma, além de solucionarmos o problema das lacunas na análise, reduzimos a complexidade computacional da transformada. Neste caso, menos instâncias da CQT para baixas frequências são necessárias, já que o número de amostras das janelas utilizadas é maior. Em contrapartida, a estrutura gerada por esse novo cálculo não pode ser organizada na forma de uma matriz, como na Eq. (2.13). Por isso, a solução adotada aqui é comprometer o tempo de processamento tornando a constante e com valores entre $0 < a \leq N_J/2$, onde N_J é o número de amostras da janela correspondente ao *bin* de maior frequência, em prol do alinhamento temporal de todos os *bins*.

2.3 Recuperação Esparsa

A ideia desta seção é apresentar os principais conceitos por trás das técnicas de recuperação esparsa de sinais de áudio. Aqui, queremos encontrar os coeficientes de expansão $\mathbf{c}$ a partir da decomposição do sinal $\mathbf{x}$ em um dicionário $\mathbf{\Phi}$ redundante. Neste caso, há infinitas possibilidades para determinar $\mathbf{c}$ e devemos definir, como mencionado anteriormente, um critério que nos ajude nessa escolha. O critério utilizado será o de esparsidade.

Agora, vamos utilizar um modelo para o sinal de áudio que admita a presença de ruído. Temos, assim,

$$\mathbf{x} = \mathbf{\Phi c} + \mathbf{e}, \tag{2.22}$$

onde $\mathbf{e}$ é o ruído aditivo. Recapitulando, temos que $\mathbf{x} \in \mathbb{R}^L$, $\mathbf{\Phi} \in \mathbb{C}^{L \times P}$, $\mathbf{c} \in \mathbb{C}^P$, com $P = JK$, e $\mathbf{e} \in \mathbb{R}^L$.

Definimos $\|\mathbf{c}\|_0 = \#\{l | c_l \neq 0\}$ como a cardinalidade l_0 de $\mathbf{c}$[6]. Esta pode ser interpretada como um critério de esparsidade, já que mede o número de componentes não-nulos de $\mathbf{c}$. Assim, quanto menor for $\|\mathbf{c}\|_0$, mais esparso será $\mathbf{c}$. Dizemos que $\mathbf{x}$

[6]Genericamente, definimos $\|\mathbf{c}\|_p = \left(\sum_l |c_l|^p \right)^{1/p}$ como sendo a norma l_p do vetor $\mathbf{c}$.

é uma recuperação esparsa no dicionário $\mathbf{\Phi}$ se $\|\mathbf{c}\|_0 \ll L$ [13].

Podemos, com isso, pensar em resolver o seguinte problema:

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{\Phi}\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_0 \right\}, \tag{2.23}$$

onde o parâmetro λ é maior que zero e controla o grau de regularização, ou seja, o compromisso entre o erro de aproximação e a esparsidade da solução. Quanto maior for λ, maior será o peso que o segundo termo irá exercer na otimização e obteremos, portanto, coeficientes $\hat{\mathbf{c}}$ mais esparsos.

Para encontrar uma solução $\hat{\mathbf{c}}$ podem ser utilizados algoritmos gananciosos, como o *Matching Pursuit* e suas variações [13]. Outra opção é relaxar o critério de esparsidade de $\mathbf{c}$, permitindo que componentes pequenos, porém não-nulos, não tenham grande influência na medida. Isso é feito substituindo-se a cardinalidade l_0 por uma função $f(\mathbf{c}) : \mathbb{C}^P \mapsto \mathbb{R}$ que meça esparsidade segundo este novo critério. A Eq. (2.23) se torna, então,

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{\Phi}\mathbf{c}\|_2^2 + \lambda f(\mathbf{c}) \right\}. \tag{2.24}$$

Normas l_p, com $0 < p \leq 1$, são boas opções para $f(\mathbf{c})$, já que estas promovem esparsidade [13]. A minimização da Eq. (2.24) para o caso específico em que $p = 1$ é conhecido na literatura como *Basis Pursuit* [17] ou LASSO (do inglês, *Least Absolute Shrinkage and Selection Operator*) [18]. Utilizar a norma ℓ_1 é conveniente pois, além de promover esparsidade, leva a um problema convexo e que, portanto, pode ser solucionado por técnicas já desenvolvidas de otimização convexa [13].

A seguir, descrevemos uma técnica chamada Limiarização Iterativa [19] que obtém os coeficientes $\hat{\mathbf{c}}$ para uma função $f(\mathbf{c})$ genérica. Analisamos, também, a solução para $p = 1$.

2.3.1 Limiarização Iterativa

Abordamos, primeiramente, o caso em que o dicionário $\mathbf{\Phi}$ forma uma base ortonormal para, em seguida, extrapolarmos o resultado obtido para o caso sobrecompleto. O primeiro passo é reescrever a Eq. (2.24) da forma

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \frac{1}{2}\|\mathbf{\Phi}(\mathbf{\Phi}^{-1}\mathbf{x} - \mathbf{c})\|_2^2 + \lambda f(\mathbf{c}) \right\}. \tag{2.25}$$

Já supondo que $\mathbf{\Phi}$ forma uma base ortonormal, utilizamos o teorema de Parseval e a igualdade $\mathbf{\Phi^{-1}} = \mathbf{\Phi}^{\mathrm{H}}$ para obter

$$\begin{aligned} \hat{\mathbf{c}} &= \arg\min_{\mathbf{c}} \left\{ \frac{1}{2} \|\mathbf{\Phi}^{\mathrm{H}}\mathbf{x} - \mathbf{c}\|_2^2 + \lambda f(\mathbf{c}) \right\} \\ &= \arg\min_{\mathbf{c}} \left\{ \sum_l \left[\frac{1}{2} (\boldsymbol{\phi}_l^{\mathrm{H}}\mathbf{x} - c_l)^2 + \lambda f(c_l) \right] \right\}, \end{aligned} \tag{2.26}$$

onde escrevemos $\mathbf{\Phi}$ em função das suas colunas $\boldsymbol{\phi}_l$ como na Eq. (2.2). Além disso, consideramos que a função $f(\mathbf{c})$ é aplicada a cada elemento c_l de $\mathbf{c}$ separadamente.

Pela Eq. (2.26), podemos perceber que o problema original de otimização dado pela Eq. (2.24) foi fatorado em problemas escalares que podem, portanto, ser solucionados individualmente. Derivando a expressão acima em relação a c_l e igualando o resultado a zero, temos

$$-\boldsymbol{\phi}_l^{\mathrm{H}}\mathbf{x} + \hat{c}_l + \lambda \frac{\mathrm{d}f(\hat{c}_l)}{\mathrm{d}\hat{c}_l} = 0, \tag{2.27}$$

o que nos leva a

$$\hat{c}_l + \lambda \frac{\mathrm{d}f(\hat{c}_l)}{\mathrm{d}\hat{c}_l} = \boldsymbol{\phi}_l^{\mathrm{H}}\mathbf{x}. \tag{2.28}$$

Aplicando a transformação $\mathbb{S}_\lambda(\cdot)$, com

$$\mathbb{S}_\lambda^{-1}(u) = u + \lambda \frac{\mathrm{d}f(u)}{\mathrm{d}u}, \tag{2.29}$$

dos dois lados da equação, obtemos

$$\hat{c}_l = \mathbb{S}_\lambda(\boldsymbol{\phi}_l^{\mathrm{H}}\mathbf{x}). \tag{2.30}$$

Para o caso em que $p = 1$, temos

$$\mathbb{S}_\lambda^{-1}(u) = u + \lambda \mathrm{sign}(u) \tag{2.31}$$

e, portanto,

$$\mathbb{S}_\lambda(u) = \mathrm{sign}(u) \max(0, |u| - \lambda). \tag{2.32}$$

Este método chama-se limiarização suave [13]. Os valores abaixo de um limiar λ são zerados e os demais são gradualmente ponderados a partir de zero, evitando descontinuidades. No limite em que $p \to 0$, a limiarização torna-se rígida, o que significa que valores acima do limar são mantidos, enquanto que os valores abaixo, descartados. Isso é ilustrado na Fig. 2.1.

Para realizar a limiarização com um dicionário sobrecompleto, um algoritmo

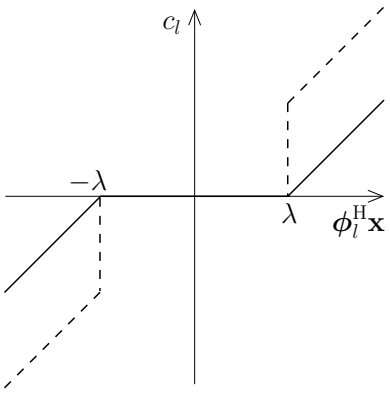


Figura 2.1: Curvas de limiarização $c_l = \mathbb{S}_\lambda(\boldsymbol{\phi}_l^{\mathrm{H}}\mathbf{x})$ para os casos $p = 1$ e $p \to 0$ (curva tracejada).

iterativo, chamado de limiarização iterativa [19], foi proposto. A ideia é, novamente, separar o problema de otimização original em problemas unidimensionais para, em seguida, tratá-los individualmente.

Assumimos que os coeficientes na n-ésima iteração são dados por $\mathbf{c}^{(n)}$ e que todos os seus elementos estão fixos, com a exceção da l-ésima entrada $c_l^{(n)}$. Queremos, então, estimar um novo valor w tal que $w = c_l^{(n+1)}$. Para isso, minimizamos

$$\hat{w} = \arg\min_w \left\{ \frac{1}{2}\|\mathbf{x} - (\boldsymbol{\Phi}\mathbf{c}^{(n)} - \boldsymbol{\phi}_l c_l^{(n)} + \boldsymbol{\phi}_l w)\|_2^2 + \lambda f(w) \right\}, \tag{2.33}$$

onde retiramos a contribuição do $l-$ésimo elemento $c_l^{(n)}$ e acrescentamos a nova estimativa w. Derivando a expressão em relação a w e igualando o resultado a zero, obtemos

$$\hat{w} + \lambda \frac{\mathrm{d}f(\hat{w})}{\mathrm{d}\hat{w}} = c_l^{(n)} + \boldsymbol{\phi}_l^{\mathrm{H}}(\mathbf{x} - \boldsymbol{\Phi}\mathbf{c}^{(n)}), \tag{2.34}$$

onde consideramos que $\|\boldsymbol{\phi}_l\|_2^2 = 1$. É possível perceber que a equação acima é da mesma forma que a Eq. (2.28) e pode, portanto, ser resolvida por um operador $\mathbb{S}_\lambda(\cdot)$ de limiarização. Para cada iteração, esse cálculo é realizado para todos os elementos l de $\mathbf{c}$. Esse processo é repetido até a convergência.

O algoritmo de limiarização iterativa discutido aqui [19] é uma das possíveis abordagens para solucionar este problema. Este, no entanto, apresenta uma convergência lenta, já que requer que cada elemento c_l seja atualizado por vez [13]. Uma solução é paralelizar o cálculo dos coeficientes de $\mathbf{c}$, como é feito em [20]. Este algoritmo chama-se ISTA (do inglês, *Iterated Soft-Thresholding Algorithm*). A sua versão rápida, chamada FISTA (do inglês, *Fast* ISTA) e ilustrada em [21], será a utilizada neste trabalho.

2.3.2 Aproximações Estruturadas

No procedimento apresentado até agora, não levamos em consideração interdependências dos elementos de $\mathbf{c}$ no processo de minimização. Este, no entanto, não é o caso para muito sinais. Para sinais de áudio, por exemplo, vemos claramente um padrão vertical (ou horizontal) na sua representação tempo-frequência quando o sinal é composto prioritariamente por parcelas aproximadamente senoidais (ou aproximadamente impulsivas, respectivamente). A ideia aqui é incorporar informações sobre propriedades estruturais dos coeficientes $\mathbf{c}$ por meio do conceito de vizinhança [12], que será apresentado no Capítulo 4.

2.4 Considerações Finais

Neste capítulo apresentamos os principais conceitos matemáticos importantes para o entendimento da dissertação em geral.

Abordamos a teoria de decomposições atômicas e descrevemos duas ferramentas amplamente utilizadas na análise de sinais de áudio dentro deste contexto. São elas a Transformada de Fourier de Tempo Curto (ou STFT) e a Transformada de Q Constante (ou CQT). Tratamos, também, das principais diferenças entre as duas transformadas.

Em seguida, introduzimos a teoria por trás da recuperação esparsa de sinais a partir de dicionários sobrecompletos. Falamos do conceito de limiarização suave e do algoritmo de limiarização iterativa para representações redundantes. Por fim, abordamos brevemente a ideia das aproximações estruturadas.

Capítulo 3

A Transformada *Fan Chirp*

3.1 Considerações Iniciais

No capítulo anterior, falamos sobre as principais técnicas utilizadas na representação tempo-frequencial de sinais de áudio e suas limitações. No caso da STFT, como foi dito, consideramos que o sinal é estacionário para intervalos curtos de tempo. Assim sendo, uma análise por meio da DFT seria capaz de representar de forma acurada as componentes frequenciais de cada trecho. Essa aproximação, no entanto, pode não ser apropriada para representar sinais de música, que podem apresentar variações rápidas significativas em frequência.

A Transformada *Fan Chirp* (FChT), inicialmente apresentada em [22], propõe modelar a frequência fundamental (juntamente com as parciais harmônicas superiores) de um sinal como uma função linear no tempo. Em [8], a FChT foi aplicada a sinais de música por meio da Transformada *Fan Chirp* de Tempo Curto (STFChT). Com ela, agora consideramos que, para intervalos de curta duração, a fundamental do sinal pode variar linearmente com o tempo. Essa mudança permite representações tempo-frequenciais mais esparsas, especialmente nas parciais superiores.

Uma extensão natural do método é permitir que a frequência fundamental do sinal varie quadraticamente com o tempo [23]. Essa mudança é especialmente relevante para sinais de áudio cujas variações em frequência são acentuadas em curtos períodos de tempo, como é o caso de vibratos rápidos e canto tirolês (iodelei, do alemão, *Jodeln*). O modelo não linear leva a uma maior esparsidade na representação e, assim, possibilita aperfeiçoar técnicas como sistemas de detecção de melodia, redução de ruído e separação de fontes.

Para que possamos modelar a fundamental como uma função (linear ou não), precisamos de informação a priori sobre a sua variação ao longo do tempo. Essa informação é obtida por meio do cálculo da saliência para cada possível fundamental presente no sinal [8]. O procedimento utilizado é descrito em [24]. A ideia aqui é,

para cada frequência candidata, acumular a energia do espectro nas frequências correspondentes às suas parciais harmônicas. Assim sendo, espera-se que os maiores picos equivalham às verdadeiras fundamentais do sinal.

Essa abordagem é a mais comumente utilizada na literatura. Porém, como utiliza informação de energia, ela apresenta sensibilidade ao timbre das fontes sonoras existentes no sinal. Em [25], o cálculo da saliência se baseia apenas nas localizações frequências das possíveis fontes, obtidas por meio de estimação de picos no espectro. Cada pico é avaliado e classificado como uma nova fundamental ou uma parcial de uma fundamental existente. A proposta aqui é adotar esta abordagem em vez da clássica utilizada na implementação original da FChT.

Para calcular a saliência das possíveis fundamentais, consideramos que as parciais correspondentes estão localizadas em múltiplos inteiros das suas frequências. No entanto, instrumentos de corda, como o piano, apresentam um leve desvio de frequência, perceptível, principalmente, nas parciais superiores. Esta inarmonicidade está relacionada às características físicas de uma corda não ideal [26]. Assim sendo, um modelo de inarmonicidade pode ser utilizado para melhorar a resolução da representação tempo-frequencial de tais instrumentos.

Neste Capítulo iremos, primeiramente, apresentar a definição formal da FChT de acordo com [8]. Em seguida, as mudanças propostas serão abordadas, sempre em conjunto com exemplos de sinais sintéticos para evidenciar as melhorias desejadas. Na ordem, temos: a extensão do método existente para permitir que a frequência fundamental do sinal varie quadraticamente com o tempo; uma nova forma para o cálculo da função de saliência que não leva em consideração o timbre do instrumento; e a inserção de um modelo de inarmonicidade para sinais de áudio. Por fim, apresentaremos os resultados obtidos para sinais de áudio reais.

3.2 Definição

A Transformada *Fan Chirp* de um sinal contínuo no tempo $x(t)$ é definida em [8] como

$$X(f,\alpha) \triangleq \int_{-\infty}^{\infty} x(t)\psi'_{\alpha}(t)\mathrm{e}^{-\mathrm{i}2\pi f\psi_{\alpha}(t)}dt, \tag{3.1}$$

onde $\psi_{\alpha}(t)$ é uma função de deformação do tempo dada por

$$\psi_{\alpha}(t) = \left(1 + \frac{1}{2}\alpha t\right)t. \tag{3.2}$$

Se fizermos a mudança de variáveis $\tau = \psi_\alpha(t)$ na Eq. (3.1), temos

$$X(f,\alpha) = \int_{-1/\alpha}^{\infty} x(\psi_\alpha^{-1}(\tau))\mathrm{e}^{-\mathrm{i}2\pi f\tau}d\tau, \tag{3.3}$$

onde $\psi_\alpha^{-1}(t)$ é dado por

$$\psi_\alpha^{-1}(t) = -\frac{1}{\alpha} + \frac{\sqrt{1+2\alpha t}}{\alpha} \tag{3.4}$$

e foi assumido que $x(t) = 0$ para $t \leq -1/\alpha$ para evitar *aliasing* [22]. Da Eq. (3.3), é possível perceber que a FChT é a Transformada de Fourier do sinal $x(t)$ deformado no tempo pela função $\psi_\alpha^{-1}(t)$. Dessa forma, após uma discretização temporal apropriada, a FChT pode ser calculada através de alguma implementação rápida da Transformada de Fourier (algoritmo FFT) [8].

Pela Eq. (3.1), podemos dizer que a FChT consiste na decomposição do sinal $x(t)$ em um dicionário composto por elementos $\{\phi_k(t)\}_k$, onde $\phi_k(t) = \psi'_\alpha(t)\mathrm{e}^{\mathrm{i}2\pi\frac{k}{T}\psi_\alpha(t)}$, com $k \in \mathbb{Z}$. Aqui, consideramos uma grade discreta de frequências e que $\phi_k(t)$ possui suporte compacto localizado em $t \in \left[\psi_\alpha^{-1}(-\frac{T}{2}), \psi_\alpha^{-1}(\frac{T}{2})\right]$. Calculando a projeção de um *chirp* linear $x_\mathrm{c}(t, f_0) = \mathrm{e}^{\mathrm{i}2\pi f_0\psi_\alpha(t)}$ em $\phi_k(t)$, obtemos[1]

$$\begin{aligned}
\langle x_\mathrm{c}(t, l/T), \phi_k(t)\rangle &= \frac{1}{T}\int_{\psi_\alpha^{-1}(-\frac{T}{2})}^{\psi_\alpha^{-1}(\frac{T}{2})} \psi'_\alpha \mathrm{e}^{\mathrm{i}2\pi\frac{l-k}{T}\psi_\alpha(t)}dt \\
&= \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} \mathrm{e}^{\mathrm{i}2\pi\frac{l-k}{T}\tau}d\tau \\
&= \delta(l-k),
\end{aligned} \tag{3.5}$$

com $l \in \mathbb{Z}$ .Isso significa que a FChT é capaz de representar *chirps* lineares de forma maximamente esparsa [8].

É importante ressaltar que quando o sinal $x(t)$ é janelado, o resultado presente na Eq. (3.5) deixa de ser válido. Como na prática é isto que ocorre, queremos que, dada uma janela $h(t)$, a Transformada *Fan Chirp* de $x_\mathrm{c}(t, f_0)$ seja igual a um impulso convoluído com a Transformada de Fourier de $h(t)$. Assim sendo, para permitir esta condição, redefinimos a FChT de um sinal $x(t)$ como [8]

$$X_h(f,\alpha) = \int_{-\infty}^{\infty} x(t)h(\psi_\alpha(t))\psi'_\alpha(t)\mathrm{e}^{-\mathrm{i}2\pi f\psi_\alpha(t)}dt. \tag{3.6}$$

Consideramos, agora, um sinal constituído por L chirps harmônicos, ou seja

[1]Note que para que esta expressão seja verdadeira, é necessário considerar que $T \leq \frac{1}{\alpha}$.

$x_{\mathrm{ch}}(t, f_0, L) = \sum_{l=1}^{L} \mathrm{e}^{\mathrm{i}2\pi l f_0 \psi_\alpha(t)}$. Percebemos que todos os componentes possuem a mesma taxa de inclinação α e, portanto, suas representações por meio da FChT serão igualmente esparsas. Assim sendo, a geometria da transformada parece adequada para representar sinais de áudio, contanto que estes sejam compostos, essencialmente, por uma fundamental e parciais harmônicas superiores.

3.3 Implementação

A implementação é feita em trechos curtos (variando entre 20 e 100 ms) consecutivos do sinal de análise amostrado, $x(n)$. Esse procedimento é denominado de Transformada *Fan Chirp* de Tempo Curto (STFChT). Assim, como já foi dito, tomamos como válida a aproximação da fundamental do sinal por um *chirp* linear dentro de cada janela.

O primeiro passo na implementação da FChT é a deformação linear causada por $\psi_\alpha^{-1}(t)$, como visto na Eq. (3.3). Agora, estamos lidando com sinais discretos no tempo e, portanto, este passo é implementado por meio de uma reamostragem não-linear. Como só temos acesso às amostras nos instantes de tempo nT_s, onde T_s é o período de amostragem, uma interpolação é realizada [8]. Note que, aqui, é necessário um valor pré-estabelecido do parâmetro α. A seguir, a FFT do sinal deformado no tempo é calculada.

A etapa de estimação do parâmetro α é essencial, já que a resolução da representação resultante depende de uma estimativa precisa dele. Caso contrário, os coeficientes da representação tendem a se espalhar, pois o modelo utilizado já não é mais válido para o (trecho do) sinal de análise. Em outras palavras, a projeção de (um trecho de) $x(n)$ sobre os *chirps* lineares não será esparsa, já que não será mais possível representá-lo com apenas um (ou poucos) elementos da base.

Para estimar α, realizamos uma busca exaustiva por valores possíveis de frequência fundamental f_0 e taxa de inclinação α. Uma função $\rho(f_0, \alpha)$, chamada de plano de saliência, é criada para auxiliar neste processo. Esta função consiste em uma acumulação da energia do espectro[2] nas frequências harmônicas de f_0 para cada par (f_0, α) possível. Se f_0 corresponder à uma fundamental existente, então a energia em suas parciais harmônicas é significativa e, consequentemente, um maior valor de $\rho(f_0, \alpha)$ é esperado. Da mesma forma, para o valor correto de α, a esparsidade do espectro será máxima e, consequentemente, o valor de $\rho(f_0, \alpha)$. Para outros valores de α, a energia nas parciais harmônicas, que antes era bem concentrada, se espalharia entre *bins* de frequência adjacentes.

O procedimento para o cálculo de α, retirado de [8], é descrito a seguir.

[2]Aqui, consideramos o espectro como sendo a magnitude da FChT do sinal de análise.

1. Primeiramente, várias instâncias da FChT para valores pré-determinados de α são calculadas.

2. A seguir, um *grid* de frequências fundamentais é definido e, para cada frequência f_0 e cada FChT calculada no passo anterior, a soma dos logaritmos das amplitudes das suas parciais harmônicas é calculada como em[3]

$$\rho(f_0, \alpha) = \frac{1}{n_{\mathrm{H}}} \sum_{i=1}^{n_{\mathrm{H}}} \log |X(if_0, \alpha)|, \tag{3.7}$$

   onde $X(f, \alpha)$ é a FChT e n_{H} é um número pré-determinado de parciais. Nesta parte, uma etapa de pós-processamento é realizada de forma a minimizar a ambiguidade gerada por múltiplos e submúltiplos das frequências fundamentais presentes no sinal [8].

3. Teremos, assim, um plano denso $\rho(f_0, \alpha)$ que irá concentrar energia em alguns pontos (f_0, α), representando a presença de uma fonte de áudio com frequência fundamental f_0 subindo (ou descendo) a uma taxa de $|\alpha|$ Hz.

4. Selecionamos, para cada frequência f_0, o maior valor de $\rho(f_0, \alpha)$, resultando em uma função de saliência $\bar{\rho}(f_0)$. Os picos de $\bar{\rho}(f_0)$ representam, como foi dito no item anterior, as fontes sonoras. Destes, podemos estimar o parâmetro α para cada fonte.

Na prática, não há um valor correto de α, já que o modelo proposto é uma aproximação de primeira ordem da frequência fundamental. Além disso, por razões de complexidade, apenas um número finito de valores é testado. Essas simplificações introduzem pequenos erros na estimação de α.

É importante ressaltar também que, no caso de um sinal polifônico, só é possível selecionar um valor de α dentre os obtidos. Dessa forma, a FChT calculada é ajustada para representar com maior precisão somente uma (a mais proeminente) das fontes harmônicas presentes, prejudicando a representação das demais [8].

A representação tempo-frequencial desejada é proporcionada pela STFChT, que é obtida pela concatenação das FChTs dos segmentos de sinal previamente calculadas. Ao concatenar as funções de saliência $\bar{\rho}(f_0)$ de cada trecho, geramos uma representação tempo-frequencial alternativa conhecida como F0grama, que mostra a evolução temporal do *pitch* de todos os tons harmônicos presentes em um sinal de música [8]. Essa representação pode fornecer informações relevantes sobre os valores de α estimados.

[3]Na prática, a implementação é feita substituindo-se o termo $\log |X(if_0, \alpha)|$ por $\log(\gamma |X(if_0, \alpha)| + 1)$, onde inserimos um controle de esparsidade por meio do parâmetro γ [8]. Os resultados reportados correspondem a $\gamma = 10$.

3.4 Deformação Não-Linear

3.4.1 Proposta

A frequência fundamental de um sinal de música pode, às vezes, apresentar rápidas flutuações em um curto período de tempo. Nesse caso, a sua aproximação por uma função linear não seria apropriada, enquanto escolher um modelo não-linear que permitisse modelar maiores variações em frequência poderia resultar em uma transformação mais esparsa. Aqui, escolhemos um polinômio de segunda ordem para aproximar a fundamental.[4]

A Fig. 3.1 mostra um exemplo da variação da frequência fundamental ao longo do tempo da melodia principal de um excerto de ópera, onde zeros representam ausência de notas. É possível notar que o sinal analisado apresenta rápidas flutuações em frequência e, como mencionado anteriormente, uma função linear pode não ser apropriada para modelar tais variações. Estas são caracterizadas, neste caso, pela presença de vibratos, como nos instantes $t \approx 2,7\,\text{s}$ e $t \approx 1,4\,\text{s}$, e mudanças contínuas entre notas, como nos instantes $t \approx 2,3\,\text{s}$ e $t \approx 5,1\,\text{s}$. Naturalmente, as variações ocasionadas pelas interrupções da melodia principal, como nos instantes $t \approx 1,3\,\text{s}$ e $t \approx 3,6\,\text{s}$, não são de interesse para a representação.

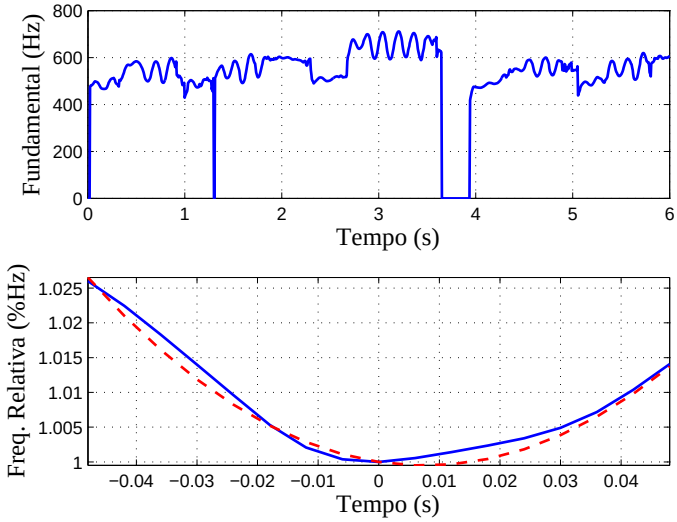


Figura 3.1: Frequência fundamental de um sinal de ópera (gráfico acima) e a aproximação quadrática, tracejada em vermelho, para um trecho de 100 ms de sinal de áudio, em azul (gráfico abaixo).

Agora, adicionamos um terceiro termo à Eq. (3.2) de forma a melhorar a representação da evolução da frequência fundamental do sinal de análise. Isso significa que as componentes frequenciais presentes no sinal poderão se comportar como um

[4] A ideia original de utilizar uma deformação não-linear se deve ao professor Pablo Cancela, docente da *Universidad de la República* (UDELAR), Uruguai. O desenvolvimento desta seção foi realizado em conjunto ao Grupo de Procesamiento de Áudio (GPA) daquela universidade.

chirp quadrático. O novo parâmetro β, chamado de curvatura, será o responsável por essa mudança. A Eq. (3.2) se torna, então:

$$\psi_{\alpha,\beta}(t) = \left(1 + \frac{1}{2}\alpha t + \frac{1}{3}\beta t^2\right) t. \tag{3.8}$$

Note que, por essa definição, a frequência instantânea passa a ser dada por

$$\nu(t) = f\frac{\mathrm{d}}{\mathrm{d}t}\psi_{\alpha,\beta}(t) = (1 + \alpha t + \beta t^2)f, \tag{3.9}$$

o que mostra que a frequência fundamental passará a ser aproximada por um polinômio de segunda ordem.

No caso da deformação linear, assumimos que o sinal $x(t) = 0$ para $t \leq \alpha$ para evitar *aliasing*. Neste caso, chegamos a tal conclusão ao analisarmos a Eq. (3.3), obtida por meio da inversão de $\psi_\alpha(t)$, dada pela Eq. (3.4). Quando passamos para a deformação não-linear, já não temos uma fórmula fechada para a expressão inversa de $\psi_{\alpha,\beta}(t)$. Futuramente devemos realizar uma análise mais detalhada das restrições impostas ao sinal $x(t)$ para este caso de forma a evitar o *aliasing*.

O próximo passo é procurar pelos valores numéricos dos parâmetros α e β que são adequados para a representação de sinais de áudio. Em seguida, devemos incorporar o modelo não-linear à implementação existente da FChT e analisar as mudanças causadas na transformação tempo-frequencial resultante.

3.4.2 Amostragem dos Parâmetros

Como foi dito, devemos primeiramente determinar quais são os possíveis valores que α e β podem assumir. Para isso, foi utilizada uma base de dados do MIREX [27] que contém excertos de áudio polifônico. Utilizamos, no total, 33 sinais desta base. Para cada sinal, a frequência fundamental da melodia principal foi extraída manualmente (um exemplo é o ilustrado na Fig. 3.1). Estes foram divididos em trechos de 100 ms[5] e, em seguida, foi realizado um ajuste polinomial para encontrar os valores de α e β. É importante ressaltar que a frequência fundamental deve ser normalizada no centro do trecho (ou seja, $t = 0\,\mathrm{s}$) antes do cálculo dos parâmetros, como pode ser visto na Fig. 3.1 (gráfico inferior).

O conjunto de pares (α, β) obtido foi utilizado para construir um histograma. O alcance dos parâmetros selecionado foi de $[-4, 4]$ para α e de $[-50, 50]$ para β e o número de *bins* foi 22 para ambos. O resultado pode ser visto na Fig. 3.2.

[5]Como estamos interessados em acompanhar variações de *pitch* típicas de sinais de áudio, um intervalo de tempo consideravelmente maior que o padrão de 20 ms foi escolhido. Assim, conseguimos encontrar valores mais significativos de β. Esperamos, com isso, ser capazes de utilizar trechos de sinal maiores para o caso da deformação não-linear, diminuindo o espalhamento causado pelo janelamento do sinal.

Percebemos que a maioria dos valores de α e β se concentra no ponto $(0, 0)$, mas que ainda há uma quantidade considerável de energia no entorno.

A partir do histograma gerado, podem ser propostos diferentes tipos de amostragem no espaço de parâmetros com o objetivo de representar os valores mais significativos de α e β. Três exemplos podem ser vistos na Fig. 3.2. No primeiro caso, a amostragem consiste em 23 pontos representando as duas principais direções do plano (α, β): $\alpha = 0$ e $\beta = 0$. No segundo caso, a amostragem consiste da primeira adicionada de 12 pontos ao redor da origem $(0, 0)$. No terceiro caso, a amostragem consiste em uma elipse de 175 pontos em torno da origem englobando 90% dos valores de α e β. Para o cálculo da FChT realizamos uma busca exaustiva entre todas as possibilidades de parâmetros e, assim, o número de pontos na amostragem está diretamente relacionada ao custo computacional. Isso determina a adoção de um número restrito de valores para ambos os parâmetros.

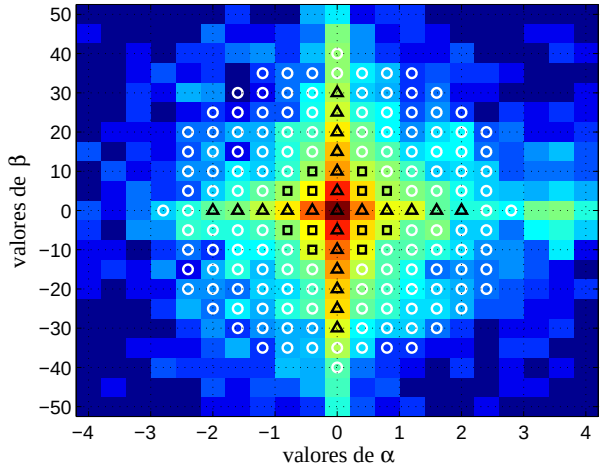


Figura 3.2: Exemplos de estratégias de amostragem: 1) a cruz formada por triângulos pretos; 2) os pontos da cruz adicionados dos pontos indicados pelos quadrados pretos; e 3) as duas primeiras amostragens adicionadas dos pontos indicados pelos círculos brancos. O histograma é exibido em escala logarítmica.

Após a escolha de uma amostragem adequada, é possível prosseguir para a implementação.

3.4.3 Implementação

A implementação da FChT com deformação não-linear continua sendo na forma de uma busca exaustiva. Agora, porém, temos que realizar a busca para um conjunto de pares (α, β). Podemos pensar que esses pares formam uma matriz $m \times n$, com $m = n = 22\, bins$ e reorganizá-la como um vetor linha de $1 \times nm$. Em seguida, realizamos o mesmo procedimento explicado na Seção 3.3, onde calculamos um plano denso $\rho(f_0, \gamma)$, sendo γ o índice do vetor linha obtido.

3.4.4 Experimentos e resultados parciais

Aqui, iremos aplicar o modelo não-linear proposto em um sinal sintético. Conhecendo o comportamento do sinal gerado, podemos analisar o resultado obtido e compará-lo ao teórico esperado. A ideia é justamente motivar o uso da modificação proposta. A aplicação em sinais reais será feita na Seção. 3.7.

Utilizamos um sinal sintético harmônico modulado em frequência por uma senoide. O sinal possui, no total, 9 parciais harmônicas. Sua frequência fundamental $f_0(t)$ é dada pela seguinte lei de formação:

$$f_0(t) = f_1(1 - 2^{1/12})\text{sen}(2\pi f_2 t) + f_1,$$

onde f_1 é a frequência central e f_2, a frequência de modulação. A expressão resultante para o sinal $x(t)$ é dada por

$$x(t) = \sum_{h=1}^{9} \cos\left(2\pi f_0(t) h\right). \tag{3.10}$$

Os seus valores foram escolhidos de forma a imitar um vibrato como geralmente encontrado em interpretações de voz cantada: 500 Hz e 6 Hz, respectivamente. A frequência de amostragem é de 44100 Hz.

A Fig. 3.3 mostra a magnitude das STFChTs do sinal escolhido para as deformações linear (segunda coluna) e não-linear (terceira coluna). A STFT para o mesmo sinal também pode ser visto (primeira coluna). Três tamanhos de janela de análise foram escolhidos para comparação: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras. Isso é feito de forma a permitir uma comparação mais fina entre os métodos.

É possível perceber o efeito do tamanho escolhido da janela na resolução das frequências altas e baixas na STFT. Aumentar o tamanho da janela resulta em uma maior resolução frequencial, notada como uma melhora na representação das parciais harmônicas inferiores, que variam mais lentamente com o tempo. Por outro lado, diminuir o tamanho da janela resulta em uma maior resolução temporal, notada como uma melhora na representação das parciais harmônicas superiores, que variam mais rapidamente com o tempo. Esta é uma questão clássica quando lidamos com a STFT[6] e, como pode ser visto, seus efeitos são suavizados ao utilizarmos a STFChT. De fato, é desejável ter o maior tamanho de janela possível quando utilizamos a STFChT, já que isso significa um menor espalhamento em frequência devido ao processo de janelamento. O limite superior é definido pelo tipo de deformação escolhido. O tamanho da janela pode aumentar contanto que as variações frequenciais presentes

[6]Como foi dito no Cap. 2, chama-se Princípio da Incerteza.

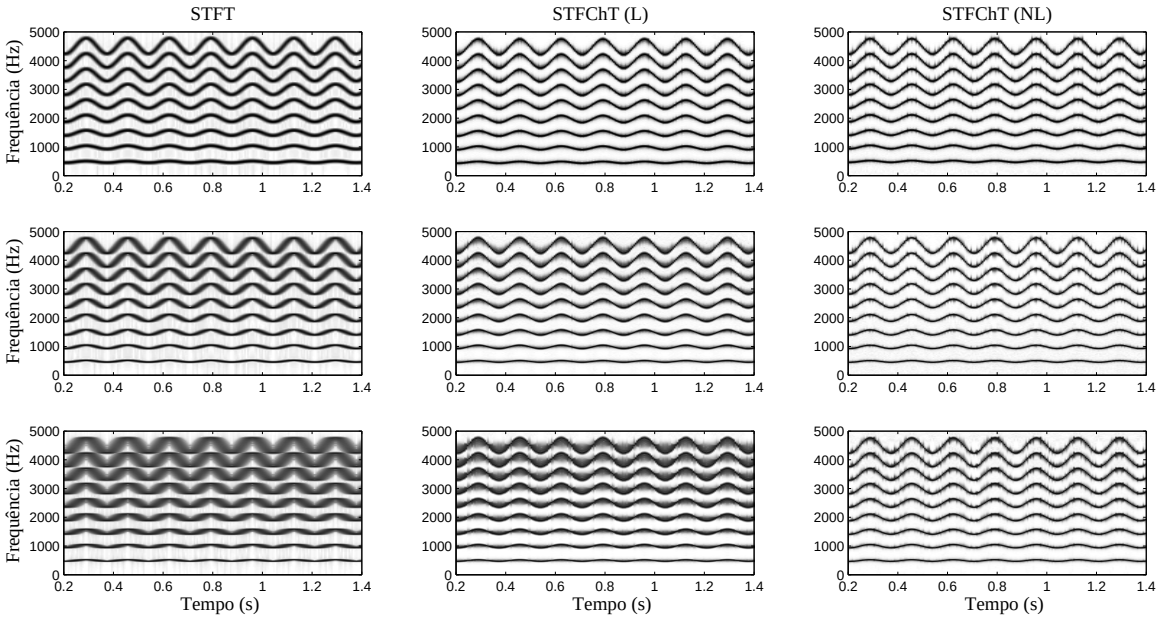


Figura 3.3: STFT (primeira coluna), STFChTs com deformação linear (segunda coluna, L) e não-linear (terceira coluna, NL) de um sinal sintético. Os seguintes tamanhos de janela de análise foram utilizados: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras.

no intervalo de tempo definido ainda sejam aproximadas de forma adequada pelo modelo selecionado. Por exemplo, quando aumentamos o tamanho da janela para 4096 amostras, a deformação linear já não é capaz de modelar a evolução das parciais harmônicas e a piora na resolução pode ser vista especialmente nas regiões de alta curvatura. Quando usamos a deformação não-linear, temos um grau de liberdade a mais para modelar as variações da frequência fundamental e podemos, assim, permitir que a janela de análise tenha um maior número de amostras. No entanto, é importante ressaltar que o cálculo da STFChT é dependente do parâmetro α (e β, para o caso não linear). Estes devem ser corretamente estimados de forma a gerar uma representação esparsa.

A representação obtida com a deformação não-linear, como vista na Fig. 3.3, possui uma maior esparsidade se comparada à obtida com a deformação linear, especialmente ao redor de picos e vales dos contornos das parciais. É evidente, assim, que um modelo de ordem maior é capaz de aproximar as variações em frequência com mais detalhe. Alguns artefatos podem ser vistos nesta representação por causa da estimação de α e β em um conjunto discreto de valores.

Tal comportamento também pode visualizado nos F0gramas correspondentes, mostrados na Fig. 3.4. Como anteriormente, as representações obtidas com deformações linear (coluna da esquerda) e não-linear (coluna da direita) foram geradas utilizando tamanhos de janela de 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras. A fundamental, estimada como sendo a frequência

correspondente ao máximo do F0grama para cada instante de tempo, é ilustrada em branco por cima do F0grama original. A resolução tempo-frequencial da representação tem um impacto considerável na acurácia da estimação da frequência fundamental. Em particular, para o caso com deformação linear e tamanho de janela de 4096 amostras, o método falha em seguir corretamente o contorno da frequência fundamental real.

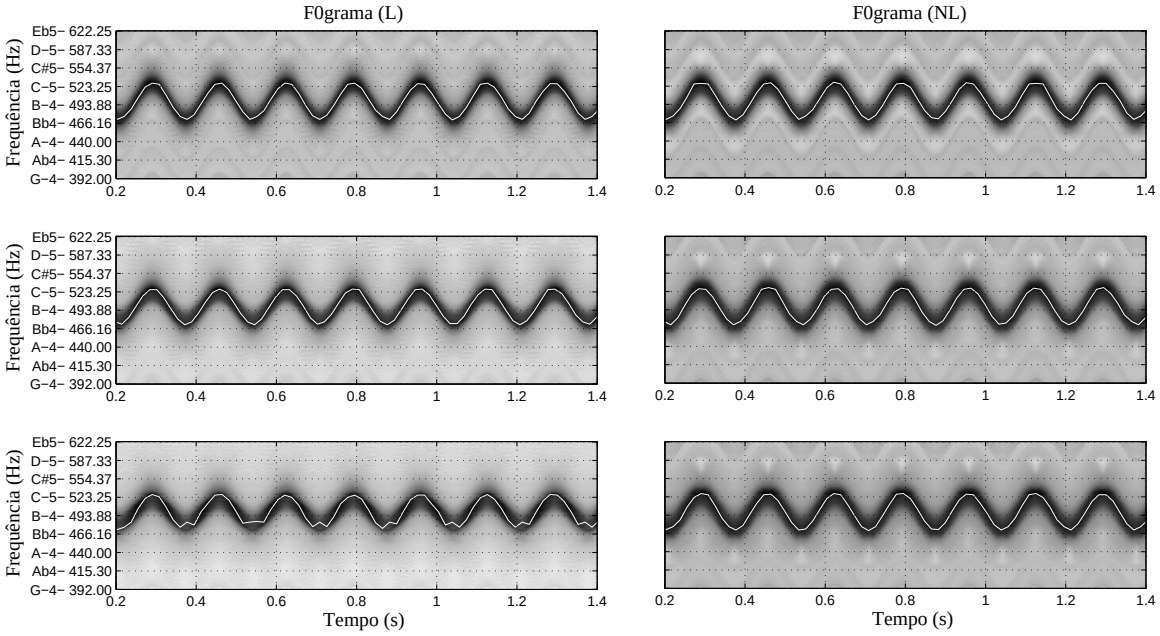


Figura 3.4: F0gramas para as STFChTs com deformação linear (primeira coluna, L) e não-linear (segunda coluna, NL) de um sinal sintético. Os seguintes tamanhos de janela de análise foram utilizados: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras.

3.5 Função de Saliência Independente do Timbre

3.5.1 Proposta

A função de saliência, dada uma taxa de inclinação α, até então utilizada é calculada pela Eq. (3.7). Como foi dito, para uma frequência fundamental f_0, o resultado será, essencialmente, a soma da energia das FChTs nas frequências harmônicas de f_0. Como consequência, o valor obtido será dependente do timbre do instrumento [25], já que este está diretamente relacionado com a energia presente nas parciais harmônicas. O logaritmo presente na expressão procura "planificar" o espectro $X(f, \alpha)$ de forma a torná-lo mais robusto às variações de timbre.

Aqui, partimos de uma função de saliência alternativa, retirada de [25]. O cálculo, agora, deixa de considerar a informação de energia e passa a utilizar informação de localização frequencial. Para isso, os picos do espectro são estimados e,

em seguida, a possibilidade de cada um deles ser uma parcial de alguma fundamental já existente é avaliada. Isso é feito definindo uma medida de desvios teóricos em relação às frequências das notas de uma escala igualmente temperada de 12 notas.

A Fig. 3.5 mostra, em linha contínua, a localização frequencial das frequências fundamentais de tal escala para uma frequência de referência de $f_{\text{ref}} = 440\,\text{Hz}$[7]. Em linha tracejada, vemos a frequência da nota C4 ($f_{\text{C4}} = 261,6\,\text{Hz}$ e número de nota MIDI $i = 60$) e suas parciais harmônicas superiores. É possível perceber que há uma diferença entre as localizações das notas que compõem a escala e das parciais harmônicas da nota C4: elas não estão perfeitamente alinhadas. Por exemplo, para a sétima parcial, notamos que ela está ligeiramente deslocada para a esquerda em relação à nota correspondente ao número MIDI $i = 95$. Da mesma forma, esse desvio está presente em quase todas as demais parciais de f_{C4}. Com isso, criamos uma sequência teórica de desvios a ser comparada com uma observada calculada a partir dos picos detectados do espectro.

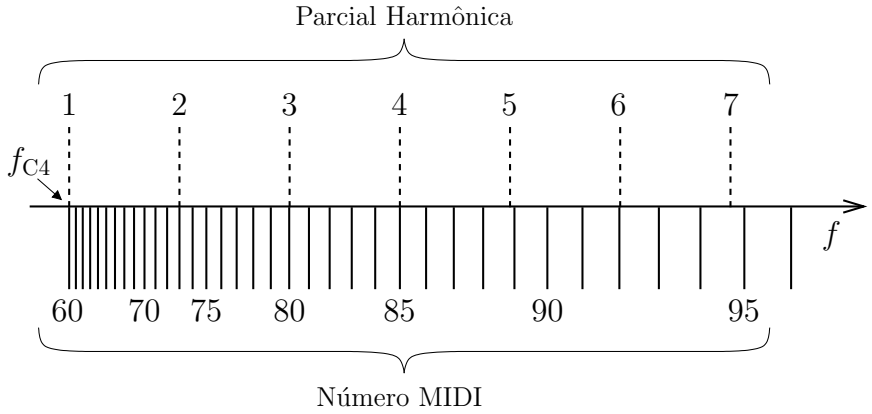


Figura 3.5: Localização frequencial das fundamentais da escala igualmente temperada para uma frequência de referência de 440 Hz (em linha contínua) e localização frequencial das parciais harmônicas para uma fundamental de 261,6 Hz (em linha tracejada).

A Fig. 3.6 mostra o diagrama de blocos do método proposto. O sinal de entrada $x(n)$, como foi dito no Capítulo 2, possui L amostras, ou seja, $n = 0, 1, \ldots, L-1$. Neste trabalho, iremos sempre considerar que $x(n)$ é um sinal discreto de áudio. A primeira etapa é calcular a transformada tempo-frequencial de $x(n)$, $X(k, j)$. Novamente, k representa o eixo temporal e adota os valores $k = 0, 1, \ldots, K-1$, e j representa a informação frequencial e adota os valores $j = 0, 1, \ldots, J-1$. O número de amostras da janela de análise $h(n)$ utilizado é N. Inicialmente, consideramos o uso de uma transformação genérica. A inclusão da função de saliência proposta na STFChT será abordada na Seção 3.5.3.

[7] As frequências são obtidas por meio da expressão $f_i = f_{\text{ref}} 2^{\frac{i}{12}}$.

O próximo passo é a estimação dos picos da transformada, resultando em $\mathcal{P}(k,p)$, com $p = 0, 1, \ldots, P-1$. Para cada quadro k_0, temos $\mathcal{P}(k_0,p) = \{(f_0(k_0), a_0(k_0)), (f_1(k_0), a_1(k_0)), \ldots, (f_{P-1}(k_0), a_{P-1}(k_0))\}$, onde $f_p(k)$ e $a_p(k)$ são a frequência e a magnitude do pico p referente ao quadro k, respectivamente. Selecionamos $P = 80$ picos por quadro. O método utilizado para a detecção de picos foi retirado de [11]. Definimos um pico como um máximo local. Assim, na etapa de detecção de picos, consideramos que a magnitude de um pico é máxima em um certo alcance de frequências e, além disso, maior que um dado limiar. Para chegar ao valor $P = 80$ escolhido, diminuímos o limiar (começando do maior valor da magnitude do espectro) até que $P \geq 80$. Selecionamos, em seguida, os 80 picos com as maiores magnitudes. Para aumentar a resolução frequencial dos picos encontrados, fazemos uma interpolação parabólica com o *bin* correspondente ao pico e os seus *bins* adjacentes [11].

A proposta de [25] para a etapa de detecção de picos se limita ao procedimento descrito acima. No entanto, a escolha dos picos como é feita leva em consideração a magnitude do espectro do sinal sob análise, o que, devido à natureza do sinal, geralmente decai com o aumento da frequência. As parciais harmônicas superiores, portanto, estão em clara desvantagem na etapa de detecção de picos. Para evitar esse problema, adicionamos um passo de pré-processamento, onde a ideia é "planificar" a magnitude do espectro, evitando o decaimento. Estimamos, inicialmente, o seu piso de ruído através do método chamado Estimador Espectral Estocástico (do inglês, *Stochastic Spectrum Estimator*, ou SSE) (detalhado em [28]) e, em seguida, o subtraímos da magnitude do espectro original (em escala logarítmica). Os picos são, por fim, estimados utilizando como entrada o resultado da operação anterior.

Os coeficientes $X(k,j)$ da transformada e os picos $\mathcal{P}(k,p)$ obtidos são utilizados para o cálculo da estimação da frequência de referência, f_{ref}. Como estamos trabalhando com uma escala igualmente temperada, precisamos estimar seu valor de forma a obter as frequências das notas que a compõem. O procedimento, retirado de [29], não será detalhado aqui.

Para calcular a função de saliência, comparamos para cada quadro k as localizações frequênciais de cada pico p e das notas referentes a uma escala igualmente temperada com frequência de referência f_{ref}. As diferenças em frequência obedecem uma determinada lei e podem, portanto, ser utilizadas como uma medida teórica para determinar quais picos correspondem às fontes sonoras e quais correspondem às parciais harmônicas. O procedimento será abordado com mais detalhes na próxima seção.

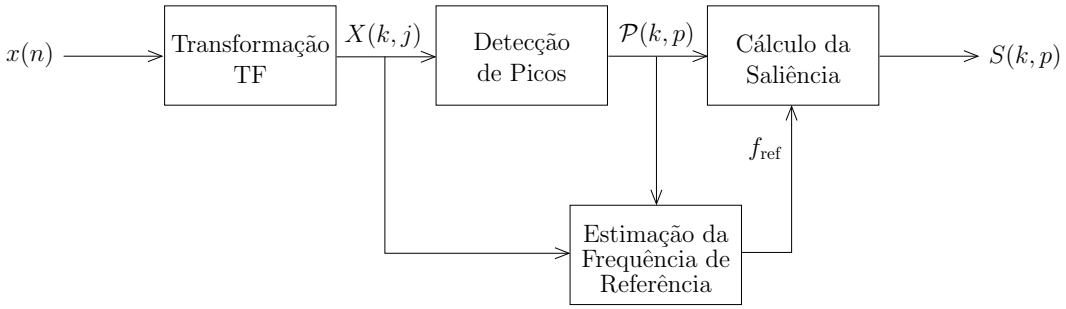


Figura 3.6: Diagrama de blocos do método proposto para o cálculo da saliência independente do timbre.

3.5.2 Cálculo da Saliência

Como foi dito, queremos definir um desvio teórico entre uma frequência f_p qualquer e uma outra pertencente à grade de frequências da escala de igual temperamento. Para isso, vamos considerar que cada semitom é dividido em 100 intervalos iguais na escala geométrica. Denominamos cada intervalo de um cent, ou seja, cada oitava é dividida em 1200 cents [30].

Como a nossa base utilizada para comparação é a escala de igual temperamento, podemos denotar cada fundamental F_i pertencente à escala como

$$F_i = f_{\text{ref}} 2^{\frac{i}{12}}, \tag{3.11}$$

com i correspondendo ao índice de nota MIDI sem o deslocamento de 69 (por simplicidade, assumimos que a nota A4 é equivalente a $i = 0$ em vez de $i = 69$).

Assumimos que o sinal de áudio $x(n)$ é constituído por fontes sonoras de frequência fundamental F_j com parciais harmônicas $f_h^{F_j}$. Dada uma fundamental F_0, as suas parciais harmônicas são dadas por:

$$f_h^{F_0} = hF_0. \tag{3.12}$$

Os desvios em cents de cada parcial $f_h^{F_0}$, com $h = 2, 3, \ldots, n_{\text{H}}$, em relação às fundamentais da escala igualmente temperada são definidos como [25]

$$d_h^{F_0} = 100 \left[12 \log_2 \left(\frac{f_h^{F_0}}{f_{\text{ref}}} \right) - \left\lfloor 12 \log_2 \left(\frac{f_h^{F_0}}{f_{\text{ref}}} \right) \right\rceil \right], \tag{3.13}$$

onde $\lfloor \cdot \rceil$ é o operador de arredondamento e f_{ref} é a frequência de sintonia A4 estimada a partir do sinal. Denotamos por $\{f_h^{F_0}\}_h$ a sequência de todas as parciais harmônicas de F_0 e por $\{d_h^{F_0}\}_h$ a sequência teórica de desvios.

Substituindo as Eqs. (3.11)e (3.12) na Eq. (3.13), obtemos

$$d_h^{F_i} = 100[12 \log_2(h) - \lfloor 12 \log_2(h) \rceil]. \tag{3.14}$$

Podemos perceber que a sequência de desvios $\{d_h^{F_i}\}_h$ não depende mais da fundamental F_i. Na técnica original, proposta em [25], o autor utiliza a Eq. (3.14) como sequência de referência. Aqui, no entanto, utilizamos a definição inicial dos desvios teóricos dada pela Eq. (3.13). Essa alteração é necessária para modelar corretamente variações nas frequências fundamentais F_j presentes no sinal, que não estão quantizadas apenas nas frequências fundamentais F_i da escala temperada.

Definimos, também, uma sequência de desvios observada $\{\hat{d}_h^{f_p}\}_h$ para cada pico p de cada quadro k. A frequência do pico p é dada por f_p. Note que, segundo a alteração proposta aqui, a frequência f_p será a utilizada no lugar de F_0 na sequência de desvios teórica $\{d_h^{F_0}\}_h$.

A saliência $S(k, p)$ é calculada para cada pico p de cada quadro k. Aqui, por simplicidade, vamos omitir a dependência de k, escrevendo apenas $S(p)$. A ideia é utilizar como medida de semelhança a correlação entre as sequências teórica $\{d_h^{f_p}\}_h$ e observada $\{\hat{d}_h^{f_p}\}_h$, ou seja,

$$S(p) = \mathrm{C}(\{d_h^{f_p}\}_h, \{\hat{d}_h^{f_p}\}_h), \tag{3.15}$$

onde $\mathrm{C}(\cdot, \cdot)$ é uma medida genérica de correlação. As duas sequências de desvios podem ser escritas na forma vetorial como $\mathbf{d}^{f_p} = [d_2^{f_p}\ d_3^{f_p}\ \ldots\ d_{n_\mathrm{H}}^{f_p}]^\mathrm{T}$ e $\hat{\mathbf{d}}^{f_p} = [\hat{d}_2^{f_p}\ \hat{d}_3^{f_p}\ \ldots\ \hat{d}_{n_\mathrm{H}}^{f_p}]^\mathrm{T}$, respectivamente. A medida de correlação utilizada no artigo é o produto interno entre as sequências. A medida que iremos utilizar aqui é o coeficiente de correlação, definido por

$$r(p) = \frac{\langle \mathbf{d}^{f_p} - \boldsymbol{\mu}_\mathrm{teo}, \hat{\mathbf{d}}^{f_p} - \boldsymbol{\mu}_\mathrm{obs} \rangle}{\sqrt{\sigma_\mathrm{teo}^2}\sqrt{\sigma_\mathrm{obs}^2}}, \tag{3.16}$$

onde μ_teo é a média do vetor $\mathbf{d}^{f_p}$ e μ_obs, a média do vetor $\hat{\mathbf{d}}^{f_p}$, com $\boldsymbol{\mu}_\mathrm{teo}$ e $\boldsymbol{\mu}_\mathrm{obs}$ sendo vetores de $n_\mathrm{H} - 1$ elementos iguais a μ_teo e μ_obs, respectivamente; σ_teo^2 e σ_obs^2 são as variâncias de $\mathbf{d}^{f_p}$ e $\hat{\mathbf{d}}^{f_p}$, respectivamente; e $\langle \cdot, \cdot \rangle$ é a operação de produto interno. Para obtermos o valor da saliência a partir do coeficiente de correlação, fazemos $S(p) = a_p r(p)$, onde a_p é a amplitude do pico p.

Definimos como $\{f_h^{f_p}\}_h$ a sequência das parciais harmônicas do pico p. Cada valor $f_h^{f_p}$ é dado pela Eq. (3.12). A sequência $\{\hat{d}_h^{f_p}\}_h$ é composta dos desvios observados correspondentes a cada frequência de $\{f_h^{f_p}\}_h$. Ela é calculada para cada pico p de cada quadro k, ou seja, consideramos todos os picos como possíveis candidatos a frequência fundamental.

Para que uma frequência f_p seja considerada uma fundamental, analisamos, dentre os demais picos $f_{p'}$ encontrados, quais correspondem a parciais harmônicas de f_p. Para isso, utilizamos uma função gaussiana G centrada na h-ésima parcial harmônica de f_p, ou seja, com média $\mu_{h,p} = hf_p$, e avaliada nos outros picos $f_{p'}$ detectados.

O desvio padrão de G é definido por $\sigma_{h,p} = \mu_{h,p}\left(2^{\frac{\kappa}{1200}} - 1\right)$, onde $\kappa = 20$ foi selecionado experimentalmente de forma a minimizar o efeito do erro introduzido na etapa de estimação de picos [25]. A ideia é permitir uma certa tolerância na posição frequencial dos picos.

A Fig. 3.7 mostra um exemplo para o caso em que o primeiro pico ($p = 0$), cuja frequência f_0 coincide com a fundamental F_0, é analisado. As frequências $f_{p'}$ dos demais picos são dadas por f_1, f_2, f_3, f_4 e f_5. Consideramos que vamos analisar até a quinta parcial harmônica, ou seja, $n_{\mathrm{H}} = 6$. Nesse caso, geramos gaussianas com médias e desvios padrões dados por $\mu_{h,0} = hf_0$ e $\sigma_{h,0} = \mu_{h,0}\left(2^{\frac{\kappa}{1200}} - 1\right)$, para $h = 2, 3, \ldots, 6$. A seguir, avaliamos os valores de G nas posições $f_{p'}$, ou seja, $G(f_{p'}; \mu_{h,0}, \sigma_{h,0})$, com $p' = 1, 2, \ldots, 5$. Na figura, vemos a função gaussiana para o caso em que $h = 3$ e o seu valor em $f_{p'}$ com $p' = 2$, ou seja, $G(f_2; \mu_{3,0}, \sigma_{3,0})$. Para os demais valores de p', $G(f_{p'}; \mu_{3,0}, \sigma_{3,0}) \to 0$. Geramos, assim, um vetor da forma

$$\mathbf{G}_{\mu_{3,0};\sigma_{3,0}} \approx [0\ G(f_2; \mu_{3,0}, \sigma_{3,0})\ 0\ 0\ 0]^{\mathrm{T}}. \tag{3.17}$$

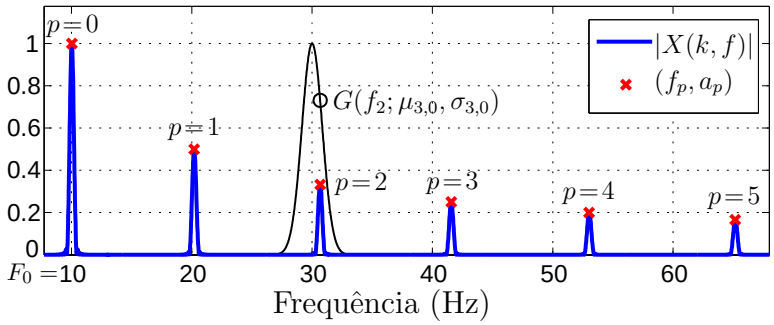


Figura 3.7: Esquema para ilustrar a seleção de picos por meio de gaussianas G com média e variância $\mu_{h,p}$ e $\sigma_{h,p}$, respectivamente, avaliadas em $f_{p'}$ (aqui, $p' = 2$, $h = 3$ e $p = 0$). A curva em azul é o espectro do $k-$ésimo quadro do sinal, $X(k, f)$; os pontos em vermelho são os picos detectados; a curva em preto é a gaussiana $G(f; \mu_{3,0}, \sigma_{3,0})$; e o ponto em preto é a gaussiana avaliada em $f = f_2$.

Neste trabalho, trabalhamos com um vetor $\hat{\mathbf{G}}_{\mu_{h,p};\sigma_{h,p}}$ que força o sinal de igualdade na Eq. (3.17). Essa medida é adotada porque em alguns casos alguns picos espúrios próximos ao pico correspondente à parcial harmônica desejada não irão resultar em valores próximos a zero quando avaliados na função gaussiana. Para isso, consideramos apenas o valor máximo de $G(f_{p'}; \mu_{h,p}, \sigma_{h,p})$ em f'_p.

O próximo passo no cálculo da sequência $\{\hat{d}_h^{f_p}\}_h$ é associar, para cada pico p', desvios $d_{p'}$ como os definidos na Eq. (3.13). Temos

$$d_{p'} = 100\left[12\log_2\left(\frac{f_{p'}}{f_{\mathrm{ref}}}\right) - \left\lfloor 12\log_2\left(\frac{f_{p'}}{f_{\mathrm{ref}}}\right)\right\rceil\right]. \tag{3.18}$$

No exemplo da Fig. 3.7, geramos o vetor $\mathbf{d}$ formado pelos desvios dos picos $p' =$

$1, 2, \ldots, 5$, ou seja, $\mathbf{d} = [d_{f_1}\ d_{f_2}\ d_{f_3}\ d_{f_4}\ d_{f_5}]^{\mathrm{T}}$.

Por fim, o desvio $\hat{d}_h^{f_p}$ correspondente à parcial harmônica h do pico p (e cuja frequência é dada por $f_h^{f_p} = hf_p$) é calculado a partir da seguinte expressão

$$\begin{aligned} \hat{d}_h^{f_p} &= \langle \mathbf{G}_{\mu_{h,p};\sigma_{h,p}}, \mathbf{d} \rangle \\ &= \sum_{p'=p+1}^{P-1} G(f_{p'}; \mu_{h,p}, \sigma_{h,p}) d_{p'}. \end{aligned} \tag{3.19}$$

Na prática, só analisamos os picos p' que se situam, frequencialmente, após o pico p. Os demais satisfazem $f_{p'} < f_p$, para $p' = 0, 1, \ldots, p-1$, e não são, portanto, candidatos plausíveis a parciais harmônicas de f_p. Além disso, só consideramos os picos p cujas frequências $f_p \leq 5000\,\mathrm{Hz}$ como possíveis fundamentais, já que, para sinais de áudio, a probabilidade de termos uma fundamental acima 5000 Hz é muito pequena.

Em alguns casos específicos, quando o módulo do desvio teórico se aproxima de 50 cents, há a possibilidade de erro no momento da quantização (função de arredondamento da Eq. (3.13)). O valor observado de um desvio que a princípio deveria ser próximo de -50 cents pode acabar sendo, na verdade, próximo a 50 cents. Um exemplo desse efeito pode ser visto na Fig. 3.8 em vermelho. Para corrigir essa falha, propusemos um pequeno ajuste no cálculo dos desvios observados. Adotamos um limiar, escolhido empiricamente como $d_\lambda = 45$ cents, e comparamos o módulo do desvio teórico para cada parcial a ele. Para todo h tal que $|d_h^{f_p}| > d_\lambda$, calculamos, agora, duas possibilidades de desvio observado: substituímos a operação de arredondamento na Eq. (3.13) pelas de piso (ou arredondamento para baixo), $\lfloor \cdot \rfloor$, e teto (ou arredondamento para cima), $\lceil \cdot \rceil$, gerando duas sequências de desvios observadas idênticas, com a exceção do ponto h (no caso da Fig. 3.8, $h = 11$). Calculamos, em seguida, a saliência para ambas e selecionamos a de maior valor (curva tracejada em preto).

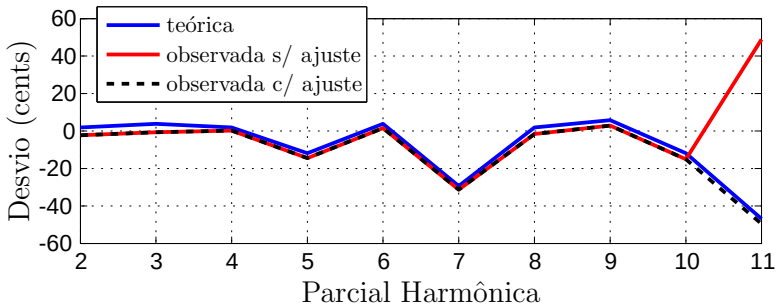


Figura 3.8: Desvios calculados para as sequências teórica (curva em azul), observada sem ajuste (curva em vermelho) e observada com ajuste (curva tracejada em preto).

Por fim, acrescentamos uma etapa de pós-processamento para retirar a ambi-

guidade gerada pelos múltiplos das frequências fundamentais presentes. O procedimento utilizado é semelhante ao apresentado em [8]. A ideia é analisar, para cada pico p correspondente à frequência f_p, se a saliência obtida nas suas frequências submúltiplas f_p/q, para $q = 1, 2, \ldots, p$, é significativa. Para isso, utilizamos, novamente, gaussianas de média $\mu_{1/q,p} = f_p/q$ e variância $\sigma_{1/q,p} = \mu_{1/q,p}(2^{\frac{\kappa}{1200}} - 1)$, com $\kappa = 20$, $G(f; \mu_{1/q,p}, \sigma_{1/q,p})$, e analisamos seus valores nos picos p' cujas frequências $f_{p'} < f_p$, $G(f_{p'}; \mu_{1/q,p}, \sigma_{1/q,p})$. Substituímos, por fim, a saliência em p, $S(p)$, por

$$\bar{S}(p) = S(p) - \max_{q \in [1,\ldots,p]} \{S(p')G(f_{p'}; \mu_{1/q,p}, \sigma_{1/q,p})\}, \tag{3.20}$$

ou seja, subtraímos de $S(p)$ a componente máxima $S(p')$, com $f_{p'} < f_p$ e tal que $f_{p'}$ seja uma frequência aproximadamente submúltipla de f_p[8]. Assim, a saliência nas frequências harmônicas de uma fundamental presente no sinal é, quase sempre, zerada.

A Fig. 3.9 mostra a saliência calculada antes (curva em azul) e depois (curva tracejada em vemelho) do pós-processamento para os picos encontrados com frequência até 3000 Hz. Nesse caso, a saliência do pico correspondente à frequência fundamental ($p = 1$) não foi alterada. Já as saliências dos picos correspondentes às suas parciais harmônicas ($p = 2$, $p = 3$ e $p = 4$, por exemplo) foram zeradas.

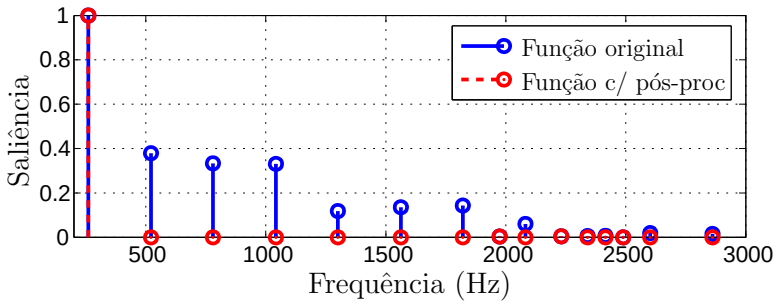


Figura 3.9: Saliência normalizada calculada para os picos encontrados até 3000 Hz sem (curva em azul) e com (curva tracejada em vermelho) pós-processamento.

3.5.3 Implementação

Para inserir a nova função de saliência apresentada na STFChT, utilizamos a mesma ideia de busca exaustiva abordada na Seção 3.3. Dividimos o sinal de entrada $x(n)$ em K quadros e, em seguida, para cada quadro k, calculamos várias instâncias da FChT para distintos valores do parâmetro α. Para cada instância $X_\alpha(k, j)$, calculamos os picos $\mathcal{P}_\alpha(k, p)$ e as saliências $S_\alpha(k, p)$ correspondentes. A função de

[8] Quando $\bar{S}(p) < 0$, fazemos $\bar{S}(p) = 0$.

referência f_{ref} utilizada é fixa e calculada somente para $\alpha = 0$ (ou seja, a instância da STFChT equivalente à STFT).

Podemos pensar que, considerando um dado quadro k, geramos uma matriz $\mathbf{S}_k \in \mathbb{R}^{N_\alpha \times P}$, onde N_α é o número de valores de α a serem testados e P, o número de picos. As componentes de $\mathbf{S}_k$ são os valores das saliências $S_\alpha(k, p)$ para cada α e p analisados. O valor de α selecionado será, portanto, o correspondente à maior componente da matriz $\mathbf{S}_k$.

A implementação da STFChT combinada com a saliência proposta foi realizada somente para o caso de deformação linear. Queremos, essencialmente, visualizar os efeitos da proposta apresentada nesta seção no cálculo da STFChT, independente da deformação escolhida. Dessa forma, utilizamos uma versão simplificada do código da STFChT para testar o método. Uma versão mais completa que inclua o caso citado é um trabalho futuro.

3.5.4 Experimentos e resultados parciais

Para verificar se a mudança proposta de função de saliência resulta em melhorias na representação tempo-frequencial final, utilizamos, novamente, um sinal sintético com frequência fundamental conhecida. Escolhemos um sinal harmônico modulado em frequência por uma senoide e com a mesma fundamental que o apresentado na Seção 3.4.4. Temos, agora, duas diferenças: utilizamos um número maior de parciais harmônicas (15 em vez de 9), e suas amplitudes decrescem de forma inversamente proporcional ao índice da parcial. Para justificar tais mudanças, analisamos a definição e a motivação da nova função de saliência, nessa ordem. A saliência proposta é uma medida de comparação, dado um número de parciais n_{H} a serem acumuladas, entre os $(n_{\text{H}} - 1)$ desvios teóricos e os $(n_{\text{H}} - 1)$ desvios correspondentes às parciais (picos detectados)[9]. Numericamente, escolhemos $n_{\text{H}} = 10$, tornando necessário, portanto, que o sinal possua ao menos 10 parciais. Para criar a possibilidade de ambiguidade com os múltiplos da fundamental (e, posteriormente, a sua eliminação com o pós-processamento), como no caso da saliência original, escolhemos gerar o sinal com um número de parciais ligeiramente superior a n_{H}. A motivação é, justamente, eliminar a dependência entre o cálculo da saliência e o timbre das fontes sonoras presentes. Assim sendo, o decaimento das amplitudes das parciais não deveria causar um impacto significativo na representação resultante. Para o caso da saliência original, apresentada no Cap. 2 e dada na Eq. (3.7), o seu cálculo está diretamente relacionado às amplitudes das parciais presentes, o que pode levar a uma leve pioria na resolução da representação obtida. Dessa forma, a contribuição provocada pela mudança da função de saliência é mais perceptível e, além disso, o

[9]Note que excluímos o desvio correspondente à frequência fundamental.

sinal de teste sintético se torna mais fiel a sinais reais de áudio.

Os experimentos realizados aqui, como já foi dito, só foram feitos para o caso da deformação linear. Os tamanhos N das janelas de análise são 1024, 2048 e 4096 amostras. Realizamos os testes para o sinal sem e com a presença de ruído branco.

As representações tempo-frequenciais do sinal sintético para o caso sem ruído podem ser vistas na Fig. 3.10. De fato, a diferença entre as STFChTs calculadas a partir das duas funções de saliência é, neste caso, pouco significativa. Plotamos, também, os valores do parâmetro α estimados para cada caso ao longo do tempo (vide Fig. 3.11). Percebemos que há apenas uma melhoria sutil na suavidade das curvas obtidas ao trocarmos a função de saliência, principalmente para o caso em que $N = 1024$. Apesar disso, há um erro claro no instante de tempo $t = 1{,}08\,\mathrm{s}$ presente na estimação do parâmetro α para o caso em que $N = 1024$. Ele reflete a forte dependência da função de saliência proposta na etapa de estimação dos picos. Neste quadro, e mais especificamente para este tamanho de janela de análise, essa etapa foi comprometida, provavelmente devido à baixa resolução frequencial da transformada.

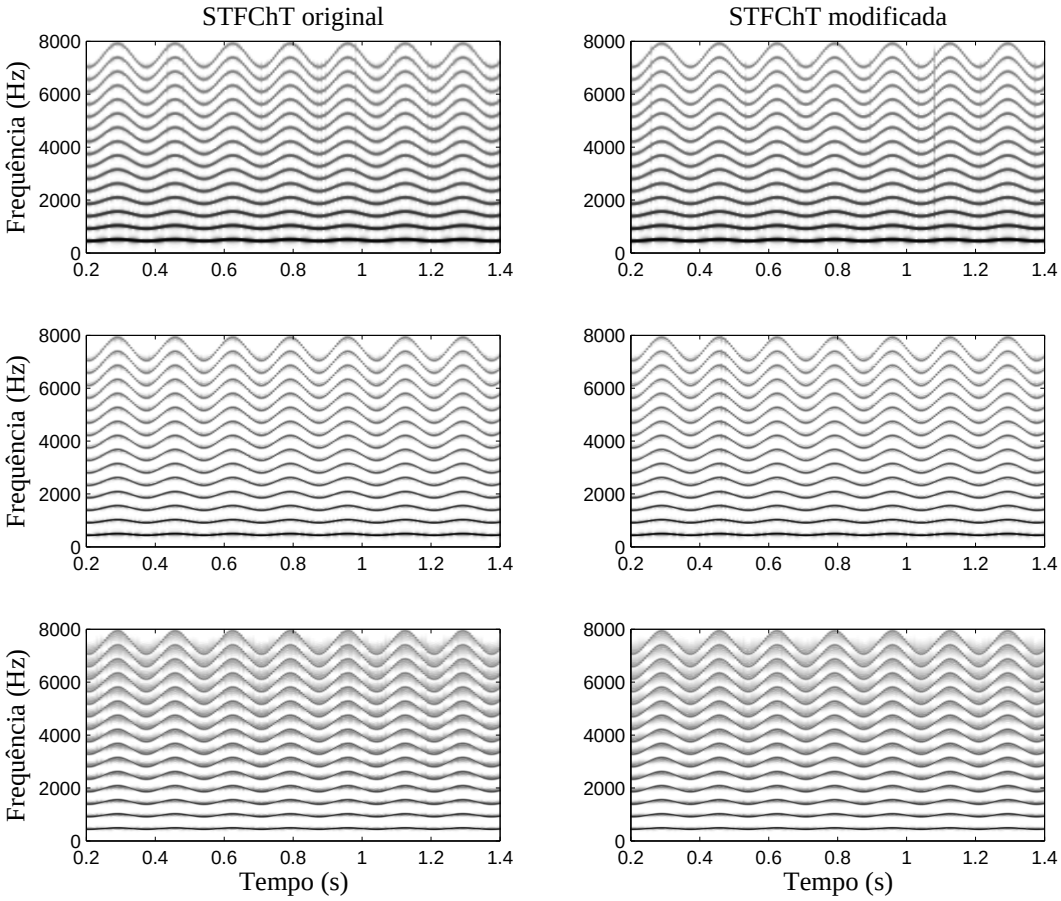


Figura 3.10: STFChTs calculadas com a saliência original (primeira coluna) e com a proposta (segunda coluna) de um sinal sintético. Os seguintes tamanhos de janela de análise foram utilizados: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras.

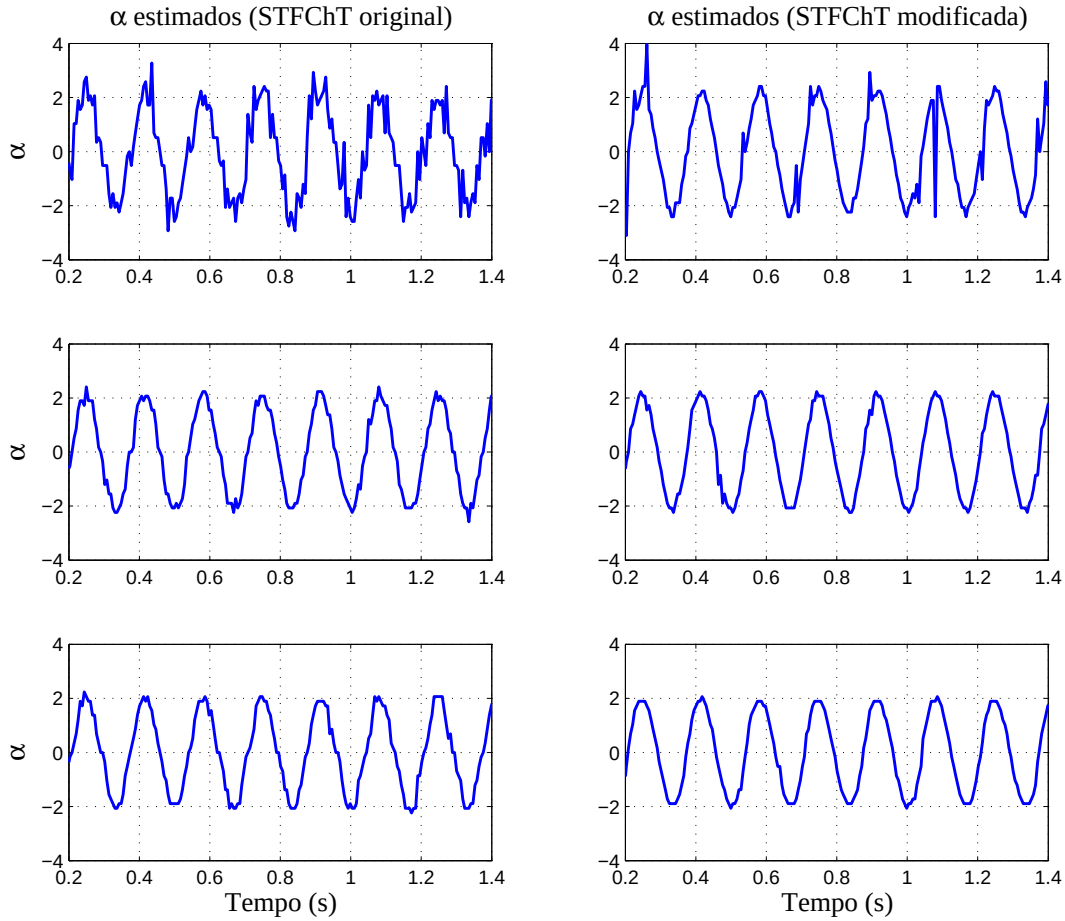


Figura 3.11: Valores estimados de α para as STFChTs calculadas com a saliência original (primeira coluna) e com a proposta (segunda coluna) de um sinal sintético. Os seguintes tamanhos de janela de análise foram utilizados: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras.

As representações tempo-frequenciais do sinal sintético para o caso com ruído (consideramos que o sinal foi corrompido a 20 dB SNR) podem ser vistas na Fig. 3.12. Aqui, a diferença entre as STFChTs se torna mais clara, principalmente para os casos em que $N = 1024$ e $N = 4096$. Vemos que boa parte dos artefatos presentes, decorrentes de erros na estimação do parâmetro α, é suavizada para a STFChT calculada a partir da nova função de saliência. Podemos, também, comprovar tal observação através dos valores estimados de α ao longo do tempo, vistos na Fig. 3.13. Novamente, para a saliência proposta, notamos curvas mais suaves, especialmente para o caso em que $N = 1024$.

Outro ponto que devemos notar é que, para a saliência proposta, não há mudanças significativas nas representações geradas para os sinais com e sem ruído. De fato, para o caso em que $N = 1024$, obtivemos um resultado ligeiramente melhor (com menos erros na estimação do parâmetro α e, consequentemente, menos espalhamento na representação obtida) para o sinal corrompido. Já para a saliência original, os resultados obtidos para o caso sem ruído são visivelmente piores.

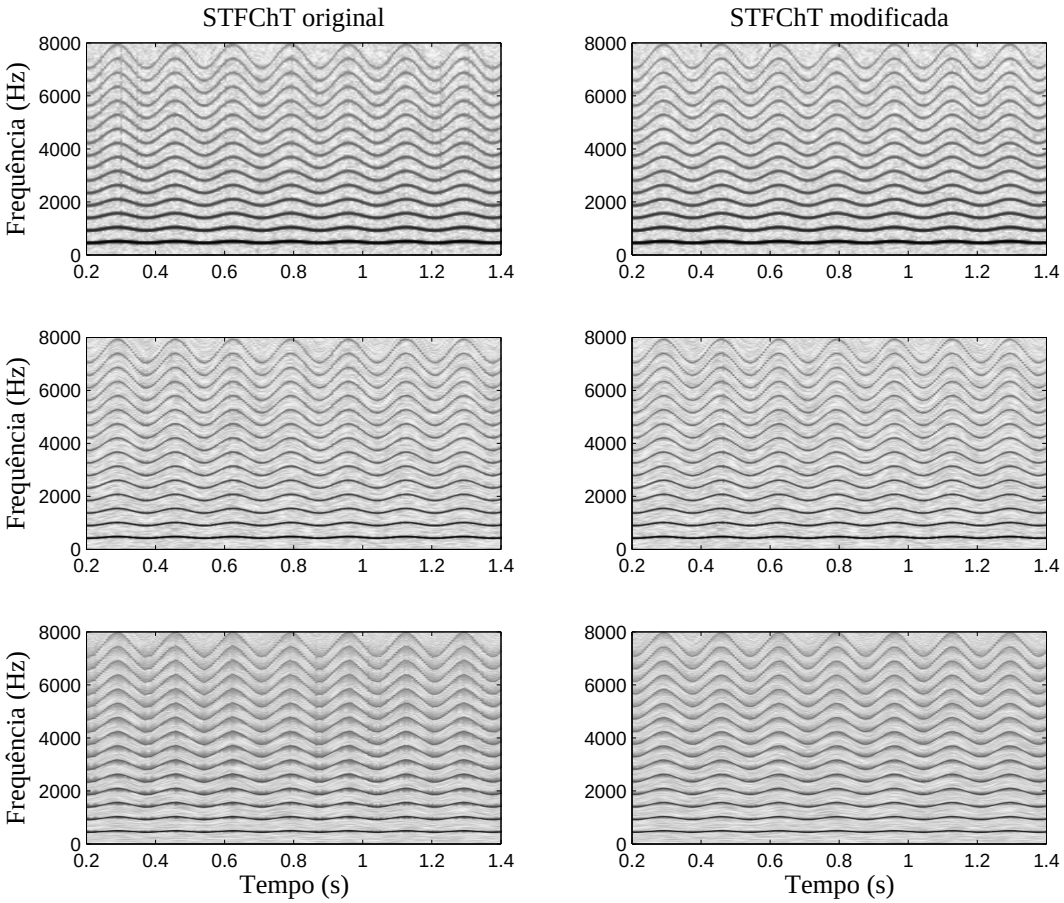


Figura 3.12: STFChTs calculadas com a saliência original (primeira coluna) e com a proposta (segunda coluna) de um sinal sintético corrompido por ruído branco. Os seguintes tamanhos de janela de análise foram utilizados: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras.

É importante ressaltar que, apesar de termos feito a comparação entre as técnicas para os mesmos tamanhos de janela, esta deve ser feita, na verdade, levando-se em consideração o melhor resultado obtido para cada saliência. Adotamos essa medida já que, para gerar a representação tempo-frequencial desejada, o usuário pode selecionar os parâmetros da forma mais conveniente possível. Tal abordagem será utilizada na Seção 3.7 para compararmos resultados, onde estes serão gerados a partir de sinais de áudio reais.

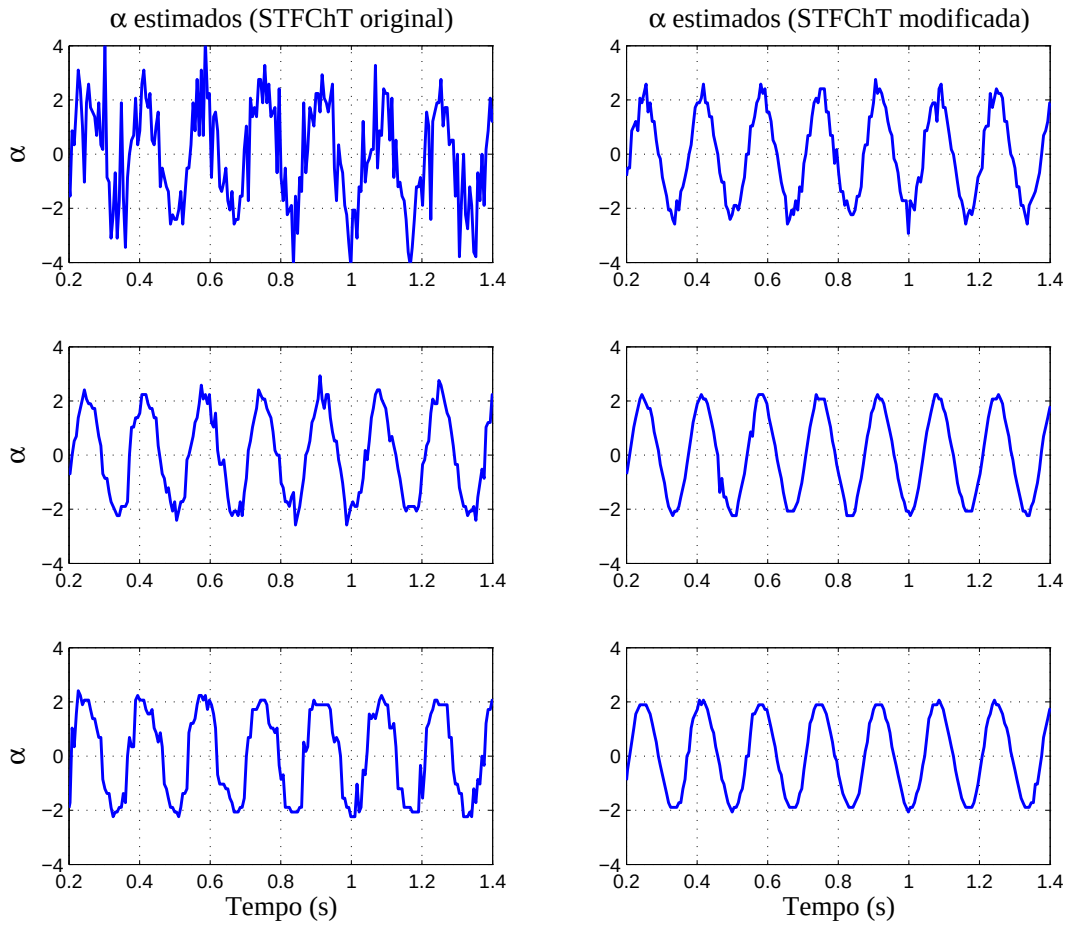


Figura 3.13: Valores estimados de α para as STFChTs calculadas com a saliência original (primeira coluna) e com a proposta (segunda coluna) de um sinal sintético corrompido por ruído branco. Os seguintes tamanhos de janela de análise foram utilizados: 1024 (primeira linha), 2048 (segunda linha) e 4096 (terceira linha) amostras.

3.6 Modelo de Inarmonicidade de Sinais de Música

3.6.1 Proposta

O modelo de inarmonicidade utilizado é dado por [25]

$$f_h^{F_0}(\eta) = hF_0\sqrt{1+\eta h^2}, \tag{3.21}$$

onde F_0 é uma frequência fundamental dada e η é o coeficiente de inarmonicidade, que está relacionado às propriedades físicas da corda. Em geral, η não é dado e devemos, portanto, estimá-lo. Os valores de η podem variar no intervalo $\mathcal{N} = \{0\} \bigcup [10^{-5}, 10^{-3}]$ [26].

A motivação é que, principalmente para instrumentos de corda, as parciais superiores estão ligeiramente deslocadas em frequência das posições que corresponderiam

aos harmônicos da fundamental em questão. Assim sendo, sua presença pode não contribuir de forma significativa no cálculo da função de saliência. Ao levarmos em consideração tal inarmonicidade, esperamos uma melhoria, mesmo que sutil, na detecção das fundamentais presentes no sinal.

A proposta é adicionar o modelo dado pela Eq. (3.21) no cálculo da nova função de saliência. Para isso, acrescentamos o modelo no cálculo dos desvios dado na Eq. (3.13), substituindo $f_h^{F_0}$ por $f_h^{F_0}(\eta)$.

3.6.2 Implementação

Aqui, vamos abordar alguns detalhes necessários para a implementação do modelo de inarmonicidade na FChT. Essencialmente, devemos procurar, para cada candidato a frequência fundamental, um valor para o coeficiente de inarmonicidade η. Assim sendo, uma nova etapa de busca exaustiva é acrescentada para estimar este parâmetro.

Os valores de η plausíveis podem variar de acordo com o instrumento de cordas utilizado. Vamos considerar o intervalo $\mathcal{N} = \{0\} \bigcup [10^{-5}, 10^{-3}]$, que é o alcance típico para uma corda de piano [26]. Selecionamos $N_\eta = 10$ valores espaçados geometricamente no intervalo $[10^{-5}, 10^{-3}]$ mais o valor $\eta = 0$ para serem analisados.

Calculamos, dados um valor de α e um pico candidato a frequência fundamental, N_η instâncias da função de saliência, $S_\alpha(p, \eta)$. Selecionamos, dentre os valores de saliência obtidos, o máximo, ou seja,

$$S_\alpha(p, \eta^*) \;=\; \max_{\eta \in \mathcal{N}} S_\alpha(p, \eta). \tag{3.22}$$

3.6.3 Experimentos e resultados parciais

Para os testes, selecionamos, novamente, o sinal harmônico sintético modulado em frequência por uma senoide. Agora, acrescentamos um grau de inarmonicidade ao sinal, dado por $\eta \approx 1,29 \times 10^{-4}$. A expressão resultante para o sinal $x(t)$ é dada por

$$x(t) = \sum_{h=1}^{15} \frac{1}{h} \cos\left(2\pi f_0(t) h \sqrt{1 + \eta h^2}\right). \tag{3.23}$$

Em seguida, rodamos o algoritmo sem e com o modelo de inarmonicidade para gerar as representações STFChTs apresentadas na Fig. 3.14. Na coluna da esquerda, mostramos o módulo da STFChT sem considerar o fator η e, na da direita, o levamos em consideração. Vemos claramente que a representação se tornou mais esparsa ao adicionarmos este modelo, já que a inarmonicidade do sinal afeta diretamente a acumulação das parciais, que, agora, não são mais harmônicas e se encontram, portanto, ligeiramente deslocadas da sua posição original.

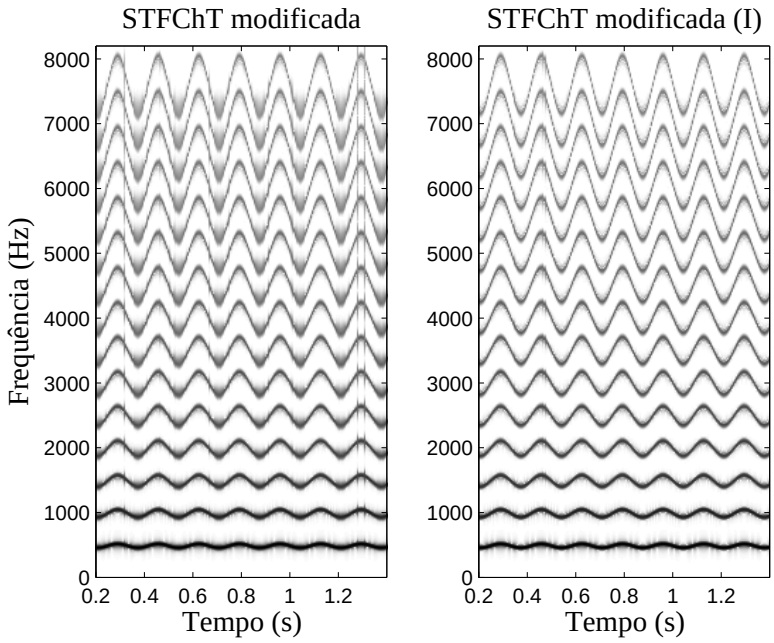


Figura 3.14: STFChTs calculadas com a saliência proposta sem o modelo de inarmonicidade (primeira coluna) e com o modelo (segunda coluna) de um sinal sintético. O tamanho de janela de análise utilizado foi de 2048 amostras.

A inarmonicidade do sinal, como já foi dito, provoca erros na estimação da taxa de inclinação α, como pode ser visto na Fig. 3.15. Na coluna da esquerda, vemos os valores de α em função do tempo estimados sem a correção no modelo. A curva ainda aparenta seguir os valores corretos do parâmetro, porém com uma certa quantidade de ruído adicionada. Esses erros acarretam, como pode ser visto na Fig. 3.14, em espalhamento na representação tempo-frequencial, especialmente nas parciais superiores.

A coluna da direita da Fig. 3.15 mostra, novamente, os valores de α estimados em função do tempo (primeira linha) para a STFTChT modificada com modelo de inarmonicidade e, também, os valores de η estimados ao longo do tempo. Percebemos que a curva de taxa de inclinação se comporta como esperado, porém o mesmo não ocorre para o parâmetro η. Para a maioria dos instantes de tempo, o fator de inarmonicidade foi estimado corretamente. Para outros, no entanto, os erros foram consideravelmente grandes. Estes não afetaram visualmente os valores de α estimados, com alguns poucos pontos de exceção. A interpretação desse resultado é a baixa sensibilidade da estimação de α em relação ao parâmetro η. O experimento indica que há ambiguidade entre os valores de η ao estimarmos α, o que significa que deveríamos mudar os valores de η sobre os quais realizamos a busca exaustiva, considerando mais amostras em regiões com maior sensibilidade e menos em regiões com menor sensibilidade.

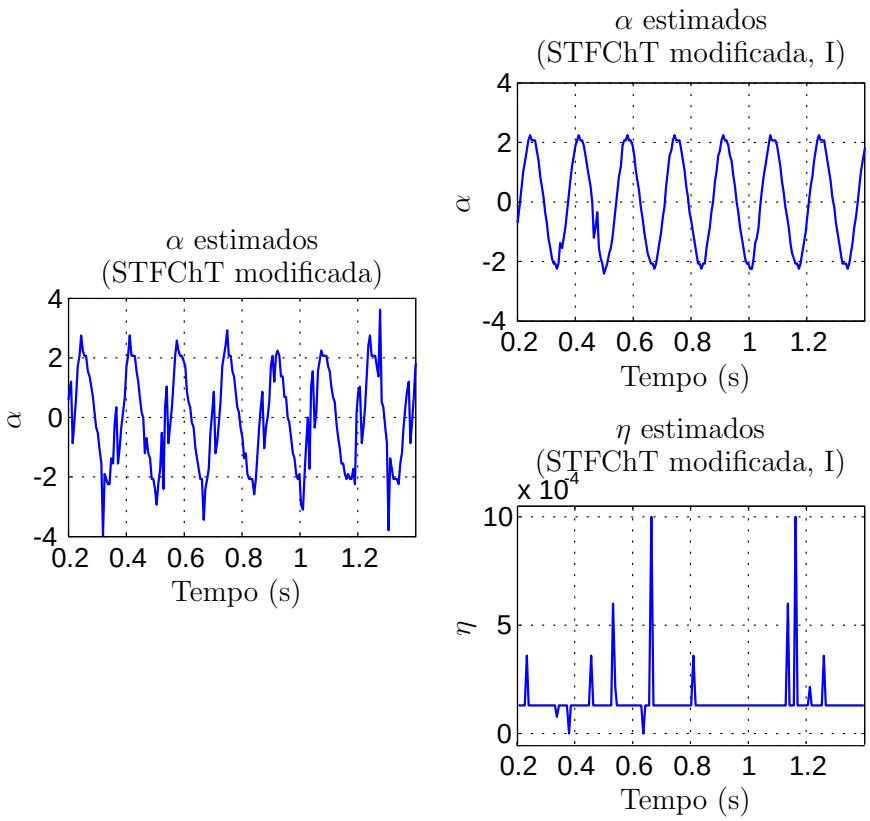


Figura 3.15: Valores de α (e η) estimados com a saliência proposta sem o modelo de inarmonicidade (primeira coluna) e com o modelo (segunda coluna) de um sinal sintético. O tamanho de janela de análise utilizado foi de 2048 amostras.

3.7 Resultados

Nesta seção, apresentaremos os resultados para as modificações propostas utilizando sinais de áudio reais. Abordamos, primeiramente, o caso da deformação não-linear. Escolhemos um sinal cujas variações frequenciais com o tempo fossem significativas. A seguir, mostramos os resultados obtidos para a nova função de saliência utilizada, onde já incluímos o modelo de inarmonicidade proposto, desta vez com outro sinal gerado por instrumento de cordas. Comparamos este método com a STFChT original.

A Fig. 3.16 mostra a magnitude da STFChT do sinal de ópera escolhido para os casos de deformação linear (centro) e não-linear (abaixo). A magnitude da STFT (acima) é mostrada para comparação. O tamanho de janela utilizado para as três imagens foi de 2048 amostras, que, considerando uma frequência de amostragem de 44100 Hz, corresponde a aproximadamente 46,4 ms. Como esse sinal apresenta flutuações em frequência suficientemente rápidas, esse quadro cujo tamanho é em torno da metade da duração proposta na Seção 3.4.2 (de 100 ms) permitiu realçar melhor as diferenças entre as três representações. Como visto na Seção 3.4.4 para sinais sintéticos, aqui também é possível perceber que a representação do gráfico

inferior (STFChT com deformação não-linear) apresenta melhor resolução.

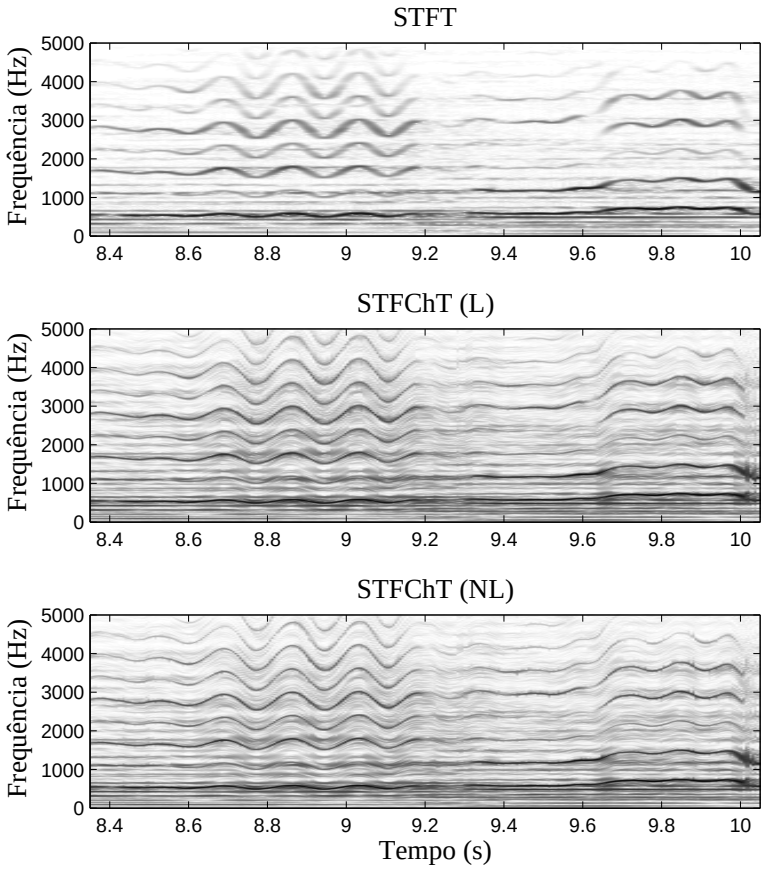


Figura 3.16: STFT e STFChTs com deformação linear (segunda linha) e não-linear (terceira linha) de um sinal de ópera.

A Fig. 3.17 mostra os F0gramas para ambas as versões de STFChT, com deformação linear (acima) e não-linear (abaixo). Os dois métodos parecem estimar corretamente a frequência fundamental do sinal (não representada aqui), mas é possível perceber que, para o caso não-linear, o gráfico é ligeiramente mais esparso.

Um outro aspecto importante a ser levado em consideração é o tempo de processamento da STFChT com deformação não-linear. Para a terceira amostragem, por exemplo, como a busca é realizada sobre um maior número de amostras, o tempo necessário para calcular a FChT é aproximadamente 5 vezes maior que no caso linear. Uma análise mais detalhada da complexidade ainda está em andamento. Essa diferença de tempo deve ser levada em consideração de forma a equilibrar esta questão com os benefícios que a deformação não-linear pode trazer para a representação tempo-frequencial de sinais de áudio com rápidas variações em frequência.

A Fig. 3.18 mostra as representações STFChT original e modificada com modelo de inarmonicidade para um sinal de violino. Vemos que há poucas diferenças perceptíveis nas representações obtidas. De fato, os resultados obtidos para a saliência original foram ligeiramente melhores, ou seja, mais esparsos. A Fig. 3.19 mostra

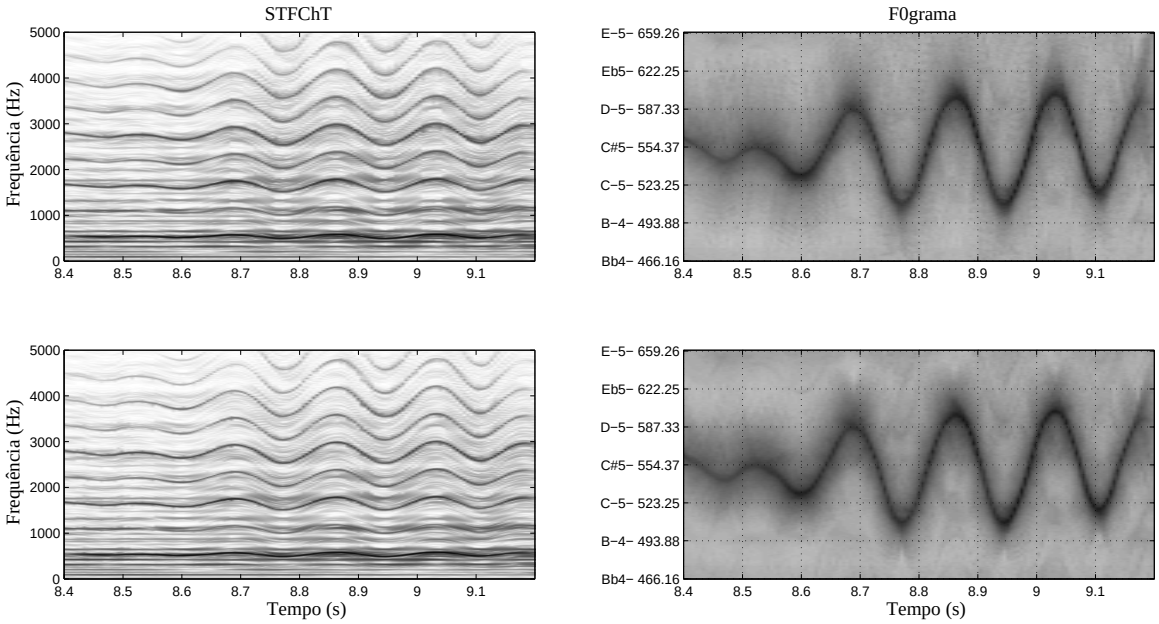


Figura 3.17: F0gramas para as STFChTs com deformação linear (primeira linha) e não-linear (segunda linha) de um sinal de ópera.

os valores de α e η estimados para ambas representações. Vemos que estes não apresentam um padrão característico, como nos casos apresentados na seção anterior. Isso indica, novamente, a presença de uma certa ambiguidade nos valores dos parâmetros.

A saliência proposta é composta de diversas etapas descritas aqui, como detecção de picos, estimação da frequência de referência e o próprio cálculo da função. Cada etapa é suscetível a pequenos erros provenientes de más escolhas dos parâmetros de entrada que, em muitos casos, podem interferir diretamente na resolução da representação final. A implementação realizada não está, nesse sentido, robusta à escolha destes parâmetros, tais como tamanho da janela, número de harmônicos a ser considerado e alcance dos valores de α, entre outros. Um ajuste mais fino poderia, possivelmente, levar a uma melhoria na esparsidade da representação de forma a alcançar a obtida pelo método original. Um estudo mais detalhado da escolha dos parâmetros de entrada deve ser feito para evitar que pequenas variações nestes acabem por acarretar representações espalhadas no plano tempo-frequencial.

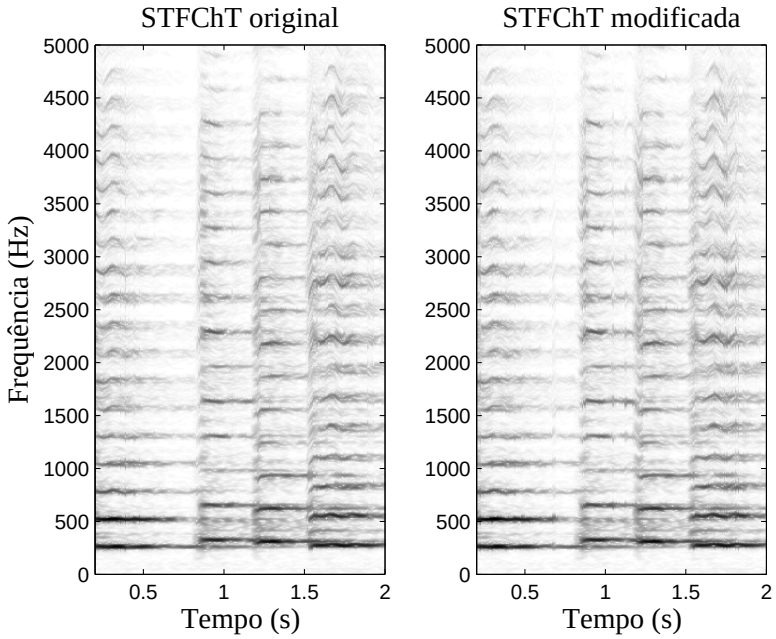


Figura 3.18: STFChTs calculadas com a saliência original (primeira coluna) e proposta com o modelo de inarmonicidade (segunda coluna) de um sinal de violino. O tamanho de janela de análise utilizado foi de 4096 amostras.

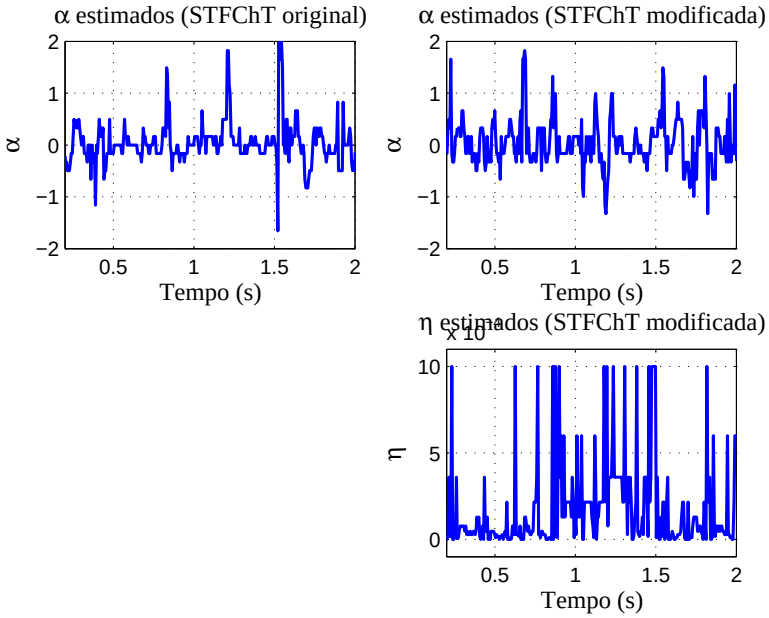


Figura 3.19: Valores de α (e η) estimados com as saliência original (primeira coluna) e proposta com o modelo de inarmonicidade (segunda coluna) de um sinal de violino. O tamanho de janela de análise utilizado foi de 4096 amostras.

3.8 Considerações Finais

Neste Capítulo, apresentamos a definição de Transformada *Fan Chirp* (FChT) e sua aplicação em sinais de áudio por meio da Transformada *Fan Chirp* de Tempo Curto (STFChT). Apresentamos duas possíveis mudanças. A primeira delas é a deformação não-linear, na qual passamos a considerar que a frequência fundamental da melodia mais proeminente de um dado trecho do sinal sob análise se comporta segundo uma curva de segundo grau. Isso nos permite aumentar o número de amostras da janela de análise e, consequentemente, melhorar a resolução da representação devido ao processo de janelamento. Por outro lado, perdemos em tempo de processamento, que aumenta em aproximadamente seis vezes com o novo método.

Outra variação do método proposta foi a troca da função de saliência até então utilizada, que leva em consideração as amplitudes das parciais harmônicas na estimação da taxa de inclinação α, por uma que leva apenas a informação de frequência. Vimos que esta proposta funcionou bem para sinais sintéticos na presença de ruído de fundo e cujas amplitudes das parciais decaem com o aumento da frequência. Se compararmos com os resultados obtidos para o mesmo tipo de sinal ao utilizarmos a STFChT original, percebemos uma melhoria na estimação de parâmetros daquela, refletida na esparsidade da representação. Ao utilizarmos um sinal real de violino, no entanto, tais melhorias não se evidenciaram.

De forma geral, percebemos que há uma grande flutuação na estimação dos parâmetros de taxa de inclinação α e fator de inarmonicidade η. O mesmo ocorre para a curvatura β, apesar de não ter sido ilustrada aqui. Tal análise nos leva à conclusão de que há, possivelmente, regiões de maior e menor sensibilidade da representação em relação a estes parâmetros. A próxima etapa deste trabalho é estudar e determinar tais regiões, de forma que a amostragem de parâmetros seja feita de acordo, ou seja, atribuindo um maior número de pontos às regiões com maior sensibilidade, e menor número às regiões menos sensíveis.

Capítulo 4

Esparsidade Estruturada

4.1 Considerações Iniciais

No Cap. 2, introduzimos o conceito de aproximações estruturadas [31] que iremos abordar aqui. Como foi dito, queremos representar um sinal de áudio $\mathbf{x}$ com coeficientes esparsos $\mathbf{c}$ utilizando um dicionário sobrecompleto $\mathbf{\Phi}$. Adotamos, para isso, uma medida de esparsidade $f(\mathbf{c})$ que, inicialmente, pode ser aplicada a cada elemento c_l, com $l = 1, 2, \ldots, P$, separadamente, já os consideramos independentes entre si. No entanto, sabemos que os coeficientes são dependentes entre si, já que as parciais que compõem $\mathbf{x}$ se manifestam no plano tempo-frequência como padrões verticais para componentes transitórias ou horizontais para componentes tonais. No primeiro caso, dizemos que o sinal apresenta esparsidade temporal e persistência frequencial e no segundo, esparsidade frequencial e persistencial temporal. Podemos, portanto, incorporar ao critério de esparsidade escolhido $f(\cdot)$ essa informação a priori sobre a estrutura de $\mathbf{c}$. Além disso, estruturas de vizinhança no domínio dos coeficientes $\mathbf{c}$ podem ser acrescentadas de forma a explorar suas propriedades de persistência [12].

Há alguns objetivos que podemos traçar dentro do contexto de esparsidade estruturada. Estamos interessados, essencialmente, em estimar os coeficientes ótimos $\hat{\mathbf{c}}$ que melhor modelam o sinal $\mathbf{x}$ pelo modelo proposto, dado por $\mathbf{\Phi}\hat{\mathbf{c}}$. Assim sendo, para um sinal ruidoso, considerando que o ruído presente é branco, ou seja, descorrelacionado, o algoritmo irá reconstruir um sinal $\hat{\mathbf{x}}$ sem a presença de ruído, o que pode ser desejável. Podemos, também, pensar em separar o sinal de áudio em três camadas: tonal, transitória e ruidosa. Para isso, o dicionário $\mathbf{\Phi}$ utilizado é formado por uma concatenação de outros dois dicionários ortogonais com diferentes resoluções tempo-frequenciais. O algoritmo utilizado se chama BCR (do inglês, *Block Coordinate Relaxation Method*) [32].

Para inserir interdependência entre os coeficientes, podemos trocar a norma ℓ_1

utilizada até então por uma norma mista $l_{p,q}$ (definida na próxima seção) e, além disso, adicionar uma ponderação de vizinhança. A solução para o problema de minimização apresentado no Cap. 2 (e dado pela Eq. (2.24)) já não pode ser obtida, para este caso, pelo método apresentado de limiarização iterativa. Para isso, é necessário utilizar operadores generalizados de limiarização suave [32]. Aqui, no entanto, escolhemos trabalhar com a norma ℓ_1, que é um caso particular da norma $\ell_{p,q}$, com $p = q = 1$, e variar a vizinhança para sinais de áudio de diferentes naturezas.

A proposta original [12] utiliza a STFT como dicionário. Aqui, utilizamos, além da STFT, a CQT. Esta transformada, como já foi dito, atribui resoluções diferentes a diferentes *bins* de frequência. Isso faz com que componentes de baixa frequência sejam analisadas com janelas maiores, resultando em maior resolução frequencial, e que componentes de alta frequência sejam analisadas com janelas menores, resultando em maior resolução temporal. Essa alteração permite lidar melhor com a limitação causada pelo Princípio da Incerteza presente na STFT. Os resultados obtidos com as duas transformadas são apresentados e comparados na Seção 4.5. Os arquivos dos sinais referentes aos experimentos realizados podem ser encontrados no site: `www.smt.ufrj.br/~isabela.apolinario`.

4.2 Definições

Definimos, inicialmente, o conceito de normas mistas. Para isso, vamos escrever o conjunto de coeficientes $\mathbf{c}$ utilizando um índice duplo $\{g, m\}$, isto é, $\mathbf{c} = \{c_{g,m}, g = 1, \ldots, G,\ m = 1, \ldots, M\}$, com $GM = P$. A norma mista $\ell_{p,q}$ de $\mathbf{c}$ é definida, então, como [32]

$$\|\mathbf{c}\|_{p,q} = \left(\sum_{g=1}^{G} \left(\sum_{m=1}^{M} |c_{g,m}|^p \right)^{q/p} \right)^{1/q}, \tag{4.1}$$

Os índices são utilizados de forma hierárquica: os coeficientes são divididos em G grupos independentes e os M coeficientes dentro de cada grupo são dependentes entre si. Sabemos que $\mathbf{c}$ é indexado, por definição, pelo índice duplo $\{k, j\}$ correspondente ao tempo e à frequência, respectivamente. Podemos, portanto, associar os índices g e m aos índices k e j de diferentes formas de acordo com o objetivo. Por exemplo, para um sinal com parcelas puramente tonais, esperamos, como foi dito, esparsidade frequencial e persistência temporal. Dessa forma, há uma forte dependência entre os coeficientes $\{c_{k,j_0}\}_k$, com $k = 0, 1, \ldots, K-1$ e $j = j_0$ fixo, e a melhor associação seria, portanto, o índice g ao j e o m ao k.

Agora, queremos solucionar o seguinte problema de minimização [32]:

$$\hat{\mathbf{c}} = \arg\min_{\mathbf{c}} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{\Phi c}\|_2^2 + \frac{\lambda}{q}\|\mathbf{c}\|_{p,q}^q \right\}. \tag{4.2}$$

A solução da equação acima, para o caso de um dicionário ortogonal, é obtida por meio de um operador generalizado de limiarização suave, definido a seguir.

Seja $\xi = \xi_\lambda : \mathbb{C}^P \mapsto \mathbb{R}^+$ uma função não-negativa, chamada função de limiarização, e Γ o conjunto de índices tempo-frequenciais, ou seja, $\gamma = \{k, j\} \in \Gamma$, para $k = 0, 1, \ldots, K-1$ e $j = 0, 1, \ldots, J-1$. Então, para $\mathbf{u} \in \mathbb{C}^P$, o operador generalizado de limiarização suave é definido componente a componente como[1] [12]

$$\mathbb{S}_{\lambda,\xi}(u_\gamma) \triangleq u_\gamma(1 - \xi_{\lambda,\gamma}(\mathbf{u}))^+ \tag{4.3}$$

e escrevemos $\mathbb{S}_\xi(\mathbf{u}) \triangleq \{\mathbb{S}_\xi(u_\gamma)\}_\gamma$. Os coeficientes são obtidos, então, como[2]

$$\hat{\mathbf{c}} = \mathbb{S}_{\lambda,\xi}(\mathbf{\Phi}^*\mathbf{x}). \tag{4.4}$$

No caso particular em que $p = q = 1$, a norma $\ell_{p,q}$ se torna a norma ℓ_1 e a solução do problema de minimização, agora chamado de LASSO, dada pela Eq (4.3) passa a ser obtida pelo operador de limiarização suave convencional. Temos, neste caso, $\xi_{\lambda,\gamma}(\mathbf{u}) = \xi^{\mathrm{L}}(u_\gamma) = \frac{\lambda}{|u_\gamma|}$ e, portanto,

$$\mathbb{S}_{\lambda,\xi}(u_\gamma) = \mathbb{S}_\lambda(u_\gamma) = \mathrm{sign}(u_\gamma)(|u_\gamma| - \lambda)^+, \tag{4.5}$$

como foi visto no Cap. 2.

4.3 Algoritmo

Até agora, consideramos que o dicionário utilizado $\mathbf{\Phi}$ forma uma base ortonormal. Para o caso mais geral, em que o dicionário utilizado é um *frame* arbitrário (ou, ainda, uma união de *frames*), o problema de minimização é solucionado por um algoritmo Landweber iterativo limiarizado [32], chamado de ISTA [12]. Aqui, utilizamos a sua versão rápida, chamada FISTA.

É possível mostrar que, para um ponto de partida arbitrário $\mathbf{c}^0 \in \mathbb{C}^P$, o resultado obtido pelo ISTA converge fortemente para a solução $\hat{\mathbf{c}}$ da Eq. (4.2):

$$\hat{\mathbf{c}} = \lim_{n\to\infty} \mathbb{S}_{\lambda,\xi}(\mathbf{c}^n + \mathbf{\Phi}^*(\mathbf{x} - \mathbf{\Phi}\mathbf{c}^n)) \tag{4.6}$$

se $\|\mathbf{\Phi}\| < \sqrt{2}$ [12]. A sua versão rápida, o FISTA, é apresentada a seguir [16], onde já consideramos que $p = q = 1$:

[1] Para $x \in \mathbb{R}$, definimos $x^+ = \max(0, x)$.

[2] Lembrando que o sinal de áudio é dado por $\mathbf{x} \in \mathbb{R}^L$.

Algoritmo 1 FISTA

1: $\mathbf{c} = \mathbf{b}^1 = \mathbf{0}$ e $t_1 = 1$
2: **repetir**
3: $\quad \mathbf{c} = \mathbb{S}_\lambda(\mathbf{b}^n + \mathbf{\Phi}^*(\mathbf{x} - \mathbf{\Phi}\mathbf{b}^n))$
4: $\quad t_{n+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_n^2})$
5: $\quad \mathbf{b}^{n+1} = \mathbf{c}^n + \frac{t_n - 1}{t_n + 1}(\mathbf{c}^n - \mathbf{c}^{n-1})$
6: **até** convergência

4.4 Vizinhança

Vimos que o problema de minimização dado na Eq. (4.2) pode ser resolvido em um único passo por meio de um operador generalizado de limiarização suave, para o caso de um dicionário $\mathbf{\Phi}$ ortonormal, ou em um processo iterativo, no qual o mesmo operador é aplicado a cada passo até a convergência. Podemos, agora, modificar a função de limiarização ξ de forma a melhor considerar estruturas internas inerentes a uma classe maior de sinais de áudio. O fato de que estes podem ser esparsos no tempo (ou frequência) e persistentes em frequência (ou no tempo, respectivamente) pode ser explorado dessa forma. A mudança é realizada por meio de um sistema de vizinhança, definido a seguir.

Definimos, inicialmente, os pesos $\mathbf{v}_\gamma$ da vizinhança tempo-frequencial no entorno do ponto γ. Estes são definidos como as sequências não-negativas $\mathbf{v}_\gamma = [v_\gamma(1)\ v_\gamma(2)\ \ldots\ v_\gamma(P)]^{\mathrm{T}}$, com $v_\gamma(\gamma') \geq 0$, $\forall \gamma, \gamma' \in \Gamma$, que satisfazem as seguintes propriedades

$$\|\mathbf{v}_\gamma\|_1 = 1,\ \sum_{\gamma' \in \Gamma} v_{\gamma'}(\gamma) \leq \tilde{C} < \infty \text{ e } v_\gamma(\gamma) > 0,\ \forall \gamma \in \Gamma, \tag{4.7}$$

onde $\tilde{C}$ é uma constante tal que $\tilde{C} > 0$. Chamamos de $N_\gamma \triangleq \mathrm{supp}(\mathbf{v}_\gamma) = \{\gamma' \in \Gamma | v_\gamma(\gamma') > 0\}$[3] a vizinhança tempo-frequencial do índice γ. A Fig. 4.1 mostra um plano tempo-frequência com $\Gamma = \{1, 2, \ldots, 30\}$ e um exemplo de vizinhança $N_\gamma = \{7, 8, 9, 12, 13, 14, 17, 18, 19, 22, 23, 24\}$ centrada em $\gamma = 18$.

Para pesos $\mathbf{v}_\gamma$ de vizinhança dados, definimos o funcional de suavização de vizinhança $\eta : \mathbb{C}^P \to \mathbb{R}_0^+$ componente a componente como

$$\eta(c_\gamma) \triangleq \left(\sum_{\gamma' \in \Gamma} v_\gamma(\gamma') |c_{\gamma'}|^2 \right)^{1/2}. \tag{4.8}$$

Para $\mathbf{c} \in \mathbb{C}^P$, estabelecemos $\eta(\mathbf{c}) \triangleq \{\eta(c_\gamma)\}_{\gamma \in \Gamma}$. O operador persistente generalizado

[3] Chamamos de $\mathrm{supp}(\mathbf{v}_\gamma)$ o suporte de $\mathbf{v}_\gamma$.

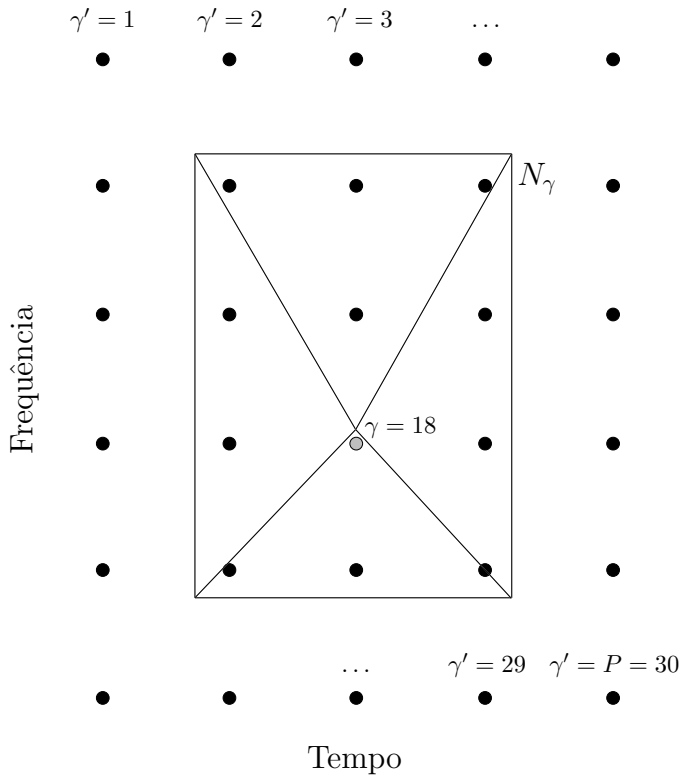


Figura 4.1: Exemplo de vizinhança.

de limiarização suave é definido, então, por

$$\mathbb{S}^*_{\lambda,\xi}(c_\gamma) \triangleq \mathbb{S}_{\lambda,\xi,v}(c_\gamma) = c_\gamma(1 - \xi^*_{\lambda,\gamma}(\mathbf{c}))^+, \tag{4.9}$$

com $\xi^*_{\lambda,\gamma} \triangleq \xi_{\lambda,\gamma} \circ \eta$, onde "$\circ$" é o operador de composição de funções, ou seja, $\xi^*_{\lambda,\gamma}(\mathbf{c}) = \xi_{\lambda,\gamma}(\eta(\mathbf{c}))$.

A primeira propriedade da Eq. (4.7) garante que a extensão global da vizinhança não interfira com a esparsidade da solução. A segunda, garante que η seja bem definido. Sem a terceira, não haveria a intuição sobre o conceito de vizinhança [12].

Para o caso que estamos analisando, o problema LASSO (no qual $p = q = 1$) passa a ser denominado *Windowed Group-Lasso* (ou WGL) e escrevemos a função de limiarização como $\xi^{\mathrm{WGL}} = \xi^{\mathrm{L}} \circ \eta$.

4.5 Aplicação em Áudio: Redução de Ruído de Fundo

4.5.1 Motivação

Aqui, assumimos o seguinte modelo para o sinal de áudio $\mathbf{x} \in \mathbb{R}^L$:

$$\mathbf{x} = \mathbf{\Phi c} + \mathbf{e}, \tag{4.10}$$

onde $\mathbf{e} \in \mathbb{R}^L$ é o ruído aditivo. Na análise, vamos considerar que este é branco gaussiano, condição suficiente para que o método proposto funcione[4] [12]. Novamente, temos $\mathbf{\Phi} \in \mathbb{C}^{L \times P}$ e $\mathbf{c} \in \mathbb{C}^P$. Queremos reduzir a quantidade de ruído na observação $\mathbf{x}$, o que significa aproximar de forma esparsa os coeficientes $\mathbf{c}$ [12]. O sinal limpo será dado, portanto, por $\hat{\mathbf{x}} = \mathbf{\Phi}\hat{\mathbf{c}}$.

4.5.2 Experimentos

Para estimar os coeficientes $\mathbf{c}$, vamos utilizar, como já foi dito, a norma ℓ_1 como critério de esparsidade e o algoritmo FISTA descrito na Seção 4.3. Além disso, utilizaremos dois dicionários gerados a partir das transformadas inicialmente propostas: a STFT e a CQT (cujo desempenho queremos aferir nesse contexto).

Para a STFT, utilizamos como janela de análise $h(n)$ uma janela de Hann de $N = 1024$ amostras e um salto temporal de $a = 256$ amostras. Utilizamos, para a CQT, a sua versão rasterizada com janela de análise igual a raiz quadrada de uma janela de Blackman-Harris. A frequência mínima desejada é dada por $f_{\min} = 20\,\text{Hz}$[5]; a máxima desejada, por $f_{\max} = 17640\,\text{Hz}$; o salto temporal, por $a = 21$; e número de *bins* por oitava, por $B = 24$. Consideramos, para os dois casos, que o sinal de entrada $x(n)$ foi amostrado a uma taxa de $f_\text{s} = 44100\,\text{Hz}$.

Os experimentos realizados aqui foram retirados de [12]. Testamos o algoritmo para, como no artigo, três sinais de áudio de naturezas diferentes: um composto de instrumentos de cordas, outro de piano e o terceiro, de instrumentos percussivos. Cada tipo de sinal de áudio gera um espectro com características próprias. Esperamos, por exemplo, que o espectro de instrumentos percussivos apresente, essencialmente, padrões verticais, os quais evidenciam os instantes de tempo no qual estes instrumentos foram tocados. Por outro lado, para instrumentos tonais, como é o caso de instrumentos de corda e do piano, esperamos padrões horizontais, que correspondem às fundamentais e aos harmônicos das notas tocadas. No caso do piano,

[4] Lembrando que vamos utilizar a norma ℓ_1 como critério de esparsidade.

[5] Na verdade, o valor de $f_{\min}$ foi reduzido até tornar o número de oitavas inteiro. Obtivemos $f_{\min} \approx 17,73\,\text{Hz}$.

espera-se, ainda, uma localização temporal marcada dos ataques bem definidos.

As Figs. 4.2, 4.3 e 4.4 mostram as magnitudes das representações STFT e CQT para os sinais de piano, cordas e percussão, respectivamente[6]. Escolhemos uma escala logarítmica para o eixo representando a frequência para ambas as transformadas. Como foi visto no Cap. 2, a CQT atribui um número maior de *bins* para baixas frequências, o que pode ser observado nos gráficos dos instrumentos tonais (piano e cordas: Figs. 4.2 e 4.3, respectivamente). Enquanto para a STFT a informação dessa faixa de frequências está espalhada ao longo do eixo frequencial, para a CQT percebemos que a energia das parcelas tonais presentes no sinal de áudio está mais bem localizada. Em compensação, a informação tonal de alta frequência está melhor representada no caso da STFT, já que, para essa região, esta transformada possui mais *bins*.

No caso do sinal de percussão, percebemos, novamente, que as parcelas tonais existentes são melhor representadas por estarem na faixa de baixas frequências. A parcela percussiva, no caso da STFT, está bem concentrada no tempo ao longo de todo o eixo frequencial. No caso da CQT, por outro lado, a esparsidade da representação é maior em altas frequências e vai diminuindo na direção da faixa de baixas frequências, onde a informação temporal destas parcelas está bastante espalhada.

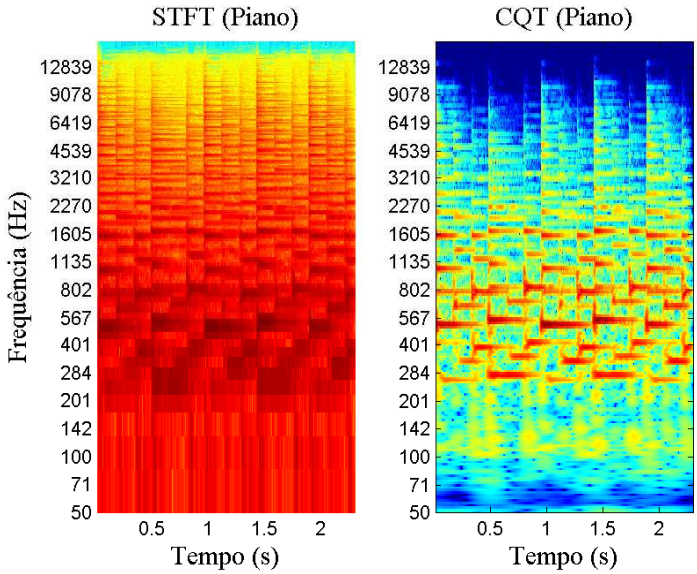


Figura 4.2: Comparação entre a STFT e a CQT de um sinal de piano sem ruído.

[6]Aqui, também utilizamos um passo de $a = 21$ para o caso da STFT, de forma a tornar mais justa a comparação entre as representações.

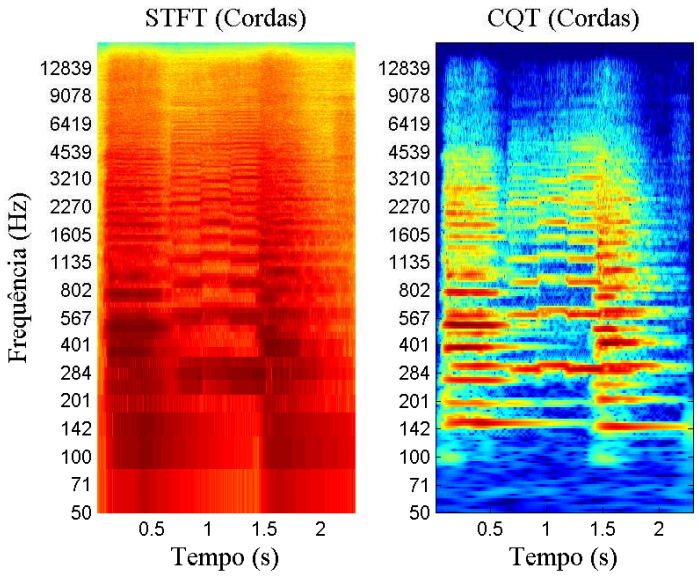


Figura 4.3: Comparação entre a STFT e a CQT de um sinal de cordas sem ruído.

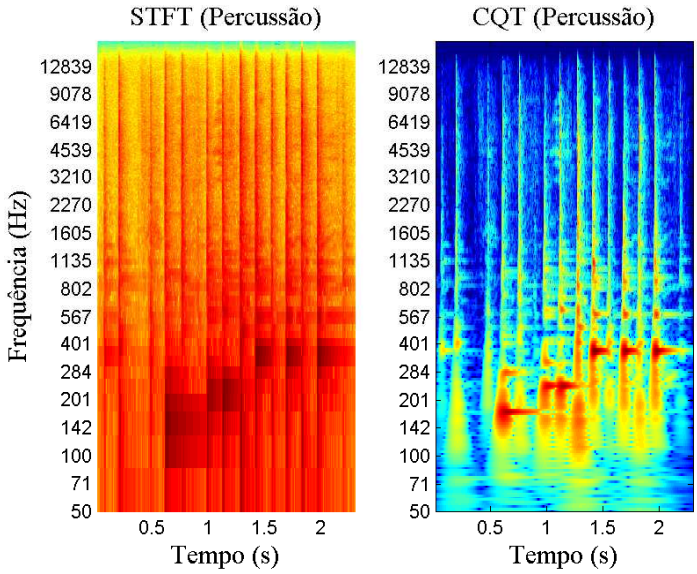


Figura 4.4: Comparação entre a STFT e a CQT de um sinal de percussão sem ruído.

O sinal de entrada $\mathbf{x}$ do algoritmo será dado por um do sinais escolhidos acima (piano, cordas e percussão) acrescentado de ruído branco. Escolhemos valores de SNR (razão sinal-ruído) iguais a 10, 20 e 30 dB. Como forma de avaliação do desempenho do algoritmo, utilizamos a SNR do sinal de saída $\hat{\mathbf{x}}$, $\mathrm{SNR}(\hat{\mathbf{x}})$, estimado como

$$\mathrm{SNR}(\hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{\|\mathbf{x}^*\|_2^2}{\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2^2} \right), \tag{4.11}$$

onde $\mathbf{x}^* \in \mathbb{R}^L$ é dado pelo sinal de áudio sem adição de ruído branco, e a nota PEAQ [33]. Esta última é uma forma de avaliação baseada em psicoacústica, na qual o sinal sem ruído estimado $\hat{\mathbf{x}}$ é comparado ao original $\mathbf{x}^*$ em um domínio perceptivo. A unidade de saída é uma ODG (do inglês, *Objective Difference Grade*)

que pode variar de -4 (degradação muito incômoda) a 0 (degradação imperceptível). A taxa de amostragem para comparação é 48 kHz e os sinais devem estar alinhados em tempo e potência.

A ideia deste trabalho é explorar diferentes tipos de vizinhança para cada sinal de áudio selecionado e cada transformada. Classificamos as vizinhanças de acordo com os seguintes aspectos: orientação, extensão e simetria. No primeiro caso, as vizinhanças podem assumir uma das três categorias a seguir: horizontal, vertical ou retangular; no segundo, as categorias são: curta, média e longa; e no terceiro, simétrica, centrada em 1/3 e assimétrica.

Dado um índice tempo-frequencial $\gamma = \{k, j\}$ e um vetor $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3, \sigma_4]^{\mathrm{T}}$, vamos definir a vizinhança N_γ como $N_\gamma(\sigma) = N_{k,j}(\sigma) = \{(k', j') | k' \in \{k - \sigma_4, k + \sigma_2\}, j' \in \{j - \sigma_3, j + \sigma_1\}\}$, ou seja, o vetor $\boldsymbol{\sigma}$ representa a extensão adicional da vizinhança na orientação $\boldsymbol{\sigma} = [\text{norte, leste, sul, oeste}]^{\mathrm{T}}$ em torno do ponto central γ. Os valores de $\boldsymbol{\sigma}$ utilizados são dados a seguir, de acordo com as classificações

- Orientação

  horizontal: $\boldsymbol{\sigma} = [0 \;\; 4 \;\; 0 \;\; 4]^{\mathrm{T}}$

  vertical: $\boldsymbol{\sigma} = [4 \;\; 0 \;\; 4 \;\; 0]^{\mathrm{T}}$

  retangular: $\boldsymbol{\sigma} = [1 \;\; 1 \;\; 1 \;\; 1]^{\mathrm{T}}$

- Extensão

  curta: $\boldsymbol{\sigma} = [0 \;\; 4 \;\; 0 \;\; 4]^{\mathrm{T}}$

  média: $\boldsymbol{\sigma} = [0 \;\; 8 \;\; 0 \;\; 8]^{\mathrm{T}}$

  longa: $\boldsymbol{\sigma} = [0 \;\; 12 \;\; 0 \;\; 12]^{\mathrm{T}}$

- Simetria

  simétrica: $\boldsymbol{\sigma} = [0 \;\; 4 \;\; 0 \;\; 4]^{\mathrm{T}}$

  centrada em 1/3: $\boldsymbol{\sigma} = [0 \;\; 2 \;\; 0 \;\; 6]^{\mathrm{T}}$

  assimétrica: $\boldsymbol{\sigma} = [0 \;\; 0 \;\; 0 \;\; 8]^{\mathrm{T}}$.

Em [12], o autor avalia, para cada sinal de áudio (piano, cordas e percussão), a SNR do sinal de saída $\hat{\mathbf{x}}$ para as classificações e subclassificações das vizinhanças. Aqui, utilizamos apenas as subclassificações que resultaram nos melhores valores de SNR. Para o sinal de piano, utilizamos vizinhança com orientação horizontal, de extensão média e centrada em 1/3. Para o sinal de cordas, vizinhança com orientação horizontal, de extensão longa e simétrica[7]. Para o sinal de percussão, vizinhança com orientação retangular, de extensão curta e centrada em 1/3.

[7]Como a vizinhança simétrica é análoga à com orientação horizontal, vamos apenas apresentar os resultados para este caso.

Resumindo, temos os seguintes parâmetros de entrada para o algoritmo de redução de ruído:

- 3 sinais de áudio: piano, cordas e percussão;
- 2 opções de transformação tempo-frequencial: STFT e CQT;
- 3 valores de SNR: 10 dB, 20 dB e 30 dB; e
- 3 classificações de vizinhança: orientação, extensão e simetria.

Realizamos, no total, 48 experimentos, já descontados os 6 relacionados ao sinal de cordas com vizinhança simétrica. A seguir, apresentamos os resultados em função dos valores de SNR e notas PEAQ em conjunto com uma análise detalhada.

4.5.3 Resultados

Nesta seção apresentaremos os resultados obtidos para os parâmetros de entrada propostos anteriormente. Mostramos as tabelas com os valores de SNR e notas PEAQ obtidos e, em seguida, analisamos os resultados.

Como foi dito, temos que garantir que os sinais a serem comparados estão alinhados temporalmente e em potência. O alinhamento temporal é mantido após o processamento. Para o alinhamento em potência, variamos a amplitude do sinal recuperado até que a nota PEAQ de saída fosse máxima. Adotamos esta como sendo a nota PEAQ atribuída ao sinal sob análise.

Outro parâmetro que devemos ajustar é o de regularização λ. Como visto no Cap. 2, ele controla o grau de esparsidade da representação obtida, ou seja, quanto maior for λ, mais esparsa será a solução. Um valor muito alto leva a perda de informação, geralmente em altas frequências, enquanto que um valor baixo irá preservar componentes relativos ao ruído. Embora numa situação prática a procura de um valor adequado para λ seja cega, já que o sinal sem ruído não é conhecido, a filosofia aqui foi descobrir qual o melhor resultado que se conseguiria em cada caso e, dessa forma, comparar diferentes configurações de forma justa. Para determinar o melhor valor de λ, utilizamos os critérios de SNR e nota PEAQ. Realizamos uma busca exaustiva com 30 valores de λ espaçados geometricamente no intervalo $[10^{-3}, 3]$. Selecionamos, então, os valores correspondentes às melhores SNR e nota PEAQ obtidas. Neste, adicionamos um critério extra: a SNR do sinal obtido deve ser, no mínimo, igual à SNR de entrada. Como resultado, obtemos λ_{SNR} e λ_{PEAQ}, respectivamente. A Fig 4.5 mostra um exemplo dessa busca, onde $\lambda_{\text{SNR}} = 0,0275$ e $\lambda_{\text{PEAQ}} = 0,0052$. Percebemos, com este caso, que os valores selecionados de λ selecionados são, em geral, diferentes.

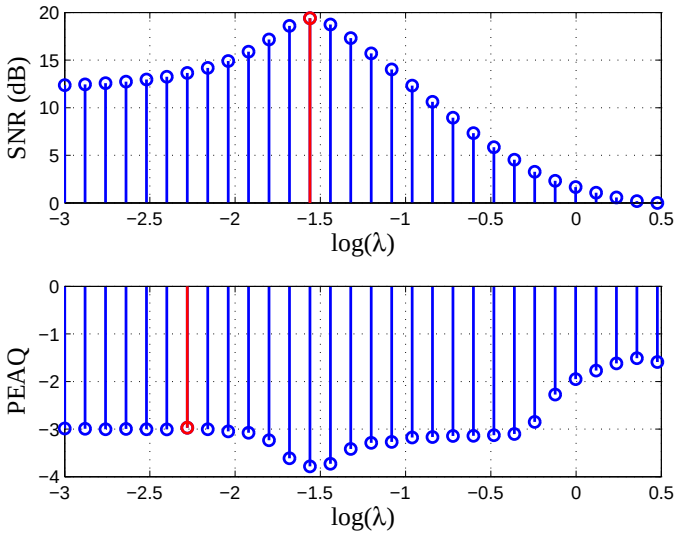


Figura 4.5: Valores de SNR (primeira linha) e notas PEAQ (segunda linha) obtidos para um sinal de piano corrompido com 10 dB de SNR. Os pontos em vermelho correspondem aos máximos de SNR e nota PEAQ.

Na Fig. 4.5, podemos notar também que a SNR obtida aumenta, a princípio, com o aumento de λ até alcançar seu máximo em $19,4$ dB. Em seguida, a SNR decai até 0 dB. A partir de $\log_{10}\lambda = -0,3622$ (que corresponde a $\lambda = 0,4343$), o valor da nota PEAQ aumenta drasticamente, indicando, supostamente, resultados perceptivamente parecidos ao sinal original sem ruído. Tal suposição é, no entanto, equivocada, pois a partir deste ponto, devido aos valores elevados do coeficiente de esparsidade λ, partes do sinal recuperado foram completamente zeradas. De alguma forma, isso impede que o PEAQ (que foi originalmente concebido para comparar diferenças leves causadas por codificação) forneça uma nota perceptiva coerente com o que se ouve.

A primeira bateria de experimentos utiliza 30 dB como SNR de entrada. Os valores de SNR e notas PEAQ obtidos para as transformadas STFT e CQT são dados nas tabelas 4.1 (SNR e STFT), 4.2 (SNR e CQT), 4.3 (PEAQ e STFT) e 4.4 (PEAQ e CQT).

Comparando os resultados mostrados nas Tabelas 4.1 a 4.4, percebemos que os valores de SNR do sinal recuperado $\hat{\mathbf{x}}$ são maiores para a transformada STFT. Essa diferença é mais acentuada para o sinal de piano, onde ela é ligeiramente superior a 2 dB. No caso das notas PEAQ, consideramos que os resultados são equivalentes para os sinais de piano e de percussão e ligeiramente piores no caso da CQT para o sinal de cordas. Percebemos que, para este valor de SNR de entrada, o algoritmo não é capaz de reduzir o ruído de forma significativa sem prejudicar o timbre do instrumento. Assim sendo, a nota PEAQ é essencialmente a mesma e o valor de λ_{PEAQ} é pequeno, o que significa que poucas alterações foram realizadas sobre o

espectro do sinal ruidoso.

A segunda bateria de experimentos utiliza 20 dB como SNR de entrada. Os valores de SNR e notas PEAQ obtidos para as transformadas STFT e CQT são dados nas tabelas 4.5 (SNR e STFT), 4.6 (SNR e CQT), 4.7 (PEAQ e STFT) e 4.8 (PEAQ e CQT).

Comparando os resultados mostrados nas Tabelas 4.5 a 4.8, percebemos que, também neste caso, os valores de SNR do sinal $\hat{\mathbf{x}}_{\mathrm{STFT}}$ são maiores que os do sinal $\hat{\mathbf{x}}_{\mathrm{CQT}}$, porém, com uma diferença maior, de quase 3 dB. Agora, as notas PEAQ obtidas com a transformada CQT foram ligeiramente piores para os sinais de cordas e de percussão e equivalentes no caso do sinal de piano. Aqui, notamos que as notas PEAQ obtidas já são melhores que as do sinal corrompido nos casos dos sinais de cordas (com a STFT) e de percussão (para ambas as transformadas).

A terceira e última bateria de experimentos utiliza 10 dB como SNR de entrada. Os valores de SNR e notas PEAQ obtidos para as transformadas STFT e CQT são dados nas tabelas 4.9 (SNR e STFT), 4.10 (SNR e CQT), 4.11 (PEAQ e STFT) e 4.12 (PEAQ e CQT).

Tabela 4.1: SNR obtida para a STFT e SNR de entrada igual a 30dB

	Cordas	Piano	Percussão
Orientação	$32,5$	$35,2$	$33,6$
Extensão	$32,6$	$35,2$	$33,0$
Simetria	—	$35,0$	$33,5$

Tabela 4.2: SNR obtida para a CQT e SNR de entrada igual a 30dB

	Cordas	Piano	Percussão
Orientação	$31,3$	$32,7$	$32,4$
Extensão	$31,4$	$32,8$	$32,5$
Simetria	—	$32,7$	$32,4$

Tabela 4.3: PEAQ obtida para a STFT e SNR de entrada igual a 30dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,4}$	$\mathbf{-2,4}$	$\mathbf{-2,9}$
Orientação	$-3,3$	$-2,4$	$-2,9$
Extensão	$-3,2$	$-2,5$	$-2,9$
Simetria	—	$-2,4$	$-2,9$

Tabela 4.4: PEAQ obtida para a CQT e SNR de entrada igual a 30dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,4}$	$\mathbf{-2,4}$	$\mathbf{-2,9}$
Orientação	$-3,4$	$-2,4$	$-2,9$
Extensão	$-3,4$	$-2,4$	$-2,9$
Simetria	—	$-2,4$	$-2,9$

Tabela 4.5: SNR obtida para a STFT e SNR de entrada igual a 20dB

	Cordas	Piano	Percussão
Orientação	$27,2$	$27,1$	$27,3$
Extensão	$27,2$	$27,2$	$26,5$
Simetria	—	$27,1$	$27,1$

Tabela 4.6: SNR obtida para a CQT e SNR de entrada igual a 20dB

	Cordas	Piano	Percussão
Orientação	$24,4$	$23,7$	$24,5$
Extensão	$24,8$	$23,9$	$24,7$
Simetria	—	$23,7$	$24,5$

Tabela 4.7: PEAQ obtida para a STFT e SNR de entrada igual a 20dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,5}$	$\mathbf{-2,7}$	$\mathbf{-3,1}$
Orientação	$-3,2$	$-2,7$	$-2,6$
Extensão	$-3,2$	$-2,7$	$-2,5$
Simetria	—	$-2,6$	$-2,6$

Tabela 4.8: PEAQ obtida para a CQT e SNR de entrada igual a 20dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,5}$	$\mathbf{-2,7}$	$\mathbf{-3,1}$
Orientação	$-3,5$	$-2,6$	$-2,8$
Extensão	$-3,4$	$-2,6$	$-2,8$
Simetria	—	$-2,6$	$-2,8$

Tabela 4.9: SNR obtida para a STFT e SNR de entrada igual a 10dB

	Cordas	Piano	Percussão
Orientação	$20,0$	$19,4$	$20,7$
Extensão	$20,5$	$19,4$	$20,2$
Simetria	—	$19,2$	$20,5$

Tabela 4.10: SNR obtida para a CQT e SNR de entrada igual a 10dB

	Cordas	Piano	Percussão
Orientação	$16,3$	$15,1$	$16,3$
Extensão	$16,8$	$15,4$	$16,5$
Simetria	—	$15,2$	$16,3$

Tabela 4.11: PEAQ obtida para a STFT e SNR de entrada igual a 10dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,7}$	$\mathbf{-3,0}$	$\mathbf{-3,6}$
Orientação	$-3,2$	$-3,0$	$-2,1$
Extensão	$-3,2$	$-3,0$	$-2,1$
Simetria	—	$-3,0$	$-2,1$

Tabela 4.12: PEAQ obtida para a CQT e SNR de entrada igual a 10dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,7}$	$\mathbf{-3,0}$	$\mathbf{-3,6}$
Orientação	$-2,6$	$-2,8$	$-2,2$
Extensão	$-2,9$	$-2,8$	$-2,2$
Simetria	—	$-2,8$	$-2,3$

Comparando, por fim, os resultados mostrados nas Tabelas 4.9 a 4.12, percebemos que, como nos casos anteriors, os valores de SNR do sinal $\hat{\mathbf{x}}_{\text{STFT}}$ são maiores que os do sinal $\hat{\mathbf{x}}_{\text{CQT}}$. Agora, a diferença é ainda maior que nos dois últimos casos: em torno de 4 dB. As notas PEAQ obtidas com a transformada CQT foram ligeiramente melhores para os sinais de cordas e de piano. Para o caso do sinal de percussão, devemos analisar com cuidado as notas PEAQ obtidas. Como foi dito anteriormente, para valores de λ suficientemente elevados, trechos do sinal são removidos e apenas os remanescentes são comparados ao sinal limpo. Neste caso, por se tratar de um sinal essencialmente percussivo com lacunas longas, a SNR do sinal obtido se manteve relativamente alta (maior ou igual à SNR de entrada, igual a 10 dB) mesmo com os cortes. Assim sendo, devemos desconsiderar tais resultados, pois, na prática, cortar trechos do sinal ruidoso não é uma alternativa válida.

De forma resumida, percebemos que os resultados obtidos por meio da STFT são maiores em termos de SNR e relativamente equivalentes em termos de nota PEAQ se comparados aos obtidos por meio da CQT. Perceptivamente, escutamos a presença de ruído de alta frequência para os sinais recuperados através da transformada CQT. Como vimos no Cap. 2, sabemos que os *bins* de frequência para esta transformada são geometricamente espaçados e que, portanto, o número de *bins* utilizados para representar altas frequências é menor que no caso da STFT, onde estes apresentam um espaçamento linear. A Fig. 4.6 mostra os módulos das representações tempo-frequenciais de um sinal composto unicamente por ruído branco $\mathbf{x}_{\text{RB}}$. É possível

perceber, claramente, que os *bins* correspondentes às altas frequências para o caso da CQT contêm mais energia que para o caso da STFT, justamente para compensar o espaçamento inerente a essa transformada.

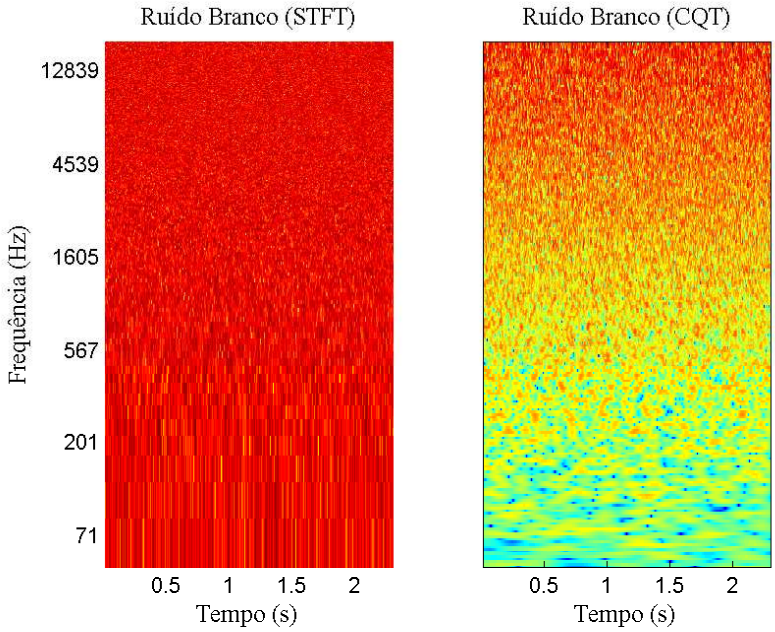
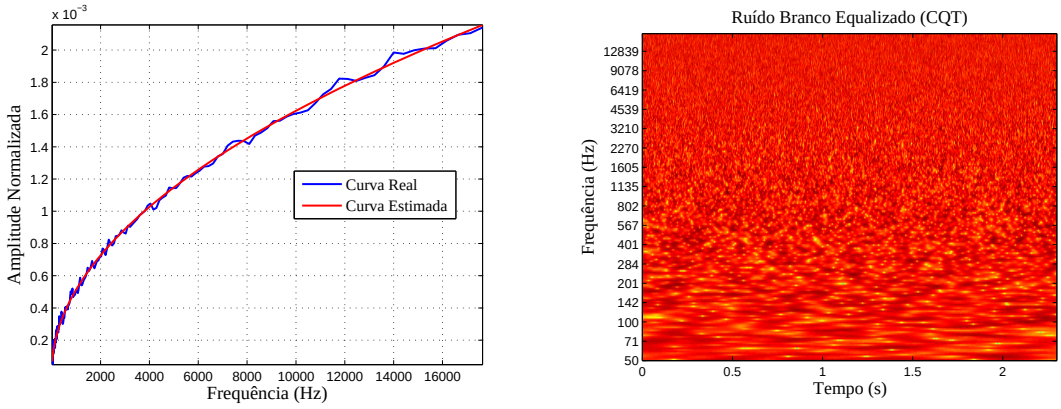


Figura 4.6: Representações de ruído branco utilizando a STFT (primeira coluna) e a CQT (segunda coluna).



(a) Curva de equalização estimada.

(b) Representação equalizada.

Figura 4.7: Curva de equalização estimada, em vermelho, a partir de uma observada, em azul (a); e representação equalizada de ruído branco utilizando a CQT (b).

O algoritmo que estamos utilizando, no entanto, não leva em consideração esse conhecimento prévio da CQT. Ele pressupõe que o sinal sob análise está corrompido com ruído branco e que este se espalha de forma uniforme sobre o plano tempo-frequencial. De forma a contornar este problema, realizamos uma operação de pré-ênfase neste plano, equalizando a energia presente em todos os *bins*. Para isso,

multiplicamos cada coluna da matriz de coeficientes por uma curva estimada[8], dada, em vermelho, na Fig. 4.7a. A curva em azul é a média dos módulos das amplitudes normalizadas dos coeficientes de cada coluna da matriz de transformação do sinal $\mathbf{x}_{\mathrm{RB}}$ calculada a partir da CQT. A Fig. 4.7b mostra a representação resultante do sinal $\mathbf{x}_{\mathrm{RB}}$, após a pré-ênfase.

Decidimos repetir uma das três baterias de testes realizados até então para a transformada CQT com equalização, chamada, por conveniência, de CQT+. Escolhemos a SNR de entrada igual a 20 dB. Os valores de SNR e notas PEAQ obtidos são dados nas tabelas 4.13 (SNR e CQT+) e 4.14 (PEAQ e CQT+).

Tabela 4.13: SNR obtida para a CQT+ e SNR de entrada igual a 20dB

	Cordas	Piano	Percussão
Orientação	$26,8$	$26,3$	$27,9$
Extensão	$27,3$	$26,5$	$27,9$
Simetria	—	$26,2$	$27,9$

Tabela 4.14: PEAQ obtida para a CQT+ e SNR de entrada igual a 20dB

	Cordas	Piano	Percussão
PEAQ original	$\mathbf{-3,5}$	$\mathbf{-2,7}$	$\mathbf{-3,1}$
Orientação	$-3,3$	$-2,7$	$-2,6$
Extensão	$-3,2$	$-2,7$	$-2,5$
Simetria	—	$-2,7$	$-2,6$

Agora, comparamos os resultados mostrados na Tabela 4.13 com os da Tabela 4.6. Vemos que os valores de SNR obtidos aumentaram entre 2,4 e 3,4 dB e podem ser considerados compatíveis aos valores apresentados na Tabela 4.5. Comparamos, também, os resultados mostrados na Tabela 4.14 com os da Tabela 4.8. Para os sinais de cordas e percussão, as notas PEAQ obtidas foram ligeiramente melhores, comparáveis, novamente, às da Tabela 4.7.

Vamos analisar auditivamente os sinais recuperados por meio da CQT+ em termos das melhores notas PEAQ. Escolhemos, inicialmente, o sinal de cordas e a vizinhança classificada segundo a sua extensão. Somos capazes de perceber diferenças consideráveis entre os resultados obtidos para os dois casos: CQT sem e com equalização. No primeiro caso, o ruído remanescente contém alguma informação de alta frequência, proveniente de parcelas de ruído com algum grau de correlação entre si. Este, portanto, não é uniforme e prejudica a qualidade do sinal recuperado.

[8] A curva estimada é da forma $K\sqrt{f}$, onde K é um valor que depende da variância do ruído presente e f é a frequência em Hz.

Por outro lado, a parcela de sinal restante é razoavelmente brilhante, pois ainda mantém parte dos harmônicos superiores. No segundo caso, o ruído é eliminado por completo, porém, em conjunto com uma parcela de alta frequência do sinal original. O mesmo ocorre para o sinal de percussão e mesma vizinhança. No caso do sinal de piano, a mudança perceptiva é apenas na natureza do ruído. Este se torna mais uniforme e soa de forma mais natural.

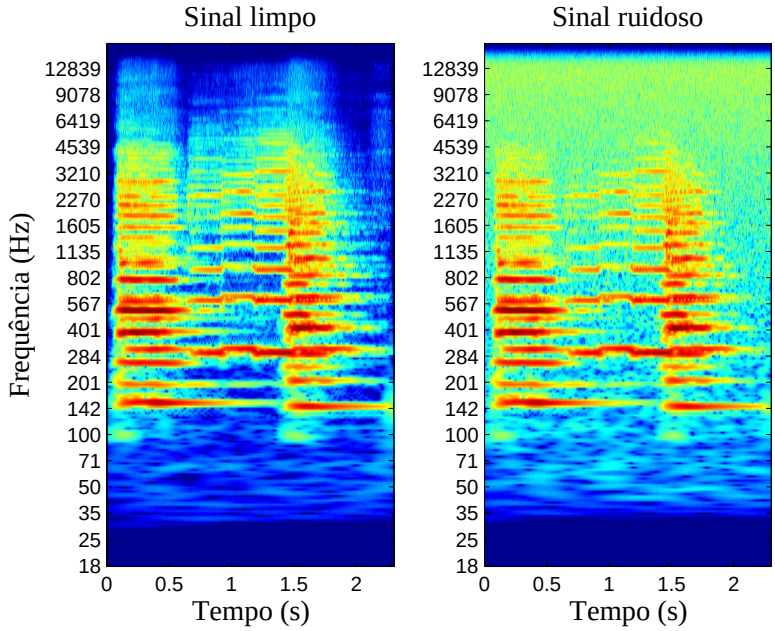


Figura 4.8: Representações do sinal de cordas sem (primeira coluna) e com a presença de ruído branco (segunda coluna) utilizando a CQT.

Escolhemos um dos sinais acima para comparar visualmente as representações geradas. Para isso, utilizamos o sinal de cordas, com a vizinhança classificada segundo a sua extensão. A Fig. 4.8 mostra as representações dos sinais original (primeira coluna) e corrompido (segunda coluna) obtidas por meio da CQT. O sinal corrompido $\mathbf{x}$ é processado e duas versões recuperadas são geradas: $\hat{\mathbf{x}}_{\text{SNR}}$ e $\hat{\mathbf{x}}_{\text{PEAQ}}$ a partir de dois graus de esparsidade λ de forma a otimizar, como foi dito, a SNR e a nota PEAQ de saída. A Fig. 4.9 mostra os valores de SNR (primeira linha) e notas PEAQ (segunda linha) de saída para diferentes λ. Os pontos máximos são dados em vemelho. A Fig. 4.10 mostra as representações de $\hat{\mathbf{x}}_{\text{SNR}}$ (primeira coluna) e $\hat{\mathbf{x}}_{\text{PEAQ}}$ (segunda coluna) após a etapa de limiarização. Por fim, a Fig. 4.11 mostra as representações obtidas dos sinais recuperados $\hat{\mathbf{x}}_{\text{SNR}}$ (primeira coluna) e $\hat{\mathbf{x}}_{\text{PEAQ}}$ (segunda coluna) no domínio no tempo, ou seja, calculamos a transformada inversa a partir das representações limiarizadas, obtendo $\hat{\mathbf{x}}_{\text{SNR}}$ e $\hat{\mathbf{x}}_{\text{PEAQ}}$, e, em seguida, calculamos a transformada novamente, resultado nas repesentações recuperadas.

Pelas Figs. 4.10 e 4.11, percebemos que, para as representações do sinal $\hat{\mathbf{x}}_{\text{SNR}}$ (primeira coluna), boa parte do ruído de alta frequência ainda está presente no sinal recuperado. Um motivo possível é o fato de que esse ruído pode não ser

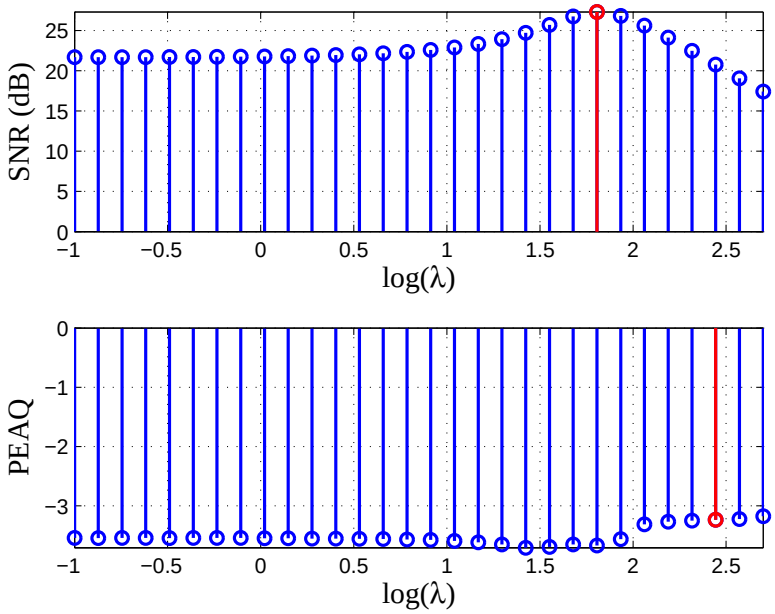


Figura 4.9: Valores de SNR (primeira linha) e notas PEAQ (segunda linha) obtidos para um sinal de cordas corrompido com 20 dB de SNR. Os pontos em vermelho correspondem aos máximos de SNR e nota PEAQ.

completamente descorrelacionado. O algoritmo analisa o grau de correlação entre os coeficientes dentro de uma dada vizinhança e leva essa informação em consideração ao limiarizá-los. Quanto maior for a interdependência entre eles, menor será o grau de limiarização. Assim sendo, coeficientes que possuem um certo grau de correlação entre si e são correspondentes às parcelas de ruído presentes no sinal podem não ser eliminados. Estes serão percebidos auditivamente na forma de ruído musical.

As representações obtidas para o sinal $\hat{\mathbf{x}}_{\text{PEAQ}}$, por outro lado, não contém informação significativa de altas frequências, o que pode ser visto, também, nas Figs. 4.10 e 4.11 (segunda coluna). Perdemos, inclusive, informações relativas ao sinal de áudio original, o que compromete o timbre e o brilho do sinal de cordas recuperado. Perceptivamente, no entanto, preferimos eliminar o ruído de alta frequência remanescente, mesmo que em troca de perda de parte da informação do sinal original.

Um outro teste que poderíamos propor é alterar o número de *bins* por oitava da CQT, até então igual a 24. Tal mudança não altera o tamanho da matriz de coeficientes gerada e, portanto, também não irá alterar significativamente o custo computacional do algoritmo. Sabemos, a partir dos parâmetros utilizados para cada transformada, que o número de coeficientes P obtidos por meio da CQT é aproximadamente seis vezes maior que para o caso da STFT. Além disso, temos que considerar que o algoritmo utilizado (FISTA) realiza, a cada iteração, operações de ida e volta para o domínio do tempo, o que requer que a transformada (e a volta) seja calculada com eficiência. De forma a melhorar o desempenho da CQT

nesse sentido, poderíamos pensar em abrir mão da sobreposição temporal utilizada, atualmente dada por $a = 21$. Tal mudança reduziria o número de coeficientes calculados e, consequentemente, a complexidade computacional.

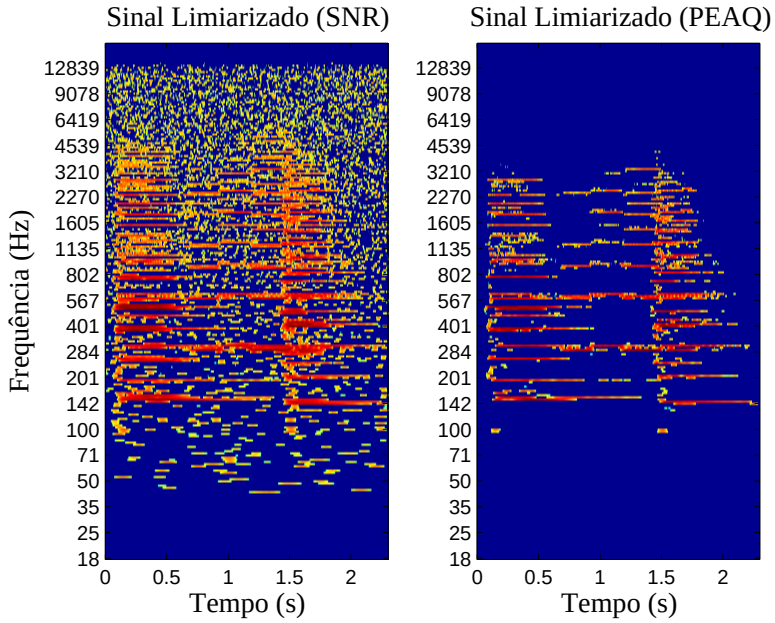


Figura 4.10: Representações limiarizadas do sinal de cordas corrompido com ruído branco a 20 dB de SNR para diferentes graus de esparsidade utilizando a CQT. Na primeira coluna, escolhemos otimizar a SNR de saída e, na segunda, a nota PEAQ obtida para o sinal recuperado.

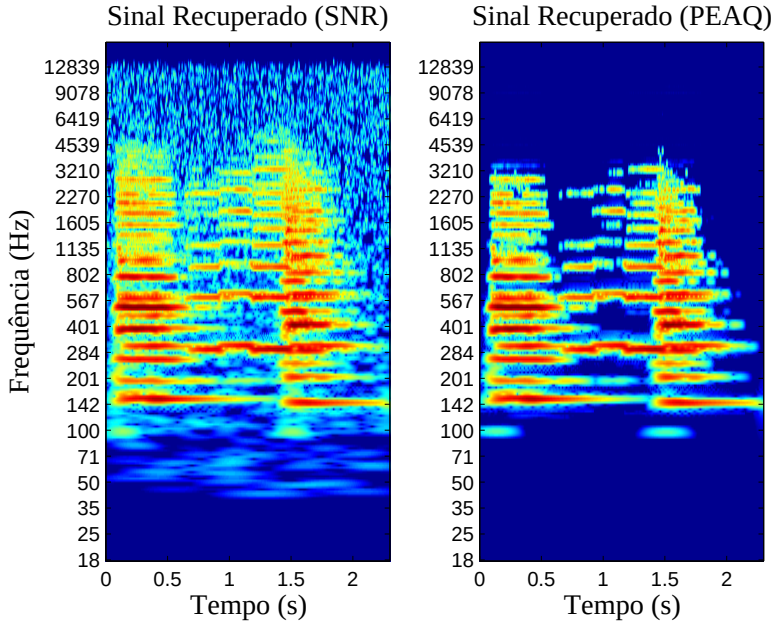


Figura 4.11: Representações recuperadas do sinal de cordas corrompido com ruído branco a 20 dB de SNR para diferentes graus de esparsidade utilizando a CQT. Na primeira coluna, escolhemos otimizar a SNR de saída e, na segunda, a nota PEAQ obtida para o sinal recuperado.

Se, por um lado, reduzir o salto temporal ao calcularmos a CQT acarretaria em uma melhora no desempenho do algoritmo, por outro, comprometeríamos a reconstrução do sinal temporal a partir da transformação inversa[9]. Uma possibilidade seria explorar técnicas mais recentes que permitem a reconstrução perfeita do sinal temporal a partir da sua transformada CQT [34]. Dessa forma, além da perda de informação a cada iteração do algoritmo FISTA, considerada irrelevante até então, poderíamos trabalhar no sentido de reduzir a sua dimensão e, assim, o tempo de processamento.

Por fim, é importante ressaltar que a questão da vizinhança não foi suficientemente explorada neste trabalho. Os resultados apresentados em termos da SNR e notas PEAQ são, para todos os casos abordados, muito semelhantes entre si, não dando margem a interpretações adicionais.

4.6 Considerações Finais

Neste Capítulo, estudamos uma técnica que utiliza informações sobre a estrutura interna do sinal sob análise. Abordamos, como principal aplicação, a redução de ruído em sinais de áudio corrompidos. Neste processo, levamos em consideração que os sinais podem ser persistentes tanto temporal, quanto frequencialmente, que é o caso dos sinais tonais e percussivos, respectivamente. Dessa forma, o ruído de fundo presente, considerado branco, é supostamente eliminado neste processo.

Os dicionários utilizados para realizar os experimentos foram a STFT e a CQT. Após os experimentos realizados há, no geral, muitos argumentos que favorecem o uso da STFT no lugar da CQT. Os valores obtidos de SNR são, de forma geral, maiores para esta transformada. Se considerarmos, no entanto, que vamos utilizar a CQT+, ainda assim podemos argumentar contra. Basta mencionar o custo computacional relativamente elevado desta transformada. O que ainda não levamos em consideração, no entanto, é o tipo de ruído presente em sinais de áudio. Este raramente ocupa toda a faixa de frequências, e sua amplitude geralmente decai com a frequência. Assim sendo, a informação de ruído de alta frequência que permanece, em boa parte dos casos, nos sinais recuperados a partir da transformada CQT, pode ser razoavelmente menor e, possivelmente, mascarada pela informação de baixa frequência.

[9]Com os parâmetros utilizamos, somos capazes de recuperar o sinal temporal original a partir da sua transformação com uma SNR de aproximadamente 37 dB.

Capítulo 5

Conclusões

Este trabalho reuniu um conjunto de técnicas utilizadas dentro do contexto de "Análise Tempo-Frequência". Primeiramente, no Cap. 2, apresentamos a teoria por trás de tais técnicas, englobando a ideia de decomposições atômicas de sinais e recuperação esparsa. No primeiro caso, falamos das representações por *frames* e as transformadas STFT e CQT. Já no segundo, exploramos o conceito de limiarização suave, expandindo-o para o algorimo de limiarização iterativa no caso de dicionários redundantes, e aproximações estruturadas.

Em seguida, no Cap. 3, definimos a FChT e falamos em detalhe sobre sua implementação. Algumas extensões e melhorias foram propostas; são elas:

- Extensão do modelo linear da fundamental para um não-linear;
- Troca da função de saliência até então utilizada por uma independente do timbre do sinal[1]; e
- Inclusão de um modelo de inarmonicidade no cálculo da nova saliência.

Aqui, ressaltamos também que, no caso da nova função de saliência utilizada, retirada de [25], algumas mudanças foram implementadas de forma a, não só integrar o uso desta saliência com a FChT, mas também melhorar o seu desempenho. Os resultados obtidos para cada etapa foram, nessa ordem, representações mais esparsas para sinais com rápidas flutuações em frequência; melhoria na estimação da taxa de inclinação α, especialmente para sinais corrompidos com ruído de fundo; e melhoria na estimação de α e representação mais esparsa para sinais com um certo grau de inarmonicidade.

No Cap. 4, por fim, abordamos o tema "Esparsidade Estruturada", onde apresentamos algumas definições, como normas mistas e vizinhança, explicamos o algoritmo utilizado e realizamos alguns experimentos de forma a reduzir o ruído de fundo em sinais de áudio. Utilizamos a minimização LASSO, ou seja, a norma ℓ_1 como critério

[1] Essa melhoria somente foi implementada para o caso da FChT com deformação linear.

de esparsidade. No artigo original [12], o dicionário utilizado é a STFT e a medida de comparação entre resultados é a SNR obtida do sinal sem ruído estimado. Aqui, porém, propusemos as seguintes modificações:

- Uso da CQT no lugar da STFT no algoritmo de redução de ruído;
- Curva de pré-equalização da CQT para uniformizar a distribuição do ruído branco nesta transformada; e
- Uso da nota PEAQ como medida auxiliar para comparar resultados.

Os resultados para as transformadas utilizadas foram analisados, como dito, em termos dos valores de SNR e notas PEAQ máximos obtidos. Tais valores foram considerados análogos para ambos os casos.

5.1 Próximas Etapas do Trabalho

Os resultados obtidos para as modificações propostas nesta dissertação nos permitiram perceber possíveis melhorias adicionais nas técnicas estudadas. As principais delas são dadas a seguir:

- Checar a presença de *aliasing* para o caso da FChT com deformação não-linear;
- Análise mais detalhada da amostragem dos seguintes parâmetros: taxa de inclinação α, curvatura β e fator de inarmonicidade η;
- Proposta de busca informada por parâmetros que substitua a atual busca exaustiva;
- Integração da nova função de saliência proposta com a FChT com deformação não-linear;
- Manipulação dos parâmetros da CQT (número de *bins* por oitava e sobreposição temporal) no método empregado para redução de ruído de fundo em sinais de áudio;
- Explorar técnicas de reconstrução perfeita do sinal temporal a partir da sua CQT e integrá-las com a técnica de redução de ruído empregada;
- Explorar o conceito apresentado de vizinhança mais a fundo;
- Realização de testes subjetivos sistemáticos e busca por outras ferramentas objetivas mais apropriadas para avaliar o sinal recuperado como, por exemplo, o PAQM (do inglês, *Perceptual Audio Quality Measure*) [35]; e

- Comparar a técnica utilizada para a redução de ruído de fundo em sinais de áudio com outras já consolidadas na literatura.

Tais e outras ideias devem ser exploradas ao longo de uma tese de DSc.

Referências Bibliográficas

[1] COHEN, L. *Time-Frequency Analysis*. Englewood Cliffs, EUA, Prentice Hall, 1995.

[2] DINIZ, P., DA SILVA, E., NETTO, S. *Processamento Digital de Sinais*. 2 ed. Porto Alegre, Brasil, Bookman, 2014.

[3] BROWN, J. C. "Calculation of a constant Q spectral transform", *Journal of the Acoustical Society of America*, v. 80, n. 1, pp. 425–434, janeiro 1991.

[4] KASHIMA, K. L., MONT-REYNAUD, B. *The bounded Q approach to time-varying spectral analysis*. Technical Report STAN-M-28, Department of Music, Standford University, Stanford, EUA, 1985.

[5] FLANDRIN, P. *Time-Frequency/Time-Scale Analysis*. San Diego, EUA, Academic Press, 1999.

[6] BURRUS, C., GOPINATH, R., GUO, H. *Introduction to Wavelets and Wavelet Transforms—A Primer*. Upper Saddle River, EUA, Prentice Hall, 1998.

[7] TYGEL, A. F., BISCAINHO, L. W. P. "Sound source separation via nonnegative matrix factor 2-D deconvolution using linearly sampled spectrum". In: *Anais do VII Congresso Brasileiro de Engenharia de Áudio*, pp. 58–65, São Paulo, Brasil, maio 2009. AES-Brasil.

[8] CANCELA, P., LÓPEZ, E., ROCAMORA, M. "Fan-chirp transform for music representation". In: *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-10)*, pp. 1–8, Graz, Áustria, setembro 2010.

[9] KLAPURI, A., DAVY, M. *Signal Processing Methods for Music Transcription*. Nova Iorque, EUA, Springer, 2006.

[10] ESQUEF, P. A. A., BISCAINHO, L. W. P. "Spectral-Based Analysis and Synthesis of Audio Signals". In: Pérez-Meana, H. (Ed.), *Advances in Audio and Speech Signal Processing: Technologies and Applications*, pp. 56–92, Hershey, EUA, fevereiro 2007. IGI Global.

[11] Zölzer, U. (Ed.). *Digital Audio Effects*. 2 ed. Chichester, Reino Unido, Wiley, 2011.

[12] SIEDENBURG, K., DORFLER, M. "Audio denoising by generalized time-frequency thresholding". In: *Proceedings of the AES 45th International Conference*, Helsinque, Finlândia, março 2012. AES.

[13] KERELIUK, C., DEPALLE, P. "Sparse atomic modeling of audio: a review". In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, pp. 81–92, Paris, França, setembro 2011.

[14] SCHORKHUBER, C., KLAPURI, A. "Constant-Q Transform Toolbox for Music Processing". In: *Proceedings of the 7th Sound and Music Computing Conference*, Barcelona, Espanha, julho 2010.

[15] KOVAČEVIĊ, J., CHEBIRA, A. "Life beyond bases: the advent of frames (part I)", *IEEE Signal Processing Magazine*, v. 24, n. 4, pp. 86–104, julho 2007.

[16] SIEDENBURG, K. *Structured Sparsity in Time-Frequency Analysis*. Tese de D.Sc., Instituto de Matemática, Universidade Humboldt de Berlim, Berlim, Alemanha, 2011.

[17] CHEN, S., DONOHO, D., SAUNDERS, M. "Atomic decomposition by basis pursuit", *SIAM Journal of Scientific Computing*, v. 43, n. 1, pp. 129–159, fevereiro 2001.

[18] TIBSHIRANI, R. "Regression shrinkage and selection via the LASSO", *Journal of the Royal Statistical Society (Series B)*, v. 58, n. 1, pp. 267–288, 1996.

[19] ELAD, M. "Why simple shrinkage is still relevant for redundant representations?" *IEEE Transactions on Information Theory*, v. 52, n. 12, pp. 5559–5569, dezembro 2006.

[20] DAUBECHIES, I., DEFRISE, M., MOL, C. D. "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint", *Communications on Pure Applied Mathematics*, v. 57, n. 11, pp. 1413–1457, novembro 2004.

[21] LORIS, I. "On the performance of algorithms for the minimization of ℓ_1-penalized functionals", *Inverse Problems*, v. 25, n. 3, pp. 1–16, março 2009.

[22] WERUAGA, L., KÉPESI, M. "The fan-chirp transform for non-stationary harmonic signals", *Signal Processing*, v. 87, n. 6, pp. 1504–1522, junho 2007.

[23] APOLINÁRIO, I. F., BISCAINHO, L. W. P., ROCAMORA, M., et al. "Fan-chirp transform with nonlinear time warping". In: *Anais do 13o. Congresso de Engenharia de Áudio*, pp. 62–68, São Paulo, Brasil, maio 2015.

[24] KÉPESI, M., WERUAGA, L. "Adaptive chirp-based time–frequency analysis of speech signals", *Speech Communication*, v. 48, n. 5, pp. 474–492, maio 2006.

[25] DEGANI, A., LEONARDI, R., MIGLIORATI, P., et al. "A pitch salience function derived from harmonic frequency deviations for polyphonic music analysis". In: *Proceedings of the 17th Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Alemanha, setembro 2014.

[26] FLETCHER, H., BLACKHAM, E. D., STRATTON, R. "Quality of piano tones", *Journal of Acoustical Society of America*, v. 34, n. 6, pp. 749–761, junho 1962.

[27] DOWNIE, J. "The music information retrieval evaluation exchange (2005-2007): a window into music information retrieval research", *Acoustical Science and Technology*, v. 28, n. 4, pp. 247–255, setembro 2008.

[28] DE O. NUNES, L., ESQUEF, P. A. A., BISCAINHO, L. W. P. "Evaluation of threshold-based algorithms for detection of spectral peaks in audio". In: *Anais do V Congresso de Engenharia de Áudio*, pp. 66–73, São Paulo, Brasil, Maio 2007. AES Brasil.

[29] DRESSLER, K., STREICH, S. "Tuning frequency estimation using circular statistics". In: *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, Viena, Áustria, setembro 2007.

[30] LOEFFLER, B. D. *Instrument Timbres and Pitch Estimation in Polyphonic Music*. Tese de M.Sc., School of Electrical and Computer Engineering, Georgia Institute of Technology, Geórgia, EUA, 2006.

[31] SIEDENBURG, K., DORFLER, M. "Structured sparsity for audio signals". In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, Paris, França, setembro 2011.

[32] KOWALSKI, M., TORRÉSANI, B. "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients", *Signal, Image, and Video Processing*, v. 3, n. 3, pp. 251–264, novembro 2009.

[33] THIEDE, T., TREURNIET, W. C., BITTO, R., et al. "PEAQ–The ITU standard for objective measurement of perceived audio quality", *Journal of the Audio Engineering Society*, v. 48, n. 1/2, pp. 3–29, janeiro/fevereiro 2000.

[34] DÖRFLER, M., HOLIGHAUS, N., GRILL, T., et al. "Constructing an invertible constant-Q transform with nonstationary gabor frames". In: *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, Paris, França, setembro 2011.

[35] BEERENDS, J. G. "Audio Quality Determination Based on Perceptual Measurement Techniques". In: Kahrs, M., Brandenburg, K. (Eds.), *Applications of Digital Signal Processing to Audio and Acoustics*, pp. 1–37, Norwell, EUA, 1998. Kluwer.