COPPE/UFRJ

TRANSCRIÇÃO MUSICAL AUTOMÁTICA USANDO REPRESENTAÇÃO FREQÜENCIAL EFICIENTE POR BANCO DE FILTROS DE ALTA SELETIVIDADE

Filipe Castello da Costa Beltrão Diniz

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

Orientadores: Luiz Wagner Pereira Biscainho
Sergio Lima Netto

Rio de Janeiro
Agosto de 2009

TRANSCRIÇÃO MUSICAL AUTOMÁTICA USANDO REPRESENTAÇÃO FREQÜENCIAL EFICIENTE POR BANCO DE FILTROS DE ALTA SELETIVIDADE

Filipe Castello da Costa Beltrão Diniz

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

Aprovada por:

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. Sergio Lima Netto, Ph.D.

Prof. Eduardo Antônio Barros da Silva, Ph.D.

Prof. Mariane Rembold Petraglia, Ph.D.

Prof. Márcio Nogueira de Souza, D.Sc.

Prof. Paulo Antonio Andrade Esquef, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2009

Diniz, Filipe Castello da Costa Beltrão

Transcrição musical automática usando representação freqüencial eficiente por banco de filtros de alta seletividade/Filipe Castello da Costa Beltrão Diniz. – Rio de Janeiro: UFRJ/COPPE, 2009.

XXIII, 170 p.: il.; 29, 7cm.

Orientadores: Luiz Wagner Pereira Biscainho

Sergio Lima Netto

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2009.

Referências Bibliográficas: p. 158 – 165.

1. Transcrição Musical Automática. 2. Representação freqüencial eficiente. 3. Banco de filtros de alta seletividade. I. Biscainho, Luiz Wagner Pereira *et al.*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Dedico esse trabalho a todos que
já me ensinaram algo na vida.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico dos Estados Unidos do Brasil (CNPq) pelo suporte financeiro no início da pesquisa.

A meus professores orientadores Sergio Lima Netto e Luiz Wagner Biscainho por terem sido sempre não apenas orientadores, mas verdadeiros amigos, o que me possibilitou crescer como pessoa e como pesquisador.

Aos membros da banca avaliadora Eduardo da Silva, Mariane Petraglia, Márcio de Souza e Paulo Esquef, por suas sugestões e correções.

À toda minha família por me apoiar na idéia de investir todo esse tempo e esforço no doutorado.

A meus amigos do Laboratório de Processamento de Sinais, da UFRJ e da Petrobras, principalmente: Michel Tcheou, Maurício Quélhas, Tadeu Ferreira, Leonardo Baltar, Ana Luisa Santos, Marcello Artimos, Carlo Marcelo, Iuri Kothe, José Marcio Faier, Helena Fernandez, Rodrigo Meirelles, João Baptista, Augusto Dantas, Augusto Santiago, Igor Isidro, William Roger, Fabio Freeland, Amaro de Lima, Lisandro Lovisolo, Cristiano Santos, Danilo Graziozi, Viviane Medeiros, Bruno Costa, Rodrigo Torres, Luis Guilherme Uzeda, Miguel Furtado, Ricardo Ramos, Luciana Leite, José Fernando Leite, Jürgen Wittmann, Andreas Mieling, Nuno Rodrigues, Bernardo Costa, Alexandre Leizor, Alan Tygel, Leonardo Nunes, Wallace Martins, Markus Lima, Rafael de Jesus, Guilherme Pinto, Luiz Fabiano, Camila Gussen, Ana Fernanda, Vanessa Sarmiento, Roberto Hori, Eduardo Vaz e Vitor Chvidchenko. Vocês não têm idéia do quanto foram essenciais para mim durante esse tempo todo. Vocês foram e sempre serão minha família! Foi um privilégio conviver com vocês, sempre me ensinando.

A todos os meus amigos que nem sabiam do que tratava minha tese. Obrigado por me lembrar que existe um mundo lá fora.

À minha esposa Maria Paula Diniz, por ter sido sempre uma grande motivação para seguir em frente e por entender as horas na frente do computador e minha ausência em épocas de submissão de artigos, de elaborações e revisões de capítulos.

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

TRANSCRIÇÃO MUSICAL AUTOMÁTICA USANDO REPRESENTAÇÃO FREQÜENCIAL EFICIENTE POR BANCO DE FILTROS DE ALTA SELETIVIDADE

Filipe Castello da Costa Beltrão Diniz

Agosto/2009

Orientadores: Luiz Wagner Pereira Biscainho
Sergio Lima Netto

Programa: Engenharia Elétrica

Transcrição musical é o processo que consiste em escrever a partitura ou alguma forma similar de notação musical para uma composição com base apenas na audição de sua execução. Considerando as diversas aplicações existentes e a viabilidade técnica para tal, abre-se a possibilidade de se realizar essa tarefa automaticamente, tanto a nível profissional quanto amador. Esse é o cenário para o nascimento da transcrição musical automática (TMA), cujo objetivo é obter a partitura de uma música a partir da gravação de sua representação acústica. Como a base de um sistema desses é, em geral, a análise espectral, foram desenvolvidas diversas ferramentas para este fim, que tentam compensar as limitações inerentes a certas famílias de técnicas, como baixa seletividade e distribuição ineficiente de canais. Contudo, isso pode exigir um maior custo computacional. Para contornar todas essas dificuldades, desenvolveu-se o *bounded-Q fast filter bank* ou BQFFB. Este trabalho visa a mostrar o impacto do BQFFB no ambiente da transcrição musical automática e a verificar sua aplicabilidade.

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

AUTOMATIC MUSIC TRANSCRIPTION USING EFFICIENT FREQUENCY REPRESENTATION WITH HIGH SELECTIVITY FILTER BANK

Filipe Castello da Costa Beltrão Diniz

August/2009

Advisors: Luiz Wagner Pereira Biscainho
Sergio Lima Netto

Department: Electrical Engineering

Music transcription consists in writing the score (or any similar way of notation) of a musical piece by listening to it. The multiplicity of applications and the technical viability open room to the automation of this task. This is the scenario for the birth of Automatic Music Transcription (AMT), the goal of which is to obtain music scores based on recordings. Since the basis of this kind of system is, usually, the spectral analysis, several tools were developed for this purpose, aiming to compensate the inherent limitations of certain families of techniques, such as low selectivity and inefficient channel distribution, although at the expenses of higher computational cost. In order to overcome these difficulties, the *bounded-Q fast filter bank* (BQFFB) was developed. This work aims at showing the impact of the BQFFB on AMT and assessing its applicability.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Publicações

[P1] DINIZ, F. C. DA C. B., KOTHE, I., BISCAINHO, L. W. P., NETTO, S. L., "A Bounded-Q Fast Filter Bank for Audio Signal Analysis". In: *Proc. of ITS 2006 - Int. Telecommunications Symp.*, pp. 971–975, Fortaleza, Brasil, Setembro 2006.

[P2] DINIZ, F. C. DA C. B., BISCAINHO, L. W. P., NETTO, S. L., "Avaliação do Bounded-Q Fast Filter Bank na Transcrição Automática de Música". In: *Anais do 5o Congresso Brasileiro de Engenharia de Áudio da AES-Brasil*, pp. 58–65, São Paulo, Brasil, Maio 2007.

[P3] DINIZ, F. C. DA C. B., BISCAINHO, L. W. P., NETTO, S. L., "Practical Design of Filter Banks for Automatic Music Transcription". In: *Proc. of International Symposium on Image, Signal Processing and Analysis*, pp. 1–5, Istambul, Turquia, Setembro 2007.

[P4] DINIZ, F. C. DA C. B., KOTHE, I., NETTO, S. L., BISCAINHO, L. W. P., "High-Selectivity Filter Banks for Spectral Analysis of Music Signals", *EURASIP Journal on Applied Signal Processing*, v. 2007, n. 94704, pp. 1–12, 2007.

Capítulo 1

Introdução

Transcrição musical é o processo de se obter a partitura ou alguma forma similar de registro ou notação musical de uma composição com base apenas na sua execução. Uma transcrição é tida como bem feita quando permite executar a composição de modo idêntico ao original[1] a partir desses registros. Uma vez transcrita, essa composição pode ser analisada, estudada e reexecutada por qualquer pessoa apta a ler os registros obtidos. A transcrição requer uma percepção musical bem apurada. Como isso é uma perícia quase que restrita a músicos experientes, os custos de profissionais qualificados para essa tarefa podem ser bastante elevados.

Existem diversas músicas ainda sem registro escrito. Isso pode acontecer tanto pelo fato de seu autor não ter uma formação musical sólida que lhe possibilite escrever a composição, quanto por não ter desejado ou podido divulgá-la em forma de partitura. A falta desses registros pode até ocasionar o total desaparecimento de algumas obras do cancioneiro popular, com impacto na manutenção e propagação de nossa cultura.

Devido ao grande volume de aplicações possíveis e à viabilidade técnica para tal, pode-se conceber realizar esta tarefa automaticamente, originando-se o chamado processo de transcrição musical automática (TMA), com um objetivo simples: obter a partitura de uma música a partir de sua representação acústica ou eletroacústica (como sua gravação, por exemplo).

Com a TMA é possível transcrever improvisações, que, por definição, não têm

[1] Está sendo desprezada, aqui, a questão da interpretação, que é de ordem subjetiva e acrescenta uma dificuldade adicional ao processo de transcrição.

partitura. Mesmo quando se conhece a partitura de uma determinada peça, um arranjo pode ter sido composto para uma determinada ocasião sem que se tenha tido a preocupação de guardar o registro escrito. A TMA permite resgatar este arranjo.

No setor do ensino de prática musical, a TMA também tem uma grande utilidade. Um estudante pode usar tal sistema para transcrever o que ele executar e, em seguida, comparar com o que está descrito na partitura. Com isso, erros seriam identificados com facilidade e, possivelmente, corrigidos.

Existe uma linha de codificação de áudio que é baseada no conceito de áudio estruturado paramétrico [1][2]. Isto, em sua forma mais pura, consiste em se codificar informações sobre o que é executado em termos de conteúdo musical (instruções similares às de uma partitura) ao invés de se codificar as formas de onda propriamente ditas. Uma maneira de se obter tais instruções é através da TMA.

A partir da TMA, é possível partir para objetivos ainda mais amplos, como a extração de informações de alto nível relativas ao conteúdo musical propriamente dito a partir da execução, gravada ou não, de uma dada composição. Esse processo é o que se convencionou chamar de *music information retrieval* ou MIR. Exemplos desse tipo de informações são: tonalidade, estilo da composição, instrumentos presentes na execução, identidade do autor etc. Além disso, através do MIR, pode-se realizar buscas por músicas em bancos de dados com base em sua melodia.

Por sua abrangência, o estudo da TMA tem despertado forte interesse científico e comercial. Essa área apresenta vários problemas que ainda estão por ser superados muito em parte devido à existência de indefinições quanto à melodia, ritmo e estrutura das composições, subjetividade quanto a início e fim de notas musicais, peculiaridades na discriminação entre gêneros musicais etc. Isso constitui, assim, um campo fértil para se verificar a aplicabilidade de novas técnicas de processamento de sinais.

1.1 Sistema típico

As informações mais básicas de uma partitura são a altura de cada nota, o instante em que estas devem ser tocadas e suas durações. Portanto, para se transcrever uma

peça, um sistema deve ter um comportamento similar ao mostrado na Figura 1.1.

Nesse fluxograma, a análise espectral é a primeira e principal etapa, de onde todas as outras derivam. Cada passo subseqüente é executado com base nos resultados dos passos anteriores, mas sempre em função do resultado da análise espectral. Em seguida, é feita a detecção dos instantes de início de cada nota, que, no campo da TMA, são chamados de *onsets*. Para cada *onset*, são identificadas as notas no espectro buscando-se suas freqüências fundamentais, denominadas individualmente de F0. Tendo-se encontrado uma dada F0 em um certo instante, faz-se o rastreamento dessa F0 no espectro a fim de se estimar o instante em que sua energia cai abaixo de um determinado limiar, o que indicaria a extinção da nota e, por conseguinte, sua duração. Por fim, com base nesses três conjuntos de informação, pode-se gerar a representação musical desejada.

1.2 Objetivo

O fundamento para sistemas de TMA é, na grande maioria das vezes, a análise espectral de sinais de áudio. A forma mais comum de se executar esse procedimento é através da transformada discreta de Fourier (DFT). Se o sinal sob análise atuar sucessivamente no tempo como entrada para a DFT, esta pode ser vista como as saídas de um banco de filtros [3] capaz de separar o sinal em bandas estreitas no domínio da freqüência. Uma discussão sobre a necessidade de soluções alternativas a DFT se encontra no Capítulo 5. O *bounded-Q fast filter bank* ou BQFFB [P4][P1] é uma ferramenta de análise espectral que apresenta alta seletividade a custo computacional reduzido(herdada do *fast filter bank* ou FFB [4]) e distribuição quase geométrica (como na *bounded-Q transform* ou BQT [5]). Esta tese visa a desenvolver um sistema de TMA com base no BQFFB.

1.3 Organização

O Capítulo 2 descreve o funcionamento típico de um sistema para TMA de forma genérica e apresenta alguns programas comerciais para este fim, como uma forma de mostrar avanços recentes na área de TMA. São abordadas questões como a precisão tanto temporal quanto freqüencial, faixa para procura de freqüências fundamentais,

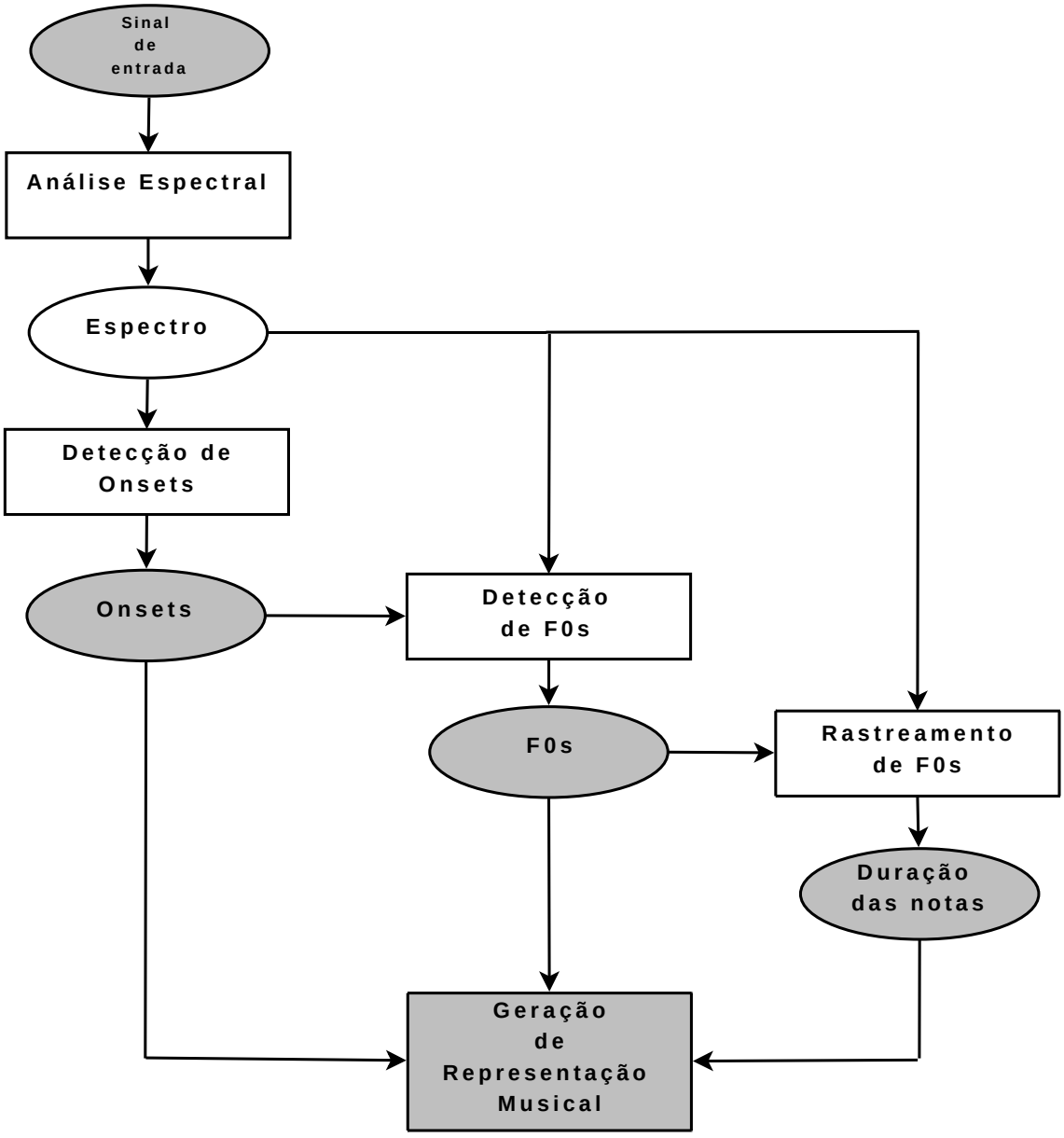


Figura 1.1: Fluxograma de um sistema típico de TMA.

interface com outros programas e formato de saída. Por fim, é feito um exame da literatura científica.

O Capítulo 3 expõe a detecção de *onsets*, que é a primeira etapa da TMA a utilizar os resultados da análise espectral, objetivando a marcação dos instantes de início de uma nota. Após algumas definições importantes para melhor compreensão do capítulo, são descritos métodos de detecção de *onsets* comuns na literatura e, quando aplicáveis, como eles são implementados no presente trabalho. Por fim, há uma discussão sobre os resultados obtidos com tais métodos.

O Capítulo 4 mostra métodos para a detecção das múltiplas freqüências fundamentais presentes no sinal. É feita uma revisão bibliográfica de métodos anteriores, fundamentando a descrição do sistema em que se baseou o módulo de detecção de freqüências fundamentais do presente trabalho. Um exemplo prático é utilizado para ilustrar cada etapa do funcionamento do sistema.

O Capítulo 5 descreve um conjunto de ferramentas para análise espectral de sinais de música caracterizadas por filtros de alta seletividade, espaçamento freqüencial mais adequado para esse tipo de sinais e complexidade computacional moderada (que não seja tão alta a ponto de ser proibitiva). Algumas características de ordem prática são discutidas e alguns testes de desempenho são realizados e analisados.

O Capítulo 6 apresenta resultados relativos ao emprego da BQFFB em cada uma das etapas do sistema de TMA, a fim de se confirmar a validade do uso dessa ferramenta para esse fim. Diversos testes são feitos, verificando-se a aplicabilidade do BQFFB nas detecções de *onsets* e de F0s. Por conta de o conceito de fim de nota musical ser algo bastante indefinido, optou-se por trabalhar apenas nas duas etapas supracitadas.

O Capítulo 7 mostra os resultados obtidos por um sistema de TMA baseado em BQFFB. É feito o detalhamento de todo o funcionamento do sistema, etapa por etapa, e por fim são mostradas graficamente algumas composições transcritas, além de análises dos experimentos.

Por fim, o Capítulo 8 aponta as contribuições do presente trabalho no campo da TMA, discute algumas questões comumente levantadas acerca desse tema e faz algumas propostas para trabalhos futuros.

O Apêndice A apresenta uma série de termos específicos de música que são

mencionados no texto, com a finalidade de auxiliar o leitor pouco familiarizado com o tema.

O Apêndice B explica os principais aspectos do funcionamento do protocolo MIDI e mostra as vantagens de seu uso na representação musical em ambiente computacional.

Capítulo 2

A transcrição musical automática

2.1 Introdução

Muito esforço já foi empregado em pesquisa na área de TMA [6][7][8][9][10]. Existem diversos programas que se propõem a realizar essa tarefa fazendo uso de tecnologias já existentes e usando como formato de saída o protocolo MIDI. Aliás, a grande maioria desses programas se caracterizam, comercialmente, como "conversores de arquivos WAV em arquivos MIDI". Esse conceito, por mais longe que esteja de uma verdadeira TMA, é largamente aceito como tal. Na Seção 2.2, são apresentados alguns programas comerciais para este fim, representando algumas soluções em termos comerciais para o problema da TMA. Também serão mostradas, na Seção 2.3, algumas soluções e valores típicos de desempenho encontrados nas pesquisas da área de TMA. Com isso, serão estabelecidos os limites aos quais se chegou nesses sistemas, abordando a precisão tanto temporal quanto freqüencial, a faixa para procura de freqüências fundamentais, a interface com outros programas e o formato de saída. A Seção 2.4 conclui o capítulo.

2.2 Programas comerciais

Existem programas comerciais que visam a executar a conversão de arquivos WAV em arquivos MIDI. Nesta seção, citam-se alguns programas com suas características, de forma a ilustrar avanços recentes nesse tema. Serão apresentados três programas: Intelliscore Ensemble®, o Neuratron AudioScore Professional 6®e o Melodyne 3.2®.

2.2.1 Intelliscore Ensemble®

Esse programa possui as seguintes funcionalidades [11]:

- Processamento de arquivos polifônicos como WAVE, MP3, WMA e CDA;
- Conversão para arquivo MIDI;
- Reconhecimento de efeitos de expressão como *tremolo* e *vibrato*;
- Alteração de notas e instrumentos;
- Reconhecimento de acordes e captura de ritmo.

Sua limitação é a incapacidade de transcrever instrumentos de percussão. Este programa tem como base três algoritmos que fazem uso de informações psicoacústicas e de treinamento feito a partir de gravações de diversos instrumentos. Contudo, essas técnicas são patenteadas e, por isso, não estão disponíveis informações mais específicas que permita reproduzir seu funcionamento. Ainda assim, este programa foi citado no presente capítulo, pois é uma das principais referências em soluções comerciais.

2.2.2 Neuratron AudioScore Professional®

Esse programa, apresentado em [12], permite que o sinal de entrada monofônico tanto esteja contido em um arquivo (modo *offline*), quanto possa ser capturado a partir de um microfone (modo *online*). É capaz de gerar notação musical a partir de um MIDI intermediário e apresenta interface com programas de edição de áudio como Sibelius®, Finale® e Cubase®.

Suas principais características, no que se refere estritamente à TMA, são as seguintes:

- Faixa de reconhecimento de *pitch* de A0 a C8 (27 Hz a 4186 Hz);
- Resolução temporal de 5,8 ms;
- Resolução freqüencial de 0,3 Hz (equivalente a um centésimo de semitom para A4).

Sua interface gráfica pode ser visualizada na Figura 2.1, onde uma partitura está sendo preenchida à medida que o arquivo de áudio é processado. Na tela superior, uma representação intermediária é usada, enquanto que a notação convencional é empregada na tela inferior.

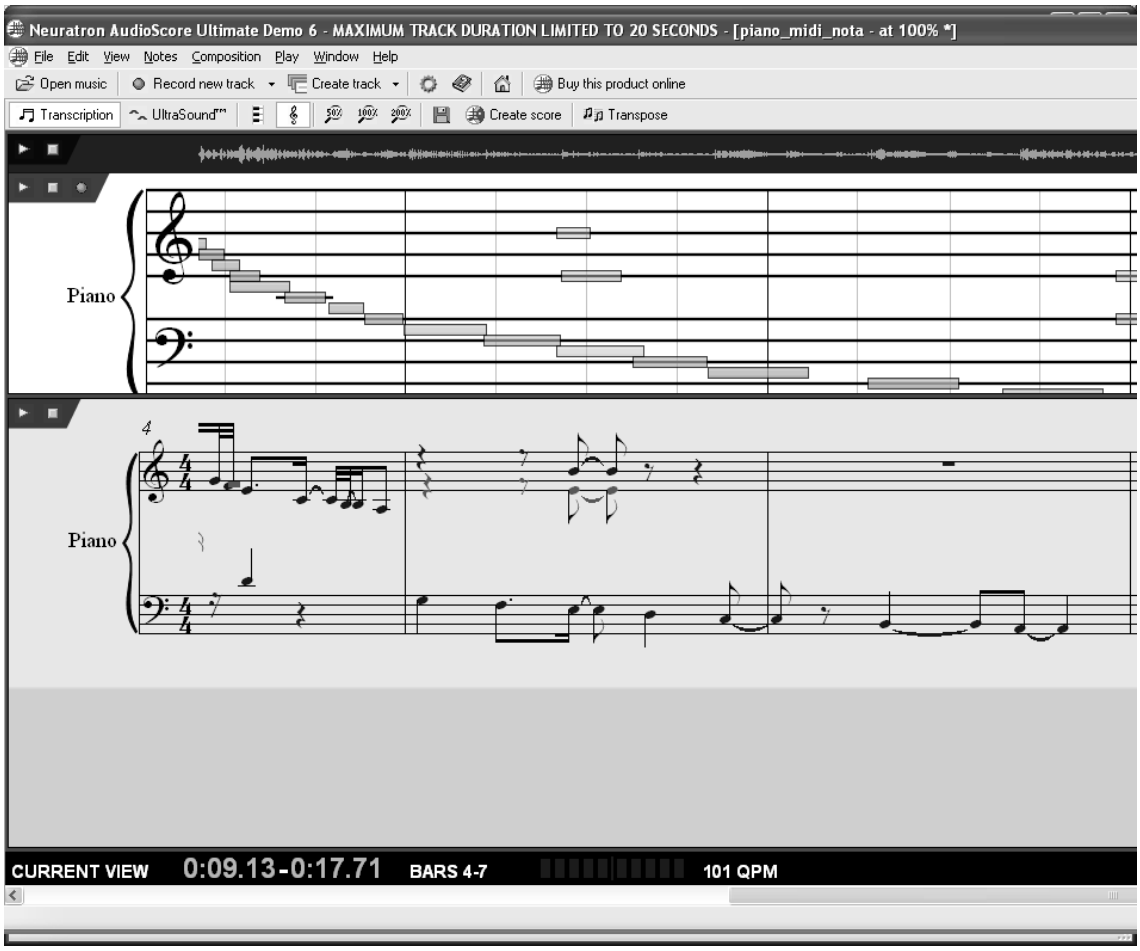


Figura 2.1: Interface gráfica do Neuratron AudioScore Professional®.

2.2.3 Melodyne®

Esse programa é do fabricante Celemony®. Possui inúmeras funcionalidades, ligadas principalmente à analise espectral e à manipulação do resultado de tal análise [13]:

- Possibilita realizar correções, concatenações e edições em geral, sem perder as características de timbre;
- Permite realizar afinações e corrigir ritmos de forma acessível;

- Mantém características dos transitórios, mesmo quando os sons são dilatados no tempo;
- Permite, a partir de um único som, gerar harmonias compostas por diversos sons de modo bastante natural;
- Possui integração com diversos programas comerciais de edição de áudio: Digidesign ProTools®, Cubase® e Nuendo® (ambos da Steinberg), Apple Logic®, Cakewalk Sonar®, Ableton Live®, MotU Digital Performer® e Propellerhead Reason®.

Uma das principais idéias por trás do Melodyne® é sua forma de representação musical. Ao invés de utilizar o formato convencional de forma de onda, cada trecho do sinal apresenta uma determinada posição vertical que varia de acordo com a altura (ver Apêndice A) da nota representada por aquele trecho. Isso pode ser visualizado na Figura 2.2.

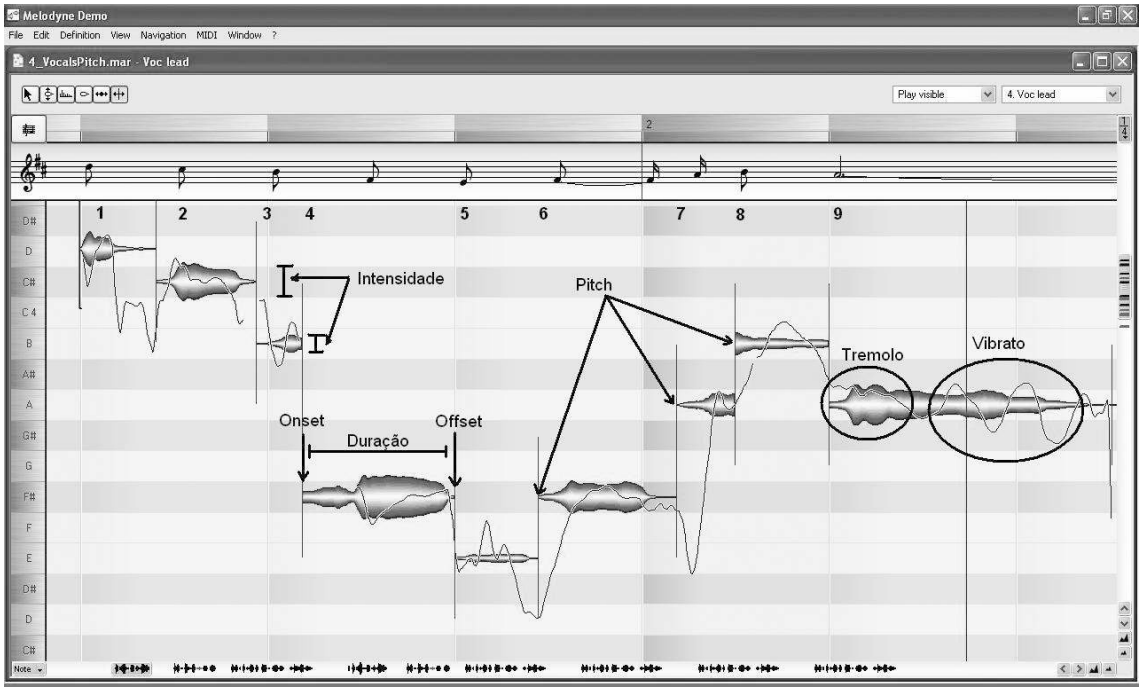


Figura 2.2: Representação de sinal de áudio utilizado pelo Melodyne, para possibilitar a visualização de *pitch*, duração, intensidade, *onset* e *offset*.

Pode-se entender essa figura como sendo a representação temporal de um sinal emitido por um instrumento que toca uma nota que diminui de intensidade continuamente e termina (divisão 1). A seguir começa uma nota mais grave, mais lenta

e de maior intensidade (divisão 2). Logo depois, vem uma nota mais rápida e de menor intensidade (divisão 3), seguida de uma nota grave e lenta com uma mudança de amplitude no meio de sua duração (divisão 4) e de outra mais grave e de intensidade mínima (divisão 5), retornando à anterior (na divisão 6). Por fim, uma nota é tocada na divisão 7, interrompida por outra mais grave na divisão 8, retornando à sua F0 original por um longo período, tendo sua intensidade caindo continuamente até zero (da divisão 9 em diante). A linha contínua que acompanha toda a melodia indica pequenas variações na freqüência. É possível perceber o *tremolo* (oscilação na amplitude da nota) e o *vibrato* (oscilação na freqüência da nota), também indicados na figura.

Pode-se ver a altura média de uma certa nota em relação às demais e a intensidade dessa nota. Arrastando-se cada trecho para cima ou para baixo, sua freqüência fundamental será modificada. Movendo-se sua borda esquerda para a esquerda ou para a direita, seu tempo de *onset* será alterado. Movendo-se sua borda direita para a esquerda ou para a direita, seu tempo de *offset* será alterado. Comprimindo-se ou expandido-se verticalmente a representação de uma dada nota, a intensidade da nota será modificada. Comprimindo-se ou expandido-se horizontalmente a representação, a duração da nota será modificada.

2.3 Principais soluções e valores típicos de desempenho na literatura na área de TMA

Na área da TMA, a avaliação do desempenho dos sistemas é bastante complexa, devido à falta de uma forma padronizada de se realizarem testes. Cada pesquisador tende a idealizar uma série de testes com certas particularidades. Cada autor usa um conjunto diferente de músicas-padrão, sendo que cada uma terá um conjunto diferente de características que representarão maior ou menor dificuldade para um determinado método. Por isso, as comparações diretas entre métodos tendem a ser difíceis ou até mesmo injustas.

Além disso, algumas linhas de pesquisa usam testes que podem ocultar as fraquezas de determinados métodos e distorcer os resultados. Muitas vezes, um método para realizar certa etapa do processo parece obter bons resultados em determinado

trabalho, mas obtém resultados desfavoráveis em outro.

Ao se pesquisarem resultados dos últimos avanços no campo da TMA, deseja-se buscar principalmente informações como percentagem de notas detectadas, percentagem de falso alarme (ou falsos positivos) na detecção de notas, precisão freqüencial e precisão temporal (além de outras figuras de mérito a serem definidas mais adiante no texto).

O sistema de TMA para piano descrito por Bello et al. [14] emprega informações originadas no domínio do tempo ou no domínio da freqüência de forma isolada, traçando comparações entre as duas abordagens. A seguir, utiliza ambos os domínios de forma conjunta. Este sistema considera os segmentos do sinal de entrada como uma soma linearmente ponderada de formas-de-onda contidas em uma base de dados de notas individuais. Assim, através da estimação das parcelas dessa soma, as notas seriam identificadas. A janela de análise apresenta comprimento de 200 ms, contudo o salto entre janelas consecutivas é de 10 ms. Para o método utilizando domínio da freqüência, a acurácia[1] do sistema está em torno de 70%, enquanto que, no domínio do tempo, esse valor sobe para 77%. A taxa de falso alarme em relação às notas varia entre 16% no domínio da freqüência e 22% no domínio do tempo.

O sistema de TMA para piano apresentado por Dixon [15] faz uso da transformada de Fourier de tempo curto (para a análise espectral), seleção adaptativa de picos no espectro (para identificar componentes freqüenciais) e rastreamento destes picos ao longo do tempo. Além disso, regras são usadas para evitar a detecção de freqüências indesejadas, como eliminar freqüências associadas a notas de duração muito curta. O autor utilizou uma margem de erro de 70 ms para a detecção do início de cada nota, acima da qual considera-se um início de nota mal identificado. A maioria dos trabalhos ligados exclusivamente à detecção de início de nota, como [16] e [17], em contrapartida, considera apenas 50 ms. A acurácia obtida foi de 70% a 80% e foi apontada como fraqueza do método de Dixon a falta de robustez em relação ao timbre do instrumento sob análise.

O trabalho de Ryynänen e Klapuri [18] trata o problema de transcrição de sinais musicais polifônicos através da modelagem de eventos que representam notas. Esses

[1] Acurácia pode ser entendida como a razão entre as notas corretamente identificadas e o total de notas transcritas pelo sistema em questão. Esta figura de mérito será detalhada no Capítulo 7.

eventos são descritos com modelos escondidos de Markov (*hidden Markov models* ou HMM). O modelo utiliza três *features* extraídos a partir de um estimador de múltiplas freqüências fundamentais. Daí é possível calcular uma medida de similaridade para diferentes notas e realizar a segmentação temporal do sinal de entrada com base nestas notas. As transições entre as notas são controladas a partir de modelos que têm por base a estimativa do tom (ver Apêndice A). A transcrição final é completada seguindo-se os caminhos apontados pelos modelos calculados. Esse sistema obteve uma acurácia de 41%. Tal resultado não é tão ruim quanto pode parecer. Se for considerado um número de, por exemplo, 12 notas passíveis de serem executadas dentro de 5 oitavas, isso significa que são 60 possibilidades de notas a serem identificadas. O resultado apresentado pelo método de Klapuri indica que, em um universo de 100 notas, 41 delas foram identificadas corretamente como sendo uma dentre as 60 possíveis.

Poucos trabalhos tratam o tema de forma integral, mas há muitas publicações sobre cada etapa do processo de TMA. Como referências para etapas do processo de TMA, pode-se citar alguns trabalhos sobre detecção de *onsets* e outros sobre detecção e identificação de F0s. No primeiro grupo, destacam-se os trabalhos de S. Dixon [19], de A. Lacoste e D. Eck [20], e também o de Yu Shiu et al [21]. No segundo grupo, destacam-se os trabalhos de A. Klapuri [22], de M. Davy et al. [23], de M. Ryynänen et al. [24], de G. Poliner et al. [25] e de A. Camacho et al [26].

S. Dixon [27] revisita alguns algoritmos clássicos da literatura para detecção de *onsets*, como aqueles que atuam sobre magnitude, fase e representação no domínio complexo, chegando-se a resultados contraditórios com outros experimentos contidos na literatura [16],[17]. Em [27], foram propostas mudanças de forma a elaborar novos métodos, como o uso de função de desvio de fase ponderada e a diferença complexa de onda retificada, que alcançaram melhores resultados. A. Lacoste [20] apresenta uma nova abordagem para detecção de *onsets* através de um algoritmo de aprendizado supervisionado que classifica os *frames* de um sinal de áudio digital como *onsets* ou *non-onsets*. Isso é feito através de uma simples rede neural ou através de um esquema de predição baseado em uma série de redes neurais treinadas com diferentes parâmetros.

A pesquisa ligada especificamente à detecção de início de nota apresenta resulta-

dos bastante discrepantes, ainda que para o mesmo método, mostrando que a forma na qual os experimentos são montados influencia determinantemente os resultados obtidos. Contudo, de uma forma geral, pode-se dizer que os melhores resultados apresentam cerca de 85% de taxa de detecção de início de nota e de 5 a 10% para a taxa de falso alarme correspondente [16].

M. Davy et al. [23] apresentam uma abordagem para detecção de F0s via representação de Gabor de sinais não-estacionários juntamente com um algoritmo baseado em MCMC (Markov-Chain Monte Carlo). Uma estrutura bayesiana é empregada para representar a informação *a priori* dos parâmetros das notas de forma a possibilitar a inferência de *pitch* e outros parâmetros das formas de onda. M. Ryynänen et al. [24] mostram um método para transcrição de linha de baixo [2] que usa um estimador de F0s seguido por modelos musicológicos. Esses modelos estimam o tom e, através de uma cadeia de Markov, determinam as probabilidades de transição entre as notas. Foi obtida uma acurácia em torno de 60%. G. Poliner et al. [25] apresentaram um modelo discriminativo para transcrição de sinais polifônicos de piano. SVMs (*support vector machines*) treinadas sobre parâmetros espectrais são usadas para classificar as instâncias das notas *frame*-a-*frame* e a saída do classificador é temporalmente controlada por HMM. Foi obtida uma acurácia de 68%. A. Klapuri [28] descreve um método para estimar múltiplas F0s consistindo de um modelo computacional do sistema auditivo humano seguido de um mecanismo de análise de periodicidade onde F0s são iterativamente identificadas e canceladas do sinal. A. Camacho et al. [26] desenvolveram um estimador de *pitch* inspirado em formas de onda dente-de-serra (SWIPE). Essa abordagem estima o *pitch* como a F0 da forma de onda dente-de-serra cujo espectro melhor combinar com o espectro do sinal de entrada. A comparação dos espectros é feita calculando-se o produto interno normalizado entre o espectro do sinal e o do cosseno modificado. Como base de comparação, deve-se citar que Klapuri [7], em sua tese de doutorado (que aqui será a base para a etapa de detecção de freqüências fundamentais), utiliza janela de análise de 190 ms. Esse valor, apesar de parecer incompatível com a não-estacionariedade do sinal de música, é necessário para a análise adequada dos sinais

[2] Linha de baixo é a seqüência de notas mais graves que representam a harmonia da composição ao longo do tempo.

de baixa freqüência.

2.4 Conclusão

Neste capítulo, três programas comerciais foram descritos, permitindo conhecer o que constitui avanços recentes neste tema. Dos três, o Melodyne® é o mais conhecido e utilizado. Apresenta resultados bastante precisos em relação à detecção de freqüências fundamentais, apesar de permitir a detecção de apenas uma nota por vez (pelo menos para a versão testada). Por isso, para se estabelecer uma referência externa, este programa servirá como base de comparação para o desempenho do sistema a ser desenvolvido no presente trabalho.

Além disso, foram apresentadas algumas das principais soluções em sistemas de TMA presentes na literatura e seus valores de desempenho. Para estes sistemas, a acurácia é, na média, de cerca de 70%, enquanto que a taxa de falso alarme gira em torno de 15%. Em relação à detecção de início de nota, a maior parte dos métodos presentes na literatura conseguem algo em torno de 85%, com falso alarme de 5 a 10%.

Capítulo 3

Detecção de *onsets*

3.1 Introdução

A detecção de *onsets* é uma etapa fundamental para se realizar a transcrição musical automática e tem como objetivo a marcação dos instantes onde há o início da execução de uma nota. Pode ser vista como a primeira etapa dentro da estrutura do sistema de TMA a utilizar os resultados da análise espectral. A Seção 3.2 apresenta algumas definições importantes para uma melhor compreensão do presente capítulo. A Seção 3.3 descreve métodos de detecção de *onsets* comuns na literatura. Por serem métodos convencionais, fazem uso da FFT para realizar a análise espectral. A Seção 3.4 discute os resultados obtidos com tais métodos e a Seção 3.5 conclui o capítulo.

3.2 Definições

O modelo de execução de uma nota musical engloba, genericamente, as seguintes etapas, como ilustrado na Figura 3.1:

- Ataque: se a nota em questão estiver em meio a silêncio, pode-se dizer que o ataque é o intervalo de tempo durante o qual a amplitude da envoltória aumenta.

- Transitório ou decaimento: é o período de tempo que vai do ataque até a região onde o sinal se desenvolve de uma forma relativamente imprevisível.

Como tudo isso ocorre de forma muito localizada no tempo, pelo princípio da incerteza, o transitório tende a apresentar uma banda muito larga no domínio da freqüência com destaque para a região das altas freqüências. Segundo [29], só é possível para o ouvido humano distinguir transitórios separados por pelo menos 10 ms.

- Sustentação: é o período onde a amplitude da nota se mantém aproximadamente constante. Nem sempre é encontrada naturalmente durante a execução de uma nota musical. Para que a sustentação ocorra, o músico deve interferir no modo de vibração do instrumento, excitando-o de modo persistente, de forma a manter a nota por um certo período de tempo. Geralmente apresenta um grau maior de previsibilidade comparado ao do transitório.

- Relaxamento: é a parte final da execução da nota, quando a amplitude da vibração sofre desvanecimento.

Define-se por *onset* o momento de início de toda a dinâmica envolvendo a execução de uma nota musical. Esse conceito fica bem claro quando se trata de uma nota precedida por silêncio, mas torna-se extremamente nebuloso quando ela é precedida por uma outra nota sem haver pausa. Neste caso, não se sabe ao certo quando uma nota parou e a nota seguinte começou.

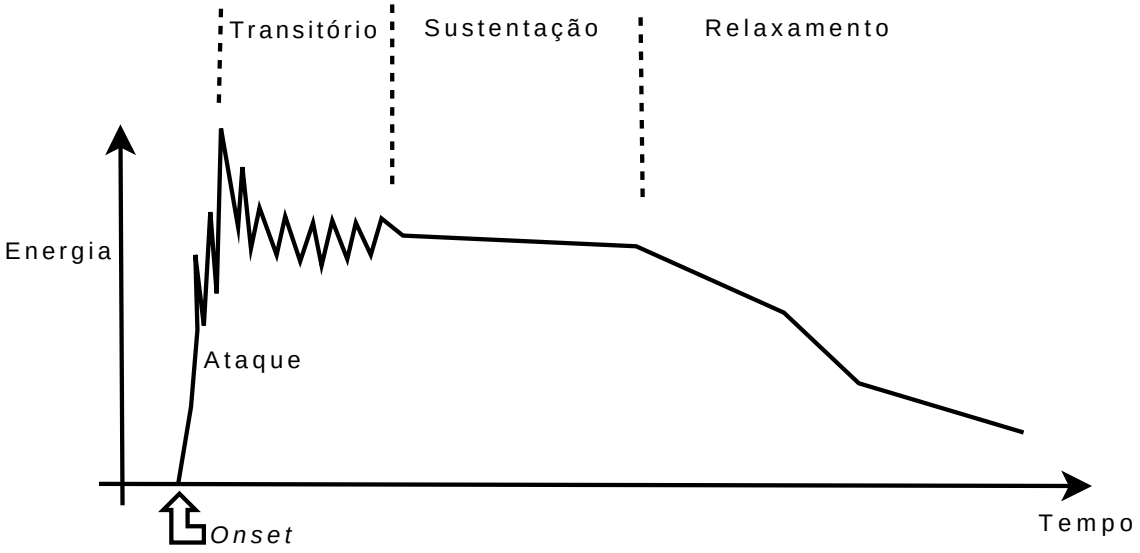


Figura 3.1: Esquema com as etapas da execução de uma nota musical.

3.2.1 Etapas para detecção de *onsets*

De uma forma geral os métodos para detecção de *onset* são organizados em três etapas: pré-processamento do sinal de entrada, definição de uma função de detecção e, por fim, a seleção dos picos da função de detecção, representando os instantes em que ela sugere a ocorrência dos *onsets*.

A etapa de pré-processamento modifica o sinal de entrada, enfatizando determinadas características suas para facilitar (ou até mesmo possibilitar) as etapas subseqüentes. Existem basicamente duas formas de pré-processar um sinal para fins de detecção de *onsets*: divisão em sub-bandas e separação transitório/regime permanente [16].

A determinação da função de detecção também é chamada de redução. A função de detecção pode ser entendida como uma função altamente sub-amostrada (pois pode ser definida apenas em alguns instantes) que traz informações quanto à ocorrência de transitórios no sinal original. Essa informação é exibida geralmente sob a forma de picos cuja localização temporal coincide com a ocorrência de um transitório no sinal sendo processado. Essa é a principal etapa de uma série de métodos de detecção de *onsets* e será tratada com mais detalhes na Seção 3.3.

Após a redução, é necessário selecionar os picos da função de detecção que representam *onsets* reais. Isso é necessário pelo fato de outros fenômenos serem capazes de gerar falsos picos na função de detecção. Entre estes fenômenos estão reverberação, ecos etc. Para isso, alguns métodos foram desenvolvidos para se determinar um limiar abaixo do qual qualquer pico é descartado. Existem basicamente duas formas de se determinar este limiar: através de um valor fixo ou de um valor dinâmico. O limiar fixo é constante para toda a extensão do sinal de música, o que não permite acompanhar as variações que a não estacionariedade do sinal provoca sobre o comportamento da função de detecção ao longo do tempo. Já o limiar dinâmico busca se adaptar a essas variações.

O limiar dinâmico pode ser obtido por meio de um filtro passa-baixas que opera sobre o sinal musical, possivelmente implementado através de uma média móvel. Contudo, tal método é passível de distorções devido a *out-liers* ou ruído. Uma solução para isso é algum tipo de processamento não-linear sobre o sinal musical, como o filtro de mediana ao invés da média, que pode fornecer um limiar $\tilde{\delta}$ na

seguinte forma:

$$\tilde{\delta}(n) = \delta + \lambda.\text{mediana}\{|d(n-M)|, ..., |d(n+M)|\}, \tag{3.1}$$

onde n é o índice que indica instantes discretos relacionados à resolução temporal da função de detecção, δ e λ são parâmetros para determinação do limiar, d é a função de detecção e $(2M+1)$ é o número de amostras da função d que entram no cálculo da mediana. Estimativas para os valores de δ e λ são feitas de forma empírica. É claro que o índice de d não pode ser negativo, e as situações onde isso acontece devem ser ignoradas.

Algum tipo de pós-processamento sobre $\tilde{\delta}$ também é possível, como filtragem passa-baixas, normalização ou a restrição de que apenas um pico seja selecionado dentro de intervalo de tempo T. O valor de T geralmente é definido como 50 ms [16]. Com isso, tem-se uma redução no número de picos espúrios.

3.3 Métodos implementados

Quatro métodos baseados em representações no domínio da freqüência (calculadas a partir da FFT) foram implementados e testados para a determinação de *onsets* em sinais musicais. O uso do domínio espectral permitirá o desenvolvimento de um novo método de determinação de *onsets* baseado na ferramenta BQFFB (ver Seção 6.2.2), cujo desempenho será comparado aos dos métodos apresentados a seguir.

3.3.1 HFC

O HFC [16], da sigla em inglês *high-frequency content*, faz uso do módulo do espectro do sinal de entrada. Ao se pensar em algum meio de se extrair informações do módulo do espectro, o mais intuitivo seria medir simplesmente a energia $E(n)$ do sinal ao longo do tempo. Se o nível de energia passasse de um certo limiar pré-estabelecido, isso indicaria o início de uma nota em um dado instante. Essa energia poderia ser descrita por

$$E(n) = \frac{1}{N} \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} \left[x(n+m)\right]^2 w(m), \tag{3.2}$$

onde $x(n)$ é o sinal musical de entrada, $w(m)$ é uma função para janelar o sinal no domínio do tempo e N é o número de amostras de $w(m)$. Novamente, devem-se ignorar as situações onde $x(n)$ tiver índice n negativo.

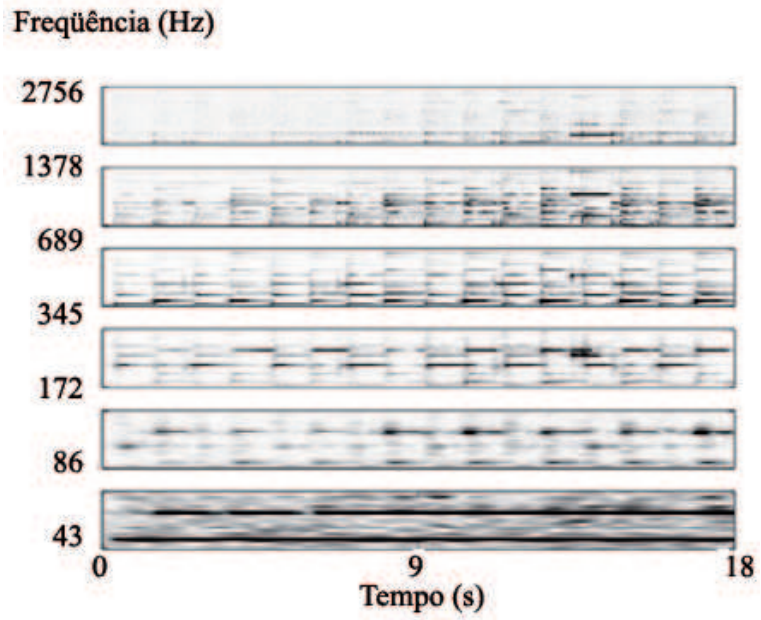


Figura 3.2: Espectro do som de um piano ao longo do tempo. Quanto mais alta é a freqüência central da banda, mais bem definidas são as manchas verticais que representam os inícios das notas.

A Figura 3.2 mostra a representação de um sinal musical típico no domínio da freqüência. É possível ver o espectrograma ao longo do tempo para 6 oitavas, cujos limites são indicados no eixo vertical. Em cada oitava, percebe-se uma série de manchas verticais. Essas manchas representam um conjunto de picos de energia e são formadas porque estes picos estão ocorrendo no mesmo instante de tempo para todos os canais de uma determinada oitava. Ao se analisar o conjunto de todas as oitavas mostradas (com exceção da última), verifica-se que essas manchas verticais estão mais nítidas (mais definidas no eixo do tempo) à medida que se aumentam as freqüências-limite das oitavas, isto é, a informação relativa às altas freqüências é mais útil para a determinação precisa dos tempos de *onset* [16]. Por isto, é interessante enfatizar a energia dos canais relativos às altas freqüências, o que pode ser feito através da inserção do vetor de pesos $W = [w_1, w_2, ..., w_k]$, na Eq. (3.2), gerando a

função de energia modificada:

$$\tilde{E}(n) = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} w_k \left| X_k(n) \right|^2 , \tag{3.3}$$

onde $X_k(n)$ representa o canal k da transformada de Fourier de tempo curto (*short-time Fourier transform* ou STFT) do sinal de entrada analisado pela janela $w(n)$ de comprimento N com avanço unitário. O tamanha da DFT é idêntico ao de um bloco do sinal de entrada. Os elementos de W podem assumir os seguintes valores, para $k = -\frac{N}{2}, ..., \frac{N}{2} - 1$:

- $w_k = 1$, o que é equivalente ao método da energia, dado pela Eq.(3.2).
- $w_k = k$ [30], ou seja, o valor de cada peso varia linearmente com o índice.
- $w_k = k^p$, ou seja, o valor de cada peso varia não-linearmente (quadraticamente se $p = 2$) com o índice.

Essas ponderações resultam em diferentes perfis para a função de detecção. O método de HFC, segundo [16], produz uma função de detecção com picos nítidos durante os transitórios e é bem-sucedido ao analisar *onsets* produzidos por instrumentos percussivos.

O algoritmo HFC é baseado na representação espectral de curta duração, o que o faz omitir qualquer consideração em relação à evolução temporal do sinal. Isso se deve ao fato de que este método atua apenas sobre um dado *frame* do sinal de forma independente. Por isso, outros métodos constroem suas funções de detecção com base na diferença entre conteúdos espectrais de um *frame* para o outro, como os que serão tratados nas próximas sub-seções.

Quanto à complexidade do método HFC, pode ser feita a seguinte análise para um espectro de N *bins*:

- Número de adições por *frame*: $N - 1$;
- Número de multiplicações por *frame*: $2N + 1$.

É importante destacar que o avanço considerado para este cálculo é unitário e não se inclui aqui o cálculo da DFT.

3.3.2 Distância espectral

Esse método atua sobre o módulo do espectro do sinal de entrada e se baseia no fato de que, quando se inicia a execução de uma nova nota, a distribuição da energia ao longo dos canais é significantemente alterada. Em contrapartida, quando uma nota ou uma pausa são sustentadas, essa distribuição se mantém.

Assim, considera-se a distribuição de energia ao longo dos canais em um determinado *frame* do sinal de entrada como as coordenadas de um ponto em um espaço N-vetorial, onde N é o número de canais. O método da distância espectral analisa a variação entre as distribuições dos *frames* nos instantes $(n-1)$ e n. Em [30], Masri utilizou a norma 1 da diferença entre estes vetores, enquanto que em [31], Duxbury sugeriu o uso da norma 2, chegando à seguinte formulação:

$$\mathrm{SD}(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \left\{ H\left(|X_k(n)| - |X_k(n-1)| \right) \right\}^2, \tag{3.4}$$

onde $H(x) = (x + |x|)/2$, ou seja, é zero para argumentos negativos.

Segundo [32], o algoritmo da distância espectral é rápido e de fácil implementação, contudo sua eficácia diminui quando aplicado a sinais não-percussivos e quando transitórios de energia se sobrepõem em misturas mais complexas.

Quanto à complexidade desse método, pode ser feita a seguinte análise para um espectro de N bandas (sem considerar o cálculo da DFT):

- Número de adições por *frame*: $3N - 1$;
- Número de multiplicações por *frame*: $2N$.

3.3.3 Variação no desvio de fase

Esse método atua sobre a fase do espectro do sinal de entrada e se baseia no fato de a freqüência ser a derivada da fase. Quando se inicia a execução de uma nova nota, a freqüência tende a se alterar, principalmente durante o transitório. Para um canal de índice k, a freqüência em um determinado *frame* n pode ser estimada a partir das fases φ_k nos *frames* $(n-1)$ e n da seguinte maneira [33]:

$$f_k(n) = \left(\frac{\varphi_k(n) - \varphi_k(n-1)}{2\pi h} \right) f_\mathrm{S}, \tag{3.5}$$

onde f_S é a freqüência de amostragem do sinal sob análise e h é o número de amostras que constituem o salto entre *frames* adjacentes. O operador $\Delta X(n)$ denota a diferença entre alguma grandeza X no instante n e esta mesma grandeza X no instante $n-1$.

Pode-se, então, calcular a diferença entre as freqüências aproximadas de dois *frames* consecutivos a partir da variação do desvio de fase [16]:

$$\begin{aligned} \Delta(\Delta\varphi_k(n)) &= \Delta\varphi_k(n) - \Delta\varphi_k(n-1) \\ &= \varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2). \end{aligned} \tag{3.6}$$

Dentro da região do transitório, a freqüência instantânea é geralmente indefinida e, por isso, $\Delta(\Delta\varphi_k(n))$ tende a ser elevada.

O método de variação no desvio de fase usa $\Delta(\Delta\varphi_k(n))$, calculada para todos os canais k dos *frames* do sinal de entrada, como base para obter a função de detecção. Quando $\Delta(\Delta\varphi_k(n))$ estiver com uma variação muito baixa em torno de zero, pode-se considerar que se está diante de um regime permanente. Caso contrário, pode-se dizer que se trata de um *onset*. A análise da variação do desvio de fase pode ser feita tanto através da curtose, a exemplo do que foi feito em [33], quanto através da média do módulo da soma das variações, assim como foi empregado em [34].

De acordo com [32], a detecção de *onsets* baseada em fase oferece uma alternativa aos métodos que atuam na energia do sinal para a detecção de *onsets* suaves. Entretanto, métodos que atuam sobre a fase são mais suscetíveis à distorção de fase e a variações introduzidas pela fase de componentes ruidosos (geralmente relacionados com valores de baixa energia). Este método não é indicado para sinais com múltiplas componentes, pois só estima uma freqüência média.

Quanto à complexidade desse método, pode ser feita a seguinte análise para um espectro de N *bins* (não incluindo o cálculo da DFT e da fase):

- Número de adições por *frame*: $3N - 1$;
- Número de multiplicações por *frame*: $N + 1$.

3.3.4 Domínio complexo

Esse método se compõe de dois outros métodos: um que atua sobre o módulo e outro que atua sobre a fase do sinal sob análise; daí dizer-se que opera sobre o

domínio complexo. A razão para se tentar tal combinação é contornar os pontos fracos e potencializar os pontos fortes de cada abordagem. Enquanto os baseados em energia (módulo) são eficazes para sons percussivos com início de nota bem definido, aqueles que se baseiam na fase são mais favorecidos por *onsets* suaves onde não existe muito contraste de energia. Além disso, segundo [32], cada um desses métodos (o que atua sobre o módulo e o que atua sobre a fase) é mais confiável em extremidades opostas do eixo das freqüências.

Para trechos de regime permanente, pode-se assumir que os valores de freqüência e amplitude se mantêm aproximadamente constantes. Dessa forma, pode-se considerar, no plano complexo, que o valor relativo ao *frame* atual é uma estimativa para o valor relativo ao próximo *frame*, sendo aquele valor referenciado como valor-alvo enquanto este é o valor medido. Se há uma diferença (maior que determinado limiar) entre o valor-alvo e o valor medido, há um indício de que existe algo de novo no sinal, o que pode ser interpretado como um *onset*.

Assume-se que, na forma polar, o valor-alvo para o canal k de uma STFT é dado por

$$\hat{S}_k(m) = \hat{R}_k(m)e^{j\hat{\phi}_k(m)}, \tag{3.7}$$

onde a amplitude-alvo $\hat{R}_k(m)$ corresponde à magnitude do *frame* anterior $|S_k(m-1)|$. A fase-alvo $\hat{\phi}_k(m)$ é dada pela soma da fase do *frame* anterior com a diferença de fase entre o *frame* atual, de índice m, e o anterior, de índice $m-1$:

$$\begin{aligned} \hat{\phi}_k(m) &= \text{princarg}\left[\tilde{\varphi}_k(m-1) + (\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2))\right] && (3.8) \\ &= \text{princarg}\left[2\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)\right], && (3.9) \end{aligned}$$

onde o operador princarg[•] se refere ao argumento principal da fase em questão.

Sendo o valor medido $S_k(m) = R_k(m)e^{j\phi_k(m)}$, pode-se calcular a distância euclidiana entre o valor alvo e o valor medido no plano complexo, obtendo-se uma estimativa da estacionariedade do canal de índice k através de

$$\Gamma_k(m) = \left\{ \left[\Re(\hat{S}_k(m)) - \Re(S_k(m))\right]^2 + \left[\Im(\hat{S}_k(m)) - \Im(S_k(m))\right]^2 \right\}^{1/2}, \tag{3.10}$$

onde os operadores $\Re$ e $\Im$ se referem, respectivamente, à parte real e à parte imaginária do número complexo em questão.

Acumulando essas medidas de estacionariedade ao longo de todos os valores de k, pode-se construir uma função de detecção da seguinte forma:

$$\eta(m) = \sum_{k=1}^{N} \Gamma_k(m). \tag{3.11}$$

Os valores de η que são maiores que um dado limiar indicam um instante onde existe uma grande probabilidade de corresponder ao início da execução de uma nota.

Quanto à complexidade desse método, podem ser feitas as seguintes considerações para um espectro de N *bins*:

- Número de adições por *frame*: $6N - 2$;
- Número de multiplicações por *frame*: $3N + 1$.

3.4 Resultados

Os experimentos computacionais ilustrando o funcionamento de cada método anteriormente descrito para detecção de *onsets* foram realizados em duas etapas: a primeira com sinais sintéticos e a segunda um sinal real, ou seja, com uma gravação da execução de uma peça musical por um musicista em um instrumento acústico. Para se considerar a detecção de um *onset* como correta, usou-se uma tolerância de 50 ms, conforme é encontrado na maioria das referências, como [16]. Seguindo esta mesma referência, os saltos entre os *frames* foram de 128 amostras (2,9 ms a uma taxa de amostragem de 44100 Hz) e a janela, de 190 ms.

3.4.1 Experimentos com sinais sintéticos

Os sinais empregados nesse tipo de experimento foram gerados a partir do protocolo MIDI (ver Apêndice B). Na verdade são sinais do tipo WAV gerados a partir de sintetizadores (*sound-fonts*[1]). Assim, tem-se o total controle do que está presente no sinal sob análise, além de se ter conhecimento preciso do instante de início de cada nota executada, como sugerido em [17], [35].

[1] *Sound-fonts* são arquivos que contêm formas de onda gravadas que podem ser ressintetizadas em diversos *pitches* (ver Apêndice A) e níveis de dinâmica. Sua qualidade, que é diretamente proporcional à qualidade do áudio previamente digitalizado, é superior à de uma síntese, mas inferior à de um som gravado.

O primeiro teste foi feito utilizando-se um sinal que apresenta a gravação de um piano em que se toca apenas uma nota por vez. A seqüência musical correspondente, que pode ser visualizada na Figura 3.3, foi planejada de modo que as durações das notas fossem cada vez menores até o meio do comprimento do sinal; do meio para o final, as durações são progressivamente maiores. Contrariamente, as freqüências fundamentais são cada vez maiores até a metade do comprimento do sinal enquanto que, do meio para o fim, são progressivamente menores.

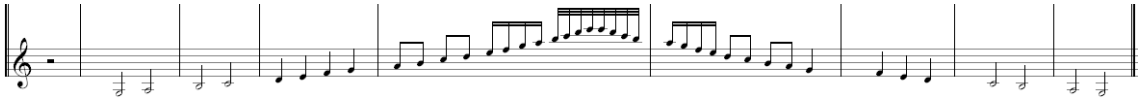


Figura 3.3: Partitura do sinal musical sintético que representa uma composição em que se executa uma nota por vez.

Para analisar o efeito da variação dos parâmetros, primeiramente fixou-se o valor de λ em 1 e variou-se o valor de δ. Verificou-se, então, qual foi o valor de δ cujo ponto na curva ROC (*Receiver Operating Characteristic*) [36] apresentava a menor distância euclidiana para o ponto ótimo (100% de detecção e 0% de falso alarme). Adotou-se este valor de δ como sendo ótimo e variou-se o valor de λ. Com esse procedimento seqüencial, formaram-se as curvas ROC da Figura 3.4 para cada um dos métodos para detecção de *onsets* descritos anteriormente quando aplicados ao sinal de teste. Pode-se ter uma noção global da eficácia de cada método e inferir que, para este caso, o método HFC é o que apresenta pior desempenho.

Para baixas taxas de falso alarme (abaixo de 30%), o HFC apresenta baixas taxas de detecção. Para valores de falso alarme acima de 30%, há uma elevação brusca da taxa de detecção, mas isso implica que qualquer valor para taxa de detecção que faça frente às obtidas pelos demais métodos já começará com pelo menos 30% de falso alarme. Isso acontece porque, para o sinal de teste em questão, uma grande parte da energia está nas baixas freqüências. Assim, quando se dá ponderação maior às altas freqüências, alguns picos da função de detecção são perdidos. O método da distância espectral, no entanto, apresentou resultados razoáveis. O método que trata a variação no desvio da fase não apresentou um bom desempenho, pois é mais apropriado para sinais que representem sons não-percussivos, como aqueles que podem ser produzidos por instrumentos de arco [33]. Ainda assim, mesmo que

não tenha obtido uma taxa de detecção elevada, apresentou taxas de falso alarme bastante reduzidas. O método que opera no domínio complexo obteve o melhor resultado dentre os testados, pois faz uso da informação de fase juntamente com a informação de módulo.

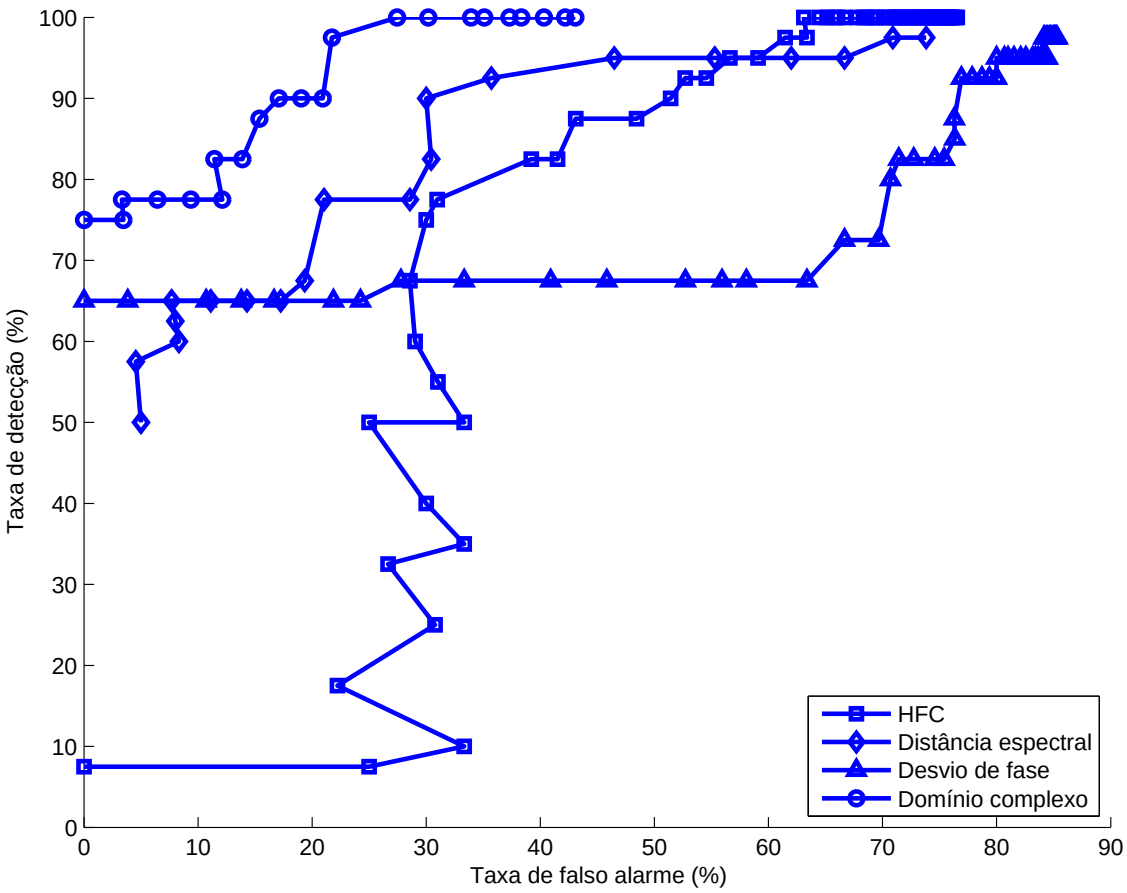


Figura 3.4: Curvas ROC referentes a cada método de detecção de *onsets* aplicados ao sinal musical sintético que representa uma composição em que se executa uma nota por vez. Os parâmetros variáveis usados para se obterem as curvas ROC foram λ e δ, contidos na Eq. (3.1).

O ponto da curva que fica mais próximo do canto superior esquerdo do gráfico, ou seja, com a menor distância euclidiana em relação ao ponto ótimo (100% de detecção e 0% de falso alarme) representa um bom compromisso entre a taxa de falso alarme e a taxa de detecção; deve-se notar, entretanto, que dessa forma se dá a mesma relevância às duas variáveis envolvidas: taxa de detecção e de falso alarme. Buscando-se um dos valores para δ e λ que definem o limiar (ver Eq. (3.1)) para atender essa restrição, foi possível montar a Tabela 3.1. Os valores numéricos

apresentados nessa tabela admitem uma margem de perda de *onsets* sem apresentar uma alta taxa de falso alarme.

Tabela 3.1: Resultados para métodos de detecção de *onsets*, relativos ao ponto da curva ROC que guarda menor distância euclidiana para o ponto ótimo, para o sinal sintético que representa uma composição em que se executa uma nota por vez.

Método	Taxa de detecção	Taxa de falso alarme	δ	λ
HFC	77,5%	30,9%	0,028	0,98
Distância espectral	77,5%	21,1%	0,005	1,00
Desvio de fase	65,0%	0,0%	0,057	0,98
Domínio complexo	90,0%	17,1%	0,010	0,86

O valor de δ varia bastante entre os métodos enquanto que λ é mais constante e igual à unidade. Isso mostra que δ tem mais influência sobre a função de detecção do que λ. O método propriamente dito dá a forma da função de detecção e essa forma é pouco alterada pelo fator multiplicador λ. No entanto, o que faz diferença é o quanto essa função será deslocada verticalmente de modo a selecionar os picos com maior probabilidade de representarem *offsets* reais. No fim das contas, é necessário unir os benefícios de uma forma apropriada com um deslocamento adequado.

O segundo teste foi feito utilizando-se a mesma idéia rítmica do primeiro exemplo. Contudo, desta vez, são executados ao piano acordes no lugar de notas simples. A partitura desse sinal pode ser visualizada na Figura 3.5.

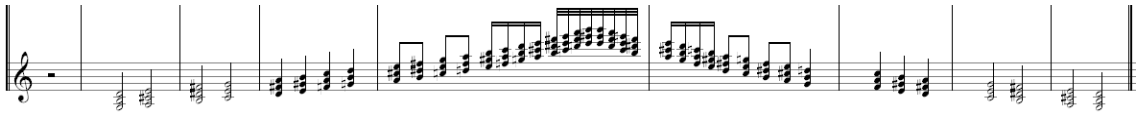


Figura 3.5: Partitura do sinal musical sintético que representa uma composição em que se executa um acorde por vez.

Seguindo a mesma metodologia que para o caso em que a composição apresenta uma nota por vez, a Figura 3.6 mostra as curvas ROC relativas a cada um dos métodos de detecção de *onsets*. A grande diferença dessas curvas para as curvas referentes ao experimento anterior é o desempenho do método que opera sobre a variação da fase. Ele apresentou uma curva bastante próxima do ponto ótimo do gráfico (canto superior esquerdo), no qual há a minimização da taxa de falso alarme

e a maximização da taxa de detecção. Isso pode ser explicado pelo fato de este método associar variações no desvio da fase como pico da função de detecção. Como existem vários sons simultâneos, existe uma chance maior de haver grandes desvios nas variações da fase de cada canal. Ainda assim, não superou o resultado do método que atua no domínio complexo.

Nesse experimento, o método da distância espectral apresentou o pior desempenho pois, neste caso, não há boa resolução freqüencial da ferramenta de análise espectral e o conteúdo de cada canal sofre muita interferência dos canais adjacentes. Isso deve ocorrer principalmente nos canais de baixa freqüência central, pois as notas referentes a esses canais são muitos mais próximas e os canais apresentam bandas muito mais largas do que seria o ideal.

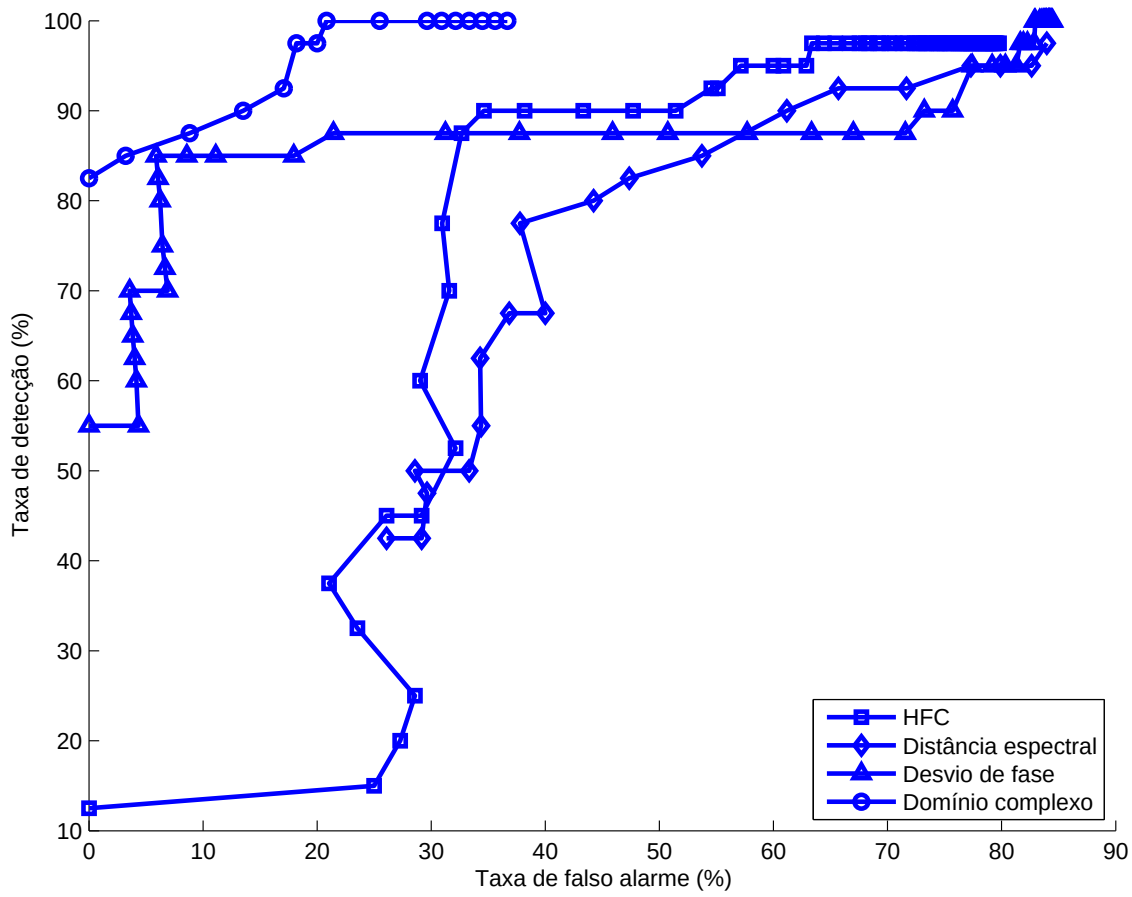


Figura 3.6: Curvas ROC referentes a cada método de detecção de *onsets* aplicados ao sinal musical sintético que representa uma composição em que se executa um acorde por vez. Os parâmetros variáveis usados para se obterem as curvas ROC foram λ e δ, contidos na Eq. (3.1).

Também em relação ao segundo experimento, a Tabela 3.2 apresenta, para cada um dos métodos testados, os melhores casos para as taxas de detecção e de falso alarme como função dos parâmetros δ e λ.

Tabela 3.2: Resultados para métodos de detecção de *onsets*, relativos ao ponto da curva ROC que guarda menor distância euclidiana para o ponto ótimo, para sinal sintético que representa uma composição em que se executa um acorde por vez.

Método	Taxa de detecção	Taxa de falso alarme	δ	λ
HFC	87,5%	32,7%	0,023	1,00
Distância espectral	77,5%	37,8%	0,008	1,00
Desvio de fase	85,0%	5,9%	0,037	1,03
Domínio complexo	87,5%	8,8%	0,010	0,89

3.4.2 Experimentos com sinal real

Para os testes com sinal real, foram usados os primeiros 25 segundos da Ária das Variações Goldberg para teclado de Bach, BWV 988, executada ao piano por Glenn Gould (gravação de estúdio de 1955).

O sinal real visa a atuar como conjunto de teste enquanto que os testes com sinais sintéticos servem para treinamento do sistema, com o objetivo de determinar o limiar de detecção para o algoritmo de seleção de picos. Para o caso do sinal real, o limiar utilizado foi o calculado através do sinal que contém acordes, pois não se sabe *a priori* o número de sons simultâneos.

A Tabela 3.3 mostra o resultado de cada um dos métodos descritos anteriormente quando aplicados ao sinal real. Este sinal precisou ser analisado auditivamente de forma bem lenta para que fossem registrados com precisão os instantes reais dos *onsets* para posterior verificação da qualidade da detecção. Novamente, foi usada uma tolerância de 50 ms para se determinar que a detecção foi correta. Por se tratar de um sinal real que, diferentemente do sintético, não está imune à presença de efeitos como ecos e reverberações, há uma tendência geral de se encontrar valores menores para a taxa de detecção e valores maiores para a taxa de falso alarme, o que é comprovado na prática.

O que se destaca na Tabela 3.3 é o fato de o método que trabalha com desvio

Tabela 3.3: Resultados para métodos de detecção de *onsets* para sinais reais com base no limiar calculado a partir de testes com sinais contendo acordes.

Método	Taxa de detecção	Taxa de falso alarme
HFC	16,9%	6,7%
Distância espectral	75,9%	8,8%
Desvio de fase	15,7%	0,0%
Domínio complexo	54,2%	18,9%

de fase ter tido um resultado muito aquém do resultado obtido no teste com sinais sintéticos. Isso se dá pois o sinal real é, como mencionado anteriormente, sujeito a efeitos como reverberações e ecos. Apesar de reverberações suavizarem transitórios (o que favoreceria o método de desvio de fase), isso provoca distorção na fase e, por isso, os métodos que atuam a partir da fase são bastante prejudicados. Isso, conseqüentemente, também afetou o método que opera no domínio complexo, ainda que de forma mais branda devido ao fato de este método também utilizar a informação de módulo. Como a informação de módulo se apresentou mais robusta a este tipo de distorção, esse método apresentou resultados melhores do que aqueles que só usam informação de fase.

O método que apresentou os melhores resultados foi o da distância espectral. Este método, embora tenha seguido a tendência prevista de não ter sua taxa de detecção muito elevada, apresentou uma taxa de falso alarme também reduzida em relação ao teste com sinal sintético. Isso se deve possivelmente ao fato de que o sinal real apresentado nesse caso tem em sua maior parte notas individualizadas, e com menor energia que o sinal usado para calibrar o limiar de detecção. O limiar, definido para um nível mais alto de sinal, acabou por ficar acima do valor que equilibraria essas duas taxas para o sinal real. Desse modo, o sistema de detecção ficou mais restritivo e a taxa de falso alarme caiu.

O mesmo não aconteceu para o método HFC, devido à sua ponderação maior para as altas freqüências. Na composição sob teste, as notas a mais são, na grande maioria das vezes, de freqüências mais baixas, pois são executadas pela mão esquerda. Como, no HFC, é dado um peso muito maior para as freqüências maiores, a energia presente nas baixas freqüências não interferiu tanto no resultado final.

3.5 Conclusão

A detecção de *onsets* é uma etapa fundamental para se realizar a TMA, tendo como objetivo a marcação dos instantes em que ocorre o início da execução de uma nota. Neste capítulo, foram descritas as implementações de diversos métodos para esse fim: *High Frequency Content*, Distância Espectral, Variação do Desvio da Fase e Detecção no Domínio Complexo.

Foram feitos experimentos computacionais em dois cenários: sinais sintéticos, gerados a partir do protocolo MIDI, e sinal real, obtido de gravação. O desempenho de todos os métodos em ambos os cenários foi analisado e seus resultados, comparados. O melhor desempenho obtido para os sinais sintéticos foi do método que atua no domínio complexo (cerca de 88% de taxa de detecção e em torno de 10% para falso alarme), enquanto que, ao se analisar o sinal real, foi do método que usa a distância espectral (cerca de 75% de taxa de detecção e em torno de 8% para falso alarme). Os experimentos mostraram que existem fatores inerentes aos sinais reais que afetam o desempenho de forma diferente cada método de detecção de *onsets*.

Devido ao fato de os métodos apresentados aqui atuarem sobre o domínio da freqüência, pode ser benéfico utilizar um espectro obtido por um banco de filtros de alta seletividade. No Capítulo 5 desta tese, serão apresentadas ferramentas de análise espectral com características que podem facilitar a detecção de *onsets*.

Capítulo 4

Identificação de múltiplas F0s

4.1 Introdução

Um sistema para TMA tem como base a detecção e a identificação[1] das múltiplas freqüências fundamentais presentes no sinal sob análise, denotadas individualmente por F0 (lê-se "efe-zero"). Esse capítulo tem como objetivo descrever métodos através dos quais esse problema é atacado na literatura e de que forma esses métodos contribuíram para o desenvolvimento deste trabalho. É importante deixar claro que a análise espectral empregada no método descrito no presente capítulo é feita através da FFT.

Na Seção 4.2, é feita uma revisão bibliográfica de métodos anteriores de estimação de F0. A Seção 4.3 descreve o sistema que serviu como base para o módulo de identificação de freqüências fundamentais do presente trabalho. Os resultados dos testes são discutidos na Seção 4.4. Questões acerca do critério de parada do método iterativo são consideradas na Seção 4.5. A Seção 4.6 discute a complexidade computacional do sistema de identificação de F0s e a Seção 4.7 conclui o capítulo.

[1] Aqui se faz distinção entre detectar, que denota reconhecer a existência de uma F0; e identificar, que seria descobrir sua identidade, ou seja, o valor da F0 ou o nome da nota a qual ela se refere.

4.2 Fundamentos para identificação de múltiplas F0s

No processo de identificação das F0s presentes em uma mistura, existem dificuldades a serem contornadas:

- Cada F0 é acompanhada por uma série de componentes ou parciais. Algumas dessas parciais, idealmente, apresentam freqüências múltiplas inteiras de F0, recebendo o nome de parciais harmônicas ou, simplesmente, harmônicos.

- Ocorre sobreposição entre harmônicos relativos a freqüências fundamentais que apresentam uma proporção praticamente racional entre si.

- Na prática, as parciais não aparecem exatamente como múltiplas inteiras de F0. Isso acontece devido tanto a erros de precisão freqüencial na análise tempo-freqüência quanto devido ao fenômeno conhecido por inarmonicidade [37]. Um dos modelos aceitos de inarmonicidade para predizer a real posição dos harmônicos é dado por

$$f = hF\sqrt{1 + (h^2 - 1)\beta}, \tag{4.1}$$

  onde f é a freqüência do harmônico, h é o índice desse harmônico, F é a freqüência fundamental e β é o fator de inarmonicidade.

Existem algumas abordagens possíveis para se resolver o problema da identificação de F0: algoritmos estatísticos e algoritmos que fazem uso da periodicidade do sinal sob análise.

O primeiro grupo é representado por métodos que definem modelos parametrizados por relações harmônicas e, então, fazem uso de algum conhecimento prévio da estatística do som a ser analisado para estimar os parâmetros. Goto [38], por exemplo, estima a probabilidade *maximum a-posteriori* (MAP) usando o algoritmo de *expectation-maximization*. Outros exemplos de algoritmos estatísticos são vistos em [39], [40] e [41].

O segundo grupo é sub-dividido em duas grandes famílias de algoritmos para identificação de F0. O que caracteriza cada uma delas é o domínio no qual eles fazem uso da periodicidade do sinal sob análise: tempo ou freqüência.

A maioria dos métodos de estimação de F0 se baseia na medição da periodicidade do sinal de entrada no domínio do tempo. A outra família de estimadores de F0, representada, por exemplo, por [42] e [43], mede a periodicidade no espectro do sinal sob análise. Esse métodos têm como base o fato de que um som harmônico tem seu espectro aproximadamente periódico. Assim, a função de autocorrelação $\rho(m)$ de um espectro de comprimento N pode ser calculada por

$$\rho(m) = \frac{2}{N} \sum_{k=0}^{N/2-m-1} |X[k]|\,|X[k+m]|\,, \tag{4.2}$$

onde $X[k]$ é uma amostra do sinal de interesse no domínio da freqüência, N é o número de amostras desse sinal e m é um número inteiro que representa o deslocamento ou diferença entre os índices de duas amostras.

Uma vantagem desse tipo de método é que os cálculos são robustos, mesmo a imperfeições quanto à harmonicidade. Isso é devido ao fato de que a diferença entre a freqüência de um dado harmônico e a freqüência do harmônico seguinte varia menos que a diferença entre a freqüência prática de um dado harmônico (que sofre de inarmonicidade) e sua posição harmônica teórica (sem inarmonicidade) [6].

Uma forma generalizada de se proceder quando se trabalha com sinais polifônicos foi, então, sugerida por de Cheveigné em [44] e [45]. Essa forma visava à estimação do valor de F0 seguida da remoção do som (correspondente à F0 detectada e identificada) da mistura e de uma posterior estimação tendo como base o sinal residual (resultado da remoção). Um efeito colateral dessa técnica inicial era o fato de essa remoção também afetar os harmônicos de outros sons que coincidiam com aqueles harmônicos cujos sons haviam sido cancelados.

Outra linha de pesquisa era formada pelos métodos que defendiam sistemas baseados em modelos matemáticos para o sistema auditivo humano. Segundo essa filosofia, Meddis e Hewitt [46] apresentaram o modelo unitário para percepção de *pitch*. Mais tarde, Karjalainen e Tolonen propuseram (em [47] e [48]) uma versão computacionalmente eficiente do modelo unitário. Esse sistema dividia o espectro em apenas duas faixas, introduzindo várias modificações que garantiam sua robustez como contra variações no timbre, o que era conseguido através de um pré-branqueamento do sinal de entrada [49].

Klapuri, então, desenvolveu um método [50][7] reunindo características desejáveis de várias das metodologias apresentadas nesta seção; dessa forma, as dificuldades citadas de cada um seriam contornadas. Isso será mostrado na próxima seção. É importante destacar que, mesmo contornando todas essas dificuldades, o problema ainda não está resolvido, pois existem outras questões abertas não levantadas pelos métodos citados aqui.

4.3 Método de Klapuri

Nessa seção, será descrito o sistema que serviu como base para o módulo de identificação de múltiplas F0s contido no sistema de transcrição musical automática sob desenvolvimento no presente trabalho. Esse módulo é fundamentado no algoritmo para identificação de múltiplas F0s apresentado na tese de doutorado de Anssi Klapuri [7] e em [50].

O sistema de Klapuri apresenta a estrutura descrita na Figura 4.1. Este pode ser implementado através das seguintes etapas:

1. Estabelecimento de um modelo para o sistema auditivo;
2. Pré-processamento do sinal de entrada;
3. Análise em sub-bandas;
4. Integração dos resultados obtidos para cada sub-banda;
5. Estimação da freqüência fundamental predominante;
6. Suavização do espectro;
7. Subtração do espectro relativo à freqüência fundamental predominante.

As sub-seções seguintes detalharão os pormenores de cada uma dessas etapas. O sistema será ilustrado através de um exemplo cujo sinal de entrada corresponde à soma de uma nota A2 (com F0 igual a 110 Hz e 40 harmônicos) com uma nota A4 (com F0 igual a 440 Hz e 10 harmônicos), ambas com fator de decaimento de 0,95 para cada harmônico e fator de inarmonicidade β de 0,0001.

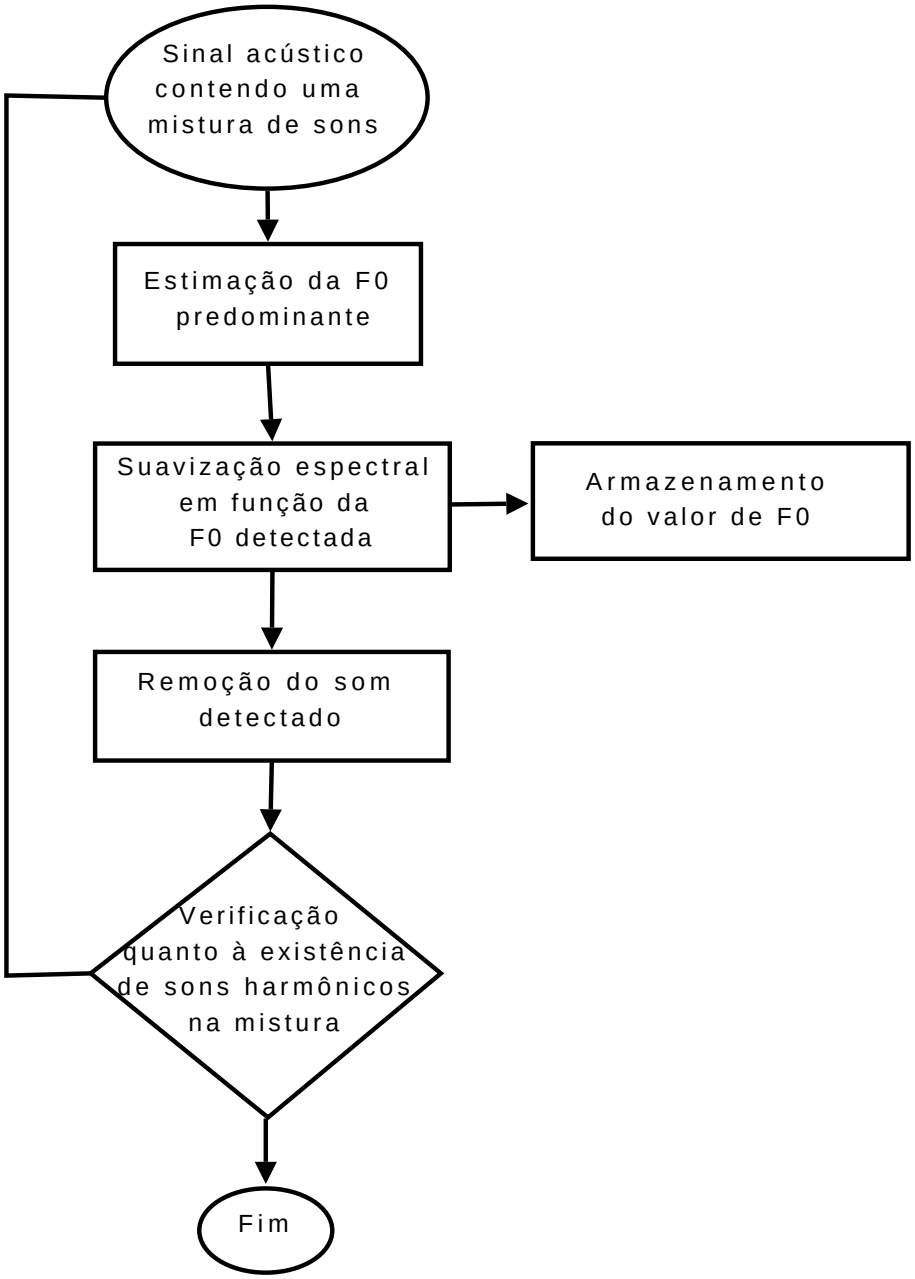


Figura 4.1: Estrutura do sistema de identificação de freqüência fundamental descrito por Klapuri.

O sinal de entrada para esse exemplo pode ser analiticamente descrito como:

$$\begin{aligned} x[k] &= \sum_{p=1}^{40}(0{,}95)^p \text{sen}\left(\frac{2\pi(110)pk}{f_S}\sqrt{1+(p^2-1)\beta}\right) \\ &+ \sum_{q=1}^{10}(0{,}95)^q \text{sen}\left(\frac{2\pi(440)qk}{f_S}\sqrt{1+(q^2-1)\beta}\right), \end{aligned} \tag{4.3}$$

onde k é o índice da amostra no domínio do tempo, a freqüência de amostragem f_S vale 44100 Hz e p e q são índices das parciais harmônicas.

Esse sinal foi, então, normalizado para que sua amplitude não ultrapassasse a unidade. Além disso, para dar mais realismo ao exemplo, foi nele inserido um ruído branco gaussiano de média zero e variância 0,1. Isso equivale a um sinal com SNR de 4,5 dB. O espectro do sinal de entrada (que pode ser visualizado na Figura 4.2) foi calculado usando-se uma DFT de mesmo tamanho que a janela de Hanning do artigo original [50], 190 ms.

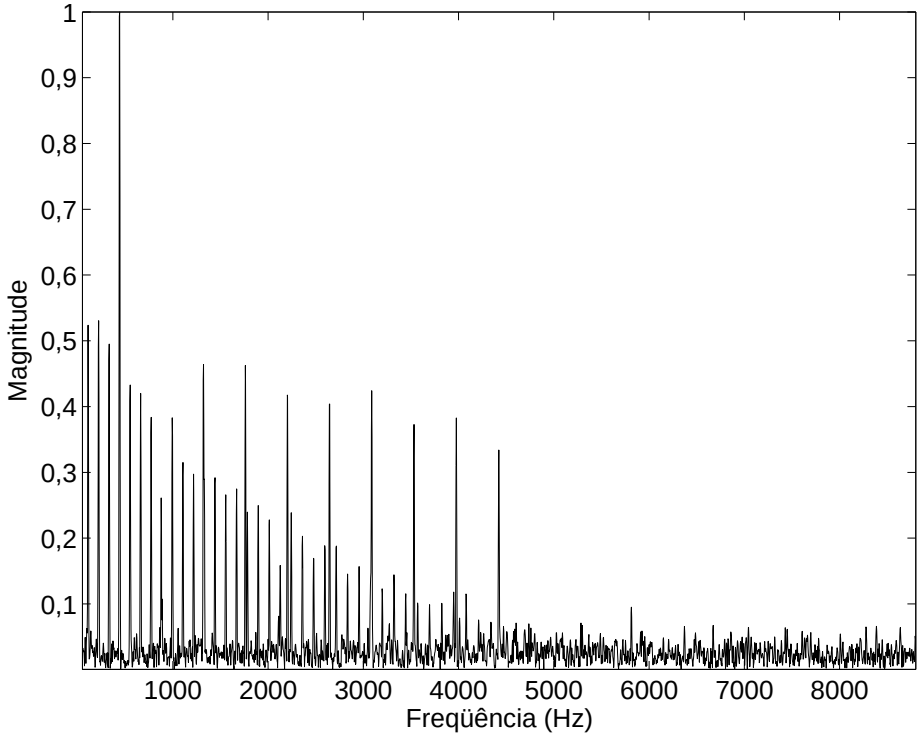


Figura 4.2: Detalhe do espectro do sinal de entrada, onde é possível perceber o "chão-de-ruído".

4.3.1 Estabelecimento de um modelo para o sistema auditivo

Segundo Klapuri, o sistema descrito em [7] tem como fundamento o sistema auditivo humano e a forma através da qual o cérebro consegue separar as freqüências fundamentais presentes em uma mistura. Tal sistema, portanto, deve realizar todo o processamento dividindo o espectro em sub-bandas. Dentro de cada sub-banda, a inarmonicidade poderia ser considerada reduzida.

O modelo, descrito em [51], consiste de 18 sub-bandas que se distribuem geometricamente entre 50 Hz e 6000 Hz. No presente trabalho, para garantir que não haja erros no fim da região de interesse, usaram-se 21 sub-bandas que se distribuem de 50 Hz a 8800 Hz. Cada banda deve obedecer às seguintes restrições:

- Conter uma região do espectro correspondente a 2/3 de oitava;
- Ter uma banda mínima de 100 Hz;
- Apresentar resposta em freqüência com forma triangular;
- Apresentar sobreposição de 50 % com as bandas adjacentes.

Cada banda desse modelo é denotada por G_b, onde b é o índice da sub-banda. Satisfazendo tais restrições, o modelo simplificado do sistema auditivo humano fica como ilustrado na Figura 4.3.

4.3.2 Pré-processamento do sinal de entrada

Como o objetivo é detectar e identificar todas as freqüências fundamentais presentes em uma mistura e é preciso considerar a possibilidade de elas terem energias de valores muito diferentes, deve-se equalizar a energia de todos os componentes freqüenciais. Isso atuaria como uma forma de branqueamento do do espectro do sinal de entrada e conseqüente equalização da intensidade da envoltória do espectro acarretando a alteração do timbre do sinal (descaracterizando-o), mas é importante lembrar que a informação de timbre não é relevante no momento, visto que o objetivo atual resume-se a detectar e identificar simplesmente as freqüências fundamentais presentes na mistura, e não que instrumento as está gerando.

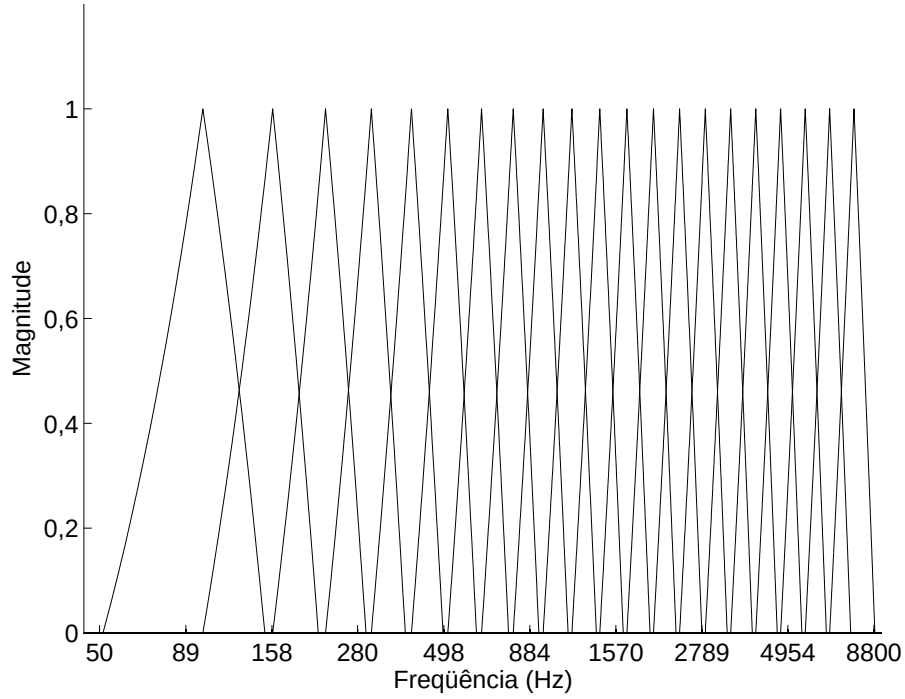


Figura 4.3: Resposta em freqüência das sub-bandas contidas no modelo de Klapuri para o sistema auditivo humano.

Para isso, Klapuri lançou mão de um procedimento conhecido como *magnitude-warping*. Isso consiste em aplicar uma transformação não-linear sobre o espectro do sinal de entrada, como a descrita por

$$Y[k] = \ln\left(1 + \frac{1}{g}X[k]\right), \tag{4.4}$$

onde $X[k]$ é o módulo do espectro do sinal de entrada, $Y[k]$ é o espectro do sinal pré-processado e g é dado por:

$$g = \left(\frac{1}{k_1 - k_0 + 1}\sum_{l=k_0}^{k_1} X^{\frac{1}{3}}[l]\right)^3. \tag{4.5}$$

Nesta equação, k_0 e k_1 denotam os limites inferior e superior impostos ao identificador de múltiplas F0s. Em [7], k_0 e k_1 são respectivamente iguais a 50 Hz e 6 kHz. Considera-se, então, que os possíveis valores para F0 estão dentro dessa banda. O próprio pré-processamento se encarrega de remover o espectro acima de 6 kHz.

Ao se submeter o espectro do sinal de entrada, mostrado na Figura 4.2, ao procedimento de *amplitude-warping*, chega-se ao espectro exibido na Figura 4.4. Comparando-se as Figuras 4.2 e 4.4, é possível perceber que os picos de menor

magnitude sofrem um aumento, enquanto que aqueles de maior magnitude têm suas amplitudes aproximadamente constantes. Para facilitar a visualização, as escalas dessas figuras não foram mantidas. De fato, não é necessário guardar tal proporção.

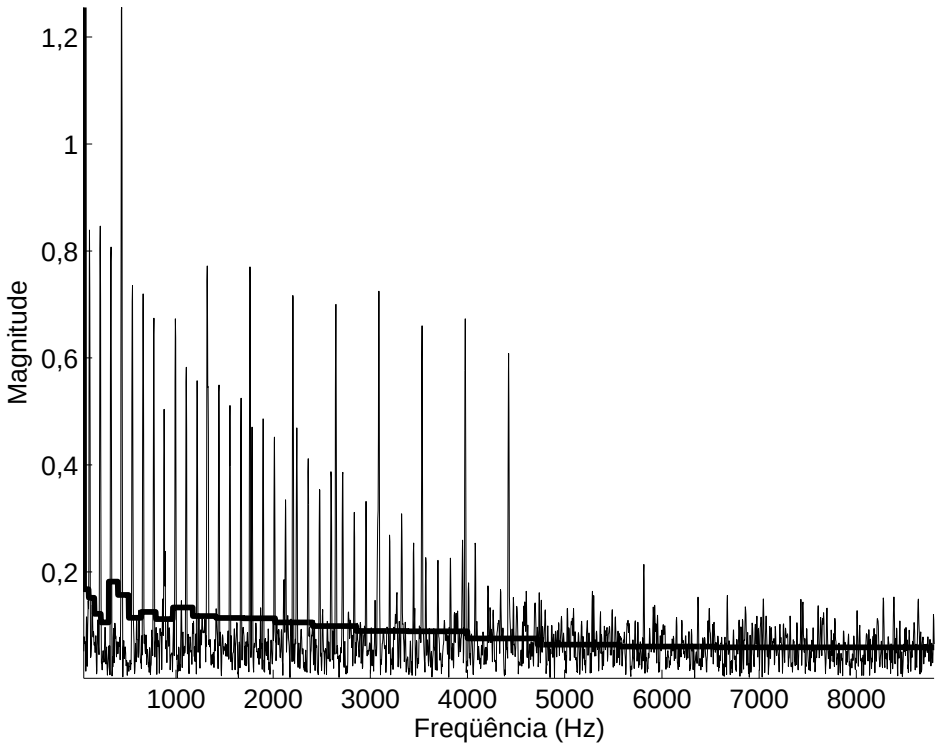


Figura 4.4: Detalhe do espectro do sinal de entrada após o procedimento de *amplitude-warping*, indicado pela linha mais fina. A linha de maior espessura representa a estimação do ruído.

Nas Figuras 4.2 e 4.4, verifica-se um "chão-de-ruído" que deve ser eliminado da representação, segundo o procedimento a seguir:

1. Definir um espectro $\hat{N}[k]$ do mesmo comprimento de $Y[k]$. Esse espectro $\hat{N}[k]$ atuará como um limiar para remoção do ruído.

2. Para cada sub-banda b do modelo do sistema auditivo, calcular a média m_b das amostras do espectro $Y[k]$ para $k_0 \leq k \leq k_1$;

3. Atribuir o valor da média calculada m_b no passo anterior para todas as amostras do espectro $\hat{N}[k]$ contidas na sub-banda b. Para a região de sobreposição das sub-bandas, usa-se a média da sub-banda de maior índice;

4. Atribuir o valor máximo do espectro $Y[k]$ para todas as amostras de $\hat{N}[k]$ contidas além da última sub-banda, para que se considere apenas o espectro até a última sub-banda;

O espectro resultante $\hat{N}[k]$ está indicado na Figura 4.4 por uma linha de maior espessura.

A supressão de ruído consiste em definir o espectro modificado:

$$Z[k] = \max\left\{0, Y[k] - \hat{N}[k]\right\}. \tag{4.6}$$

Para o exemplo em análise, o espectro resultante da supressão de ruído é mostrado pela Figura 4.5. Pode-se ver que o ruído de baixa amplitude visível na Figura 4.4 foi bastante reduzido, principalmente em baixas freqüências.

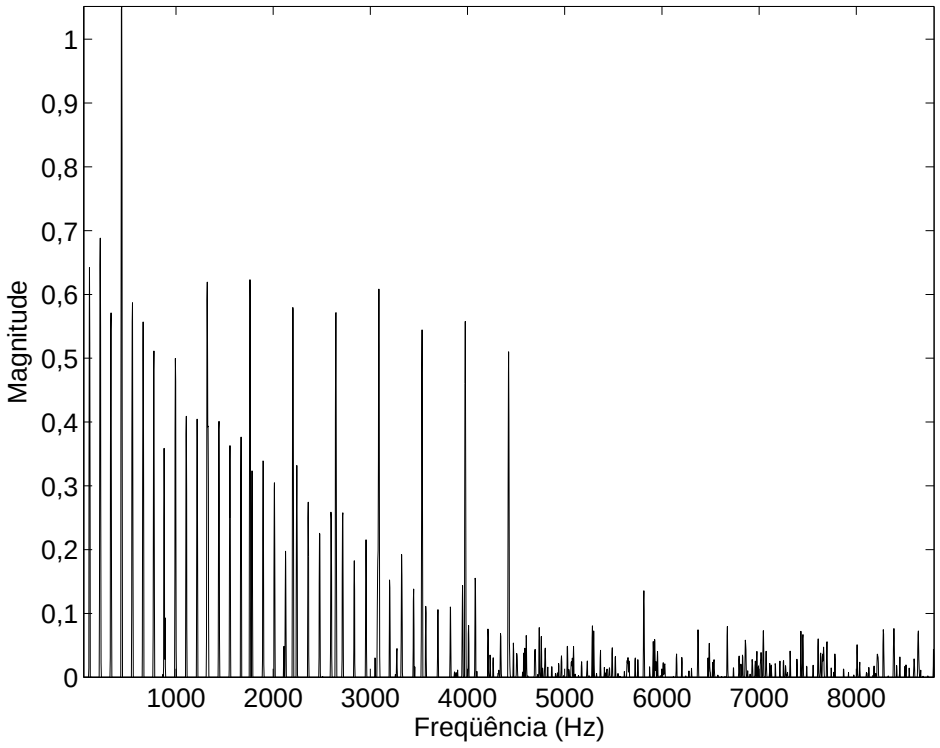


Figura 4.5: Detalhe do espectro do sinal de entrada após os procedimentos de *amplitude-warping* (linha fina) e supressão de ruído. A linha grossa é a estimativa do ruído.

4.3.3 Análise em sub-bandas

A análise em sub-bandas é baseada no conceito da seleção harmônica [52]. Este consiste em definir um critério para selecionar amostras do espectro, dependente

de uma série de parâmetros. A seguir, deve-se somar as amostras selecionadas por esse critério para diversos valores dos parâmetros. Por fim, deve ser possível reconhecer uma relação entre o valor da soma e os parâmetros usados para definir o critério de seleção. Ao se realizar a seleção harmônica, apenas aquelas componentes múltiplas inteiras da candidata a F0 poderão interferir positiva ou negativamente na sua identificação. Todo este procedimento é detalhado a seguir.

Uma forma bastante simplificada de se representar a seleção harmônica é através da equação

$$L(b,n) = \sum_{j=0}^{J(n)-1} Z_b[k_b + nj]. \tag{4.7}$$

A Eq. (4.7) objetiva somar as magnitudes dos $J(n)$ harmônicos da n-ésima candidata a F0, iniciando-se na primeira posição, de índice k_b, no espectro Z_b definido dentro da sub-banda b. O espectro Z_b é obtido submetendo-se o espectro pré-processado Z pelo filtro referente à sub-banda b, denominado G_b, ou seja, $Z_b = Z \times G_b$. Isso é feito para cada valor de b e para cada F0 candidata (denotada por n). A razão de se iniciar a soma na posição k_b é o fato de que, em uma dada sub-banda, existem parciais harmônicas relativas a todas as freqüências maiores ou iguais à freqüência associada a k_b e menores que a largura desta sub-banda.

Contudo, não se conhece a posição real de cada harmônico, que ainda pode sofrer o efeito da inarmonicidade, explicada anteriormente. As possibilidades de deslocamento freqüencial são contempladas pela inserção da variável m pertencente ao intervalo $0 \leq m \leq n-1$, fazendo com que a Eq. (4.7) seja alterada para

$$L(b,n,m) = \sum_{j=0}^{J(m,n)-1} Z_b[k_b + m + nj]. \tag{4.8}$$

Aqui, $J(m,n)$ é o número de múltiplos inteiros da candidata a F0 denotada por n dentro da atual sub-banda, e é dado por

$$J(m,n) \;\; = \;\; \left\lfloor \frac{K_B - m}{n} \right\rfloor, \tag{4.9}$$

onde K_B é o número de amostras da sub-banda em questão e o operador $\lfloor \bullet \rfloor$ denota o maior inteiro menor que ou igual ao argumento. Como existe o *offset* m, a região restante da sub-banda em questão (do índice m até o final) passa a ter $K_B - m$

amostras. Dividindo-se essa região pelos n intervalos e tomando-se a parte inteira do resultado, obtém-se J.

É necessário, então, buscar o valor máximo de $L(b, n, m)$ em relação a m

$$L(b,n) = \max_{0 \leq m \leq n-1} \left\{ \sum_{j=0}^{J(m,n)-1} Z_b[k_b + m + nj] \right\}. \tag{4.10}$$

A condição $m \leq (n-1)$ evita possível sobreposição de harmônicos, o que poderia ocorrer se $m = n$, prejudicando conclusões subseqüentes. Contudo, pode-se restringir ainda mais o valor máximo de m, calculando-se m para o maior valor do fator de inarmonicidade β relativo ao maior valor de n e l_b, dentro da sub-banda b. Em [50], o valor máximo de β é igual a 0,01. Assim, chega-se ao maior valor de m para o qual se deve testar a função custo em busca se seu máximo. Para se realizar este cálculo, faz-se

$$m_{max} = l_b \left(\sqrt{1 + 0,01 \left[(l_b/n)^2 - 1 \right]} - 1 \right). \tag{4.11}$$

A Eq. (4.11) tem base na Eq. (4.1) e (l_b/n) é uma aproximação para o índice h da parcial harmônica. De acordo com esta equação, se $h = 1$, m_{max} se torna igual a zero, ou seja, a Eq. (4.1) indica que não inarmonicidade. Contudo, para valores de h muito maiores que 1, todos os valores de m devem ser considerados, isto é, de zero a m_{max} (diferente de zero).

Assim, para uma faixa de freqüência onde se considera haver apenas a primeira parcial de uma F0 correspondente ao índice n, a inarmonicidade, segundo o modelo adotado, não ocorre, e j poderá assumir apenas o valor 1, reduzindo a Eq. (4.10) a

$$L(b,n) = Z_b(n). \tag{4.12}$$

Do mesmo modo, se na sub-banda em questão estiver contido apenas o segundo harmônico da uma F0 correspondente ao índice n, por se tratar de um índice h pequeno, a inarmonicidade também não será permitida, j poderá assumir apenas o valor 2 e a Eq. (4.10) se reduzirá a

$$L(b,n) = Z_b(2n). \tag{4.13}$$

Quando o índice do harmônico for alto (para valores de n que variam de k_1 a K_B), a inarmonicidade será permitida e a Eq. (4.10) será empregada em sua forma original.

Por exemplo, se a sub-banda em questão estiver definida de 1000 Hz (k_b) a 1200 Hz (l_b), os valores de n a serem testados variam de 50 Hz (correspondente a k_1) a 200 Hz (correspondente a K_B). Assim, a sub-banda em questão contém as parciais de índices 20 a 24 para o valor de n relativo a 50 Hz e as parciais de índices 5 e 6 para o valor de n referente a 200 Hz. Para todos os valores de n contidos neste intervalo, haverá mais de uma parcial dentro da sub-banda considerada. Desta maneira, segundo [50], pode-se entender que o algoritmo combina o uso de posições espectrais para os harmônicos de índice baixo com o uso de intervalos espectrais para harmônicos de índice alto.

A Eq.(4.10) ainda necessita de uma normalização, denotada pela função $c(m,n)$, que é utilizada também em [50] e descrita em [7]. Esta, segundo [7], foi obtida a partir de métodos de aprendizagem computacional e visa a equalizar o efeito final de cada candidata a F0. Se não houvesse normalização, valores mais baixos de n implicariam um valor mais elevado para J e, por isso, um valor maior no somatório que representa a seleção harmônica. O valor de $c(m,n)$ varia entre 0,25 (para valores altos de J) e 1 (para J igual a 1). Com isso, chega-se à forma final da equação que implementa a seleção harmônica:

$$L(b,n) = \max_{0 \leq m \leq n-1} \left\{ c(m,n) \sum_{j=0}^{J(m,n)-1} Z_b[k_b + m + nj] \right\}, \tag{4.14}$$

onde

$$c(m,n) \;=\; \frac{0{,}75}{J(m,n)} + 0{,}25 \tag{4.15}$$

Assim, é possível reconhecer uma relação entre uma candidata a F0 predominante e a soma de seus harmônicos dentro de uma mesma sub-banda. Uma vantagem clara da seleção harmônica em sub-bandas é o fato de se limitar o número de componentes freqüenciais que podem interferir na identificação de uma certa freqüência fundamental. Como se viu, ao se realizar a seleção harmônica, apenas as componentes múltiplas inteiras da candidata a F0 poderão interferir na sua identificação. Na verdade, com a presença do termo m, a possibilidade de haver inarmonicidade é também levada em consideração, de modo que a soma desses harmônicos seja computada sobre amostras mais próximas de suas posições reais.

Ao fim da análise em sub-bandas, ter-se-á uma matriz $\mathbf{L}$, de B linhas e N colunas, onde B é o número de sub-bandas e N é o número de amostras do espectro dentro dos

limites inferior e superior definidos no modelo do sistema auditivo. Cada elemento n de cada linha da matriz representará o valor máximo da seleção harmônica para a candidata a F0 de índice n.

Seguindo com o exemplo proposto, a matriz resultante dessa etapa após uma iteração é mostrada na Figura 4.6. Essa figura exibe a matriz no formato de uma imagem em que cada píxel representa um elemento, e o nível de cinza associado a esse píxel corresponde ao valor desse elemento. Nessa escala, branco indica que o valor é igual a zero. Nessa matriz, é possível reconhecer as 21 linhas correspondentes a cada uma das 21 sub-bandas definidas para o modelo dos filtros auditivos entre 50 Hz e 8800 Hz (ver Seção 4.3.1). As colunas representam cada componente freqüencial n. Pode-se também ver que a extensão da parte cinzenta em cada linha varia com o índice da sub-banda, o que faz sentido, visto que as bandas apresentam larguras que guardam uma razão geométrica entre si. Mais explicações serão fornecidas ao fim da seção seguinte.

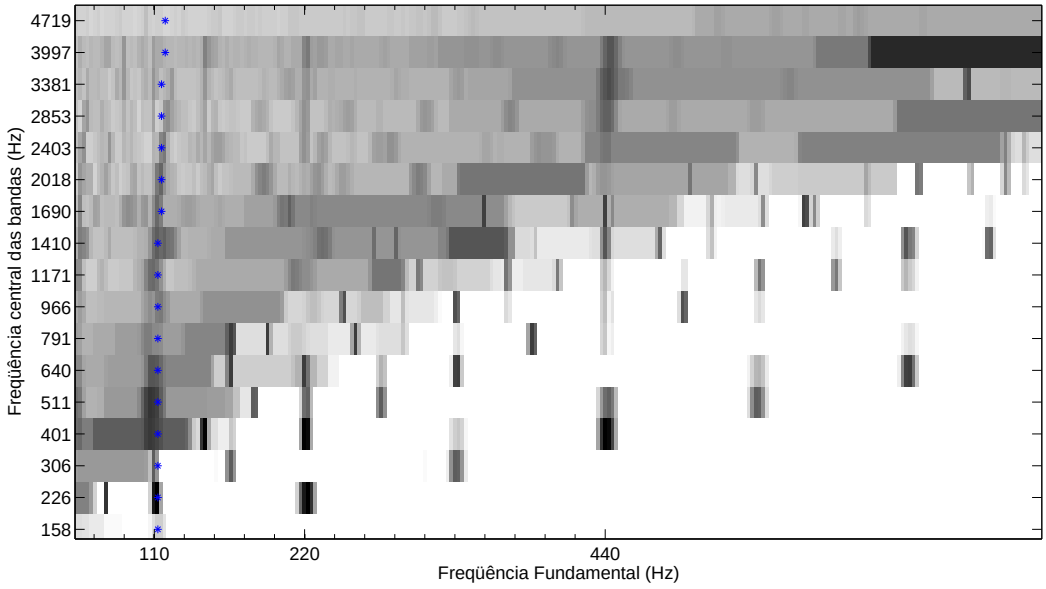


Figura 4.6: Detalhe da matriz resultante da análise em sub-bandas após uma iteração aplicada ao sinal descrito na Eq. (4.3). Os asteriscos representam os elementos da matriz L selecionados a partir do modelo de inarmonidade utilizado na fase da integração dos resultados de cada sub-banda.

4.3.4 Integração dos resultados obtidos para cada sub-banda

Nesta etapa, as curvas $L(b, n)$ obtidas individualmente para cada sub-banda serão integradas em um único vetor. Isso é feito testando-se, para cada candidata a F0, isto é, para cada coluna da matriz construída na etapa anterior, uma série de valores para o fator de inarmonicidade β. Aquele valor de β que definir a posição real do maior número de harmônicos será o valor escolhido. Se as fontes que compõem o sinal tiverem fatores de inarmonicidade diferentes, a integração será feita apenas para uma das fontes. Uma outra fonte com um fator diferente irá apenas atrapalhar este processo. Daí, pode-se ver que um segundo produto dessa etapa é o valor do fator de inarmonicidade para cada n.

Isso pode ser feito da seguinte forma:

1. Define-se um vetor $\mathbf{u}$ de possíveis valores para β. Esse vetor deve conter o elemento 0 (para o caso de não haver inarmonicidade) e uma faixa de valores que, por experimentação, foi adotada como sendo, neste trabalho, de 10 valores logaritmicamente espaçados de 10^{-5} a 10^{-4}.

2. Para cada n e cada um dos elementos de $\mathbf{u}$:

   (a) Calculam-se as diferenças entre as posições harmônicas e as posições com inarmonicidade em cada banda b segundo o modelo da Eq. (4.1), definindo o vetor $\mathbf{h}_n(b)$, considerando o fator de inarmonicidade β em questão, usando a Eq. (4.1);

   (b) Define-se

   $$\hat{L}(n) = \sum_b \left[L(b, h_n(b))\right]^2; \tag{4.16}$$

3. Armazena-se o valor máximo de $\hat{L}(n)$, para cada elemento do vetor $\mathbf{u}$, definindo-se assim a função global $L_{\mathrm{g}}(n)$, que denota a energia da série de harmônicos (destacada pela seleção harmônica) relativa à n-ésima candidata a F0.

Continuando com o exemplo proposto, a Figura 4.7 representa a função L_g obtida ao final desta etapa. Pode-se ver, na Figura 4.6, caminhos formados por regiões

mais escuras traçados pelos elementos de valores mais elevados da matriz L. Estes caminhos apresentam uma curvatura para a direita provocada pela inarmonicidade e pelo fato de a largura de cada sub-banda ser crescente. Para se realizar a integração dos resultados de cada sub-banda, deve-se acompanhar esses caminhos e, para isso, emprega-se o modelo de inarmonicidade definido pela Eq. (4.1).

Entretanto, a Eq. (4.1) serve para definir posições de parciais harmônicas. Deve-se, então, adaptar este modelo para que seja capaz de modelar a curvatura do caminho imposto pelos valores mais elevados da matriz L. Se todas as sub-bandas apresentassem a mesma largura, o modelo da Eq. (4.1) já seria adequado para definir as posições desses valores mais elevados. Mas como as diversas sub-bandas têm larguras cada vez maiores, essa proporção deve ser levada em consideração para o cálculo da curvatura resultante. Para este ajuste, o valor de β a ser testado numa sub-banda deverá ser normalizado, através de um fator dado pela razão entre a largura da sub-banda em questão e a largura da primeira sub-banda.

Valores maiores para o fator de inarmonicidade β têm a capacidade de dar maior curvatura ao caminho. Ao se testar uma série de valores de β, testa-se também uma série de curvaturas. Aquela que modelar o caminho de forma mais próxima, resultará em um valor mais alto para a integração e indicará o valor de β a ser usado. A precisão deste modelo pode ser verificada através da marcação com astericos sobre a matriz apresentada na Figura 4.6. Cada asterisco representa a posição do harmônico relativo a dado n em cada sub-banda, e a integração deve ser feita sobre essas posições.

4.3.5 Estimação da freqüência fundamental predominante

Nesta etapa, o vetor construído na etapa anterior, L_g, é analisado. O elemento desse vetor que apresentar o máximo valor tende a estar associado à F0 predominante. Contudo, para estimar a F0 predominante de forma mais precisa, deve-se lançar mão da seguinte heurística:

1. Verifica-se qual é o elemento p_1 da seqüência de pesos globais L_g que apresenta o maior valor.

2. Executa-se uma suavização (explicado na seção seguinte) do espectro do sinal

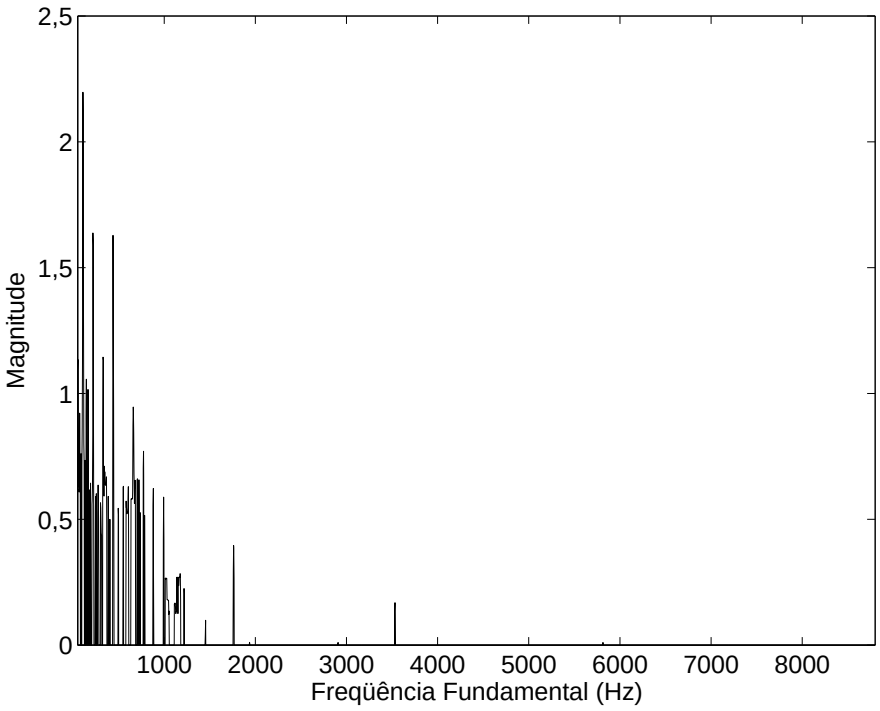


Figura 4.7: Vetor resultante da integração da análise de todas as sub-bandas após uma iteração no exemplo do sinal descrito na Eq. (4.3), onde o maior pico representa a freqüência de 110 Hz.

de entrada em função da F0 indicada por esse elemento de maior valor. Esse espectro suavizado denota a influência dessa F0 e de suas parciais harmônicas no espectro como um todo.

3. Recalcula-se o peso global para o elemento em questão levando-se em consideração o espectro suavizado ao invés do espectro original de entrada usado na análise em sub-bandas, chegando-se a $\tilde{p}_1$. Isso deve ser feito utilizando-se os algoritmos descritos nas Seções 4.3.3 e 4.3.4 apenas para o índice n que representa a F0 cujo peso é p_1.

4. Se o novo peso recalculado for menor que o segundo maior peso p_2 anterior à suavização, ou seja, se $\tilde{p}_1 < p_2$, então guarda-se a F0 candidata referente a p_1 e os cálculos são refeitos para o segundo maior elemento, p_2.

5. Os quatro passos anteriores devem ser refeitos até se encontrar algum elemento cujo peso $\tilde{p}_n$, após a suavização, seja maior que o maior peso menor que p_n antes da suavização, p_{n+1}.

6. O maior peso recalculado após a suavização indica a F0 predominante.

Um outro critério de parada usado para impedir que haja um número muito elevado de iterações de estimação é limitar o número de iterações. No presente trabalho, empregou-se um número máximo de 10 iterações de estimação.

Pode-se entender a suavização espectral como sendo o processo de obtenção do espectro que representa a influência de uma determinada F0 juntamente com suas parciais harmônicas. Logo, se uma dada F0 não estivesse presente, após a suavização, o peso global $\tilde{p}_1$ do elemento que denota essa F0 em questão sofreria uma grande redução, fazendo com que $\tilde{p}_1$ fosse menor do que p_2.

O novo cálculo do peso global após a suavização do espectro não compromete o custo computacional do sistema, pois apenas um peso $\tilde{p}_n$ é calculado e não a seqüência L_g completa de pesos globais. Para o exemplo em questão, como pode ser visto na Figura 4.7, a F0 predominante identificada é de 110 Hz.

4.3.6 Suavização espectral

Um dos maiores problemas para um sistema que se propõe a detectar e identificar múltiplas freqüências fundamentais é a sobreposição de harmônicos. Como explicado anteriormente, isso acontece sempre que houver duas séries de harmônicos cujos valores para F0 guardam uma proporção racional.

Uma idéia largamente utilizada é identificar a F0 predominante e removê-la, juntamente com todos os seus harmônicos, do espectro sob análise. Dessa forma, em uma segunda etapa, uma nova tentativa de detecção e identificação seria realizada sobre o espectro residual.

Dados esses dois pontos, surge a pergunta: como fazer para remover um harmônico relativo a uma F0 sem que a detecção e a identificação de outra F0 fiquem prejudicadas devido à remoção indevida de harmônicos comuns a ambas? Em [7], esse problema é abordado através do procedimento conhecido como suavização espectral. Tal procedimento é realizado antes da remoção do conjunto de harmônicos relativos à F0 predominante. O único valor que é necessário conhecer é o da F0 mais recentemente identificada.

O procedimento é o seguinte:

1. Localizar a posição de cada harmônico relativo à F0 predominante. É importante lembrar que, para se estimar a posição de cada harmônico, deve-se usar o valor do fator de inarmonicidade encontrado na etapa de integração dos resultados das sub-bandas. Contudo, para aumentar a precisão desta estimação (como uma forma de refinamento *a posteriori* da estimativa), o valor do fator de inarmonicidade é multiplicado por um fator λ_m que pode assumir algum dos N_m valores logaritmicamente espaçados entre 0,1 e 10. Para este trabalho, foi adotado, por experimentação, $N_m = 100$. Na Eq. (4.1), substitui-se β por $\lambda_m\beta$. O valor de λ_m que maximizar a soma das magnitudes dos harmônicos (apontados pela Eq. (4.1) modificada) será escolhido para fornecer as posições dos harmônicos. Então, um novo valor para o fator de inarmonicidade é definido multiplicando-se λ_m pelo valor anteriormente encontrado;

2. Definir janelas triangulares cujas freqüências centrais são as posições práticas dos harmônicos (com inarmonicidade) dessa F0 predominante, com largura de uma oitava e amostras definidas apenas para os valores de freqüência do espectro onde estão os harmônicos;

3. Usar cada janela como um vetor de pesos para se calcular a média ponderada dos harmônicos que estão dentro de seus limites;

4. Guardar o espectro suavizado (resultante), formado pelo conjunto dessas médias ponderadas, e definido apenas para os valores de freqüência onde estão os harmônicos. (Para a próxima etapa, é importante que se guardem esses valores.)

No exemplo, o resultado da suavização espectral na primeira iteração é ilustrado na Figura 4.8, onde a linha envoltória denota o espectro suavizado.

4.3.7 Subtração do espectro relativo à freqüência fundamental predominante

Esta etapa é a continuação natural e finalidade da etapa anterior. É necessária para que uma nova iteração possa ter início em busca de uma nova freqüência fundamental predominante.

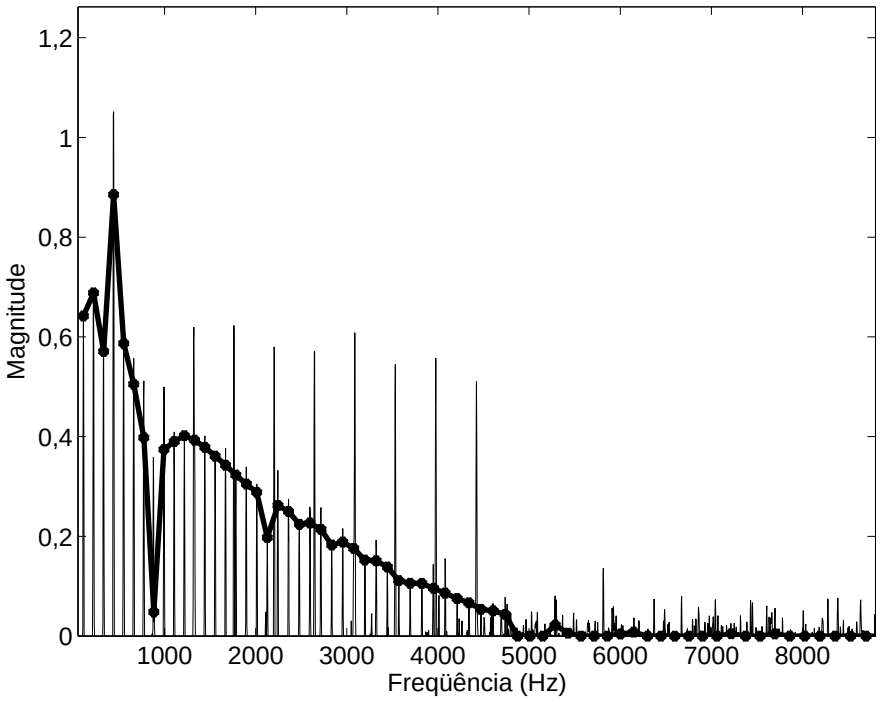


Figura 4.8: Suavização espectral na primeira iteração no exemplo do sinal descrito na Eq. (4.3).

No exemplo, a remoção é ilustrada na Figura 4.9, onde se pode ver que a série de harmônicos relativos à F0 predominante detectada e identificada na presente iteração, e constituída pelo espectro suavizado, foi removida. A eficácia dessa remoção pode ser melhor observada se a Figura 4.9 for comparada com o espectro do sinal de entrada, na Figura 4.2.

Um dos fatores que mais influenciam na qualidade da remoção do espectro relativo à F0 predominante é considerar, além das amostras relativas a cada harmônico do espectro suavizado, uma certa quantidade de amostras vizinhas. Isso deve ser feito porque, como há sobreposição entre os filtros dos diversos canais da DFT, cada harmônico é representado por mais de um canal. Se apenas a amostra central de cada harmônico for subtraída, isso não será suficiente para remover a influência da F0 em questão. Por experimentação, o número de amostras vizinhas foi definido como 12 (isto é, 6 de cada lado da amostra original).

A seguir, a segunda iteração do algoritmo tem início considerando o espectro resultante da presente etapa como espectro de entrada. A análise em sub-bandas é executada, gerando-se a matriz exibida na Figura 4.10. A análise relativa a cada

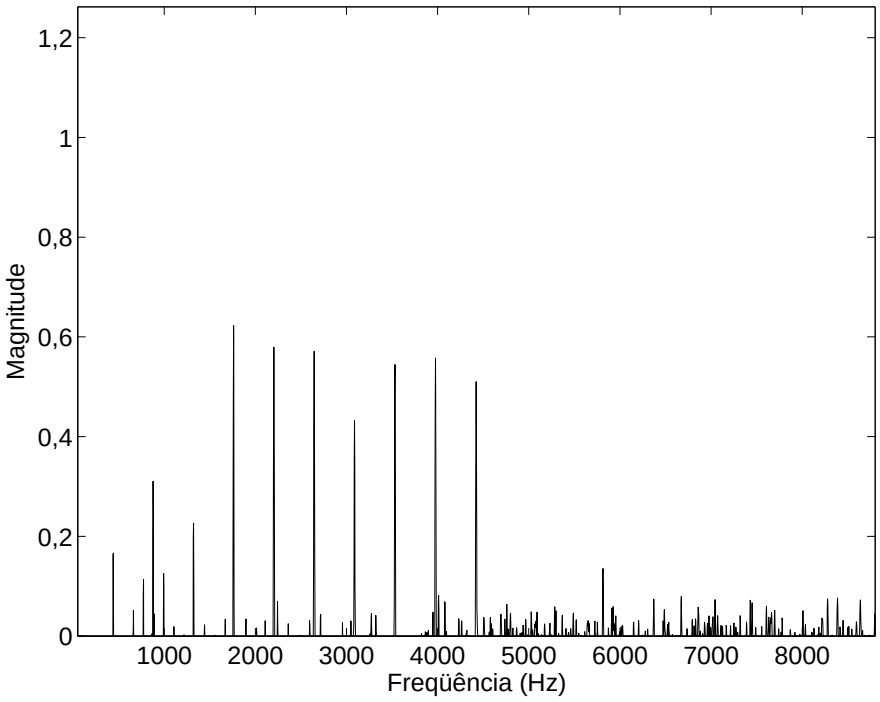


Figura 4.9: Resultado da remoção do espectro relativo à F0 predominante na primeira iteração.

uma das sub-bandas é integrada, chegando-se a uma análise global, mostrada na Figura 4.11. A partir dessa análise global, é possível estimar o valor da F0 predominante como sendo igual a 440 Hz. Executando-se, então, a suavização espectral, realiza-se a remoção da série de harmônicos relativos a esta F0 predominante. A remoção resulta em um espectro que é praticamente formado por ruído, o que indica o fim das iterações do método.

4.4 Aspectos práticos

Esta seção objetiva apresentar alguns aspectos práticos sobre o que foi exposto no presente capítulo. Primeiramente, será discutida a metodologia dos experimentos. Em seguida, será abordado o desempenho do sistema de identificação de F0s para a situação em que a quantidade de notas é informada ao sistema.

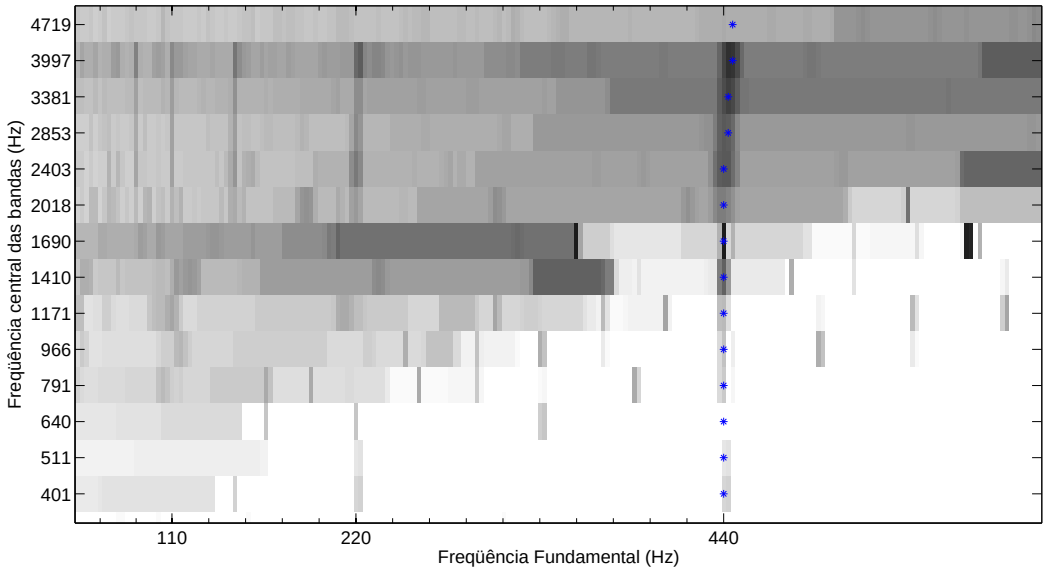


Figura 4.10: Detalhe da matriz resultante da análise em sub-bandas na segunda iteração.

4.4.1 Descrição dos sinais de teste

Os testes com sinais reais foram feitos a partir de sons de piano contidos na base McGill University Master Samples (MUMS) [53]. Foram usados os arquivos referentes ao piano gravado através do sistema *Multi-Point Pickup* (MPP), pois são os que apresentam melhor qualidade de gravação. Existem 3 tipos de dinâmica (ver Apêndice A): forte (*forte* - contidas na pasta LOUD), média (*mezzo* - na pasta MEDIUM) e suave (*piano* - na pasta SOFT). Cada pasta tem todas as notas desde A0 a B7. Cada arquivo contido na referida pasta contém o som produzido ao se pressionar com uma dada intensidade uma tecla específica do piano. Nos testes de desempenho em relação à taxa de identificação de notas, as três dinâmicas serão empregadas. Cada sinal foi analisado com uma janela de 190 ms tomada 100 ms após o início da nota.

4.4.2 Desempenho quanto à identificação de notas

Além de se testar o sistema de identificação com cada arquivo de piano (de C2 a B6) da base MUMS separadamente, testaram-se também misturas de 2, 3 e 4 sons.

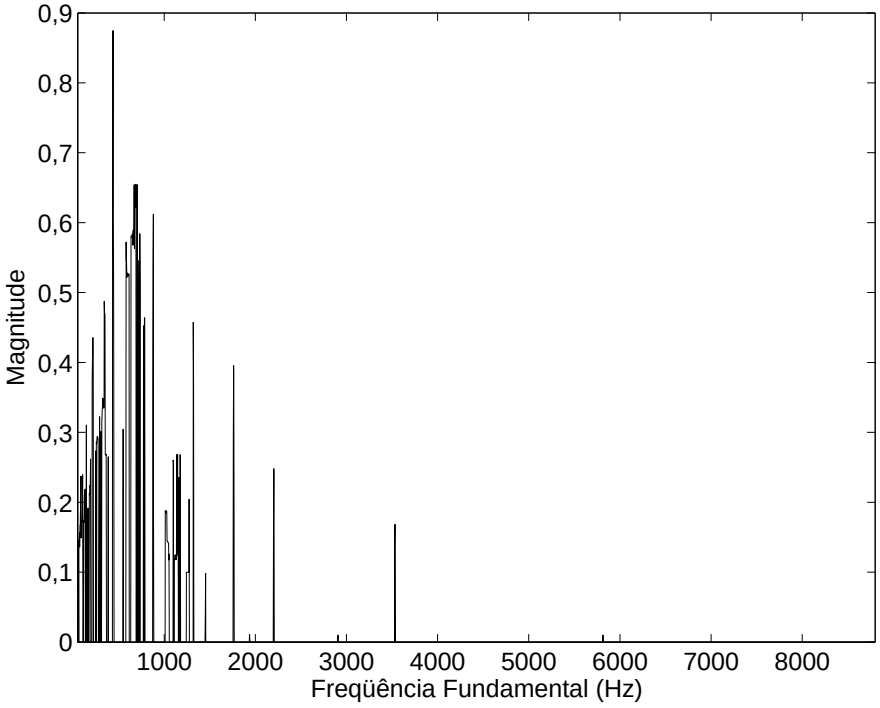


Figura 4.11: Vetor resultante da integração da análise de todas as sub-bandas na segunda iteração.

A fim de se criar uma estrutura de testes perfeitamente reprodutível, optou-se por misturas musicais, em oposição a misturas aleatórias. Misturas musicais são aquelas em que os sons apresentam relações freqüenciais comumente encontradas na música ocidental. Klapuri mostra em [51] que sons presentes em misturas aleatórias são mais fáceis de se identificar automaticamente do que sons presentes em misturas musicais. Isso se deve em parte ao fato de que, nesse segundo tipo de mistura, os harmônicos apresentam maior sobreposição.

Dessa maneira, as misturas que contêm dois sons foram construídas somando-se os sinais referentes a duas notas, sendo a primeira nota igual a cada uma das notas de C2 a B6 e a segunda nota uma quinta acima da primeira. Isso equivale a dizer que a F0 da segunda nota é uma vez e meia a F0 da primeira. As misturas que contêm três sons foram montadas somando-se os sinais das três notas que formam acordes maiores, sendo que a primeira nota é cada uma das notas de C2 a B6. Do mesmo modo, as misturas com quatro sons foram concebidas de forma a compor acordes maiores com sétima. Assim, pode-se resumir as regras de formação das misturas através da Tabela 4.1. Segundo essas regras, para cada tipo de mistura, existem 60

sons, onde a nota de freqüência mais baixa é uma das notas de C2 a B6.

Tabela 4.1: Regras de formação das misturas de sons para testes, onde s indica 1 semi-tom (ver Apêndice A).

Número de sons	Regra de formação
1	C2 a B6
2	1a nota: n 2a nota: $n + 7s$
3	1a nota: n 2a nota: $n + 4s$ 3a nota: $n + 7s$
4	1a nota: n 2a nota: $n + 4s$ 3a nota: $n + 7s$ 4a nota: $n + 11s$

Sinal contendo apenas uma F0

Nesta seção serão apresentados os resultados referentes aos testes feitos quando o sinal de entrada apresenta apenas uma F0. A configuração da DFT foi a mesma do exemplo ilustrativo. Uma identificação é dada como correta quando o sistema acerta tanto a nota quanto a oitava. Os testes foram realizados para as três intensidades disponíveis: suave, média e forte (que são identificadas, na base MUMS, como *soft*, *medium* e *loud*, respectivamente).

A Tabela 4.2 resume os resultados encontrados. Para as três intensidades, os resultados ficaram em pelo menos 86% de acerto. Os erros encontrados estavam localizados nas notas com F0 dentro das segunda e sexta oitava do piano, sendo em sua grande maioria erros de oitava. Os erros na segunda oitava se deram basicamente porque as notas apresentavam pouca energia devido ao comprimento longo dos filtros contidos nesta oitava. As notas erroneamente identificadas na oitava 6 eram identificados como sendo da oitava 5. Isso porque as notas presentes nesta oitava apresentam poucos harmônicos, já que esta é a última oitava a ser considerada no conjunto de filtros auditivos empregado. Assim, os modelos definidos para as

notas da oitava anterior acabam por apresentar maior energia, resultando em erro.

Tabela 4.2: Resultados obtidos nos testes quando o sinal de entrada apresenta apenas uma F0.

Intensidade	Identificação	Número de erros / Total de casos
suave	88,3 %	7/60
média	86,7 %	8/60
forte	90,0 %	6/60

Mistura contendo mais de uma F0

Nesta seção serão apresentados os resultados referentes aos testes feitos quando o sinal de entrada apresenta mais de uma F0 (mais especificamente, com 2, 3 ou 4 notas, como explicado anteriormente). Os testes foram realizados para as três intensidades disponíveis, sendo que a mesma dinâmica estava presente dentro de um intervalo ou acorde: suave, média e forte (que são identificadas, na base MUMS, como *soft*, *medium* e *loud*, respectivamente).

A Tabela 4.3 resume os resultados encontrados. Há uma tendência em se obter resultados piores à medida que se aumenta o número de notas presentes no sinal de entrada.

Tabela 4.3: Resultados obtidos nos testes quando o sinal de entrada apresenta mais de uma F0.

Intensidade	Nº de Notas	% Identificação	Número de erros / Total de casos
suave	2	70,0 %	36/120
	3	60,0 %	72/180
	4	56,7 %	104/240
média	2	70,0 %	36/120
	3	60,6 %	71/180
	4	58,8 %	99/240
forte	2	67,5 %	39/120
	3	63,9 %	65/180
	4	63,8 %	87/240

Parte dos problemas encontrados nesses casos decorrem dos erros vistos para o caso de haver apenas uma F0. Contudo, a grande dificuldade aqui é a sobreposição de parciais harmônicas. Isso é o que se tenta atacar com o método de subtração do espectro relativo à F0 identificada, mas ainda assim constitui a maior fonte de erros.

Outros erros cometidos pelo sistema nesses casos foram erros de oitavas: a nota propriamente dita era corretamente identificada, mas em outra oitava que não a correta. Este foi o erro mais freqüente. Essa confusão se deve principalmente ao fato de que, ao se aplicar o algoritmo para estimação da F0 predominante apresentado na Seção 4.3, o pico da F0 correta decorrente do espectro suavizado cai abaixo do pico de uma parcial.

Os resultados descritos aqui não estão inteiramente compatíveis com aqueles apresentados em [50], pois aqui não houve ajuste de parâmetros inerentes ao algoritmo sugerido por Klapuri.

4.5 Critério de parada

Em todos os testes feitos até aqui, a quantidade de F0s presentes era conhecida. Por isso, deve ser elaborada uma forma automática de se estimar essa quantidade. Como o método é iterativo, uma alternativa seria definir um critério de parada que, quando atendido, indicaria que a última F0 presente no resíduo foi detectada e nada resta a ser feito.

4.5.1 Determinação da figura de mérito

É necessário selecionar uma figura de mérito que, uma vez avaliada, permita decidir se se deve proceder ou não à próxima iteração do algoritmo. Os candidatos a figura de mérito são:

- Máximo peso global do resíduo: essa é a candidata indicada pelo trabalho original de Klapuri [51] e guarda alta correlação com o número de F0s a serem detectadas.

- Energia do espectro residual: essa é a candidata mais intuitiva. Equivale ao quadrado da norma 2 do espectro residual e pode ser calculada através do

cálculo da variância do vetor que representa esse espectro. De forma análoga, pode-se também considerar a norma 1, que é a soma dos módulos dos elementos, ou até mesmo a norma infinita, que é o valor máximo contido no vetor do espectro.

- Energia do espectro suavizado: poderia ser vista como a energia relativa ao espectro da última F0 detectada.

- Versão normalizada do máximo peso global: é similar à primeira figura de mérito, normalizada pelo valor do máximo peso global na primeira iteração.

- Versão normalizada da energia do espectro residual: é similar à segunda figura de mérito, normalizada pelo valor da energia do espectro residual na primeira iteração.

- Versão normalizada da energia do espectro suavizado: é similar à terceira figura de mérito, normalizada pelo valor da energia do espectro suavizado na primeira iteração.

Para eleger uma figura de mérito, foi feito o seguinte procedimento:

1. Refazer, usando-se apenas as notas de intensidade média, o teste para 1, 2, 3 e 4 notas presentes no sinal de entrada, mas informando ao sistema que existem 8 notas. Com isso, ter-se-á informação acerca das candidatas à figura de mérito ao longo de 8 iterações;

2. Para cada iteração, registrar o valor de F0 relativo a cada uma das candidatas;

3. Analisar a eficiência de cada uma em reconhecer se existem F0s a serem detectadas. Isso será avaliado através de uma curva *receiver operating characteristic* ou ROC (ver [36]) verificando-se o ponto que estiver mais próximo ao ponto ótimo onde se tem 100% de detecção e 0% de falso alarme;

4. Selecionar aquela que apresentar a maior eficiência.

O estimação do limiar será feito apenas com base nas notas relativas à intensidade média; com isso, os sons com esta intensidade terão a função de atuar como conjunto de treinamento, enquanto que o conjunto de teste será formado pelos sons

com intensidade suave e forte. A Figura 4.12 mostra as curvas ROC tendo como parâmetro do detector cada uma das candidatas a figura de mérito sobre a qual será avaliado o critério de parada. O ponto ótimo do gráfico é aquele que maximiza a taxa de detecção e minimiza a taxa de falso alarme em relação ao reconhecimento do melhor instante para se parar as iterações, ou seja, o canto superior esquerdo. Quanto mais uma curva se aproximar desse ponto, maior é a sua eficiência para indicar a presença de F0s. Portanto, a melhor curva dentre as que foram avaliadas é a de energia do espectro suavizado, pois é a que apresenta menor distância ao ponto ótimo.

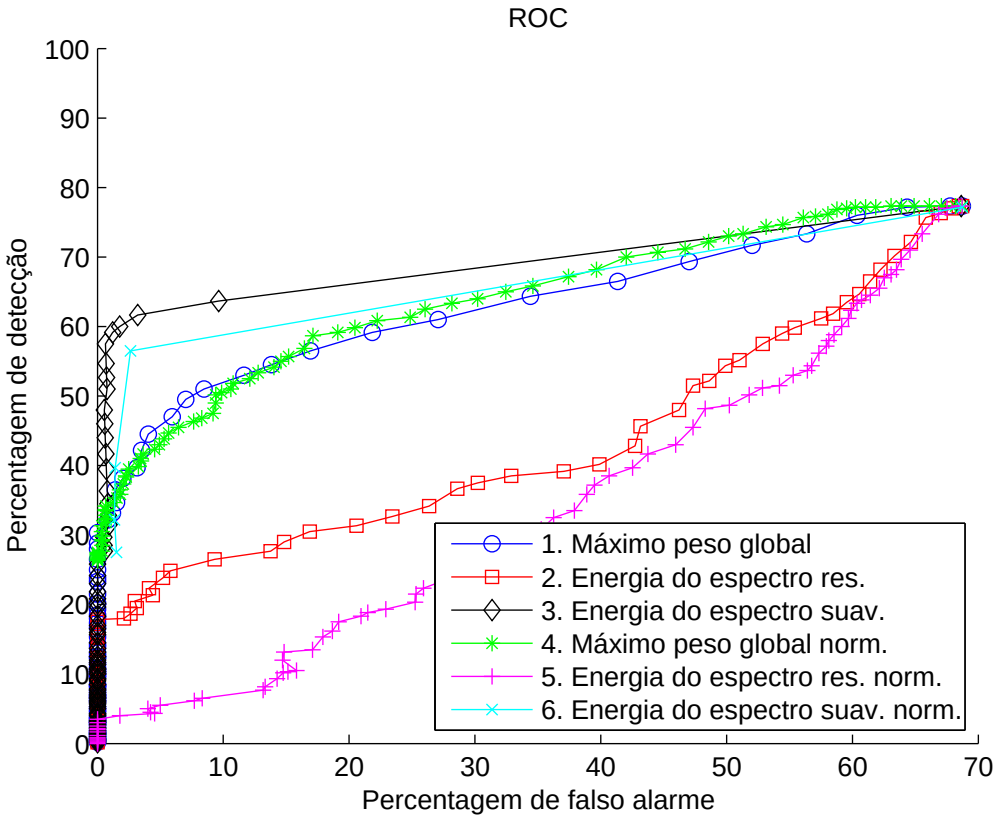


Figura 4.12: *Receiver operating curves* relativas a cada uma das candidatas à figura de mérito para avaliação do critério de parada do algoritmo iterativo de detecção de F0s. Os valores de percentagem de detecção e de falso alarme são relativos ao correto reconhecimento de que ainda existem F0s presentes na mistura. O parâmetro que se varia para formar as curvas é o valor do limiar de detecção para cada figura de mérito. A faixa de variação foi do máximo ao mínimo valor possível de cada figura de mérito.

4.5.2 Determinação do limiar

Tendo-se estabelecido qual deve ser a figura de mérito sobre a qual é avaliado o critério de parada, deve-se definir o valor para o limiar que indicará a presença ou não de F0s ainda por serem detectadas. O valor desse limiar terá grande efeito sobre as taxas de detecção e falso alarme. Já que existe um compromisso entre essas duas taxas, uma forma para se determinar o limiar poderia ser observar a curva ROC e selecionar algum valor que maximizasse a taxa de detecção sem elevar muito a taxa de falso alarme. Se a aplicação for a ressíntese do sinal analisado, segundo Klapuri [51], a melhor opção é minimizar o falso alarme, mesmo que isso resulte em reduzir a taxa de detecção. Isso se justifica pelo fato de que, sob o prisma da psicoacústica, um som falsamente detectado pode causar um desconforto maior do que a perda de um som que realmente exista na mistura. Se o objetivo for a transcrição musical automática, essa decisão dependerá do usuário. Pode-se optar por ter maior taxa de detecção, mesmo que isso implique maior taxa de falso alarme, já que seria preferível descartar notas espúrias a tentar transcrever manualmente o que o sistema não foi capaz de detectar.

Uma outra taxa que se pode considerar na determinação do valor do limiar de detecção é a de perda, que é calculada como a razão entre o número de notas que não foram detectadas pelo sistema devido a uma interrupção precoce do número de iterações e o número total de notas de referência. Analogamente, a taxa de falso alarme é calculada como a razão entre o número de notas que foram erroneamente detectadas pelo sistema devido a uma interrupção tardia do processo iterativo e o número total de notas de referência. Por fim, a taxa de substituição é a razão entre as notas corretamente detectadas, mas incorretamente identificadas, e o número total de notas de referência. Assim, somando-se a taxa de detecção, a de perda e a de substituição, chega-se a 100%. Pode-se estabelecer um compromisso envolvendo as taxas de falso alarme e de perda, escolhendo-se um valor para o limiar de detecção que garanta que essas duas taxas sejam as mais próximas possíveis.

O procedimento utilizado para se estimar esse limiar é:

1. Para todas as experiências de todos os casos de polifonia (1 a 4), identificar 8 notas.

2. Constituir uma matriz na qual cada linha é uma experiência. Cada uma das 8 colunas terá zeros e uns. Os zeros indicam notas incorretamente detectadas ou identificadas enquanto que os uns constituem detecções e identificações corretas.

3. Com base na figura de mérito selecionada, variar o valor do limiar de detecção. Cada valor de limiar indicará, para cada linha, em que coluna o algoritmo deve parar.

4. Verificar quantas notas corretas são identificadas, armazenando os valores para taxas de identificação de notas; quantas notas ficam depois do momento da parada, constituindo a taxa de perda; e quantas notas incorretas existem antes do momento da parada, totalizando a taxa de falso alarme.

5. Através da curva ROC, verificar qual valor de limiar gera a taxa de perda igual à de falso alarme.

Sob esse critério, o valor calculado para ambas as taxas foi de 10%, o que produz, segundo a curva ROC em questão, uma taxa de identificação de 64%.

4.5.3 Resultados

Refazendo-se os testes acoplando-se esse módulo para avaliação do critério de parada, obtêm-se os resultados quanto à identificação de F0s resumidos na Tabela 4.4. É importante lembrar que os valores calculados para as taxas de falso alarme e de perda foram estimados com base em todos os casos de polifonia juntos. Assim, ao se considerar todos os eventos de falso alarme e se dividir pelo total de notas de referência, obter-se-á um valor mais próximo ao que se obtém ao dividir o número de eventos perdidos pelo total de notas.

Observando-se a Tabela 4.4, percebe-se que, para quantidades pequenas de notas presentes no sinal de entrada, o falso alarme é bastante alto e a taxa de perda é baixa, enquanto que para quantidades mais altas de notas no sinal de entrada a situação se inverte. Isso é justificado pelo fato de que o limiar de detecção do critério de parada é calculado a partir de todas as ocorrências (todos os casos esperados de polifonia).

Como previsto, os resultados obtidos levando-se em consideração o critério de parada são quantitativamente inferiores aos da situação anterior (onde se sabia *a*

priori o número de notas na mistura). Quando o momento da parada é subestimado, ocorre perda de notas, o que é refletido na redução da taxa de detecção. Quando este momento é superestimado, ocorre falso alarme, que diminui a confiabilidade do sistema. Contudo, só agora, com a definição de um critério de parada, é possível levar o sistema para situações reais de uso, permitindo que ele faça parte do sistema de TMA.

Tabela 4.4: Resultados obtidos nos testes utilizando-se o critério de parada.

Intensidade	Nº de Notas	Identificação	Falso Alarme	Substituição	Perda
suave	1	91,7 %	47,8 %	8,3 %	0,0 %
	2	69,2 %	26,9 %	25,8 %	5,0 %
	3	65,3 %	27,0 %	24,5 %	10,2 %
	4	60,8 %	17,3 %	32,9 %	17,3 %
forte	1	95,0 %	65,5 %	5,0 %	0,0 %
	2	76,7 %	47,1 %	21,7 %	1,7 %
	3	67,2 %	28,6 %	31,1 %	1,7 %
	4	67,1 %	20,6 %	30,8 %	2,1 %

4.6 Complexidade

O sistema para identificação de múltiplas F0s descrito por Klapuri atua sobre o espectro de uma janela do sinal de entrada. Portanto, a complexidade do sistema deverá ser calculada como função número de canais desse espectro, definido como N.

Cada etapa do sistema deverá, então, contribuir com uma parcela da complexidade. Por complexidade, entende-se o número de adições N_A, de multiplicações N_M e de exponenciações N_E. Cada parcela também será função de uma série de constantes. Estas são o número de sub-bandas B e o número de possíveis valores para o fator de inarmonicidade, u.

Para a etapa de pré-processamento (sem o cálculo do espectro), têm-se as seguintes sub-etapas:

1. Calcular a raiz cúbica do espectro de entrada: $N_E = N$

2. Calcular g: $N_A = N$

3. Estimar espectro do ruído: $N_A : B(N/B) = N$, onde foi feita uma aproximação considerando cada sub-banda como tendo o mesmo número de canais.

4. Subtração espectral: $N_A = N$

Portanto, o pré-processamento tem complexidade de $N_A = 3N$ e $N_E = N$.

Para a análise em sub-bandas, têm-se as seguintes sub-etapas (que devem ser feitas para cada sub-banda):

1. Calcular sub-banda: $N_M = N$

2. Seleção harmônica: $N_A = (N/B)n\frac{N/B}{n} = \left(\frac{N}{B}\right)^2$. O valor considerado para o cálculo foi a média dos possíveis valores. O número médio de amostras de uma sub-banda é N/B. O somatório das cerca de $\frac{N/B}{n}$ amostras de cada sub-banda feito para cada valor de m. O valor de m varia de 1 a n, onde n é o índice associado à F0. Isso deve ser feito para todas as amostras do espectro até aquela que representa a largura da banda em questão, o que, na média, representa N/B amostras.

Portanto, a análise em sub-bandas tem complexidade de $N_A = \sum_{b=1}^{B}\left(\frac{N}{B}\right)^2 = \frac{N^2}{B}$ e $N_M = BN$.

Para a integração das sub-bandas, tem-se apenas que executar um somatório (que deve ser feito para cada um dos N elementos do espectro e para cada um dos elementos de $\mathbf{u}$). Portanto, a integração das sub-bandas tem complexidade de $N_A = L_u BN$, onde L_u é o número de elementos do vetor $\mathbf{u}$.

A etapa de estimação da F0 predominante, segundo o artigo original de Klapuri [50], apresenta uma complexidade desprezível, visto que os cálculos são feitos apenas para um peso e não para toda a matriz de pesos, como é feito nas etapas de análise e integração.

Para a etapa de suavização espectral, tem-se apenas que calcular a média ponderada das amostras do espectro dentro de cada uma das janelas definidas através das sub-bandas do modelo auditivo. O refinamento do valor do fator de inarmonicidade não tem grande influência na complexidade final da etapa. Portanto, a complexidade computacional da suavização espectral é de $N_A = 2BN$ e $N_M = BN$. A etapa da subtração espectral corresponde a $N_A = N$.

Totalizando, tem-se a complexidade total do sistema de identificação de F0 descrito por Klapuri como sendo:

$$\begin{aligned} N_A &= 3N + \frac{N^2}{B} + uBN + 2BN + N, \\ &= \frac{N^2}{B} + N(4 + B(u + 2B)), \end{aligned} \tag{4.17}$$

$$N_M = BN + BN = 2BN, \tag{4.18}$$

$$N_E = N. \tag{4.19}$$

4.7 Conclusão

Nesse capítulo, foi apresentado o sistema de identificação de F0 com base no método de Klapuri [50]. Foi feito um estudo sobre o desempenho do sistema. Por fim, foi exposto um critério para se definir o momento de parada das iterações do algoritmo e foi feita uma análise da complexidade do método.

Ao se informar a quantidade de F0s ao sistema, obteve-se a taxa de identificação máxima de cerca de 90% (quando existia apenas uma F0 no sinal de entrada) e a taxa de identificação mínima de cerca de 56% (quando existiam 4 F0s no sinal de entrada). Ao se empregar o critério de parada, a taxa de identificação máxima chegou a 95% e a mínima foi em torno de 67%. Como conseqüência do uso do critério de parada, o falso alarme foi de no máximo 65%, enquanto que a substituição foi de no máximo 39% e a perda máxima foi de 17%. Os valores relativos a erros foram elevados devido ao fato de que o conjunto de treino não representava com fidelidade o conjunto de teste, por não haver muitos sinais disponíveis. Melhores resultados seriam obtidos se ambos os conjuntos apresentassem sons com as três intensidades possíveis. Os resultados descritos aqui não estão inteiramente compatíveis com aqueles apresentados em [50], pois aqui não houve ajuste de parâmetros inerentes ao algoritmo sugerido por Klapuri.

O método desenvolvido por Klapuri e descrito neste capítulo é perfeitamente passível de ser melhorado utilizando-se uma ferramenta de análise espectral diferente da FFT (que é seu fundamento). Como a FFT apresenta um espaçamento linear entre seus canais, é necessário que estes sejam em grande quantidade para que possam cobrir todo o espectro de interesse efetivamente. Uma ferramenta com espaçamento não-linear é, provavelmente, mais apropriada para aplicações musicais.

Além disso, os filtros de FFT apresentam baixa seletividade, o que pode acarretar interferência entre os canais e maior dificuldade na distinção das notas executadas. Logo, uma ferramenta baseada em filtros mais seletivos poderia ser favorável a uma estimação mais acurada das notas.

Capítulo 5

Banco de filtros para análise espectral

5.1 Introdução

Esse capítulo visa a descrever ferramentas para análise espectral de sinais de música que são caracterizadas por pelo menos uma das seguintes características: filtros de alta seletividade, espaçamento freqüencial geométrico ou quase geométrico (que tendem a ser mais adequados para sinais musicais) e complexidade computacional que não comprometa a viabilidade de sua utilização. A Seção 5.2 faz uma revisão das ferramentas disponíveis na literatura padrão para análise espectral. A Seção 5.3 descreve métodos de análise espectral com espaçamento freqüencial linear, enquanto que a Seção 5.4 mostra métodos com espaçamento freqüencial geométrico. A seguir, os métodos com espaçamento freqüencial linear por partes são explicados na Seção 5.5 e as questões sobre a implementação prática, na Seção 5.6. Por fim, experimentos computacionais para avaliar o desempenho dessas ferramentas são descritos na Seção 5.7 e o capítulo é concluído na Seção 5.8.

A maior parte do texto deste capítulo consta, com algumas adaptações, nas referências [P1], [P4], [P2] e [P3]. Estes trabalhos apresentam algumas das contribuições originais descritas no presente texto.

5.2 Análise espectral

A ferramenta padrão para análise espectral é a transformanda rápida de Fourier (*fast Fourier transform*, FFT), que é o algoritmo rápido para a transformada discreta de Fourier (*discrete Fourier transform*, DFT). A FFT é largamente utilizada em diversas aplicações devido à sua simplicidade [54]. Considerando a FFT como um banco de filtros, pode-se interpretar que tal simplicidade decorre, parcialmente, do uso de filtros *kernel*[1] de baixa ordem, o que resulta em canais de baixa seletividade, o que será mais detalhado na Seção 5.3.

Uma tentativa de se resolver esse problema foi feita por Lim e Farhang-Boroujeny [4], que tomaram a estrutura em árvore da FFT trocando apenas os filtros *kernel* por filtros de ordem mais alta, chegando, então no *fast filter bank* (FFB). A complexidade do FFB é ligeiramente mais alta do que a da FFT, mas permite que a seletividade dos canais seja bem mais alta.

Os canais do FFB e da FFT são uniformemente distribuídos ao longo das freqüências, o que significa que todos os canais apresentam a mesma largura de banda, independente de suas freqüências centrais. Dependendo da aplicação em questão, essa disposição, mostrada na Figura 5.1(a), pode não ser eficiente para sinais de música, devido ao igual temperamento empregado na música ocidental [55]. Para aproveitar melhor o fato de as notas usadas no temperamento igual são geometricamente espaçadas no domínio da freqüência, Brown [56] criou, com base na DFT, a *constant-Q transform* (CQT), na qual a largura de banda δf de cada canal varia proporcionalmente em relação à sua freqüência central f_0 (como visto na Figura 5.1(b)), mantendo assim seu fator de qualidade $Q = f_0/\delta f$ constante. Ao se considerar o problema de identificação de notas musicais, a CQT se mostra como uma representação espectral mais apropriada, devido ao fato de seus canais serem geometricamente espaçados.

Mesmo a implementação rápida da CQT [57] apresenta uma complexidade computacional extremamente elevada, se comparada com a da FFT. A resposta para essa questão foi aproximar o eixo com espaçamento geométrico por um que tenha um

[1]Filtros *kernel* são os filtros que servem de base para a definição dos filtros protótipos, os quais são calculados realizando-se modulações na freqüência (para se especificar sua localização no eixo freqüencial) e decimações (para definir a largura de sua banda).

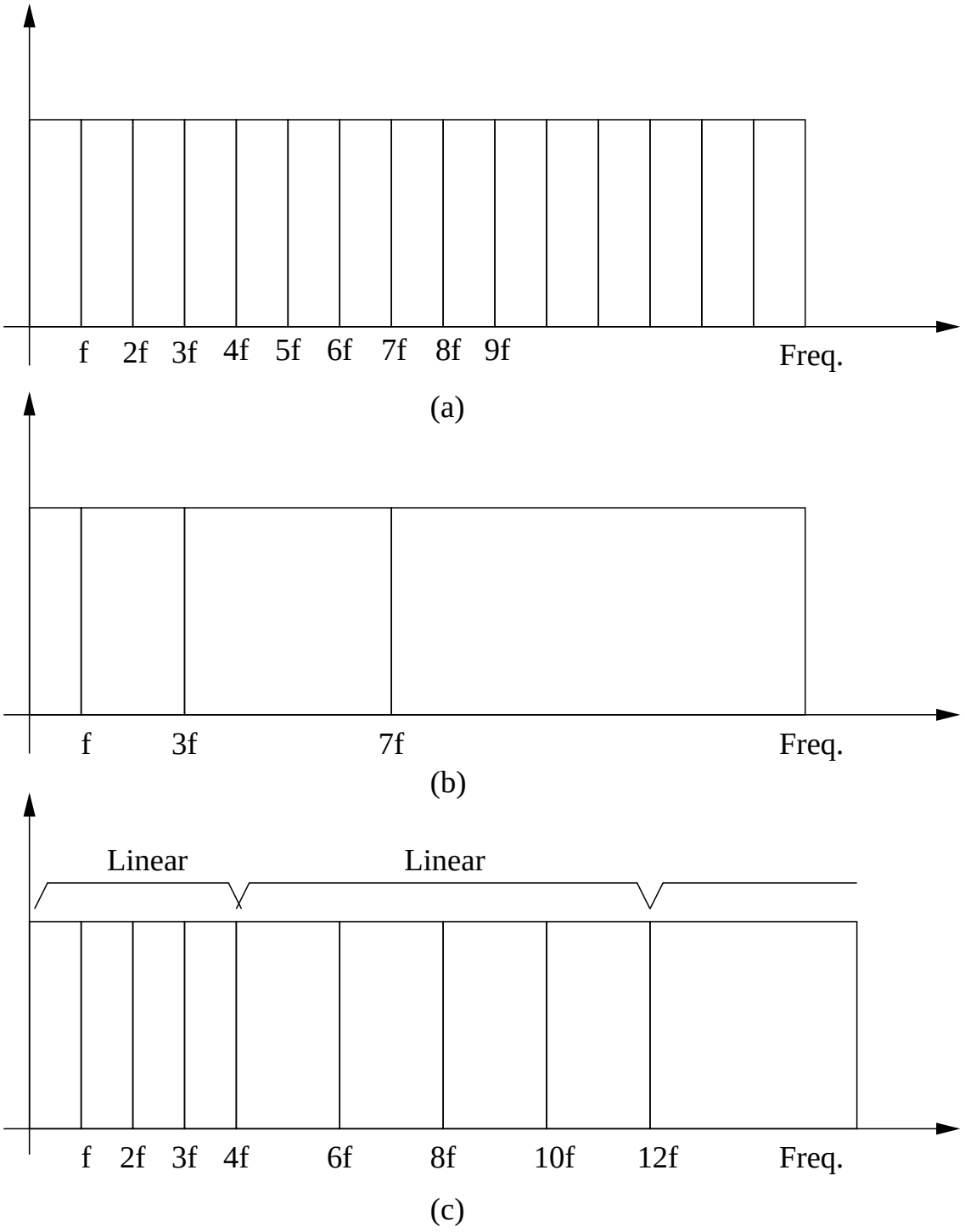


Figura 5.1: Métodos para análise espectral de sinais de música: (a) espaçamento freqüencial linear; (b) espaçamento freqüencial geométrico; (c) espaçamento freqüencial linear por partes.

espaçamento linear por partes, o que foi proposto na *bounded-Q transform* (BQT) [5], também baseada na DFT. Essa abordagem, em que o espectro é dividido em oitavas e os canais dentro de uma oitava são linearmente espaçados, é exibida na Figura 5.1(c).

As ferramentas anteriores não são capazes de combinar todas as características desejáveis para a análise espectral de sinais de áudio:

- Alta seletividade para cada canal, individualmente;
- Distribuição eficiente de canais na freqüência, no sentido de aproveitar melhor o espaçamento das freqüências que podem estar contidas no sinal sob análise;
- Complexidade computacional reduzida.

Para esse fim, o *constant-Q fast filter bank* (CQFFB) e o *bounded-Q fast filter bank* (BQFFB) foram desenvolvidos. O CQFFB [58][59] pode ser visto como uma versão de alta seletividade da CQT, da qual herda sua alta complexidade computacional. A seguir, o BQFFB é introduzido como a ferramenta de maior eficiência, combinando a complexidade reduzida herdada da FFT, a distribuição quase-geométrica da BQT e a alta resolução do FFB. O conceito original da BQFFB foi apresentado em [P1], mas tal ferramenta foi aperfeiçoada em [P4], com uma implementação que evitava filtros decimadores.

5.3 Métodos com espaçamento freqüencial linear

5.3.1 *Fast Fourier transform*

A *short-time DFT* é definida por

$$X(p,k) = \frac{1}{N}\sum_{n=0}^{N-1} w(n)x(n-p)e^{-j2\pi kn/N}, \tag{5.1}$$

onde $x(n)$ é a n-ésima amostra do sinal de entrada, $2\pi k/N$ é a freqüência digital normalizada em radianos por amostra (a duração do sinal é de N amostras), $0 \leq k \leq (N-1)$ é o índice da raia, e $w(n)$ é uma função janela, como a janela de Hamming [60]. Ao deslocar uma janela retangular $w(n)$ ao longo de $x(n)$ em saltos de S amostras, considera-se a DFT como uma transformada em blocos. O índice p

denota a posição da janela. A FFT é a família de algoritmos rápidos para DFT [54]. Essa rapidez decorrente da sua baixa complexidade é a responsável por sua grande utilização [61]. O tipo mais popular é a *radix-2* FFT, que é baseada numa simples estrutura treliça modular.

Fazendo-se $S = 1$, chega-se à *sliding* FFT (sFFT), que pode ser vista como um banco de filtros de N canais [3], com $N = 2^L$, onde L é um inteiro, organizado sob a forma de uma estrutura em árvore, como na Figura 5.2. O filtro relativo a cada canal é composto por uma cascata de L sub-filtros. Em cada nó do diagrama, encontra-se um filtro juntamente com seu complementar. Cada filtro é complexo e representa uma versão modulada e interpolada de um mesmo filtro protótipo

$$H(z) = 1 + z^{-1}, \tag{5.2}$$

com apenas dois coeficientes não nulos. Um filtro $H_{l,b}(z)$ é construído substituindo-se z em $H(z)$ por [4]

$$W_N^{-\tilde{b}} z^{2^{L-l-1}} = \left(e^{-j2\pi/N}\right)^{\tilde{b}} z^{2^{L-l-1}}, \tag{5.3}$$

onde $l = 0, ..., (L-1)$ é o índice do nível, $b = 0, ..., (2^l - 1)$ é o índice do filtro dentro de cada nível, e $\tilde{b}$ é a representação com reversão de *bits* do inteiro b. Assim, cada nó divide o espectro em duas partes de igual largura. A reordenação dos canais de saída é decorrente do algoritmo da FFT, que permite utilizar apenas metade do número de multiplicações.

A configuração de filtragem como um todo pode ser vista na Figura 5.3, a qual ilustra como o filtro do canal 0 é formado numa sFFT de 8 canais.

Pode-se mostrar que a complexidade da FFT é $N \log_2 N$ multiplicações complexas para uma seqüência de entrada de comprimento N [60], se nenhuma outra simplificação é levada em consideração. Contudo, a sFFT descrita anteriormente, por sua vez, requer $C_{\mathrm{FFT}} = 1$ multiplicação complexa por amostra de entrada por canal [62].

5.3.2 *Fast filter bank*

Visando a evitar o *trade-off* entre rejeição no lobo secundário e largura do lobo principal que é inerente à solução da FFT janelada, Lim e Farhang-Boroujeny [4]

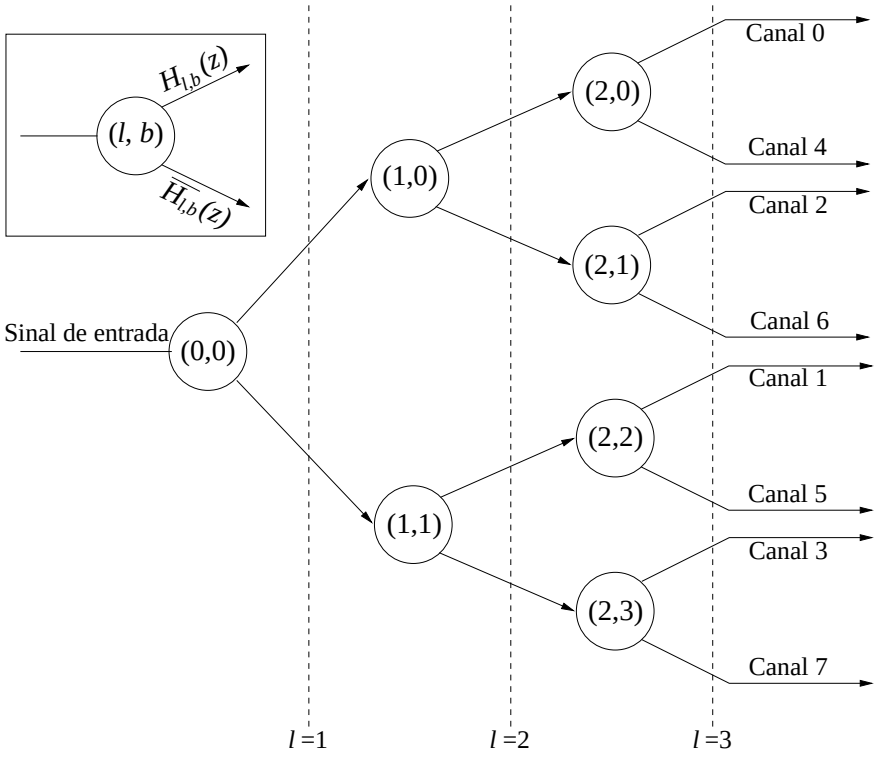


Figura 5.2: Representação em árvore para a sFFT e o FFB, permitindo a ambos os algoritmos apresentar uma implementação modular rápida. Cada nó no diagrama é composto por um par de filtros (protótipo e complementar) para gerar os sinais que servirão de entrada para os filtros do próximo nível.

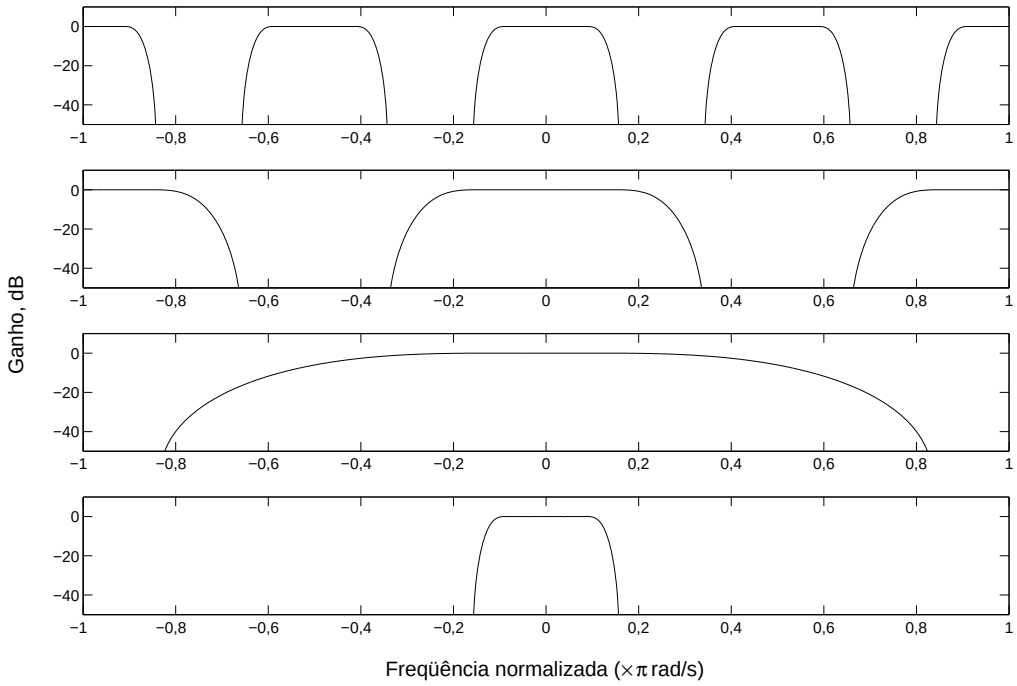


Figura 5.3: Construção do filtro do canal 0 em uma sFFT ou em um FFB de 8 canais, a partir de versões modificadas dos filtros *kernel*. De cima para baixo, os gráficos mostram a resposta em freqüência hipotética dos filtros protótipos (0, 0), (1, 0) e (2, 0), e do filtro do canal 0 resultante (ver Figura 5.2).

propuseram a associação da estrutura em árvore da FFT com filtros *kernel* mais longos. A idéia é tirar proveito da implementação modular da FFT e usá-la com filtros com transição estreita entre banda de passagem e banda de rejeição. O conceito por trás dessa idéia é a abordagem conhecida como *frequency response masking* (FRM) [63]. Utilizando-se este mesmo conceito, considerou-se também trabalhar com banco de filtro de cosseno modulado (CMFB) e a sua versão de alta seletividade (FRM-CMFB). Contudo, optou-se pelo *fast filter bank* por apresentar uma complexidade reduzida, já que o FFB aproveita a estrutura em árvore da FFT.

O FRM tem como objetivo projetar filtros digitais de banda de transição bem estreita com baixa complexidade. Seu ponto de partida é a observação de que a resposta em freqüência de um filtro interpolado na forma $H(z^L)$ é composta por réplicas periódicas da resposta em freqüência de $H(z)$ comprimidas por um fator igual a L. Cada réplica apresenta uma banda de transição L vezes mais estreita que a de $H(z)$. Um filtro $G(z)$ (que não precisa ser tão seletivo) pode ser projetado para suprimir as réplicas indesejáveis. Já que o número de coeficientes não-nulos de $H(z^L)$ é L vezes menor que sua ordem e as especificações de $G(z)$ não precisam ser muito restritivas, o filtro resultante apresenta uma complexidade muito baixa. O projeto como um todo é feito através de procedimentos de otimização devidamente selecionados.

O banco de filtros da FFT descrito na Seção 5.3.1 (ver Figuras 5.2 e 5.3) é estruturalmente adequado para o projeto de FRM, já que tem como base filtros interpolados em cascata. A principal modificação necessária para transformar os canais originais da FFT em filtros de alta seletividade é empregar filtros *kernel* de ordens mais altas em cada nível l da estrutura no lugar do filtro *kernel* (único) de ordem baixa da FFT, dado pela Eq. (5.2). Um procedimento derivado do FRM pode ser utilizado para gerar os filtros necessários para cada parte da estrutura, de modo que cada filtro interpolado é mascarado pelos filtros subseqüentes em cascata. A técnica resultante é denominada *fast filter bank* (FFB), sobre o qual se encontram mais detalhes em [64] e [65]. Assim, as características de fase linear da estrutura da FFT são mantidas, evitando distorções de fase no sinal. É importante notar que, assim como os filtros da FFT, os filtros do FFB também serão complexos.

No presente trabalho, a menos que se especifique explicitamente de outra forma,

o FFB segue as mesmas especificações que em [4], mantendo assim as mesmas ordens para os filtros, como mostrado na Tabela 5.1. A Figura 5.4 faz uma comparação entre a seletividade do FFB e a da FFT. Considerando o nível de rejeição mínimo como no mais alto lobo secundário em cada caso, os filtros da FFT apresentam uma rejeição de cerca de 13 dB entre canais adjacentes [4], enquanto que os filtros do FFB, em sua forma original especificada em [4] (cuja ordem pode ser vista na Tabela 5.1), apresentam uma atenuação de cerca de 56 dB.

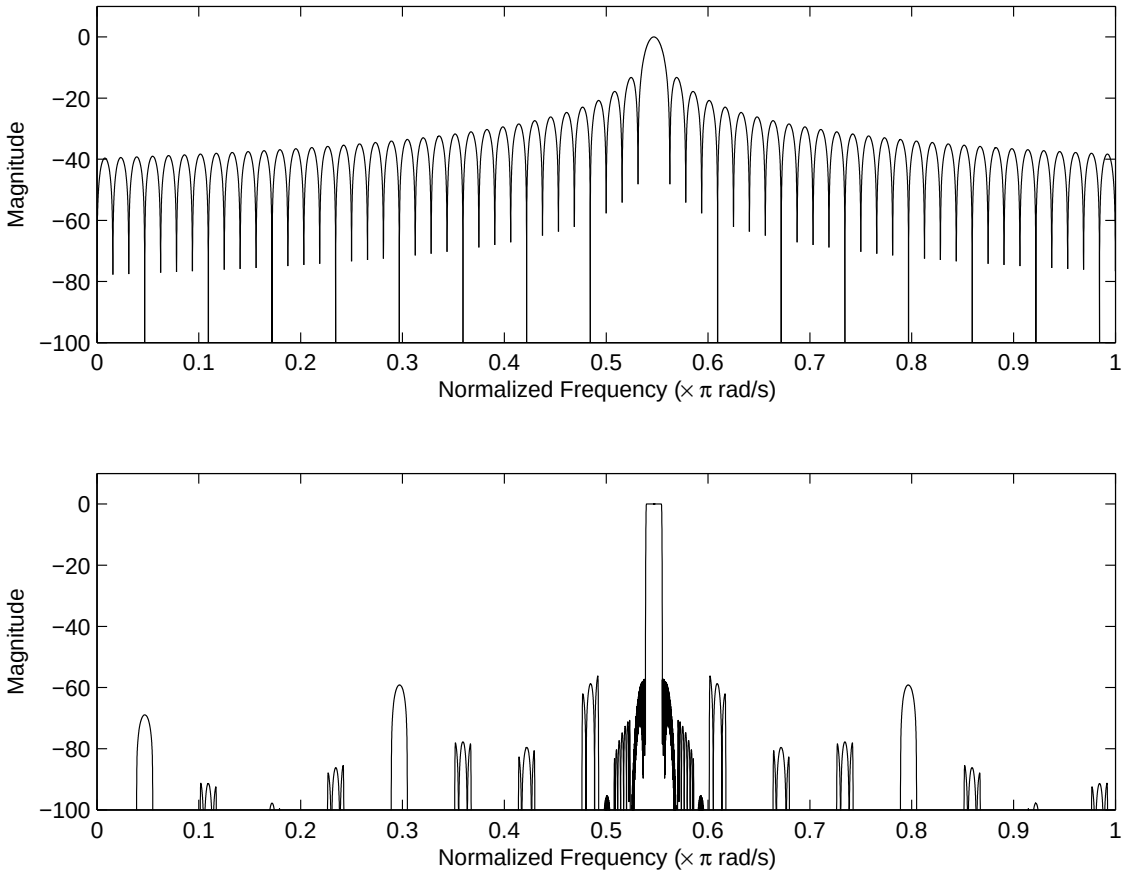


Figura 5.4: Resposta em freqüência do canal 35 de um banco de filtros de 128 canais: (a) FFT; (b) FFB.

O número de coeficientes do filtro por nível da FFB é apresentado na Tabela 5.1. Pode-se ver que a quantidade acumulada de coeficientes distintos e não-nulos até um dado nível na cascata $l \leq 5$ é $(2N + 23)$. Isso leva a um número de multiplicações complexas por canal por amostra de entrada igual a

$$C_{\mathrm{FFB}}(l) = C(l) = \frac{2N + 23}{N} \approx 2, \tag{5.4}$$

onde N é o número de canais da FFB. Isso é aproximadamente duas vezes o custo

computacional da *radix-2* FFT. Assim, um grande aumento na seletividade é obtido pelo FFB em troca de um ligeiro aumento na complexidade.

Tabela 5.1: Número de coeficientes não-nulos por nível da estrutura de sub-filtros do FFB.

Nível da cascata (l)	Coeficientes distintos por filtro	Filtros protótipos	Coeficientes por nível	Coeficientes acumulados $C(l)$
1	7	1	7	7
2	6	2	12	19
3	3	4	12	31
4	3	8	24	55
5	2	16	32	87
6	2	32	64	151
7	2	64	128	279
8	2	128	256	535
⋮	⋮	⋮	⋮	⋮
$\log_2 N$	2	$N/2$	N	$2N + 23$

É importante lembrar que as ferramentas com espaçamento linear na freqüência descritas até aqui apresentam uma resolução freqüencial constante ao longo do espectro. A seções seguintes lidam com ferramentas de resolução variável.

5.4 Métodos com espaçamento freqüencial geométrico

Apesar da alta seletividade do FFB, seus canais ainda são distribuídos uniformemente ao longo das freqüências, como na FFT. Entretanto, as freqüências das notas presentes na música ocidental (com o temperamento igual, ver Apêndice A) são ge-

ometricamente espaçadas [55]. Assim, notas mais graves são mais próximas entre si em Hz do que as notas mais agudas. Como conseqüência, na análise espectral de sinais de música, se o espaçamento entre canais for linear, uma resolução suficiente para distinguir notas mais graves implicaria uma resolução freqüencial desnecessariamente alta para as notas mais agudas, enquanto que uma boa resolução para as mais agudas seria insuficiente para as mais graves. Pode-se, então, contornar esse problema distribuindo-se geometricamente os canais ao longo das freqüências, empregando-se um número reduzido de canais.

O objetivo da *constant-Q transform* (CQT) [56], que tem como base a DFT, é fornecer esse espaçamento geométrico na freqüência. Isso é obtido variando-se o espaçamento e a banda dos canais de forma proporcional à freqüência central de cada canal, fazendo com que a razão entre esses dois valores permaneça constante. Dado o número de canais por oitava, pode-se definir esse fator de qualidade constante como

$$Q = \frac{f(k)}{\delta f(k)}, \tag{5.5}$$

onde $f(k)$ é a freqüência do k-ésimo canal e $\delta f(k)$ é o espaçamento entre os canais k e $(k+1)$. Assim, à medida que $f(k)$ aumenta geometricamente, um Q constante é obtido por um aumento similar na largura de banda $\delta f(k)$, de modo que os filtros preencham todo o espectro.

Na Eq. (5.1), atribuir um valor fixo a Q é equivalente a escolher uma janela de comprimento diferente para cada componente espectral, transformando N em

$$N_k = \frac{f_\mathrm{s}}{\delta f(k)} = \frac{f_\mathrm{s}}{f(k)} Q, \tag{5.6}$$

onde f_s é a freqüência de amostragem.

Com as definições acima, é possível obter a expressão para o k-ésimo componente espectral da CQT

$$X_{\mathrm{C}Q}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} w_k(n) x(n) e^{-j2\pi kn/N_k}. \tag{5.7}$$

Um discussão detalhada sobre a escolha de uma janela $w_k(n)$ adequada para cada canal k pode ser encontrada em [56]. Mas um ponto importante a ser levantado sobre esse tipo de transformada é que ela não permite inversão, o que faz com que não haja reconstrução perfeita.

5.4.1 *Constant-Q fast filter bank*

O CQFFB [58] combina a alta seletividade do FFB com o conceito do fator de qualidade Q constante da CQT. A idéia é empregar a distribuição freqüencial da CQT no espaçamento dos filtros do banco. O comprimento variável da janela N_k da CQT é substituído pelas larguras variáveis dos filtros. As freqüências das raias da CQT se tornam as freqüências centrais dos filtros correspondentes do CQFFB, enquanto que a distância entre duas raias adjacentes da CQT é substituída pela largura de um filtro do CQFFB. Naturalmente, a melhora na seletividade implica um aumento no custo computacional.

A seguir, duas implementações diferentes do CQFFB serão apresentadas. A primeira consiste no seguinte procedimento:

1. De posse do valor para Q para se obter o nível necessário de detalhamento freqüencial, projeta-se um FFB com um número L de canais escolhido de forma a satisfazer $N = 2^L \leq 2Q$; toma-se, então, o filtro correspondente ao canal $\lceil Q \rceil$;

2. Para cada canal k do CQFFB,

   (a) Reamostra-se o sinal de entrada de forma que a nova freqüência de amostragem seja

   $$f_{\mathrm{s}}(k) = \frac{N}{Q} f_{\min} r^{k-1}, \tag{5.8}$$

   onde

   $$r = \frac{2 + 1/Q^2 + (1/Q)\sqrt{4 + 1/Q^2}}{2} \tag{5.9}$$

   é a razão entre freqüências centrais de canais contíguos e $f_{\min}$ é a freqüência central do canal de índice $k = 1$;

   (b) Submete-se a versão reamostrada do sinal de entrada ao filtro FFB escolhido no primeiro passo.

Reamostrar o sinal de entrada para $f_{\mathrm{s}}(k)$ faz com que a faixa de freqüência desejada do sinal de entrada se mova para a banda de passagem do filtro FFB escolhido. A principal desvantagem dessa abordagem é a necessidade de se executar

um elevado número de cálculos para realizar as reamostragens do sinal de entrada. Além disso, são ainda necessárias filtragens *anti-aliasing*. A complexidade para um dado canal k, em função do número de multiplicações complexas por amostra de entrada, é dada por

$$C_{\mathrm{C}Q\mathrm{FFB}}(k) = C_{\mathrm{R}}(k) + (C_Q + 1)\gamma(k), \tag{5.10}$$

onde $C_{\mathrm{R}}(k)$ é o custo da reamostragem, $\gamma(k)$ é o fator de reamostragem (ambos para o canal de índice k) e C_Q é o custo relativo ao filtro FFB selecionado no primeiro passo do procedimento.

Uma implementação alternativa visa a reamostrar os filtros ao invés de reamostrar o sinal de entrada [59]. O procedimento para isso é o seguinte:

1. De posse do valor para Q para se obter o nível necessário de detalhamento freqüencial, projeta-se um FFB com um número L de canais escolhido de forma a satisfazer $N = 2^L \leq 2Q$; toma-se, então, o filtro correspondente ao canal $\lceil Q \rceil$;

2. Para cada canal k do CQFFB,

   (a) Reamostra-se a resposta impulsional do filtro escolhido no primeiro passo de acordo com a Eq. (5.8);

   (b) Submete-se o sinal de entrada ao filtro modificado no passo anterior.

Reamostrar a resposta impulsional do filtro FFB selecionado para $f_{\mathrm{s}}(k)$ faz com que a banda de passagem da filtro se mova para a faixa de freqüência desejada do sinal de entrada. Isso faz com que a filtragem seja mais complexa, já que o banco de filtros perde uma importante característica dos filtros FFB: a grande quantidade de coeficientes nulos. Em contrapartida, os cálculos para obtenção dos filtros são feitos apenas uma vez, de forma *off-line*. A complexidade para um dado canal k é, então,

$$C_{\mathrm{C}Q\mathrm{FFB}}(k) = (C_Q + 1)\gamma(k). \tag{5.11}$$

As Eqs. (5.11) e (5.12) mostram que essa segunda implementação do CQFFB é menos custosa, visto que não inclui a parcela relativa à reamostragem, que é executada somente na primeira implementação. A complexidade como um todo é

dada por

$$C_{\mathrm{CQFFB,Total}} = \sum_{k=q_1}^{q2} \left(C_Q r^{-k} + 1 \right), \tag{5.12}$$

onde $q_1 = \left\lfloor \log_r \left(2^{-D}(N/2Q) \right) \right\rfloor$, $q_2 = \lfloor \log_r (N/2Q) \rfloor$, D é o número de oitavas e o operador $\lfloor \bullet \rfloor$ denota o maior inteiro menor que ou igual ao seu argumento.

Esse tipo de ferramenta pode ser muito útil, por exemplo, em transcrição automática de música para instrumentos perfeitamente afinados e de afinação fixa. Contudo, sua complexidade computacional é muitíssimo elevada, o que motivou o desenvolvimento dos métodos descritos a seguir.

5.5 Métodos com espaçamento freqüencial linear por partes

Para reduzir a alta complexidade inerente à CQT, a *Bounded-Q Transform* (BQT) foi proposta em [5]. Nessa ferramenta de análise, apenas as oitavas são geometricamente espaçadas, enquanto que dentro de cada oitava, as raias são linearmente espaçadas, como visto na Figura 5.1(c). Essa distribuição de raias pode ser considerada uma boa aproximação para uma distribuição puramente geométrica (se o número de canais por oitava for cuidadosamente escolhido).

Um método de *constant-Q* projetado para R canais por oitava dividiria uma oitava (começando na freqüência f_0) em bandas com larguras dadas por

$$\delta f_{CQ}(k) = f_0 \left[\left(\sqrt[R]{2} \right)^k - \left(\sqrt[R]{2} \right)^{k-1} \right], \tag{5.13}$$

onde $k = 1, ..., R$ são os índices dos canais. Analogamente, um método de *bounded-Q* projetado para $N = 2^L$ canais por oitava, com L sendo um inteiro, resultaria em bandas de larguras dadas por

$$\delta f_{\mathrm{B}Q} = \frac{f_0}{N}. \tag{5.14}$$

Fazendo, para uma mesma oitava, a largura da banda do canal da ferramenta do tipo BQ igual à largura da banda do primeiro canal da ferramenta do tipo CQ, isto é, $\mathrm{BW}_{\mathrm{B}Q} = \mathrm{BW}_{\mathrm{C}Q}(1)$, e resolvendo para N, obtém-se o número mínimo de canais

bounded-Q por oitava que garantirá larguras de bandas iguais à menor largura de banda do tipo *constant-Q*:

$$N_{\min} = 2^{\lceil \log_2(1/(\sqrt[R]{2}-1)) \rceil}. \tag{5.15}$$

5.5.1 *Bounded-Q fast filter bank*

O *Bounded-Q Fast Filter Bank* (BQFFB), apresentado em [P1], [P4] e [66], combina o espaçamento linear por partes da BQT com a alta seletividade do FFB. Isso pode ser alcançado usando-se um CQFFB para separar o sinal de entrada em oitavas e, a seguir, aplicar um FFB dentro de cada oitava para obter um espaçamento linear. Nesse esquema, o CQFFB requer apenas dez canais de saída, correspondendo ao limite de dez oitavas do sistema auditivo humano, o que não demanda muito em termos de custo computacional. Este CQFFB é composto por filtros projetados de acordo com o seguinte procedimento:

1. Obtenha o filtro da oitava mais alta (indexada por D) a partir do segundo filtro de um FFB de dois canais;

2. Obtenha o filtro de cada oitava seguinte (indexada por $d = (D-1), ..., 1$) como a cascata do segundo filtro de um FFN de 2^{D-d+1} canais com o primeiro filtro de um FFB de 2^{D-d} canais.

Este procedimento é descrito pelo Algoritmo 5.1. Neste algoritmo, o filtro separador da oitava de índice d é denotado por $CQFFB(d)$, D indica o número de oitavas e o filtro de índice b de um FFB de c canais é representado por $FFB(c, b)$.

Algoritmo 5.1 Algoritmo para formação dos filtros CQFFB separadores de oitava.

Para $d = D$ **até** 1

Se $d = D$

$CQFFB(d) \leftarrow FFB(2, 2)$

senão

$CQFFB(d) \leftarrow FFB(2^{D-d+1}, 2) * FFB(2^{D-d}, 2)$

Fim do **"Se"**

Fim do laço **"Para"**

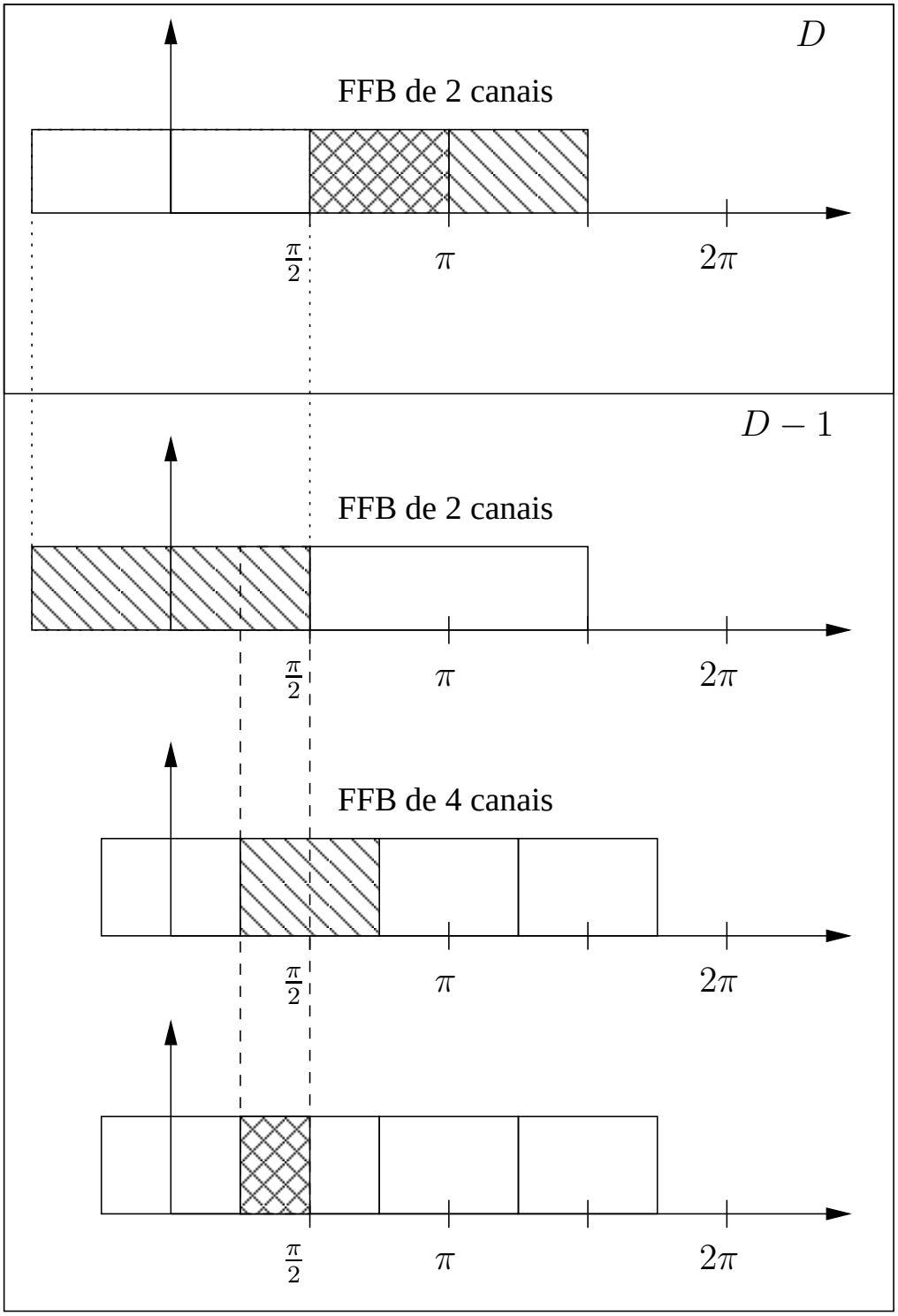


Figura 5.5: Procedimento para construir os filtros CQFFB para separar oitavas em um BQFFB. As regiões hachuradas em apenas uma direção indicam as bandas de passagem dos filtros que compõem um filtro separador de oitava e as regiões hachuradas em duas direções denotam as bandas de passagem dos filtros separadores de oitavas resultantes. É mostrada apenas a passagem da oitava D para a oitava $D-1$ para ilustração.

Como ilustração adicional, a Figura 5.5 mostra a passagem da oitava D para a oitava $D-1$. Nesta figura, exitem regiões hachuradas em apenas uma direção (que indicam filtros que compõem um filtro separador de oitava) e regiões hachuradas em duas direções (que denotam os filtros separadores de oitavas resultantes). O procedimento pode ser entendido da seguinte forma:

- O primeiro gráfico mostra a formação do filtro relativo à última oitava, de índice D. Esse filtro é formado pelo segundo filtro de um FFB de 2 canais. O primeiro filtro desse FFB se estende de $-\pi/2$ a $\pi/2$. O segundo filtro se estende de $\pi/2$ a $3\pi/2$, que é a região hachurada. Como o sinal de entrada é real, ou seja, o espectro estende-se de 0 a π, a interseção entre a banda do sinal de entrada e a banda de passagem do segundo filtro do FFB de 2 canais pode ser vista como a banda de passagem do filtro separador de oitavas relativo à oitava D, hachurada em duas direções. A razão para esta primeira etapa do procedimento é o fato de que o filtro associado à oitava mais alta (indexada por D), cuja banda de passagem vai de $\pi/2$ a π, é realizado pelo segundo filtro de um FFB de 2 canais. A banda resultante é limitada pela borda superior do espectro do sinal (ver parte superior da Figura 5.5).

- O segundo gráfico mostra um FFB de 2 canais. Novamente, o primeiro filtro desse FFB se estende de $-\pi/2$ a $\pi/2$. Para se obter o filtro relativo à oitava $D-1$, necessita-se deste filtro, cuja região está hachurada.

- O terceiro gráfico mostra um FFB de 4 canais em que o segundo canal está hachurado. O filtro relativo a este segundo canal também é usado para constituir o filtro relativo à oitava $D-1$.

- O quarto gráfico mostra a interseção entre a região hachurada no segundo gráfico e a região hachurada no terceiro gráfico, que resulta no filtro relativo à oitava de índice $D-1$, segundo o Algoritmo 5.1. Para a oitava $(D-1)$, o filtro deve ser projetado de modo que seu limite inferior seja $\pi/4$ e seu limite superior seja $\pi/2$. Esses limites podem ser alcançados combinando-se o primeiro filtro (passa-baixas) do primeiro gráfico com o segundo filtro (passa-banda) do segundo gráfico.

Esse processo é repetido iterativamente até que se chegue à oitava mais baixa (estendida até DC), permitindo construir os demais níveis do banco de filtros. Usando-se os filtros anteriormente mencionados na Seção 5.3.2 (isto é, com as mesmas ordens descritas em [62]) para separação de oitavas, o total de coeficientes não-nulos necessários para tal procedimento é dado pela Tabela 5.2.

Tabela 5.2: Número acumulado de coeficientes não-nulos dos filtros CQFFB separadores de oitava usados no BQFFB, onde $d = D$ é a oitava mais alta.

Número octaves (D)	Índice da oitava (d)	Coeficientes na oitava d	Coeficientes acumulados $F(D)$
1	D	7	7
2	$D-1$	6	13
3	$D-2$	3	16
4	$D-3$	3	19
5	$D-4$	2	21
6	$D-5$	2	23
7	$D-6$	2	25
8	$D-7$	2	27
9	$D-8$	2	29
10	$D-9$	2	31

Depois que as oitavas tiverem sido separadas na etapa de *constant-Q*, cada uma deve ser dividida em N canais linearmente espaçados através do seguinte procedimento, também visualizado na Figura 5.6. Nessa figura, estão mostradas as 3 oitavas superiores, de índices D, $D-1$ e $D-2$.

1. Para $d = 1, ..., D$, sub-amostrar o sinal de saída do filtro relativo à oitava d por um fator igual a 2^{D-d+1}. É importante reparar que a oitava de índice D também é sub-amostrada;

2. Submeter cada sinal sub-amostrado a um FFB de $2N$ canais, obtendo os canais

da oitava d separados.

Esse procedimento pode também ser descrito através do Algoritmo 5.2, no qual D é o número de oitavas, X é o sinal de entrada, $CQFFB(d)$ é o filtro CQFFB separador de oitavas relativo à oitava d, FFB_{2N} é o FFB de $2N$ canais, Y_{BQFFB} é a saída do BQFFB, o operador $\otimes$ denota convolução e $\downarrow$ indica sub-amostragem:

Algoritmo 5.2 Algoritmo para implementação do BQFFB através de um CQFFB seguido por um FFB.

Para $d = 1$ **até** D

$Y = CQFFB(d) \otimes X$

$Y_S = Y \downarrow 2^{D-d+1}$

$B(d) = FFB_{2N} \otimes Y_S$

Fim do laço **"Para"**

$Y_{BQFFB} = \sum_{d=1}^{D} B(d)$

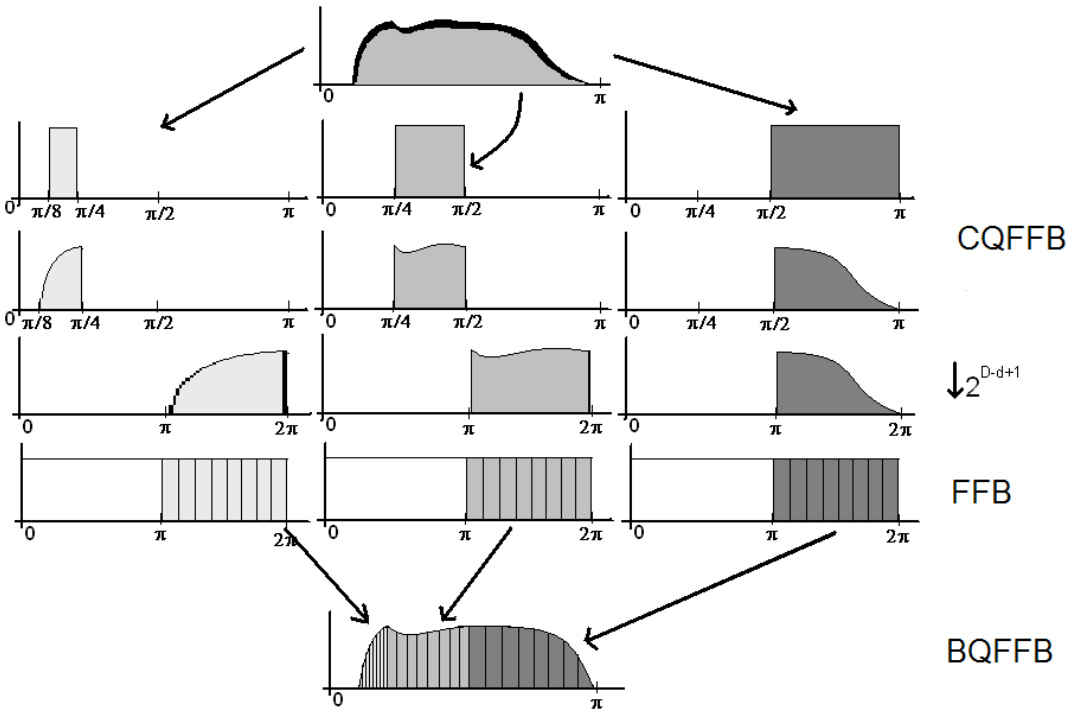


Figura 5.6: Implementação do BQFFB baseada em um CQFFB seguido de um FFB.

A sub-amostragem do sinal de cada oitava faz com que seu espectro fique mais largo (de 0 a 2π), sem necessitar do filtro de decimação, já que os filtros FFB de alta seletividade empregados na etapa de separação de oitavas eram suficientes para

evitar *aliasing*. É importante notar que o FFB usado dentro de cada oitava deve apresentar o dobro do número de canais que se deseja separar, visto que a parte negativa das respostas em freqüência dos filtros também é gerada. A Tabela 5.2 mostra o número acumulado $F(D)$ de coeficientes não-nulos para os filtros separadores de oitavas calculados em função do número de oitavas, denotado por D. Assim, o número de multiplicações complexas por amostra de entrada para o BQFFB é dado por:

$$C_{\mathrm{B}Q\mathrm{FFB,Total}} = (F(D) + D) + 2C(l)D, \tag{5.16}$$

onde $C(l)$ é obtido da Tabela 5.1.

A Tabela 5.3 resume as principais características de todas as ferramentas de análise espectral vistas até aqui. Por fim, é importante lembrar que, diferentemente da FFT e do FFB, nem o CQFFB nem o BQFFB são estruturalmente inversíveis. A ressíntese direta do sinal analisado através desses métodos requer um banco de filtros de síntese que possa aproximar a reconstrução perfeita. Esse fato resulta da não-inversibilidade da CQT [56].

Tabela 5.3: Comparação entre as diferentes ferramentas de análise espectral. O asterisco (*) indica as ferramentas de alta seletividade baseadas em FFB, as quais tendem a ser mais complexas que as baseadas em FFT.

Ferramenta de análise	**Distribuição na freqüência**	**Seletividade do canal**	**Complexidade Computacional**
FFT	linear	baixa	baixa
FFB	linear	alta	baixa (*)
CQT	geométrica	baixa	alta
CQFFB	geométrica	alta	alta (*)
BQT	linear por partes	baixa	média
BQFFB	linear por partes	alta	média (*)

5.6 Questões práticas acerca do BQFFB

A seguir, alguns aspectos relativos a implementações práticas e aplicações serão discutidos. A Sub-seção 5.6.1 discute a seleção dos valores dos parâmetros que controlam o desempenho do BQFFB. A Sub-seção 5.6.2 comenta sobre ajustes que devem ser feitos na implementação para que algumas dificuldades inerentes ao algoritmo do BQFFB sejam contornadas. Uma comparação entre o custo computacional do BQFFB e do CQFFB é feita na Sub-seção 5.6.3 mostrando o ganho que se obtém ao se utilizar o primeiro, enquanto que na Sub-seção 5.6.4 são feitos comentários acerca da dificuldade em se definir uma grade freqüencial para análise espectral.

5.6.1 Escolha para o valor dos parâmetros

Um primeiro problema a ser levado em consideração é a resolução do banco de filtros. Em aplicações musicais, é em geral preferível lidar com a organização da escala igualmente temperada usada na música ocidental: cada oitava é dividida em 12 notas musicais que seguem uma progressão geométrica de razão igual a $\sqrt[12]{2} \approx 1,06$. Essa razão é conhecida como semitom. Para detectar essa variação de um semitom, a resolução deve ser pelo menos igual à raiz quadrada desse valor, isto é, $\sqrt[24]{2} \approx 1,03$ (que é equivalente a um quarto de tom).

No caso de se desejar usar canais do tipo *constant-Q*, como no CQFFB [56], o fator de qualidade correspondente é dado por

$$Q = \frac{f(k)}{\delta f_{CQ}(k)} = \frac{f(k)}{(2^{1/48} - 2^{-1/48}) f(k)} \approx \frac{1}{0{,}0289} \approx 34{,}6, \tag{5.17}$$

onde $f(k)$ é a freqüência central (do ponto de vista geométrico) e $\delta f_{CQ}(k)$ é a largura da banda de uma dado canal k. Para simplificar os cálculos, o valor a ser utilizado para o fator de qualidade será 35.

O separação de quarto de tom corresponde a 24 canais geométricos dentro de uma oitava. Usando a Eq. (5.15), a solução do tipo *bounded-Q* deve empregar, no mínimo, $N_{\min} = 64$ canais por oitava, fazendo com que todos esses canais sejam mais estreitos que os canais do tipo *constant-Q*. Contudo, na prática, $N = 32$ pode ser usado, visto que somente 3 dos 24 canais CQFFB seriam mais estreitos que seus correspondentes BQFFB.

5.6.2 Ajustes na implementação

Esta sub-seção discute alguns pontos que devem ser considerados na implementação do BQFFB para evitar o aparecimento de falhas na análise tempo-freqüência. Esses ajustes foram apresentados em [P3] e são relativos a: lacuna no fim de oitava, compensação do filtro de oitava e mapeamento para escala de temperamento igual. Contudo, as soluções publicadas para os dois primeiros itens foram posteriormente revisadas, levando à abordagem integrada apresentada neste trabalho.

A metade superior do espectro de cada oitava, de 0 a π, deve ser submetida aos canais apropriados de um FFB para a segunda etapa do algoritmo. Entretanto, tomar apenas a metade dos canais cria uma lacuna no extremo superior da oitava. Para ilustrar esse problema de forma simples, um FFB de $N = 8$ canais é mostrado na Figura 5.7(a). Ao se considerar apenas os $N/2$ canais cujas freqüências centrais são maiores ou iguais a π (seguindo o algoritmo para construção do BQFFB mostrado na Seção 5.5), percebe-se que há uma lacuna no fim da oitava (ver Figura 5.7(b)). É importante lembrar que nesta figura está representado somente o que está contido dentro de uma oitava.

Para se resolver esse problema, foi proposto em [P3] que se considerasse, juntamente com os $N/2$ canais de freqüências centrais mais altas da Figura 5.7(b), o primeiro canal dentre aqueles de freqüência central mais baixa (da Figura 5.7(c)), resultando no espectro representado na Figura 5.7(d). Com isso, a lacuna seria preenchida.

Logicamente, o CQFFB que separa o espectro em oitavas não é ideal como mostrado na Figura 5.6. Se os filtros das oitavas apresentassem respostas muito próximas da ideal, não seria possível para o sistema detectar transitórios rápidos nas oitavas mais baixas. Isso pode ser ilustrado submetendo-se ao banco um conjunto de senóides de amplitude unitária (130 Hz, 150 Hz, 260 Hz, 300 Hz, 520 Hz, 600 Hz, 1040 Hz, 1200 Hz, 2080 Hz e 2400 Hz) ao BQFFB, utilizando-se uma janela de 190 ms. Os filtros separadores de oitava são definidos de acordo com a Seção 5.5 e têm suas respostas em freqüência mostradas na Figura 5.8(a). Percebe-se, primeiramente, que as bandas de passagem desses filtros não apresentam o aspecto plano desejado. Isso faz com que, mesmo dentro de uma mesma oitava, as saídas dos canais do BQFFB não apresentem amplitudes unitárias ou pelo menos iguais entre

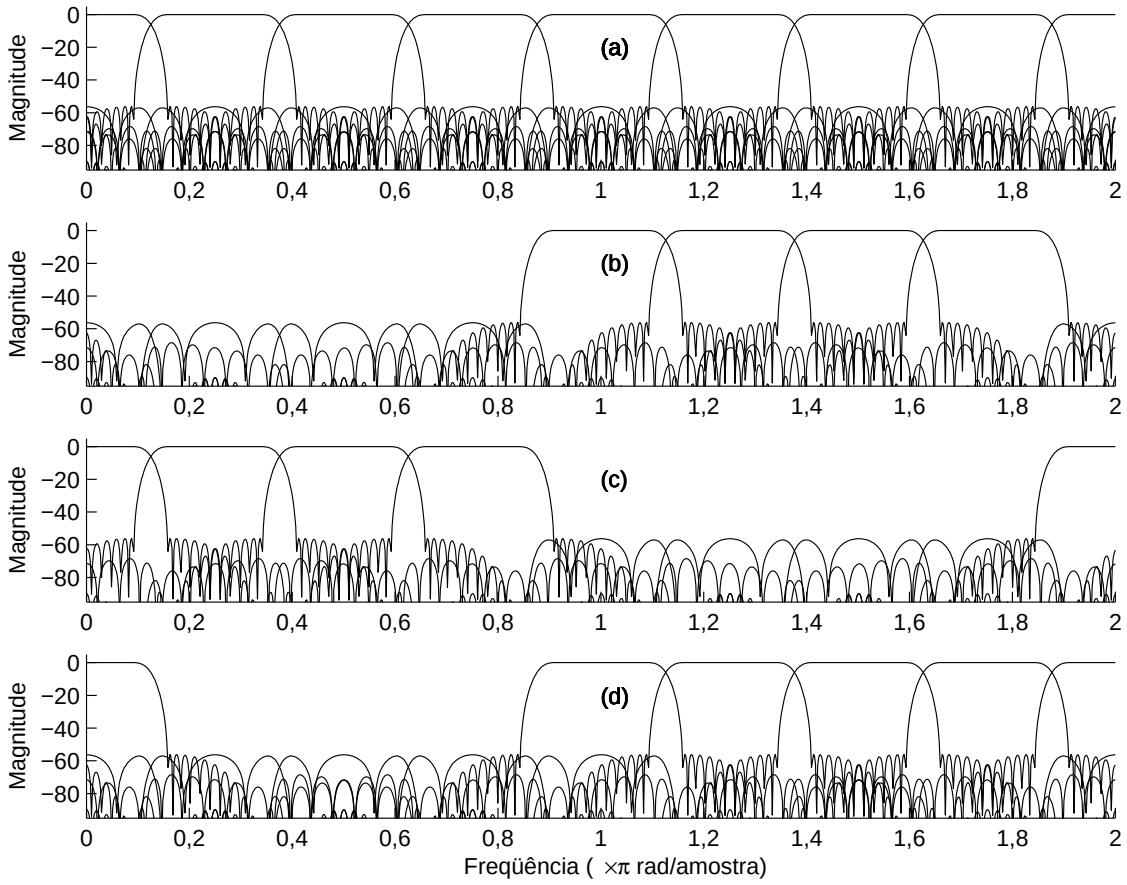


Figura 5.7: Resposta em magnitude (em dB) dos canais de um FFB de $N = 8$ canais: (a) Banco completo; (b) $N/2$ canais mais altos; (c) $N/2$ canais mais baixos; (d) $N/2$ canais mais altos mais o primeiro canal.

si, como visto na Figura 5.8(b). Além disso, existe uma alta taxa de sobreposição entre as oitavas.

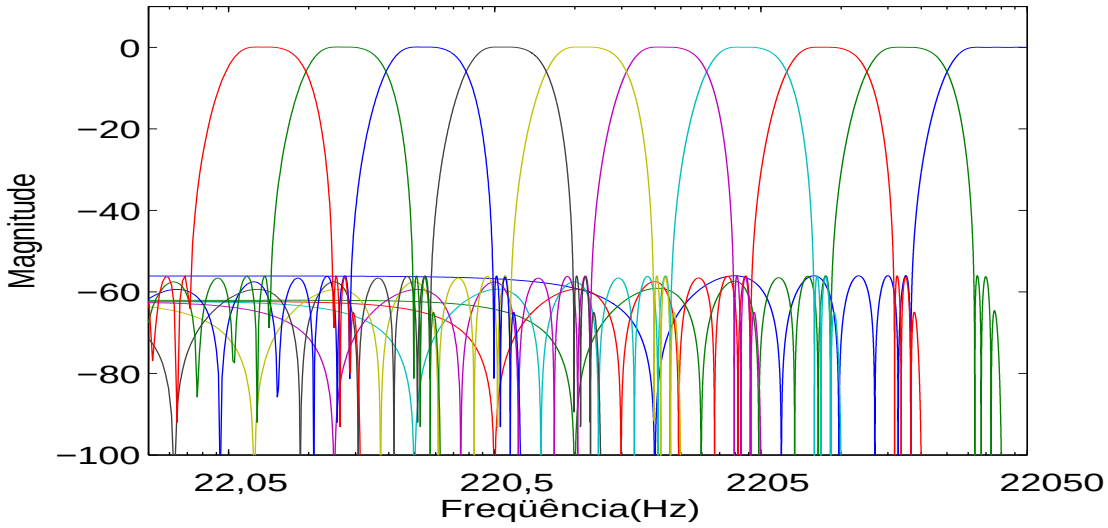
Aumentando-se a seletividade dos filtros protótipos, obtêm-se os filtros separadores de oitava mostrados na Figura 5.9(a). É possível ver que tais filtros apresentam aspecto mais plano na banda de passagem, o que faz com que as saídas dos canais apresentem amplitudes unitárias e mais próximas entre si. Além disso, a sobreposição entre as oitavas é reduzida. O problema dessa abordagem reside no fato de os filtros relativos às oitavas mais baixas apresentarem comprimentos excessivamente elevados. Isso faz com que a janela de 190 ms não seja suficientemente longa para que as senóides de baixa freqüência (particularmente as de 130 e 150 Hz) excitem os filtros de oitava de forma significativa, o que conseqüentemente faz com que as respostas do filtros a estas senóides sejam tão menores que as respostas às outras senóides que nem é possível visualizá-las na Figura 5.9(b).

Para resolver tal problema, adotou-se uma configuração híbrida para os filtros separadores de oitava, da seguinte forma (ver Figura 5.10(a)):

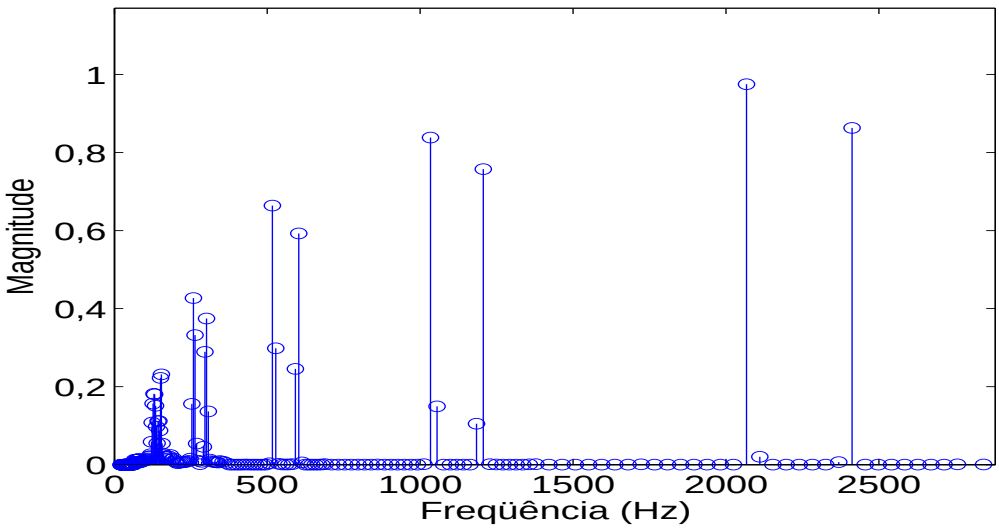
- As oitavas de 1 a 3 apresentarão a seletividade original (segundo descrito em [4]);
- As oitavas de 5 a 10 apresentarão a seletividade aumentada (reduzindo-se a banda de passagem em relação às especificações originais);
- A oitava 4 apresentará uma seletividade híbrida.

A seletividade híbrida é obtida empregando-se na composição do CQFFB separador de oitava filtros oriundos de FFBs com seletividades diferentes: um FFB com seletividade original e outro com seletividade aumentada (ver algoritmo ilustrado pela Figura 5.5). Uma oitava com este tipo de seletividade é necessária para servir de interface entre as demais seletividades de modo a garantir a complementaridade entre os filtros. Aplicando-se esta configuração, as senóides de baixa freqüência passam a ser passíveis de detecção, como pode ser visto na Figura 5.10(b).

É necessário, então, trabalhar com um banco de filtros de respostas não ideais e tratá-las para que se obtenha um resultado coerente do que manipular filtros muito seletivos para separação de oitavas.

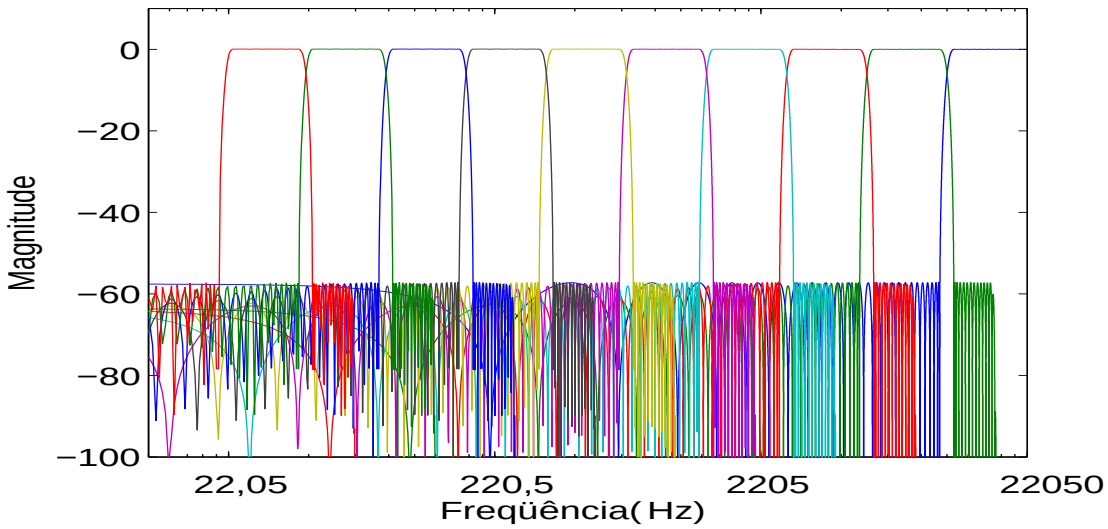


(a)

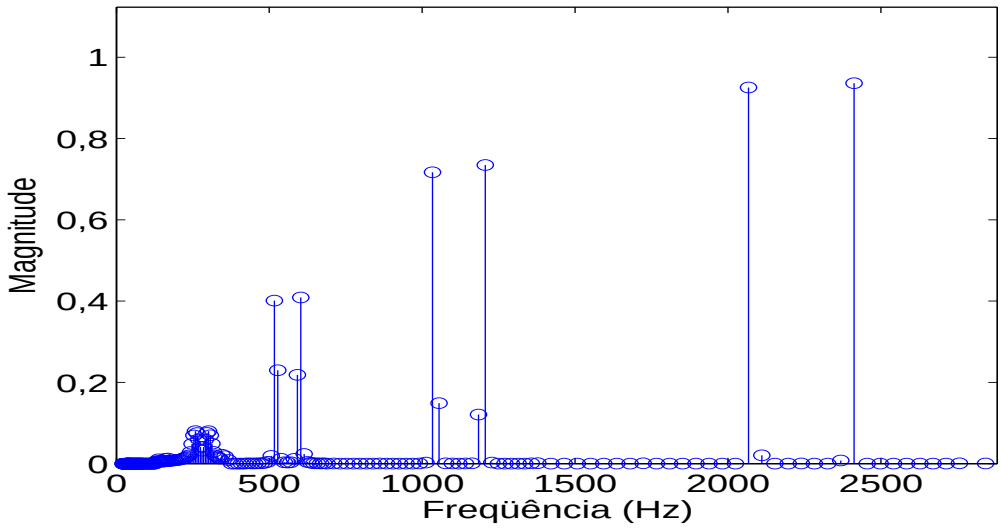


(b)

Figura 5.8: CQFFB representando 10 canais geometricamente espaçados por um fator de 2: (a) Resposta em freqüência dos filtros separadores de oitavas com a seletividade original [P4] (magnitude em dB); (b) Saída do BQFFB para entradas senoidais de freqüências 130 Hz, 150 Hz, 260 Hz, 300 Hz, 520 Hz, 600 Hz, 1040 Hz, 1200 Hz, 2080 Hz e 2400 Hz e amplitude unitária com base nos separadores de oitava com a seletividade original [P4].

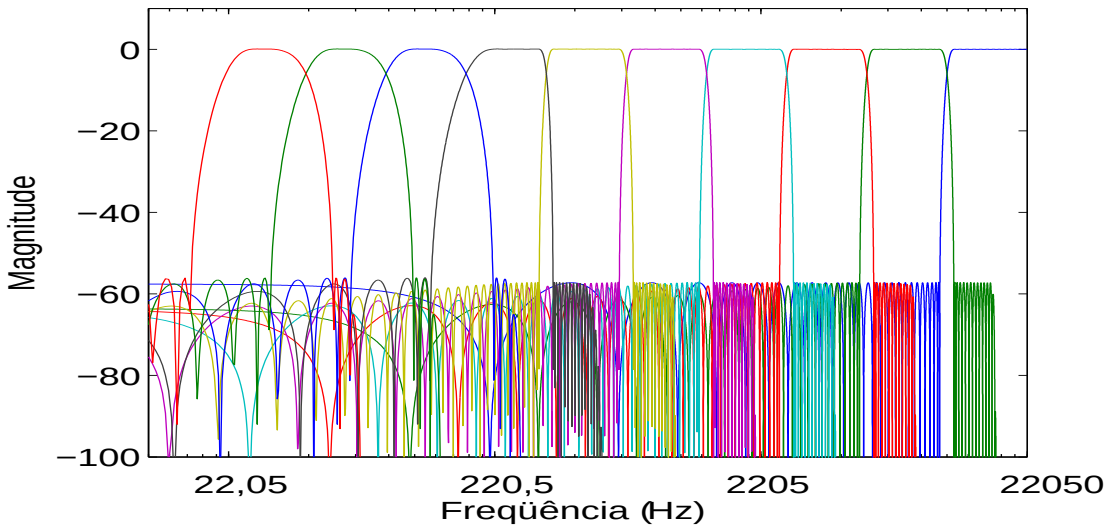


(a)

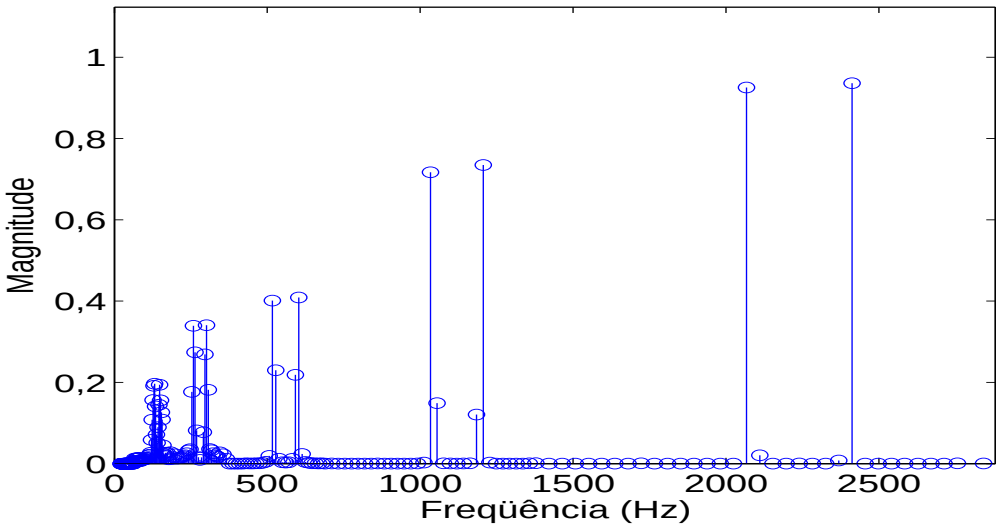


(b)

Figura 5.9: CQFFB representando 10 canais geometricamente espaçados por um fator de 2: (a) Resposta em freqüência dos filtros separadores de oitavas com seletividade aumentada (magnitude em dB); (b) Saída do BQFFB para entradas senoidais de freqüências 130 Hz, 150 Hz, 260 Hz, 300 Hz, 520 Hz, 600 Hz, 1040 Hz, 1200 Hz, 2080 Hz e 2400 Hz e amplitude unitária com base nos separadores de oitava com seletividade aumentada.



(a)



(b)

Figura 5.10: CQFFB representando 10 canais geometricamente espaçados por um fator de 2: (a) Resposta em freqüência dos filtros separadores de oitavas com seletividade híbrida (magnitude em dB); (b) Saída do BQFFB para entradas senoidais de freqüências 130 Hz, 150 Hz, 260 Hz, 300 Hz, 520 Hz, 600 Hz, 1040 Hz, 1200 Hz, 2080 Hz e 2400 Hz e amplitude unitária com base nos separadores de oitava com seletividade híbrida.

A partir desse ponto, já se está considerando o número de canais obtido na Seção 5.6.1, ou seja, $N = 32$ canais. Isso implica o emprego de um banco de filtros de $2N = 64$ canais. Somando-se as respostas de todos os filtros, obtém-se um valor unitário constante. E isso acontece porque cada filtro possui a faixa de transição direita com inclinação igual à do filtro da oitava imediatamente superior e apresenta a faixa de transição esquerda com inclinação igual à do filtro relativo à oitava imediatamente inferior.

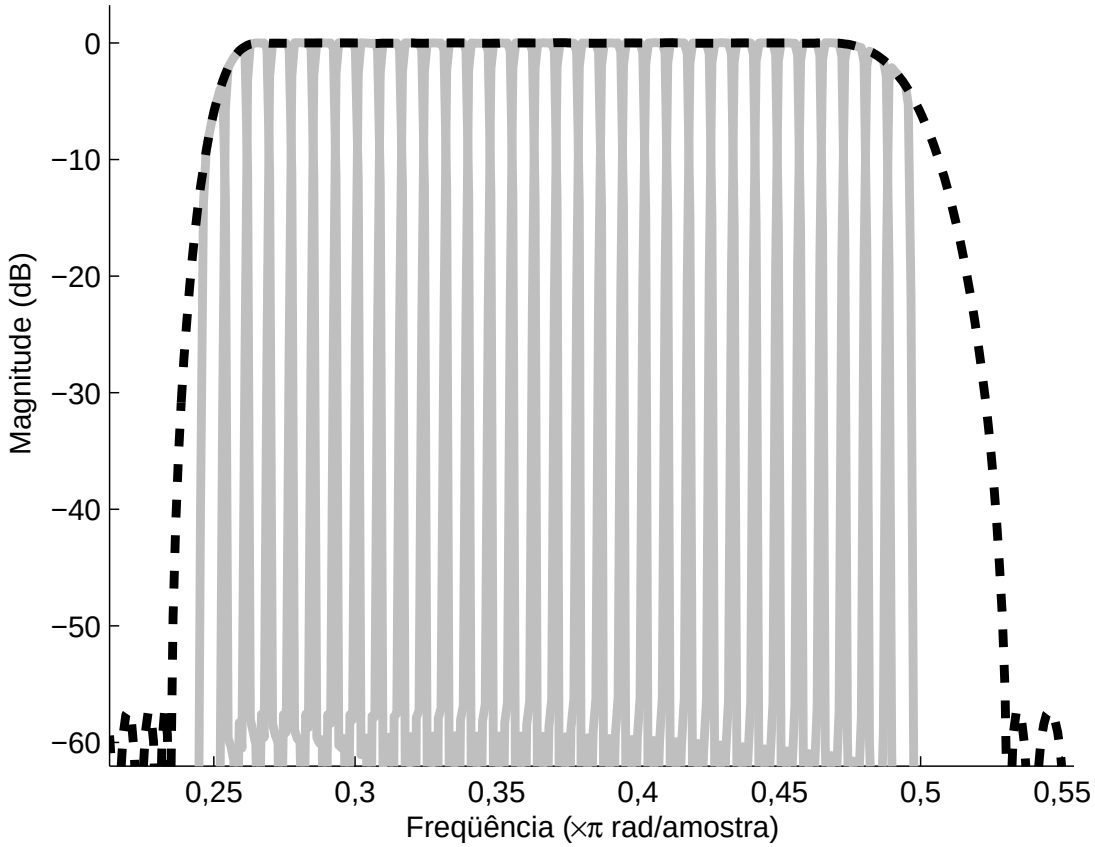


Figura 5.11: Resposta em freqüência de canais (linhas cheias) usados originalmente no BQFFB, que não preenchem toda a região delimitada pelo filtro separador de oitavas (linha tracejada), incluindo banda de passagem e faixa de transição.

Contudo, os filtros de canais não preenchem toda a região delimitada por cada filtro de oitava (ver Figura 5.11). Assim, ao se somar os canais, não se obtém um valor unitário constante, como mostrado na Figura 5.12. Para contornar essa dificuldade, optou-se por considerar, além dos 32 canais originais (de índices 33 a

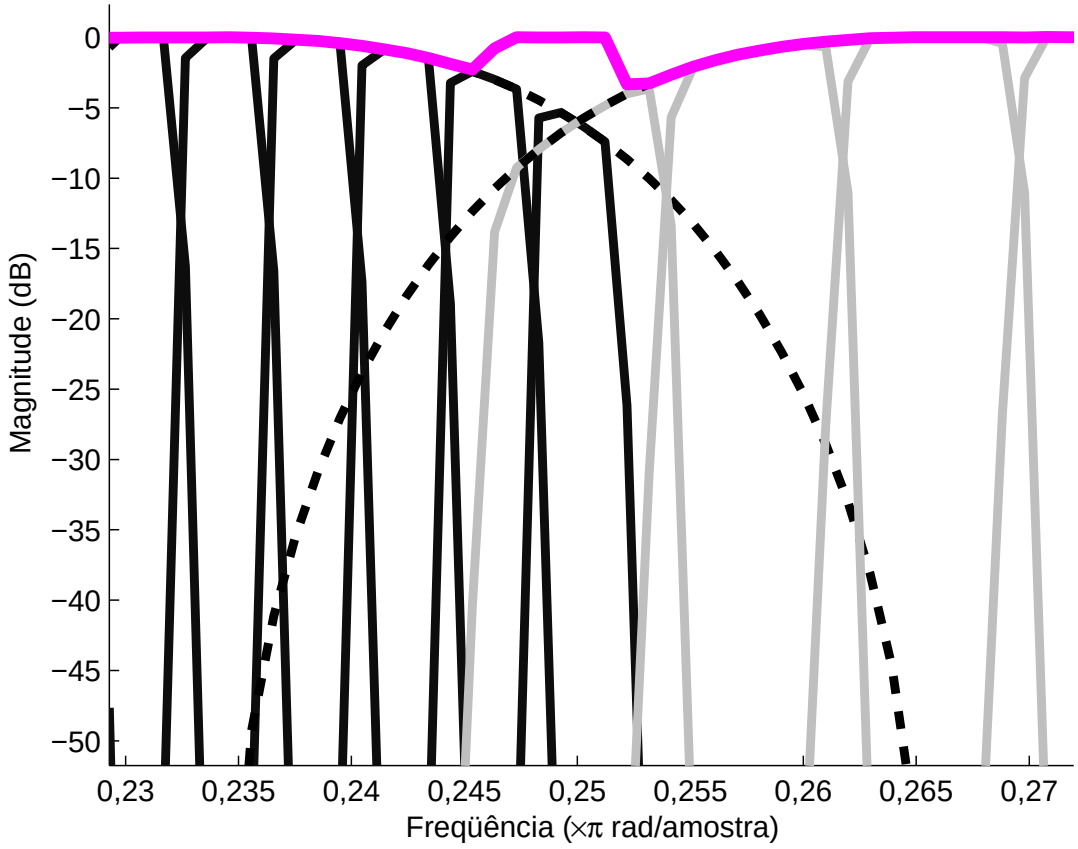


Figura 5.12: Somando-se os canais usados originalmente no BQFFB, não se obtém um valor unitário constante para a magnitude da resposta em freqüência resultante. O filtros separadores de oitava estão representados por linhas tracejadas e os filtros de canais são representados por linhas pretas contínuas (para a oitava inferior) e cinzenta (para a oitava superior). A linha cheia mais grossa indica a soma da magnitude das respostas em freqüência.

64), os de índices 32, 1, 2, 3 e 4. O canal de índice 32 fica à esquerda dos 32 canais originais e os demais ficam à direita dos 32 canais originais. Assim, há um quase total preenchimento da região delimitada pelo filtro separador de oitava, o que é exibido na Figura 5.13.

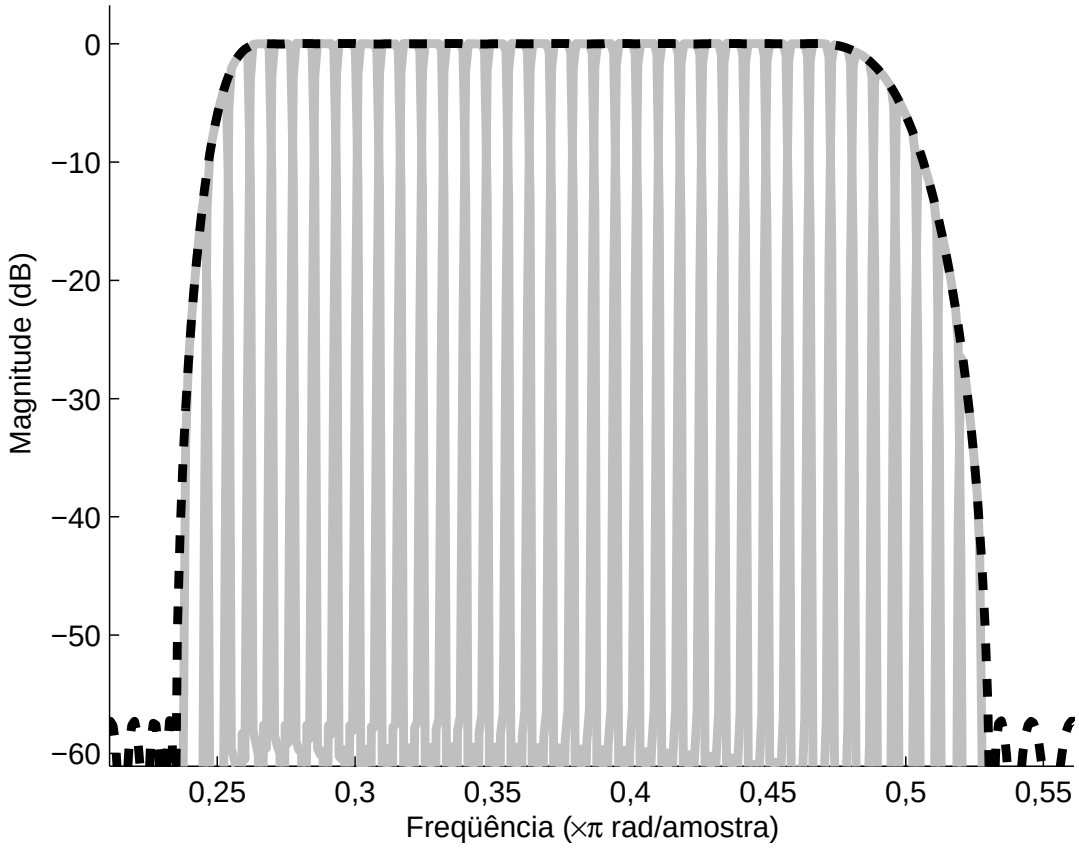


Figura 5.13: Resposta em freqüência de canais adicionais (linhas cheias) usados juntamente com os originais do BQFFB para preencherem toda a região delimitada pelo filtro separador de oitava (linha tracejada), incluindo banda de passagem e faixa de transição.

Com esses filtros a mais, a soma de todos os canais resulta em um valor unitário constante, como mostrado na Figura 5.14. Contudo, com o uso desses novos filtros, há uma sobreposição de informação nas regiões de transição entre as oitavas: os filtros no fim de uma oitava operam sobre a região delimitada pelos primeiros filtros da oitava seguinte. Para se distribuir essa informação entre as oitavas, foi elaborado um sistema de endereçamento que indica para qual oitava e canal BQFFB deve ir a

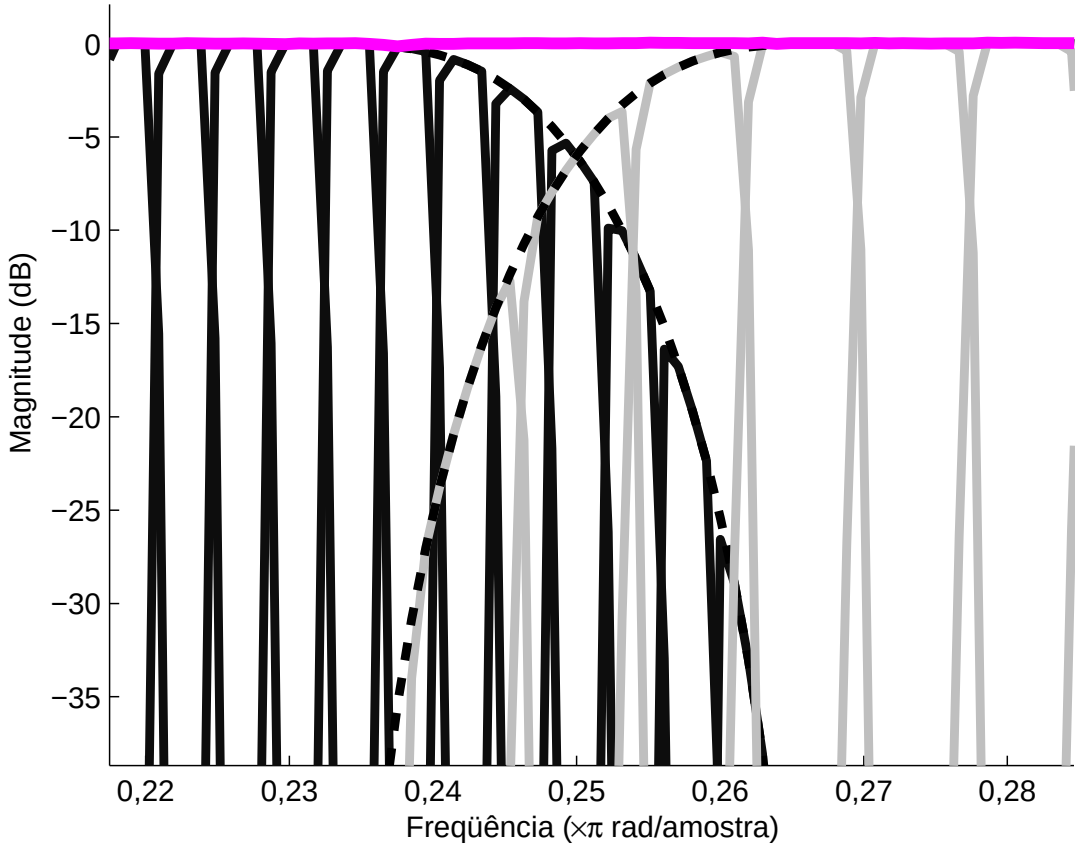


Figura 5.14: Soma dos canais originais do BQFFB com os adicionais usados para permitir o preenchimento da região delimitada pelo filtro separador de oitava, que resulta em valor unitário constante. O filtros separadores de oitava estão representados por linhas tracejadas e os filtros de canais são representados por linhas contínuas pretas (para a oitava inferior) e cinzentas (para a oitava superior). A linha cheia mais grossa mostra a soma dos filtros de canais.

informação relativa a cada filtro. O critério para esse endereçamento foi direcionar para o mesmo canal os filtros que operam sobre a mesma região somando a saída deste grupo de filtros e formando a saída do canal. É claro que isso faria com que filtros de larguras diferentes fossem direcionados para o mesmo canal. Mas, como se trata de uma região de transição, isso faz sentido, porque a informação contida nos canais presentes nesta região é composta pelo conteúdo de canais de larguras diferentes.

Os procedimentos anteriores permitem separar o espectro inteiro em oitavas. Essas oitavas são definidas como função da freqüência de amostragem normalizada 2π da seguinte maneira: a oitava mais alta é definida de $\pi/2$ a π; a segunda mais alta vai de $\pi/4$ a $\pi/2$; e assim por diante. A seguir, cada uma das $M = 10$ oitavas é dividida em N canais igualmente espaçados.

Uma simples estratégia pode ser estabelecida para se mapear os canais BQFFB em candidatas a freqüências fundamentais pertencentes à escala de temperamento igual:

1. Calcular as freqüências fundamentais teóricas das notas musicais e definir uma banda de meio tom t_l em torno da F0 de cada nota l (por exemplo, para a nota A4 = 440 Hz, ou seja, $l = 440$, a banda t_l seria definida de 427,5 a 452,9 Hz);

2. Uma vez que o número de oitavas M a serem analisadas pelo BQFFB e o número de canais N dentro de cada oitava tenham sido definidos, estabelecer os limites inferior e superior para cada canal BQFFB;

3. Compor cada candidata a freqüência fundamental l somando os canais BQFFB contidos em t_l ponderados por sua percentagem de interseção.

Para ilustrar esse procedimento, considera-se o exemplo em que $M = 10$ e $N = 32$. Os canais BQFFB associados a A4 são: A (que vai de 425,2 Hz a 436,0 Hz), B (que vai de 436,0 Hz a 446,8 Hz), C (que vai de 446,8 Hz a 456,6 Hz). A soma ponderada deve, então, ser 79 % para o canal A, 100 % para o canal B e 56 % para o canal C. Esse processo é ilustrado na Figura 5.15.

Para se ilustrar esses ajustes, foi montado um experimento no qual é usada como sinal de entrada uma escala ascendente cromática (uma progressão de notas onde

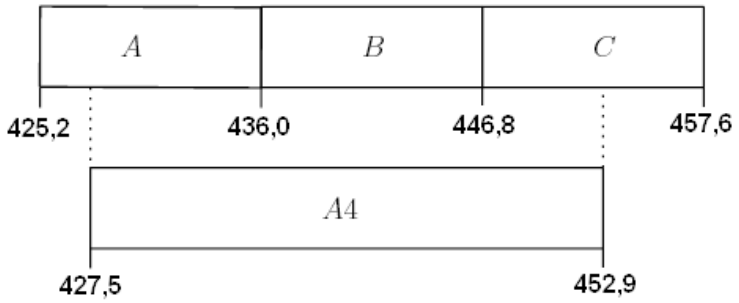


Figura 5.15: Exemplo de mapeamento entre canais BQFFB e canais relativos a notas da escala temperada.

cada uma está um 1 semi-tom acima da anterior) que se estende de C5 (523 Hz) a E8 (5274 Hz). Esse sinal é então submetido a um BQFFB com $M = 10$ oitavas, $N = 32$ canais por oitava, sem qualquer ajuste, gerando a saída mostrada na Figura 5.16.

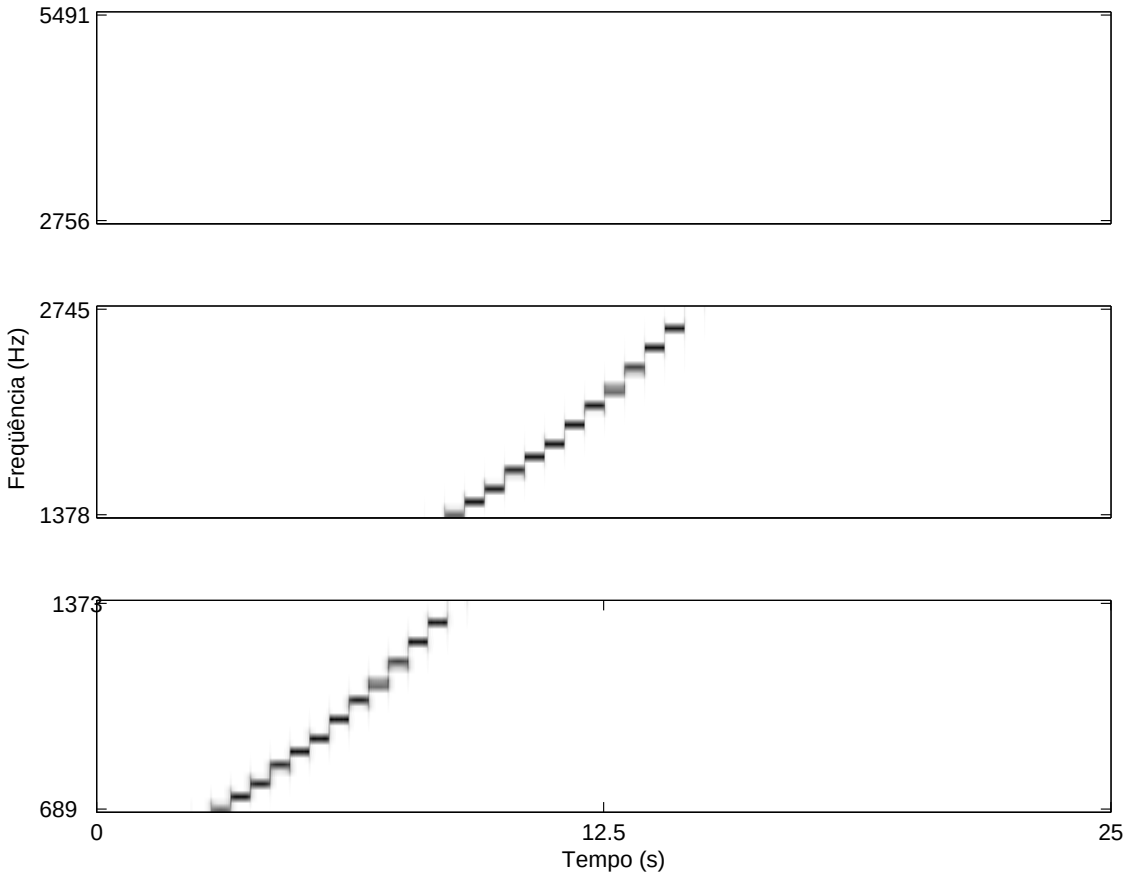


Figura 5.16: Saída do BQFFB para um sinal de entrada contendo uma escala cromática. Pode-se ver que algumas notas são representadas por mais de um canal BQFFB.

Aplicando o procedimento de mapeamento para escala de igual temperamento, as oitavas do BQFFB com 32 filtros de canais são mapeados nas oitavas musicais com 12 canais de notas cada. O resultado é visto na Figura 5.17, onde todos os tons são devidamente distinguíveis em freqüência e no tempo.

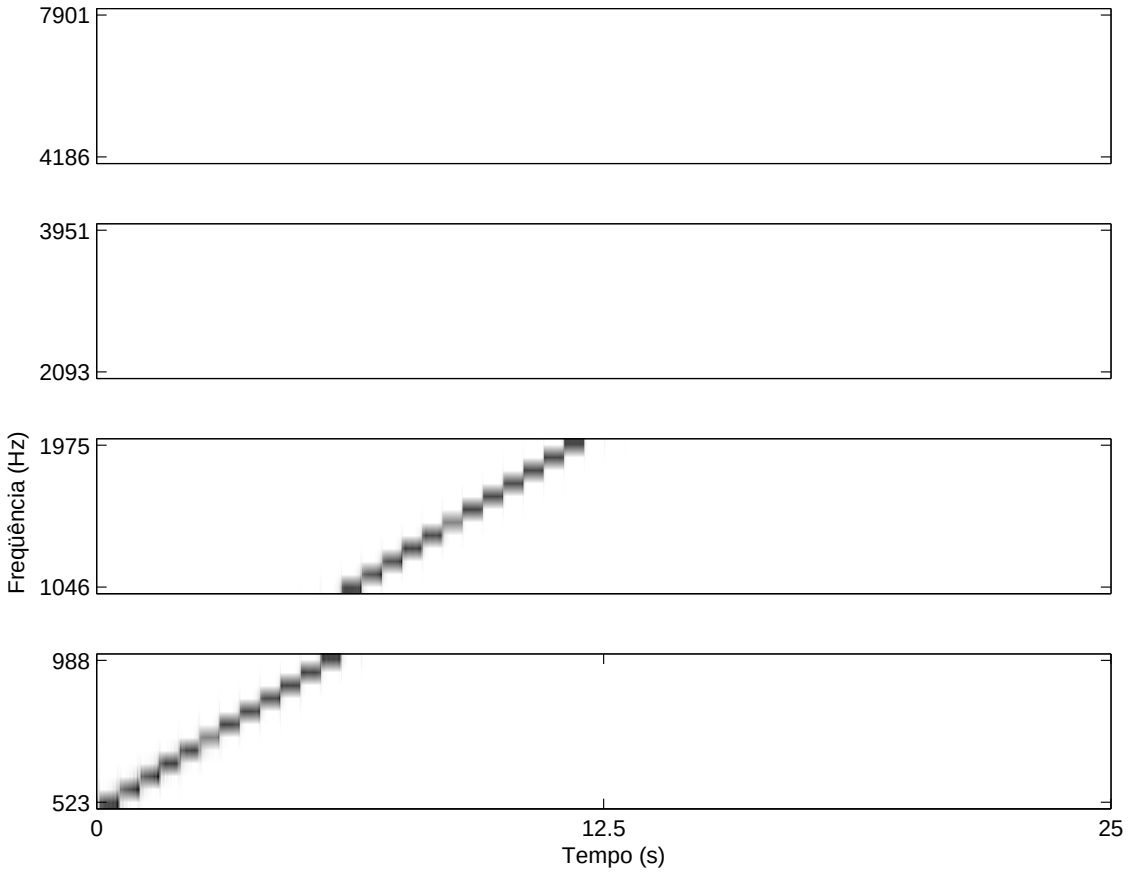


Figura 5.17: Saída do BQFFB para um sinal de entrada contendo uma escala cromática depois do mapeamento para escala temperada.

Apesar de a Figura 5.17 apresentar tons melhor definidos que os da Figura 5.16 (no sentido de que cada nota é representada por um canal), os diferentes tons (que foram gerados com a mesma intensidade) ainda são exibidos com intensidades diferentes, um efeito da separação não ideal das oitavas. Isso pode ser melhor visualizado na Figura 5.18(a), que mostra a soma direta das saídas dos filtros.

Já a Figura 5.18(b) mostra a mesma soma quando se utiliza a composição de canais em oitavas de seletividade híbrida [P3]. Pode-se verificar que os ajustes foram bem sucedidos para equalizar os filtros separadores de oitava. Os picos observados em ambos os gráficos são associados aos transitórios dos filtros e aos ataques

das notas, que foram propositalmente mantidos para servir como marcadores das transições entre as notas.

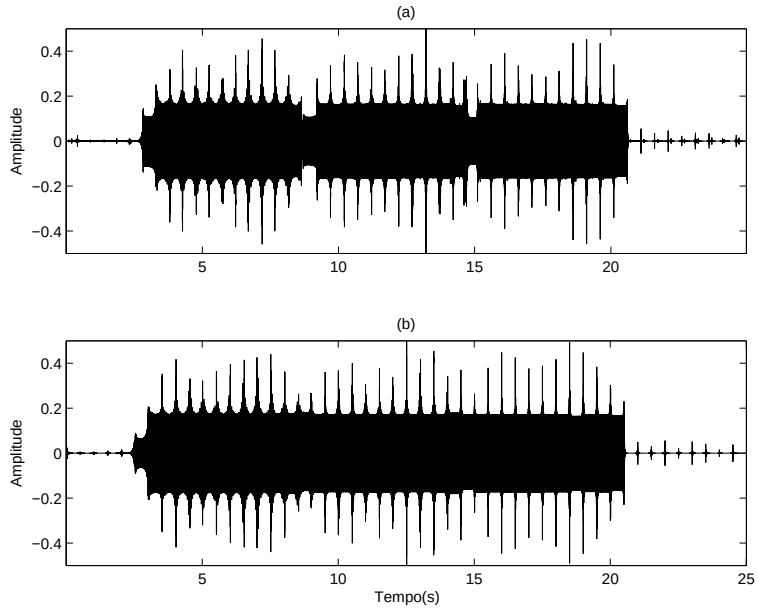


Figura 5.18: Saída do BQFFB para um sinal de entrada contendo uma escala cromática. (a) Antes dos ajustes. (b) Depois dos ajustes.

5.6.3 Comparação de complexidade

Para comparar os custos computacionais do CQFFB e do BQFFB, a Figura 5.19 mostra o número de multiplicações complexas necessárias para analisar um espectro de 10 oitavas em função do número de canais. Essas curvas são regidas pela Eq. (5.12) para o caso do CQFFB e pela Eq. (5.16) para o caso do BQFFB. Fica claro que o BQFFB apresenta uma complexidade computacional bem mais atrativa. Em aplicações típicas, usando de 100 a 320 canais, a diferença é da ordem de 5 ordens de magnitude em favor do BQFFB.

5.6.4 Requisitos e aplicações

No contexto da transcrição automática de música, mesmo para o simples caso de um sinal monofônico, a identificação de notas musicais ainda enfrenta uma série de problemas:

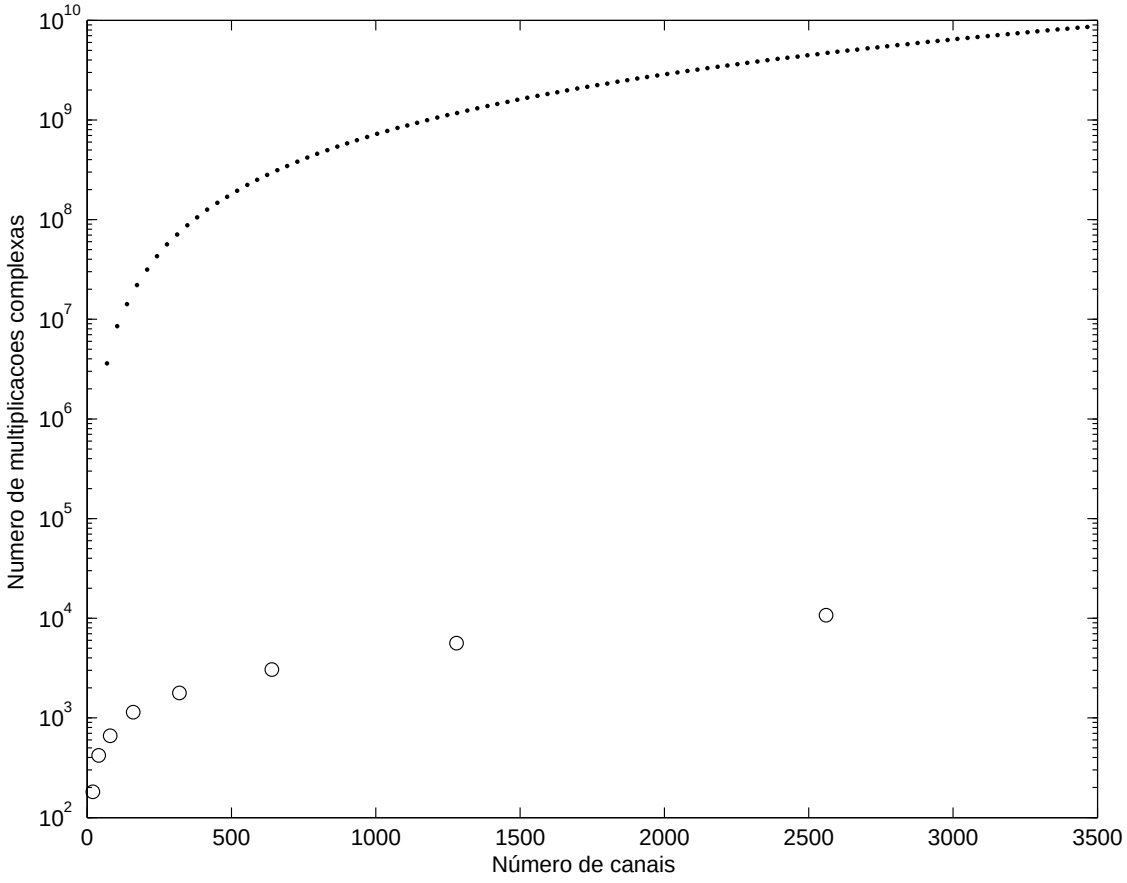


Figura 5.19: Comparação entre a complexidade computacional do CQFFB, representado por pontos (.), e do BQFFB, representado por círculos (o), como função do número de canais.

- A afinação absoluta, a partir de uma nota de referência como A4 = 440 Hz, não é sempre garantida;
- Instrumentos podem estar simplesmente desafinados, deslocando as notas arbitrariamente na freqüência;
- Instrumentos (como sinos) podem apresentar inarmonicidade, e nesse caso o *pitch* percebido não é necessariamente associado à freqüência fundamental;
- A maior parte dos instrumentos emite notas continuamente variáveis (como um violino, por exemplo, em contraste com o piano).
- Além disso, no contexto mais usual de sinais polifônicos, quando o problema de notas superpostas no espectro deve ser resolvido, os harmônicos devem ser cuidadosamente considerados. Contudo, estes apresentam espaçamento linear na freqüência.

Todas essas considerações podem ser resumidas em uma frase: não há uma grade de freqüências ideal para análise espectral de sinais musicais. De fato, dependendo da aplicação, diversas soluções devem ser empregadas de forma simultânea. Sob essa perspectiva, a economia provida pelos métodos *bounded-Q* de 5 ordens de magnitude em relação aos métodos *constant-Q* faz com que os primeiros sejam bem mais atrativos, em geral. O espaçamento linear dos harmônicos não é tão preocupante se houver granularidade suficiente. É claro que essa alta granularidade deve estar acompanhada por uma boa capacidade de separação de canais, e é essa a importância da inclusão de filtros FFB na estrutura proposta.

Em geral, os métodos propostos podem ser vistos como representações tempo-freqüência direcionadas a aplicações musicais e podem fornecer o par (magnitude, freqüência) $\times$ tempo como parâmetros para sistema de extração de informações musicais (ou *music information retrieval - MIR*). Esses parâmetros podem, então, ser analisados por camadas de processamento de alto nível. Como esse tipo de aplicação, em geral, manipula uma grande quantidade de dados, um número reduzido de canais (e um número reduzido de amostras de saída) é um aspecto importante tanto do CQFFB quanto do BQFFB.

5.7 Experimentos computacionais

Nesta seção, alguns experimentos simples serão descritos para avaliar o desempenho dos métodos de alta seletividade e de resolução variável, usando-se os métodos com espaçamento linear como referência. Os resultados obtidos nesta seção foram decorrência de sistemas que já apresentam os ajustes de implementação descritos na Sub-seção 5.6.2, como preenchimento da lacuna ao fim da oitava (compensando o filtro separador de oitava), emprego de banco de filtros com seletividade híbrida e mapeamento para notas da escala temperada.

Primeiro, considera-se um sinal de teste de 1 segundo de duração formado pela soma de 8 tons puros de magnitude unitária. Os primeiros tons estão nas freqüências 263 Hz e 295 Hz, as quais correspondem às notas C4 e D4 ligeiramente desafinadas em relação à afinação absoluta, de modo a simular uma condição realística. Os próximos três harmônicos de cada tom também estão incluídos. Como o principal objetivo desse experimento é a detecção visual de notas, as magnitudes dos componentes foram estabelecidas como sendo unitárias para facilitar a visualização dos resultados.

A resolução freqüencial adotada na simulação do CQFFB foi com $Q = 35$, como mostrado na Eq. (5.17), e também serve como referência para as escolhas do número de canais para os demais métodos. Para fazer uma comparação justa, o canal com a melhor resolução nas ferramentas de espaçamento linear por partes deve satisfazer a condição de ter a mesma largura de um canal de uma ferramenta de espaçamento linear. Essa restrição se aplica ao canal relativo à freqüência mais grave e satisfazendo-a, está se garantindo que a resolução máxima da ferramenta com resolução variável será a mesma da ferramenta de espaçamento linear. Para satisfazer tais restrições, tanto a FFT quanto o FFB dividem o espectro em 4096 canais de 0 a 22050 Hz (assumindo uma freqüência de amostragem de 44100 Hz), cada com largura igual a 5,38 Hz.

O BQFFB, por sua vez, divide o espectro (a partir de seu limite superior) em 7 oitavas mais a sub-banda restante relativa às freqüências mais baixas. Cada uma dessas 8 sub-bandas é uniformemente dividida em 32 canais, mantendo, assim, na banda mais grave, o mesmo espaçamento da FFT e do FFB.

As Figuras 5.20 a 5.23 mostram as respostas da FFT, do FFB, do CQFFB e

do BQFFB ao sinal de teste, respectivamente. A partir dessas figuras, fica evidente que a FFT causa uma alta interferência aparente em torno dos tons de teste. Isso é devido à baixa seletividade dos filtros associados à FFT. Esse efeito pode ser um grande problema em algumas aplicações e chegar a mascarar alguns componentes de sinal próximos a esses picos. Em contrapartida, o FFB é capaz de detectar os picos claramente, usando, porém, um número desnecessariamente alto de canais. O CQFFB identifica os tons com menos canais, contudo ao custo de uma carga computacional elevadíssima. De fato, o BQFFB apresenta um desempenho comparável ao do FFB, mas com uma complexidade de cerca de 5 ordens de magnitude menor que a do CQFFB. A Figura 5.23 já inclui os ajustes propostos na Sub-seção 5.6.2. Pode-se ver que, quando há um super-dimensionamento do número de canais, há uma maior probabilidade de um tom ser representado por mais de uma raia fazendo com que a energia seja dividida e a envoltória sofra distorções, como ocorreu com a FFT e com o FFB. A pequena variação na amplitude dos picos detectados pelo BQFFB decorre das aproximações ao se calcular a compensação do filtro separador de oitava. Mas essa variação ainda é menor do que a que ocorreria se não houvesse tal compensação.

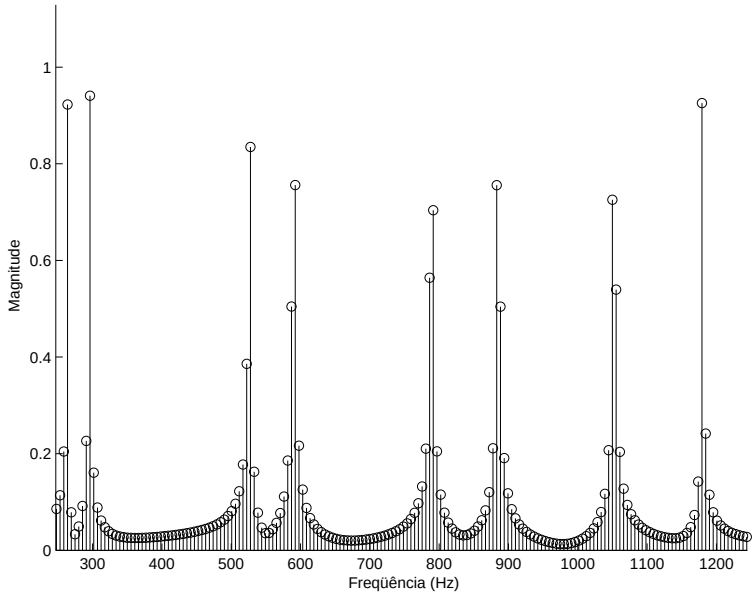


Figura 5.20: Análise por FFT do sinal de teste formado por dois tons. A distribuição freqüencial linear da FFT faz com que a resolução seja constante ao longo de todo o espectro, e a reduzida atenuação do lobo secundário gera um aparente "chão de ruído" que pode mascarar alguns sinais de menor potência.

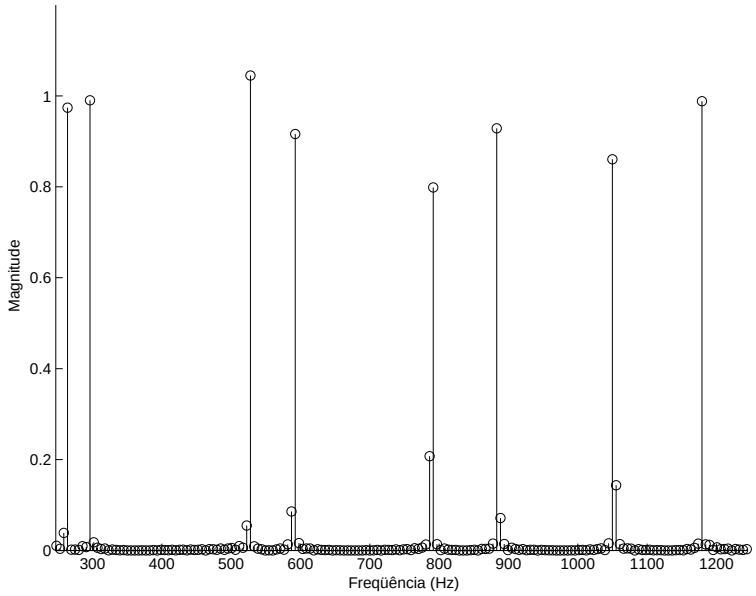


Figura 5.21: Análise por FFB do sinal de teste formado por dois tons. A distribuição freqüencial linear do FFB faz com que a resolução seja constante ao longo de todo o espectro, onde a alta seletividade do FFB evita o surgimento do "chão de ruído". Informação sobre as ordens dos filtros FFB pode ser encontrada na Tabela 5.1.

Também foram realizados testes para investigar a capacidade de análise do BQFFB para sinais musicais reais de características variadas. Tais resultados foram apresentados em [P2].

5.8 Conclusão

Nesse capítulo, foram apresentados diversos algoritmos para análise espectral de sinais de música. O FFB foi visto como uma versão de alta seletividade do algoritmo padrão FFT. A CQT e a BQT foram vistas como variações da FFT que possuem uma distribuição mais eficiente dos canais ao longo das freqüências. A CQT usa uma separação geométrica para emular a organização usualmente empregada na música ocidental (escala temperada). Enquanto isso, a BQT usa uma separação linear por partes para permitir uma implementação rápida do algoritmo sem sacrificar sua habilidade de discriminação de notas musicais. Duas novas ferramentas são apresentadas: o CQFFB e o BQFFB, os quais são versões de alta seletividade da CQT e da BQT, respectivamente.

O BQFFB é uma ferramenta eficiente para análise espectral, pois combina redu-

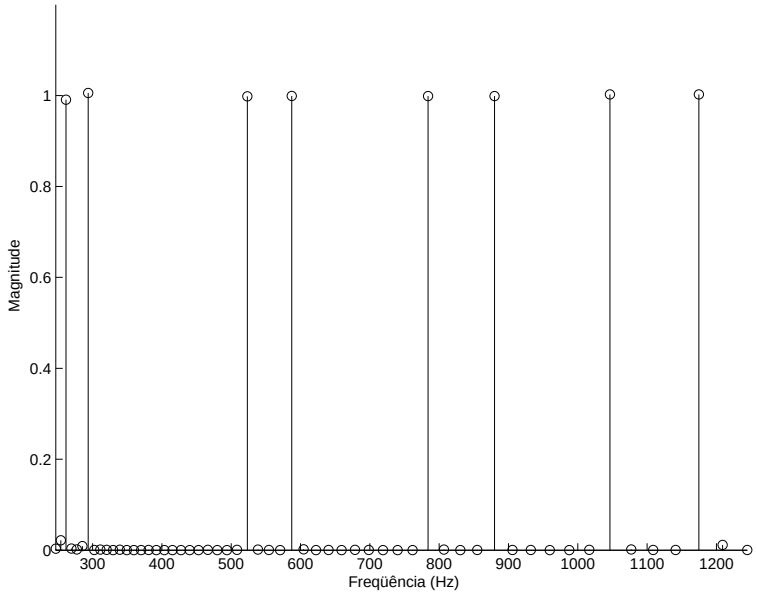


Figura 5.22: Análise por CQFFB do sinal de teste formado por dois tons. A distribuição freqüencial geométrica do CQFFT posiciona os canais de forma mais eficiente, e a alta seletividade do FFB evita o surgimento do "chão de ruído", mas com altíssima complexidade computacional.

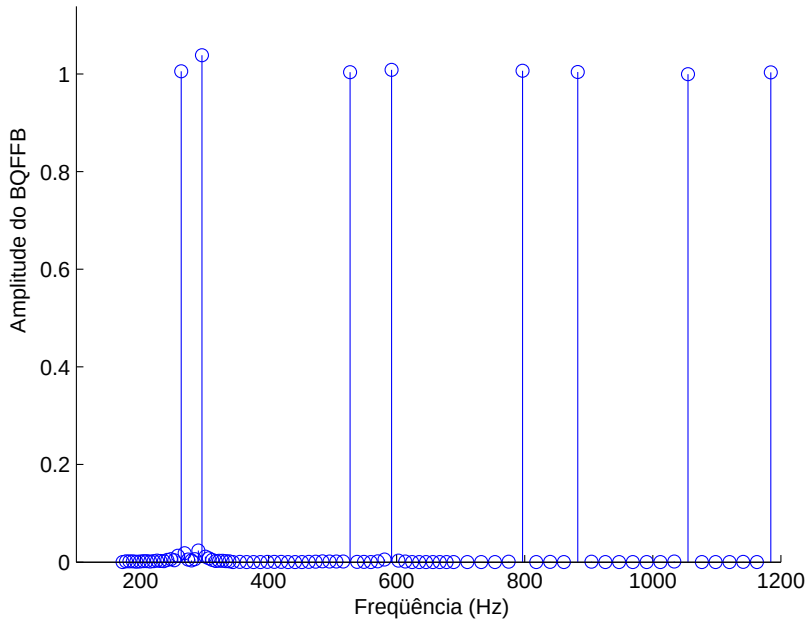


Figura 5.23: Análise por BQFFB do sinal de teste formado por dois tons. A distribuição freqüencial linear por partes do BQFFT posiciona os canais de forma eficiente, e a alta seletividade do FFB evita o surgimento do "chão de ruído", com a vantagem de apresentar complexidade computacional reduzida.

zido custo computacional, distribuição adequada de seus canais e alta seletividade entre canais adjacentes. Esta divide cada oitava do espectro linearmente em um número fixo de canais, mantendo assim uma distribuição quase-geométrica, o que seria adequado para aplicações musicais. Quanto à seletividade, esse método é comparável ao FFB. Além disso, o BQFFB emprega um número muito menor de canais do que os métodos de espaçamento puramente linear. Quanto ao desempenho, o BQFFB é comparável ao CQFFB, sobre o qual ainda apresenta uma complexidade cerca de 5 ordens de magnitude menor. O resultado é uma ferramenta robusta e eficiente para aplicações relacionadas à análise de sinais de música.

Foi feita uma análise qualitativa do desempenho do BQFFB como base de um sistema de TMA. Ao sistema foram submetidos sinais musicais com características escolhidas para avaliar sua resposta a problemas específicos envolvidos em TMA. Foi possível verificar que o BQFFB apresenta um bom desempenho nas altas freqüências, permitindo detectar visualmente o início e o fim das notas (a detecção automática de *onset* será implementada no sistema completo de TMA). Contudo, também foi visto que o desempenho em baixas freqüências fica prejudicado, em alguns casos específicos, dado que as estreitas bandas de passagem dos filtros com freqüências centrais baixas não são capazes de comportar as largas bandas dos transitórios. Efeitos como o trinado e o *glissando* foram bem analisados, com resultados bastante positivos.

Capítulo 6

Detecção de *onsets* e identificação de F0s usando o BQFFB

6.1 Introdução

No início do presente trabalho, foram apresentadas as etapas para a transcrição musical automática. No Capítulo 3, foram descritos métodos para a detecção de *onsets*, enquanto que no Capítulo 4, foi detalhado um sistema para detecção de múltiplas F0s. Com isso, seria possível obter-se uma forma de notação musical. Contudo, essas duas etapas sofrem do mesmo problema: as técnicas envolvidas trabalham no domínio da freqüência, atuando sobre canais pouco seletivos (devido ao fato de os filtros apresentarem ordens baixas) e muito numerosos (por serem linearmente espaçados).

Como uma solução para esse problema, foi apresentado, no Capítulo 5, o BQFFB. O fato de o BQFFB apresentar um número constante de canais de alta seletividade por oitava faz com que se obtenha uma resolução mais adequada para cada região do espectro. Isso fornece uma melhor definição do conteúdo de cada canal, o que pode favorecer os métodos que atuam sobre o domínio da freqüência.

Então, para se confirmar essa ferramenta como um recurso válido para fins de TMA, é feita uma série de testes descritos no presente capítulo. A Seção 6.2 mostra os resultados dos testes feitos ao se aplicar o BQFFB na etapa de detecção de *onsets*. A Seção 6.3 exibe os resultados dos testes realizados quando se aplicou o BQFFB na detecção de F0s e a Seção 6.4 conclui o capítulo.

6.2 Aplicação do BQFFB ao método de detecção de *onsets*

6.2.1 Metodologia

Para validar o emprego do BQFFB na detecção de *onsets*, deve-se aplicar os métodos descritos no Capítulo 3 substituindo a FFT pelo BQFFB. Todos os métodos já estudados para detecção de *onsets* serão reavaliados, por generalidade. O BQFFB utilizado neste capítulo é calculado para as oitavas de 2 a 9 e apresenta 32 canais por oitava (segundo o que foi especificado no Capítulo 5).

Como estabelecido antes, os métodos escolhidos para compor o teste de detecção de *onsets* foram os que atuam no domínio da freqüência, pois é neste domínio que o BQFFB apresenta vantagens. Métodos para detecção de *onsets* como, por exemplo, o de Klapuri [67], que divide o sinal de entrada em poucas bandas para posterior análise no domínio no tempo, não apresentam conexão com a ferramenta em questão.

Portanto, serão analisados os métodos que usam HFC, distância espectral, variação da fase e plano complexo, empregando tanto a FFT como o BQFFB. Para os testes, serão usados os mesmos sinais empregados anteriormente: os sinais sintéticos (uma nota por vez e um acorde por vez) e o sinal real definidos nas Seções 3.4.1 e 3.4.2, respectivamente.

Além disso, será feita uma análise da complexidade de cada um dos métodos de detecção de *onsets* através da contabilização das operações necessárias, tanto para a FFT como para o BQFFB, tendo como base o número de canais.

O sistema foi projetado de modo que permitisse uma fácil substituição da ferramenta de análise espectral. Os únicos parâmetros a serem reajustados de uma configuração (baseada em FFT) para a outra (baseada em BQFFB) são aqueles que controlam o desempenho da seleção de picos (ver Seção 3.2.1).

6.2.2 Resultados

O primeiro teste, portanto, consiste em empregar a FFT e o BQFFB aos métodos previamente descritos no Capítulo 3 utilizando como sinal de entrada um sinal sintético gerado a partir do protocolo MIDI satisfazendo a restrição de que ape-

nas uma nota seja executada por vez. Para esse teste, as curvas ROC relativas aos métodos de detecção de *onsets* a partir do BQFFB são mostradas na Figura 6.1 e comparadas com as curvas ROC geradas a partir desses mesmos métodos de detecção de *onsets* com base em FFT. Novamente, os parâmetros a serem variados para se obter a curva ROC foram λ e δ. Variou-se primeiramente δ fixando-se λ em 1. Os valores de δ usados foram aqueles que permitissem a taxa de detecção variar de zero a algo em torno de 100%. Quando o valor da taxa de detecção parava de variar indicava que não adiantava mais variar o valor de δ. Assim o valor de δ que fizesse o ponto da curva ROC se aproximar do ponto ótimo era fixado e, em seguida, variava-se o valor de λ. Com isso, formava-se a curva ROC. Deve-se deixar claro que a tolerância usada foi novamente de 50 ms. Os valores ótimos para taxa de detecção e taxa de falso alarme obtidos a partir das curvas ROC apresentadas na Figura 6.1 são mostrados na Tabela 6.1. Os valores apresentados nesta tabela foram obtidos através do mesmo procedimento para busca do ponto ótimo da curva ROC utilizado na Seção 3.4. Cabe aqui ressaltar que diferentemente do que é esperado na teoria de detecção de forma geral, aqui as curvas ROC podem não atingir 100% por duas razões: o limiar de detecção não é constante e a função de detecção pode não apresentar pico algum em determinado *onset*.

O método mais favorável ao emprego do BQFFB foi aquele baseado em distância espectral, no qual foi possível observar um sensível aumento na taxa de detecção, acompanhado de redução na taxa de falso alarme. Esse método é extremamente favorecido pela eficiente distribuição de canais e também pelo aumento da seletividade promovidos pelo BQFFB. Tendo uma definição maior de cada canal, causada pela redução da interferência entre eles, as distâncias espectrais são calculadas com maior acurácia.

É importante destacar que o desempenho do sistema de detecção de *onsets* a partir do BQFFB utilizando a distância espectral foi superior ao melhor resultado obtido com base em FFT através do método do domínio complexo, apresentando igual taxa de detecção e menor taxa de falso alarme. Enquanto o sistema com base em FFT obtinha $90,0\%$ de detecção e $17,1\%$ de falso alarme, com o BQFFB, obteve $92,5\%$ para a taxa de detecção e $10,3\%$ para o falso alarme. Para se ter uma idéia mais ampla da qualidade dos melhores métodos de detecção de *onsets* para a

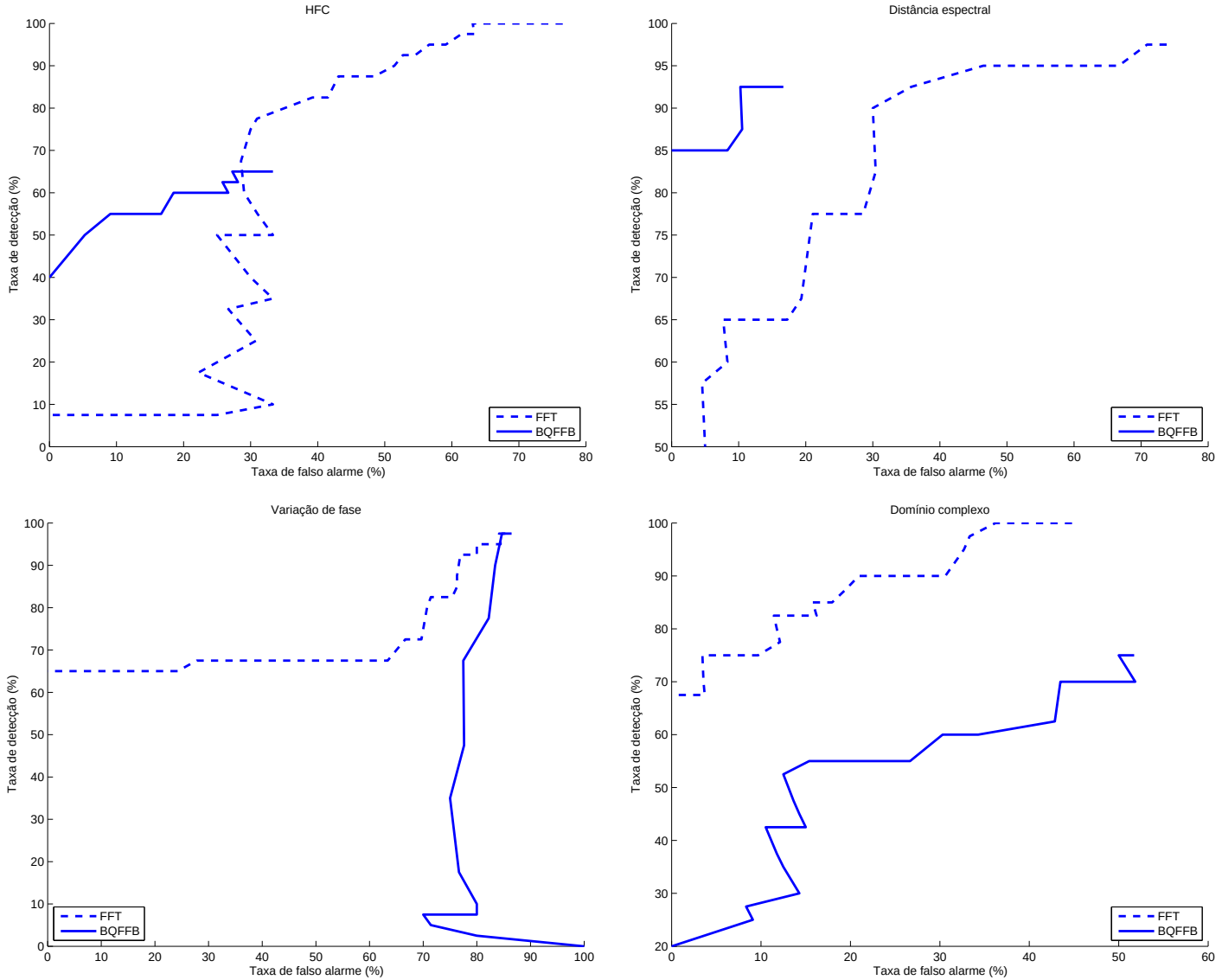


Figura 6.1: Comparação entre as curvas ROC relativas aos métodos de detecção de *onsets* a partir da FFT e do B*Q*FFB para o experimento sobre um sinal sintético em que se executa uma nota por vez. Os parâmetros variáveis usados para se obterem as curvas ROC foram λ e δ, contidos na Eq. (3.1).

Tabela 6.1: Comparação dos valores ótimos de taxa de detecção e de falso alarme para métodos de detecção de *onsets* em sinais sintéticos com uma nota por vez, utilizando FFT e B*Q*FFB.

Método	Taxa de detecção	Taxa de falso alarme
HFC com FFT	77,5%	30,9%
HFC com B*Q*FFB	60,0%	18,5%
Distância espectral com FFT	77,5%	21,1%
Distância espectral com B*Q*FFB	**92,5**%	**10,3**%
Desvio de fase com FFT	65,0%	0,0%
Desvio de fase com B*Q*FFB	90,0%	83,4%
Domínio complexo com FFT	**90,0**%	**17,1**%
Domínio complexo com B*Q*FFB	70,0%	23,5%

FFT e para o BQFFB, as curvas ROC relativas ao método da distância espectral (para BQFFB) e ao método do domínio complexo (para a FFT) são mostradas na Figura 6.2.

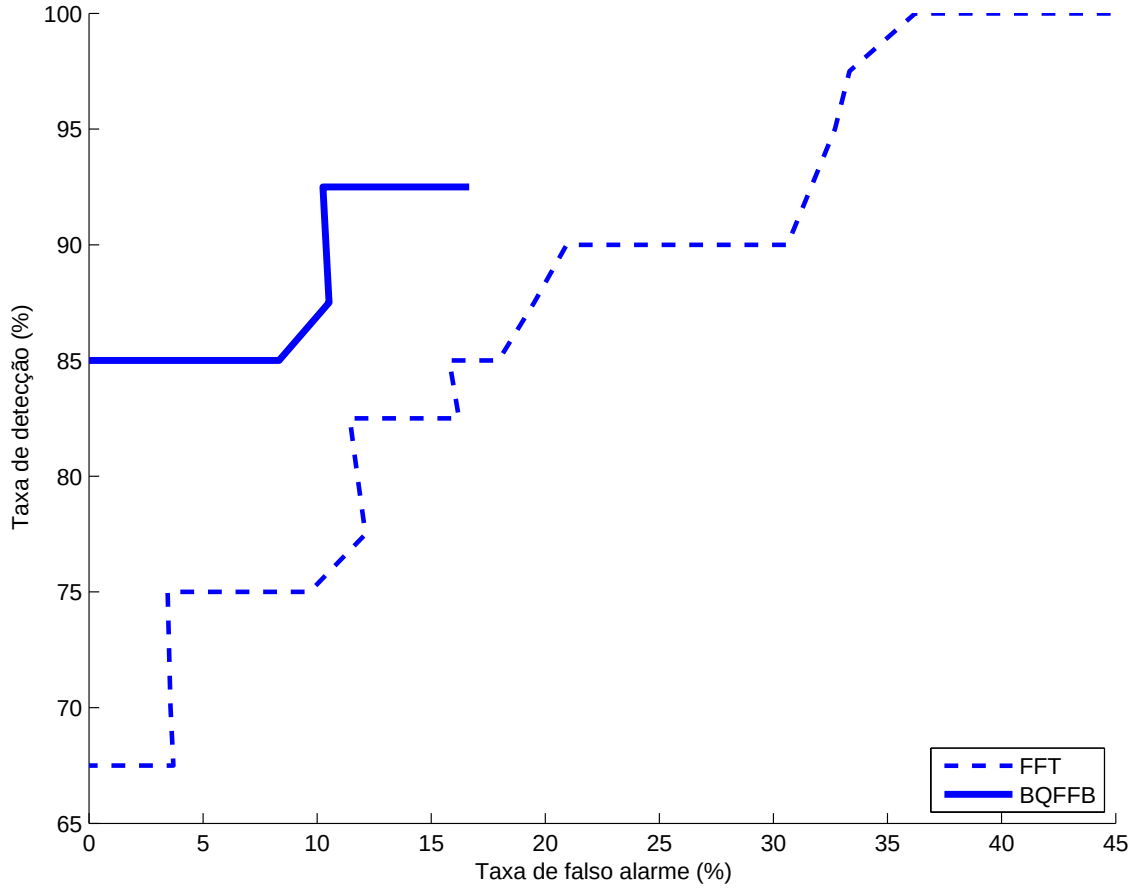


Figura 6.2: Comparação entre as curvas ROC relativas aos métodos que apresentam os melhores desempenhos para detecção de *onsets* a partir da FFT (domínio complexo) e do BQFFB (distância espectral) para o experimento sobre um sinal sintético em que se executa uma nota por vez. Os parâmetros variáveis usados para se obterem as curvas ROC foram λ e δ, contidos na Eq. (3.1).

A abordagem através da variação no desvio da fase já não havia apresentado um resultado favorável com a FFT devido à natureza do instrumento em questão. Com o emprego do BQFFB, o desempenho foi inferior, devido ao fato de ser necessária uma série de decimações para gerar a multi-resolução inerente do BQFFB (ver Figura 5.6). O problema é que essas decimações acabam por remover a linearidade da fase (o que foi inclusive comprovado por experimentação). Isso gera incompatibilidade com os métodos de detecção de *onsets* que fazem uso da fase. Cálculos pelo

plano complexo, conseqüentemente, também não obtiveram bons resultados.

O método baseado no conteúdo de alta freqüência, HFC, mostrou-se desfavorável à atual implementação do BQFFB. Isso pode ser entendido como uma conseqüência do fato de o BQFFB apresentar menos canais na região de altas freqüências, o que atuaria como uma atenuação da ponderação feita pelo vetor W usado. Poder-se-ia tentar formas mais agressivas de ponderação para sobrepujar essa atenuação; porém, como o método de HFC já não apresentava resultados tão bons quanto os obtidos através da distância espectral, julgou-se que isso não seria necessário.

A Figura 6.3 e a Tabela 6.2 referem-se, analogamente, ao caso de sinais sintéticos em que se executa um acorde por vez.

Tabela 6.2: Comparação dos valores ótimos de taxa de detecção e de falso alarme para métodos de detecção de *onsets* em sinais sintéticos com um acorde por vez, utilizando FFT e BQFFB.

Método	Taxa de detecção	Taxa de falso alarme
HFC com FFT	87,5%	32,7%
HFC com BQFFB	77,5%	14,3%
Distância espectral com FFT	77,5%	37,8%
Distância espectral com BQFFB	**87,5**%	**5,6**%
Desvio de fase com FFT	85,0%	5,9%
Desvio de fase com BQFFB	95,0%	85,2%
Domínio complexo com FFT	**87,5**%	**8,8**%
Domínio complexo com BQFFB	72,5%	27,8%

Novamente, o método mais favorável ao emprego do BQFFB foi aquele baseado em distância espectral, no qual foi possível observar um sensível aumento na taxa de detecção, acompanhado de redução na taxa de falso alarme. Esse método é extre-

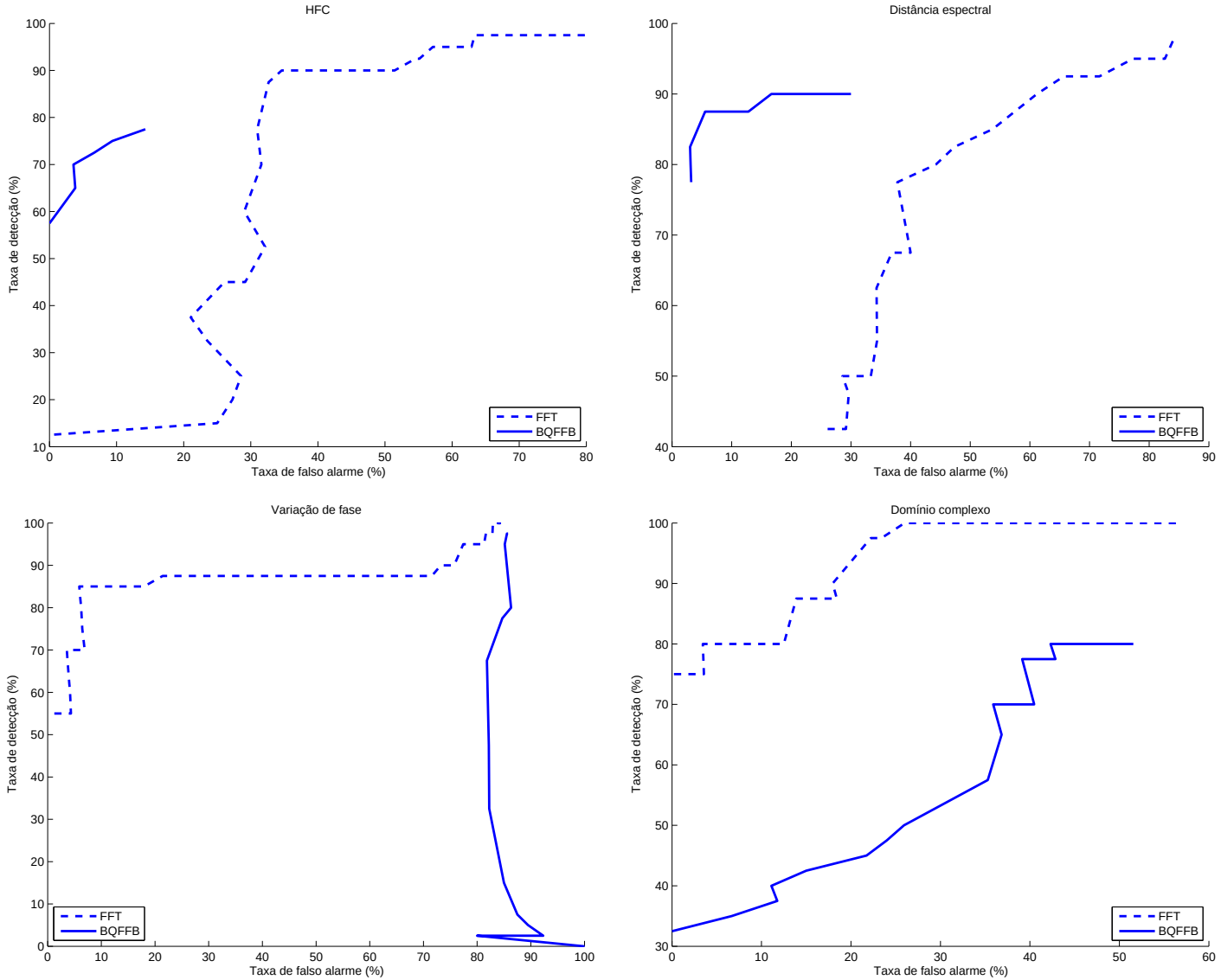


Figura 6.3: Comparação entre as curvas ROC relativas aos métodos de detecção de *onsets* a partir da FFT e do B*Q*FFB para o experimento sobre um sinal sintético em que se executa um acorde por vez. Os parâmetros variáveis usados para se obterem as curvas ROC foram λ e δ, contidos na Eq. (3.1).

mamente favorecido pela eficiente distribuição de canais assim como pelo aumento da seletividade promovidos pelo BQFFB. O desempenho do sistema de detecção de *onsets* a partir do BQFFB utilizando a distância espectral foi superior ao melhor resultado obtido com base em FFT através do método do domínio complexo, apresentando maior taxa de detecção e menor taxa de falso alarme. Enquanto o sistema com base em FFT obtinha $87,5\%$ de detecção e $8,8\%$ de falso alarme, com o BQFFB obteve 87,5% de detecção e 5,6% de falso alarme.

Para se ter uma idéia mais ampla da qualidade dos melhores métodos de detecção de *onsets* para o caso de um acorde por vez ao se usar a FFT e o BQFFB, as curvas ROC relativas ao método da distância espectral (para BQFFB) e ao método do domínio complexo (para a FFT) são mostradas na Figura 6.4.

Assim como foi mencionado no caso onde se executa uma nota por vez, a abordagem através da variação no desvio da fase obteve desempenho inferior, devido ao fato de as decimações necessárias para gerar a multi-resolução removerem a linearidade da fase. Cálculos pelo plano complexo, conseqüentemente, também não obtiveram bons resultados. Mais uma vez, o método baseado no conteúdo de alta freqüência, HFC, mostrou-se desfavorável à atual implementação do BQFFB.

Por fim, a Tabela 6.3 mostra os resultados da aplicação dos métodos de detecção de *onsets* comparando o emprego da FFT com o do BQFFB quando se analisa um sinal real. O sinal é da mesma gravação utilizada no Capítulo 3: a Ária das Variações de Goldberg executada pelo pianista Glenn Gould (gravação de estúdio de 1955).

No caso real, não houve ajuste dos parâmetros de seleção de pico δ e λ (apresentados na Seção 3.2.1) do modo como foi feito nos casos com sinais sintéticos. Os resultados do caso real foram obtidos com os mesmos limiares calculados para o caso sintético de um acorde por vez. Enquanto que, a partir da FFT, os métodos de HFC e variação de fase apresentaram taxa de detecção em torno de 15%, nenhum método obteve menos de 70% quando baseou-se na BQFFB. Daí, pode-se dizer que os métodos para detecção de *onsets* com base em BQFFB podem ser mais estáveis, permitindo maior grau de previsibilidade. Contudo, a taxa de falso alarme obtida com base no BQFFB foi mais alta do que a obtida com a FFT.

Para o caso específico do método da distância espectral, se as taxas de detecção de ambas as ferramentas forem equalizadas para permitir comparação, a taxa de

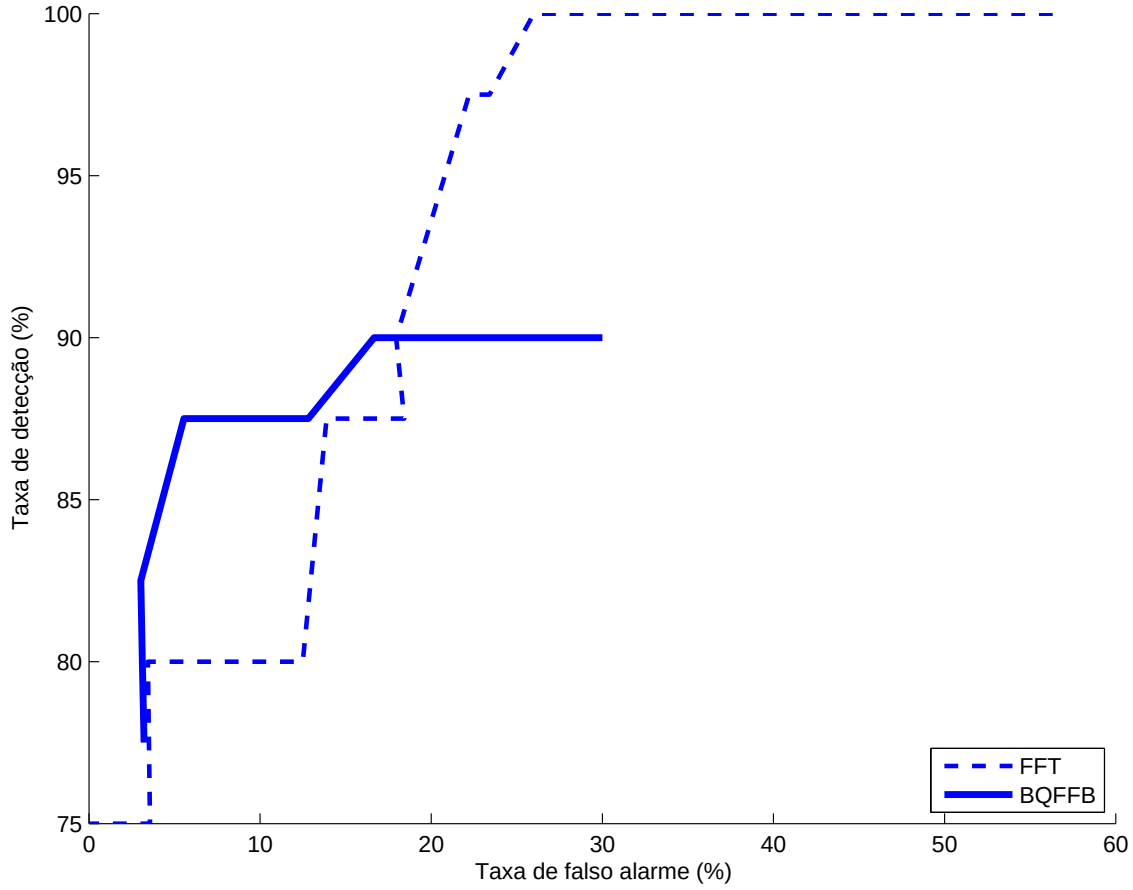


Figura 6.4: Comparação entre as curvas ROC relativas aos métodos que apresentam os melhores desempenhos para detecção de *onsets* a partir da FFT (domínio complexo) e do BQFFB (distância espectral) para o experimento sobre um sinal sintético em que se executa um acorde por vez. Os parâmetros variáveis usados para se obterem as curvas ROC foram λ e δ, contidos na Eq. (3.1).

Tabela 6.3: Comparação dos resultados para métodos de detecção de *onsets* em sinais reais, utilizando FFT e BQFFB.

Método	Taxa de detecção	Taxa de falso alarme
HFC com FFT	16,9%	6,7%
HFC com BQFFB	71,1%	42,1%
Distância espectral com FFT	**75**,**9**%	**8**,**8**%
Distância espectral com BQFFB	**80**,**7**%	**31**,**5**%
Desvio de fase com FFT	15,7%	0,0%
Desvio de fase com BQFFB	97,6%	81,8%
Domínio complexo com FFT	54,2%	18,9%
Domínio complexo com BQFFB	84,3%	34,6%

falso alarme para a FFT é aproximadamente a metade da obtida pelo BQFFB. Essa diferença é provavelmente devida à detecção de picos espúrios provocados por ecos e reverberações. Como, ao se usar o BQFFB, tem-se maior definição do conteúdo de cada canal devido à alta seletividade, fica-se mais sujeito a falsos alarmes.

Para contornar esse problema, pode-se lançar mão de pré-processamento para remoção de ecos, reverberações e ruídos. Com isso seria possível um melhor desempenho do BQFFB para detecção de *onsets*, pois se estaria trabalhando com uma ferramenta de alta seletividade sobre sinais que foram submetidos à supressão de ruído, obtendo-se um sinal resultante com melhor resolução (devido à alta seletividade) e menos corrompido. De qualquer forma, em relação ao desempenho final do sistema de TMA, um *onset* falsamente detectado pode ser tratado posteriormente com algum tipo de pós-processamento. Portanto, neste caso, uma perda seria mais grave que um falso alarme.

Na verdade, o ideal para se verificar o desempenho do sistema de TMA proposto sobre um sinal real seria ajustar os parâmetros do sistema usando-se sinais reais. Dessa forma, o sistema estaria preparado para quaisquer efeitos de ecos, reverberações e ruídos presentes nos sinais sob análise, obtendo-se um desempenho melhor. Contudo, como apontado por [15], existe nessa área de pesquisa uma grande dificuldade para se obterem sinais de teste reais acompanhados por transcrições feitas de forma não-automática. Por esse motivo, para se ter uma noção do desempenho ótimo do sistema para o sinal real descrito nesta seção, pode-se propor ajustar os parâmetros da detecção de *onsets* para o próprio sinal.

Com esse ajuste, o método de detecção de *onsets* usando-se distância espectral a partir da FFT conseguiu a taxa de detecção de 87% e a taxa de falso alarme de 11%. Já, a partir do BQFFB, foi obtida a taxa de detecção de 89% e falso alarme de 16%. Isso mostra desempenhos similares ao se compararem as duas ferramentas de análise espectral.

6.2.3 Análise de complexidade

A Tabela 6.4 mostra a comparação entre a complexidade dos métodos para detecção de *onsets* ao se empregar FFT e BQFFB.

Pode-se ver que todos os métodos têm a complexidade extremamente reduzida ao

Tabela 6.4: Análise de complexidade para métodos de detecção de *onsets*, utilizando FFT e BQFFB.

Método	No. de adições por *frame*	No. de multiplicações por *frame*
HFC com FFT	1104	2211
HFC com BQFFB	287	577
Distância espectral com FFT	3314	2210
Distância espectral com BQFFB	863	576
Desvio de fase com FFT	3314	1106
Desvio de fase com BQFFB	863	289
Domínio complexo com FFT	6628	3316
Domínio complexo com BQFFB	1726	865

se trocar a FFT pelo BQFFB. Isso acontece pelo fato de essa ferramenta trabalhar com um número muito menor de canais. Mesmo sabendo-se que a FFT por si só é muito menos complexa que o BQFFB, o ganho que se obtém com a redução do número de canais faz com que o custo computacional como um todo seja reduzido. Contudo, a detecção dos *onsets* representa a menor parte dos cálculos envolvidos em um sistema de TMA, sendo a maior parte representada pela detecção das F0s. Este assunto será tratado na próxima seção, na qual haverá um estudo do tempo de execução.

6.3 Aplicação do BQFFB ao método de identificação de F0s

6.3.1 Metodologia

Nesta seção, será apresentado o emprego do algoritmo de Klapuri sobre o espectro gerado pelo BQFFB e serão discutidos os resultados referentes aos testes feitos sobre sinais de entrada apresentando de 1 a 4 tons. Os testes foram realizados para as três intensidades disponíveis: suave, média e forte (que são identificadas, na base MUMS como *soft*, *medium* e *loud*, respectivamente).

Um questionamento válido seria sobre a necessidade de se recorrer a um banco de filtros de alta-seletividade, caso já se empregasse uma distribuição mais eficiente para os canais no domínio da freqüência. Somente essa nova distribuição já não seria suficiente para se obter o mesmo desempenho de uma análise espectral baseada em FFT, porém com uma grande redução na complexidade? Para responder a esse tipo de questão, serão também apresentados testes envolvendo a BQT, que, como foi explicado anteriormente, representa uma versão de baixa seletividade do BQFFB. Dessa forma, é possível verificar isoladamente o efeito das duas facetas do BQFFB: a distribuição eficiente de canais no domínio da freqüência e a alta seletividade.

Os testes feitos serão os mesmos utilizados para se testar o método original de Klapuri no Capítulo 4. Primeiramente, serão feitos os testes para BQT e BQFFB informando-se ao sistema a quantidade de tons presentes no sinal de entrada, ou seja, sem se aplicar nenhum critério de parada para o algoritmo iterativo de identificação

de F0s. Por fim, serão apresentados os resultados decorrentes da aplicação do critério de parada.

6.3.2 Resultados

Os experimentos foram os mesmos feitos para o caso da FFT no Capítulo 4. Por conveniência, os resultados relativos à FFT serão repetidos nas Tabelas 6.5 e 6.6. A

Tabela 6.5: Resultados obtidos com base em FFT nos testes quando o sinal de entrada apresenta mais de uma F0.

Intensidade	N^o de Notas	% Identificação
suave	1	88,3 %
	2	70,0 %
	3	60,0 %
	4	56,7 %
média	1	86,7 %
	2	70,0 %
	3	60,6 %
	4	58,8 %
forte	1	90,0 %
	2	67,5 %
	3	63,9 %
	4	63,8 %

compatibilização do processo de seleção harmônica para as ferramentas do tipo BQ foi possível porque os cálculos (para se selecionar canais que seriam somados) foram feitos com base nos valores de freqüência. Na verdade, em vez de fazer cálculos com base no índice do canal (obtendo-se um índice), calculavam-se os canais a serem selecionados com base em um valor de freqüência (obtendo-se uma freqüência). Tendo-se esta freqüência, buscava-se o índice do canal cuja freqüência central era a mais próxima daquela freqüência obtida. Não foi usado um procedimento de ponderação análogo ao da Seção 5.6.2 (que mapeava os canais do BQFFB em notas da escala de temperamento igual), pois a complexidade se tornaria muito elevada. O processo de integração das sub-bandas se deu da mesma forma: calculavam-se as

Tabela 6.6: Resultados obtidos com base em FFT nos testes utilizando-se o critério de parada.

Intensidade	Nº de Notas	Identificação	Falso Alarme	Substituição	Perda
suave	1	91,7 %	47,8 %	8,3 %	0,0 %
	2	69,2 %	26,9 %	25,8 %	5,0 %
	3	65,3 %	27,0 %	24,5 %	10,2 %
	4	60,8 %	17,3 %	32,9 %	17,3 %
forte	1	95,0 %	65,5 %	5,0 %	0,0 %
	2	76,7 %	47,1 %	21,7 %	1,7 %
	3	67,2 %	28,6 %	31,1 %	1,7 %
	4	67,1 %	20,6 %	30,8 %	2,1 %

freqüências das parciais a serem somadas e, depois, buscavam-se os índices.

A Tabela 6.7 resume os resultados encontrados empregando-se uma versão do método de Klapuri para identificação de F0s baseado em BQT, enquanto que a Tabela 6.8 é relativa ao emprego deste mesmo método com o BQFFB. Pode-se ver,

Tabela 6.7: Resultados obtidos nos testes da aplicação da BQT ao método de Klapuri quando o sinal de entrada apresenta de 1 a 4 F0s, sem critério de parada.

Intensidade	Nº de Notas	Identificação
suave	1	40,0 %
	2	25,0 %
	3	19,4 %
	4	17,9 %
média	1	41,7 %
	2	26,7 %
	3	21,1 %
	4	17,5 %
forte	1	45,0 %
	2	30,0 %
	3	21,1 %
	4	18,8 %

Tabela 6.8: Resultados obtidos nos testes da aplicação do BQFFB ao método de Klapuri quando o sinal de entrada apresenta de 1 a 4 F0s, sem critério de parada.

Intensidade	N^o de Notas	% Identificação
suave	1	98,3 %
	2	68,3 %
	3	46,7 %
	4	37,9 %
média	1	96,7 %
	2	74,2 %
	3	50,6 %
	4	42,9 %
forte	1	96,7 %
	2	73,3 %
	3	54,4 %
	4	41,2 %

na Tabela 6.7, que os resultados utilizando-se apenas a BQT foram inferiores aos resultados apresentados empregando-se a FFT (na forma original do método), que foram apresentados na Tabela 6.5. Isso acontece porque, de acordo com o método de identificação de F0s de Klapuri, a identificação correta de uma F0 depende muito dos canais cuja freqüência central é mais alta. O que acontece nos canais mais altos tem grande influência sobre a capacidade de se identificar F0s de menor freqüência, já que é feita uma soma envolvendo os canais relativos a cada parcial harmônica. Na BQT, as oitavas mais altas apresentam a mesma seletividade da FFT, porém com resolução muito mais baixa. Assim, a detecção de F0s nas oitavas mais baixas é prejudicada pela grande quantidade de interferência inter-canal e baixa resolução das oitavas mais altas. Contudo, como foi explicado, os maiores erros acontecem nas oitavas mais altas devido ao fato de estas apresentarem, ao mesmo tempo, baixa resolução e baixa seletividade.

A solução para esse problema apresentado pelo uso da BQT é se recorrer a uma ferramenta de análise espectral que apresente boa seletividade. Ao se utilizar o BQFFB, sua alta seletividade compensou o efeito adverso causado pelo aumento

sucessivo da largura dos canais. Isso, então, possibilitou maior eficiência na identificação das notas, o que pode ser visto na Tabela 6.8.

A taxa de identificação do BQFFB, para o caso onde existe apenas uma nota presente na mistura, foi cerca de 10% superior ao da FFT. Para 2 notas presentes na mistura, os desempenhos do BQFFB e da FFT foram similares. A primeira nota da mistura é geralmente identificada corretamente, enquanto que a identificação da segunda nota é prejudicada. O fato de que, de 1 para 2 notas, há uma queda maior de desempenho no BQFFB (ao se comparar com a FFT) indica que há uma dificuldade extra na etapa da remoção do espectro associado à F0 mais recentemente identificada. Um detalhe importante em relação à remoção do espectro suavizado é a vizinhança em torno dos pontos do espectro suavizado, o que é ilustrado na Figura 6.5. Na Figura 6.5(a), são mostrados dois pontos do espectro suavizado. Este espectro deve ser subtraído do espectro sob análise. Na Figura 6.5(b), é mostrado o que deve ser feito no caso da FFT. Para a FFT, a remoção do espectro relativo à F0 mais recentemente identificada deve ser feita usando-se uma vizinhança constante de 6 amostras em torno de cada ponto do espectro suavizado. Para o caso do BQFFB, exibido na Figura 6.5(c), esta vizinhança é variável devido à resolução variável desta ferramenta. Para as oitavas 2, 3 e 4, a vizinhança é de 3 amostras; para a oitava 5, a vinhança é de 2 amostras; e para as demais oitavas, a vizinhança é de 1 amostra. Contudo, em vez de a vizinhança ser constante (como no caso da FFT), para o BQFFB, é mais indicado que seja multiplicada por uma janela de suavização (usou-se a de Hanning de tamanho igual ao indicado para a vizinhança) em torno de cada ponto de interesse do espectro suavizado. Chegou-se a estes valores por experimentação.

Para 3 e 4 notas na mistura do sinal de entrada, o BQFFB apresentou desempenho inferior ao da FFT porque o erro cometido na remoção do espectro das F0s identificadas era propagado para as demais iterações do algoritmo. Deverá ser feito um estudo mais aprofundado da utilização da vizinhança em torno dos pontos do espectro suavizado para o BQFFB. Entretanto, a superioridade do desempenho do BQFFB sobre sinais com apenas 1 nota e a similaridade de desempenho para 2 notas mostram o potencial do BQFFB para identificação de F0s.

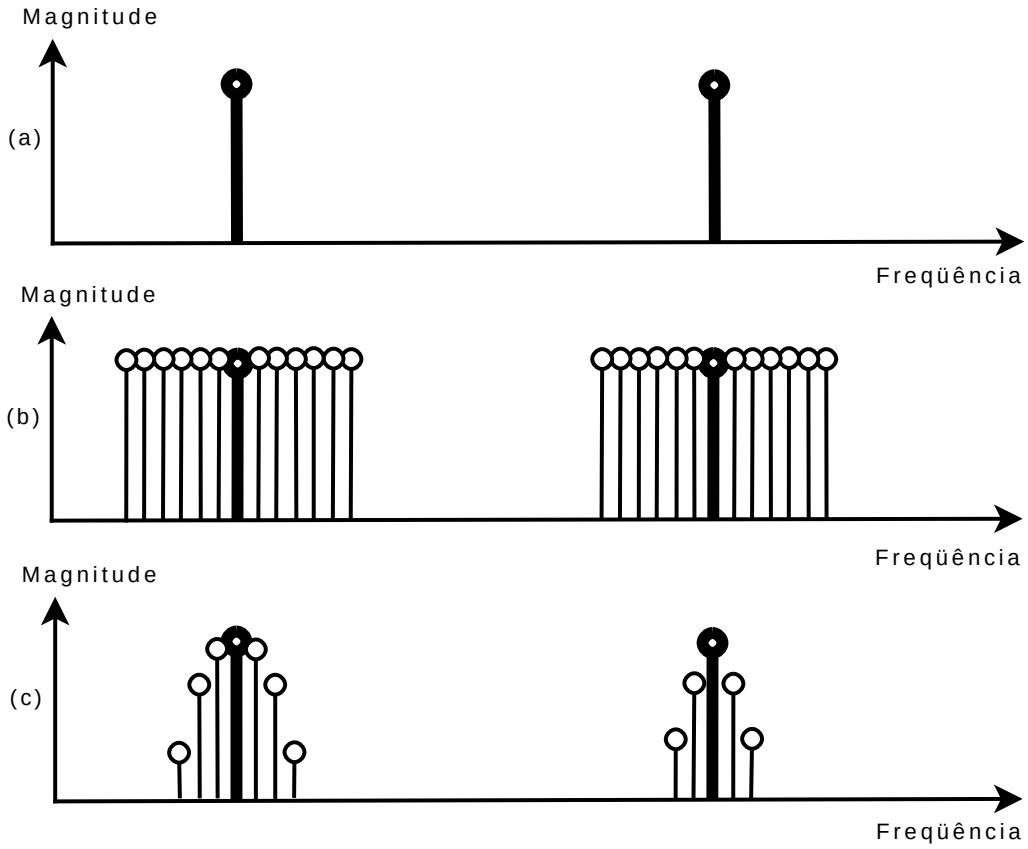


Figura 6.5: Preparação do espectro suavizado através da consideração de vizinhança em torno dos pontos definidos para o espectro suavizado. (a) Espectro suavizado original. (b) Vizinhança utilizada no caso da FFT. (c) Vizinhança utilizada no caso do BQFFB.

6.3.3 Critério de parada

Até este ponto, foram feitos testes para verificar a capacidade de ferramentas de análise espectral do tipo BQ identificarem múltiplas F0s informando-se ao sistema a quantidade de tons presentes no sinal de entrada. Agora se torna necessário analisar como esse sistema se comporta se essa informação não for fornecida. Para isso, deve-se, assim como foi feito no Capítulo 4, adotar um critério de parada para o algoritmo iterativo de detecção de F0s. O ajuste do limiar de detecção será feito com base nas notas de intensidade média, enquanto que o teste do sistema será feito com as intensidades suave e forte.

Mas, antes disso, assim como foi feito no Capítulo 4, deve-se definir para cada ferramenta o valor do limiar de detecção que indicará o momento em que não existem mais F0s a serem detectadas. É importante lembrar que a figura de mérito a ser utilizada para cálculo desse critério de parada é o peso global máximo normalizado, descrito na Seção 4.5. Executando-se o mesmo tipo de teste feito anteriormente, chegou-se, para o sistema baseado em BQFFB, ao valor de limiar igual a 0,18687. Isso foi feito igualando-se a taxa de falso alarme à taxa de perda, chegando-se a um valor estimado de 11% para ambas as grandezas. Isso, segundo a curva ROC em questão, produz uma taxa de detecção para o sistema baseado em BQFFB igual a 54%. As taxas de falso alarme e de perda para a FFT ficaram em torno de 10%, enquanto que, para as ferramentas de alta seletividade, essas taxas ficaram em torno de 12%. A taxa de detecção está mais baixa que aquela obtida pela FFT devido ao desempenho do sistema com base em BQFFB para os casos de 3 e 4 notas. Para esses casos, a taxa de detecção do BQFFB sofreu as conseqüências do desempenho do sistema de remoção do espectro relativo à F0 mais recentemente detectada. Isso pode ser concluído com base no forte desempenho do sistema para sinais de entrada contendo um tom em face ao fraco desempenho para maiores graus de polifonia.

Deve-se lembrar que esses valores foram calculados sobre todos os valores possíveis de polifonia (de 1 a 4 F0s). Portanto, cada caso individual apresentará outros valores para taxa de perda e falso alarme, cuja média (considerando a totalidade dos casos) é igual ao valor estimado. A Tabela 6.9 resume os resultados encontrados empregando-se o método de Klapuri baseado em BQFFB com critério de parada.

Contudo, ao se comparar com a FFT, percebe-se que os resultados obtidos com o BQFFB para 1 nota presente no sinal de entrada são superiores aos da FFT. A FFT chegou a resultados próximos a 93% nos casos monofônicos ao custo de uma alta taxa de falso alarme (48% para a intensidade suave e 65% para a intensidade forte). Além de se obter uma taxa de 100% com o BQFFB, a taxa de falso alarme é reduzida nas ferramentas de alta seletividade, onde se obteve 50% para a intensidade forte mas apenas 29% para a intensidade suave, representando uma melhora em cerca de 15%. Isso, no entanto, é uma conseqüência direta do ponto em que se calibrou o limiar de detecção para o critério de parada.

Para o caso de 2 notas presentes no sinal de entrada, a FFT e o BQFFB apresentam resultados similares quanto à taxa de identificação, em torno de 70%. Porém foi verificada uma vantagem do BQFFB quanto às taxas de falso alarme, que foram reduzidas a quase a metade dos valores obtidos pela FFT.

Tabela 6.9: Resultados obtidos nos testes da aplicação do BQFFB ao método de Klapuri quando o sinal de entrada apresenta de 1 a 4 F0s, com critério de parada.

Intensidade	Nº de Notas	Identificação	Falso Alarme	Substituição	Perda
suave	1	100,0%	29,4%	0,0%	0,0%
	2	64,2%	11,6%	25,0%	10,8%
	3	42,2%	6,8%	41,7%	16,1%
	4	32,1%	2,2%	43,8%	24,2%
forte	1	98,3%	49,6%	1,7%	0,0%
	2	70,8%	27,8%	24,2%	5,0%
	3	52,8%	13,9%	40,0%	7,2%
	4	38,7%	5,0%	49,2%	12,1%

6.3.4 Análise da complexidade

Ao se substituir o número de canais usados em cada caso nas equações de complexidade do método, é possível montar a Tabela 6.10. Observa-se que a FFT, apesar de apresentar uma complexidade bastante reduzida, faz com que o sistema de identificação de F0s precise trabalhar com uma quantidade muito elevada de canais. Como o custo computacional do método de Klapuri é proporcional ao quadrado

do número de canais, essa característica se torna fundamental na determinação da complexidade total de cada caso apresentado na Tabela 6.10.

Analisando esta tabela, pode surgir um questionamento sobre o fato de os métodos BQ necessitarem de menos multiplicações que a FFT. Isso se dá principalmente porque com a FFT é necessário executar a análise espectral sobre todo o espectro, enquanto que, como as ferramentas do tipo BQ operam com base na divisão do espectro em oitavas, permitem realizar os cálculos com base em uma faixa mais estreita (apenas nas oitavas de interesse).

Tabela 6.10: Análise comparativa de complexidade (número de multiplicações) do método de Klapuri para identificação de múltiplas F0s ao se variar a ferramenta de análise espectral.

Base do método	Análise espectral	Identificação de F0s	Total
FFT	8192	137313	145505
BQT	2632	10752	13384
BQFFB	3072	10752	13824

Nota-se que há uma grande vantagem no uso de ferramentas do tipo BQ de alta seletividade no que se refere à complexidade computacional. Pode-se dizer que é possível reduzir em cerca de dez vezes a complexidade computacional do sistema de identificação de F0s.

6.4 Conclusão

Nesse capítulo, foi visto que a troca da FFT pelo BQFFB em ambas as etapas da transcrição (detecção de *onsets* e na identificação de freqüências fundamentais) traz ganhos em relação à redução do custo computacional (que é cerca de 10 vezes menor) e apresenta um desempenho comparável ao sistema baseado em FFT (para 1 e 2 notas presentes no sinal de entrada) apesar de apresentar um desempenho inferior para os graus de polifonia 3 e 4.

Na detecção de *onsets*, para sinais sintéticos, ambas as ferramentas permitiram atingir uma taxa de detecção de cerca de 90% e uma taxa de falso alarme de cerca de 10%. Para o sinal real em questão, foram obtidas taxas de detecção comparáveis

para ambas as ferramentas (em torno de 80%), enquanto que o BQFFB apresentou maior taxa de falso alarme.

Na identificação de F0s, ao se informar a quantidade de F0s ao sistema, o sistema com base em BQFFB obteve a taxa de identificação máxima de 98% (em comparação com 88% do sistema baseado em FFT). Ao se empregar o critério de parada, o sistema com base em BQFFB obteve a taxa de identificação máxima de 100% (em comparação com a taxa de 95% da FFT). Como conseqüência do uso do critério de parada, o falso alarme foi de, no máximo, 50% (em comparação com 65% do sistema baseado em FFT).

Mesmo com o desempenho do BQFFB mais fraco que o da FFT para 3 e 4 notas, existe ainda um real potencial de melhora, visto que o desempenho do BQFFB para 1 nota é acentuadamente superior. Isso indica que o problema pode estar concentrado na forma da remoção do espectro suavizado relativo à F0 mais recentemente identificada, o que será alvo de estudo futuro para melhoria do sistema. Pode-se realizar testes para identificação de F0s através de outros algoritmos que não façam uso de remoções iterativas do espectro relativo a F0 identificada. O bom desempenho nos casos onde há apenas 1 nota contida no sinal de entrada sugere que se busquem outros tipos de abordagem para identificação de F0s mais favoráveis ao BQFFB.

De qualquer forma, o objetivo do capítulo era verificar se seria válida a aplicação do BQFFB como base para a transcrição musical automática. Tendo validado a utilização dessa ferramenta (pelo menos para polifonias de até 2 graus) para análise espectral em cada etapa individualmente, deve-se agora testar o sistema como um todo. Isso será feito no próximo capítulo.

Capítulo 7

Sistema de transcrição musical automática com base em BQFFB

7.1 Introdução

O Capítulo 6 mostrou que o BQFFB é uma ferramenta válida para análise espectral e suas vantagens e potenciais de melhoria foram evidenciados pelos experimentos realizados. Este capítulo visa a apresentar um sistema completo de TMA com base no BQFFB. A Seção 7.2 explica o funcionamento do sistema implementado. Na Seção 7.3, o ajuste dos parâmetros que controlam o desempenho do sistema é discutido e os resultados obtidos nos experimentos computacionais são apresentados e analisados, enquanto que na Seção 7.4 são feitas comparações com o desempenho de um sistema comercial. A Seção 7.5 conclui o presente capítulo.

7.2 Método implementado

Esta seção objetiva mostrar o funcionamento do sistema de TMA implementado. A entrada do sistema é um arquivo WAV contendo um sinal de áudio em mono (1 canal amostrado a 44100 Hz e 16 *bits*) e a saída é um arquivo MIDI (ver Apêndice B).

O funcionamento do sistema é um desdobramento do que foi apresentado na Seção 1.1 e é ilustrado na Figura 1.1. Contudo, algumas etapas foram adicionadas a fim de tratar problemas de ordem prática. Isso pode ser visto no fluxograma da Figura 7.1, que mostra o sistema proposto dividido em três módulos: um para

análise espectral, outro para detecção de *onsets* e um terceiro para detecção de F0s. O primeiro módulo é o responsável por gerar a representação tempo-freqüência que, por sua vez, servirá como matéria-prima para os dois outros módulos. A seguir, serão detalhadas as etapas que formam cada módulo. Esse módulo foi ajustado de forma específica para piano, que é um instrumento com ataques de nota bem definidos e que tipicamente decaem em intensidade ao longo do tempo.

7.2.1 Módulo para análise espectral

Este módulo tem como entrada o sinal contido em um arquivo de áudio em formato WAV e, como mencionado anteriormente, tem como produto a representação tempo-freqüência utilizada pelos demais módulos.

A primeira etapa desse módulo é a normalização do sinal de entrada. Isso é feito dividindo-se cada amostra pelo valor da máxima magnitude desse sinal. Com isso, há melhor aproveitamento da faixa dinâmica e se garante maior previsibilidade dos valores usados pelos algoritmos envolvidos para que haja coerência com os ajustes feitos na configuração do sistema.

Em seguida, submete-se o sinal normalizado ao BQFFB, cuja implementação é detalhada no Capítulo 5. Contudo, a representação tempo-freqüência neste ponto contém, para cada filtro, um sinal de saída que apresenta um transitório temporal.

Para que seja possível a correta equalização das bases de tempo de cada oitava, esse transitório deve ser removido. Isso é feito retirando-se T_t amostras do início e do fim do sinal de saída de cada filtro, onde T_t é igual à metade do comprimento da resposta impulsional do filtro.

7.2.2 Módulo para detecção de *onsets*

Este módulo tem como entrada a representação tempo-freqüência gerada pelo módulo de análise espectral e tem como produto o instante de *onset* de cada nota presente no sinal de entrada. Para que isso seja possível, é necessário o ajuste de alguns parâmetros:

- Tipo do método de detecção de *onsets*: implementado segundo uma das quatro opções apresentadas no Capítulo 3, ou seja, HFC, distância espectral, variação

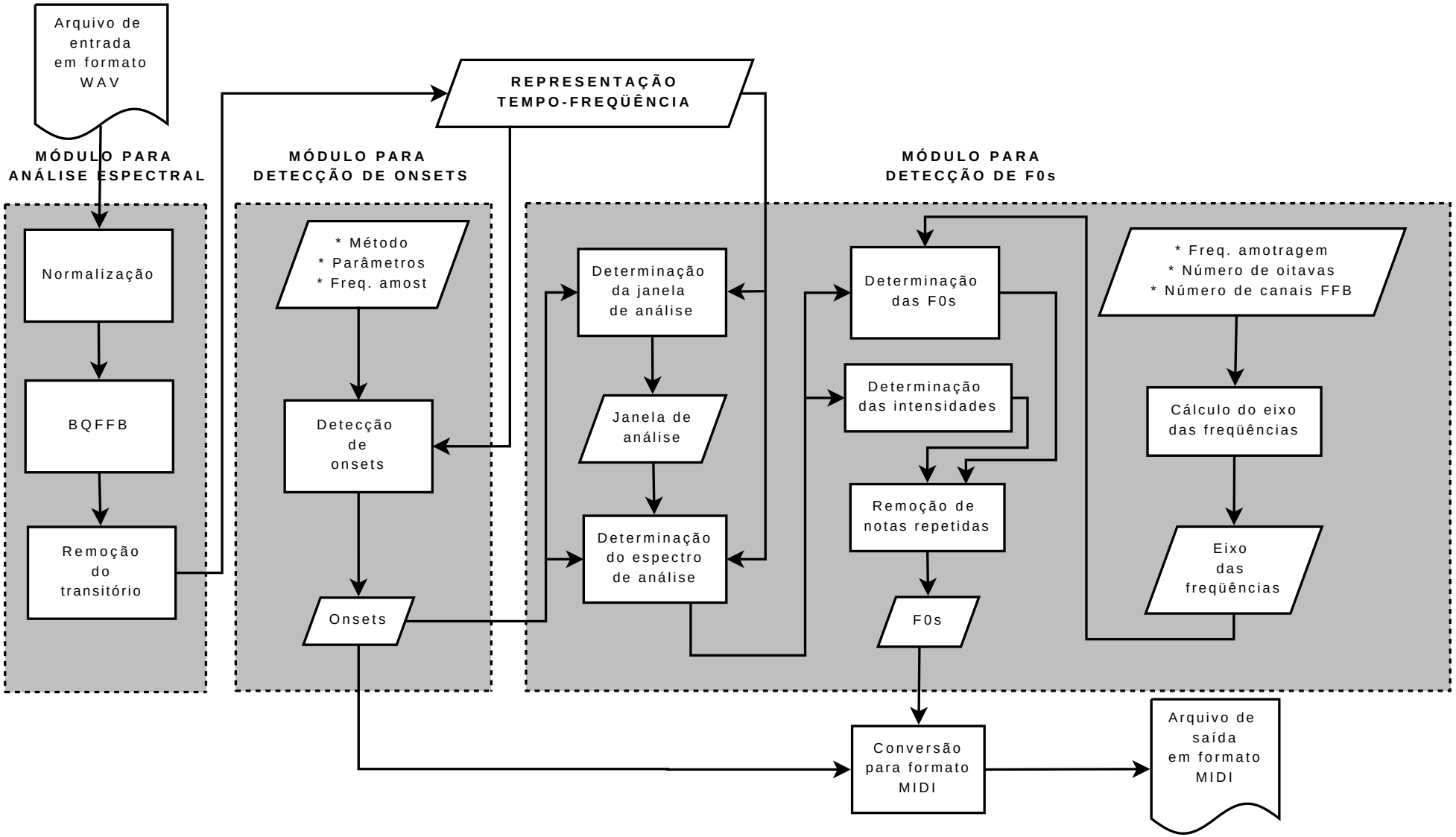


Figura 7.1: Fluxograma de funcionamento do sistema de TMA baseado em BQFFB.

no desvio de fase ou domínio complexo.

- Parâmetros: é o conjunto das variáveis que controlam as etapas da detecção de *onsets*, englobando os parâmetros da seleção de picos λ, δ, M e T (definidos na Seção 3.2.1), além dos parâmetros relativos a cada método.

- Freqüência de amostragem: é necessária para se calcular o valor de outros parâmetros baseados em limiares psico-acústicos.

A informação sobre os instantes dos *onsets* é empregada tanto para se determinar sobre que parte da representação tempo-freqüência deve ser feita a detecção de F0s quanto para gerar o arquivo de saída em formato MIDI.

7.2.3 Módulo para detecção de F0s

Este módulo tem como entrada a representação tempo-freqüência e os instantes dos *onsets*. Seu produto é o conjunto de F0s identificadas para cada *onset* que, juntamente com o conjunto dos instantes de *onsets*, faz com que seja possível montar o arquivo de saída em formato MIDI.

O primeiro passo é determinar a janela de análise. Como foi visto na Seção 3.2, o instante em que existe maior estabilidade e previsibilidade do *pitch* da nota é durante a sustentação ou o decaimento. É nessa região que deve ser obtido o espectro sobre o qual se detectarão as F0s. A Figura 7.2 mostra um perfil de energia comumente encontrado para uma nota. O sinal exibido nesta figura tem início em um *onset* e termina no *onset* seguinte. Após o pico, tem-se a fase do decaimento. Resta definir, dentro dessa região, por meio de experimentos, onde estará a janela de observação dentro da qual se calculará o espectro que servirá de entrada para a detecção de F0s.

Para isso, divide-se o sinal em N_t intervalos de tempo e calcula-se a energia em cada um deles, formando-se o vetor V_t. O elemento de V_t que corresponde ao máximo de energia é denominado $V_t(t_p)$. Considerando a diferença d_{jan} entre a amplitude do pico e amplitude mínima após o pico, a janela deve ser definida começando na amostra de V_t (de índice maior do que t_p) correspondente a $I.d_{\text{jan}}$, denominada $V_t(t_i)$, e terminando naquela correspondente a $F.d_{\text{jan}}$, denominada $V_t(t_f)$, sendo $F < I$ (que

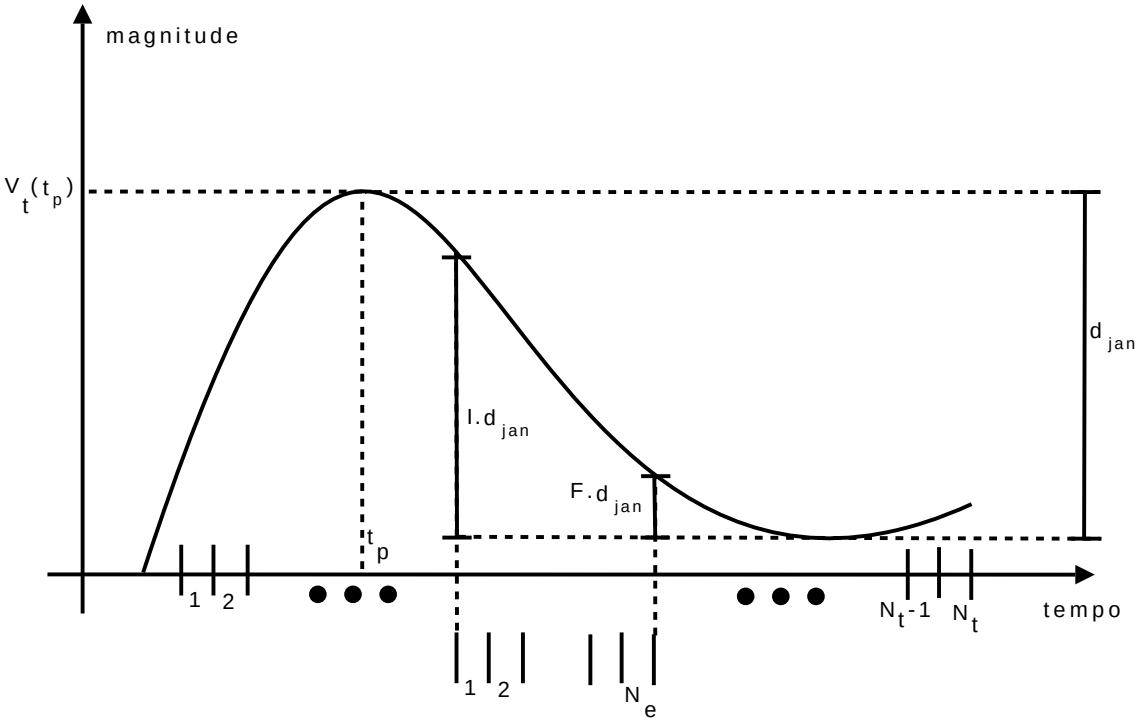


Figura 7.2: Perfil de uma nota comum sem sustentação para determinação da janela de análise.

terão seus valores definidos mais adiante). V_e é o vetor de $N_e < N_t$ elementos de V_t cujos índices estão entre t_i e t_f.

O espectro que será analisado é determinado, então, calculando-se a média dos N_e espectros correspondentes aos elementos de V_e. Esse espectro de análise será usado, juntamente com o eixo das freqüências correspondente, para determinar as F0s e, como produto secundário, a intensidade relativa a cada F0. Com referência ao algoritmo descrito na Seção 4.3, a intensidade é calculada como a energia do espectro suavizado. Como explicado na Seção 4.3.6, o espectro suavizado é calculado com base na F0 mais recentemente detectada e representa a influência dessa F0 no espectro sob análise. Portanto, a energia do espectro suavizado é proporcional à intensidade da nota correspondente à F0 em questão.

Até esse ponto, tem-se, para cada *onset*, um conjunto de F0s. Contudo, o fato de se detectar a presença de uma F0 f_A em um determinado *onset* $t(1)$ não significa necessariamente que a execução de f_A tenha tido início em $t(1)$. A execução de f_A em questão pode ter tido início em um instante $t(0)$ anterior a $t(1)$, mas tal nota ter continuado soando até $t(1)$. Portanto, detectar f_A em $t(1)$ seria um erro, pois,

na verdade, f_A deveria estar associada a $t(0)$.

Como é possível distinguir a situação que f_A é executada em $t(0)$ e soa até $t(1)$ daquela onde f_A é executada em $t(0)$ e novamente é executada em $t(1)$? Para desfazer essa ambigüidade, lança-se mão da seguinte heurística (ilustrada pela Figura 7.3):

1. Detectam-se os *onsets* (como na Figura 7.3(a)).

2. Detecta-se a F0 f_A no instante $t(n)$, armazenando sua intensidade $I(f_A, t(n))$ (como na Figura 7.3(b));

3. Se, ao se buscar por F0s no instante $t(n+1)$, for encontrada f_A, deve-se verificar a intensidade $I(f_A, t(n+1))$;

4. Se $I(f_A, t(n+1)) < I(f_A, t(n))$, significa que f_A começou a ser executada em $t(n)$ e sua intensidade foi caindo ao longo do tempo até $t(n+1)$, quando sua intensidade passou a ser $I(f_A, t(n+1))$ (como na Figura 7.3(c)). Nesse caso, a detecção de f_A em $t(n+1)$ foi um erro, e a associação de tal F0 ao instante $t(n+1)$ precisa ser descartada;

5. No entanto, se $I(f_A, t(n+1)) \geq I(f_A, t(n))$, houve um ganho de energia, o que indica que f_A foi novamente executada no instante $t(n+1)$, e a associação de tal F0 ao instante $t(n+1)$ não deve ser descartada (como nas Figuras 7.3(d-e)).

O algoritmo de remoção de notas repetidas pode falhar no caso de o músico executar notas rebatidas (constituídas por uma série de repetições de uma dada nota), sendo possível inclusive que as intensidades sejam de forma proposital progressivamente reduzidas.

7.3 Experimentos

Esta seção objetiva analisar os resultados de experimentos computacionais realizados com o sistema de transcrição musical com base em BQFFB, proposto nesta tese.

7.3.1 Metodologia

Para se verificar o desempenho do sistema, será utilizado um sinal sintético, no qual o número de notas executadas por vez varia de 1 a 4 e a estrutura rítmica é a mesma

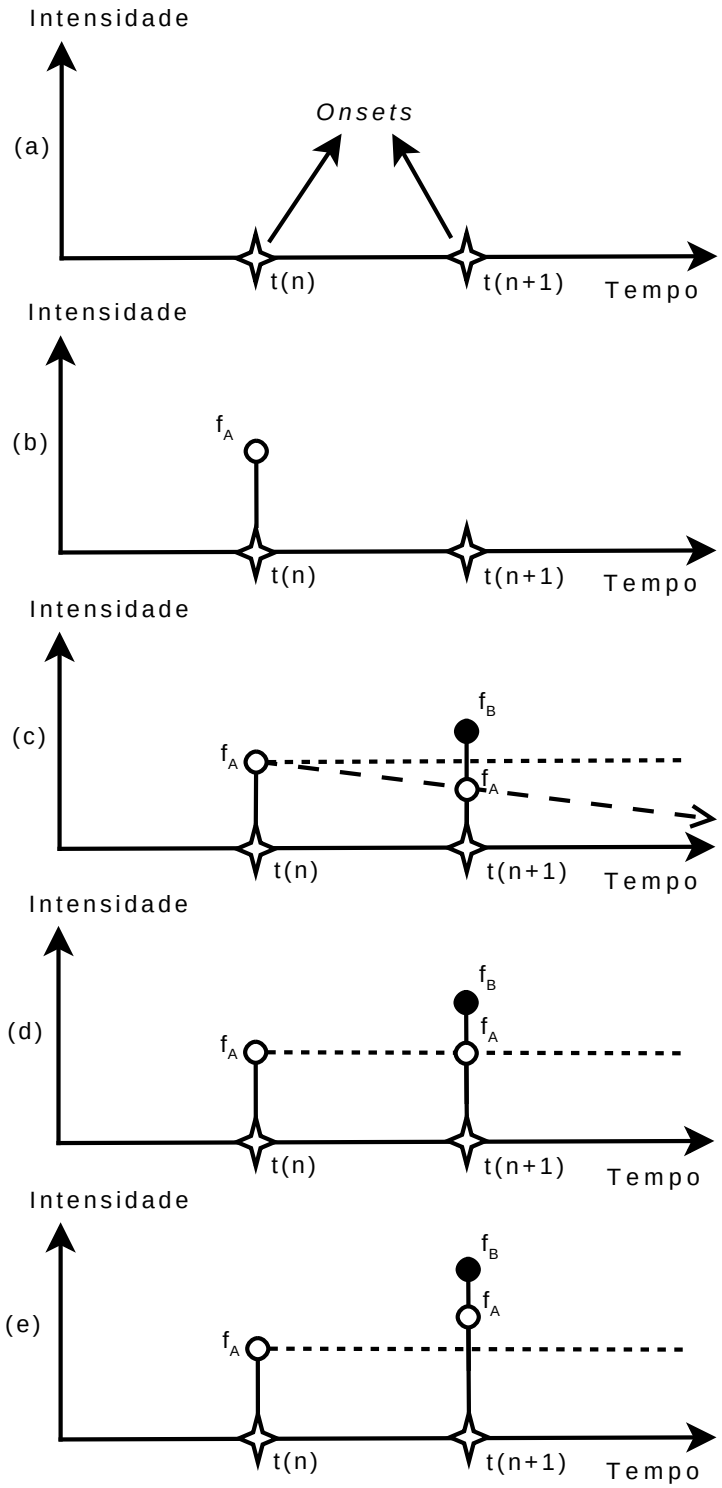


Figura 7.3: Esquema para a remoção de notas repetidas. (a) Detecção dos *onsets*; (b) Indetificação de f_A no instante $t(n)$; (c) Identificação de f_A e f_B no instante $t(n+1)$, sendo que f_A foi tocada apenas no instante $t(n)$ e soou até o instante $t(n+1)$; (d)-(e) Identificação de f_A e f_B no instante $t(n+1)$, sendo que f_A foi tocada novamente no instante $t(n+1)$

dos sinais sintéticos descritos no Capítulo 3. A partitura referente a este sinal com número variável de notas por vez pode ser visualizada na Figura 7.4. Além desse sinal, serão feitos testes utilizando-se o mesmo sinal real empregado no Capítulo 3 e outro sinal que contém os 20 primeiros segundos de uma gravação da Fuga número 2 do Livro 1 do Teclado Bem Temperado de Bach, BWV 847.

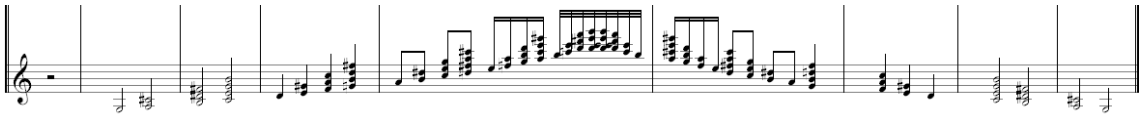


Figura 7.4: Partitura do sinal musical sintético que representa uma composição em que se executa de 1 a 4 notas por vez.

As duas figuras de mérito utilizadas como avaliação serão:

- *Note error rate*: abreviada como NER, representa a taxa de erro de notas e é dada por:

$$\text{NER} = 100\frac{FA + P + S}{N}, \tag{7.1}$$

  onde FA é o número de falsos alarmes (notas detectadas excedendo o número de notas de referência), P é o número de perdas (notas não detectadas faltando para completar o número de notas de referência), S é o número de substituições (notas que foram identificadas de forma errada) e N é o número de notas de referência. Essa figura de mérito não apresenta limite superior e, para um desempenho ideal, apresenta valor 0%.

- Acurácia: definida em [15] e [68], representa a razão entre notas corretamente identificadas e o número total de notas. É dada por:

$$A = 100\frac{H}{FA + P + S + H}, \tag{7.2}$$

  onde H é o número de notas corretamente identificadas. Essa figura de mérito, para um desempenho ideal, apresenta valor 100% e tem, portanto, uma faixa limitada, diferentemente da NER.

A fim de se obter uma análise mais ampla do desempenho do sistema, será uti-

lizada também uma figura do tipo *piano-roll*[1]. Assim, será possível comparar visualmente o *piano-roll* relativo à composição original com o *piano-roll* que representa o arquivo de saída em formato MIDI.

7.3.2 Ajuste de parâmetros

Esta seção objetiva tratar da configuração dos parâmetros livres que controlam o desempenho do sistema de transcrição musical com base em BQFFB.

Os parâmetros que controlam a detecção de *onset* e de F0s foram objeto dos testes realizados no Capítulo 6, portanto os valores lá determinados serão os considerados nos testes do presente capítulo. Novamente, a faixa de tolerância para detecção de *onsets* é de 50 ms e, para se ter uma identificação de F0s como correta, deve-se identificar corretamente a nota e a oitava.

Como foi explicado na Seção 7.2.3, dados um *onset* e o seu sucessor, é necessário estimar em que região se deve estabelecer a janela de análise, a qual é dependente de dois parâmetros: I e F, definidos na Seção 7.2.3 e ilustrados na Figura 7.2. Para isso, utilizaram-se os mesmos sinais de teste empregados no Capítulo 3: um em que há a execução de uma nota por vez e outro no qual há a execução de um acorde de três notas por vez. Suas partituras podem ser visualizadas nas Figuras 3.3 e 3.5. Considerando $N_t = 100$ (definido na Seção 7.2.3), avaliou-se a taxa de detecção do sistema, bem como sua acurácia e a NER (ver definição da Seção 7.3.1) para diversos valores de I e F, buscando-se aqueles que resultam em maximizar a taxa de detecção. Tais valores são iguais a 25% e 0, respectivamente. Tendo-se verificado os valores ótimos para I e F, deve-se buscar o valor de N_t que maximiza a taxa de detecção e a acurácia enquanto minimiza a NER. Tal valor foi de $N_t = 100$ intervalos.

[1] *Piano-roll* é um gráfico que apresenta o tempo no eixo das abscissas e as notas em forma de números no eixo das ordenadas. A referência para essa escala é a nota Dó da quarta oitava, que é representada pelo número 60. O nome *piano-roll* é uma alusão aos rolos de papel perfurado usados para automatizar a execução de músicas em pianolas.

7.3.3 Experimentos com sinais sintéticos

Nesta seção, serão feitos testes utilizando-se o sistema de TMA com base em FFT e o sistema de TMA com base em BQFFB. O objetivo é mostrar que os sistemas apresentam desempenhos comparáveis.

O primeiro teste será analisar o sinal sintético em que se executa de 1 a 4 notas por vez, que é descrito na Seção 7.3.1 e ilustrado pela Figura 7.4. O sistema com base em FFT gerou, então, o gráfico do tipo *piano-roll* exibido na Figura 7.5. Esta figura mostra três gráficos: o primeiro representa as notas realmente executadas e contidas no sinal original; o segundo representa as notas detectadas e identificadas pelo sistema com base em FFT; o terceiro representa as notas detectadas e identificadas pelo sistema com base em BQFFB.

Comparando-se as Figuras 7.5(a) e 7.5(b), verifica-se que as notas de maior duração são detectadas e identificadas com maior precisão enquanto que as notas mais rápidas são, muitas vezes, perdidas ou substituídas. Além disso, a probabilidade de falsos alarmes também é inversamente proporcional à duração das notas.

Do mesmo modo, este sinal sintético também foi submetido ao sistema com base em BQFFB e foi gerado o *piano-roll* mostrado na Figura 7.5(c). Pode-se ver que foi possível, com o BQFFB, modelar de forma mais precisa a figura melódica representada no sinal de entrada, principalmente, em relação às notas mais rápidas. Isso se deve em grande parte a uma melhor detecção de *onsets*. É também possível verificar que as notas detectadas a partir do BQFFB, na Figura 7.5(c), estão mais próximas das notas de referência, na Figura 7.5(a).

Como uma forma de avaliar a qualidade da transcrição evidenciando cada etapa e comparar o desempenho da FFT com o do BQFFB, pode-se resumir as características desta experiência na Tabela 7.1.

Para se analisar a Tabela 7.1, deve-se lembrar que, diferentemente das experiências realizadas nos Capítulos 3, 4 e 6, as características são calculadas levando-se em consideração o funcionamento global do sistema. Assim, um *onset* que tenha sido erroneamente detectado constituirá um falso alarme de *onset*. As notas que tiverem sido detectadas para esse *onset* serão, por definição, consideradas como falsos alarmes de F0s. Além disso, uma F0 que tenha sido erroneamente detectada, mas eliminada pelo procedimento de remoção de notas repetidas (ilustrado na

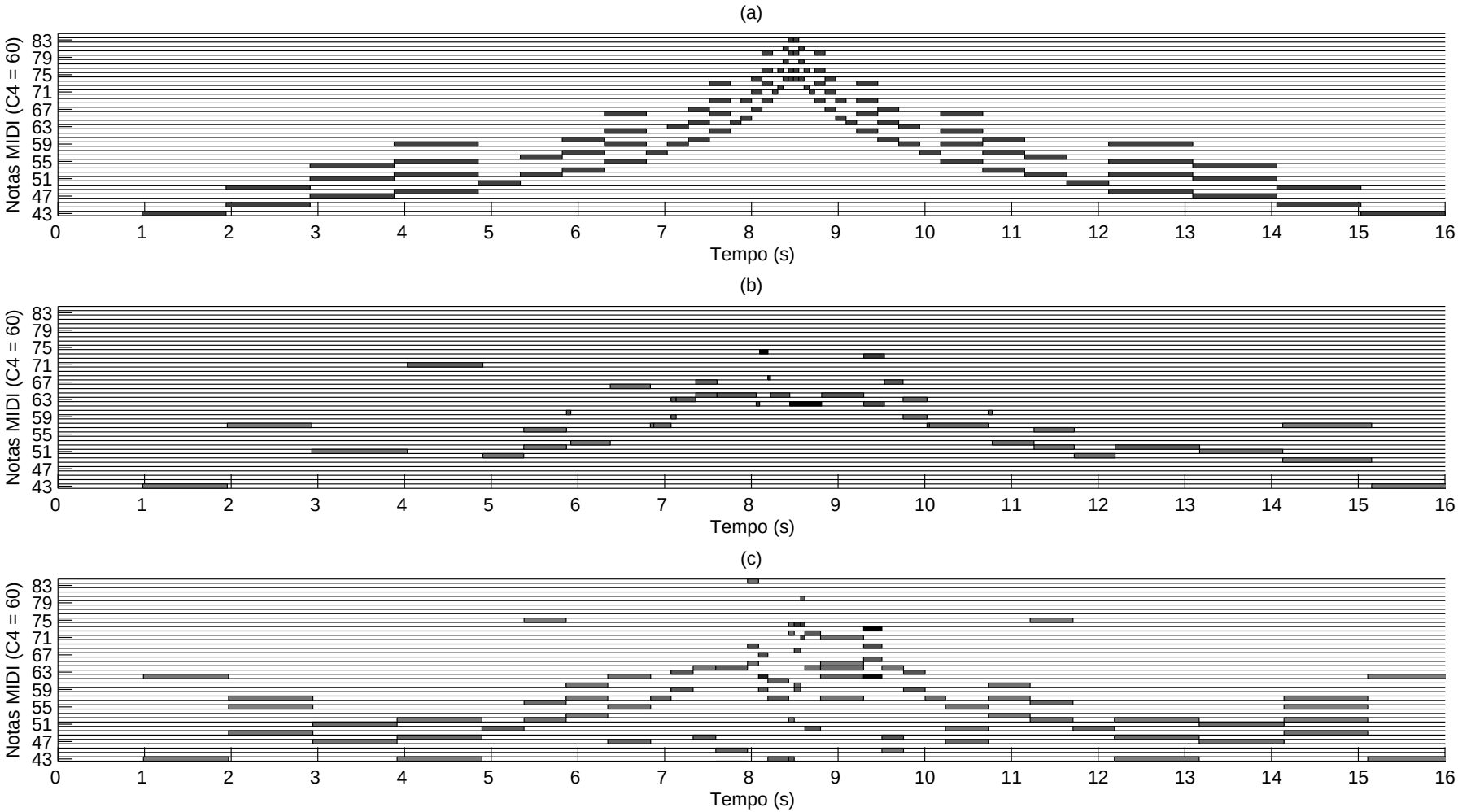


Figura 7.5: Comparação entre (a) o *piano-roll* referente ao sinal de entrada sintético que apresenta de 1 a 4 notas por vez, (b) o *piano-roll* que representa o arquivo MIDI gerado pelo sistema com base em FFT e (c) o *piano-roll* que representa o arquivo MIDI gerado pelo sistema com base em BQFFB.

Tabela 7.1: Comparação do desempenho de sistema de TMA com base em FFT com sistema com base em BQFFB para sinal sintético contendo de 1 a 4 notas por vez.

Característica	FFT	BQFFB
Detecção de *onsets*	72,5%	80,0%
Taxa de falso alarme para *onsets*	36,4%	3,1%
Detecção de F0s	20,0%	44,0%
Taxa de falso alarme para F0s	34,1%	14,8%
Taxa de perda para F0s	59,0%	21,0%
Taxa de substituição para F0s	7,0%	31,0%
NER	80,0%	65,0%
Acurácia	20,0%	40,4%

Figura 7.3), não será considerada para cômputo da taxa de falso alarme.

O sistema de TMA com base no BQFFB apresentou maiores taxas de detecção de F0s e *onsets*, reduzindo bastante as taxas de falso alarme (aumentando, porém, a taxa de substituição). Para se comparar o desempenho do sistema com base em FFT com o do sistema com base em BQFFB de forma global, recorre-se às figuras de mérito descritas na Seção 7.3.1. Isso pode ser resumido nas duas últimas linhas da Tabela 7.1, onde pode-se ver que a NER relativa ao sistema com base em BQFFB representa uma redução de quase 20% em relação à do sistema com base em FFT. Do mesmo modo, o sistema com base em BQFFB apresenta uma acurácia duas vezes maior. Daí, pode-se deduzir que o sistema baseado em BQFFB apresentou um melhor desempenho.

7.3.4 Experimentos com sinais reais

Os testes com sinal real objetivam comprovar a eficácia do sistema proposto em um ambiente com as dificuldades que um sistema comercial deve tratar, como reverberações, ecos e ruídos. O primeiro experimento a ser realizado tem como sinal de entrada os primeiros 25 segundos da Ária das Variações de Goldberg executada pelo pianista Glenn Gould (gravação de estúdio de 1955). Os *piano-rolls* correspondentes (obtidos por um músico extremamente bem treinado) são mostrados na Figura 7.6, e os resultados da transcrição são apresentados na Tabela 7.2. Como os

sinais sintéticos que foram usados para ajustar os parâmetros da detecção de *onsets* não representavam muito bem este sinal (como foi comprovado pelos resultados do Capítulo 3), optou-se por realizar um teste que visava a mostrar o melhor desempenho possível para este sinal. Isso foi feito ajustando-se os parâmetros da detecção de *onsets* pelo próprio sinal de teste.

A partir da Figura 7.6, pode-se dizer que nenhuma das duas ferramentas de análise espectral foi capaz de promover um bom desempenho para o sistema de TMA. Isso se dá principalmente devido à grande quantidade de notas executadas em alta velocidade. Esse fato é ainda confirmado pelo bom desempenho na detecção das notas mais lentas, identificadas nos *piano-rolls* por linhas horizontais mais longas.

Tabela 7.2: Comparação do desempenho de sistema de TMA com base em FFT com sistema com base em BQFFB para o sinal real contendo a Ária das Variações Goldberg para teclado de Bach, BWV 988, executada ao piano.

Característica	FFT	BQFFB
Detecção de *onsets*	83,13%	73,5%
Taxa de falso alarme para *onsets*	9,45%	21,0%
Detecção de F0s	30,7%	36,0%
Taxa de falso alarme para F0s	19,0%	50,0%
Taxa de perda para F0s	55,3%	45,6%
Taxa de substituição para F0s	28,9%	36,8%
NER	98,2%	155,3%
Acurácia	23,8%	18,8%

Para se comparar o desempenho do sistema com base em FFT com o do sistema com base em BQFFB de forma global, recorre-se às figuras de mérito descritas na Seção 7.3.1. Isso pode ser resumido nas duas últimas linhas da Tabela 7.2. Analisando-se o resultado dos testes com sinal real, pode-se ver que, apesar de ter havido uma redução na taxa de detecção de *onsets* para o BQFFB, houve um aumento na taxa de detecção de F0s. Isso fez com que a acurácia para o BQFFB caísse em cerca de 14%.

Esse sinal levou o sistema a condições extremas, pois não apresenta notas tão bem definidas em certos pontos devido à alta velocidade de execução. Para ilustrar

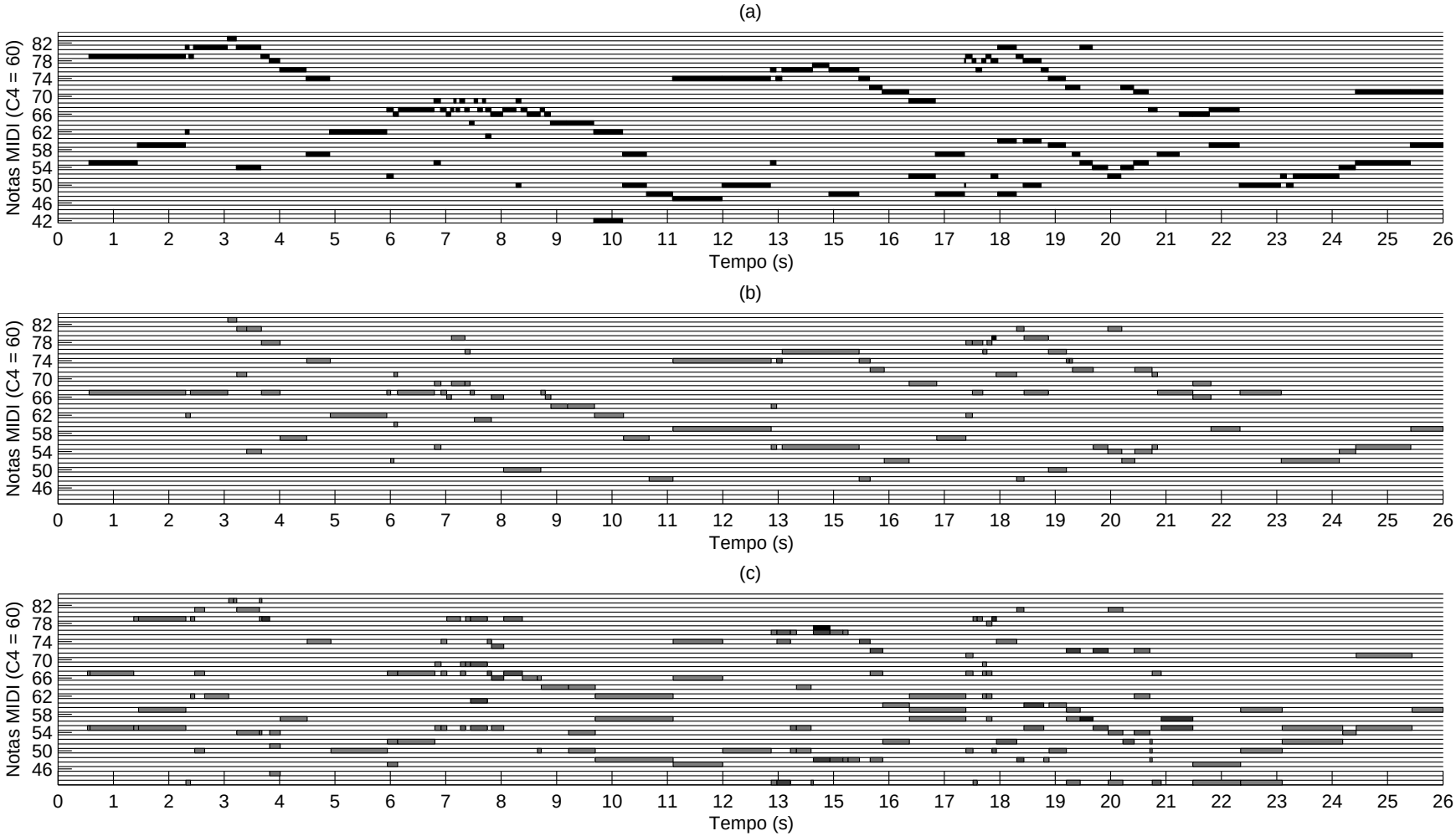


Figura 7.6: Comparação entre (a) o *piano-roll* referente ao sinal de entrada que contém a Ária das Variações de Goldberg, (b) o *piano-roll* que representa o arquivo MIDI gerado pelo sistema com base em FFT e (c) o *piano-roll* que representa o arquivo MIDI gerado pelo sistema com base em B*Q*FFB.

o comportamento do sistema frente a um sinal com características mais bem definidas em relação ao ataque das notas, utilizou-se como segundo experimento a Fuga número 2 do Livro 1 do Teclado Bem Temperado de Bach, BWV 847, executada pelo pianista Glenn Gould (gravação de estúdio de 1962). Para este experimento, foram usados os valores dos ajustes feitos para o caso sintético.

Como uma forma de avaliar a qualidade da transcrição evidenciando cada etapa e comparar o desempenho da FFT com o do BQFFB, pode-se recorrer aos *piano-rolls* mostrados na Figura 7.7 e às características da experiência resumidas na Tabela 7.3.

A partir da Figura 7.7, já é possível ver que os desenhos melódicos e harmônicos fornecidos pelo sistema baseado em BQFFB, na Figura 7.7(c), foram modelados de forma mais próxima do *piano-roll* de referência (obtidos através de uma partitura e verificados por um músico bem treinado), na Figura 7.7(a), do que aqueles gerados pelo sistema baseado em FFT, na Figura 7.7(b).

Executando as análises de forma relativa, referindo-se primeiramente a *onsets*, o sistema com base em BQFFB apresentou taxa de detecção cerca de 10% mais alta, com falso alarme reduzido praticamente à metade. Devido ao melhor desempenho quanto à identificação de *onsets* e à maior seletividade, o sistema baseado em BQFFB conseguiu um aumento de cerca de 20% na taxa de detecção de F0s em relação à taxa obtida pela FFT, reduzindo a perda em cerca de 50%.

Tabela 7.3: Comparação do desempenho de sistema de TMA com base em FFT com sistema com base em BQFFB para o sinal real contendo a Fuga número 2 de Bach, BWV 847, executada ao piano.

Característica	FFT	BQFFB
Detecção de *onsets*	68,9%	76,7%
Taxa de falso alarme para *onsets*	19,5%	10,4%
Detecção de F0s	42,3%	50,4%
Taxa de falso alarme para F0s	14,8%	28,2%
Taxa de perda para F0s	41,6%	23,4%
Taxa de substituição para F0s	16,1%	27,6%
NER	68,6%	81,8%
Acurácia	38,2%	38,2%

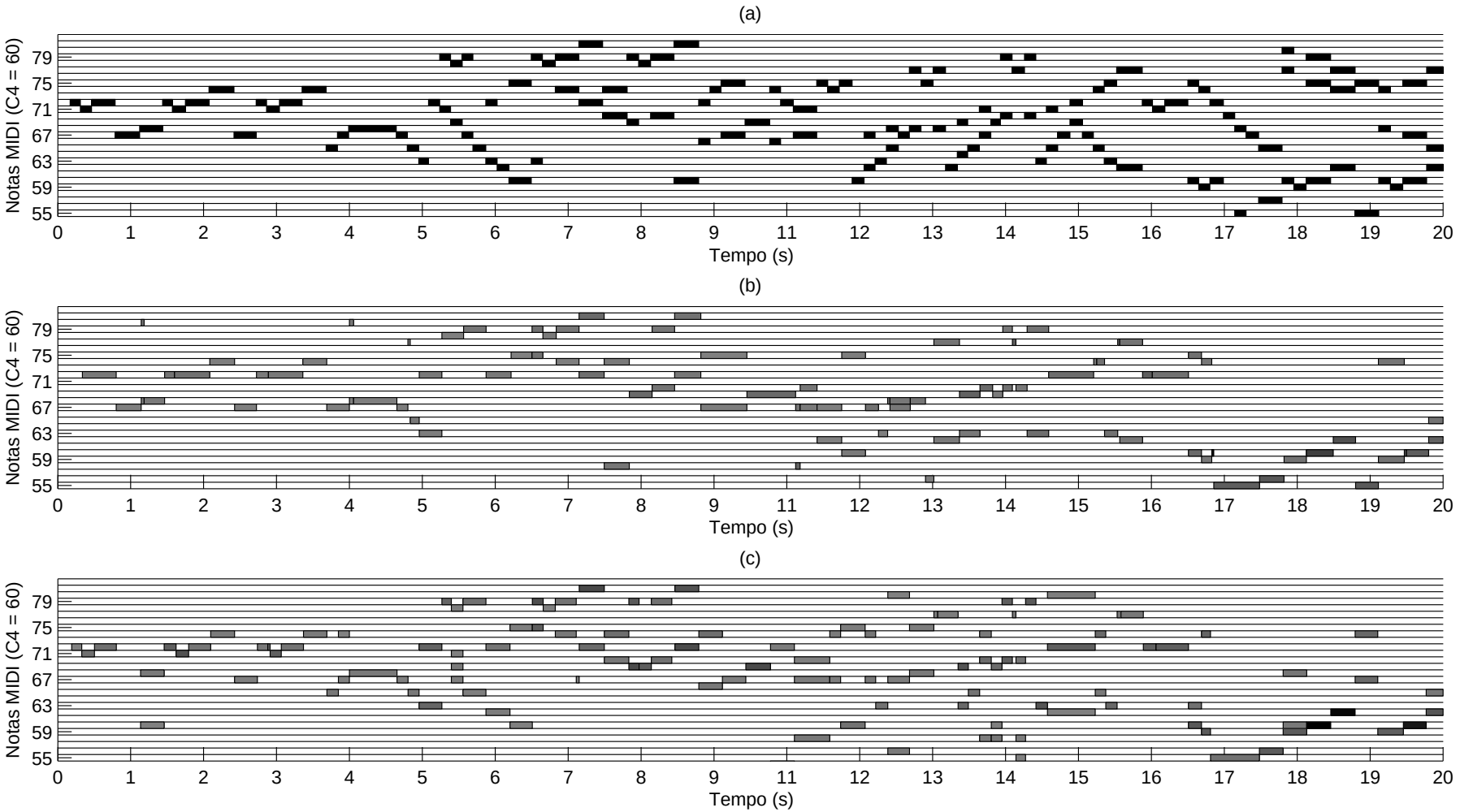


Figura 7.7: Comparação entre (a) o *piano-roll* referente ao sinal de entrada que contém a Fuga número 2 do Livro 1 do Teclado Bem Temperado de Bach, BWV 847, (b) o *piano-roll* que representa o arquivo MIDI gerado pelo sistema com base em FFT e (c) o *piano-roll* que representa o arquivo MIDI gerado pelo sistema com base em BQFFB.

Para se comparar o desempenho do sistema com base em FFT com o do sistema com base em BQFFB de forma global, recorre-se às figuras de mérito descritas na Seção 7.3.1. Isso pode ser resumido nas duas últimas linhas da Tabela 7.3. Ao se migrar do sistema baseado em FFT para aquele com base em BQFFB, a NER subiu em torno de 14% enquanto que a acurácia permaneceu a mesma.

7.4 Comparação de desempenho

Para se ter uma referência externa do desempenho do sistema proposto, pode-se compará-lo com o desempenho de um sistema comercial. Para isso, foi selecionado o Melodyne®, um dos *software* mais reconhecidos na área. Ao se utilizar este programa, ilustrado pela Figura 7.8, foram obtidos os resultados apresentados na Tabela 7.4. É importante lembrar que todos os parâmetros de detecção passíveis de ajuste no Melodyne foram configurados com valores automaticamente definidos.

Como essa versão do Melodyne trabalha apenas com sinais monofônicos, o único sinal de teste que é passível de comparação com o sistema proposto é o que apresenta uma nota por vez, ilustrado e descrito na Seção 3.4.1.

Para que a comparação seja justa, estabeleceu-se que o máximo número de notas que o sistema proposto poderia estimar para cada *onset* identificado é 1. Como o Melodyne busca invariavelmente uma nota por *onset*, essa restrição deve ser também imposta ao sistema proposto para que as taxas de falso alarme não sejam alteradas devido a uma má estimativa do número de notas. O resultado deste teste utilizando-se o sistema proposto pode ser visto na Figura 7.9, onde se pode ver a similaridade entre a linha melódica de referência, exibida na Figura 7.9(a), a provida pelo BQFFB, na Figura 7.9(b), e a do Melodyne, representada na Figura 7.8.

A sensibilidade do Melodyne pode ser aumentada de modo a aproximar as taxas de detecção para *onsets* do sistema com base em BQFFB do Melodyne e, dessa forma, ter uma comparação justa. Fazendo-se isso, tem-se os resultados mostrados na quinta coluna da Tabela 7.4.

Analisando-se esta tabela, percebe-se que o sistema proposto apresenta maior taxa de detecção de *onsets*, mantendo, mesmo em relação ao Melodyne com sensibilidade aumentada, uma taxa mais baixa de falso alarme. A taxa de detecção de

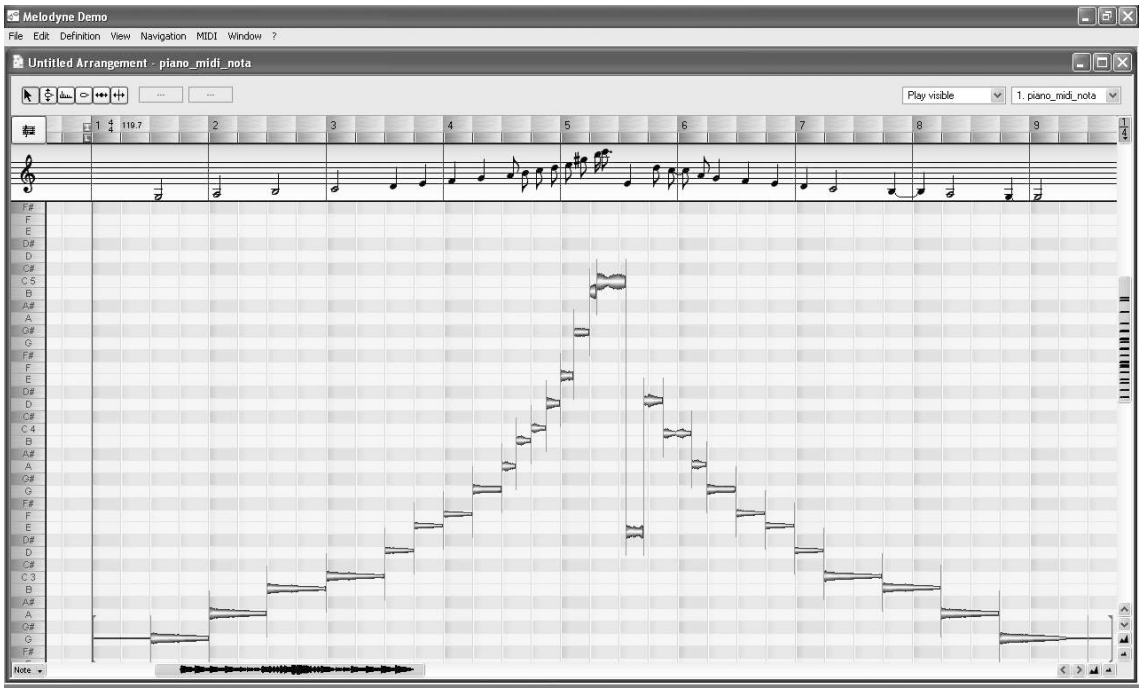


Figura 7.8: Interface do Melodyne ao se processar o sinal onde se executa uma nota por vez.

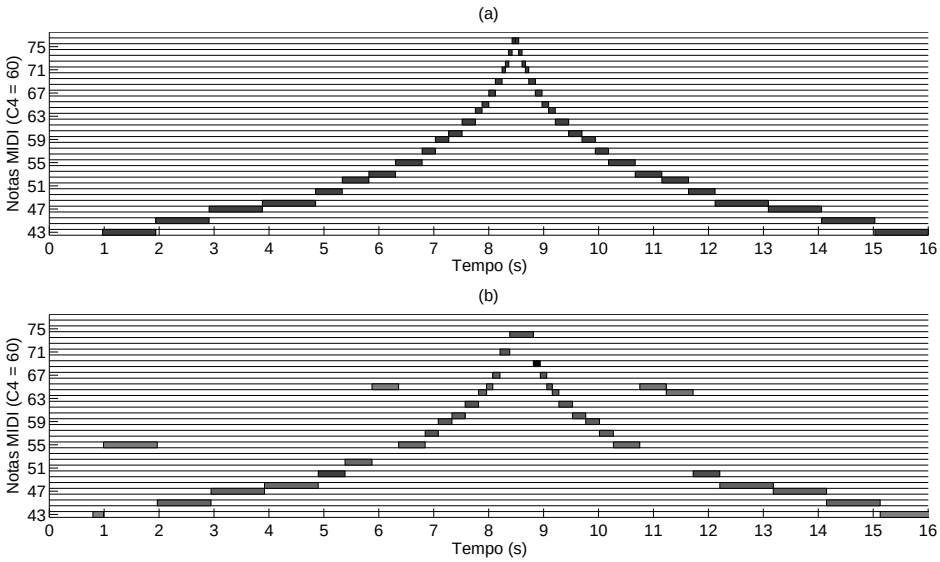


Figura 7.9: Resultado do sistema com base em BQFFB quando se analisa um sinal de entrada em que se executa uma nota por vez com a restrição de se identificar apenas uma nota para cada *onset* detectado. (a) Notas de referência (b) Notas identificadas pelo sistema.

Tabela 7.4: Comparação por etapa do desempenho do Melodyne (configuração padrão e configuração com sensibilidade aumentada) com sistema baseado em BQFFB para sinal sintético onde se executa uma nota por vez.

Sistema	BQFFB	Melodyne	Melodyne
Característica	-	*default*	sintonizado
Detecção de *onsets*	85,0%	75,0%	82,5%
Falso alarme para *onsets*	2,9%	6,7%	6,1%
Detecção de F0s	67,5%	70,0%	80,0%
Falso alarme para F0s	2,9%	6,7%	6,1%
Perda para F0s	17,5%	25,0%	17,5%
Substituição para F0s	15,0%	5,0%	2,5%
NER	35,0%	35,0%	25,0%
Acurácia	65,8%	67,0%	76,0%

F0s para o sistema com base em BQFFB é similar à do Melodyne nas condições padrão, mas é menor se comparada à configuração com sensibilidade aumentada. O que mais diferencia o Melodyne dos demais concorrentes é a taxa de erros de substituição.

Para se comparar o desempenho do Melodyne com o do sistema proposto de forma integral, recorre-se às figuras de mérito descritas na Seção 7.3.1. Isso pode ser resumido nas duas últimas linhas da Tabela 7.4. O desempenho geral do sistema proposto é bastante similar ao do Melodyne em condições padrão.

Como visto anteriormente, a vantagem do Melodyne com sensibilidade aumentada está em sua reduzida taxa de substituição. Apesar disso, o sistema proposto, em sua primeira versão, com ajustes padrão, apresenta uma acurácia superior a 85% da acurácia obtida pelo Melodyne com sensibilidade aumentada. Assim, pode-se afirmar que o sistema de TMA com base em BQFFB está próximo ao satisfatório dentro do que o mercado oferece comercialmente como tecnologia de ponta. Maiores avanços poderão ser feitos em versões futuras do sistema.

7.5 Conclusão

Este capítulo apresentou o sistema de TMA com base no BQFFB. Foram analisados os resultados obtidos com sinais sintéticos e reais. Em ambos os casos, em relação ao sistema com base em FFT, o sistema proposto foi capaz de obter um desempenho similar. O sistema com base em FFT apresenta acurácia média de 27% e NER média de 82%; enquanto que o sistema com base em BQFFB exibe acurácia média de 32% e NER média de 100%.

Para se ter uma referência externa, comparou-se o sistema proposto com um dos programas para análise espectral, transcrição e edição de sinais musicais mais utilizados mais utilizados no mercado de áudio profissional: o Melodyne. A configuração padrão do Melodyne apresentou um desempenho bastante similar ao do sistema proposto.

Embora se tenham obtido resultados relativamente bons, comprovando a real eficácia do BQFFB em aplicações de TMA, ainda existe espaço para melhorias, que podem ser obtidas tanto utilizando-se outros sistemas para implementar os módulos de detecção de F0s e *onsets* quanto no pós-processamento das informações obtidas por esses módulos. Um exemplo de pós-processamento é a predição do tom (ver Apêndice A) em que se encontra a composição para que não se identifiquem notas que não fazem sentido harmônico.

Capítulo 8

Conclusão

Este trabalho mostrou a implementação de um sistema de transcrição musical automática, ou TMA, com base no *bounded-Q fast filter bank* (BQFFB). A vantagem de se empregar o BQFFB é o fato de consistir em uma ferramenta de análise espectral que apresenta distribuição quase-geométrica de canais de alta seletividade, o que é bastante eficiente quando se analisam sinais musicais. Além disso, o BQFFB apresenta custo computacional moderadamente maior que o custo associado à FFT.

O Capítulo 2 apresentou as principais etapas para se implementar um típico sistema de TMA: a detecção de *onsets* e a identificação das freqüências fundamentais, ou F0s, presentes no sinal de entrada. O Capítulo 3 descreveu métodos convencionais para a detecção de *onsets* que atuam no domínio da freqüência e que têm como fundamento a análise espectral realizada a partir da FFT. No Capítulo 4, foi detalhado um sistema, com base em FFT, para identificação de F0s. Apesar de bons resultados terem sido alcançados, havia margem para melhorias, principalmente referentes à análise espectral.

Então, no Capítulo 5, foi apresentado o BQFFB com seus detalhes e ajustes de implementação, além de experimentos para avaliar seu comportamento. Contudo, para validar o BQFFB como uma ferramenta de análise espectral adequada para TMA, no Capítulo 6, foram realizados testes em cada uma das etapas de TMA. Uma vez estabelecido o BQFFB como uma base fundamentada para TMA, o Capítulo 7 detalhou a implementação de um sistema de TMA cujo alicerce é o BQFFB. Foram feitos testes para se avaliar seu desempenho e comparações, tanto em relação ao mesmo sistema com base em FFT como em relação a sistemas comerciais de

larga utilização no mercado de áudio profissional. Em todos os cenários, o sistema proposto se mostrou competitivo.

O presente capítulo visa a concluir o trabalho descrito nos demais capítulos. A Seção 8.1 enumera as contribuições da pesquisa envolvida nesse trabalho para o campo da TMA enquanto que a Seção 8.2 fornece uma análise sobre este trabalho e a Seção 8.3 fornece temas para trabalhos futuros para essa linha de pesquisa fundamentados no atual trabalho.

8.1 Contribuições

O presente trabalho de pesquisa apresentou as seguintes contribuições para o campo da TMA:

- Foi descrito o BQFFB, implementado como a cascata de um banco CQFFB separador de oitavas, onde cada oitava é, então, analisada por um FFB. Seu desempenho foi avaliado na Seção 5.7 e em [P2] e sua complexidade computacional foi analisada na Sub-seção 5.6.3.

- Foram detalhados os ajustes na implementação do BQFFB na Sub-seção 5.6.2, principalmente em relação à separação de oitavas e como deve ser feita a compensação dos canais que se localizam nas fronteiras das oitavas.

- Foi proposto um banco de filtros de seletividade híbrida para realizar a separação das oitavas na Sub-seção 5.6.2, de modo que as oitavas relativas às freqüências mais baixas fossem separadas por um filtro de menor seletividade para que o transitório ficasse melhor caracterizado.

- Foi proposto o emprego do BQFFB no método de distância espectral para detecção de *onsets* na Seção 6.2. Isso promoveu um desempenho superior para o método de detecção de *onsets* para os sinais sintéticos (em relação ao mesmo método com base em FFT).

- Foi proposto o emprego do BQFFB como base de um sistema iterativo para identificação de F0s na Seção 6.3. Isso fez com que o desempenho do sistema de detecção de F0s apresentasse um desempenho similar ao mesmo sistema com

base em FFT, o que foi comprovado pelos testes realizados sobre as gravações da base MUMS.

- Foi proposto um sistema de TMA com base em BQFFB na Seção 7.2. Os testes realizados na Seção 7.3 com sinais sintéticos mostraram vantagens ao se utilizar o BQFFB, com desempenho comparável ao de sistemas que usam FFT e a de sistemas comerciais como o Melodyne 8.0®, utilizado para fins de análise espectral e transcrição em áudio profissional (o que foi feito na Seção 7.4).

8.2 Análise crítica

A transcrição musical automática é um problema sem solução fechada. Existem muitos trabalhos sobre cada etapa do processo de TMA, mas poucos tratam o tema de forma integral. Além disso, devido à grande subjetividade e diversidade que sempre caracterizaram o universo musical, é extremamente difícil se chegar a uma solução geral que inclua todos os timbres e dinâmicas que podem estar presentes em uma gravação ou execução de uma composição.

Pode-se afirmar que o presente trabalho obteve resultados significativos no que se refere à representação espectral de sinais musicais. Demonstrou eficácia na detecção de até duas notas simultâneas e na detecção de *onsets*. Existem, no entanto, claras oportunidades de melhoria na forma de se aplicar a ferramenta desenvolvida em TMA, principalmente em relação à detecção e identificação de F0s. Para isso, há duas frentes de trabalho: alterar alguma etapa do método utilizado neste trabalho ou investigar outros métodos para detecção de F0s. O resultado do sistema de TMA como um todo também poderia ser melhorado testando-se outros métodos de processamento da informação sobre F0s e *onsets*.

Assim como um transcritor humano, antes de começar o processo de transcrição propriamente dito, o sistema deve estar familiarizado com o estilo musical ao qual a composição pertence e conhecer os timbres dos instrumentos envolvidos. Daí pode-se afirmar que a TMA deveria ser uma das últimas etapas de MIR, porque as informações sobre estilo e timbre são de grande valia para se determinar F0s e *onsets*. Ter essas informações *a priori* significaria um melhor ajuste de uma série de parâmetros que controlam o desempenho dos métodos utilizados.

Além disso, em comparação com soluções comerciais, os métodos de detecção e identificação de *onsets* e F0s apresentam um tempo de execução elevado. Isso na verdade é devido à alta complexidade dos métodos que atuam sobre a representação espectral. No entanto, visto que, com o BQFFB, há uma distribuição geométrica de canais, há indícios que qualquer método de TMA que utilize os canais resultantes da representação espectral tenderá a ter a execução mais rápida se atuar sobre o BQFFB (em comparação com a DFT). Em suma, este trabalho serviu para demonstrar que o BQFFB é uma ferramenta de análise espectral que tem um grande valor para tratar sinais musicais para TMA e o desempenho dos métodos que atuaram sobre sua representação é comparável com o método de referência, que é a DFT.

8.3 Trabalhos futuros

O ponto atual do desenvolvimento da pesquisa envolvida no presente trabalho dá margem aos seguintes temas para trabalhos futuros:

- Testar outros métodos de remoção do espectro relativo à F0 mais recentemente detectada, ao se utilizar o algoritmo de Klapuri para detecção de múltiplas F0s.

- Testar o desempenho do sistema de TMA para freqüências de amostragem reduzidas. Isso seria feito para tentar reduzir o tempo de execução do sistema sem piora da qualidade.

- Atualmente, o sistema de TMA proposto funciona com base em um sinal de entrada que é, em toda sua extensão, submetido ao BQFFB para análise e posterior processamento. Para que se obtenha maior flexibilidade em relação à extensão do sinal contido no arquivo de entrada, seria necessário que o sistema fosse migrado para uma estrutura em *buffers* em que cada *buffer* seria processado por todo o sistema de forma serial. Isso, futuramente, serviria de base para um sistema em tempo real.

- Para melhorar o desempenho do sistema de TMA proposto, pode-se executar, na forma de um pré-processamento, a remoção de ruído através de métodos de restauração de sinais de áudio. Isso pode ser feito através de técnicas como filtragem de Wiener, *wavelet shrinkage* ou filtragem adaptativa [69]. Com

isso, pode-se conseguir a redução de interferências, cliques (falsos alarmes para detecção de *onsets*) e de freqüências espúrias (falsos alarmes para identificação de F0s).

- Para se detectar acordes de forma mais eficiente, pode-se manter um banco de dados com modelos de acordes. Cada modelo indicaria que canais do BQFFB estariam ativos para determinado acorde. Assim, ao se identificar a F0 predominante, tentar-se-ia encaixar em um dos modelos de acorde que contivessem essa nota. Aquele que apresentasse maior correlação com o sinal sob análise seria identificado como o acorde executado.

- Outra proposta seria detectar algumas F0s predominantes, usando-se o sistema baseado em BQFFB, e assumir que a F0 de mais baixa freqüência é a tônica do acorde (ver Apêndice A). Dessa forma, como um pós-processamento, os modelos a serem testados, segundo a técnica proposta no item anterior, seriam aqueles modelos conhecidos de acordes referentes a essa tônica. Assim haveria um universo mais reduzido de acordes entre os quais seria buscado o ótimo.

- Ainda para fins de detecção de notas presentes, poderia ser empregado algum sistema baseado em redes neurais que operasse sobre os canais do BQFFB. Este tema está sendo explorado por outros membros da equipe de pesquisa no qual o presente trabalho se insere [70].

- Para ser possível utilizar o sistema de TMA proposto no Capítulo 7 em sinais com mais de um instrumento, pode-se, através algum tipo de técnica de análise-por-síntese, separar os sinais relativos a cada instrumento e processar posteriormente para transcrição. Para isso, deveria existir um dicionário contendo uma série de sinais de referência para cada instrumento que pudesse existir no sinal em questão. Assim, através de correlação, calcular-se-ia que trechos contidos no dicionário descrevessem de forma mais acurada a parcela do sinal de entrada relativa a determinado instrumento.

- Também para fins de separação de sinais para posterior transcrição automática de cada instrumento, poderia ser empregado o método de separação de fontes sonoras via deconvolução 2D por fatoração em matrizes não negativas usando

espectro linearmente espaçado. Este tema está sendo explorado por outros membros da equipe de pesquisa no qual o presente trabalho se insere [71].

- Como a maior dificuldade dessa área de pesquisa é a falta de uma base de dados consistente para testes e ajustes de parâmetros, seria interessante que houvesse um esforço neste sentido. Deveria ser feita a seleção das gravações mais propícias e classificá-las quanto aos instrumentos presentes, ao nível de polifonia, ao estilo e à velocidade de execução das notas. A seguir, cada gravação deveria ser transcrita de forma manual. Assim, o sinal de cada gravação seria acompanhado de duas seqüências de informação para cada instrumento presente: *onsets* e notas executadas em cada *onset*. Desse modo, seria implementada uma interface de testes em que o usuário teria a possibilidade de selecionar o banco completo ou apenas uma parte segundo os critérios descritos acima.

Referências Bibliográficas

[1] BROSSIER, P. M., SANDLER, M. B., PLUMBLEY, M. D., "Real Time Object Based Coding". In: *114th Audio Engineering Society Convention*, pp. 1–6, Amsterdam, Holanda, Maio 2003.

[2] VINCENT, E., PLUMBLEY, M. D., "Low Bit-Rate Object Coding of Musical Audio Using Bayesian Harmonic Models", *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, n. 4, pp. 1273–1282, Abril 2007.

[3] VAIDYANATHAN, P. P., *Multirate Systems and Filter Banks*. Prentice-Hall: Upper Saddle River, NJ, EUA, 1992.

[4] 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, Abril 1992.

[5] KASHIMA, K. L., MONT-REYNAUD, B., *The Bounded-Q Approach to Time-Varying Spectral Analysis*, Tech. Rep. STAN-M-28, Stanford University, Dep. of Music, Stanford, CA, EUA, 1985.

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

[7] KLAPURI, A., *Signal Processing Methods for the Automatic Transcription of Music*, Tese de doutorado, Tampere University of Technology, Institute of Signal Processing, Finlândia, Abril 2004.

[8] MARTIN, K. D., *Automatic Transcription of Simple Polyphonic Music: Robust Front-End Processing. Technical Report 399*. MIT Media Library Perceptual Computing Section, 1996.

[9] GODSMARK, D., BROWN, G. J., "A Blackboard Architecture for Computational Auditory Scene Analysis", *Speech Communication*, v. 27, n. 3-4, pp. 351–366, Abril 1999.

[10] ELLIS, D. P. W., *Prediction-Driven Computational Auditory Scene Analysis*, Tese de doutorado, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, EUA, 1996.

[11] "http://www.intelliscore.net", Site do programa intelliScore Ensemble, Acessado em 21/12/2007, Fabricante: Innovative Music Systems.

[12] "http://www.neuratron.com/audioscore.htm", Site do programa Neuratron AudioScore Professional, Acessado em 10/10/2009, Fabricante: Neuratron.

[13] "http://www.celemony.com", Site do Programa Melodyne, Acessado em 10/10/2009, Fabricante: Celemony.

[14] BELLO, J. P., DAUDET, L., SANDLER, M. B., "Automatic Piano Transcription Using Frequency and Time-Domain Information", *IEEE Transactions on Audio, Speech and Language Processing*, v. 14, n. 6, pp. 2242–2251, Novembro 2006.

[15] DIXON, S., "On the Computer Recognition of Solo Piano Music". In: *Australian Computer Music Conference*, Queensland, Austrália, 2000.

[16] BELLO, J. P., DAUDET, L., ABDALLAH, S., et al., "A Tutorial on Onset Detection in Music Signals", *IEEE Transactions on Speech and Audio Processing*, v. 13, n. 5, pp. 1035–1047, Setembro 2005.

[17] COLLINS, N., "A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions". In: *118th Audio Engineering Society Convention*, pp. 1–12, Barcelona, Espanha, Maio 2005.

[18] RYYNÄNEN, M. P., KLAPURI, A., "Polyphonic Music Transcription using Note Event Modeling". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, EUA, Outubro 2005.

[19] DIXON, S., "Learning to Detect Onsets of Acoustic Piano Tones". In: *MOSART Workshop on Current Research Directions in Computer Music*, pp. 1–5, Barcelona, Espanha, Novembro 2001.

[20] LACOSTE, A., ECK, D., "A Supervised Classification Algorithm for Note Onset Detection", *EURASIP Journal on Applied Signal Processing*, v. 2007, n. 43745, pp. 1–13, 2007.

[21] SHIU, Y., CHO, N., CHANG, P.-C., et al., "Robust On-line Beat Tracking with Kalman Filtering and Probabilistic Data Association (KF-PDA)", *IEEE Transactions on Consumer Electronics*, v. 54, n. 3, pp. 1–10, Agosto 2008.

[22] KLAPURI, A., "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes". In: *International Conference on Music Information*, pp. 216–221, Victoria, Canadá, Outubro 2001.

[23] DAVY, M., GODSILL, S., IDIER, J., "Bayesian Analysis of Polyphonic Western Tonal Music", *Journal of Acoustical Society of America*, v. 1119, n. 4, pp. 2498–2517, Abril 2006.

[24] RYYNÄNEN, M., KLAPURI, A., "Automatic Bass Line Transcription From Streaming Polyphonic Audio". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, v. 4, pp. 1437–1440, Hawaí, EUA, Abril 2007.

[25] POLINER, G. E., ELLIS, D. P. W., "A Discriminative Model for Polyphonic Piano Transcription", *EURASIP Journal on Applied Signal Processing*, v. 2007, n. 48317, pp. 1–9, 2007.

[26] CAMACHO, A., HARRIS, J. G., "A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music", *Journal of Acoustical Society of America*, v. 124, n. 3, pp. 1638–1652, Setembro 2008.

[27] DIXON, S., "Onset Detection Revisited". In: *Conference on Digital Audio Effects (DAFx-06)*, v. 9, pp. 133–137, Montreal, Canadá, Setembro 2006.

[28] KLAPURI, A., "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model", *IEEE Transactions on Audio, Speech and Language Processing*, v. 16, n. 2, pp. 255–266, Fevereiro 2008.

[29] MOORE, B. C. J., *An Introduction to the Psychology of Hearing*. New York: Academic: EUA, 1997.

[30] MASRI, P., *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*, Tese de doutorado, University of Bristol, Reino Unido, Dezembro 1996.

[31] DUXBURY, C., SANDLER, M., DAVIES, M., "A Hybrid Approach to Musical Note Onset Detection". In: *Digital Audio Effects Conference*, pp. 33–38, Hamburgo, Alemanha, 2002.

[32] BELLO, J. P., DUXBURY, C., DAVIES, M., et al., "On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain", *IEEE Signal Processing Letters*, v. 11, n. 6, pp. 553–556, Junho 2004.

[33] BELLO, J. P., SANDLER, M., "Phase-based Note Onset Detection for Music Signals". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 441–444, Hong Kong, Abril 2003.

[34] DUXBURY, C., BELLO, J. P., DAVIES, M., et al., "A Combined Phase and Amplitude Based Approach to Onset Detection for Audio Segmentation". In: *4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)*, pp. 275–280, Londres, Reino Unido, Abril 2003.

[35] DUXBURY, C., BELLO, J. P., SANDLER, M., et al., "A Comparison Between Fixed and Multiresolution Analysis for Onset Detection in Musical Signals". In: *Digital Audio Effects Workshop (DAFx*, pp. 207–211, Nápoles, Itália, Outubro 2004.

[36] TREES, H. L. V., *Detection, Estimation, and Modulation Theory, Part I*. 1st ed. John Wiley & Sons: Nova Iorque, EUA, 1968.

[37] FLETCHER, N. F., ROSSING, T. D., *The Physics of Musical Instruments*. 2nd ed. Springer - Verlag: Nova Iorque, EUA, 1998.

[38] GOTO, M., "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, pp. 3365–3368, Salt Lake City, Estados Unidos, Maio 2001.

[39] GOTO, M., MURAOKA, Y., "Real-time Beat Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions", *Speech Communication*, v. 27, n. 3-4, pp. 311–335, 1999.

[40] GOTO, M., "A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals". In: *18th International Congress on Acoustics*, pp. 1085–1088, Kyoto, Japão, 2004.

[41] GOTO, M., "An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds", *Journal of New Music Research*, v. 30, n. 2, pp. 159–171, 2001.

[42] LAHAT, M., NIEDERJOHN, R., KRUBSACK, D. A., "Spectral Autocorrelation Method for Measurement of the Fundamental Frequency of Noise-Corrupted Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, v. 6, n. 6, pp. 741–750, Junho 1987.

[43] KUNIEDA, N., SHIMAMURA, T., SUZUKI, J., "Robust Method of Measurement of Fundamental Frequency by ACLOS - Autocorrelation of Log Spectrum", *IEEE Transactions on Acoustics, Speech and Signal Processing*, v. 1, n. 1, pp. 232–235, Maio 1996.

[44] DE CHEVEIGNÉ, A., "Separation of Concurrent Harmonic Sounds: Fundamental Frequency Estimation and a Time Domain Cancellation Model for Auditory Processing", *Journal of the Acoustical Society of America*, v. 93, n. 6, pp. 3271–3290, Junho 1993.

[45] DE CHEVEIGNÉ, A., "Concurrent Vowel Identification. III. A Neural Model of Harmonic Interference Cancellation", *Journal of the Acoustical Society of America*, v. 101, n. 5, pp. 2857–2865, Maio 1997.

[46] MEDDIS, R., HEWITT, M. J., "Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery", *Journal of the Acoustical Society of America*, v. 89, n. 6, pp. 2866–2894, Junho 1991.

[47] TOLONEN, T., KARJALAINEN, M., "A Computationally Efficient Multipitch Analysis Model", *IEEE Transactions on Speech and Audio Processing*, v. 8, n. 6, pp. 708–716, Novembro 2000.

[48] KARJALAINEN, M., TOLONEN, T., "Multi-Pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis". In: *IEEE Transactions on Acoustics, Speech and Signal Processing*, v. 2, n. 2, *1*, pp. 929–932, Phoenix, Março 1999.

[49] HÄRMÄ, A., KARJALAINEN, M., SAVIOJA, L., et al., "Frequency-Warped Signal Processing for Audio Applications", *Journal of the Audio Engineering Society*, v. 48, n. 11, pp. 1011–1031, Novembro 2000.

[50] KLAPURI, A., "Multiple Fundamental Frequency Estimation by Harmonicity and Spectral Smoothness", *IEEE Trans. Speech and Audio Processing*, v. 11, n. 6, pp. 804–816, Novembro 2003.

[51] KLAPURI, A., "Multipitch Estimation and Sound Separation by the Spectral Smoothness Principle". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, v. 5, pp. 3381–3384, Salt Lake City, EUA, Maio 2001.

[52] PARSONS, T. W., "Separation of Speech from Interfering Speech by Means of Harmonic Selection", *Journal of the Acoustical Society of America*, v. 60, n. 4, pp. 911–918, Outubro 1976.

[53] "http://www.music.mcgill.ca/resources/mums/html", Site da McGill University Master Samples, Acessado em 26/02/2009, McGill University.

[54] COOLEY, J. W., TUKEY, J. W., "An Algorithm for the Machine Computation of Complex Fourier Series", *Mathematics of Computation*, v. 19, n. 90, pp. 297–301, Abril 1965.

[55] WILLIAM, D., BROWN, E., *Theoretical Foundations of Music*. Wadsworth: Belmont, EUA, 1978.

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

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

[58] GRAZIOSI, D. B., DOS SANTOS, C. N., NETTO, S. L., et al., "A Constant-Q Spectral Transformation with Improved Frequency Response". In: *Proc. of ISCAS 2004 - Int. Symp. on Circuits and Systems*, v. 5, pp. 544–547, Vancouver, Canadá, Maio 2004.

[59] DOS SANTOS, C. N., NETTO, S. L., BISCAINHO, L. W. P., et al., "A Modified Constant-Q Transform for Audio Signals". In: *Proc. of ICASSP 2004 - International Conference on Audio, Speech, and Signal Processing*, v. 2, pp. 469–472, Montreal, Canadá, Maio 2004.

[60] DINIZ, P. S. R., DA SILVA, E. A. B., NETTO, S. L., *Digital Signal Processing: System Analysis and Design*. Cambridge University Press: Cambridge, Reino Unido, 2002.

[61] HAYKIN, S., VAN VEEN, B., *Signals and Systems*. 2nd ed. Wiley: Hoboken, EUA, 2002.

[62] FARHANG-BOROUJENY, B., LIM, Y. C., "A Comment on the 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.

[63] LIM, Y. C., "Frequency-Response Masking Approach for the Synthesis of Sharp Linear Phase Digital Filters", *IEEE Transactions on Circuits and Systems*, v. 33, n. 4, pp. 357–364, 1986.

[64] LIM, Y. C., FARHANG-BOROUJENY, B., "Analysis and Optimum Design of the FFB". In: *Proc. of ISCAS 1994 - IEEE Int. Symp. on Circuits and Systems*, v. 2, pp. 519–512, Londres, Reino Unido, Junho 1994.

[65] WEI, L. J., LIM, Y. C., "Designing the Fast Filter Bank with Minimum Complexity Criterion". In: *Proc. of ISSPA 2003 - Int. Symp. on Signal Processing Applications*, v. 2, pp. 279–282, Paris, França, Julho 2003.

[66] KOTHE, I., *Técnicas de Análise Espectral de Linhas Musicais*, Dissertação de mestrado, Universidade Federal do Rio de Janeiro, PEE / COPPE, Rio de Janeiro, Brasil, 2006.

[67] KLAPURI, A., *Automatic Transcription of Music*, Tese de mestrado, Tampere University of Technology, Institute of Signal Processing, Finlândia, 1998.

[68] DANIEL, A., EMIYA, V., DAVID, B., "Perceptually-Based Evaluation of the Errors Usually Made When Automatically Transcribing Music". In: *ISMIR*, pp. 550–555, Pensilvânia, EUA, 2008.

[69] BISCAINHO, L. W. P., *Restauração Digital de Sinais de Áudio Provenientes de Gravações Musicais Degradadas*, Tese de doutorado, Universidade Federal do Rio de Janeiro, PEE / COPPE, Rio de Janeiro, Brasil, 2001.

[70] SZCZUPAK, A. L., BISCAINHO, L. W. P., "Identificação de Notas Musicais em Registros de Violão Solo". In: *Anais do 7o Congresso Brasileiro de Engenharia de Áudio da AES-Brasil*, São Paulo, Brasil, Maio 2009.

[71] TYGEL, A., BISCAINHO, L. W. P., "Sound Source Separation via Nonnegative Matrix Factor 2-D Deconvolution Using Linearly Sampled Spectrum". In: *Anais do 7o Congresso Brasileiro de Engenharia de Áudio da AES-Brasil*, São Paulo, Brasil, Maio 2009.

[72] MED, B., *Teoria da Música*. 4th ed. Musimed: Brasília, Brasil, 1996.

[73] "http://www.midi.org", Site do MIDI Manufacturers Association, Acessado em 25/05/2009.

Apêndice A

Conceitos básicos de música

Neste apêndice, serão apresentados alguns conceitos básicos sobre música que são utilizados ao longo da dissertação. Para uma referência mais detalhada sobre o tema, o leitor deve se referir a [72]. Os conceitos aqui apresentados serão tratados na ordem em que aparecem no texto

- **Oitava musical** é conjunto de 12 notas: Dó♯, Dó, Ré, Ré♯, Mi, Fá, Fá♯, Sol, Sol♯, Lá, Lá♯ e Si. Essas notas são definidas de modo que a freqüência fundamental de uma nota seja igual à freqüência fundamental multiplicada por $\sqrt[12]{2}$. Essa razão é definida como a distância de 1 semi-tom. Essa estrutura é a base da música ocidental e é conhecida como **temperamento igual**.

- **Notação por cifras** é composta por uma letra (de A a G), um símbolo opcional e um número (de 0 a 9). A letra define a nota segundo a Tabela A.1. O símbolo pode ser ♯, que denota o sustenido; ou ♭, que significa bemol. O sustenido indica que a altura da nota deve ser aumentada em um semi-tom, enquanto que o bemol indica que a altura da nota deve ser reduzida em um semi-tom. O número indica a oitava musical em que a nota se encontra.

- ***Pitch*** é a freqüência fundamental de um som como é percebida pelo ouvido humano. De acordo com a terminologia acústica da ANSI, é o atributo auditivo de um som de acordo com o qual os sons poderiam ser ordenados em uma escala de baixo para alto. A **altura** da nota denota o ***pitch***.

- **Intensidade** ou **dinâmica** da nota se refere ao quão forte a nota deve ser

Tabela A.1: Associação entre notas e cifras.

Notas	Cifras
Dó	C
Ré	D
Mi	E
Fá	F
Sol	G
Lá	A
Si	B

executada, e pode ser classificada, de modo crescente, como *piano*, *mezzo* e *forte*. O conjunto das variações de intensidades é chamado de **dinâmica**.

- **Partitura** é uma representação escrita de música padronizada mundialmente. Tal como qualquer outro sistema de escrita, dispõe de símbolos próprios (tradicionalmente, notas musicais desenhadas sobre um pentagrama e notação adicional) indicando a altura e a duração de cada som, além de recursos de expressividade solicitados pelo autor.

- ***Tremolo*** é uma oscilação controlada na amplitude de uma nota musical.

- ***Vibrato*** é uma oscilação controlada no *pitch* de uma nota musical.

- ***Trinado*** é a alternância rápida entre duas notas próximas.

- ***Glissando*** é a seqüência ascendente ou descendente de notas executadas em rápida sucessão, de modo a não serem individualmente articuladas, mas sim tocadas em um movimento contínuo.

- **Acorde** é um conjunto de três ou mais notas executadas simultaneamente. A nota de mais baixa freqüência geralmente é a que define o acorde e é chamada de tônica.

- **Tonalidade** é o sistema de sons baseados nas escalas onde os seus graus são observados de acordo com sua função dentro da harmonia. Cada um dos graus de uma escala desempenha funções próprias na formação e concatenação dos acordes.

- **Transposição** é o processo de se modificar a altura de uma nota ou coleção de notas por um intervalo constante. Quando se transpõe uma música, automaticamente modifica-se a tonalidade em que se encontra.

- **Harmonicidade** é a medida de quão harmônicas são as parciais de um som, isto é, de quão exatas são suas posições em relação às posições harmônicas ideais.

Apêndice B

O protocolo MIDI

O MIDI (Musical Instrument Digital Interface) [73] é um protocolo que possibilita que instrumentos musicais eletrônicos, computadores, sintetizadores, placas de som, *samplers* e *drum-machines* se comuniquem, se sincronizem um com o outro e controlem uns aos outros. Ele não transmite sinal de áudio ou mídia. Ao invés disso, transmite mensagens que representam eventos como o *pitch* e a intensidade de notas musicais a serem tocadas, sinais de controle para parâmetros como volume, *vibrato* e sinais de *clock* para determinar o tempo da música. Por isso, pode ser considerado como uma espécie de partitura a ser lida por dispositivos eletrônicos específicos. Como um protocolo eletrônico, é conhecido por ser largamente empregado na indústria.

B.1 Mensagens

Todos os controladores compatíveis com MIDI, instrumentos musicais e *softwares* compatíveis com MIDI seguem a mesma especificação MIDI 1.0. Por isso, interpretam qualquer mensagem MIDI da mesma forma e podem, assim, se comunicar. Por exemplo, se uma nota é tocada em um controlador MIDI, ela irá soar no *pitch* certo no instrumento MIDI cujo conector MIDI-IN é conectado ao conector MIDI-OUT do controlador.

Quando uma música é tocada em um instrumento MIDI (ou controlador), há a transmissão de mensagens em um canal MIDI a partir do conector MIDI-OUT. Uma típica seqüência de mensagens por um canal MIDI correspondendo a uma tecla

sendo pressionada e liberada num teclado consiste das etapas abaixo:

1. O usuário pressiona uma tecla com uma velocidade específica (que é geralmente traduzida no volume de uma nota, mas pode ser também usada pelo sintetizador para definir características do timbre). Isso faz com que o instrumento envie uma mensagem de *note-on*.

2. O usuário modifica a pressão aplicada na tecla enquanto a está pressionando, o que é uma técnica denominada *aftertouch*. O instrumento envia uma ou mais mensagens de *aftertouch*.

3. O usuário, então, libera a tecla, novamente com a possibilidade de a velocidade de liberação da tecla controlar alguns parâmetros. Isso faz com que o instrumento envie uma mensagem de *note-off*.

Note-on, *aftertouch* e *note-off* são todas mensagens de canal MIDI. Para mensagens do tipo *note-on* e *note-off*, a especificação MIDI define um número (de 0 a 127) para todos os possíveis *pitches* (C, C♯, D etc), e esse número é incluído na mensagem.

Outros parâmetros podem ser transmitidos nas mensagens de canal também. Por exemplo, se o usuário usa a roda-de-*pitch* de um instrumento, presente em grande parte dos teclados de hoje em dia, esse gesto é transmitido através do MIDI usando a série de mensagens *pitch bend*, que também é uma mensagem de canal. O instrumento musical gera mensagens de forma autônoma. Tudo que o músico deve fazer é tocar as notas (ou fazer algum gesto que produza mensagens MIDI). Essa abstração consistente e automatizada de gestos musicais poderia ser considerada o coração do padrão MIDI.