REPRESENTAÇÃO ESPECTRAL DE SINAIS PARA
TRANSCRIÇÃO MUSICAL AUTOMÁTICA

Cristiano Nogueira dos Santos

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. Sergio Lima Netto, Ph.D.

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. Marcio Nogueira de Souza, D.Sc.

Prof. Jacques Szczupak, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
DEZEMBRO DE 2004

DOS SANTOS, CRISTIANO NOGUEIRA

Representação espectral de sinais para transcrição musical automática [Rio de Janeiro] 2004

XII, 73 pp., 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2004)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1.Transcrição musical 2.Bancos de filtros 3.FFB 4.CQFFB 5.BQFFB

I.COPPE/UFRJ II.Título (série)

Agradecimentos

Agradeço aos meus pais, Danilo Sacramento dos Santos e Célia Nogueira dos Santos, à minha namorada Helena Hiroko Otsuka, à minha avó Jandira Sacramento, à minha irmã Simone Nogueira dos Santos Neves e ao meu cunhado Vinicius Couto Neves pelo carinho, compreensão e auxílio nos momentos mais difíceis. Sem eles este trabalho não se realizaria. A todos os meus parentes e amigos agradeço pelo apoio indispensável.

Agradeço ainda aos meus colegas de trabalho que, sendo fontes inesgotáveis de ajuda e amizade, foram co-responsáveis pelo bom andamento deste trabalho. Em especial, agradeço a Luiz Wagner Pereira Biscainho, Sergio Lima Netto, Fábio Pacheco Freeland e Paulo Antônio Andrade Esquef.

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

REPRESENTAÇÃO ESPECTRAL DE SINAIS PARA TRANSCRIÇÃO MUSICAL AUTOMÁTICA

Cristiano Nogueira dos Santos

Dezembro/2004

Orientadores: Sergio Lima Netto
Luiz Wagner Pereira Biscainho

Programa: Engenharia Elétrica

A transcrição musical automática (TMA) tem como objetivo identificar as notas musicais presentes em um sinal de música. Entre as aplicações da TMA, estão o estudo de composições, a edição de gravações e a codificação de sinais de música.

Para a obtenção do espectro do sinal, é comum utilizar-se a DFT (*discrete Fourier transform*) ou outras formas alternativas dela derivadas, como a CQT (*constant*-Q *transform*) e a BQT (*bounded*-Q *transform*). A CQT e a BQT apresentam resoluções na freqüência variáveis, decrescentes ao longo do espectro, o que é importante para sinais musicais. Mas a baixa seletividade de seus canais resulta na inacurácia das informações presentes no espectro.

O presente trabalho busca no FFB (*fast filter bank*) alternativas a essas ferramentas citadas. Para tal, são criadas duas novas ferramentas: o CQFFB (*constant*-Q *fast filter bank*) e o BQFFB (*bounded*-Q *fast filter bank*), que oferecem seletividade superior e resoluções na freqüência, respectivamente, similares à da CQT e à da BQT.

Faz-se um estudo detalhado da implementação do FFB, do CQFFB e do BQFFB e comparamos seus desempenhos. A importância da seletividade e da resolução na freqüência é discutida na comparação de espectros obtidos de sinais com senóides e notas musicais. Em seguida, desenvolve-se um algoritmo básico de identificação de notas através de espectros de sinais de música. O trabalho é encerrado com uma análise dos resultados, que se mostraram positivos e promissores.

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

SPECTRAL REPRESENTATION OF SIGNALS IN
AUTOMATIC MUSIC TRANSCRIPTION

Cristiano Nogueira dos Santos

December/2004

Advisors: Sergio Lima Netto
Luiz Wagner Pereira Biscainho

Department: Electrical Engineering

Automatic music transcription (AMT) aims at finding every note of a given musical signal. Among its utilities, AMT helps composition study, recording edition and music signal coding.

The DFT (discrete Fourier transform) is frequently used to obtain a music signal spectrum, as well as other DFT-based tools like the CQT (constant-Q transform) and the BQT (bounded-Q transform). These two last tools work as music-oriented DFTs, since their frequency resolutions decrease along the frequency scale. Despite these advantages, these tools exhibit poor selectivity, thus providing an inaccurate description of the spectrum.

Our work proposes, through the use of the FFB (fast filter bank), alternatives for the DFT-based tools. In this thesis, two new tools are created: the CQFFB (constant-Q fast filter bank) and the BQFFB (bounded-Q fast filter bank). Their selectivity is superior while their frequency resolutions are similar to those of DFT-based tools like the CQT and the BQT, respectively.

A detailed study on the implementation of the FFB, CQFFB and BQFFB has been done and their performances have been compared. The importance of selectivity and frequency resolution is discussed via comparisons among their spectra calculated over sinusoids and music signals. A basic algorithm for musical note identification is presented and used to test and illustrate the application of the BQFFB. We conclude by analyzing the results, which were quite promising.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Introdução à transcrição musical

A transcrição musical é o ato de representar um sinal de música de forma a permitir sua análise visual e sua reprodução. De forma geral, a nota é a unidade básica de execução de uma peça musical. Portanto, espera-se a seqüência de notas musicais de uma gravação na saída de um sistema básico de transcrição musical automática (TMA[1]). As notas devem ser descritas com as informações de instante de tempo de início, duração e freqüência fundamental.

O objetivo deste trabalho é realizar o estudo de sinais musicais e ferramentas que permitam especificar e desenvolver um transcritor, livre dos padrões tradicionais de notação musical e aplicável a sinais musicais mono (com apenas um canal de gravação).

É importante ter em mente que uma transcrição pode envolver diferentes graus de detalhamento, indo desde a simples identificação das notas até a descrição de efeitos ornamentais, próprios a um estilo de interpretação musical. Quanto maior for o grau de detalhamento da transcrição, maior será a possibilidade de se obter um alto grau de similaridade sonora entre a reprodução, a partir da transcrição, e o som original. Como um primeiro passo, este trabalho aborda essencialmente a identificação das notas.

Na área musical, tanto profissionais quanto amadores e estudantes costumam utilizar transcrições para analisar composições. Essa forma de representação

[1]A sigla TMA servirá neste texto tanto para transcrição musical automática, quanto para transcritor musical automático. A diferenciação se dará pelo contexto.

permite-lhes visualizar a evolução das notas de toda uma peça musical. Isto evita que seja necessário ouvir repetidas vezes a seqüência musical sob análise, o que, muitas vezes, é um processo complicado e lento.

Em casos de maior complexidade, a análise da estrutura musical pode exigir esforço e tempo desconfortáveis para o interessado, pois, neste caso, é preciso que, primeiro, seja ouvido, compreendido e escrito o conjunto de notas presentes para que, em seguida, seja possível sua análise. O esforço necessário para uma transcrição musical varia de acordo com a música e com a experiência do músico em questão, podendo tornar a tarefa praticamente inviável.

A utilização da TMA pode auxiliar tanto no estudo de uma obra quanto no processo de composição, permitindo ao compositor que toque a música livremente, podendo observar o que acabou de tocar, sem que precise interromper a execução para transcrever as notas. E caso ele não se interesse pela gravação sonora, mas somente pela transcrição, pode ainda economizar espaço em seu computador.

A edição musical também pode se beneficiar da TMA. É possível reproduzir a música transcrita através de aplicativos de síntese musical, usando o padrão MIDI, alterando-se os instrumentos usados, assim como o estilo musical. A edição pode ser feita ainda com o uso de som sintetizado, obedecendo às condições da gravação original e misturando-se a ela, como permite a proposta do padrão de áudio do MPEG-4 [2].

Ainda na linha dos padrões, o MPEG-7 áudio [2] surge para estabelecer critérios de descrição do ambiente sonoro de uma gravação, o que pode ser diretamente atendido pela TMA. O MPEG-7 pode, com a descrição, facilitar a codificação seletiva, o uso de bibliotecas de música, a escolha de uma estação de rádio, a edição de áudio, entre outros aspectos.

1.2 Histórico da TMA

É interessante comparar o reconhecimento musical com o de fala. Apesar de o primeiro ser formado apenas por notas musicais, um sinal de música, freqüentemente, tem a complexidade de diferentes notas (ou vozes) soando simultaneamente, enquanto, nos sinais de fala considerados para processamento, admite-se apenas uma voz a cada instante. A presença de apenas uma nota a cada instante em uma música

(monofonia) não é comum, sendo que para esses casos já existem soluções de TMA consideradas satisfatórias [3].

Recentes contribuições de alto nível vêm surgindo ultimamente para a polifonia, procurando aplicar regras de percepção humana aos algoritmos de identificação de notas. Note-se, porém, que, na maioria dos casos, os sistemas vêm sendo desenvolvidos em duas linhas de atuação distintas quanto ao conteúdo musical. O conteúdo pode consistir na exclusiva presença de instrumentos percussivos de som indeterminado ou na exclusiva presença de notas musicais. O presente trabalho se dedica essencialmente ao segundo caso.

Um bom resumo da evolução dos sistemas de transcrição musical é apresentado em [3, 4]. O primeiro sistema polifônico conhecido é o de Moorer. Ele desenvolveu um algoritmo com o objetivo de transcrever duetos. As limitações impunham que as duas vozes concorrentes ocupassem faixas de notas distintas. O sistema foi ainda aprimorado por Chafe e por Maher, sempre mantendo o limite de apenas duas vozes. Em 1989, Katayose apresentou seu transcritor, que poderia lidar com cinco vozes simultâneas, mas passando a apresentar maior taxa de erro.

Até essa época, não era comum o uso de regras de percepção. A identificação de notas era feita apenas de forma heurística. Desde então, os principais sistemas conhecidos passaram a adotar novas técnicas de percepção musical, como modelos de timbre, previsão estatística de transição de acordes e arquitetura "quadro negro" (*blackboard*). Esses últimos avanços foram testados por Kashino [5] e Martin [6, 7], e significaram uma sensível evolução para estes sistemas, segundo Klapuri [3].

Os sistemas de Foo e Lee [8, 9] enfatizam a forma de comprovar, em uma etapa posterior, a existência das notas identificadas. Eles recriam as notas, a partir de modelos previamente gravados, e comparam com o trecho de música em análise. Isto, segundo eles, é a simulação do processo de percepção musical de um aluno, testando as notas até considerar casado o som que está tocando com o som gravado que está tentando reproduzir.

Quanto ao desempenho destes sistemas, o sistema de Kashino trabalha com sucesso com três notas simultâneas e diferentes instrumentos, como flauta, violão e piano. Já os sistemas de Martin e Foo, com cerca de quatro notas, focalizam seus testes apenas em músicas com piano.

Em 2000, Goto [10] desenvolveu um algoritmo que estima, com bastante confiabilidade, a freqüência fundamental predominante em um sinal de música com fundo harmônico e/ou ruidoso. Este algoritmo foi usado por Goto para identificar a melodia e a linha de baixo em sinais de áudio complexos. Com essas informações, é mais fácil obter as demais notas presentes nos sinais sob análise.

Em todos os casos, é possível notar que ainda não chegamos a um sistema com resultados satisfatórios para uso comercial. É preciso ainda avançar no maior número de notas e instrumentos simultâneos.

1.3 Estrutura da tese

No Capítulo 2, o leitor será apresentado a alguns conceitos de música, envolvendo as notas musicais, o uso de suas combinações entre si e sua representação no domínio da freqüência. Isto permitirá compreender os termos e exemplos utilizados nos capítulos a seguir.

O Capítulo 3 trata do processo de TMA em si. Neste capítulo estão descritas as etapas em que podemos dividir o processo de TMA, com seus objetivos, dificuldades e a diversidade com que podem ser implementadas. Aqui são apresentados os obstáculos principais a serem vencidos para que se obtenham resultados satisfatórios na TMA.

No Capítulo 4, é apresentado o *fast filter bank* (FFB), cuja estrutura é baseada no algoritmo da FFT (*fast Fourier transform*), unindo a baixa complexidade desta a uma alta seletividade. Propõe-se, aqui, utilizar o FFB como base de novas ferramentas, o CQFFB (*constant*-Q *fast filter bank*) e o BQFFB (*bounded*-Q *fast filter bank*), para serem usadas na etapa de análise espectral da TMA.

Aplicamos, no Capítulo 5, o BQFFB em exemplos de sinais polifônicos típicos. É usado ainda um algoritmo heurístico para a estimação de freqüências fundamentais, obtendo as notas musicais. Este capítulo visa a avaliar a aplicação do BQFFB na TMA.

O sexto e último capítulo conclui com uma análise da contribuição desta tese e a apresentação de propostas para sua continuação.

Capítulo 2

Breve introdução à estrutura musical

2.1 Introdução

Para compreender os objetivos de um TMA, é preciso conhecer os elementos que constituem uma música, as características desses elementos e sua nomenclatura. O presente capítulo apresenta o som e a música, de forma resumida e objetiva, com o intuito de esclarecer as dificuldades a serem enfrentadas no processo de TMA.

A Seção 2.2 introduz os conceitos de vibrações periódicas, séries harmônicas, além das principais propriedades sonoras: altura, intensidade e timbre. A Seção 2.3 parte das séries harmônicas para explicar as oitavas e as notas musicais, seguindo com a apresentação dos intervalos e da escala ocidental. A Seção 2.4 mostra como se estrutura uma peça musical sobre notas e intervalos, chegando à polifonia e aos acordes. A Seção 2.5 conclui o capítulo indicando os harmônicos, as notas e os acordes como objetos de análise da TMA.

2.2 Som

A percepção das vibrações sonoras pelo ser humano pode ser descrita como na figura 2.1. Esta figura mostra curvas médias de audibilidade humana, em função da intensidade e da freqüência de estímulo do som. As vibrações, quando são irregulares no tempo, formam sons ruidosos. Quando periodicamente regulares, dão origem

a um tipo de som que chamaremos de tonal, tomando certa liberdade. A maior parte dos instrumentos de corda e de sopro emite sons tonais, que podem ser bem aproximados pelo modelo senoidal, como na seguinte equação:

$$s(t) = \sum_n A_n(t) cos(2\pi n f_0 t + \phi_n), \tag{2.1}$$

onde o sinal $s(t)$ é a soma de seus componentes harmônicos h_n e f_0 é a freqüência fundamental do sinal. E, ainda, cada harmônico é representado por sua freqüência nf_0, fase ϕ_n e função de amplitude no tempo $A_n(t)$.

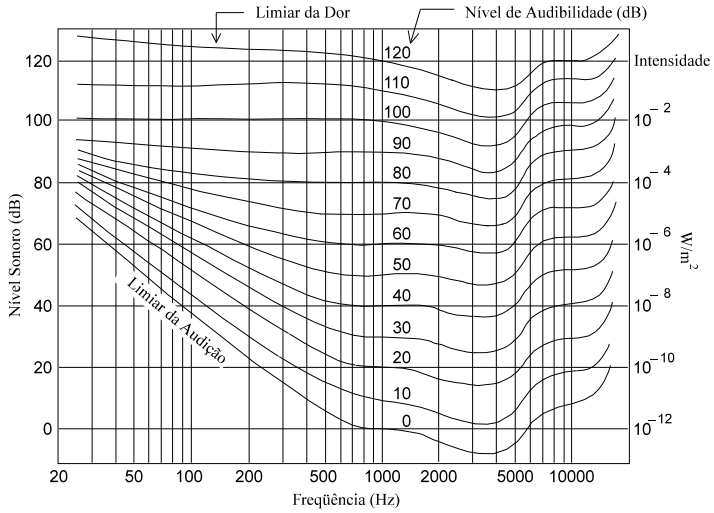


Figura 2.1: Curvas de nível de audibilidade. Esta figura, baseada em [1], mostra curvas médias de audibilidade humana em função da intensidade e da freqüência de estímulo.

2.2.1 Série harmônica

É interessante considerar o exemplo de uma corda vibrante de comprimento finito e presa em ambas as extremidades [11]. Neste exemplo podem ser observados infinitos modos de vibração. Constituem os harmônicos do som o primeiro modo de vibração, conhecido como modo fundamental, juntamente com os infinitos modos de vibração parciais (segundo, terceiro, quarto, ...). O primeiro modo de vibração é aquele em que os únicos nodos[1] são os extremos da corda, levando a um único

[1]Nodos são os pontos da corda em que a velocidade de oscilação é nula.

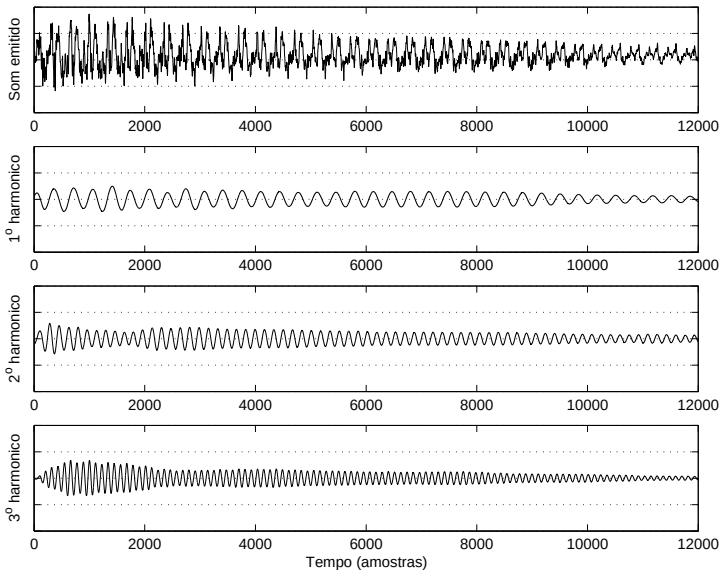


Figura 2.2: Amplitude ao longo do tempo de um a) Sinal de nota Dó3 de piano, com taxa de amostragem 44,1 kHz, além de seus b) primeiro harmônico, c) segundo harmônico e d) terceiro harmônico, que foram obtidos através de filtragem do sinal em a).

segmento de corda, que vibra com comprimento de onda λ. No segundo modo de vibração, um nodo central divide a corda ao meio, dividindo-a em dois segmentos de corda, que oscilam, cada um, com comprimento de onda $\lambda/2$. E, assim por diante, o n-ésimo modo de vibração, por sua vez, apresenta oscilações com comprimento de onda λ/n.

A freqüência do modo fundamental chamamos de freqüência fundamental $f_0 = v/\lambda$, onde v é a velocidade de oscilação da corda. As vibrações parciais, por sua vez, possuem freqüências de oscilação múltiplas de f_0. Assim, o primeiro harmônico h_1 tem freqüência igual a $f(h_1) = f_0$, o segundo harmônico tem $f(h_2) = 2f_0$ e o n-ésimo harmônico tem freqüência $f(h_n) = nf_0$. O conjunto de vibrações harmônicas de um som é conhecido como série harmônica, e é bem representado pelo modelo senoidal da equação (2.1). A figura 2.2 mostra a amplitude do sinal em função do tempo de uma nota de piano, assim como de seus três primeiros harmônicos. As amplitudes dos harmônicos foram obtidas por meio de filtragem do sinal. A figura 2.3 mostra o mesmo sinal, agora com a envoltória espectral evoluindo ao longo do tempo.

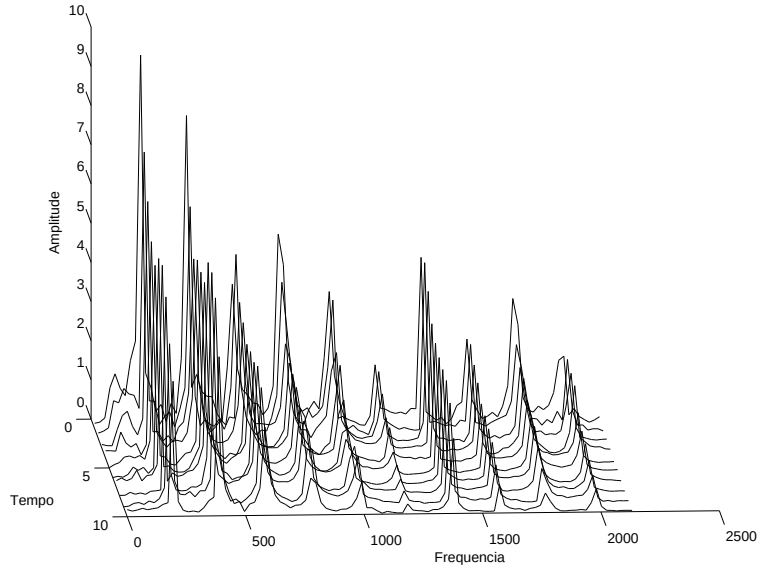


Figura 2.3: Evolução ao longo do tempo da envoltória espectral do sinal da figura 2.2, amostrado a 44,1 kHz.

2.2.2 Propriedades do som

O som dito tonal pode ser caracterizado por três principais propriedades: altura, intensidade e timbre. A altura está ligada à sua freqüência de vibração. Quanto maior for sua freqüência fundamental, mais alto ou agudo é o som. E quanto menor for sua freqüência fundamental, mais baixo ou grave é o som.

A intensidade do som depende da amplitude de vibração. Ela distingue o som forte do fraco.

Por fim, o timbre é o elemento pelo qual distinguimos uma voz humana de outra, ou um instrumento musical de outro. É freqüente atribuírem-se qualidades subjetivas ao som, como dizer que determinado som é opaco e outro é colorido. Essas denominações são referentes ao timbre do som, também conhecido como "cor" do som. Entre os diferentes fatores responsáveis pelo timbre [12, 13] estão a) a forma de onda ou envoltória do sinal e b) as amplitudes relativas dos harmônicos entre si [14]. Esses dois fatores podem ser tratados como um único fator: a evolução no tempo da envoltória espectral. A figura 2.4 apresenta a evolução do espectro ao longo do tempo de uma nota de viola. A mesma nota tocada por piano na figura 2.3 pode servir de comparação entre diferentes timbres.

Observando as figuras 2.2 e 2.3, pode-se perceber que as vibrações parciais misturam-se à vibração fundamental, com amplitudes que podem ser de mesma

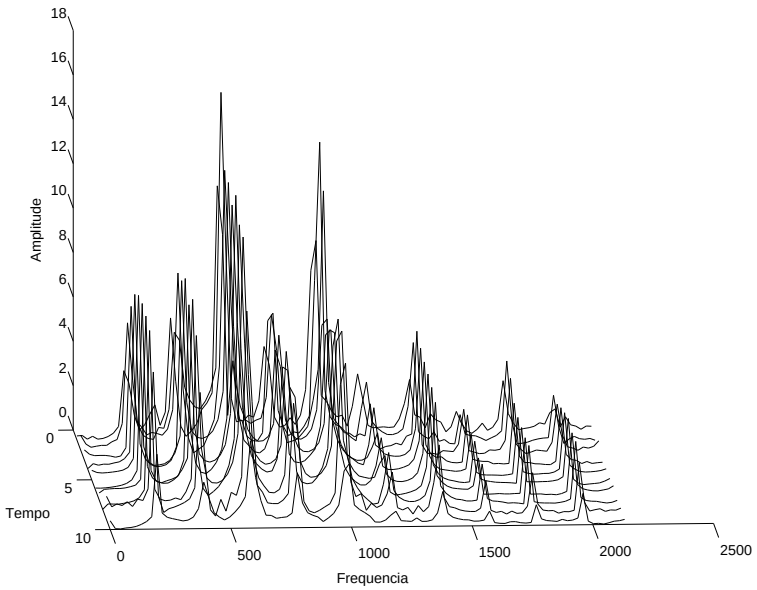


Figura 2.4: Evolução ao longo do tempo da envoltória espectral de um sinal de uma nota Dó3 de viola, amostrado a 44,1 kHz.

ordem de grandeza. Apesar disso, grosso modo, a altura percebida do som tonal é a mesma altura percebida de uma senóide cuja freqüência seja igual à freqüência fundamental deste som tonal[2]. O que diferencia estes dois sons de mesma altura é o timbre, conforme dito anteriormente.

2.3 Notas musicais

2.3.1 Oitava

Da série harmônica, vimos que um determinado som A tem as freqüências de seus harmônicos f_{n_A} como múltiplas de sua freqüência fundamental f_{0_A}, ou seja, $(f_{n_A} = nf_{0_A})$, onde $n = \{1, 2, 3, ...\}$. Assim, se considerarmos um som B com freqüência fundamental $f_{0_B} = 2f_{0_A}$, todas as freqüências dos harmônicos de B serão também freqüências dos harmônicos de A ($f_{n_B} = nf_{0_B} = 2nf_{0_A}$). Se A e B ocorrerem ao mesmo tempo, haverá sobreposição total dos harmônicos de B pelos de A. Os dois sons são perfeitamente consonantes (isto é, todos os harmônicos de uma nota

[2]Isso simplifica, propositalmente, o conceito de *pitch*, freqüência percebida, que não necessariamente se confunde com a fundamental. Uma boa referência para o conceito de *pitch* pode ser encontrada em [15]

coincidem com os da outra), e na notação musical são reconhecidos como a mesma nota musical. Isto significa que se a freqüência fundamental de Dó0 é $f_{0_{Do0}}$, também será Dó toda nota cuja freqüência fundamental for $f = 2^n f_{0_{Do0}}$.

Na música ocidental, as notas musicais foram criadas dividindo esse intervalo de freqüência entre uma nota e sua repetição. Foram criadas sete diferentes notas: Dó, Ré, Mi, Fá, Sol, Lá e Si. Note que a oitava nota seria a repetição da primeira. Entende-se também como oitava o intervalo entre duas freqüências quaisquer f_A e f_B, se $f_B = 2f_A$.

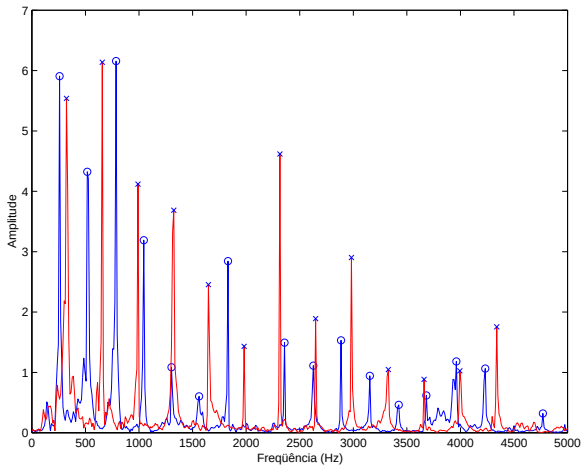
2.3.2 Intervalo

Intervalo é a diferença de altura entre duas notas. O intervalo entre um Dó e um Mi subseqüente na escala é uma terça, e entre um Dó e um Sol é uma quinta, já que o Mi e o Sol são, respectivamente, a terceira e a quinta notas a contar-se a partir do Dó.

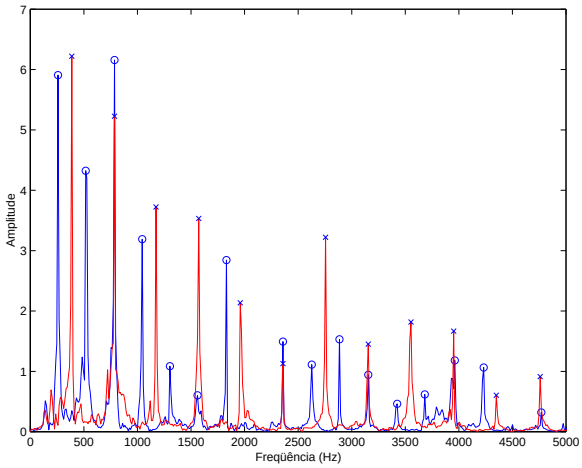
Os intervalos são também classificados segundo o grau de consonância entre as notas que o compõem. O grau de consonância de um intervalo está relacionado ao grau de sobreposição entre os seus harmônicos. No caso do intervalo de oitava, trata-se do grau máximo de sobreposição, dado que todos os harmônicos de um dos sons são sobrepostos pelos do outro. Outros exemplos de consonância são os observados a seguir. Na figura 2.5a, encontra-se um intervalo de terça[3], representado pelas notas Dó e Mi de um piano. Observe que os harmônicos de Dó múltiplos de 5 ($f = 5nf_{0_{Do}}$) estão sobrepostos pelos harmônicos de Mi múltiplos de 4 ($f = 4nf_{0_{Mi}}$). Na figura 2.5b, tem-se um intervalo de quinta[4], onde as notas Dó e Sol apresentam os harmônicos $f = 3nf_{0_{Do}}$ sobrepostos pelos harmônicos $f = 2nf_{0_{Sol}}$. Um exemplo de intervalo dissonante é o de sétima, ilustrado na figura 2.5c, em que aparece o espectro das notas Dó e Si tocadas simultaneamente, e é menor a ocorrência de harmônicos sobrepostos.

[3]Terça maior, na definição física, corresponde à freqüência $f_2 = 5/4f_1$.

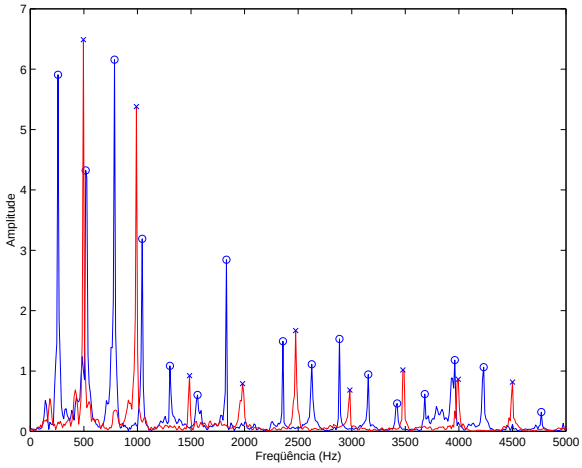
[4]Quinta justa, na definição física, corresponde à freqüência $f_2 = 3/2f_1$.



(a)



(b)



(c)

Figura 2.5: Exemplos com notas de piano amostrados com 11,025 kHz. Magnitude espectral de sinais de notas a) Dó3 (‘o’) e Mi3 (‘x’), b) Dó3 (‘o’) e Sol3 (‘x’) e c) Dó3 (‘o’) e Si3 (‘x’).

2.3.3 Escala de igual temperamento

Depois de diversas adaptações na cultura musical ocidental, no século XVIII [16], foi introduzido o sistema de escala de igual temperamento. Este sistema dividiu uma oitava em doze faixas de freqüência, cujas freqüências centrais encontram-se em perfeita progressão geométrica. Essas faixas correspondem a uma faixa (ou intervalo) de semitom[5]. Se o intervalo de uma oitava corresponde a dois sons tais que $f_2 = 2f_1$, o intervalo de um semitom equivale a dois sons tais que $f_2 = 2^{1/12} f_1$ (ou $f_2 \approx 1,06 f_1$).

Esta divisão da escala musical é utilizada pelos instrumentos temperados ou de som fixo, como o violão e o piano, que tocam sempre a mesma freqüência para cada nota. Esta adaptação, porém, resulta em notas que formam intervalos entre si ligeiramente diferentes dos intervalos da escala diatônica maior "natural" [11]. Isto faz com que a sobreposição dos harmônicos não seja exata em intervalos consonantes que não sejam múltiplos inteiros, como na tabela 2.1.

Tabela 2.1: Comparação, entre as escalas "natural" e temperada, das razões de freqüências presentes em alguns intervalos.

Intervalo	Escala "natural"	Escala temperada
Segunda maior	$9/8 = 1,125$	$2^{2/12} \approx 1,122$
Terça maior	$5/4 = 1,25$	$2^{4/12} \approx 1,260$
Quarta justa	$4/3 \approx 1,333$	$2^{5/12} \approx 1,335$
Quinta justa	$3/2 = 1,5$	$2^{7/12} \approx 1,498$
Sexta maior	$5/3 \approx 1,667$	$2^{9/12} \approx 1,682$
Sétima maior	$15/8 = 1,875$	$2^{11/12} \approx 1,888$
Oitava justa	2	$2^{12/12} = 2$

De acordo com a escala que abrange o grande órgão, as notas musicais compreendem um intervalo de oito oitavas, do Dó0 ao Dó8, como convencionamos chamar[6]. Para conhecer as freqüências das notas musicais, segundo a escala temperada, basta utilizar uma freqüência de referência, o Lá da quinta oitava: 440 Hz. Se o

[5]Na escala de igual temperamento, podemos chamar os semitons assim: Dó, Dó# ou Ré♭, Ré, Ré# ou Mi♭, Mi, Fá, Fá# ou Sol♭, Sol, Sol# ou Lá♭, Lá, Lá# ou Si♭ e Si.

[6]Existem ainda outras convenções.

intervalo entre a nota de interesse e a nota de referência (Lá5) for igual a k semitons, a freqüência fundamental daquela nota pode ser obtida através de $f = 440 \times 2^{k/12}$. Uma outra referência é o Dó4, que, nesta convenção, corresponde ao chamado Dó central do piano.

2.4 Estrutura da música ocidental

Conhecidas as características do som e as notas musicais, cumpre-nos conhecer as partes mais internas da estrutura de uma peça musical. A música possui três características básicas [17]: melodia, ritmo e harmonia. A melodia é a sucessão de sons formando sentido musical. O ritmo regula os sons organizando-os no tempo quanto ao início, duração e intensidade. A harmonia, por sua vez, é elemento enriquecedor do conteúdo musical, mas não essencial. Ela consiste na combinação de sons simultâneos. Grosso modo, a melodia é o que orienta a seqüência de notas ao longo do tempo. Em paralelo, a harmonia costuma servir de acompanhamento e sustentação musical.

É possível encontrar padrões na seqüência de notas em melodias populares e mesmo eruditas, dependendo isto principalmente do estilo. Existem similaridades entre estas seqüências em melodias de um mesmo estilo musical. Já em uma mesma melodia, é comum encontrar repetições literais dessas seqüências.

Entre os intervalos melódicos (de notas subseqüentes) são percebidos mais padrões do que regras de formação. No caso da harmonia, os intervalos harmônicos (de notas simultâneas) costumam obedecer a regras de formação quanto a padrões de transição entre um acorde e o seguinte. Os acordes são combinações de notas simultâneas, ou intervalos harmônicos. No ritmo também são observadas regras de formação. São comuns os compassos binários, ternários e quaternários. E dentro de cada compasso existem variações típicas de um estilo musical próprio.

Uma música pode ainda ser classificada de acordo com sua tessitura. Segundo Roy Bennett, em [18], há três maneiras básicas de um compositor "tecer" uma música:

- monofônica: constituída por uma única linha melódica;
- polifônica: com duas ou mais linhas melódicas entretecidas ao mesmo tempo;

- homofônica: uma única linha melódica acompanhada de acordes.

Chamaremos de polifônicos também os sinais homofônicos, para fins de simplificação.

Para enriquecer a música, em casos polifônicos, utilizam-se acompanhamentos à melodia principal com intervalos harmônicos (notas simultâneas). Esses intervalos podem ser mais ou menos consonantes. Quanto maior a consonância, maior a capacidade de "casar" os sons das diferentes notas, devido à maior sobreposição dos harmônicos.

Inicialmente, os acompanhamentos utilizavam apenas o intervalo de oitava, de maior consonância. Com o tempo, foram acrescentados novos intervalos aos considerados consonantes, e ainda os dissonantes passaram a ser mais comumente ouvidos [18]. Esse conjunto de notas executadas simultaneamente, observadas as regras de harmonia, chama-se acorde.

Os acordes podem variar bastante. O uso de acordes com intervalos consonantes, bastante comum em músicas populares e determinados estilos eruditos, implica alto grau de sobreposição de harmônicos. E isto é tido como uma das maiores dificuldades da TMA, pois complica a identificação de notas com intervalos de oitavas, além do reconhecimento de timbres, fortemente baseado nas amplitudes relativas dos harmônicos.

2.5 Conclusão

A notação musical e seus detalhes são de grande importância na análise feita por um TMA, pois o conhecimento da música ocidental, com suas regras de formação, é um dos principais fatores usados pelos homens quando desejam compreender quais notas estão presentes em um dado trecho de música.

A nota musical é o elemento essencial da TMA, como se vê no capítulo 1, e tem entre suas características principais a altura (freqüência fundamental) e o timbre (evolução temporal da envoltória espectral). A freqüência fundamental deve ser associada a uma faixa de semitom, indicando a nota musical propriamente dita, segundo a nomenclatura da escala de igual temperamento. Por sua vez, o timbre deve ser associado a um intrumento musical.

Tanto no caso da nota (ou da altura) quanto no caso do timbre, os harmônicos são os elementos da TMA que deverão ser analisados, pois eles são os objetos de mais fácil observação no espectro, e a partir deles podem-se estimar as eventuais notas presentes e seus respectivos timbres. Por isso, é interessante utilizar-se o modelo senoidal, conforme a equação 2.1. A estrutura musical polifônica, porém, típica tanto em músicas populares quanto eruditas, complica a TMA, principalmente no reconhecimento de timbres, pois a consonância presente em acordes implica sobreposição de harmônicos no espectro, misturando informações de diferentes notas musicais.

O presente capítulo apresentou os elementos musicais, objetos de análise da TMA, sendo os principais a nota musical e o timbre. A partir deste ponto, pode-se compreender melhor o capítulo 3, que trata das etapas e técnicas usadas na TMA.

Capítulo 3

Etapas da TMA

A TMA consiste em um encadeamento de análises interdependentes, em que são observados os sinais musicais nos domínios do tempo e da freqüência. Para uma maior organização e eficiência, a TMA costuma ser dividida em etapas bem definidas, que trocam informações entre si. O presente capítulo descreve o processo de TMA, apresentando suas etapas.

Na seção 3.1, há uma breve introdução de como se dá a transcrição musical por uma pessoa. A seção 3.2 introduz as técnicas usadas na TMA, indicando as influências e as diferenças da percepção humana para o processo de TMA. As seções 3.3 a 3.6 detalham as etapas de um TMA, contendo os objetivos, as técnicas aplicadas e os obstáculos enfrentados por cada uma delas.

3.1 Transcrição musical feita pelo homem (TMH)

Fizemos uma pequena pesquisa informal, envolvendo músicos profissionais e amadores. Verificou-se que, para transcrever uma música a partir de sua audição, costuma-se, em primeiro lugar, escolher um intrumento musical cuja participação na música será transcrita. Escolhida a fonte sonora, o músico ouve repetidas vezes um trecho musical que seja longo o suficiente para que possa perceber o que está sendo tocado e curto o suficiente para que possa ser gravado na memória.

A partir da audição do intervalo musical, ainda que haja outros instrumentos interferindo, o músico não apresenta grandes dificuldades para identificar o instrumento de interesse e observá-lo. No caso de o instrumento estar tocando um acorde,

porém, a identificação de notas é mais complicada e ele, então, reconhece primeiro o acorde como um todo, para, em seguida, buscar as notas. A identificação do acorde é feita a partir de características auditivas particulares, geradas pelos intervalos presentes, sendo elas mais evidentes nos intervalos dissonantes.

No reconhecimento das notas e dos acordes, o músico pode facilitar o processo utilizando um instrumento e comparando o som da gravação sonora com a reprodução das notas e/ou acordes em seu instrumento. Nesta etapa de refinamento da transcrição, são reduzidos os enganos e acrescentados os detalhes mais difíceis de se identificar na etapa inicial.

3.2 TMA

A nota musical deve ser entendida como elemento essencial da TMA. É a partir das notas transcritas que o músico pode reproduzir uma peça musical. As informações básicas que descrevem uma nota musical na transcrição são a) o tempo de início e b) a freqüência fundamental. Outra informação que pode ser acrescentada às notas na TMA é o timbre, que pode ser caracterizado não só pelo instrumento musical como pelo estilo de tocar.

Na literatura [6, 5, 3, 8], a TMA está dividida em diferentes números de etapas. Neste trabalho, a TMA está dividida em quatro principais etapas: 1) detecção de tempo de início das notas musicais e segmentação do sinal, 2) identificação das notas em cada segmento de sinal, 3) reconhecimento do timbre de cada nota identificada na segunda etapa e 4) análise dos resultados. Veja na figura 3.1.

Numa primeira etapa, assim como na TMH, a TMA precisa segmentar o sinal no domínio do tempo. Para facilitar a análise do espectro, na segunda etapa, é interessante que os segmentos de sinal contenham apenas conjuntos de notas simultâneas, evitando que notas subseqüentes misturem suas informações no espectro. Quanto mais isoladas estiverem as notas, mais fácil será sua identificação. Na primeira etapa, é obtido o tempo de início de cada nota.

Em seguida à segmentação do sinal, chega-se à identificação de notas presentes em cada segmento. Com mais esta etapa, já foram obtidas as informações de tempo inicial (na primeira etapa) e freqüência fundamental. Basta, então, iden-

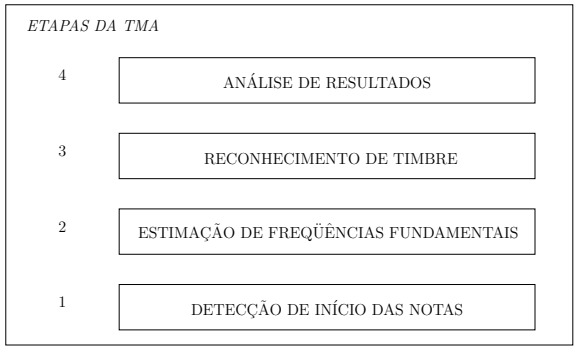


Figura 3.1: Etapas de uma TMA. Elas podem ser seguidas no sentido de baixo para cima (*bottom-up*) ou de cima para baixo (*top-down*). No sentido de baixo para cima, as informações vão sendo colhidas do sinal nas etapas inferiores para serem analisadas nas etapas superiores. No sentido de cima para baixo, as informações das etapas superiores influenciam na busca das informações das etapas inferiores.

tificar as notas musicais associadas às faixas de semitom em que se encontrem as freqüências fundamentais.

Uma terceira etapa se encarrega de reconhecer os timbres das notas musicais identificadas anteriormente. O timbre de uma nota a associa a uma fonte sonora ou instrumento musical. O reconhecimento de um instrumento presente na peça musical pode auxiliar a identificação das próximas notas, permitindo uma previsão de seu ritmo característico na peça, além da faixa de freqüência a que estão limitadas suas notas.

Por fim, na quarta etapa deve-se avaliar a coerência dos resultados obtidos. Isto pode ser feito através da ressíntese do sinal. A ressíntese pode ser feita a partir das informações das notas obtidas nas etapas inferiores. Esta etapa da TMA corresponde à etapa da TMH em que o músico reproduz o que foi percebido para sua confirmação. O segmento de sinal ressintetizado pode ser comparado ao segmento original.

As etapas são dispostas em direção vertical para indicar que existe uma hierarquia entre elas. Quanto mais baixo o nível da etapa, mais concreto é o tipo da informação recolhida. Isto significa que as informações estão mais próximas dos sinais em forma bruta. E quanto mais alto o nível da etapa, mais abstrata ela será. Suas informações estão mais próximas do significado musical do som para o homem.

O percurso destas etapas pode ser feito em dois sentidos. No sentido de baixo para cima (*bottom-up*), cada etapa em posição superior dá seqüência ao processo com os resultados obtidos na etapa inferior. Por exemplo, as informações da envoltória espectral de uma notas musical obtidas na segunda etapa orientam a de reconhecimento de timbre, na terceira etapa.

O sentido de cima para baixo (*top-down*) é o inverso. Neste caso, as informações vão dos níveis de análise superiores para os de níveis inferiores, utilizando conhecimentos mais próximos de conceitos de teoria musical. Um exemplo disto ocorre quando a identificação de um instrumento musical leva à expectativa de notas com determinadas características de tempo e freqüência.

3.2.1 Comparando a TMA com a TMH

Antes de estudarmos as etapas, uma por uma, é importante comparar algumas características da percepção humana com as da etapa de identificação de notas da TMA.

Existem diferentes métodos descritos na literatura para estimar freqüências fundamentais em sinais de música. Esses métodos podem ser reunidos em dois principais grupos: aqueles que usam a) o domínio do tempo e b) o domínio da freqüência.

As análises feitas no domínio do tempo costumam observar a periodicidade de trechos de sinais através da autocorrelação. Quando aplicadas a sinais monofônicos, as técnicas associadas à autocorrelação obtêm bons resultados. Quando aplicadas a sinais polifônicos, a observação de periodicidade do sinal mistura informações, podendo levar a conclusões errôneas, como a identificação de notas falsas [3]. Isto se explica da seguinte forma: a composição de harmônicos de duas notas pode ser suficiente para indicar a periodicidade de uma terceira nota "fantasma", que teria sido induzida pela periodicidade dos harmônicos em comum entre as duas notas presentes.

Algumas técnicas baseadas na autocorrelação utilizam procedimentos que se aproximam da forma como se dá a percepção musical humana. O homem, porém, tem algumas dificuldades na identificação de notas em situações como, por exemplo, quando um ou mais instrumentos tocam um acorde. Neste caso, ele costuma, numa

primeira abordagem, perceber o acorde, mas não precisamente identificar cada uma das notas presentes [19].

Por causa dessa dificuldade para distinguir notas em sinais polifônicos, segundo Klapuri [3], não devemos buscar aplicar a forma humana de analisar os sons. Essa dificuldade vem do fato de que, na tradição musical ocidental, conjugam-se notas musicais com certo grau de consonância com a finalidade de levar o ouvinte a perceber um acorde como um conjunto sonoro único, indivisível.

Os métodos que observam o domínio da freqüência, por sua vez, podem analisar o espectro com a ajuda de transformadas, como a DFT, por exemplo. Buscam-se, nestes casos, os pontos de máxima energia identificados no espectro como candidatos a harmônicos do trecho de sinal. A análise no domínio da freqüência é descrita com maiores detalhes na seção 3.4.

3.3 Etapa 1: Detecção de início das notas

3.3.1 Descrição

Nesta primeira etapa, três diferentes objetivos podem ser atingidos: 1) a própria detecção dos instantes de início das notas musicais, b) a segmentação do sinal em intervalos limitados por esses tempos e, ainda, 3) a identificação de sua estrutura rítmica[1].

Esta primeira etapa da TMA pode ser, então, dividida também em sub-etapas, ainda segundo o conceito hierárquico de níveis de informação. A figura 3.2 ilustra esta estrutura. A detecção de tempos das notas, como está mais ligada à forma de onda do sinal, aparece num nível inferior, enquanto a sub-etapa de identificação do ritmo fica num nível superior. Na subida, as informações de tempo das notas levam à identificação do ritmo, enquanto na descida, o ritmo pode auxiliar na detecção de início de notas mais fracas.

A partir dos tempos das notas em um primeiro trecho de sinal analisado, chega-se a um ritmo estimado e, a partir deste, podem-se estimar os próximos inícios de notas, em trechos de sinal a seguir. Além disso, em casos de incerteza quanto

[1]Para identificar o ritmo de um sinal de música existem ainda técnicas especificamente desenvolvidas [20].

à autenticidade de certo tempo de início, o ritmo pode ser um fator auxiliar na decisão.

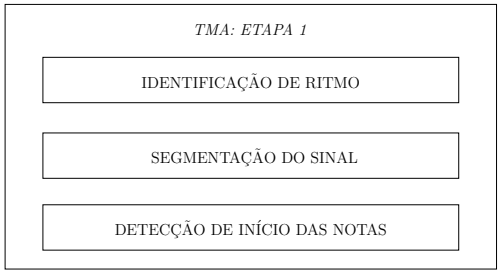


Figura 3.2: A primeira etapa da TMA é responsável pela análise dos tempos das notas, o que orienta a segmentação do sinal e a identificação de sua estrututa rítmica.

A identificação da estrutura rítmica do sinal é também relevante na TMA. Primeiramente, algumas formas de transcrição musical, como a partitura, exigem a definição do ritmo da peça descrita.

A detecção de início de sons não serve somente à TMA. Ela também facilita a edição de gravações musicais e o sincronismo de áudio e vídeo em filmagens, por exemplo.

A segmentação do sinal é de grande importância para a etapa seguinte, onde são processados os algoritmos de estimação de freqüências fundamentais. A análise do trecho de sinal segmentado garante um mínimo de isolamento no espectro sob observação. Isto evita que as freqüências presentes nos trechos adjacentes confundam a análise do trecho corrente, fazendo-se passar por informações ruidosas.

A segmentação traz uma outra vantagem em casos de segmentos muito longos. Nestas situações, pode-se optar por uma análise mais eficiente, trabalhando-se apenas com um número mínimo de amostras. Sabendo-se, antecipadamente, que o restante do intervalo contém informação redundante, pode-se ignorá-lo.

3.3.2 Técnicas

Segundo Klapuri, em [20], os primeiros sistemas de detecção de novas notas baseavam-se na envoltória do sinal como um todo. Devido à pouca eficiência deste método, ao longo do tempo essa busca passou a ser feita sobre faixas de freqüência, com bancos de filtros. Scheirer teria sido o primeiro a declarar que os sistemas de

detecção deveriam seguir o modelo de audição humana, que trata o sinal separadamente em faixas de freqüência, combinando os resultados no final [21]. Bilmes teria ido na mesma direção, mas usando apenas duas faixas, alta e baixa, não sendo tão eficiente.

A maior parte dos sistemas desenvolvidos [21, 22, 23] para detectar o início de notas envolve a estimação do ritmo musical. Nestes casos, usa-se a autocorrelação de longos trechos de sinal para eliminar erros e ajustar a sensibilidade do processo de detecção de baixo nível, que utiliza a envoltória do sinal.

3.3.3 Obstáculos

Um ponto de grande importância é a difícil análise em cima de alterações graduais ou modulações de amplitude ou de freqüência em uma nota. Essas alterações podem ser encaradas erroneamente como entradas de novas notas. Veja exemplo na figura 3.3.

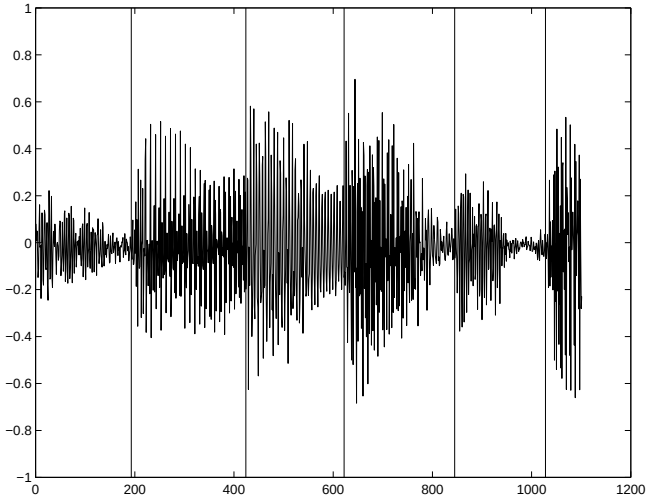


Figura 3.3: Sinal segmentado entre tempos de início das notas de um piano. Repare a modulação de amplitude da envoltória da segunda e quinta notas. As modulações de amplitude podem confundir, induzindo a detecção enganosa de novas notas.

Entre os intrumentos com ataque sensível é mais simples o trabalho de detecção. As maiores dificuldades surgem com a detecção de entrada de instrumentos de sopro e arco, como a flauta e a viola. A envoltória de seus transitórios são mais suaves (ver figura 3.4), além de suas envoltórias apresentarem modulações durante

a sustentação da nota. Isto dificulta o que poderia ser uma simples detecção por análise de diferença abrupta de amplitude na envoltória do sinal ao longo do tempo.

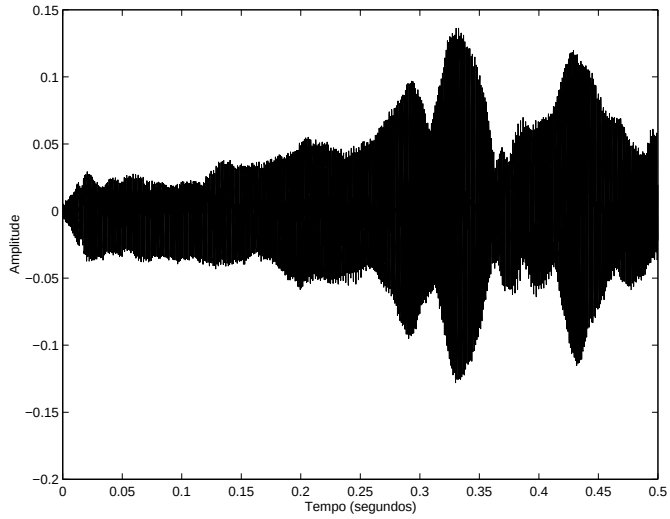


Figura 3.4: Sinal de flauta. Exemplo de forma de onda de uma nota Lá#6. Neste caso, a TMA deveria indicar uma única nota sustentada ao longo do tempo, mas as modulações presentes tenderiam a confundir a análise.

3.3.4 Resumo

A detecção de início das notas é exigência na identificação de notas na TMA. Ela ainda permite a segmentação do sinal, levando a um processamento mais eficiente na estimação de freqüências fundamentais, com o isolamento de notas subseqüentes. Além disso, a detecção de novas notas auxilia a identificação do ritmo musical, que, por sua vez, pode realimentar a detecção de novas notas.

3.4 Etapa 2: Identificação de notas musicais

3.4.1 Descrição

A identificação de notas musicais presentes num segmento de sinal é feita através da estimação de freqüências fundamentais deste segmento. As freqüências fundamentais indicam a altura dessas notas. Cada nota contém um conjunto de componentes harmônicos, segundo a sua série harmônica, que são observados como

picos na envoltória espectral. Em casos de música monofônica, a identificação das notas é mais simples, mas, em sinais polifônicos, o aumento do número de notas simultâneas dificulta bastante essa análise.

O conceito de observar os sinais de música através dos picos do espectro segue o modelo senoidal, que representa analiticamente o sinal harmônico. Este modelo se contrapõe aos modelos de percepção, que procuram reconhecer de forma sintética as notas musicais. No modelo senoidal, cada harmônico é analisado como uma senóide, com suas propriedades individuais para, em seguida, serem associados a uma nota musical, se for o caso.

A associação dos harmônicos a uma nota musical deve levar em consideração, além das freqüências múltiplas de uma mesma fundamental, a semelhança de aspectos como o tempo de início e modulações em amplitude ou freqüência.

Observando a figura 3.5, pode-se compreender o procedimento básico utilizado para a obtenção das notas musicais. O primeiro passo é escolher a ferramenta de análise espectral, como a FFT, por exemplo. Com a envoltória de amplitude espectral, são encontrados os picos candidatos a harmônicos. A análise do espectro ao longo do tempo indica quais desses picos se mantêm candidatos, formando "linhas" (*tracks*, em inglês) no espectrograma de picos [24], veja na figura 3.6. Das linhas às freqüências fundamentais, buscam-se as semelhanças (de tempo e freqüência, por exemplo) dos harmônicos entre si, formando a série harmônica de cada nota. Como cada nota musical está associada a uma faixa de semitom no domínio da freqüência na escala musical, as freqüências fundamentais irão identificar as notas musicais. E, finalmente, das notas chega-se aos acordes e destes até o tom da música, se for o caso.

O processo descrito no parágrafo anterior é o processo de subida. O processo de descida parte dos acordes ou ainda do tom encontrado para prever os elementos dos níveis inferiores. Se tomarmos o exemplo de um processo de subida que identificou duas notas presentes: um Dó e um Sol, chega-se ao intervalo de quinta e pode-se esperar que este intervalo formasse o acorde Dó Maior. Este acorde, porém, é formado também pela nota Mi. O processo de descida vai buscar a presença da nota Mi no segmento.

A DFT é de grande importância nesta etapa da TMA. Ela é a transformada

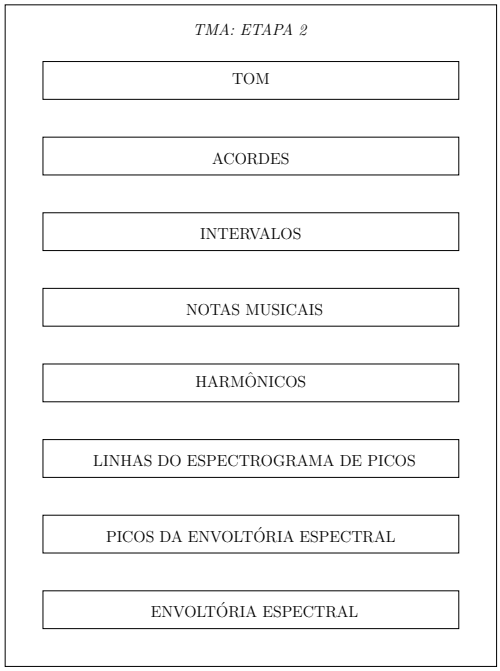


Figura 3.5: Arquitetura da etapa de identificação de notas musicais.

mais popular na observação de espectrogramas, e ainda serve de base para outras transformadas. Como a resolução da DFT é constante em todo o espectro, ela se torna ineficiente na análise de notas no espectro, que estão distribuídas logaritmicamente. Por isso, foram criadas a CQT (*constant*-Q *transform*) [25] e a BQT (*bounded*-Q *transform*) [3], que possuem resolução na freqüência, respectivamente, proporcional e progressiva em relação à faixa de freqüência. Ou seja, em baixas freqüências há mais pontos do que em altas freqüências para representar uma faixa de mesma largura em Hertz. Outra ferramenta usada nesta etapa é o FFB [26, 9], que será abordada no Capítulo 4.

É importante notar que as ferramentas descritas acima são derivadas da FFT e, por isso mesmo, possuem saídas complexas. Transformadas ou bancos de filtros que tenham saídas reais não são ideais para a análise da envoltória de sinais. Saídas complexas permitem a obtenção da envoltória do sinal com facilidade, bastando calcular o módulo do sinal. É simples compreender isso com o exemplo de uma senóide. Se a saída da análise espectral for complexa, ela terá módulo constante. Caso a saída seja real, teremos o módulo sinuoso, com vales e cristas, dificultando a análise [27].

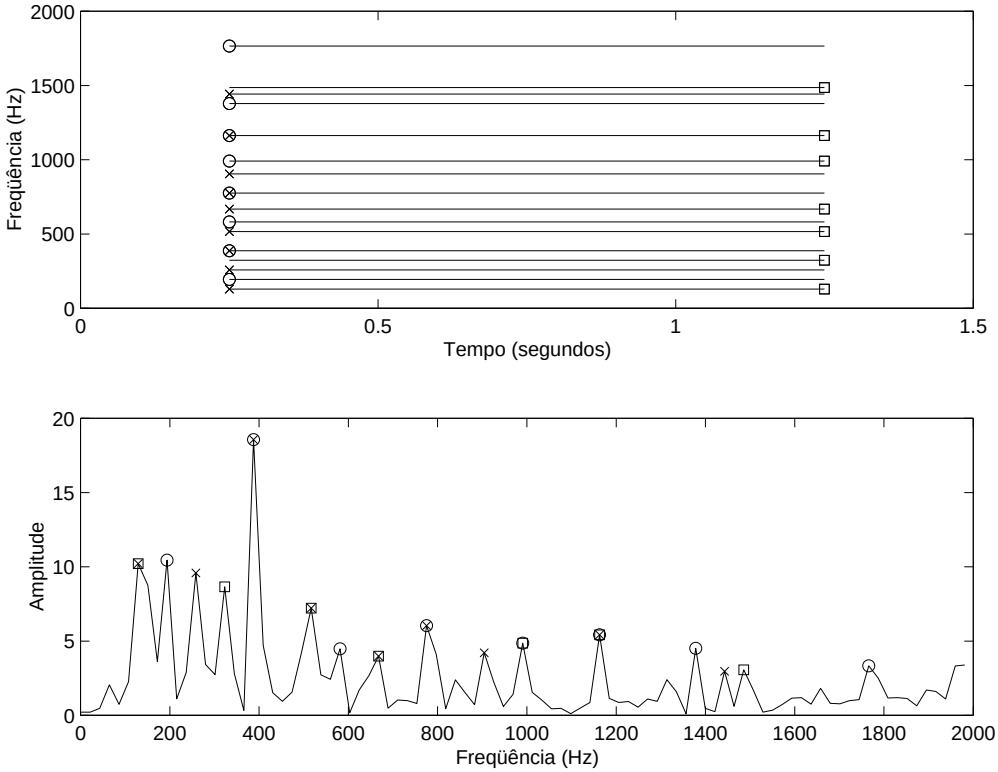


Figura 3.6: Exemplificando a segunda etapa da TMA. O gráfico inferior apresenta a envoltória espectral de um acorde Dó-Mi-Sol. O gráfico superior, com as linhas espectrais, mostra um espectrograma de picos, onde apenas picos com amplitude acima de um certo patamar foram escolhidos como candidatos a harmônicos. Ambos os gráficos mostram as duas séries harmônicas identificadas na sub-etapa de harmônicos: Dó e Sol, marcados com 'x' e 'o', respectivamente. A série harmônica de Mi é identificada no gráfico com um quadrado. Esta nota ainda não teria sido identificada no exemplo por haver sobreposição de sua fundamental com a da nota Dó.

3.4.2 Obstáculos

Cada nota musical está diretamente ligada a uma freqüência fundamental e seus respectivos harmônicos. Apesar disso, não é simples a tarefa de identificar as notas presentes. A dificuldade de se encontrar as notas tem três principais causas: 1) nem todos os conjuntos harmônicos representam notas musicais, 2) há sempre alta irregularidade na envoltória espectral e 3) ocorrem harmônicos sobrepostos de diferentes notas.

Uma série harmônica encontrada em um espectrograma pode ser entendida como uma nota musical, ressalvando-se algumas exceções. Essas exceções podem vir de interferências ruidosas ou ainda de ressonâncias provocadas pelo soar de uma ou mais notas. Um exemplo deste segundo caso é uma das cordas de um piano vibrar influenciada pela vibração de uma nota, como mostra a figura 3.7.

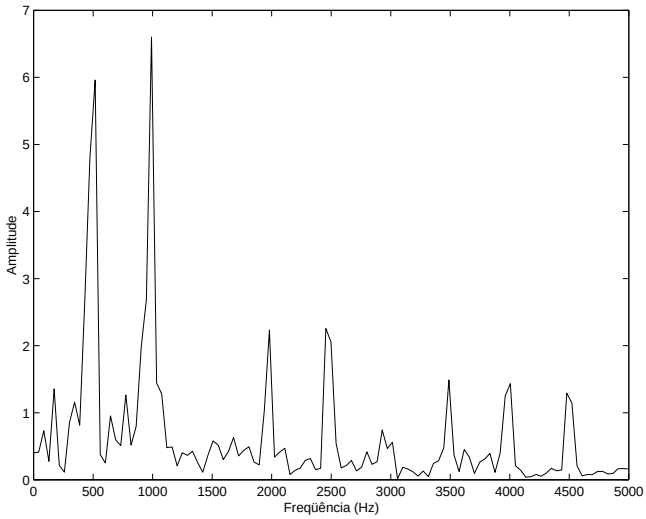


Figura 3.7: Exemplo de nota Si de piano com ressonância uma oitava abaixo.

As irregularidades na envoltória espectral podem ser vistas como picos espúrios ou aumento da amplitude real de uma faixa, seja um pico ou um vale no espectro. Elas costumam ter origem no ruído produzido pelos próprios instrumentos ou pelo ambiente de gravação musical. Outra fonte de irregularidade pode vir da ferramenta de análise espectral, como a seletividade pobre da FFT, por exemplo. Esses picos podem se sobrepor aos harmônicos e se tornar incômodos para a análise.

A sobreposição de harmônicos é, porém, o maior obstáculo a ser vencido. Os intervalos musicais consonantes usam notas cujos harmônicos, em grande número, se sobrepõem (ou quase) e confundem a análise de notas musicais.

O mais complicado caso de sobreposição de harmônicos é o intervalo de uma ou mais oitavas inteiras. Numa oitava, uma das notas tem todos os seus harmônicos coincidentes com a outra, já que o primeiro harmônico da mais alta tem a mesma freqüência do segundo harmônico da mais baixa. O intervalo de quinta também tem alta sobreposição de harmônicos, apresentando 1/2 dos harmônicos da quinta sobrepostos e 1/3 dos harmônicos da tônica sobrepostos, uns pelos outros. Lembrando,

no entanto, que a sobreposição não é exata entre instrumentos temperados ou de som fixo, como descrito na seção 2.3.3.

3.4.3 Técnicas

Entre as principais técnicas utilizadas na estimação de freqüências fundamentais, descrevemos aqui quatro: 1) a arquitetura "quadro negro" (em inglês, *blackboard*), 2) a estimação das linhas de melodia e de baixo (acompanhamento de baixa freqüência), 3) a subtração iterativa de notas do espectro e 4) a alta resolução espectral.

A arquitetura quadro negro foi utilizada na TMA por Kashino [5] e Martin [6]. Seu nome se deve à idéia de um grupo de alunos que vão ao quadro negro resolver um problema, cada aluno contribuindo com seu conhecimento específico para a solução. Com esta arquitetura, Kashino e Martin partiram da análise espectral ao tom da música, utilizando os conhecimentos específicos ("alunos") sobre intervalos comumente encontrados num acorde, probabilidade de transição de um acorde para outro, modelos de timbre, entre outros. Com esta arquitetura facilitaram o processo de subida e descida das etapas da identificação de notas.

O trabalho de Goto [10] seguiu outro rumo por se preocupar em identificar as linhas melódica e de baixo de uma peça musical. Seu sistema divide o espectro em duas faixas: altas e baixas freqüências. Nas duas faixas em paralelo ele busca a estrutura harmônica predominante. Ele ressalva ainda que não importa que a freqüência fundamental da linha melódica não esteja na faixa mais alta. Este método pode não ser completo por não chegar a todas as notas, mas serve de passo inicial importante por serem as outras notas, do acompanhamento, bastante correlacionadas à melodia.

Klapuri [3], por sua vez, desenvolveu a técnica de subtração iterativa de notas do espectro. Ao identificar uma nota, sua contribuição na envoltória espectral deve ser subtraída para facilitar a identificação de outras eventuais notas presentes. Note-se que o método de Goto pode ser de grande utilidade aqui. Para subtrair uma nota, sem que se percam as informações do espcetro sobrepostas por esta nota, deve-se utilizar um modelo de timbre adequado [28].

Por fim, outra técnica utilizada na estimação de freqüências fundamentais é a de alta resolução espectral. Foo e Lee [9] utilizaram o FFB de Lim [26] para obter

um espectrograma com resolução de 5,4 Hz, na faixa de 0 a 10 kHz. Isto permitiria observar harmônicos separados, que em menor resolução estariam sobrepostos. Isto é possível devido à diferença entre a freqüência de uma nota nas escalas "natural" e temperada (ver seção 2.3.3).

Além das quatro técnicas mencionadas acima, podem ser acrescentados métodos para estimar as freqüências fundamentais com maior acurácia. Uma delas [6] sugere usar as freqüências dos harmônicos, que sejam da mesma série harmônica, para se aproximar da freqüência fundamental real. Outros métodos buscam se aproximar da freqüência real de um pico através da compensação de distorções causadas pelos cálculos do janelamento e da DFT: usando as amplitudes dos canais (*bins*) DFT vizinhos ao pico sob análise [3] ou usando a DFT de derivadas do sinal [29].

Entre os sistemas estudados acima, os resultados obtidos com sucesso na TMA envolveram até quatro notas musicais simultâneas.

3.4.4 Resumo

Essa é a etapa mais importante e complexa de um TMA. A sobreposição de harmônicos e um grande número de notas simultâneas são as maiores dificuldades encontradas para a identificação das notas musicais.

3.5 Etapa 3: Reconhecimento de timbre

Encontrada uma nota musical, pode-se associá-la a um instrumento musical, que a teria gerado. A utilidade do reconhecimento do timbre está associada a dois principais fatores. O primeiro é que cada instrumento, numa peça musical, deve ter a transcrição de sua parte, independentemente dos outros instrumentos. O segundo fator é que o conhecimento do instrumento torna possível a ressíntese do som para a etapa de análise dos resultados obtidos.

Para identificar o instrumento, é necessário conhecer as características dos harmônicos da nota musical em análise. Com o problema da sobreposição de harmônicos, esta etapa se torna complexa, pois não se podem obter as características individuais de cada um dos dois harmônicos sobrepostos em uma mesma faixa de freqüência.

Entre as principais características que identificam o instrumento estão 1) o ataque da nota e 2) sua composição harmônica. O ataque deve ser analisado com a envoltória de subida da nota. Sua composição harmônica pode ser observada nas amplitudes relativas dos harmônicos. Outras características também relevantes são relacionadas em [12, 13, 14].

Para estudar a envoltória de subida, é preciso ter resolução na freqüência suficiente para separar os harmônicos uns dos outros e resolução no tempo suficiente para poder observar a envoltória com suas peculiaridades, como variação brusca de amplitude devido ao ataque.

Para analisar a amplitude relativa dos harmônicos sobrepostos, pode-se utilizar a técnica de subtração de notas do espectro ou a de separação de harmônicos muito próximos através de alta resolução na freqüência. Existe ainda um método de reconhecimento de timbre desenvolvido por Kashino [30]. Este método segue o princípio de filtro casado para identificar a envoltória espectral de um determinado timbre.

Para reconhecer os instrumentos, é necessário ter um banco de dados com os modelos de instrumentos com suas características principais. É importante ainda que esses modelos sejam parametrizáveis, variando de acordo com a freqüência (altura da nota), a fase, a duração e o estilo de tocar.

Com a nota e o timbre, a primeira parte da transcrição se encerra. Segue-se, a partir daqui, a análise dos resultados. A figura 3.8 mostra o que se espera da saída de um TMA. Trata-se de dois instrumentos, cada um com as notas transcritas de um trecho de música.

3.6 Etapa 4: Análise dos resultados

Ao final da primeira parte da transcrição, representada pelas etapas anteriores, devem-se conferir os resultados. Como muitos dos harmônicos presentes foram sobrepostos, podemos ter acumulado erros e chegado a conclusões equivocadas quanto às notas e aos instrumentos. É interessante que sejam usados os modelos de instrumentos e notas musicais para ressintetizar os sinais e comparar os resultados [8].

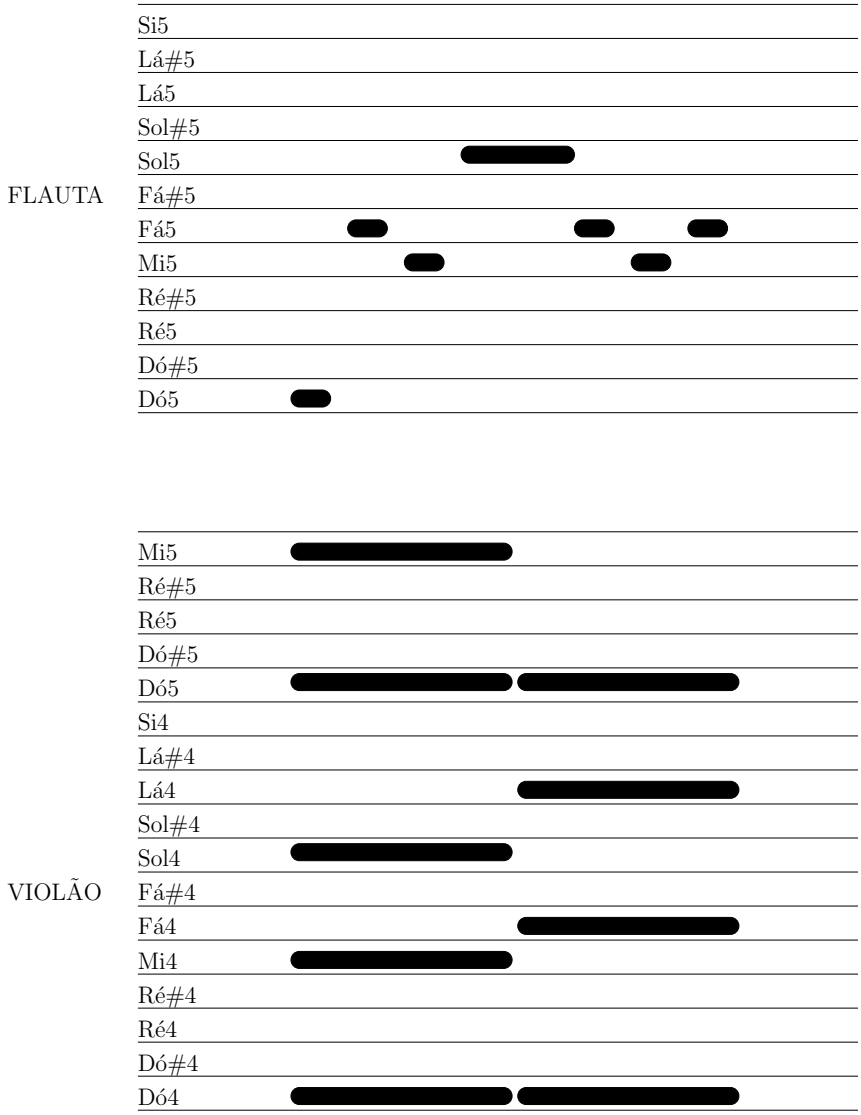


Figura 3.8: Exemplo de saída esperada de um TMA. A figura apresenta as participações de dois intrumentos, violão e flauta, com as notas musicais na vertical e o eixo do tempo na horizontal.

Essa "prova real" a que se submetem os resultados das etapas anteriores pode percorrer tanto uma análise de sentido musical segundo, por exemplo, as teorias de harmonia ou particularidades de determinado estilo musical, como uma análise direta sobre os dados, segundo as técnicas utilizadas no processamento dos sinais.

A ressíntese do sinal musical, através das informações retiradas da análise, serve como uma última verificação e pode levar a acréscimos ou retiradas de notas musicais de acordo com a comparação feita entre o sinal recriado e o sinal sob análise.

A ressíntese envolve algumas complicações como, por exemplo, a fase do sinal

a ser recriado. As diferentes fases em que podem se misturar os sinais podem levar a somas destrutivas entre harmônicos de diferentes notas.

Existem observações que são feitas pelo homem que podem auxiliar de forma significativa a análise dos resultados da TMA. Entre elas estão a restauração de trechos danificados [19] ou ainda de trechos de uma nota mascarada por outra. O processo de restauração pode ser entendido com o exemplo de uma pessoa não ouvir uma sílaba na palavra, mas compreendê-la por reconhecimento do contexto.

Podem-se ainda observar as regras de harmonia, se são seguidas ou não pelos resultados. Trechos similares, ou mesmo repetidos, como costuma ocorrer, podem ser reconhecidos em diferentes partes do sinal, reduzindo a complexidade da análise nos casos de recorrência. Fontes sonoras com movimentos similares de notas pela escala musical podem ser encontradas. Podem-se ainda utilizar conhecimentos de padrão estilístico de cada instrumento e sua função temática na melodia ou harmonia.

Enfim, a etapa de análise dos resultados é de grande importância para a TMA. Diferentes técnicas podem ser utilizadas, tornando mais confiável a transcrição obtida. Entre as técnicas mais importantes merecem menção a) a análise por ressíntese e b) a previsão de eventos baseado nas informações já obtidas.

3.7 Conclusão

A TMA divide-se em etapas complexas, em que ainda há bastante campo para desenvolvimento de novas técnicas. Na etapa de segmentação do sinal, ainda merecem estudos os instrumentos com envoltória mais suave, além da mistura de instrumentos com ligeira diferença de tempo de início [20]. A etapa 3 ainda tem dificuldades no que se refere a sinais polifônicos. A etapa 4 precisa ainda de estudos e orientação de outras áreas de pesquisa como a psicoacústica, por exemplo. Por fim, a etapa de identificação de notas constitui ainda foco das atenções de estudos, visto ser ela de importância central na TMA.

Para aprimorar a análise feita na segunda etapa da TMA, desenvolvemos uma nova ferramenta para calcular a envoltória espectral do segmento de sinal. Esta ferramenta propõe-se como alternativa à DFT e outras ferramentas também utilizadas, e é apresentada no próximo capítulo.

Capítulo 4

Análise espectral de sinais de música

4.1 Introdução

A análise espectral de sinais digitais é comumente associada à transformada discreta de Fourier: DFT (*discrete Fourier transform*) ou FFT (*fast Fourier transform*), a versão de implementação rápida da DFT. Mas a análise espectral de sinais musicais através da FFT apresenta limitações no que se refere a dois aspectos principais: a) a distribuição linear de amostras no domínio da freqüência, em contraste com a distribuição logarítmica de notas na escala musical e b) a pobre seletividade dos canais da FFT, que implica a alta interferência de informação entre as faixas de freqüência representadas por essas amostras. Este capítulo apresenta algumas ferramentas, encontradas na literatura, desenvolvidas para melhorar a análise espectral de sinais de música e, ainda, propõe novas ferramentas para este mesmo fim.

O capítulo se divide assim. A Seção 4.2 apresenta a FFT em duas partes, a distribuição de amostras no espectro e o banco de filtros sFFT (*sliding* FFT) [31], com sua estrutura e seletividade. As Seções 4.3 e 4.4 tratam, respectivamente, da CQT (*constant*-Q *transform*) [25] e da BQT (*bounded*-Q *transform*) [3], desenvolvidas para otimizar a distribuição de amostras da DFT no espectro, segundo a escala musical. Na Seção 4.5, descreve-se o FFB (*fast filter bank*) [26], um banco de filtros baseado na FFT, aproveitando desta a estrutura FRM (*frequency-response masking*) de baixa complexidade computacional e incrementando a seletividade de seus canais.

Nas Seções 4.6 e 4.7 são apresentadas as novas ferramentas, o CQFFB (*constant*-Q *fast filter bank*) [32] [33] e o BQFFB (*bounded*-Q *fast filter bank*), que reúnem, respectivamente, a distribuição de amostras da CQT e da BQT com a seletividade do FFB.

4.2 FFT

4.2.1 Distribuição de amostras no espectro

As notas musicais distribuem-se no espectro de acordo com os semitons da escala musical ocidental de 12 tons. Como a seqüência de freqüências centrais das faixas de semitons (onde se localizam as notas) obedece a uma progressão geométrica, pode-se obter uma análise espectral mais eficiente desses sinais com amostras que também estejam distribuídas segundo uma progressão geométrica.

A FFT divide o espectro em faixas de mesma largura, ou seja, segundo uma progressão aritmética, de acordo sua definição:

$$X_{DFT}(k) = \frac{1}{N}\sum_{n=0}^{N-1} win(n)x(n)exp(-j2\pi kn/N), \tag{4.1}$$

onde $x(n)$ é a n-ésima amostra do segmento de sinal sob análise, $X_{DFT}(k)$ é a amostra referente ao canal k da transformada de Fourier discreta do sinal $x(n)$, $win(n)$ é a função de janelamento e N é o total de canais utilizado na DFT. Esta distribuição, porém, não é otimizada para um TMA analisar o espectro de um sinal musical.

Para se ter uma idéia, da freqüência de 16,35 Hz (freqüência de Dó0) a 22,05 kHz (metade da taxa de amostragem do padrão das gravações de CD), existem 124 freqüências representando notas musicais. Uma resolução suficiente para distinguir duas notas quaisquer seria a metade do menor semitom [$16,35$ Hz$\times(1/2)(2^{1/12}-1) \approx 0,49$ Hz]. Se utilizarmos esta resolução numa FFT para descrever o espectro audível completo, serão necessárias mais de 90000 amostras. E a escolha de qualquer resolução pior que essa será insuficiente para as baixas freqüências, além de mais do que suficiente para atender às altas freqüências. Daí surge o interesse por ferramentas que otimizem a distribuição de amostras no espectro, descritas nas Seções 4.3 e 4.4.

4.2.2 Estrutura do banco de filtros FFT

Pensando na FFT como uma ferramenta que divide o espectro em faixas de freqüência, pode-se entendê-la como um banco de filtros. O banco de filtros FFT é também conhecido como *sliding* FFT [31] (sFFT), e sua complexidade computacional é estudada em [34].

Observe a figura 4.1, onde se apresenta, de duas formas, a estrutura do banco de filtros FFT com quatro canais de saída. Para encontrar a função de transferência dos canais da FFT, pode-se começar acompanhando a formação dos filtros na figura 4.1a.

Começando pela função de transferência do primeiro canal, queremos: $\frac{X_0(z)}{X(z)}$, onde $X(z)$ é a transformada z de $x(n)$. Pela figura 4.1a:

$$x_0(n) = a(n) + W_4^0 a(n-1) \tag{4.2}$$

$$a(n) = x(n) + W_4^0 x(n-2), \tag{4.3}$$

que são equivalentes, no domínio da transformada z, a

$$X_0(z) = (1 + W_4^0 z^{-1})A(z) \tag{4.4}$$

$$A(z) = (1 + W_4^0 z^{-2})X(z). \tag{4.5}$$

Conclui-se, assim, que

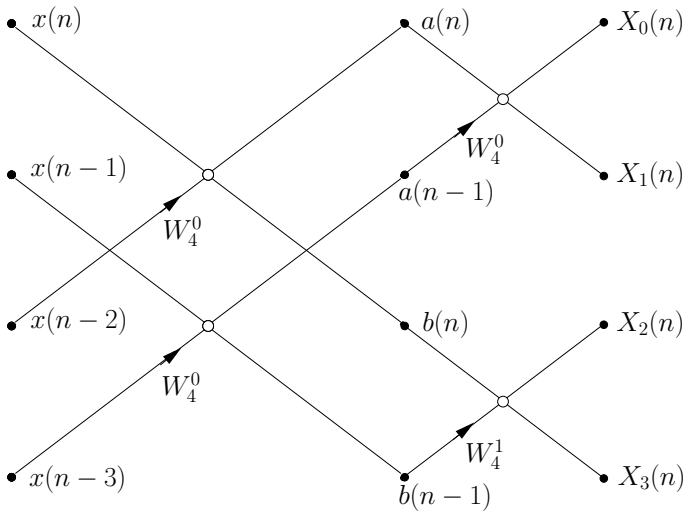
$$\frac{X_0(z)}{X(z)} = (1 + W_4^0 z^{-1})(1 + W_4^0 z^{-2}). \tag{4.6}$$

Continuando a observação dos filtros, chegamos aos demais canais

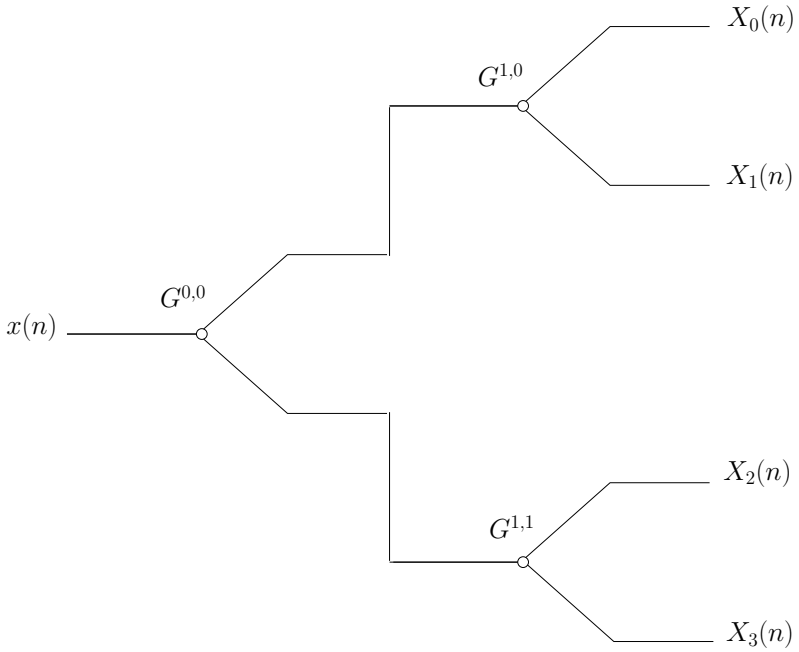
$$\begin{aligned} \frac{X_1(z)}{X(z)} &= (1 + W_4^1 z^{-1})(1 - W_4^0 z^{-2}) \\ \frac{X_2(z)}{X(z)} &= (1 - W_4^0 z^{-1})(1 + W_4^0 z^{-2}) \\ \frac{X_3(z)}{X(z)} &= (1 - W_4^1 z^{-1})(1 - W_4^0 z^{-2}). \end{aligned} \tag{4.7}$$

Para obtermos uma fórmula geral simples da função de transferência dos canais da sFFT, é interessante aplicar a substituição:

$$-W_N^k = W_N^{k+\frac{N}{2}} \tag{4.8}$$



(a)



(b)

Figura 4.1: a) Estrutura borboleta da FFT de 4 saídas; b) Estrutura de banco de filtros FRM da sFFT.

e apresentar as equações (4.7) da forma a seguir:

$$\begin{aligned}
\frac{X_1(z)}{X(z)} &= (1 + W_4^1 z^{-1})(1 + W_4^2 z^{-2}) \\
\frac{X_2(z)}{X(z)} &= (1 + W_4^2 z^{-1})(1 + W_4^0 z^{-2}) \\
\frac{X_3(z)}{X(z)} &= (1 + W_4^3 z^{-1})(1 + W_4^2 z^{-2}). \qquad (4.9)
\end{aligned}$$

Numa forma mais compacta, as equações (4.4) e (4.9) podem ser representadas assim:

$$\begin{aligned}
\frac{X_0(z)}{X(z)} &= G^{1,0}G^{0,0} \\
\frac{X_1(z)}{X(z)} &= G^{1,1}G^{0,2} \\
\frac{X_2(z)}{X(z)} &= G^{1,2}G^{0,0} \\
\frac{X_3(z)}{X(z)} &= G^{1,3}G^{0,2},
\end{aligned} \tag{4.10}$$

o que equivale à nomenclatura da figura 4.1b, onde

$$G^{i,j} = 1 + z^{-2i}W_N^j \tag{4.11}$$

e

$$j = [(2^i k) \bmod N]. \tag{4.12}$$

Observe que os filtros $G^{i,j}$ e $G^{i,j+N/2}$ são complementares, de modo que

$$G^{i,j} + G^{i,j+N/2} = 2, \tag{4.13}$$

o que reduz o número de multiplicações, pois se $A(z) = X(z)G^{i,j}$, então $B(z) = 2X(z) - A(z)$.

No exemplo acima, onde $N = 4$, pôde-se visualizar como são formados os sub-filtros da sFFT. Extrapolando $N = 2^L$, onde L é natural não nulo, chega-se à fórmula geral:

$$sFFT(z) = \frac{X_k(z)}{X(z)} = \prod_{i=0}^{\log_2 N} [1 + (z^{-1}W_N^k)^{2^i}]. \tag{4.14}$$

4.2.3 Demonstração da fórmula da sFFT

Para demonstrar a fórmula geral em (4.14), pode-se tomar a seguinte linha de raciocínio. O cálculo da sFFT é definido como descrito abaixo [26]:

$$X(k) = \sum_{i=0}^{N-1} x(n-i)W_N^{ki} = (\sum_{i=0}^{N-1} q^{-i}W_N^{ki})x(n), \tag{4.15}$$

onde $q^{-i}\{x(n)\} = x(n-i)$ e $W_N = e^{-j\frac{2\pi}{N}}$. Portanto, no domínio z, a sFFT pode ser expressa como

$$\text{sFFT}(z) = \sum_{i=0}^{N-1} z^{-i}W_N^{ki}. \tag{4.16}$$

Aplicando a soma de uma seqüência finita em progressão geométrica sobre a equação (4.16), obtém-se:

$$\text{sFFT}(z) = \frac{1 - \left(z^{-1} W_N^k\right)^N}{1 - z^{-1} W_N^k}. \tag{4.17}$$

Para $N = 2$, a equação (4.17) torna-se

$$\text{sFFT}(z) = \frac{1 - \left(z^{-1} W_2^k\right)^2}{1 - z^{-1} W_2^k}, \tag{4.18}$$

e a equação (4.14) torna-se

$$\text{sFFT}(z) = 1 + z^{-1} W_2^k, \tag{4.19}$$

que são equivalentes. Agora, assumindo que as equações (4.17) e (4.14) são equivalentes para todo L, onde $L = \log_2 N$, então

$$\frac{1 - \left(z^{-1} W_{2^L}^k\right)^{2^L}}{1 - z^{-1} W_{2^L}^k} = \prod_{i=0}^{L-1} \left[1 + \left(z^{-1} W_{2^L}^k\right)^{2^i}\right]. \tag{4.20}$$

Multiplicando os dois lados por $1 + \left(z^{-1} W_{2^L}^k\right)^{2^L}$, temos

$$\frac{1 - \left(z^{-1} W_{2^L}^k\right)^{2^{L+1}}}{1 - z^{-1} W_{2^L}^k} = \prod_{i=0}^{L} \left[1 + \left(z^{-1} W_{2^L}^k\right)^{2^i}\right], \tag{4.21}$$

que é a mesma equação de (4.20) para $L' = L + 1$. Isto completa a demonstração.

4.2.4 Seletividade

Usando a equação (4.16), tomamos como exemplo o cálculo da função de transferência do canal 34 de uma sFFT de 256 canais, como descrito abaixo.

$$\begin{aligned} H_{34}(z) =& \prod_{i=0}^{7} \left[1 + \left(z^{-1} W_{256}^{34}\right)^{2^i}\right] \\ =& G^{0,34} G^{1,68} G^{2,136} G^{3,16} G^{4,32} G^{5,64} G^{6,128} G^{7,0}, \end{aligned} \tag{4.22}$$

onde

$$\begin{cases} G^{0,34} = 1 + z^{-1} W_{256}^{34}; & G^{1,68} = 1 + z^{-2} W_{256}^{68} \\ G^{2,136} = 1 + z^{-4} W_{256}^{136}; & G^{3,16} = 1 + z^{-8} W_{256}^{16} \\ G^{4,32} = 1 + z^{-16} W_{256}^{32}; & G^{5,64} = 1 + z^{-32} W_{256}^{64} \\ G^{6,128} = 1 + z^{-64} W_{256}^{128}; & G^{7,0} = 1 + z^{-128} W_{256}^{0}, \end{cases} \tag{4.23}$$

O módulo da resposta correspondente ao canal 34 é ilustrado na figura 4.2, onde pode-se notar uma resposta pobre na banda passante e que os primeiros lobos laterais chegam a estar apenas 13 dB abaixo do patamar da banda passante, tendo, portanto, baixa atenuação na banda de rejeição.

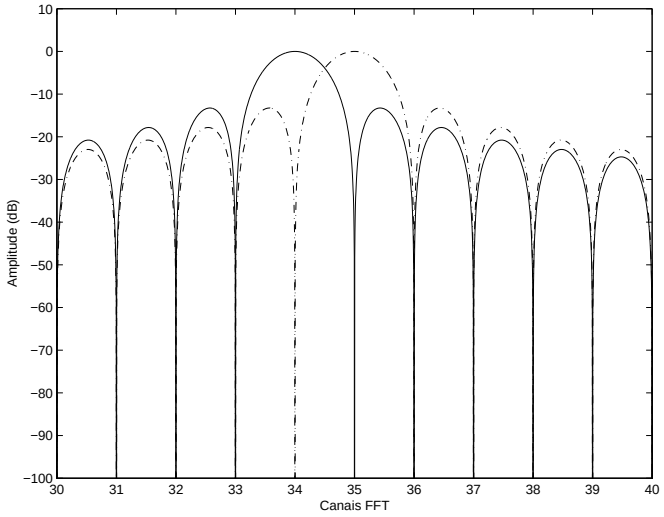


Figura 4.2: Módulos das respostas dos canais 34 ($|H_{34}(z)|$, em linha contínua) e 35 ($|H_{35}(z)|$, em linha tracejada) de uma sFFT de 256 bandas.

4.2.5 FFT com janelamento

É preciso acrescentar, porém, que, apesar da possibilidade de analisar os sinais de áudio com a *sliding* FFT, é comum que esses sinais sejam analisados com a STFT (*short-term Fourier transform*). O sinal é segmentado, e cada segmento tem sua amplitude no tempo suavizada nas primeiras e últimas amostras. Essa suavização é obtida através do janelamento dos segmentos [24].

A resposta na freqüência dos canais sFFT, observada no exemplo da figura 4.2, é alterada quando ocorre o janelamento. Para entender o efeito do janelamento sobre o espectro do sinal, pode-se analisar o efeito do janelamento sobre as respostas na freqüência dos filtros $H_{34}(z)$ e $H_{35}(z)$. A figura 4.3 mostra o resultado do janelamento Hanning sobre os canais 34 e 35 de uma FFT. Os canais janelados $h'_{34}(n)$ e $h'_{35}(n)$ foram obtidos com o produto entre a função janela de Hanning $w_H(n)$ e $h_{34}(n)$ e $h_{35}(n)$, respectivamente:

$$h'_C(n) = w_H(n)h_C(n), \tag{4.24}$$

onde C corresponde ao canal da FFT.

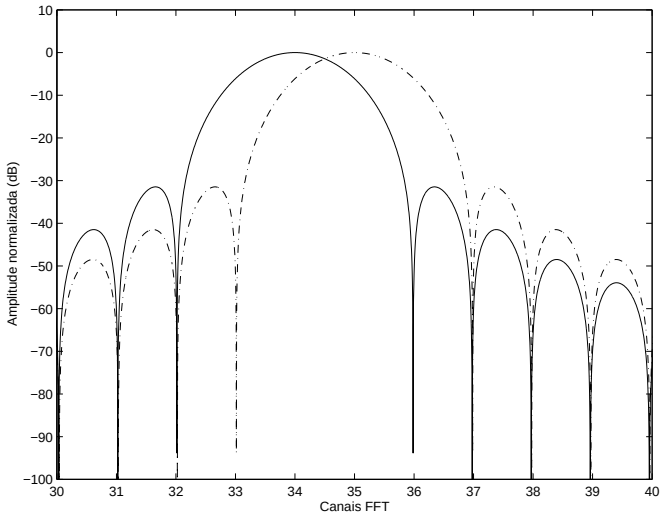


Figura 4.3: Módulos das respostas dos canais 34 ($|H'_{34}(z)|$, em linha contínua) e 35 ($|H'_{35}(z)|$, em linha tracejada) de uma sFFT de 256 bandas, com janelamento Hanning.

Comparando as figuras 4.2 e 4.3, observa-se que o janelamento altera a atenuação na banda de rejeição com ligeira melhora. Mas, em contrapartida, o lobo central sofre alargamento, implicando maior intersecção entre as bandas passantes de filtros adjacentes. Isto prejudica a separação de informação entre as faixas de interesse.

Essa seletividade pobre da FFT levou Lim e Farhang-Boroujeny a desenvolver o FFB (*fast filter bank*) [26], aproveitando a estrutura rápida da FFT e enriquecendo sua seletividade, como será visto na Seção 4.5.

4.3 CQT

A CQT (*constant*-Q *transform*) é uma adaptação da DFT, e se diferencia desta última por apresentar distribuição logarítmica das freqüências no domínio da transformada. O termo "Q constante" refere-se ao fator de qualidade Q de cada faixa de freqüência k de saída da transformada:

$$Q = \frac{f_k}{\Delta f_k}, \tag{4.25}$$

onde f_k e Δf_k são, respectivamente, a freqüência central e a largura da faixa a ser determinada no projeto.

Na DFT, a largura Δf é constante, sendo $\Delta f = f_a/N$, onde f_a é a freqüência de amostragem e N é o número de amostras do segmento de sinal sob análise. Portanto, na DFT, Q não é constante, mas diretamente proporcional a f_k. Com isso, para obter Q constante, é preciso que N seja variável, tornando-se uma função de k, isto é, $N(k)$. Desta forma a CQT é gerada a partir de uma modificação da equação (4.1) da forma:

$$X_{CQT}(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} win(k,n)x(n)exp(-j2\pi Qn/N(k)), \tag{4.26}$$

onde $win(k,n)$ é a função janela, que tem seu comprimento inversamente proporcional ao canal k. Repare que a mudança fundamental entre os cálculos da DFT e da CQT vem da diferença no número de amostras de entrada. Isto acarreta um período $T_{CQT} = N(k)/Q$, que na DFT seria $T_{DFT} = N/k$.

Segundo Brown [25], é interessante trabalhar com uma resolução na CQT que possa distinguir dois semitons; para tal, a razão entre as freqüências centrais de dois canais adjacentes deveria ser $2^{1/24}$, que é a raiz quadrada da razão entre as freqüências de dois semitons adjacentes. Deve-se optar pela raiz quadrada ao invés da metade por se tratar de freqüências em progressão geométrica, não em progressão aritmética. Então,

$$Q = \frac{f_k}{(\Delta f)_{\mathrm{CQT}}} = \frac{f_k}{(2^{1/24}-1)f_k} \approx \frac{1}{0{,}0293} \approx 34{,}127. \tag{4.27}$$

Tomamos, aqui, a liberdade de chamar esta resolução de "quarto-de-tom".

Um algoritmo eficiente de implementação da CQT é descrito em [35]. Para implementar a CQT, basta escolher as freqüências mínima f_{min} e máxima f_{max}, além do fator Q desejado.

4.4 BQT

Outra ferramenta interessante, elaborada para adequar a distribuição de faixas de freqüência da FFT à música, é a BQT (*bounded*-Q *transform*), citada por Klapuri em [3]. Nesta ferramenta, o objetivo é dividir o espectro em oitavas e, em

cada uma dessas oitavas, aplicar a FFT com a resolução adequada. Com esse procedimento, chega-se a uma espécie de FFT por partes, onde a resolução cresce das oitavas mais altas até as mais baixas, mas se mantém fixa dentro de cada oitava.

No cálculo da BQT, toma-se um segmento de sinal de N amostras, representando um intervalo de tempo T, para análise e calcula-se sua FFT, com resolução $1/T$. Guardam-se, da FFT, apenas as amostras da oitava mais alta, isto é, da metade superior da faixa de freqüência, e ignoram-se as amostras restantes. Em seguida, toma-se o segmento de N amostras usado na etapa anterior acrescido das N amostras seguintes do sinal, obtendo assim um segmento de $2N$ amostras. Este novo segmento é decimado por 2, tornando-se um novo segmento de N amostras, representando, desta vez, um intervalo de tempo igual a $2T$. Calcula-se a FFT do novo segmento e aproveitam-se, novamente, apenas as amostras da oitava mais alta. A nova resolução é igual a $1/(2T)$. E, assim por diante, este método permite também que seja otimizada a relação entre resolução no tempo e na freqüência para sinais de música.

Naturalmente, como qualquer outra transformada em blocos, a CQT e a BQT poderiam ser descritas como bancos de filtros multitaxa [36]. Essa abordagem, porém, não foi explorada nesta tese por simplicidade.

4.5 *Fast filter bank* - FFB

A sFFT é um banco de filtros que consome apenas uma multiplicação complexa por canal por amostra. Mas isso ocorre ao custo da baixa seletividade de seus filtros, o que para algumas aplicações pode comprometer a eficiência da análise espectral. Aproveitando a estrutura do algoritmo da sFFT, foi criado o FFB (*fast filter bank*) [26], que tem complexidade ligeiramente maior que a sFFT e seletividade significativamente maior.

A seletividade pobre dos canais da sFFT, vista nas seções anteriores, deve-se ao fato de que seu sub-filtro-protótipo é de primeira ordem: $G_{\mathrm{sFFT}_p}(z) = 1 + z^{-1}$, alterando-se de acordo com o conjunto de fatores (i, j). Assim, cada sub-filtro da sFFT, $G^{i,j}(z) = 1 + (z^{-1}W_N^k)^{2^i}$, é resultado da substituição de cada z por $(zW_N^{-k})^{2^i}$, ou seja, $G^{i,j}(z) = G_{\mathrm{sFFT}_p}((zW_N^{-k})^{2^i})$. Procedimento similar ocorre com os sub-filtros

do FFB.

O FFB foi desenvolvido para utilizar a estrutura da sFFT, porém substituindo o sub-filtro $G_{\mathrm{sFFT}_p}(z)$ por filtros de ordem maior, para obter maior seletividade. Mas é importante notar que, enquanto o sFFT tem apenas um sub-filtro-protótipo $G_{\mathrm{sFFT}_p}(z)$, o FFB foi projetado em [26] com um sub-filtro-protótipo diferente $G^i_{\mathrm{FFB}_p}(z)$ para cada nível i, sendo que as ordens dos sub-filtros se reduzem ao longo do percurso do nível de entrada para o nível de saída. E para se obter os sub-filtros FFB, faz-se $G^{i,j}_{\mathrm{FFB}}(z) = G^i_{\mathrm{FFB}_p}(zW_N^{-k})$.

Para compensar o aumento das ordens dos sub-filtros, reduz-se a complexidade computacional, fazendo com que os sub-filtros-protótipos do FFB, $G^i_{\mathrm{FFB}_p}$, sejam do tipo meia-banda [26] e tenham resposta ao impulso simétrica e de comprimento ímpar, o que leva a

$$G_{\mathrm{FFB}}(z) = \sum_{n=-\infty}^{\infty} g_{\mathrm{FFB}_p}(n) z^{-n} \tag{4.28}$$

$$g_{\mathrm{FFB}}(n) = g_{\mathrm{FFB}_p}(-n) \tag{4.29}$$

$$g_{\mathrm{FFB}}(n) = 0, \text{se } n \neq 0 \text{ e } n \text{ par} \tag{4.30}$$

$$g_{\mathrm{FFB}}(0) = 1. \tag{4.31}$$

Os coeficientes iguais a zero que resultarem do projeto meia-banda e a simetria ajudam a reduzir a complexidade computacional de cada sub-filtro em cerca de quatro vezes.

Possivelmente, o FFB pode ser implementado de forma eficiente usando estruturas alternativas como a decomposição polifásica [36].

Em [9], são listados os sub-filtros-protótipos $G^i_{\mathrm{FFB}_p}$ escolhidos para um projeto de FFB. Aproveitando deste os sete primeiros sub-filtros para obter um FFB

com 256 canais, chegamos à lista a seguir.

$$
\begin{aligned}
G^{0,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,6275(z + z^{-1}) - 0,1862(z^3 + z^{-3}) + 0,0878(z^5 + z^{-5}) - \\
&- 0,0426(z^7 + z^{-7}) + 0,0186(z^9 + z^{-9}) - 0,0067(z^{11} + z^{-11}) \\
G^{1,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,6209(z + z^{-1}) - 0,1688(z^3 + z^{-3}) + 0,0659(z^5 + z^{-5}) - \\
&- 0,0229(z^7 + z^{-7}) + 0,0055(z^9 + z^{-9}) \\
G^{2,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,5738(z + z^{-1}) - 0,0753(z^3 + z^{-3}) \\
G^{3,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,5654(z + z^{-1}) - 0,0654(z^3 + z^{-3}) \\
G^{4,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,5013(z + z^{-1}) \\
G^{5,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,5003(z + z^{-1}) \\
G^{6,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,5001(z + z^{-1}) \\
G^{7,j}_{\mathrm{FFB}_p}(z) &= 1 + 0,5000(z + z^{-1})
\end{aligned}
$$

Os módulos das respostas na freqüência dos canais 34 e 35 de um FFB de 256 canais são mostradas na figura 4.4. Como principais características podem-se notar a atenuação mínima de 56 dB na banda de rejeição e a banda passante extremamente plana.

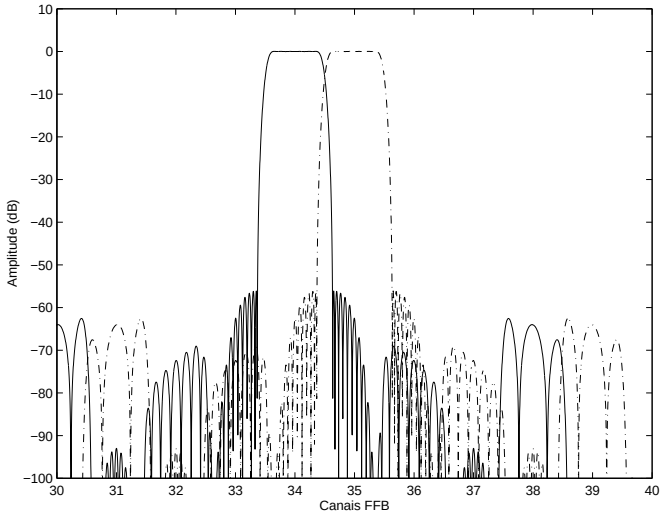


Figura 4.4: Módulos das respostas dos canais 34 ($|H_{\mathrm{FFB}_34}(z)|$, em linha contínua) e 35 ($|H_{\mathrm{FFB}_35}(z)|$, em linha tracejada) de uma sFFT de 256 bandas.

4.5.1 Canais FFB com janelamento

Se tratarmos o FFB como transformada, podemos aplicar uma espécie de STFFB (*short-term fast filter bank*). Neste caso, se aplicarmos janelas aos segmentos dos sinais de entrada, teremos uma nova função de transferência, que podemos observar janelando o filtro resultante de um dos canais do FFB.

A figura 4.5 mostra o resultado desses janelamentos. A seletividade dos canais FFB não sofre tanta alteração quanto a dos canais FFT, pois na banda de rejeição a atenuação é bastante alta, o que impede que o janelamento cause grandes alterações. Isto é mais uma vantagem do FFB, mantendo sua alta seletividade.

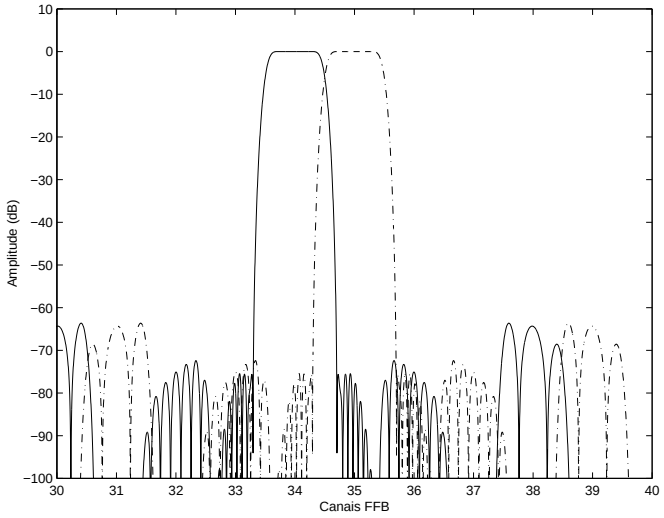


Figura 4.5: Módulos das respostas dos canais 34 ($|H'_{\mathrm{FFB}_34}(z)|$, em linha contínua) e 35 ($|H'_{\mathrm{FFB}_35}(z)|$, em linha tracejada) de uma sFFT de 256 bandas, com janelamento Hanning.

4.5.2 Complexidade computacional do FFB

Para se ter uma idéia do custo computacional do FFB, segue nesta sub-seção um estudo sobre o número de multiplicações complexas por amostra por canal, dado em [34] como ligeiramente maior que 1.

Baseada no projeto de filtros de [9], a tabela 4.1 apresenta uma relação com o número de coeficientes por nível de FRM (*frequency response masking*)[37] da estrutura de sub-filtros FFB. Note, pela tabela, que a partir de certo nível, os filtros passam a ter apenas um coeficiente simétrico diferente de zero e de um. Com isso,

é possível estabelecer um cálculo seguro da complexidade do FFB.

Tabela 4.1: Relação com quantidade de coeficientes por nível na estrutura de sub-filtros FFB.

Nível	Coeficientes por filtro	Filtros	Coeficientes por nível	Total de coeficientes
$N-1$	6	1	6	6
$N-2$	5	2	10	16
$N-3$	2	4	8	24
$N-4$	2	8	16	40
$N-5$	1	16	16	56
$N-6$	1	32	32	88
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	1	2^{N-2}	2^{N-2}	$N+24$

A tabela 4.2 apresenta o cálculo da complexidade computacional do FFB para alguns valores de N, indicando ainda o cálculo para um N genérico.

Tabela 4.2: Complexidade computacional do FFB: número de multiplicações complexas por amostra por canal.

N	Total de coeficientes	Multiplicações complexas
2	6	6/2=3
4	16	16/4=4
8	24	24/8=3
16	16+24	40/16=2,5
32	32+24	56/32=1,75
64	64+24	88/64=1,375
$\vdots$	$\vdots$	$\vdots$
N	$N+24$	$(N+24)/N \approx 1,0$

A partir da terceira coluna da tabela 4.2 temos que a complexidade do FFB é igual a $(N+24)/N$ multiplicações complexas por amostra por canal, para $N \geq 16$. Essa complexidade decresce à medida que aumentamos N, tendendo a se aproximar de 1,0.

4.6 CQFFB

O FFB mostra-se um banco de filtros rápido e de alta seletividade, porém o problema da resolução no domínio da freqüência ainda não está resolvido para os sinais de música.

Nesta seção e na seguinte, novas ferramentas são apresentadas, visando a adaptar o FFB aos sinais musicais, tornando logarítmica a distribuição de faixas de freqüência dos canais de saída.

O CQFFB está para o FFB assim como a CQT está para a FFT. O CQFFB usa os sub-filtros do FFB, adequando-os a um banco com fator de qualidade Q constante. Para tanto, é preciso apenas definir o Q e as freqüências mínima f_{min} e máxima f_{max} desejados. As implementações previstas para o CQFFB estão descritas nas Seções 4.6.1 e 4.6.2, a seguir.

4.6.1 Implementação 1: CQFFB

Esta foi a primeira implementação desenvolvida para o CQFFB [32]. Envolve apenas um canal FFB, sem que este sofra alterações. Isto garante a baixa complexidade computacional do FFB.

Dado o Q desejado, escolhe-se o canal Q de um FFB de N canais de tal forma que $N/2$ seja o menor inteiro maior que ou igual a Q, onde Q é um número inteiro. Escolher o canal Q sem poder alterá-lo mantém as características de simetria e meia-banda dos sub-filtros. Isto implica ainda que esta implementação seja usada apenas para os casos em que o Q desejado seja inteiro ou possa ser bem aproximado por um inteiro, o que não é uma restrição muito significativa, o que será visto mais adiante.

Se o canal for reamostrado, ele manterá seu fator Q inalterado, porém perde as características de simetria e meia-banda, tornando-se menos eficiente. Portanto, segue-se o caminho inverso, ou seja, é preciso reamostrar o sinal de entrada.

A estrutura do CQFFB é mostrada na figura 4.6. Ela obedece ao algoritmo descrito a seguir.

- Definir um Q inteiro, f_{min} e f_{max};
- Calcular o conjunto de $f_k = f_{min}\{1, Q, Q^2, Q^3, ..., f_{max}/f_{min}\}$;
- Definir o canal $H_{\mathrm{FFB}_Q}(z)$;
- Obter os sinais $X'_k(z)$ resultantes da reamostragem do sinal de entrada $X(z)$ pelos fatores $R = f_k/f_Q \approx R_{I_k}/R_{D_k}$, onde f_Q é a freqüência central da banda passante de $H_{\mathrm{FFB}_Q}(z)$, R_{I_k} e R_{D_k} são, respectivamente, os fatores de interpolação e decimação, que devem ser inteiros;
- Filtrar os sinais $X_k(z)$ com o filtro $H_{\mathrm{FFB}_Q}(z)$.

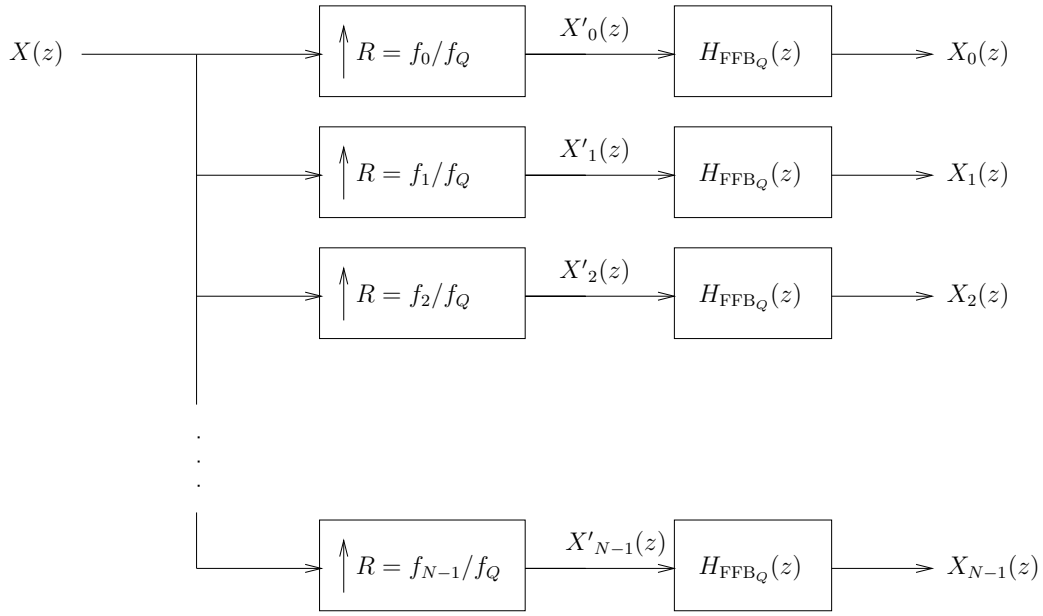


Figura 4.6: Estrutura do CQFFB. Neste diagrama, os blocos de reamostragem incluem ainda os filtros *anti-aliasing*.

Para o cálculo da complexidade, devem-se computar as multiplicações complexas presentes na filtragem pelo canal e as multiplicações simples referentes às reamostragens do sinal.

A partir do resultado da equação (4.27), o canal 34 é o melhor para obter resolução de um quarto-de-tom. Este canal deve ser escolhido de um FFB de 128 canais, que contém seis níveis de sub-filtros em sua estrutura. Segundo a tabela 4.1 e lembrando que para um único canal cada nível tem apenas um sub-filtro, o total

de coeficientes é igual a 18 (soma dos sete primeiros elementos da segunda coluna da tabela 4.1). Assim, cada filtragem com $H_{\mathrm{FFB}_{34}}(z)$ tem 18 multiplicações complexas envolvidas.

Nas reamostragens, utilizamos um filtro FIR *anti-aliasing* espectral com $20max(R_{I_k}, R_{D_k})$ coeficientes, onde $max(R_{I_k}, R_{D_k})$ indica o valor máximo entre R_{I_k} e R_{D_k}. Para que a precisão do fator de reamostragem resultante R fosse boa, optamos por fazer $R_{I_k} = R/1000$ e $R_{D_k} = 1000$, assim $max(R_{I_k}, R_{D_k}) = 1000$ para canais CQFFB onde $R_{D_k} > R_{I_k}$. Estes correspondem a quase a metade dos canais, pois 34 é ligeiramente maior que 64 ($N/2$). Para os canais onde $R_{D_k} < R_{I_k}$, isto é, $R_{I_k} > 1000$, $1000 < max(R_{I_k}, R_{D_k}) < 2000$. Isto leva a cerca de 20000 multiplicações simples por amostra por reamostragem do sinal de entrada para o cálculo em freqüências $f_k < f_Q$. Nas freqüências $f_k > f_Q$, a complexidade vai de 20000 a 40000. Podemos estimar uma média de 25000 para cada uma das freqüências f_k.

Porém, este alto valor é gasto apenas na primeira oitava calculada. Nas oitavas seguintes, podem-se aproveitar as reamostragens anteriores. Isto significa que nas oitavas seguintes, como a reamostragem é de oitava para oitava entre f_k e $2f_k$, tem-se $max(R_{I_k}, R_{D_k}) = 2$, levando a 40 multiplicações simples por amostra por reamostragem do sinal de entrada.

4.6.2 Implementação 2: mCQFFB

A segunda implementação desenvolvida foi chamada de mCQFFB (*modified*-CQFFB) [33]. Aqui, é preciso escolher um filtro-protótipo $H_{\mathrm{mC}Q\mathrm{FFB}_p}(z)$ com o Q desejado e reamostrá-lo para obter todos os demais filtros. Não existe mais a preocupação de se perder as características FFB como a reduzida complexidade, pois sua aplicação é atender à utilização de um fator real Q qualquer, diferente de inteiro.

Para a obtenção de $H_{\mathrm{mC}Q\mathrm{FFB}_p}(z)$, dado um Q, escolhe-se um canal $H_{\mathrm{FFB}_Q}(z)$ com fator de qualidade mais próximo de Q e altera-se o filtro resultante $H(z)$ deste canal para que tenha o fator de qualidade igual a Q. Para tanto, é necessário modular $H_{\mathrm{FFB}_Q}(z)$.

Nesta implementação, em vez de reamostrar o sinal de entrada, deve-se reamostrar $H_{\mathrm{mC}Q\mathrm{FFB}_p}(z)$ para obter os filtros $H_{\mathrm{mC}Q\mathrm{FFB}_k}(z)$ de acordo com as freqüências

centrais f_k. Os filtros perdem as características de baixa complexidade do FFB, mas, por outro lado, são reamostrados apenas no projeto, não na execução.

O algoritmo da implementação segue abaixo e a estrutura mCQFFB se apresenta na figura 4.7.

- Definir Q, f_{min} e f_{max};
- Calcular o conjunto de $f_k = f_{min}\{1, Q, Q^2, Q^3, ..., f_{max}/f_{min}\}$;
- Definir o filtro-protótipo $H_{\mathrm{mC}Q\mathrm{FFB}_p}(z)$ e obter f_Q, a freqüência central de sua banda passante ;
- Obter os filtros $H_{\mathrm{mC}Q\mathrm{FFB}_k}(z)$, através da reamostragem de $H_{\mathrm{mC}Q\mathrm{FFB}_p}(z)$ pelo fator $R = f_Q/f_k$;
- Filtrar o sinal de entrada com $H_{\mathrm{mC}Q\mathrm{FFB}_k}(z)$.

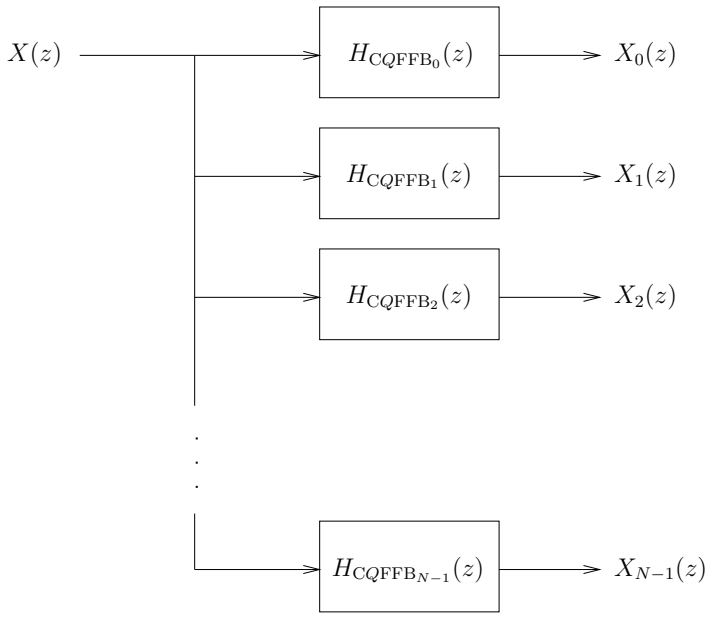


Figura 4.7: Estrutura do mCQFFB.

Para calcular a complexidade por amostra do mCQFFB, basta somar os comprimentos dos filtros mCQFFB para o valor de Q desejado. Para a complexidade por amostra por canal deve-se dividir o total pelo número de canais de saída.

4.7 BQFFB

O CQFFB mostra-se muito eficaz na análise espectral, porém carrega ainda uma alta complexidade computacional. O BQFFB (*bounded*-Q *fast filter bank*) objetiva reduzir a complexidade, buscando manter uma resolução decrescente com a freqüência.

O BQFFB está para o FFB assim como a BQT está para a FFT. O BQFFB divide o espectro em oitavas e aplica a cada oitava um FFB de N canais. Esse método mantém a resolução constante dentro de uma mesma oitava, porém uma oitava mais alta tem resolução igual à metade da de sua adjacente mais baixa.

O filtro *anti-aliasing* aplicado às decimações do BQFFB pode ser um passa-baixas com freqüência de quebra f_p, freqüência de corte $f_r = 0{,}5$ e freqüência de amostragem $f_a = 2$. Assim, as últimas amostras de saída de cada FFB, que estiverem sobre a banda de transição (de f_p a 0,5) devem ser descartadas, conforme se vê na figura 4.8.

A figura 4.8 apresenta a estrutura do BQFFB. O seu algoritmo de implementação é descrito abaixo:

- Definir o sinal de entrada, cuja taxa de amostragem chamaremos de f_a;
- Definir a resolução Res_{max}, a ser aplicada nas freqüências mais baixas;
- Definir N, o número de canais de saída do FFB a cada oitava;
- Calcular quantos passos P são necessários para atingir a resolução Res_{max}: $P > \log_2(\frac{f_a/N}{Res_{max}})$;
- Executar o procedimento abaixo P vezes, até atingir a resolução Res_{max};
- Aplicar o FFB de análise ao sinal de entrada;
- Dos canais de saída indo de 0 a $N-1$, guardar os seguintes canais:

  $\lfloor f_p N/4 \rfloor$ a $N/2 - 1$, para o primeiro passo;

  $\lfloor f_p N/4 \rfloor$ a $\lfloor f_p N/2 \rfloor - 1$, para os passos intermediários;

  0 a $\lfloor f_p N/2 \rfloor - 1$, para o último passo;

- Decimar o sinal de entrada por 2.

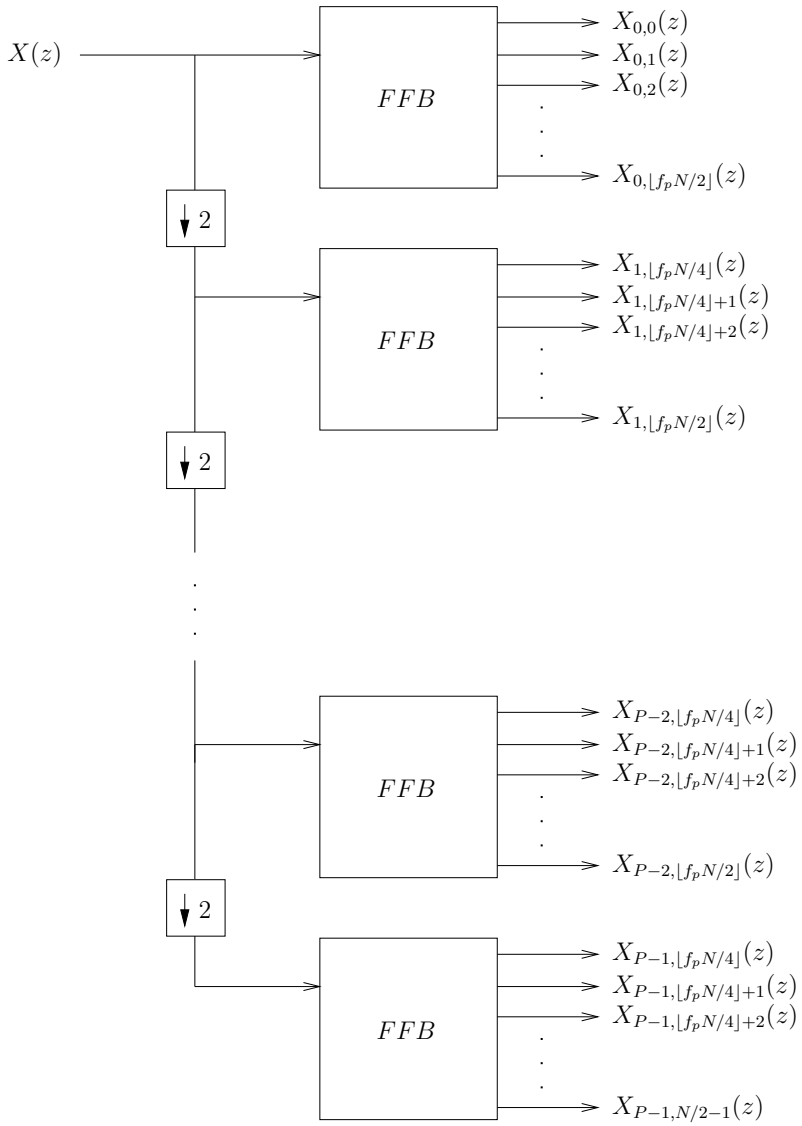


Figura 4.8: Estrutura do BQFFB.

A complexidade computacional por amostra do BQFFB deve ser calculada em duas partes: 1) as multiplicações complexas relativas aos cálculos de FFB e 2) as multiplicações simples devidas às reamostragens do sinal.

O número de multiplicações complexas é igual ao número de aplicações de FFB multiplicado pela complexidade por amostra do FFB de N canais, isto é, $P(N+24)$, para $N > 64$. Como o total de canais de saída do BQFFB é igual a PN, a complexidade por amostra por canal do BQFFB é igual à do FFB: $1 + 24/N$, para $N > 64$.

O número de multiplicações simples é igual a $(P-1) \times 40$, isto é, o número de decimações por 2, que é $(P-1)$, multiplicado pelo comprimento do filtro *anti-aliasing*

espectral.

4.8 Exemplos

4.8.1 Exemplo CQFFB

Para exemplificar a utilização do CQFFB, fizemos um teste com um sinal $x(n)$ formado por seis senóides de diferentes freqüências: 185,0, 196,0, 587,3, 622,3, 1046,5 e 1108,7 Hz (correspondendo a F#4, G4, D5, D#5, C6 e C#6, respectivamente), amostrado a uma taxa de 44,1 kHz durante 1 segundo. Este sinal foi processado via FFT, FFB, CQT e CQFFB, todas com 100 bandas indo de cerca de 130,8 a 2282,4 Hz para garantir uniformidade. Com isso, a resolução na freqüência para a FFT e o FFB, que distribuem suas saídas linearmente no domínio da freqüência, foi de

$$\Delta f_{\text{FFT}} = \Delta f_{\text{FFB}} = \frac{2282{,}4 - 130{,}8}{100} = 21{,}5 \text{ Hz}. \tag{4.32}$$

Os módulos das saídas das ferramentas de análise para o sinal $x(n)$ estão na figura 4.9. A análise espectral feita pela FFT está representada na figura 4.9a. Note que não há amostras iguais a zero e não foi possível distinguir entre as duas senóides de baixa freqüência. A resposta do FFB, na figura 4.9c, mostra a capacidade da ferramenta para mostrar as senóides puras em média e alta freqüencias, deixando, por outro lado, difusa a informação em baixa freqüência. Tanto na FFT quanto no FFB, as duas senóides de baixa freqüência estão reunidas na mesma banda e, em vez de as saídas dessa banda terem amplitude maior, apresentam baixa amplitude. Isto se deve a um vale na envoltória da saída deste canal, gerado pelo batimento resultante da proximidade das freqüências das duas senóides.

Na figura 4.9b, a CQT tem sucesso em distinguir as componentes senoidais claramente, enquanto deixa parecer que há presença de ruído. O CQFFB mostra, na figura 4.9d, que, com sua resolução logarítmica e alta seletividade, consegue distinguir claramente as senóides além de zerar as freqüências onde não há informação.

A figura 4.10 compara as saídas das transformadas para um sinal com as notas Mi3, Sol3 e Dó4. Os espectros obtidos com FFT e FFB não permitem distinguir a fundamental do Sol, "mascarada" pela fundamental do Mi. Além disso, os efeitos da baixa seletividade da FFT e da CQT podem ser observados nas regiões de baixa

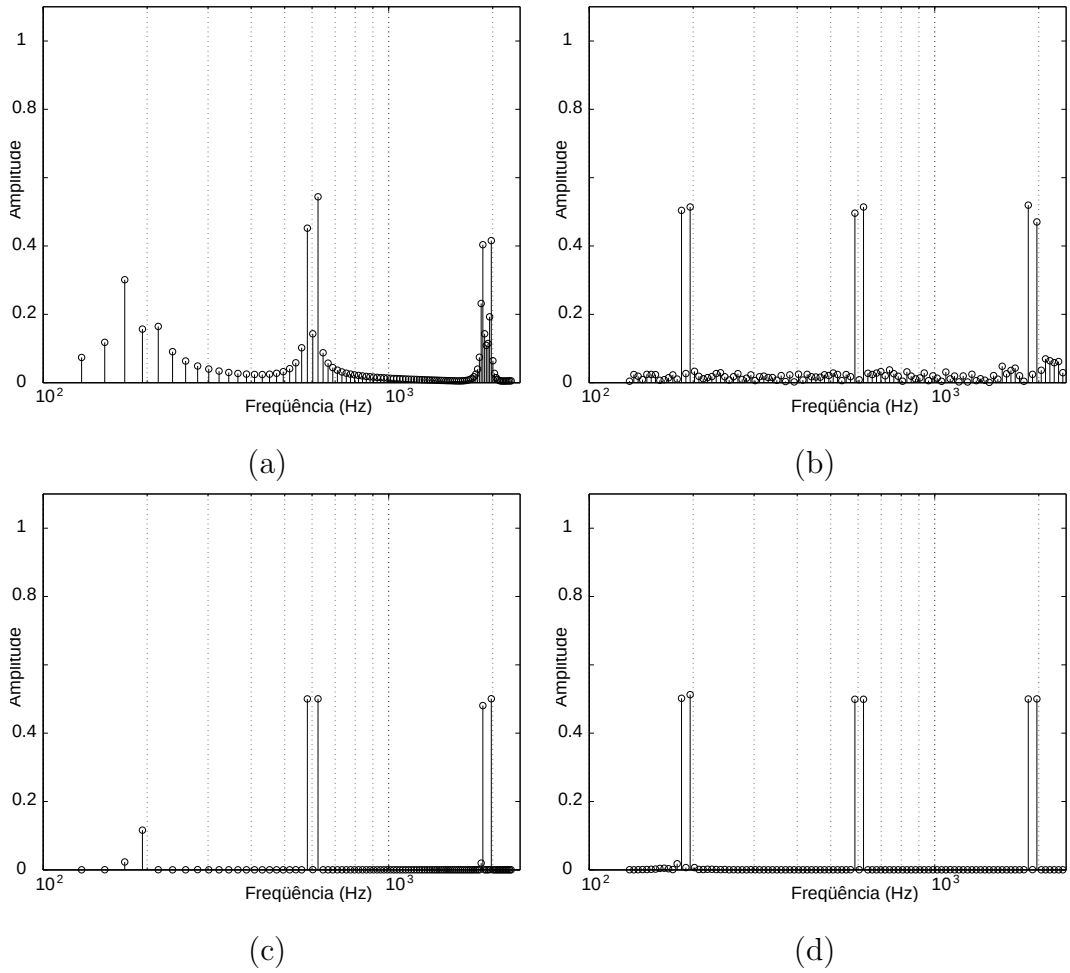


Figura 4.9: Exemplo: Módulo das respostas de transformadas de 100 amostras de um sinal de entrada composto de seis senóides: (a) FFT; (b) CQT; (c) FFB e (d) CQFFB.

amplitude espectral. A vantagem do CQFFB sobre a CQT, por sua vez, pode ser observada na faixa de freqüência entre 700 Hz e 2 kHz. O CQFFB mostra mais picos e mais bem definidos que a CQT. Dentre essas opções, o CQFFB apresenta a melhor opção de observação do espectro.

4.8.2 Exemplo BQFFB

Para entender a vantagem de usar o BQFFB, podemos comparar o uso das três ferramentas, FFB, CQFFB e BQFFB, com duas condições: 1) que todas as ferramentas apresentem resolução suficiente para distinguir um semitom na freqüência de Dó0 em 16,35 Hz e 2) que todas elas possam abranger toda a faixa entre 16,35 Hz e 22,05 kHz, para um sinal amostrado com 44,1 kHz.

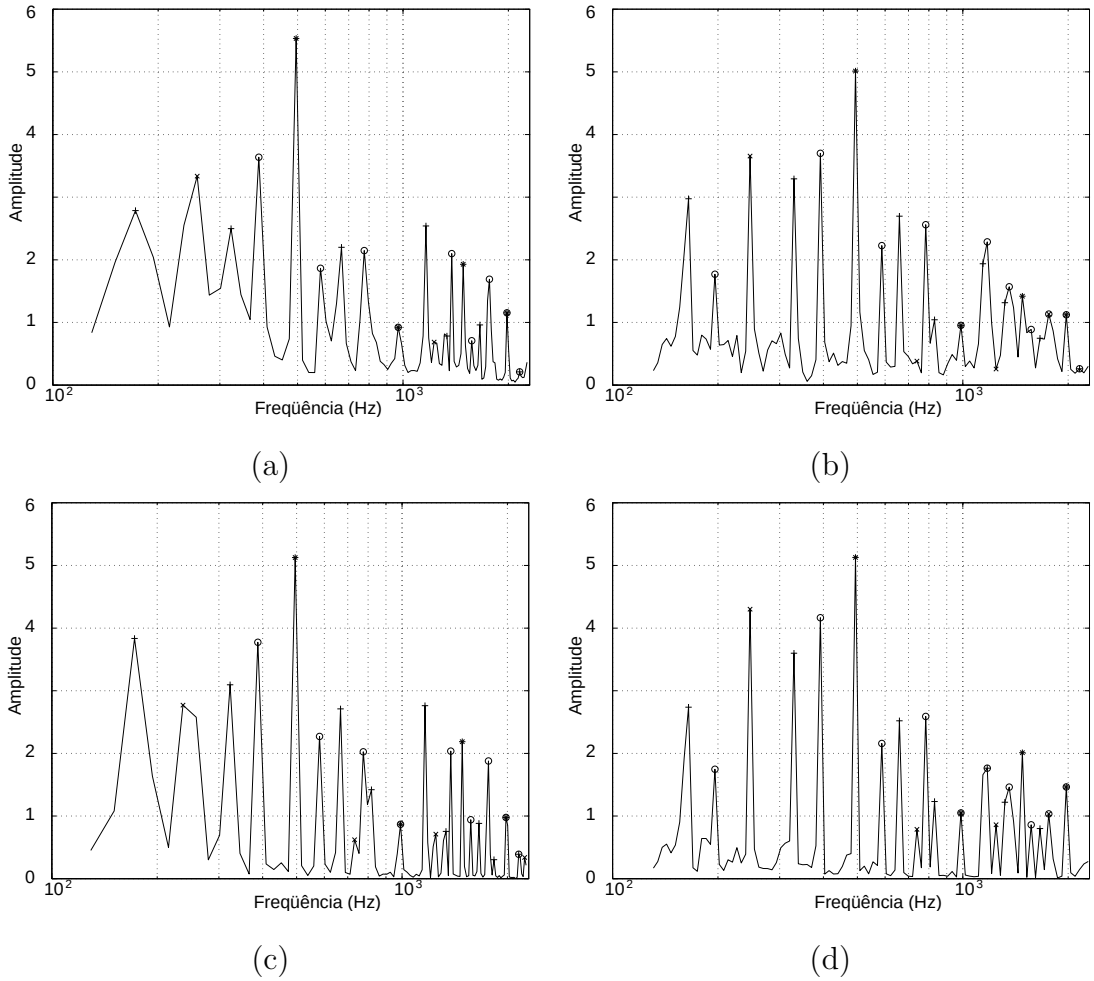


Figura 4.10: Exemplo: Módulo das respostas de transformadas de 100 amostras de um sinal de entrada composto de três notas musicais Mi3('+'), Sol3('o') e Dó4('x'): (a) FFT; (b) CQT; (c) FFB e (d) CQFFB.

Para atender às condições, foi designada a resolução de quarto-de-tom, ou seja, $Q = 34$ para o CQFFB, que ficou com 250 canais de saída. Para as reamostragens do sinal no CQFFB, utilizamos o fator de decimação $R_{D_k} = 1000$. O FFB precisou de 2^{17} canais de saída e o BQFFB precisou de 10 níveis de FFB de 256 canais, ou seja, $P + 1 = 10$ e $N = 256$.

O FFB exigiu muitos canais pois para obter a resolução fixa maior que meio semitom de 16,35 Hz ($16,35 \times (2^{1/24} - 1) \approx 0,48$ Hz), são necessários mais de 90000 canais de saída (44100/0,48 = 91875). A resolução do FFB ficou aproximadamente igual a 0,34 Hz ($44100/2^{17} \approx 0,34$). A configuração B$Q$FFB ficou com resolução aproximada de 172,27 Hz na oitava mais alta e resolução aproximada de 0,34 Hz na região de mais baixas freqüências.

Para a comparação dos custos computacionais das três ferramentas neste exemplo, veja a tabela 4.3:

Tabela 4.3: Comparação entre as custos computacionais referentes ao exemplo 4.8.2.

Ferramenta	N	Resolução	Multiplicações complexas	Multiplicações simples
FFB	2^{17}=131072	0,34 Hz	$N + 24 = 131096$	0
CQFFB	250 (semitons)	0,5 a 627,7 Hz	$17N = 4250$	≈ 609040
BQFFB	$10(256) = 2560$	0,3 a 172,3 Hz	$P(N + 24) = 2800$	$40(P - 1) = 320$

4.9 Conclusão

Na literatura, podem-se encontrar a FFT, a CQT, a BQT e o FFB, em trabalhos de transcrição musical. Aqui, neste capítulo, foram descritas e comparadas estas ferramentas de análise espectral com os novos CQFFB e BQFFB. Estas duas novas ferramentas são apresentadas, trazendo a vantagem de unir a) a distribuição de faixas de freqüência de saída segundo uma progressão geométrica e b) a alta seletividade do FFB. O BQFFB, porém, é o escolhido para servir de base aos testes descritos no próximo capítulo por contar ainda com a mais baixa complexidade computacional, sendo, por isso, a ferramenta mais eficiente.

Capítulo 5

Testes

5.1 Introdução

O Capítulo 4 descreve a ferramenta BQFFB como um banco de filtros apto a representar de forma eficiente o espectro de um sinal de música para a TMA. No presente capítulo, utilizaremos o BQFFB para obter o espectro de alguns trechos musicais e aplicaremos sobre esses espectros um algoritmo simplificado de identificação de notas. Esse algoritmo foi desenvolvido por nós e será apresentado na Seção 5.2. O objetivo, aqui, é demonstrar a eficácia do BQFFB na obtenção do espectro no problema de TMA.

Para essa avaliação do BQFFB, usamos três conjuntos de testes: 1) sinais midi com 12 notas formando 12 semitons adjacentes; 2) sinais midi com acordes de 7 e 4 notas, respectivamente, nos formatos Dó-Mi-Sol-Si-Ré-Fá-Lá e Dó-Mi-Sol-Si; e 3) gravação de uma peça de Chopin.

Todos os sinais são compostos por notas de piano. A escolha deste intrumento se deveu a ele ser o mais utilizado nos experimentos de TMA descritos na literatura. Isto nos ajuda a manter um bom parâmetro de comparação.

Os sinais midi foram obtidos do sintetizador Midi Orchestrator, de uma placa de som da Turtle Beach Systems (Pinnacle Project Studio). Primeiro, foi gerado um arquivo midi para cada nota desejada. Depois, cada arquivo midi foi gravado em formato ".wav". Esses sinais foram gravados com 44100 Hz de freqüência de amostragem, tendo comprimento médio de 20000 amostras. Para formar os acordes, os sinais das notas foram somados.

O primeiro conjunto de testes, com 12 notas, apresentado na Seção 5.3, serve para avaliar, na prática, o limite da resolução do BQFFB. Neste caso, os sinais ficam com as máximas quantidade e aproximação possíveis de freqüências fundamentais de diferentes notas musicais no espectro.

O segundo conjunto de testes, com acordes, permite verificar a envoltória espectral do BQFFB diante de intervalos musicais mais freqüentemente encontrados. A Seção 5.4 descreve os resultados destes testes.

Os dois primeiros conjuntos, com sinais midi, permitem manipular facilmente sua formação para montar combinações interessantes de notas.

A Seção 5.5 apresenta o terceiro conjunto de testes, que utiliza acordes de gravação com piano real. Neste caso, para evitar problemas com o transitório, foi escolhida uma peça de andamento lento.

A Seção 5.6 encerra este capítulo, analisando os resultados dos testes e avaliando a contribuição do BQFFB à TMA.

5.2 Algoritmo

O algoritmo desenvolvido para a identificação de notas se baseia na observação de picos na envoltória espectral. Optamos por não filtrar os sinais com o BQFFB, mas calcular os espectros por segmentos de sinal. Desta forma, esses espectros equivalem aos obtidos por uma transformada, por isso, chamamos o processo de BQFFBT (*bounded*-Q *fast filter bank transform*).

O cálculo da BQFFBT foi feito da seguinte forma. Todos os níveis de FFBT deveriam ser calculados sobre segmentos que tivessem, aproximadamente, o mesmo instante de tempo central. Escolhemos, então, 9 instantes de tempo que dividissem os sinais em 10 segmentos de iguail comprimento.

O passo-a-passo do algoritmo segue abaixo:

- Segmentar o sinal entre instantes de tempo que definam os inícios de duas notas ou acordes subseqüentes (em nosso trabalho fizemos a segmentação manualmente);

- Calcular os espectros usando BQFFBT com 8 níveis de FFBT de 256 canais de saída, ou seja, $P = 8$, $N = 256$ e cada canal de saída resultante corresponde

a um filtro de ordem 4286;

- Designar como candidatos a freqüências fundamentais os picos da envoltória espectral que estiverem acima de um limiar de amplitude, a ser determinado;
- Designar como candidatos a harmônicos os picos e suas amostras adjacentes na envoltória espectral que estiverem acima de um limiar de amplitude, a ser determinado;
- Buscar os harmônicos de um candidato a fundamental entre as freqüências múltiplas de sua freqüência fundamental;
- Exigir de um candidato a fundamental que sejam encontrados ao menos os 3 primeiros harmônicos (valor definido empiricamente), para evitar que picos espúrios sejam identificados como notas falsas;
- Eliminar os candidatos a fundamental que sejam harmônicos de outras notas, evitando intervalos de oitavas;

Algumas questões sobre este algoritmo merecem destaque. Para melhor compreensão do objetivo e do método, os pontos abaixo indicam que o algoritmo:

- identifica notas com a condição de haver picos na envoltória espectral para cada freqüência fundamental;
- baseia-se na observação dos picos do espectro para a identificação dos harmônicos;
- interrompe a busca de harmônicos de uma nota (ou série harmônica) se deixar de encontrar dois harmônicos seguidos;
- admite que um mesmo pico seja definido como harmônico de mais de uma nota, devido à sobreposição de harmônicos entre notas simultâneas;
- não identifica intervalos de oitava, ou seja, notas cujas fundamentais sejam múltiplas de fundamentais de outras notas;
- trabalha apenas com as notas da escala musical entre Dó2 e Dó8 (aproximadamente entre 65 e 4186 Hz).

A configuração escolhida para o BQFFB ($P = 8$ e $N = 256$ canais) permite que cada oitava, da oitava 2 à mais alta, tenha resolução suficiente para distinguir duas freqüências que estejam distantes uma da outra de mais do que a metade do menor semitom. Se quiséssemos, por exemplo, abranger até a oitava 0 com esta resolução, seria necessário usar $P = 10$.

Um dos itens do algoritmo inclui, além dos picos, as amostras adjacentes a eles como candidatos a harmônicos [3]. Isto é importante, pois há casos em que o pico de um harmônico "mascara" outro harmônico.

O piano tem notas mais graves que o Dó2, mas estas notas possuem harmônicos fundamentais com amplitude muito baixa, confundindo-se com informações ruidosas. Essas notas e os intervalos de oitava representam casos à parte das demais e merecem um tratamento especial, como o reconhecimento da nota pelo modelo de timbre, fora do escopo desta tese.

As seções a seguir descrevem os testes e trazem tabelas com os resultados obtidos.

5.3 Teste 1: sinais com 12 semitons

Nosso primeiro conjunto de testes para o algoritmo da Seção 5.2 teve por objetivo avaliar a capacidade do BQFFB de distinguir semitons.

Foram observados 13 sinais, cada um com 12 notas adjacentes. A disposição das notas foi feita seguindo o modelo do primeiro sinal: Dó2-Dó#2-Ré2-Ré#2-Mi2-Fá2-Fá#2-Sol2-Sol#2-Lá2-Lá#2-Si2. O segundo sinal foi do Dó#2 ao Dó3 e o último sinal foi do Dó3 ao Si3. A tabela 5.1 mostra os resultados encontrados.

Os resultados indicam que poucas notas foram perdidas. Essas notas tiveram as amplitudes de seus harmônicos fundamentais "mascaradas" por picos vizinhos. Isso ocorreu com apenas 4 notas (Ré#2, Mi2, Sol#2 e Lá2), dentre as 24 notas utilizadas.

O sinal com as notas Ré#2 a Ré3 contou com 3 notas "mascaradas" e tem as primeiras freqüências do espectro do quinto segmento mostradas na figura 5.1. As 3 notas foram o Mi2, o Sol#2 e o Lá2. Essas notas não tiveram suas freqüências fundamentais identificadas no espectro da BQFFBT devido à interferência destrutiva

Tabela 5.1: Resultado da identificação de notas entre os sinais com 12 notas simultâneas, com máximo, média e mínimo de notas perdidas por segmento de sinal. Os sinais foram divididos em 9 segmentos.

Notas	Máximo	Média	Mínimo
Dó2 a Si2	3	2,00	0
Dó#2 a Dó3	4	2,33	0
Ré2 a Dó#3	4	2,56	1
Ré#2 a Ré3	3	1,67	0
Mi2 a Ré#3	2	1,00	0
Fá2 a Mi3	2	1,00	0
Fá#2 a Fá3	3	0,56	0
Sol2 a Fá#3	2	0,56	0
Sol#2 a Sol3	1	0,22	0
Lá2 a Sol#3	1	0,22	0
Lá#2 a Lá3	0	0,00	0
Si2 a Lá#3	0	0,00	0
Dó3 a Si3	0	0,00	0

de outras notas presentes no espectro.

O gráfico mostra as notas identificadas: Ré#2 (‘o’), Fá2 (‘x’), Fá#2 (‘+’), Sol2 (‘*’), Lá#2 (‘quadrado’), Si2 (‘losango’), Dó3 (‘triângulo-V’), Dó#3 (‘estrela’) e Ré3(‘.’). Além dessas notas, o algoritmo identificou o Mi3 (‘Δ’), Sol#3 (‘>’) e Lá3 (‘<’). Essas 3 últimas notas foram identificadas na oitava 3 e não na oitava 2 porque seus harmônicos fundamentais foram “mascarados”.

O fato de que essas notas foram percebidas na oitava superior mostra que o BQFFB pôde distinguir os demais harmônicos dessas notas. Assim, um algoritmo um pouco mais sofisticado poderia ser suficiente para identificar as notas que não tiveram suas freqüências fundamentais encontradas.

5.4 Teste 2: sinais com acordes de 7 e 4 notas

O segundo conjunto de testes do algoritmo foi dividido em dois sub-conjuntos de acordes com 7 e 4 notas. O primeiro conjunto deu origem a 3 sinais e o segundo,

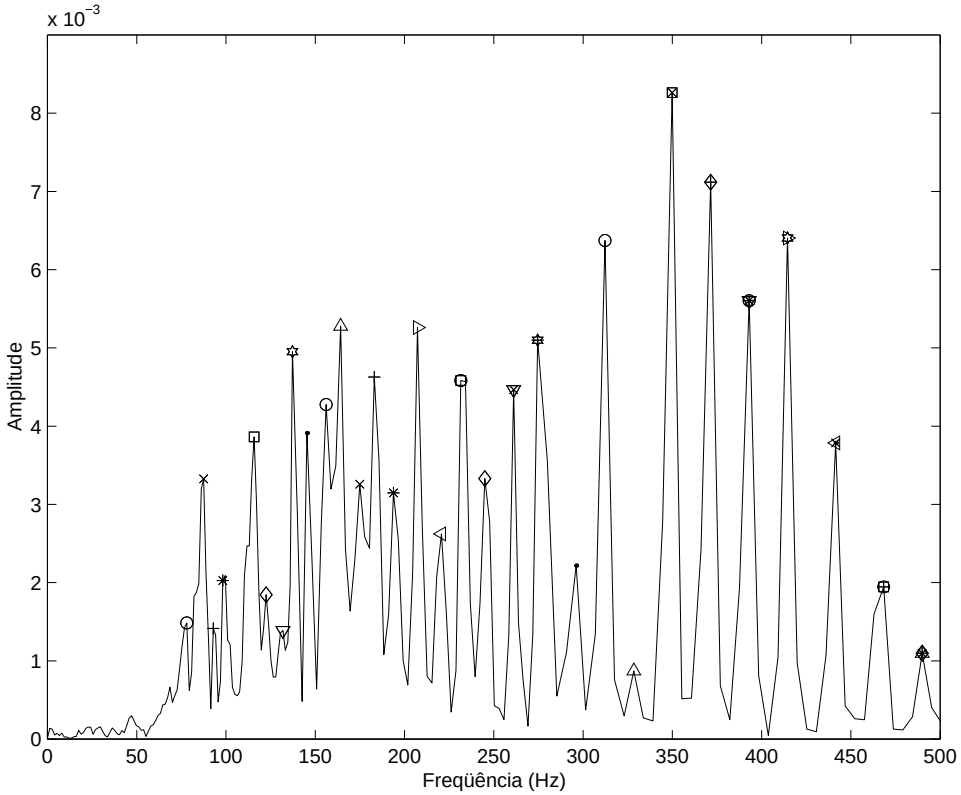


Figura 5.1: Espectro BQFFB de sinal composto pelas seguintes notas tocadas por piano sintetizado: Ré#2 (‘o’), Fá2 (‘x’), Fá#2 (‘+’), Sol2 (‘*’), Lá#2 (‘quadrado’), Si2 (‘losango’), Dó3 (‘V’), Dó#3 (‘estrela’) e Ré3(‘.’). As notas Mi3 (‘Δ’), Sol#3 (‘$>$’) e Lá3 (‘$<$’) foram identificadas com uma oitava a mais do que o devido por haver “mascaramento” de suas freqüências fundamentais.

a 13 sinais.

Neste teste, todos os segmentos tiveram o mesmo resultado: nenhuma das notas foi perdida nos dois conjuntos de acordes. Dentre esses conjuntos, dois casos são ilustrados nas figuras 5.2 e 5.3, respectivamente, com 7 e 4 notas.

A figura 5.2 mostra que, além das notas identificadas, há picos em faixas de freqüência de alta amplitude na envoltória espectral em torno das freqüências 87,31 Hz (Fá2), 110 Hz (Lá2) e 116,54 Hz (Lá#2). Estas notas falsas foram detectadas pelo algoritmo. As faixas de freqüência com alta amplitude mencionadas tiveram origem no modelo do sintetizador utilizado, tendo sido observadas mesmo nos espectros de algumas notas formadoras do acorde, quando analisadas individualmente. Na figura 5.3, há ainda o caso da nota Ré4 identificada, apesar de ausente no acorde.

O número de notas falsas encontradas no espectro pode ser reduzido com um processo de TMA que utilize modelos de notas previamente gravados. Com eles, é possível verificar se os picos encontrados podem ser associados ao modelo das notas correspondentes.

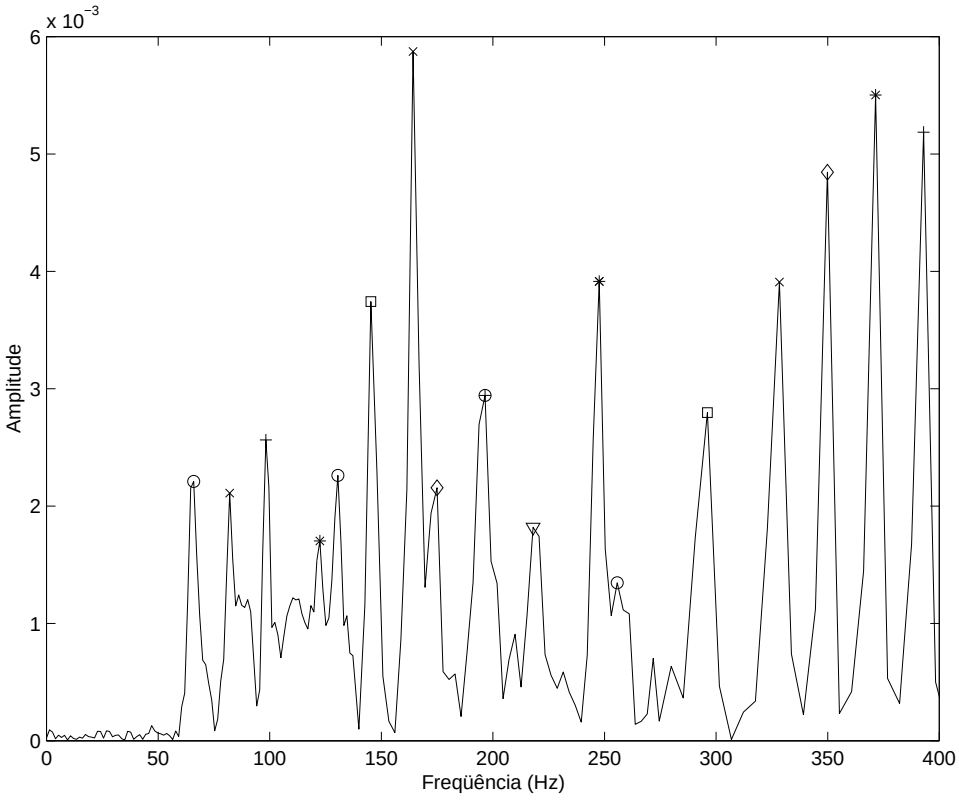


Figura 5.2: Espectro BQFFB de sinal composto pelas seguintes notas tocadas por piano sintetizado: Dó2 (‘o’), Mi2 (‘x’), Sol2 (‘+’), Si2 (‘*’), Ré3 (‘quadrado’), Fá3 (‘losango’) e Lá3 (‘V’).

5.5 Teste 3: peça de Chopin

Após a avaliação do algoritmo com notas sintetizadas, escolhemos um conjunto de testes que utiliza uma gravação de piano real. A peça musical escolhida foi o “Prelúdio Opus 28 No. 20 (‘Marcha Fúnebre’)”, em Dó Menor, de Frédéric Chopin. Foram testados seus 10 primeiros acordes, que são descritos na tabela 5.2.

Nestes acordes, foram comumente encontradas notas da oitava 1 e intervalos de uma ou mais oitavas. Nestes casos, as notas válidas para o nosso algoritmo seriam

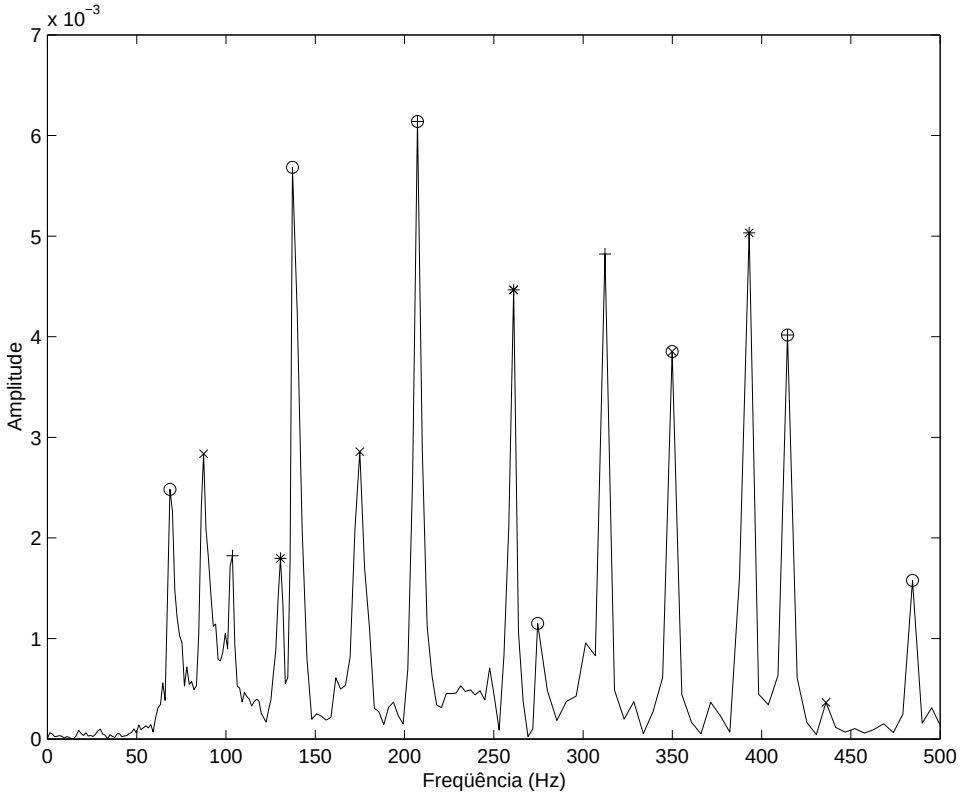


Figura 5.3: Espectro BQFFB de sinal composto pelas seguintes notas tocadas por piano sintetizado: Dó#2 (‘o’), Fá2 (‘x’), Sol#2 (‘+’) e Dó3 (‘*’).

apenas as notas mais graves dos intervalos de oitava e aquelas que não fossem da oitava 1.

Assim como no teste 2, todos os segmentos tiveram o mesmo resultado: nenhuma das notas foi perdida. Neste teste, os acordes tinham, em média, 2 ou 3 notas válidas, e estas foram todas identificadas. Quanto às notas falsas identificadas, elas ocorreram em maior número. Para estas, a mesma ressalva do teste anterior deve ser feita, isto é, um algoritmo que compare os picos encontrados com modelos das notas pode reduzir o número de notas falsas. Os erros serão reduzidos ainda mais caso estes modelos não contemplem somente um espectro, mas levem em consideração a evolução do espectro ao longo do tempo.

Tabela 5.2: Dez primeiros acordes da peça de Chopin para piano, usados no teste 3.

Acorde	Notas
1	Dó2, Dó3, Sol3, Dó4, Mi♭4, Sol4
2	Fá1, Fá2, Lá♭3, Dó4, Mi♭4, Lá♭4
3	Sol1, Sol2, Sol3, Si4, Mi♭4, Sol4
4	Ré4, Fá4
5	Dó2, Sol2, Dó3, Mi♭3, Sol3, Dó4, Mi♭4
6	Lá♭1, Lá♭2, Mi♭3, Lá♭3, Dó4, Mi♭4
7	Ré♭1, Ré♭2, Fá3, Lá♭3, Ré♭4, Fá4
8	Mi♭1, Mi♭2, Ré♭3, Mi♭3, Sol3, Dó4, Mi♭4
9	Si♭3, Ré♭4
10	Lá♭1, Lá♭2, Dó3, Mi♭3, Lá♭3, Dó4

5.6 Conclusão

Os testes apresentados, neste capítulo, tiveram como objetivo principal avaliar a aplicabilidade do BQFFB na TMA. Para tanto, foi utilizado um algoritmo simples de identificação de notas através da análise da envoltória espectral de um sinal de música.

Na oitava 1, que vai aproximadamente de 32 a 64 Hz, porém, há dificuldades especiais ligadas à má formação dos harmônicos fundamentais, dificilmente identificados no espectro, exigindo esforços além de um simples algoritmo de identificação através de picos. Em [3], há a sugestão de inserir artificialmente picos no espectro sobre freqüências graves para ajudar o algoritmo a identificar eventuais notas, porém usar este artifício pode levar à incorreta identificação de uma nota desta oitava, pois pode haver outras notas cujas fundamentais sirvam à nota mais grave como seus harmônicos. Outro método interessante seria utilizar o modelo de timbre dessas notas e acordes para identificar a nota através do exercício de comparação com análise por ressíntese.

Com notas nas oitavas 2 e 3 (aproximadamente de 65 a 260 Hz), conseguimos bons resultados em acordes com quatro e sete notas, além dos casos com doze notas adjacentes simultâneas.

Os resultados obtidos permitem inferir que o BQFFB pode ser uma boa

alternativa por facilitar a observação de notas mais graves, permitindo alta resolução. Os testes com 12 notas adjacentes ou 12 semitons, que é o caso com o maior número de notas possível, mostrou que o número de notas em si não é um problema para esta ferramenta, que permite a distinção de intervalos de semitons na maior parte dos casos. Este tipo de teste não foi encontrado em outras publicações.

As vantagens demonstradas do BQFFB são a baixa complexidade computacional entre as ferramentas com canais de alta seletividade e que, em baixas freqüências, ele permite a observação dos picos de harmônicos com boa resolução, assim como em altas freqüências, otimizando a relação entre resoluções no tempo e na freqüência.

Capítulo 6

Conclusão

6.1 Nossa contribuição

Esta tese veio apresentar as bases e a riqueza potencial da pesquisa em transcrição musical automática (TMA), que envolve desde a segmentação do sinal no domínio do tempo até o reconhecimento de instrumentos musicais em sinais mono.

O Capítulo 1 mostra que os sistemas bem-sucedidos em TMA limitam-se a sinais com pequeno número de notas concorrentes, e não contam com a presença de instrumentos percussivos de som indeterminado. Estamos, portanto, ainda aquém da realidade das gravações de música popular comumente encontradas.

O Capítulo 2 fez uma revisão de teoria musical. O modelo senoidal foi apresentado como eficiente na representação dos harmônicos de notas musicais. Foram apresentados das notas musicais aos acordes, com ênfase nos intervalos harmônicos consonantes, responsáveis pela sobreposição de harmônicos pela maior dificuldade da TMA em sinais polifônicos.

O problema da TMA foi apresentado no Capítulo 3. As análises presentes na TMA foram divididas em quatro principais etapas: 1) a segmentação do sinal no domínio do tempo ou detecção de início de notas; 2) a identificação de notas no domínio da freqüência, a principal etapa; 3) reconhecimento de instrumentos musicais e 4) análise por ressíntese. Cada etapa foi descrita com seus obstáculos e soluções encontradas na literatura.

A maior contribuição desta tese se encontra no Capítulo 4, onde é feito um estudo das ferramentas de cálculo de espectro utilizadas em sinais de música. Parte-

se da idéia de que as transformadas DFT, CQT e BQT são bancos de filtros e que na análise da envoltória espectral são necessários bancos de filtros complexos, o que facilita a obtenção da envoltória no tempo dos sinais de saída.

O estudo mostra também o FFB, um banco de filtros pouco encontrado na literatura, utilizado em [9]. Este banco é baseado no banco de filtros *sliding* FFT, aproveitando sua estrutura rápida, mas enriquecendo sua seletividade. A tese mostra como podemos chegar da *sliding* FFT ao FFB e como partir do FFB para obter suas modificações CQFFB e BQFFB. Estas têm a vantagem de distribuir os canais de saída, respectivamente, de forma proporcional ou progressiva em relação à freqüência.

Um algoritmo básico de identificação de notas foi proposto no Capítulo 5 para avaliar, na prática, o uso do BQFFB na TMA. Os resultados obtidos foram positivos, aceitando grande número de notas simultâneas de piano. O algoritmo foi treinado com notas de um sintetizador de piano e foi testado com sucesso em um sinal de música real com média de quatro notas simultâneas.

6.2 Possível extensão da pesquisa

A presente tese está longe de esgotar o tema da TMA. Dentre algumas possibilidades de extensões do trabalho aqui apresentado, podemos citar:

- Explorar modificações sobre a seletividade e a implementação dos sub-filtros componentes do CQFFB e do BQFFB;
- Explorar variações de seletividade e resolução ao longo do espectro;
- Aprimorar o algoritmo de identificação de notas musicais;
- Comparar as ferramentas utilizadas neste trabalho com *wavelets* [36];
- Avaliar a TMA de gravações antigas na presença de ruído, que pode servir para a ressíntese do sinal musical com maior qualidade sonora;
- Desenvolver novas soluções para as demais etapas do processo, como a segmentação de sinais, o reconhecimento de instrumentos musicais e a análise por ressíntese;

- Estender o problema da TMA para auxiliar a implementação de técnicas de separação de instrumentos ou vozes, marcação de ritmo para, por exemplo, edição audiovisual etc.

Esperamos que esta tese motive a continuação das pesquisas nesta linha de trabalho.

Referências Bibliográficas

[1] EARGLE, J., *Sound Recording*. Editora Van Nostrand Reinhold, 1976.

[2] MESSERSCHMITT, D. G., VARIAN, H., *INFOSYS 224: Strategic computing and communications technology*. http://www.sims.berkeley.edu/courses/is224-/s99/GroupG/report1.html, School of Information Management and Systems, University of California, Berkeley, EUA, 1999.

[3] KLAPURI, A., *Automatic transcription of music*. M.sc. thesis, Tampere University of Technology, Departament of Information Technology, Signal Processing Laboratory, Finlândia, Abril 1998.

[4] KLAPURI, A., *Literature review on polyphonic music transcription*. http://www.cs.tut.fi/~klap/iiro/overview2001/literature.html, Departament of Information Technology, Signal Processing Laboratory, Tampere University of Technology, Finlândia, 2001.

[5] KASHINO, K., NAKADAI, K., KINOSHITA, T., *et al.*, "Application of bayesian probability network to music scene analysis". In: *Proceedings of the International Joint Conference on AI, CASA workshop*, 1995.

[6] MARTIN, K., *A blackboard system for automatic transcription of simple polyphonic music*, Technical Report 399, MIT Media Lab., Cambridge, MA, EUA, 1996.

[7] MARTIN, K., *Automatic transcription of simple polyphonic music: robust front end processing*, Technical Report 385, MIT Media Lab., Cambridge, MA, EUA, 1996.

[8] LEE, E., FOO, S. W., "An innovative approach to transcription of polyphonic signals". In: *3rd IEEE International Conference on Information, Communication and Signal Processing (ICICS 2001)*, Cingapura, Outubro 2001.

[9] LEE, E., FOO, S. W., "Transcription of polyphonic signals using fast filter bank". In: *IEEE International Symposium on Circuits and Systems - ISCAS*, pp. III.241–III.244, Scottsdale, EUA, Maio 2003.

[10] GOTO, M., "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, Istambul, Turquia, Junho 2000.

[11] NUSSENZVEIG, M., *Curso de Física Básica*. Editora Edgard Blücher, 1981.

[12] MARTIN, K., "Musical instrument identification: a pattern-recognition approach". In: *136th meeting of the Acoustical Society of America*, Outubro 1998.

[13] KLAPURI, A., ERONEN, A., "Musical instrument recognition using cepstral coefficients and temporal featues". In: *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, Istambul, Turquia, 2000.

[14] DOS SANTOS, C. N., CALôBA, L. P., BISCAINHO, L. W. P., "Discriminação neural de instrumentos musicais baseada no espectro". In: *Congresso Brasileiro de Redes Neurais*, pp. 337–341, Rio de Janeiro, Abril 2001.

[15] MEDDIS, R., HEWITT, M., "Virtual pitch and phase sensitivity of a computer model of the auditory periphery I: pitch identification", *J. Acoustic Society of America - JASA*, v. 89, n. 1, pp. 2866–2882, Junho 1991.

[16] ISAACS, A., MARTIN, E., *Dicionário de Música*. Zahar Editores, 1985.

[17] MED, B., *Teoria da Música*. Editora Musimed, 1996.

[18] BENNET, R., *Uma Breve História da Música*. Jorge Zahar, 1986.

[19] SCHEIRER, E., *Bregman's chimerae: music perception as auditory scene analysis*, Technical report, MIT Media Lab., Cambridge, MA, EUA, 1996.

[20] KLAPURI, A., "Sound onset detection by applying psychoacoustic knowledge". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1999.

[21] SCHEIRER, E., *Tempo and beat analysis of acoustic musical signals*, Technical report, MIT Media Lab., Cambridge, MA, EUA, 1996.

[22] GOTO, M., MURAOKA, Y., "A real-time beat tracking system for audio signals". In: *Proceedings of the 1995 International Computer Music Conference*, Setembro 1995.

[23] GOTO, M., MURAOKA, Y., "A beat tracking based on multiple-agent architecture - A real-time beat tracking system for audio signals". In: *Proceedings of the Second International Conference on Multiagent Systems*, 1996.

[24] MCAULAY, R. J., QUATIERI, T. F., "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 34, n. 4, Junho 1986.

[25] BROWN, J., "Calculation of a constant Q spectral transform", *J. Acoustic Society of America - JASA*, v. 89, n. 1, pp. 425–434, Janeiro 1991.

[26] LIM, Y. C., FARHANG-BOROUJENY, B., "Fast filter bank (FFB)", *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, v. 39, n. 5, pp. 316–318, Maio 1992.

[27] DOS SANTOS, C. N., BISCAINHO, L. W. P., NETTO, S. L., "Transcrição musical automática com bancos de filtros". In: *Anais da VII Convenção Nacional da Sociedade Brasileira de Áudio*, v. 1, pp. 31–35, São Paulo, Maio 2003.

[28] KLAPURI, A., "Multipitch estimation and sound separation by the spectral smoothness principle". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, Salt Lake City, EUA, 2001.

[29] DESAINTE-CATHERINE, M., MARCHAND, S., "High-precision Fourier analysis of sounds using signal derivates", *Journal of the Audio Engineering Society*, v. 48, n. 7/8, Julho 2000.

[30] KASHINO, K., MURASE, H., "A sound source identification system for ensemble music based on template adaptation and music stream extraction", *Speech Communication*, v. 27, pp. 337–349, Setembro 1999.

[31] RABINER, L. R., GOLD, B., *Theory and Application of Digital Signal Processing*. Editora Prentice-Hall, 1975.

[32] GRAZIOSI, D. B., DOS SANTOS, C. N., NETTO, S. L., *et al.*, "A constant-Q spectral transformation with improved frequency response". In: *IEEE International Symposium on Circuits and Systems - ISCAS*, Vancouver, Canadá, Maio 2004.

[33] DOS SANTOS, C. N., NETTO, S. L., BISCAINHO, L. W. P., *et al.*, "A modified constant-Q transform for audio signals". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, Montreal, Canadá, Maio 2004.

[34] FARHANG-BOROUJENY, B., LIM, Y. C., "A comment on computational complexity of sliding FFT", *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, v. 39, n. 12, pp. 875–876, Dezembro 1992.

[35] BROWN, J., "An efficient algorithm for the calculation of a constant Q transform", *J. Acoustic Society of America - JASA*, v. 92, n. 5, pp. 2698–2701, Novembro 1992.

[36] VAIDYANATHAN, P. P., *Multirate Systems and Filterbanks*. Prentice Hall Signal Processing Series, 1995.

[37] DINIZ, P. S. R., DA SILVA, E. A. B., NETTO, S. L., *Processamento Digital de Sinais: Projeto e Análise de Sistemas*. Bookman, 2004.