Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

Trabalho de Formatura Supervisionado do
Bacharelado em Ciência da Computação
Monografia de Iniciação Científica

Sistema Automático de Transcrição Melódica

por

Adriano Brito Mitre

Supervisor: Prof. Dr. Marcelo Gomes de Queiroz

São Paulo, 2005

À minha mãe, pela arte. Ao meu pai, pela ciência.

Sumário

Parte I

Parte Técnica

Capítulo 1

Introdução

1.1 A Transcrição Melódica

O processo de transcrição melódica pode ser entendido como a obtenção de uma representação simbólica de uma melodia a partir de sua audição.

Músicos, em geral, são capazes de ouvir uma peça e, em seguida, escrever a partitura (Fig. 1.1) correspondente. Partituras são a maneira usual de se representar música simbolicamente.

Figura 1.1: Exemplo de partitura.

Uma partitura é constituída essencialmente de notas. Notas são eventos musicais discretos, identificados por uma série de parâmetros, como intensidade, duração e altura (qualidade de grave ou agudo). Informações relativas ao instrumento que executa a nota, bem como à maneira precisa de fazê-lo também podem ser consideradas atributos de uma nota. Partituras geralmente carregam estas informações.

A principal finalidade de uma partitura é permitir que um músico, ao lê-la, seja capaz de executar a música ali representada. Logo, uma partitura deve carregar informações suficientes para tanto.

O objetivo deste trabalho não é propriamente a obtenção de partituras, mas de uma representação simbólica do tipo MIDI*. Neste formato, as principais informações referentes às notas, tais como altura e duração, são preservadas.

Expresso de maneira formal, transcrição significará, para efeitos do presente trabalho, a obtenção de um conjunto de quádruplas da forma $(t_0, n, \Delta t, v)$, onde n denota a altura da nota, v sua intensidade, t_0 seu início e Δt sua duração. No restante do texto iremos nos referir a essas quádruplas simplesmente como notas.

Convém esclarecer que, mesmo não tendo sido este o foco, o sistema apresentado neste trabalho pode ser utilizado, junto com sistemas como o *Lilypond*†, para produzir partituras. Neste caso o usuário deve fornecer ao sistema certas informações sobre a melodia de entrada, como o andamento e a forma de compasso adotados, para que as partituras obtidas sejam inteligíveis.

1.2 Aplicações de uma Ferramenta Automática

Uma ferramenta computacional capaz de automatizar o processo de transcrição musical é desejável por muitas razões. Entre suas diversas aplicações destacam-se:

- Geração automática de partituras.
- Codificação estruturada de áudio – Os formatos de multimídia mais recentes — MPEG-7, por exemplo — visam a padronizar o armazenamento de informação simbólica junto aos sinais de áudio e vídeo.
- Manipulação simbólica e resíntese – Obter a representação MIDI de uma gravação acústica possibilita, por exemplo, resintetizar com timbre de saxofone uma melodia gravada ao clarinete.
- Busca de informação musical – A transcrição representa parte importante de sistemas de *query-by-humming* (por exemplo [López e Rocamora, 2005]). Nestes sistemas, o usuário realiza consultas em uma base dados musicais entoando um trecho de sua melodia.
- Análise de improvisação musical – A representação simbólica de uma performance permite que se responda, de maneira automatizada, perguntas como “qual o padrão rítmico mais freqüente?”
- Processamento musicalmente significativo de áudio – Tendo acesso à descrição simbólica do conteúdo de um sinal de áudio, o programa de edição pode, por exemplo, remover automaticamente todas as notas dó de uma performance.

*Acrônimo para *Musical Instrument Digital Interface*. Formato padrão para troca de dados relativos à performance musical e parâmetros entre dispositivos musicais eletrônicos.

†idem, http://lilypond.org/web

- Aprendizado musical auxiliado por computador – Torna-se possível realizar exercícios para o "ouvido musical" utilizando o próprio instrumento acústico, o que até então era privilégio dos instrumentos eletrônicos.

1.3 Cenário Atual

O problema da transcrição automática de melodias, embora estudado há vários anos (vide [Roads, 1996, cap.12] para uma taxonomia), dificilmente pode ser considerado um problema fechado. Uma variedade de aspectos técnicos e cognitivos faz corresponder a cada solução proposta um novo conjunto de situações e exemplos (sintéticos ou naturais) em que a solução não fornece a resposta correta ou não fornece a resposta a tempo. Estes dois aspectos - robustez e eficiência - são fundamentais, por exemplo, durante uma performance interativa.

Dentre as dificuldades técnicas destaca-se o problema da resolução temporal versus resolução de freqüência, que será discutido em detalhes no Capítulo 6. Do lado cognitivo pode-se mencionar que a determinação da freqüência fundamental em um fragmento de áudio é um problema que em geral não admite solução objetiva e única, mesmo de posse do espectro contínuo (com precisão infinita) do fragmento em questão. Todo método de solução é heurístico por natureza e tem maior ou menor grau de acerto de acordo com a afinidade do fragmento analisado com as hipóteses implícitas no método.

Isto posto, pode-se compreender o sistema aqui descrito como uma proposta robusta de solução do problema, fornecendo ao usuário diversas opções de técnicas de refinamento espectral e uma heurística inédita, baseada em estimativas confiáveis de parciais, para a determinação de freqüência fundamental.

Ressalte-se que o trabalho não se restringiu ao estudo do ferramental teórico necessário à solução do problema, mas consistiu também na modelagem e implementação do sistema que será apresentado.

Capítulo 2

Fundamentos

2.1 Sons Harmônicos

Um som é classificado como harmônico se a maior parte de sua energia for devida à oscilações com freqüência $n \cdot f$, para alguma freqüência f. Denominar-se-á freqüência fundamental de um som ao maior valor f que verificar esta condição. Por exemplo, o som cujo espectro é visto na Figura 2.1 possui freqüência fundamental de $466,16$ Hz.

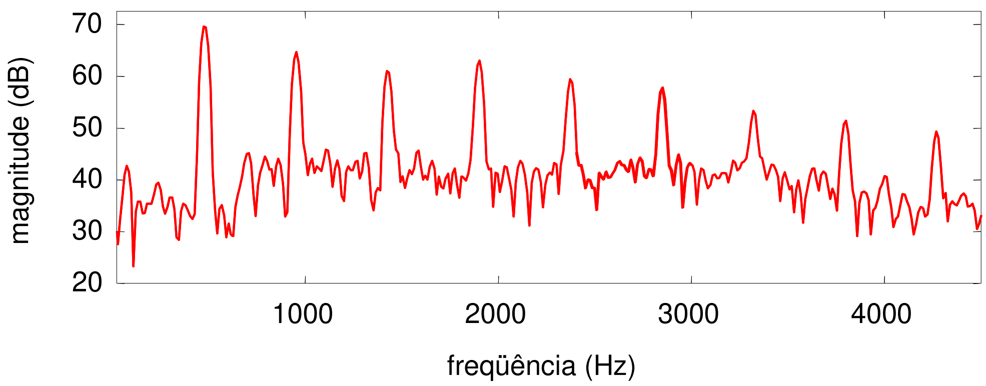


Figura 2.1: Exemplo de espectro de som harmônico.

Além da definição dada acima, existe uma segunda forma de definir freqüência fundamental: "denomina-se freqüência fundamental de um som à freqüência do som puro (senóide) que produz mesma sensação de altura". Convém notar que esta segunda definição não é restrita a sons harmônicos, mas se aplica a qualquer som com altura bem definida. Altura, no contexto deste trabalho, designará a qualidade que distingue um som grave de um som agudo e, portanto, não tem relação com o volume de um som. A língua inglesa possui o termo *pitch*, que não provoca esta ambigüidade.

Uma propriedade importante de sons harmônicos é o fato de sua altura ser determinada por sua freqüência fundamental (no sentido da primeira definição). Posto de maneira mais precisa, para sons harmônicos as duas definições são equivalentes.

Quase todos os instrumentos instrumentos melódicos* empregados na música ocidental produzem sons harmônicos. As razões para isso se encontram na física dos instrumentos, baseados em cordas ou em tubos cônicos ou cilíndricos. Os modos de vibração destes instrumentos constituem séries aproximadamente harmônicas. Exceções importantes são os instrumentos tonais baseados em membranas, como o xilofone a marimba e o vibrafone.

2.2 Modelo do Sinal

O sistema espera como entrada um sinal constituído por uma sucessão de notas e pausas (i.e., períodos de silêncio). A evolução dinâmica de uma nota musical pode ser descrita, na maior parte das entradas naturais, pelo envelope ADSR, que consiste em uma sucessão de quatro estágios: ataque, amortecimento, sustentação e dissipação. A inspeção de certos aspectos destes estágios é muito importante no projeto de um sistema transcritor. A Tabela 2.1 apresenta um resumo dessas características.

Tabela 2.1: Modelo evolutivo de uma nota típica.

Estágio	Energia	Caráter
Ataque	Cresce rapidamente	Transiente
Amortecimento	Decresce	Em estabilização
Sustentação	Permanece constante ou decresce lentamente	Periódico (som harmônico)
Dissipação	Decresce rapidamente	Periódico (som harmônico)

O modelo de sinal adotado no artigo baseia-se em [Serra e Smith III, 1990], mais especificamente na adaptação feita por [Desainte-Catherine e Marchand, 2000]. Essencialmente o sinal de áudio é decomposto como

$$s(t) = d(t) + e(t) \tag{2.1}$$

onde

*Instrumentos musicais com capacidade de produzir melodias. Por exemplo, o piano e a flauta são instrumentos melódicos; o pandeiro não.

- $d(t)$ é a parte determinística (componente de banda-estreita)
- $e(t)$ é a parte estocástica (componente de banda-larga)

A parte determinística é constituída por uma soma de osciladores senoidais com freqüências e amplitudes que variam lentamente. Um oscilador de variação lenta é geralmente chamado de parcial, senóide ou componente senoidal. Portanto, a parte determinística pode ser formalizada como

$$d(t) = \sum_{p=1}^{P} osc\left(f_p(t), a_p(t)\right) \tag{2.2}$$

onde P é o número de parciais, dados por

$$osc\left(f_p(t), a_p(t)\right) = a_p(t) \cdot \cos\left(\phi_p(t)\right) \tag{2.3}$$

com

$$\frac{d\phi_p}{dt}(t) = 2\pi f_p(t) \quad \text{i.e.} \quad \phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u) \cdot du \tag{2.4}$$

Como esperamos sons harmônicos, sabemos que as freqüências dos parciais relacionam-se de acordo com

$$f_p(t) = p \cdot f_1(t) \tag{2.5}$$

Deste modo, temos a garantia de que, para todo instante t, os parciais estarão distantes, no domínio da freqüência, de pelo menos $f_1(t)$. Formalmente:

$$\min_{1 \le i < j \le P} \left\{ \left| f_i(t) - f_j(t) \right| \right\} \ge f_1(t) \tag{2.6}$$

A garantia é expressa como desigualdade, e não igualdade, por duas razões. A primeira é que alguns instrumentos produzem séries harmônicas incompletas, como um sintetizador de onda quadrada, que produz apenas harmônicos ímpares. A segunda razão é que em algumas famílias de instrumentos as freqüências dos parciais desviam de maneira sistemática dos múltiplos da fundamental.

Nos instrumentos de corda, por exemplo, a freqüência dos parciais é dada pela fórmula $f_n = n \cdot f_1 \sqrt{1 + B \cdot (n-1)^2}$, onde B é o coeficiente de inarmonicidade. Este fenômeno decorre do fato de a velocidade do som em uma corda não ser independente da freqüência. O efeito psicoacústico da inarmonicidade em instrumentos de corda foi estudado em [Järveläinen et al., 2000].

Os instrumentos melódicos de percussão em membranas, como a marimba, o xilofone ou o vibrafone, divergem ainda mais acentuadamente do modelo harmônico, tanto pela freqüência dos parciais quanto pela densidade das séries harmônicas. Apesar disso, experimentos indicam que nosso sistema é capaz de transcrever melodias executadas com estes instrumentos. Em primeiro lugar porque essas divergências não invalidam o espaçamento mínimo entre parciais no âmbito da freqüência, necessário à precisa estimação dos mesmos. Em segundo, porque princípios psicoacústicos e de seletividade harmônica guiaram a construção de nosso estimador de F0 (freqüência fundamental), tornando-o mais condizente com a percepção de altura.

A parte estocástica é o que resta do sinal quando se lhe subtrai a parte determinística. Este sinal residual é plenamente descrito por sua densidade espectral de potência, que fornece a potência esperada em cada banda crítica. A justificação dessa afirmação foge ao escopo deste texto e pode ser encontrada em [Serra e Smith III, 1990].

Como se observa na Tabela 2.1, durante o ataque de uma nota, o sinal tem caráter transiente. A Transformação de Fourier, por sua vez, é apropriada para sinais periódicos. No entanto, o que poderia constituir um problema foi revertido em favor do sistema, que se utiliza de métricas de aperiodicidade do sinal para determinar, com maior exatidão, o início das notas (*note onsets*). Técnicas de detecção de ínicio de nota são apresentadas no Capítulo 9.

2.3 Abordagem Escolhida e Justificação

A abordagem adotada neste trabalho privilegiou a estimação de freqüência fundamental. Esta decisão foi baseada na hipótese, confirmada pelos resultados deste trabalho, de que o contorno de F0 (Figura 2.3) é tipicamente suficiente para determinar as notas presentes em um sinal de áudio.

Entretanto, as notas só apresentam característica tonal (e portanto F0) alguns centésimos após seu início (vide §9.1). Por conta disso, fez-se vantajosa a construção de um módulo específico para detecção de início de notas.

Por fim, foi concebido o módulo de produção de notas com a finalidade de integrar as informações provenientes de cada um dos fluxos. O resultado desta abordagem são transcrições com notas corretas não apenas em termos de altura, mas também de localização temporal e, conseqüentemente, em termos rítmicos.

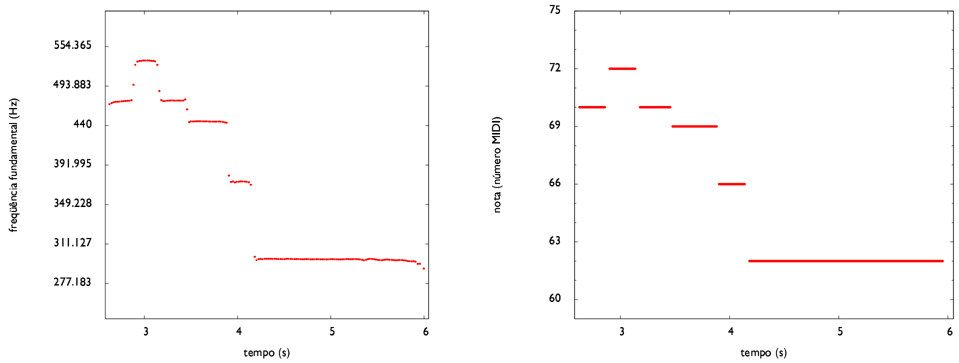


Figura 2.2: O contorno de F0 é tipicamente suficiente para determinar as notas presentes no sinal.

Capítulo 3

Arquitetura do Sistema

O sistema tem como entrada um sinal de áudio monofônico (por exemplo uma gravação acústica) e produz uma transcrição de seu conteúdo melódico.

Uma parcela significativa dos sistemas de transcrição encontrados na literatura mescla tratamento minucioso de alguns aspectos do problema e hipersimplificação de outros. O cuidado devotado a um certo aspecto é tipicamente determinado por sua proximidade com o foco da pesquisa. Como a praxe é apresentar resultados apenas do sistema como um todo, não é possível dimensionar o impacto individual de cada decisão de projeto.

Em vista disso, nosso sistema foi projetado como um conjunto de módulos independentes e desacoplados, ilustrados na Figura 3.1. Além de oferecer maior liberdade de experimentação, a modularidade de nosso sistema torna mais fácil identificar a causa de uma transcrição mal-sucedida. Por fim, amplia sua funcionalidade, permitindo fácil integração com outros sistemas por meio de *pipes*. Com efeito, a comunicação entre os diferentes módulos de nosso sistema pode ser realizada com *pipes*.

O sistema foi codificado na linguagem C padrão ANSI, escolhida em função de sua portabilidade e eficiência. Apesar de não ser uma linguagem orientada a objetos, conceitos deste paradigma foram freqüentemente considerados durante o desenvolvimento do sistema, além das boas práticas de programação [Kernighan e Pike, 2000].

Os capítulo 4 a 10 se ocupam de apresentar, em detalhes, cada um dos módulos do sistema de transcrição.

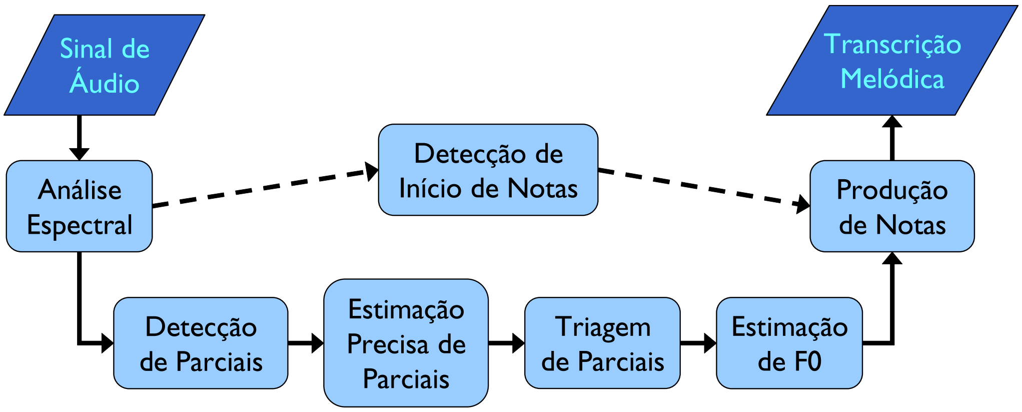


Figura 3.1: Módulos do sistema e seu fluxo de dados.

Capítulo 4

Análise Espectral

O sistema obtém estimativas espectrais por meio da Transformada Discreta de Fourier de Curto Tempo (abreviada em inglês por STDFT). O leitor não familiarizado com a análise de Fourier pode consultar, por exemplo, [Smith III, 2003]. O espectro associado a um dado instante t é obtido da seguinte forma. Seja F_s^{-1} o período interamostral do sinal de entrada, medido em segundos. A duração de cada quadro (*frame*) a ser analisado será determinada pelo usuário, respeitando $\tau = N \cdot F_s^{-1}$ para algum inteiro positivo N. Então a seqüência de N amostras contidas no intervalo de tempo $[t - \tau/2, t + \tau/2]$ será multiplicada ponto a ponto por uma função de suavização (também chamada de janela) e o resultado deste produto será por fim transformado pela STDFT. Se o valor de N for apropriado, o sistema realizará o cálculo por meio do algoritmo FFT (*Fast Fourier Transform*), que apresenta desempenho assintótico $O(N \lg N)$, contra $O(N^2)$ da implementação trivial da DFT.

Para um sinal de entrada de duração T segundos, o sistema calculará o espectro associado a todos os instantes $t \in [0, \mathrm{T}]$ que satisfaçam, para algum inteiro m, a condição $t = t_0 + m \cdot \tau \cdot (1 - \alpha)$, onde $\alpha \in [0, 1)$ denota o fator de sobreposição escolhido pelo usuário. A parcela t_0 compensa o viés de quadros com número par de amostras, associando-os não ao instante da $N/2$-ésima amostra, mas ao ponto médio entre as amostras $N/2$ e $N/2 + 1$. Portanto, se N for par, $t_0 = F_s^{-1}/2$. Caso contrário, $t_0 = 0$. Para os casos em que $t < \tau/2$ ou $t > \mathrm{T} - \tau/2$ são requeridas amostras fora da extensão do sinal de entrada. O sistema lida com a questão interpretando o sinal como precedido e sucedido por silêncio (amostras de valor nulo).

Os módulos descritos nas seções 5 a 8 operam em regime *por-quadro*. Por conta disso, informações referentes ao tempo serão omitidas nessas seções.

Capítulo 5

Detecção de Parciais

O módulo anterior fornece, para cada quadro, seu espectro complexo, cujo k-ésimo escaninho (*bin*) será denotado S_k. Ao escaninho de índice k corresponde a freqüência $2\pi k/N$ radianos ou, equivalentemente, $2\pi F_s k/N$ Hertz. Para a estimação de F0, contudo, interessam apenas informações referentes aos parciais, e não todo o espectro.

Em condições razoáveis, espera-se que cada parcial no sinal de entrada produza um máximo local no espectro de magnitude. A recíproca, no entanto, não é verdadeira. Por conta disso, propuseram-se diversas heurísticas para discriminar entre máximos locais decorrentes de ruído daqueles produzidos por parciais. Uma estratégia adotada com freqüência em sistemas de análise/resíntese é o acompanhamento evolutivo de parciais (*partial tracking*), como em [Lagrange et al., 2003], que não pôde ser incorporado ao sistema proposto por não satisfazer a exigência de operação em regime *por-quadro*. É fácil, contudo, acrescentar ao sistema um módulo de filtragem *offline* baseado nessa estratégia.

A heurística utilizada no sistema exige que cada máximo local no espectro de magnitude seja superior aos mínimos locais que lhe são adjacentes em pelo menos $\delta_{\min}$ decibéis. Formalmente, dado um $k \in \{2, \ldots, N/2-1\}$, S_k será classificado como parcial somente se verificar

$$20\cdot\log_{10}\left(|S_k|/|S_{k^+}|\right) \geq \delta_{\min} \quad \text{onde} \quad k^+ = \max\left\{j > k : |S_k| \geq |S_{k+1}| \geq \ldots \geq |S_j|\right\}$$

e também

$$20\cdot\log_{10}\left(|S_k|/|S_{k^-}|\right) \geq \delta_{\min} \quad \text{onde} \quad k^- = \min\left\{j < k : |S_j| \leq |S_{j+1}| \leq \ \ldots \leq |S_k|\right\}.$$

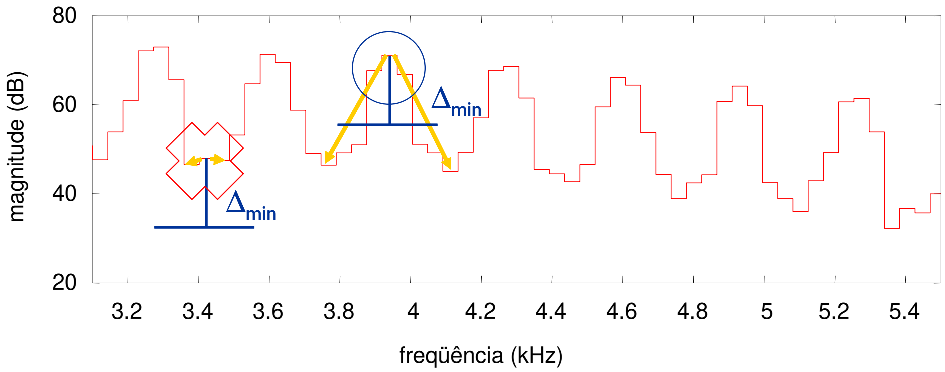


Figura 5.1: Exemplo de funcionamento da heurística de detecção de parciais.

Capítulo 6

Estimação Precisa de Parciais

6.1 Introdução

Para a correta determinação de notas a partir das freqüências fundamentais é necessária, na escala de doze notas com temperamento igual (i.e., o sistema de afinação padrão no ocidente), uma exatidão de freqüência de pelo menos $\mathrm{F0}_{\mathrm{min}} \cdot \left(\sqrt[12]{2}-1\right)$ Hz, onde $\mathrm{F0}_{\mathrm{min}}$ denota a freqüência fundamental da nota mais grave da seqüência melódica de entrada. Caso se deseje extrair nuances expressivas como vibratos e glissandos, a precisão necessária deverá ser muito maior.

A resolução (i.e., capacidade de discriminação) de freqüência da STDFT é de $L_w \cdot \tau^{-1}$ Hz, onde L_w denota a largura do lobo central da resposta de freqüência da função de suavização, medida em escaninhos. A exatidão de freqüência da STDFT é, por sua vez, dada por τ^{-1} e, portanto, difere da resolução apenas por um fator pequeno (tipicamente vale que $L_w \leq 4$). Deste modo, para que seja possível distinguir as notas de um violão de 6 cordas, cuja corda mais grave é convencionalmente afinada em 82,4 Hz, é necessário um quadro de pelo menos $\left(82,4 \cdot \left(\sqrt[12]{2}-1\right)\right)^{-1} \simeq 204$ ms.

Sinais musicais raramente possuem característica estacionária por tanto tempo. Assim, a utilização de quadros muito longos prejudica a resolução temporal, de modo que o conteúdo harmônico de duas ou mais notas consecutivas pode se misturar no espectro de um quadro. Além disso, uma precisão temporal de 20 ms exigiria, no caso do violão, um fator de sobreposição de quadros de 90% e, portanto, uma carga computacional dez vezes maior. Em suma, quadros longos não são a abordagem correta.

Como o sinal de entrada é suposto monofônico, o espaçamento mínimo de freqüência entre dois parciais quaisquer será de pelo menos $\mathrm{F0}_{\mathrm{min}}$ Hz. Existem, para situações como essa, técnicas para obter estimativas mais precisas a partir de quadros com duração de $L_w \cdot (\mathrm{F0}_{\mathrm{min}})^{-1}$ segundos (um exemplo pode ser visto na Fig. 6.1). Isto representa uma grande redução na exigência de tamanho do quadro. No exemplo do violão, um quadro de $2 \cdot (82,4)^{-1} \simeq 24$ ms passa a ser suficiente, considerando-se a janela de Hamming.

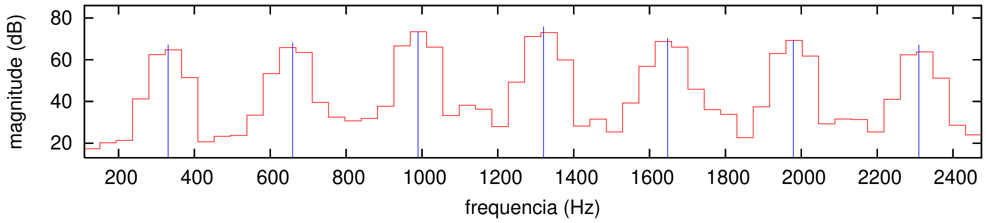


Figura 6.1: Espectro de magnitude e estimativas precisas dos parciais.

A seguir serão apresentadas as principais técnicas de estimação precisa de parciais aplicáveis a sinais musicais. Técnicas desenvolvidas exclusivamente para a janela retangular são imprópria à análise de sinais musicais, que tipicamente possuem grande densidade de componentes. Por conta disso, estas técnias não serão apresentadas neste texto. Alguma dentre estas serão citadas na Seção 6.3.3.

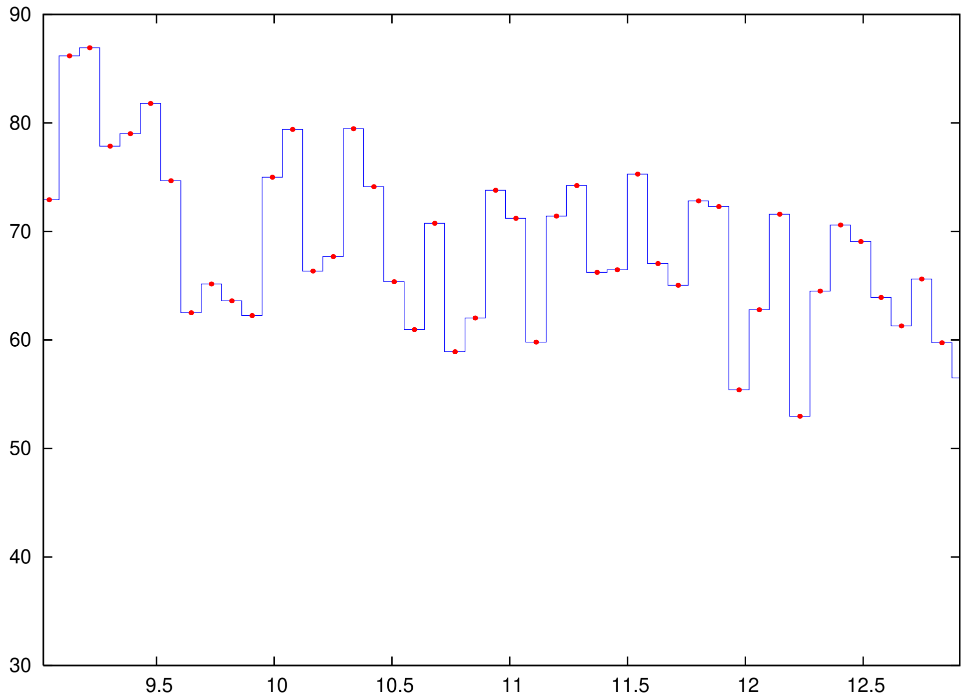
6.2 Superamostragem Espectral

A superamostragem espectral, popularmente referida como *zero-padding*, é técnica de interpolação "perfeita", pois obtém pontos que de fato pertencem ao espectro do sinal suavizado. O leitor interessado em uma demonstração deste fato deve consultar [Smith III, 2003, Ideal Spectral Interpolation]. Como o nome sugere, este método consiste no aumento da taxa de amostragem do espectro, o que pode ser alcançado por meio do cálculo da STDFT do sinal suavizado (i.e., já multiplicado pela função de suavização) concatenado a uma seqüência de zeros.

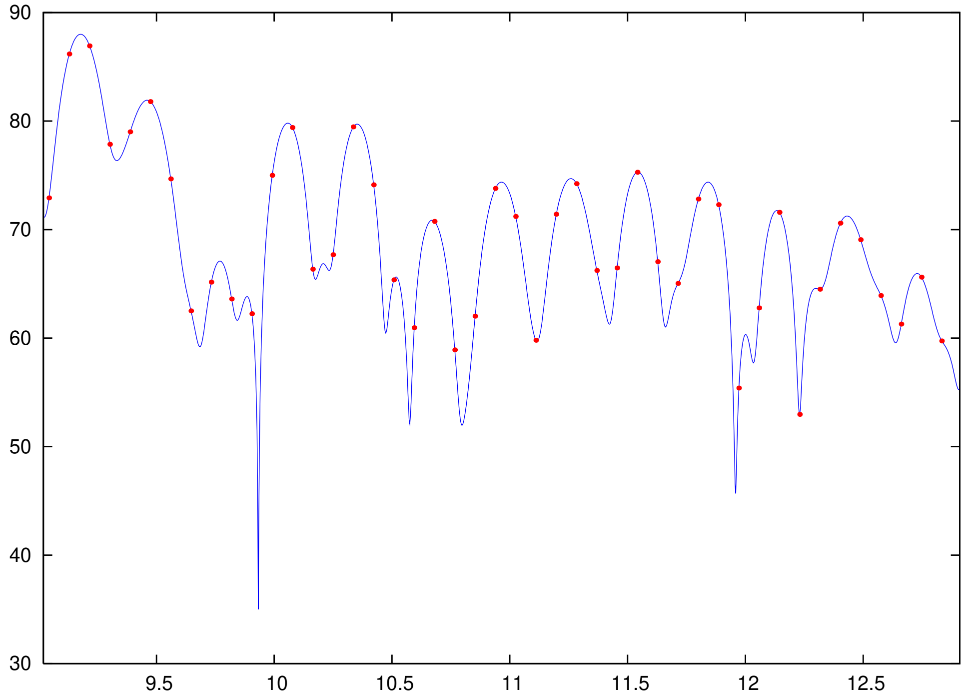
Esta técnica obtém resultados de grande exatidão, mas apresenta uma severa contrapartida: seu custo computacional. Para aumentar a precisão em m vezes é necessário calcular uma STDFT com tamanho mN. A Figura 6.2 compara o espectro de magnitude obtido pela STDFT com sua verão superamostrada.

6.3 Métodos Baseados no Espectro de Magnitude

Os métodos baseados no espectro de magnitude, também chamados de métodos de interpolação, obtêm estimativas mais precisas dos parciais por meio de interpolação com coeficientes espectrais adjacentes.



(a) Espectro de magnitude sem superamostragem.



(b) Espectro de magnitude com 32x de superamostragem.

Figura 6.2: Superamostragem: a resolução é determinada pelo função de suavização, mas a definição pode ser aumentada arbitrariamente.

6.3.1 Interpolação Parabólica

A técnica da interpolação parabólica, também referida como quadrática, beneficia-se do fato de que o lobo central da resposta de freqüência da maioria das funções de suavização assemelha-se, em escala logarítmica, a uma parábola. Com os coeficientes de magnitude do pico e seus dois vizinhos, determina-se um polinômio do segundo grau, cujo ponto máximo dará origem à nova estimativa do parcial.

A interpolaçõa parabólica é freqüentemente reforçada pelo uso concomitante de superamostragem. É possível obter um desempenho ainda melhor com a aplicação de funções de suavização específicas, calculadas pela transformada inversa de uma parábola.

6.3.2 Algoritmo de Grandke

Especificamente para a janela de Hann[*], um algoritmo de interpolação que utiliza o coeficiente vizinho de maior magnitude para obter estimativas mais consistentes do que as obtidas pela interpolação parabólica foi desenvolvido em [Grandke, 1983].

Grandke(S, k)
if $|S_{k+1}| > |S_{k-1}|$
  then $a \leftarrow |S_k|/|S_{k-1}|$
    $d \leftarrow (2a-1)/(a+1)$
    return $k - 1 + d$
  else $a \leftarrow |S_{k+1}|/|S_k|$
    $d \leftarrow (2a-1)/(a+1)$
    return $k + d$

6.3.3 Outras técnicas

Quinn, Triângulo, Baricêntrico, Agrez.

6.4 Métodos Baseados no Espectro de Fase

Há técnicas mais sofisticadas de estimação de parciais, que consideram, além da magnitude, a informação de fase presente no espectro. Essas técnicas apresentam desempenho superior, mas possuem a desvantagem de exigirem o cômputo de STDFT adicionais.

[*]não obstante seja um tributo ao metereologista austríaco Julius von Hann, alguns autores referem-se à janela como sendo de Hanning

6.4.1 Phase-Vocoder

O método do phase-vocoder faz uso, de maneira direta, da noção de que a freqüência instantânea de um oscilador harmônico é dada por sua derivada da fase. Denotando por $\tilde{\varphi}_k(n)$ a fase acumulada do k-ésimo escaninho do espectro de Fourier calculado em um quadro centrado no instante t, definimos a *freqüência instantânea* do parcial contido neste escaninho como sendo

$$\hat{f}_k(t) = \frac{1}{2\pi}\frac{d\tilde{\varphi}_k(t)}{dt} \tag{6.1}$$

Como não se tem acesso à derivada analítica da fase, uma aproximação é obtida com base na diferença da fase calculada em dois quadros próximos:

$$\hat{f}_k(t) = \frac{\tilde{\varphi}_k(t) - \tilde{\varphi}_k(t-\Delta)}{2\pi\Delta} \tag{6.2}$$

Tipicamente usa-se $\Delta = F_s^{-1}$, ou seja, os valores de fase são calculados para dois quadros deslocados de apenas uma amostra.

6.4.2 Método da Derivada

O método da derivada, apresentado inicialmente em [Desainte-Catherine e Marchand, 1998, Desainte-Catherine e Marchand, 2000], beneficia-se do fato de que a derivação opera como um filtro linear com resposta de freqüência conhecida. Assim, a razão entre a amplitude de uma senóide e de sua derivada pode ser usada para estimar sua freqüência.

Neste texto apresentaremos apenas o desenvolvimento discreto do método da derivada, restrito à primeira derivada. O leitor interessado em uma exposição analítica e generalizada para n-ésima derivada deve consultar [Marchand, 2000, §3.3].

Como na prática se trabalha com sinais cuja derivada não é conhecida, uma aproximação pela esquerda é obtida por meio do filtro dado por

$$a'_-[n] = (a[n] - a[n-1])\,F_s \tag{6.3}$$

cuja transformada Z é dada por

$$H(z) = F_s\left(1 - z^{-1}\right) \tag{6.4}$$

A partir da transformada Z, calcula-se a resposta de freqüência do filtro:

$$|H(z)| = 2F_s \sin\left(\frac{\omega}{2}\right) \tag{6.5}$$

Considerando um sinal constituído apenas por uma senóide “contida” no escaninho k, i.e., de freqüência f_k tal que $\lfloor f_k N/F_s + 0.5 \rfloor = k$, é fácil ver que

$$\frac{|S'_k|}{|S_k|} = 2F_s \sin\left(\frac{\omega}{2}\right) \tag{6.6}$$

onde S'_k denota o espectro do sinal filtrado a'. Lembrando que $\omega = 2\pi f/F_s$, é possível reescrever a Equação 6.6 como

$$\frac{|S'_k|}{|S_k|} = 2F_s \sin\left(\frac{\pi f_k}{F_s}\right) \tag{6.7}$$

A Equação 6.7 expressa a relação entre os valores S'_k e S_k, conhecidos, como função do parâmetro f_k, que desejamos estimar. É necessário, portanto, manipulá-la a fim de que f_k seja expressa em função de S'_k e S_k. Logo, o estimador da derivada, obtido a partir da Equação 6.7, é dado por

$$f_k = \frac{F_s}{\pi} \arcsin\left(\frac{|S'_k|}{2F_s\,|S_k|}\right) \tag{6.8}$$

Conforme se esperaria, a Equação 6.8 vale mesmo para sinais com múltiplos componentes, desde que haja entre estes suficiente espaçamento no domínio da freqüência.

6.4.3 Método Trigonométrico

O método trigonométrico, introduzido em [Lagrange et al., 2005], é um aperfeiçoamento do método da derivada. Este último possui como defeito seu mal desempenho com parciais de freqüências elevadas. É tão sofrível para freqüências próximas à de Nyquist que o método se fez virtualmente inútil para estimar parciais muito agudos.

A idéia central do método trigonométrico é a construção de um filtro com propriedades simétricas ao utilizado pelo método da derivada, dado pela Equação 6.3. O filtro obtido permite que se desenvolva um estimador que apresenta excelente desempenho nos agudos e que, portanto, complementa o estimador da derivada. Um estimador híbrido é então construído com os dois. A decisão de qual usar é feita com base na freqüência quantizada do parcial que se deseja estimar com precisão.

Neste texto, apresentaremos apenas o desenvolvimento analítico do método trigonométrico. O leitor interessado na justificação da validade do resultado analítico no domínio discreto deve consultar a referência original.

Considere-se um sinal e sua versão deslocada no tempo:

$$s(t) = a\cos(\omega t + \phi) \tag{6.9}$$

$$s(t-\Delta) = a\cos(\omega(t-\Delta) + \phi) \tag{6.10}$$

Somando um oscilador harmônico à sua versão deslocada no tempo, com fase normal e invertida, os seguintes sinais interferidos são obtidos:

$$s(t) - s(t-\Delta) = a'\cos\left(\omega t + \phi'\right) \tag{6.11}$$

$$s(t) + s(t-\Delta) = a''\cos\left(\omega t + \phi''\right) \tag{6.12}$$

onde

$$a' = 2a \sin(\omega\Delta/2) \tag{6.13}$$
$$a'' = 2a \cos(\omega\Delta/2) \tag{6.14}$$

Substituindo ω por $2\pi f/F_s$ e agrupando as amplitudes, obtém-se

$$\frac{a'}{a} = 2\sin\left(\frac{\pi f\Delta}{F_s}\right) \tag{6.15}$$
$$\frac{a''}{a} = 2\cos\left(\frac{\pi f\Delta}{F_s}\right) \tag{6.16}$$

Analogamente ao que foi feito da Equação 6.7 para a Equação 6.8, manipula-se a Equação 6.15 no intuito de expressar f em função das demais variáveis:

$$f^- = \frac{1}{\pi\Delta}\arcsin\left(\frac{a'}{2a}\right) \tag{6.17}$$
$$f^+ = \frac{1}{\pi\Delta}\arccos\left(\frac{a''}{2a}\right) \tag{6.18}$$

Os estimadores f^- e f^+ possuem comportamentos simétricos. O estimador f^- é equivalente ao estimador da derivada. Por conta disso, f^- apresenta excelente acurácia para freqüências abaixo de $F_s/4$, mas desempenho pífio para frequências próximas de $F_s/2$. Já o estimador f^+ exibe grande precisão para freqüências acima de $F_s/4$ e desempenho sofrível para freqüências graves.

Essa complementaridade sugere que a solução esteja em um estimador híbrido, que acumularia os benefícios de f^- e f^+. Lembremos que o picos cuja freqüência se deseja estimar já possuem uma estimativa de freqüência grosseira, associada à freqüência do centro do escaninho em que o parcial está contido. Apesar de incerta em até $F_s/(2 \cdot N)$ Hz, esta estimativa é suficientemente precisa para determinar qual dos estimadores deve ser usado. Assim, tem-se o estimador trigonométrico:

$$\hat{f}_p = \begin{cases} f_p^-, & \text{se} \;\; k_p < N/4 \\ f_p^+, & \text{c.c.} \end{cases} \tag{6.19}$$

cuja versão discreta é dada por

$$\hat{f}_p = \begin{cases} \frac{F_s}{\pi}\arcsin\left(\frac{|S_0[k_p]-S_{-1}[k_p]|}{2|S_0[k_p]|}\right), & \text{se} \;\; k_p < N/4 \\ \frac{F_s}{\pi}\arccos\left(\frac{|S_0[k_p]+S_{-1}[k_p]|}{2|S_0[k_p]|}\right), & \text{c.c.} \end{cases} \tag{6.20}$$

Apesar de utilizar implicitamente dois estimadores, o método trigonométrico requer apenas o cômputo de uma STDFT adicional. Logo, não há aumento em termos de carga computacional em relação ao método da derivada.

6.4.4 Reatribuição Espectral

O método da reatribuição espectral, introduzida em [Kodera et al., 1978] e desenvolvida em [Auger e Flandrin, 1995], visa a redispor os pontos do reticulado tempo-freqüência de acordo com os centros de energia de cada célula. Apesar de o método da reatribuição espectral incluir também uma fórmula para estimar a localização temporal de transientes, esta não será apresentada por não pertencer ao escopo deste módulo.[†] O leitor interessado na reatribuicão no tempo deve consultar, por exemplo, [Auger e Flandrin, 1995]. Um estudo abrangente do método, incluindo resultados teóricos e experimentais, foi realizado em [Hainsworth, 2004, cap.3].

A idéia central do método da reatribuição espectral é utilizar o conhecimento da derivada analítica da função de suavização — e não do sinal, como no método da derivada — para obter estimativas precisas de freqüência.

Seja a STFT do sinal s no instante t dada por

$$\mathrm{S}TFT_{s,w}(t,\Omega) = \int_{-\infty}^{+\infty} s(t+\theta) w(\theta)\, e^{-i\Omega\theta} d\theta \tag{6.21}$$

onde w denota a função de suavização utilizada. Introduzindo $\tau = t + \theta$ temos

$$\mathrm{S}TFT_{s,w}(t,\Omega) = \int_{-\infty}^{+\infty} s(\tau) w(\tau - t)\, e^{-i\Omega(\tau-t)} d\tau \tag{6.22}$$

O espectro complexo resultante é dado por

$$STFT_{s,w}(t,\Omega) = A_{s,w}(t,\Omega)\, e^{i\varphi_{s,w}(t,\Omega)} \tag{6.23}$$

Detenhamo-nos na "freqüência instantânea":

$$\Omega_{s,w}(t,\Omega) = \frac{d\varphi_{s,w}(t,\Omega)}{dt} \tag{6.24}$$

Se considerarmos a parte imaginária da derivada do logaritmo da STFT, com as equações 6.23 e 6.24, nós temos

$$\begin{aligned} \mathrm{I}m\left[\frac{d}{dt}\ln\left(\mathrm{S}TFT_{s,w}(t,\Omega)\right)\right] &= \mathrm{I}m\left[\frac{d}{dt}\ln\left(A_{s,w}(t,\Omega)\, e^{i\varphi_{s,w}(t,\Omega)}\right)\right] \\ \cdots &= \mathrm{I}m\left[\frac{d}{dt}\left(\ln\left(A_{s,w}(t,\Omega)\right) + \ln\left(e^{i\varphi_{s,w}(t,\Omega)}\right)\right)\right] \\ \cdots &= \mathrm{Im}\left[\overbrace{\frac{d}{dt}\ln\left(A_{s,w}(t,\Omega)\right)}^{\in\mathbb{R}} + i\frac{d}{dt}\varphi_{s,w}(t,\Omega)\right] \\ \cdots &= \frac{d\varphi_{s,w}(t,\Omega)}{dt} \end{aligned}$$

[†] É questionável se sua utilização no módulo de detecção de início de notas traria benefícios práticos. De todo modo, a técnica não foi implementada em nosso sistema.

Usando a Equação 6.24 obtemos a seguinte igualdade

$$\mathrm{Im}\left[\frac{d}{dt}\ln\left(\mathrm{S}TFT_{s,w}\left(t,\Omega\right)\right)\right] = \Omega_{s,w}\left(t,\Omega\right) \tag{6.25}$$

Por outro lado, pela Regra da Cadeia temos:

$$\mathrm{Im}\left[\frac{d}{dt}\ln\left(\mathrm{S}TFT_{s,w}\left(t,\Omega\right)\right)\right] = \mathrm{Im}\left[\frac{\frac{d}{dt}\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}{\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}\right] \tag{6.26}$$

É possível desenvolver o numerador utilizando a Equação 6.22 e, em seguida, aplicando a Regra de Leibniz:

$$\begin{aligned}\frac{d}{dt}\mathrm{S}TFT_{s,w}\left(t,\Omega\right) = \frac{d}{dt}\int_{-\infty}^{+\infty} s\left(\tau\right)w\left(\tau-t\right)e^{-i\Omega(\tau-t)}d\tau \\ \cdots = \int_{-\infty}^{+\infty}\frac{d}{dt}\left[s\left(\tau\right)w\left(\tau-t\right)e^{-i\Omega(\tau-t)}\right]d\tau \\ \cdots = \int_{-\infty}^{+\infty} s\left(\tau\right)\frac{d}{dt}\left[w\left(\tau-t\right)e^{-i\Omega(\tau-t)}\right]d\tau\end{aligned}$$

Aplicando a Regra do Produto ao termo entre colchetes obtemos

$$\begin{aligned}\frac{d}{dt}w\left(\tau-t\right)e^{-i\Omega(\tau-t)} = w\left(\tau-t\right)\frac{d}{dt}e^{-i\Omega(\tau-t)} + e^{-i\Omega(\tau-t)}\frac{d}{dt}w\left(\tau-t\right) \\ \cdots = i\Omega w\left(\tau-t\right)e^{-i\Omega(\tau-t)} - w'\left(\tau-t\right)e^{-i\Omega(\tau-t)}\end{aligned}$$

E com isso retomamos o desenvolvimento do numerador da Equação 6.26:

$$\begin{aligned}\int_{-\infty}^{+\infty} s\left(\tau\right)\frac{d}{dt}\left[w\left(\tau-t\right)e^{-i\Omega(\tau-t)}\right]d\tau = \cdots \\ \cdots = \int_{-\infty}^{+\infty} s\left(\tau\right)\left[i\Omega w\left(\tau-t\right)e^{-i\Omega(\tau-t)} - w'\left(\tau-t\right)e^{-i\Omega(\tau-t)}\right]d\tau \\ \cdots = i\Omega\int_{-\infty}^{+\infty} s\left(\tau\right)w\left(\tau-t\right)e^{-i\Omega(\tau-t)}d\tau - \int_{-\infty}^{+\infty} s\left(\tau\right)w'\left(\tau-t\right)e^{-i\Omega(\tau-t)}d\tau \\ \cdots = i\Omega\mathrm{S}TFT_{s,w}\left(t,\Omega\right) - \mathrm{S}TFT_{s,w'}\left(t,\Omega\right)\end{aligned}$$

Tendo concluído o desenvolvimento do numerador, retornamos à Equação 6.26:

$$\begin{aligned}\mathrm{Im}\left[\frac{\frac{d}{dt}\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}{\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}\right] = \mathrm{Im}\left[\frac{i\Omega\mathrm{S}TFT_{s,w}\left(t,\Omega\right) - \mathrm{S}TFT_{s,w'}\left(t,\Omega\right)}{\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}\right] \\ \cdots = \Omega - \mathrm{Im}\left[\frac{\mathrm{S}TFT_{s,w'}\left(t,\Omega\right)}{\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}\right]\end{aligned}$$

Logo,

$$\mathrm{Im}\left[\frac{d}{dt}\ln\left(\mathrm{S}TFT_{s,w}\left(t,\Omega\right)\right)\right]=\Omega-\mathrm{Im}\left[\frac{\mathrm{S}TFT_{s,w'}\left(t,\Omega\right)}{\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}\right] \tag{6.27}$$

Por fim, o estimador de freqüência da Reatribuição Espectral é obtido combinando-se as equações 6.25 e 6.27:

$$\Omega_{s,w}\left(t,\Omega\right)=\Omega-\mathrm{Im}\left[\frac{\mathrm{S}TFT_{s,w'}\left(t,\Omega\right)}{\mathrm{S}TFT_{s,w}\left(t,\Omega\right)}\right] \tag{6.28}$$

Note que com isso a "freqüência instantânea" foi expressada em função de duas STFT. Na primeira, emprega-se a função de suavização w e, na segunda, sua derivada analítica w'.

6.5 Estimação da Amplitude

Com exceção da interpolação parabólica e da superamostragem espectral, que obtêm diretamente estimativas de freqüência e amplitude, as técnicas apresentadas estimam apenas a freqüência dos parciais. Entretanto, conhecendo-se a resposta de freqüência da função de suavização utilizada, é possível usar os estimadores de freqüência para aperfeiçoar a estimativa de amplitude dos parciais.

Seja $\hat{f}_k$ a estimativa precisa de freqüência de um parcial "contido" no escaninho k, cuja freqüência associada é f_k, e seja W a resposta de freqüência da função de suavização, então a amplitude do parcial pode ser estimada pela fórmula

$$\hat{a}_k=\frac{a_k}{W\left(\left|\hat{f}_k-f_k\right|\right)} \tag{6.29}$$

Foi mostrado em [Nuttall, 1981] que a resposta de freqüência das funções de suavização da família *raised-cosine* pode ser obtida pelas fórmulas

$$W\left(f\right)=\frac{Lf}{\pi}\sin\left(\pi Lf\right)\sum_{k=0}^{K}\frac{\left(-1\right)^k a_k}{L^2f^2-k^2}\ ,\ \forall f \tag{6.30}$$

$$W\left(n/L\right)=\begin{cases} a_0\ ,\ \text{se } n=0 \\ \frac{1}{2}a_{|n|}\ ,\ \text{c.c.}\end{cases} \tag{6.31}$$

onde L denota a duração da janela. Uma função de suavização $w\left(t\right)$ pertence à família *raised-cosine* se existirem constantes reais $\left\{a_k\right\}_0^K$ tais que

$$w\left(t\right)=\frac{1}{L}\sum_{k=0}^{K}a_k\cos\left(2\pi kt/L\right)\ \text{para}\ \left|t\right|\leq L/2 \tag{6.32}$$

É importante observar que a família *raised-cosine* engloba muitas das principais funções de suavização utilizadas na análise de sinais musicais, tais como Hann, Hamming e Blackman.

6.6 Desempenho dos Estimadores

Mototsugu Abe e Julius O. Smith III realizaram estudos minuciosos sobre as propriedades do estimador parabólico, investigando seu comportamento em circunstâncias adversas tais como presença de ruído, modulação em freqüência ou amplitude e interferência entre componentes. Em [Abe e Smith III, 2004] pode ser encontrado um compêndio que estabelece, teórica e experimentalmente, as condições necessárias (por exemplo, fator mínimo de "zero-padding") para que haja a garantia de determinada qualidade da estimativa, i.e., para que o erro máximo esteja confinado a uma faixa pré-estabelecida.

6.6.1 Eficiência e aplicabilidade

A Tabela 6.1 lista as principais técnicas de estimação de parciais aplicáveis a sinais musicais e algumas de suas características.

Tabela 6.1: Estimadores de parciais e suas características.

Técnica	Custo mínimo por quadro	Custo adicional por pico	Função de suavização
Superamostragem de fator m	$O(mN \lg mN)$	$O(m)$	Qualquer
Trigonométrico	$O(N \lg N)$	$O(1)$	Qualquer
Reatribuição	$O(N \lg N)$	$O(1)$	Qualquer*
Phase-Vocoder	$O(N \lg N)$	$O(1)$	Qualquer
Derivada	$O(N \lg N)$	$O(1)$	Qualquer
Interpolação de Grandke	$O(1)$	$O(1)$	Hann
Interpolação Parabólica	$O(1)$	$O(1)$	Melhor se específica

Amostras externas ao quadro em análise

Os estimadores da derivada, trigonométrico e phase-vocoder requerem uma amostra fora do quadro cujos parciais se deseja estimar (no caso do phase-vocoder mais de uma amostra pode ser necessária). Não obstante isso represente uma desvantagem potencial, é possível contornar essa limitação. Basta considerar, em lugar de um quadro com N amostras, dois quadros com $N-1$ amostras: o primeiro com as amostras $s[1], \dots, s[N-1]$ e o segundo com as amostras $s[2], \dots, s[N]$.

*A rigor apenas funções de suavização cuja derivada pode ser obtida analiticamente. Na literatura não foram encontrados relatos sobre o uso de métodos aproximativos.

Neste ponto poder-se-ia objetar que a redução no tamanho do quadro poderia prejudicar a capacidade de discriminação em freqüência. No entanto, como tipicamente $N \geq 64$, subtrair o quadro de uma amostra não afeta apreciavelmente a resolução de freqüência da análise. Outra objeção possível seria a de que, ao ser reduzido, o tamanho do quadro perderia a propriedade de ser produto de pequenas potência de primos e, com isso, não poderia mais ter sua transformação calculada pela FFT. Contudo basta usar uma amostra de *zero-padding* para que o tamanho total do quadro, i.e. o tamanho aparente para a FFT, seja preservado.

Apesar destas considerações, o uso de uma amostra externa ao quadro não impôs dificuldades adicionais ao nosso sistema. Isto se deve ao fato de seu *buffer* ter sido projetado para lidar com métodos de detecção de início de nota que exigem não apenas uma, mas muitas amostras anteriores ao quadro em análise.

6.6.2 Eficácia

Estudos comparativos que investigam o erro médio, a variância e o viés dos principais estimadores de parciais podem ser encontrados em [Keiler e Marchand, 2002], [Hainsworth e Macleod, 2003a] e [Hofbauer, 2004].

Conforme demonstrado em [Hainsworth, 2004, §3.4.3], a reatribuição espectral é um estimador viesado. No entanto, como a magnitude do viés depende do tamanho e da forma da janela, é possível forjar circunstâncias em que o viés é controladamente pequeno.

Influência da função de suavização

Muitos testes foram realizados pelo autor a fim de comparar os estimadores, tanto com sinais artificiais, quanto com gravações acústicas. Nestes testes experimentaram-se outras funções de suavização que não as de Hann, Hamming e Blackman. Estudos sobre funções de suavização podem ser encontrados em [Harris, 1978] e [Nuttall, 1981].

A função de suavização apresentada em [Nuttall, 1981, Fig. 10] se mostrou particularmente apropriada ao método trigonométrico. Obtivemos, com esta combinação, resultados muito superiores aos publicados em [Lagrange et al., 2005], bastando que a janela tivesse tamanho suficiente para que $\Delta f_{\min} \geq 6\mathcal{B}$, onde $\Delta f_{\min} = \min_{i<j} \{|f_i - f_j|\}$ e $\mathcal{B} = F_s/N$ Hz. Com a janela Blackman-Harris também obtivemos resultados superiores aos publicados quando $\Delta f_{\min} \geq 4\mathcal{B}$.

Estes resultados surpreendentes e positivos indicam ser pertinente a realização de um estudo sistemático sobre a influência da função de suavização no desempenho dos estimadores, em particular com os métodos de reatribuição espectral e trigonométrico. O autor do presente trabalho não tem conhecimento da publicação de nenhum estudo com este propósito. Estudos clássicos sobre funções de suavização, por exemplo , podem ser utilizados como ponto de partida

Sumário

Levando em conta os estudos comparativos e os resultados de testes realizados pelo próprio autor é possível resumir as observações acerca da eficácia dos estimadores da seguinte forma:

- O método trigonométrico é, no geral, o melhor estimador. Para $\Delta f_{\min} \in [2\mathcal{B}, 4\mathcal{B})$, a janela de Hann é a mais adequada. Para $\Delta f_{\min} \in [4\mathcal{B}, 6\mathcal{B})$ a função de Blackman-Harris produz melhores resultados. Por fim, se $\Delta f_{\min} \geq 6\mathcal{B}$, a função mais indicada é a apresentada em [Nuttall, 1981, Fig. 10].
- O segundo melhor estimador é o método da reatribuição espectral.
- O algoritmo de Grandke é o melhor dos estimadores baseados no espectro de magnitude e deve ser usado em situações que requeiram economia de recursos computacionais.

6.7 Exemplo de Resultado

Como se pode ver na Figura 6.3, a quantização do STDFT torna as trajetórias dos parciais descontínuas, dificultando por exemplo sua identificação visual. Já na Figura 6.4 os picos espectrais foram estimados com precisão muito maior, de modo que, mesmo havendo algum ruído, é fácil identificar visualmente as trajetórias dos parciais devido a seu caráter contínuo.

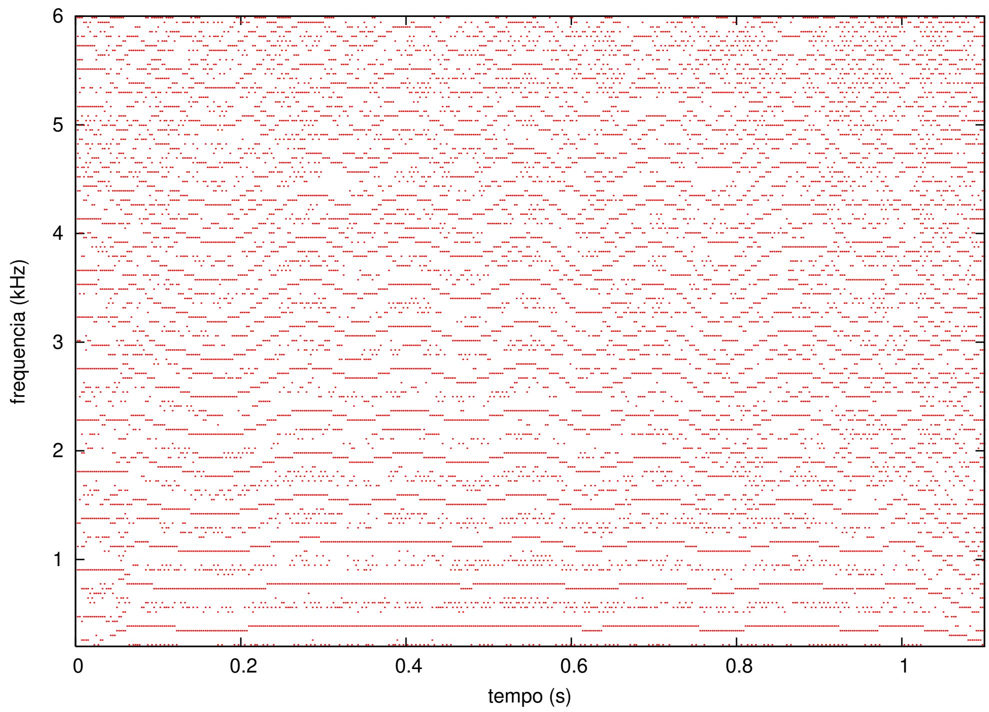


Figura 6.3: Picos espectrais de voz com vibrato tal como fornecidos pela STDFT.

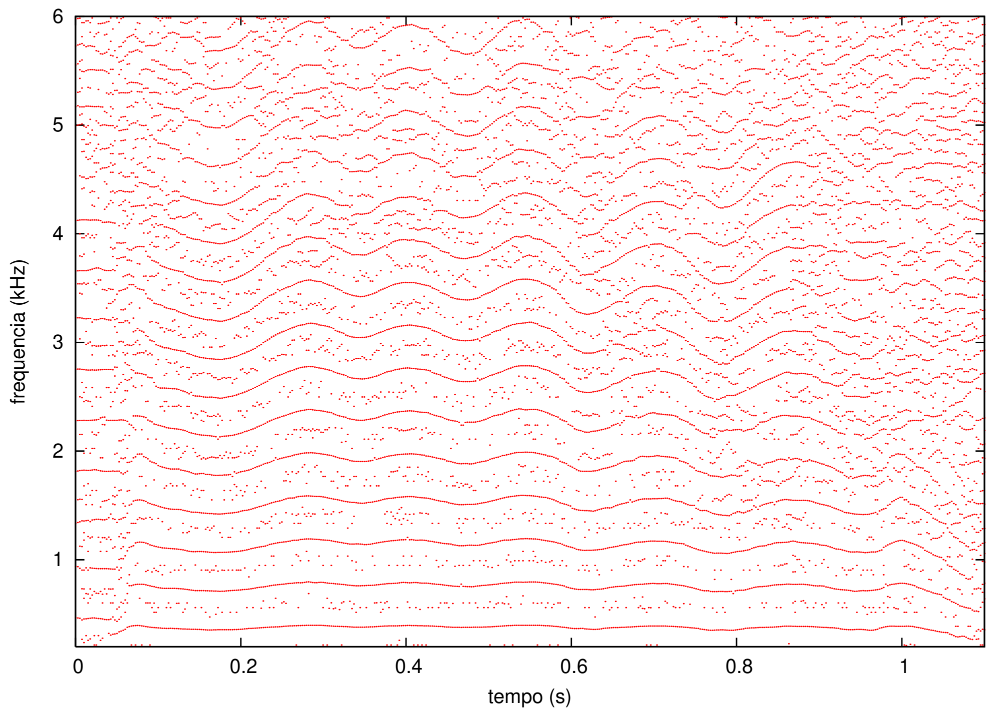


Figura 6.4: Estimativa precisa dos picos espectrais de voz com vibrato.

Capítulo 7

Triagem de Parciais

A quantidade de parciais produzidos pelo módulo anterior é tipicamente elevada, não raro da ordem de centenas. No entanto, nem todos precisam ou devem ser considerados para a estimação de F0. Por conta disso criou-se este módulo, cuja finalidade é filtrar parciais irrelevantes e picos espúrios, i.e., produzidos por ruído. O efeito deste processo pode ser apreciado no exemplo da Figura 3.

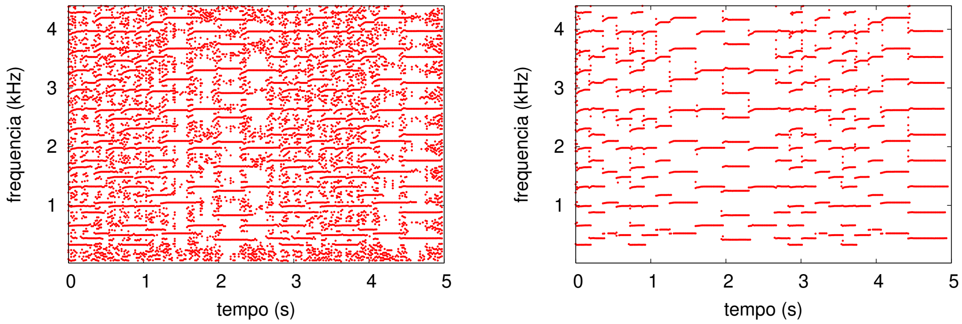


Figura 7.1: Exemplo do efeito da triagem de parciais.

Denotaremos por (f_i, a_i) o i-ésimo parcial, com freqüência f_i Hertz e magnitude a_i decibéis. O conjunto de todos parciais será representado por Q. Por fim, $P \subseteq Q$ denotará o subconjunto dos parciais principais, obtidos após a triagem.

A triagem se dará em dois passos. O primeiro critério exige que o parcial tenha magnitude igual ou superior a $\alpha_{\min}$ e não inferior à do maior pico do quadro em mais do que $\Delta_{\max}$. Além disso, exige-se que tenha freqüência entre $\phi_{\min}$ e $\phi_{\max}$. Os valores para $\alpha_{\min}$, $\Delta_{\max}$, $\phi_{\min}$ e $\phi_{\max}$ são fornecidos pelo usuário. Segue a formalização desse critério.

$$\text{Se}\,(f_i, a_i) \in P, \text{ então} \begin{cases} f_i \in [\phi_{\min}, \phi_{\max}] \\ a_i \geq \alpha_{\min} \\ a_\kappa - a_i \leq \Delta_{\max} \end{cases} \quad \text{onde} \quad \kappa = \min\{i : a_i \geq a_j, \forall (f_j, a_j) \in Q\}$$

O segundo critério de filtragem foi baseado na observação de que a maior parte dos instrumentos produz séries harmônicas com envoltórias aproximadamente contínuas, ou seja, séries em que a diferença de magnitude entre harmônicos consecutivos é pequena. Denotando por $\sigma_{\max}$ a variação máxima permitida entre harmônicos sucessivos, segue a formalização deste critério.

$$\text{Se}\,(f_i, a_i) \in P, \text{ então} \begin{cases} i > \kappa \Rightarrow a_i + \sigma_{\max} \geq a_{i^-} \\ i < \kappa \Rightarrow a_i + \sigma_{\max} \geq a_{i^+} \end{cases}$$

$$\textit{onde} \quad i^- = \max\{j < i : (f_j, a_j) \in P\} \quad \textit{e} \quad i^+ = \min\{j > i : (f_j, a_j) \in P\}$$

Capítulo 8

Estimação de F0

O método que será proposto parte da suposição de que o parcial de máxima magnitude pertence à série harmônica principal, ou seja, é um múltiplo da freqüência fundamental. Denotando por f_κ a freqüência do parcial de máxima magnitude, nosso conjunto de candidatos à freqüência fundamental é, então, dado por

$$C = \left\{ c_n \stackrel{def}{=} \frac{f_\kappa}{n} : 1 \leq n \leq \left\lfloor \frac{f_\kappa}{\mathrm{F0}_{\min}} \right\rfloor \right\} \tag{8.1}$$

É preciso, neste ponto, estabelecer formas de comparar essas dezenas, eventualmente centenas de candidatos. De acordo com o princípio da seletividade harmônica, o primeiro passo neste sentido é coletar as séries harmônicas correspondentes a cada candidato. Denotando com uma seta para a esquerda a operação de atribuição (por exemplo, $x \leftarrow 7$ significa a variável x recebe o valor 7), apresentamos um algoritmo que cumpre a tarefa.

```
SÉRIE-HARMÔNICA(f_cand, P)
for each (f, a) ∈ P
      do i ← ⌊f/f_cand + 0.5⌋
         d ← DESVIO-RELATIVO(f_cand · i, H[i]_freq)
         d̂ ← DESVIO-RELATIVO(f_cand · i, f)
         if d̂ ≤ min{d, ²⁴√2}
            then if d̂ < d or a > H[i]_mag
                    then H[i]_freq ← f
                         H[i]_mag ← a
return H
```

Se o i-ésimo harmônico do n-ésimo candidato à F0 estiver presente no espectro, sua magnitude e freqüência estarão, ao fim da execução do algoritmo, em $H[n][i]_{mag}$ e $H[n][i]_{freq}$. Caso contrário valerá que $H[n][i]_{mag} = H[n][i]_{freq} = 0$.

Faz-se necessário, então, um modo de quantificar a proeminência de cada candidato em função de sua série harmônica. Isto será feito levando em conta princípios psicoacústicos, em particular o conceito de banda crítica [Roederer, 2002, §3.4].

As funções Φ e Ψ, definidas a seguir, são adaptações do modelo de soma harmônica apresentado em [Klapuri, 2004, §6.3.3]. A motivação psicoacústica das fórmulas pode ser encontrada na referência e não será reproduzida aqui.

Formalmente, a proeminência do n-ésimo candidato será dada por

$$\Phi\left(n\right) = \sum_{i=1}^{I(n)} H[n][i]_{mag} \cdot \Psi\left(i\right) \tag{8.2}$$

$$onde \quad I(n) = \max\left\{j \in \{1, \ldots, |H|\} : H[n][j]_{mag} > 0\right\} \tag{8.3}$$

e $\Psi\left(i\right)$ denota a fração de banda crítica que corresponde a cada harmônico em função de seu índice, calculada como

$$\Psi\left(i\right) = \begin{cases} 1, \text{ se } \; i = 1 \\ \min\left\{\log_{2^{1/3}}\left(i \cdot \sqrt{\frac{i+1}{i}}\right) - \log_{2^{1/3}}\left((i-1) \cdot \sqrt{\frac{i}{(i-1)}}\right), 1\right\}, \text{ c.c.} \end{cases} \tag{8.4}$$

De posse da proeminência dos candidatos, a determinação de F0 será feita em dois estágios. Primeiramente, os candidato com proeminência relativa à máxima de pelo menos $\beta \in [0, 1]$, pré-fixado, são selecionados. Dentre estes, aquele que possuir a maior média ponderada χ, definida a seguir, é eleito F0. Caso haja empate, vence o candidato de menor freqüência. Formalmente, será considerado F0 o candidato de índice φ dado pelas seguintes equações

$$\varphi = \min\left\{n : c_n \in C^{\Phi} \quad e \quad \chi(n) \geq \chi(m), \forall c_m \in C^{\Phi}\right\} \tag{8.5}$$

$$\chi(n) = \frac{\sum\limits_{i=1}^{I(n)} H[n][i]_{mag} \cdot \Psi(i)}{\sum\limits_{i=1}^{I(n)} \Psi(i)} \tag{8.6}$$

$$C^{\Phi} = \left\{c_n \in C : \Phi(n) \geq \beta \cdot \max_{m|c_m \in C}\left\{\Phi(m)\right\}\right\} \tag{8.7}$$

Com isso, já se tem uma estimativa da freqüência fundamental. No entanto, seu valor exato baseou-se na estimativa de freqüência de um único parcial, o de máxima magnitude. Ainda que este tenha a estimativa isoladamente mais confiável, um estimador de F0 ainda mais robusto pode ser obtido combinando-se as estimativas dos diversos harmônicos.

Como os estimadores de parciais são essencialmente não-viesados, espera-se que os erros individuais se anulem ao serem acumulados. Sabe-se que a qualidade da estimativa de cada parcial depende, dentre outros fatores, de sua razão sinal/ruído (abreviada em inglês por SNR) e de quão estável encontra-se sua freqüência absoluta. Por conta disso, devem ser privilegiados os harmônicos de maior magnitude, que possuem maior SNR, e de menor índice, uma vez que variações na freqüência fundamental são multiplicadas pelo índice do harmônico. Considerando estes princípios, desenvolvemos uma fórmula para re-estimação da freqüência fundamental:

$$\mathrm{F0}' = \frac{\sum_{i=1}^{I(n)} H[i]_{freq}/i \cdot H\left[i\right]_{mag} \cdot \Psi\left(i\right)}{\sum_{i=1}^{I(n)} H\left[i\right]_{mag} \cdot \Psi\left(i\right)} \tag{8.8}$$

onde $H[i]$ denota o i-ésimo parcial da série harmônica de c_φ, ou seja, $H[i] \stackrel{def}{=} H[\varphi][i]$.

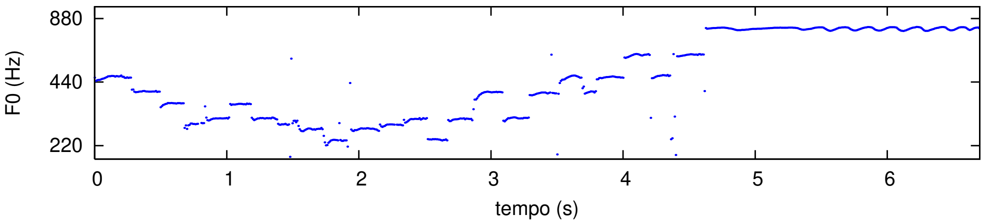


Figura 8.1: Exemplo de saída do módulo de estimação de F0.

Capítulo 9

Detecção de Início de Notas

9.1 Introdução

Nosso sistema de transcrição foi concebido como um conjunto de módulos independentes e desacoplados, visando ampliar a liberdade de experimentação e facilitar a integração com outros sistemas. Por conta disso, o módulo de Produção de Notas foi projetado de forma a requerer apenas a freqüência fundamental estimada em cada instante, como é feito em grande parte dos sistemas descritos na literatura. No entanto, fornecer-lhe adicionalmente uma lista com os instantes precisos de início de cada nota é bastante desejável, como ficará claro a seguir.

Características físicas dos instrumentos musicais fazem com que haja um intervalo apreciável desde o instante em que lhes é dado o impulso inicial — quando, por exemplo, a corda de um violão é dedilhada — até que adquiram comportamento estável e periódico. A percepção de início de nota contudo, está associada ao instante em que o instrumento começa a soar, mesmo antes de possuir característica tonal bem definida. Por conta disso, associar o ínicio de cada nota ao primeiro instante em que sua F0 foi detectada é sujeitar-se a produzir notas que começam sistematicamente atrasadas em alguns centésimos de segundo. Como não há razão para que o final das notas seja também atrasado, o resultado é a produção de notas encurtadas e ritmicamente imprecisas. Por conta disso torna-se interessante empregar técnicas *ad hoc* para determinar com precisão o início de cada nota.

A evolução de uma nota musical típica pode ser descrita, na maior parte das entradas naturais, pelo modelo evolutivo ADSR (Figura 9.1). Este modelo consiste em uma sucessão de quatro estágios: ataque (Attack), amortecimento (Decay), sustentação (Sustain) e decaimento (Release). A inspeção de cada um destes estágios é muito importante para a construção de um detector de ínicio de nota (Tabela 9.1). Uma discussão mais extensa sobre as características de início de notas pode ser encontrada em [Bello, 2003, §4.1].

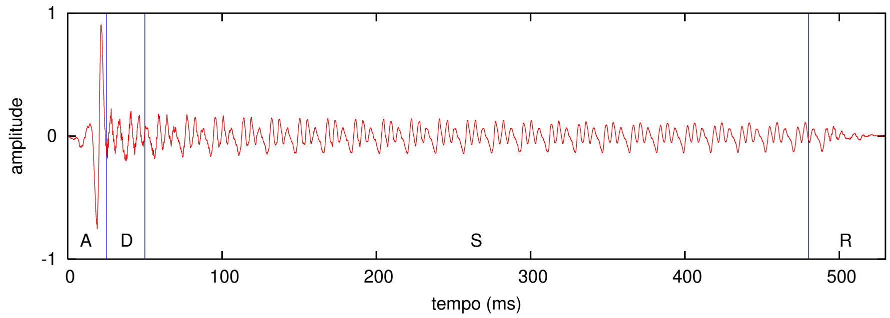


Figura 9.1: Exemplo de nota que segue o envelope ADSR.

Estágio	Energia	Caráter
Ataque	Cresce rapidamente	Transiente
Amortecimento	Decresce	Em estabilização
Sustentação	Permanece constante ou decresce lentamente	Periódico (som harmônico)
Decaimento	Decresce rapidamente	Som harmônico

Tabela 9.1: Características do Modelo Evolutivo ADSR.

9.2 Estratégia de Detecção

Quase todos os métodos de detecção de início de notas podem ser esquematizadas como ilustrado na Figura 9.2. A estratégia consiste nos seguintes passos.

Primeiramente, produz-se, a partir do sinal de áudio, uma *função de detecção*. Esta função deve conter informação sobre alguma característica importante para a detecção de início, por exemplo o perfil energético do sinal. Tipicamente a função de detecção é amostrada com taxa muito inferior à do sinal original, usualmente entre 20 e 100 Hz.

Em seguida, os picos

A Figura 9.2 representa graficamente esta estratégia.

Nas seções a seguir discutiremos em detalhe as operações de redução e coleta de picos indicadas na Figura 9.2, bem como a escolha de funções de detecção adequadas.

9.3 Funções de Detecção

Chamamos de função de detecção a uma função que transforma o sinal de áudio em um outro sinal, com freqüência de amostragem muito inferior, no qual alguma

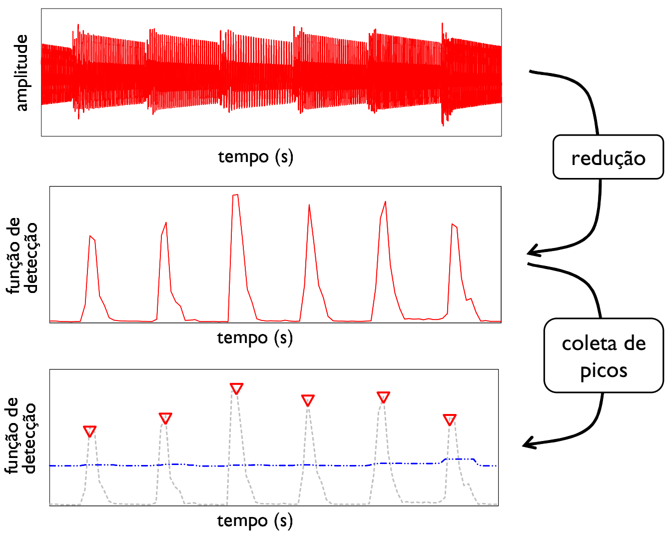


Figura 9.2: Estratégia usual para detecção de início de notas.

característica facilmente identificável indicará os inícios de eventos do sinal original. Um estudo taxonômico de técnicas de detecção de início de nota pode ser encontrado em [Bello et al., 2005]. A eficácia das principais técnicas foi avaliada e comparada em [Collins, 2004].

9.3.1 Funções Baseadas na Energia

Funções de detecção que buscam evidenciar crescimentos bruscos de energia classificam-se como baseadas na energia. As funções deste tipo funcionam especialmente bem para detectar início de notas de instrumentos percussivos, por exemplo piano ou tamborim.

Seja $S_k(m)$ o k-ésimo escaninho, obtido pela STDFT em um quadro de $N = 2K$ amostras centrado no instante mhF_s^{-1}, onde h é o salto (em amostras) entre quadros sucessivos, e seja w_k o k-ésimo coeficiente da função de suavização, então definimos a energia ponderada pela freqüência como

$$\overline{E}_N^w(m) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} w_k \cdot |S_k(m)|^2} \tag{9.1}$$

Note que $\overline{E}_N^w(m)$ é proporcional à energia RMSe um quadro com $2K$ amostras centrado na amostra m do sinal original processado por um filtro com resposta de freqüência, em potência, dada por w. Não é difícil demonstrar esta afirmação utilizando o Teorema de Parseval. Além disso, se para todo k valer que $w_k = 1$, então $\overline{E}_N^w(m)$ se reduz à energia RMS do sinal original.

Por fim, obtém-se a função de detecção por meio de uma aproximação pela esquerda da derivada da energia ponderada pela freqüência, ou seja,

$$D(n) = \overline{E}(n) - \overline{E}(n-1) \tag{9.2}$$

Em lugar de analisar o sinal como um todo, é possível realizar o processo de análise da energia por banda de freqüência. Nesta abordagem, empreendida em [Klapuri, 1999], o sinal é subdividido em 21 bandas, que são processadas individualmente para. Ao final, o resultado das diversas bandas é combinado para dar origem à função de detecção. Além da divisão em bandas, o autor procurou modelar outros processos psicoacústicos envolvidos na percepção de volume, como a influência da freqüência na percepção de volume. Em nosso sistema foi implementada uma simplificação desta técnica, a fim de adaptá-la ao arcabouço do STDFT.

9.3.2 Funções Baseadas na Fase

Funções de detecção que buscam evidenciar variações de freqüência no espectro, em particular no interior de cada escaninho, se classificam como baseadas na fase. As funções deste tipo funcionam especialmente bem para detectar início de notas de instrumentos tonais não percussivos, por exemplo o violino.

Um oscilador harmônico com freqüência e amplitude constantes produz pela STDFT um pico espectral com amplitude constante e fase variando linearmente com o tempo. Quando a freqüência do oscilador varia, a evolução da fase do escaninho associado deixa de ser linear. Assim, a segunda derivada da fase em relação ao tempo pode ser usada para identificar variações de freqüência: quanto mais distante de zero for seu valor, maior terá sido a variação.

Seja $\tilde{\varphi}(m)$ a fase acumulada de um certo escaninho do espectro obtida pela STDFT a partir de uma janela centrada no instante mhF_s^{-1}, então seu desvio de fase é dado por

$$d_\varphi(m,k) = princarg\left[\tilde{\varphi}_k(m) - 2\tilde{\varphi}_k(m-1) + \tilde{\varphi}_k(m-2)\right] \tag{9.3}$$

onde

$$princarg(\varphi) = \begin{cases} \varphi, \; se \; 0 \le \varphi < 2\pi \\ princarg\left[\varphi - 2\pi \cdot sign(\varphi)\right], \; c.c. \end{cases}$$

com

$$sign(x) = x/|x|$$

Uma função de detecção simples pode ser construída simplesmente somando-se o desvios de todos os escaninhos.

$$\eta(m) = \frac{1}{K}\sum_{k=1}^{K} |d_\varphi(m,k)| \tag{9.4}$$

Funções de detecção mais sofisticadas, baseadas na dispersão estatística dos desvios de fase, foram empregadas em [Bello e Sandler, 2003].

9.3.3 Funções Baseadas no Conteúdo Harmônico

Em [Hainsworth e Macleod, 2003b] é apresentada uma metodologia para detectar mudanças no conteúdo harmônico. O autor faz uso, dentre outras métricas, da distância de Kullback-Liebler modificada:

$$D_{MKL}(n) = \sum_{k=1}^{K} \max\left\{\log_2\left(\frac{|S_k(m)|}{|S_k(m-1)|}\right), 0\right\} \tag{9.5}$$

9.3.4 Funções Híbridas

Abordagens híbridas foram desenvolvidas no intuito de unir as vantagens das técnicas baseadas na energia aos benefícios das técnicas baseadas na fase. Por um lado, as funções baseadas na energia do sinal permitem excelente acurácia na detecção de notas de instrumentos percussivos, tais como o piano. Por outro lado, as técnicas baseadas na fase são muito mais sensíveis a inícios suaves, marcados por mudanças de freqüência fundamental mas não por crescimento apreciável de energia. Inícios deste tipo são produzidos, por exemplo, por um violino arqueado ininterruptamente durante uma sucessão de notas.

O problema de detecção de início de notas recebeu primeiramente abordagem híbrida em [Duxbury et al., 2002]. Mais recentemente, foi proposta em [Duxbury et al., 2003, Bello et al., 2004] uma funcão de detecção híbrida baseada no domínio dos complexos.

Considere-se um oscilador harmônico de variação lenta ou nula. Como sua amplitude é constante, espera-se que a magnitude do escaninho associado seja a mesma em um quadro e em seu antecessor. Como a freqüência do oscilador é suposta constante e a derivada da fase de uma senóide em relação ao tempo é sua freqüência instantânea, espera-se que a fase do oscilador varie linearmente com o tempo. Logo, observando a fase de um parcial em dois instantes sucessivos é possível determinar sua fase no instante seguinte. Considerando que a fase e a amplitude determinam o espectro complexo de um parcial, é possível então determinar o espectro complexo de um parcial no quadro m a partir da observação de seus valores nos quadros $m-1$ e $m-2$.

Considerando um parcial contido no k-ésimo escaninho, temos que o valor previsto para seu espectro é dado por

$$\hat{S}_k(m) = \hat{A}_k(m)\, e^{i\hat{\phi}_k(m)} \tag{9.6}$$

onde a previsão de amplitude é dada por

$$\hat{A}_k(m) = |S_k(m-1)| \tag{9.7}$$

e a previsão de fase,

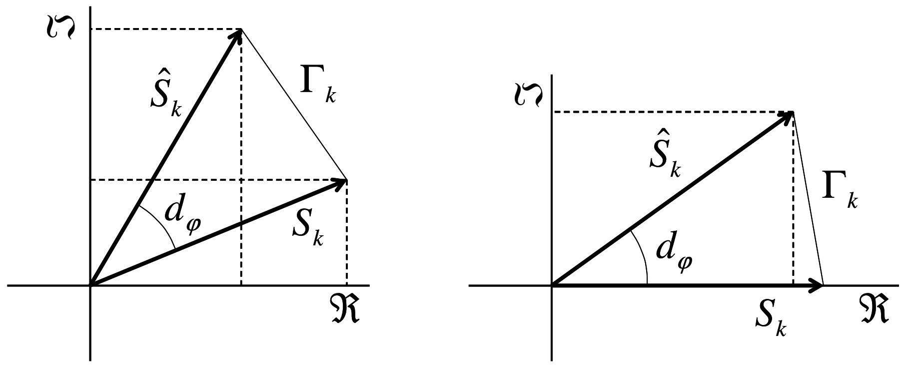
$$\hat{\phi}_k(m) = princarg\left[2\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)\right] \tag{9.8}$$

Medindo-se a "distância" entre o valor previsto e o espectro aferido, é possível definir o grau de imprevisibilidade do k-éstimo escaninho como

$$\Gamma_k(m) = \left\| \hat{S}_k(m) - S_k(m) \right\|_2 \tag{9.9}$$

Esta medida é ilustrada na Figura 9.3.4.

Com a finalidade de tornar o cálculo de Γ_k mais eficiente, a distância entre S_k e $\hat{S}_k$ será calculada pelo diagrama rotacionado (Figura 9.3.4).



(a) diagrama normal (b) diagrama rotacionado

Figura 9.3: Diagrama no domínio dos complexos ilustrando a diferença de fase e a distância euclidiana entre o vetor predito e o vetor aferido de um escaninho do espectro de Fourier

Tendo um dos vetores alinhado axialmente, é possível calcular a distância entre os dois fazendo apenas uma chamada de função trigonométrica, contra três do cálculo direto. Assim, podemos calcular a Equação 9.9 por meio da fórmula

$$\Gamma_k(m) = \left\{ \hat{R}_k(m) + R_k(m) - 2\hat{R}_k(m) R_k(m) \cos(d_{\varphi,k}(m)) \right\}^{1/2} \tag{9.10}$$

Como chamadas a funções trigonométricas são computacionalmente caras, esta otimização tem impacto apreciável no desempenho do sistema de detecção de início de notas, mesmo sem alterar a complexidade do algoritmo.

Por fim, a função de detecção é construída somando-se o grau de imprevisibilidade de todos os escaninhos.

$$\eta(m) = \sum_{k=1}^{K} \Gamma_k(m) \tag{9.11}$$

9.4 Coleta de Picos

De posse de uma função de detecção adequada, espera-se que inícios de nota produzam características identificáveis e bem localizadas na função de detecção.

Tipicamente a característica associada a inícios de notas são máximos locais (i.e., picos), em geral com alguma variabilidade em termos de forma e magnitude e sujeitos a contaminação por ruído, o qual pode advir de ruído no sinal ou de aspectos do sinal não relacionados a inícios de nota, como vibrato. Portanto, um algoritmo robusto de coleta de picos se faz necessário para estimar os instantes de início das notas contidas no sinal.

9.4.1 Pós-Processamento

O pós-processamento é uma etapa opcional que depende do método de redução usado para produzir a função de detecção. O propósito do pós-processamento é facilitar as tarefas de limiarização coleta de picos aumentando a uniformidade e consistência das características relacionadas a eventos na função de detecção, idealmente transformando-as em máximos locais isolados e facilmente detectáveis. Nesta categoria incluem-se processos que visam a reduzir os efeitos de ruído (por exemplo suavização) e processos necessários à seleção bem-sucedida de parâmetros de limiarização para uma ampla gama de sinais (por exemplo normalização e remoção de corrente contínua).

Uma técnica de suavização bastante empregada em funções de detecção é a convolução com uma janela de Hann com largura do lobo central igual ao intervalo mínimo esperado entre dois eventos distintos. Nosso sistema, no entanto, não faz uso de pós-processamento.

9.4.2 Limiarização

Como se vê no exemplo da Figura 9.4, ser máximo local da função de detecção não é suficiente para identificar um início de nota. Formalmente, o conjunto dado por

$$T_0 = \{n : f(n) > \max\{f(n-1), f(n+1)\}\} \tag{9.12}$$

tipicamente contém índices de muitos quadros que não correspondem a início de notas.

Limiarização Estática

Uma melhora significativa pode ser obtida acrescentando uma condição de valor mínimo.

$$T_e = \{n \in T_0 : f(n) > \delta\} \tag{9.13}$$

No entanto, eventos musicais em instantes diferentes de um mesmo arquivo de áudio podem ser muito diferentes entre si. Por exemplo, em uma mesma melodia há passagens mais brandas e outras mais intensas.

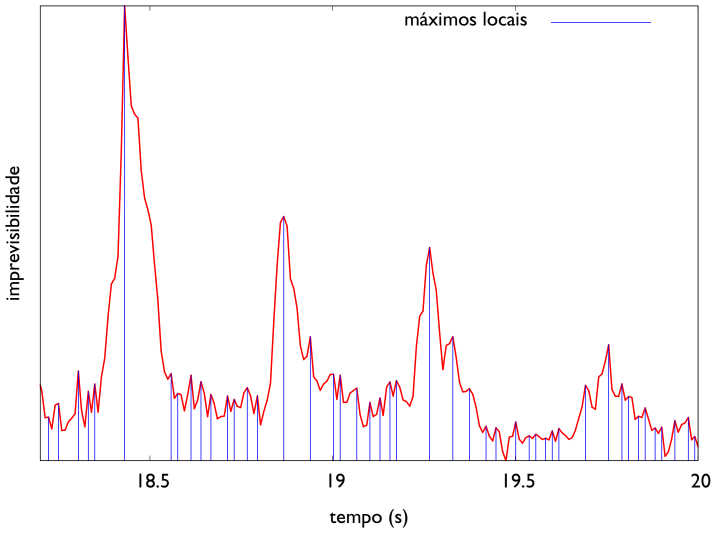


Figura 9.4: Como fica claro no exemplo, não é razoável assumir que todos os máximos locais da função de detecção correspondem a inícios de nota.

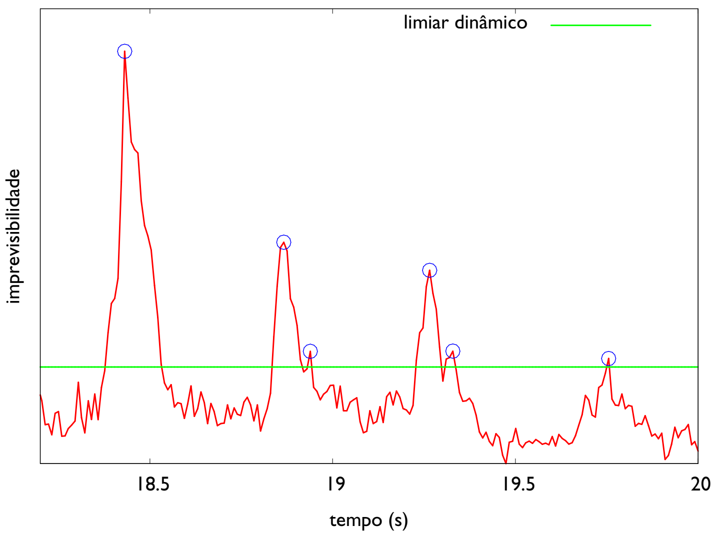


Figura 9.5: O uso de limiarização estática reduz o número de falsos positivos, mas não completamente. A partir de determinado ponto a eliminação de falsos positivos acarreta em falsos negativos.

Limiarização Dinâmica

Solução adequada: limiar adaptativo, dado por uma versão suavizada da própria função de detecção. Há vários filtros de suavização, por exemplo o filtro da mediana (9.4.2) ou curvas de Bézier.

$$T_d = \left\{ n \in T_0 : f(n) > \tilde{\delta}(n) \right\} \tag{9.14}$$

$$\tilde{\delta}(n) = \delta + \lambda \cdot \text{mediana} \{ f(n-M), \ldots, f(n+M) \} \tag{9.15}$$

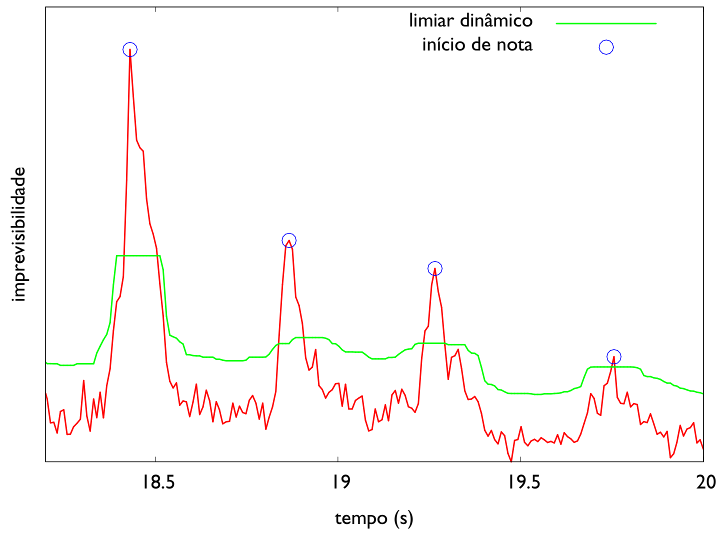


Figura 9.6: Sinais musicais freqüentemente contém eventos com características muito distintas entre si, por exemplo em termos de intensidade sonora. Por conta disso, é desejável que o limiar seja adaptativo.

O filtro da mediana possui algumas propriedades interessantes (vide [Regalia, 1993] para um estudo abrangente), em particular a preservação de seqüências monotônicas e a remoção de impulsos, que o tornam adequado para a limiarização adaptiva desejada.

Denotando por $s_{(m)}$ o m-ésimo menor termo da seqüência $s_1, \ldots, s_n$, definimos

$$\text{mediana} \{ s_1, \ldots, s_{2n+1} \} = s_{(n)} \tag{9.16}$$

Nosso sistema implementa o filtro da mediana de acordo com [Astola e Campbell, 1989], consumindo tempo $O(n \lg m)$, em contraste com $O(nm \lg m)$ da implementação trivial, onde n denota o comprimento do sinal e m, o tamanho da janela

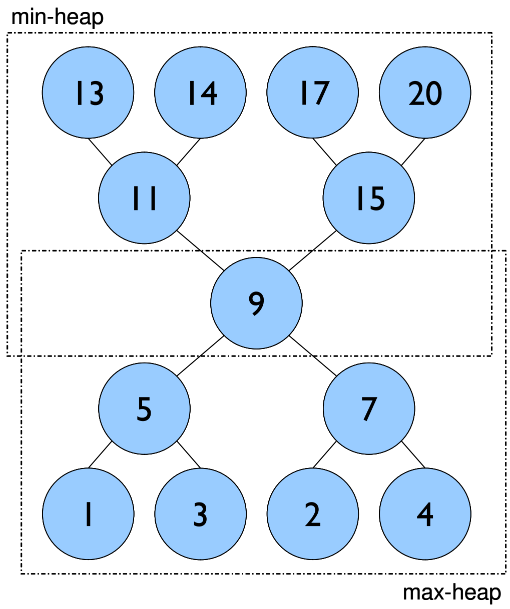


Figura 9.7: Exemplo de heap composto usado na implementação eficiente do filtro da mediana. O nó compartilhado pelos max-heap e min-heap é a mediana.

Capítulo 10

Produção de Notas

Informações referentes ao tempo foram omitidas nas seções 5 a 8, em que se descreviam operações realizáveis individualmente para cada quadro. A produção de notas, no entanto, não faz sentido em regime *por-quadro*. Conseqüentemente reaparecerão, nesta seção, informações temporais.

O funcionamento do módulo é descrito pelo algoritmo *Produz-Notas*, em que $freq[Fund[i]]$ denotará a i-ésima estimativa de F0, associada a um quadro centrado no instante $inst[Fund[i]]$. Além disso, $intens[Fund[i]]$ denotará a proeminência da i-ésima estimativa, calculada pela Equação 8.2. Além de uma seqüência de estimativas de F0 (Figura 8.1), o módulo de produção de notas recebe o conjunto T dos instantes de início de nota.

O atraso máximo na detecção de F0, i.e., o tempo máximo decorrido entre um início de nota e a detecção da F0 que lhe corresponde, será denotado por $\sigma_{\max}$. O maior intervalo aceitável sem constatações de F0 no "interior" de uma nota será representado por $\gamma_{\max}$. Finalmente, $\Delta_{\min}$ denotará a duração da nota mais curta permitida e $\nu_{\min}$, a intensidade mínima aceitável.

Durante o desenvolvimento do módulo de produção de notas, autor teve em mente três objetivos. O primeiro foi evitar que irregularidades na estimação de F0 ocasionem notas espúrias. O segundo foi eliminar o atraso decorrente do fato de as notas muitas vezes não possuírem, em seus estágios iniciais, altura (e, portanto, F0) bem definida. O terceiro foi impedir que uma seqüência de notas repetidas seja interpretada como uma única nota.

```
Produz-Notas(Fund, T)
for i ← 1 to tamanho[Fund]
     do n ← ⌊log_{12√2}(freq[Fund[i]]/440) + 0.5⌋ + 69        ▷ número de nota MIDI
        if ultimaConstatacao[Nota[n]] ≠ INDEFINIDA
          then if inst[Fund[i]] − ultimaConstatacao[Nota[n]] > γ_max
                 then if intens[Nota[n]] > ν_min and duracao[Nota] > Δ_min
                        then Imprime(Nota[n])
                      Reinicia(Nota[n])
        if inicio[Nota[n]] = INDEFINIDO
          then t ← max {T[i] ≤ inst[Fund[i]]}
               if inst[Fund[i]] − t ≤ σ_max
                 then inicio[Nota[n]] ← t
                 else inicio[Nota[n]] ← inst[Fund[i]]
        ultimaConstatacao[Nota[n]] ← inst[Fund[i]]
        intens[Nota[n]] ← max {intesidade[Fund[i]], intens[Nota[n]]}
        duracao[Nota[n]] ← ultimaConstatacao[Nota] − inicio[Nota]
        for m ← 1 to tamanho[Nota]
            do if inst[Fund[i]] − inicio[Nota[m]] > γ_max or i = tamanho[Fund]
                 then if intens[Nota[m]] > ν_min and duracao[Nota[m]] > Δ_min
                        then Imprime(Nota[m])
                      Reinicia(Nota[m])
```

Capítulo 11

Resultados

Como o sistema teve sua implementação concluída recentemente, os experimentos realizados até o momento não foram muito numerosos e visaram a uma análise essencialmente qualitativa. Nesses experimentos constatamos que os parâmetros fornecidos pelo usuário têm influência decisiva sobre os resultados, de modo que a realização de testes estatisticamente relevantes pressupõe sua otimização.

Não obstante, os experimentos apresentaram resultados animadores. Testes realizados com sinais sintéticos foram extremamente bem sucedidos, muitos dos quais resultaram em transcrições perfeitas. As melodias foram sintetizadas com diversos timbres, indicando que as hipóteses adotadas pelo sistema não estão vinculadas a peculiaridades de determinada família de instrumentos.

Previsivelmente, os testes com gravações acústicas reais foram mais desafiadores. Recursos expressivos, por exemplo inflexões, não raro levam a cenários cuja desambiguação depende de conhecimentos musicológicos. Considerando que o sistema baseou-se apenas em técnicas de processamento de sinais, seu desempenho em tais circunstâncias mostrou-se bastante satisfatório.

Constatamos ainda que a presença de efeitos, naturais ou artificiais, degrada o resultado das transcrições. Em gravações com forte reverberação ou eco, há situações de polifonia artificial, em que uma nota soa simultaneamente com rastros das que a antecederam.

O sistema foi compilado e executado com igual sucesso em ambientes Linux e Windows. Além disso, foi capaz de produzir transcrições em tempo inferior à duração das gravações, demonstrando potencial para aplicações de tempo-real.

Infelizmente, não há uma base de dados universalmente aceita para testar sistemas de transcrição automáticos, o que dificulta a comparação com outras soluções encontradas na literatura. Apesar disso, pretendemos realizar testes sistemáticos em um corpo representativo de gravações acústicas e sinais sintetizados, divulgando os resultados tão logo estejam disponíveis.

Capítulo 12

Discussão e Perspectivas Futuras

12.1 Discussão dos resultados

As transcrições produzidas foram, em geral, satisfatórias. Merece destaque o desempenho do sistema em situações tipicamente consideradas difíceis, como notas graves ou muito curtas.

A utilização de técnicas de ponta para estimação de parciais teve importância central no sucesso da metodologia proposta. O emprego destas técnicas permitiu ao sistema trabalhar com janelas mais curtas, o que resultou em uma melhor resolução temporal e, por consegüinte, em uma reduzida carga computacional.

É importante salientar, no entanto, que, não fosse a nova heurística para determinação de F0, não teria sido possível tirar proveito destas técnicas. Além disso, a eficácia no reconhecimento de F0 foi determinante para o bom desempenho do sistema.

As heurísticas para detecção e triagem de parciais, apesar de produzirem gráficos mais "limpos", de maneira geral não foram capazes de aumentar a taxa de acerto da estimação de F0. Na maior parte dos testes usou-se apenas filtragem mínima, i.e., filtragem de parciais com magnitude negativa ou freqüência fora do espectro audível.

A desvinculação entre estimação de F0 e detecção de início de nota mostrou-se desejável por duas razões. Por um lado, esta independência foi responsável pela robustez do sistema em circunstâncias em que o detector de início de notas apresentou comportamento irregular. Por outro lado, a utilização de técnicas específicas para detecção de início de notas resultou em notas melhor localizadas no tempo, o que em última análise se traduz em transcrições ritmicamente mais fiéis. A diferença é nítida.

12.2 Limitações e Perspetivas futuras

Apesar do bom desempenho, o sistema apresenta algumas limitações. Uma destas limitações é o ajuste manual de seus diversos parâmetros. Ainda que possua valores padrão para a maioria dos parâmetros, o tamanho do quadro utilizado na análise espectral tem de ser fornecido pelo usuário.

Uma melhoria que poderia ser feita ao sistema, portanto, é a incorporação de um método que determinasse de maneira automática e dinâmica o tamanho "ideal" do quadro para análise.

Além disso, os valores padrão foram determinados empiricamente. A otimização automática dos parâmetros poderia aumentar significamente a acurácia do sistema. Seria interessante um estudo neste sentido.

Por ser modular, o sistema se presta tanto a aplicações típicas de sistemas de transcrição, por exemplo *query-by-humming*, quanto à integração em sistemas maiores, por exemplo de análise-manipulação-resíntese.

Graças à modularidade do sistema, um desenvolvedor interessado terá facilidade em acrescentar novas técnicas de refinamento espectral, heurísticas para determinação de fundamental, interface gráfica, etc.

Capítulo 13

Conclusão

Tendo como base os resultados do presente estudo, foi possível concluir que a metodologia adotada é bastante adequada à tarefa da transcrição melódica. Os resultados produzidos foram em geral muito bons, com destaque para situações tipicamente consideradas difíceis.

Por fim, testes realizados com versões demonstrativas de aplicativos comerciais, notadamente Solo Explorer* e TS-AudioToMIDI†, sugerem superioridade do sistema apresentado em diversos aspectos. O sistema proposto freqüente e consistentemente produziu resultados melhores, em termos tonais e rítmicos.

Adicionalmente, o sistema pode fornecer resultados de análise espectral, estimativas de parciais e de freqüência fundamental, entre outros, em formato textual próprio para ferramentas de impressão de gráficos ponto-a-ponto, tal como o *gnuplot*. Com isso, torna-se útil na produção de, por exemplo, sonogramas e F0-gramas. Em particular, os gráficos do presente trabalho foram produzidos desta forma.

O sistema está disponível em www.mitre.com.br/asymut/, onde se encontram também exemplos em que se pode comparar gravações originais com versões re-sintetizadas a partir das transcrições obtidas pelo sistema.

*disponível em www.recognisoft.com

†idem, http://audioto.com/

Referências Bibliográficas

[Abe e Smith III, 2004] Abe, M. e Smith III, J. O. (2004). Design Criteria for Simple Sinusoidal Parameter Estimation based on Quadratic Interpolation of FFT Magnitude Peaks. In *Proc. of Audio Engineering Society 117th Convention*, San Francisco, USA. [cited at p. 29]

[Astola e Campbell, 1989] Astola, J. e Campbell, T. G. (1989). On Computation of the Running Median. In *IEEE Transactions on Acoustics, Speech and Signal Processing - VOL. 37 - NO. 4*, páginas 572–574. [cited at p. 47]

[Auger e Flandrin, 1995] Auger, F. e Flandrin, P. (1995). Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089. [cited at p. 26]

[Bello, 2003] Bello, J. P. (2003). *Towards the Automated Analysis of Simple Polyphonic Music: A Knowledge-based Approach.* tese de PhD, Department of Electronic Engineering, Queen Mary, University of London. [cited at p. 40]

[Bello et al., 2005] Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., e Sandler, M. B. (2005). A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13(5.2):1035–1047. [cited at p. 41]

[Bello et al., 2004] Bello, J. P., Duxbury, C., Davies, M., e Sandler, M. (2004). On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain. *IEEE Signal Procesing Letters*, 11(6):553–556. [cited at p. 43]

[Bello e Sandler, 2003] Bello, J. P. e Sandler, M. (2003). Phase-based note onset detection for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong. [cited at p. 42]

[Collins, 2004] Collins, N. (2004). A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions. In *Proc. of Audio Engineering Society 118th Convention*, Barcelona, Spain. [cited at p. 41]

[Desainte-Catherine e Marchand, 1998] Desainte-Catherine, M. e Marchand, S. (1998). High Precision Fourier Analysis of Sounds Using Signal Derivatives. Research Report 120498, LaBRI, Université Bordeaux. [cited at p. 23]

[Desainte-Catherine e Marchand, 2000] Desainte-Catherine, M. e Marchand, S. (2000). High Precision Fourier Analysis of Sounds Using Signal Derivatives. *Journal of the Audio Engineering Society*, 48(7/8):654–667. [cited at p. 8, 23]

[Duxbury et al., 2003] Duxbury, C., Bello, J. P., Davies, M., e Sandler, M. (2003). Complex Domain Onset Detection for Musical Signals. In *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03)*, Londres, Reino Unido. [cited at p. 43]

[Duxbury et al., 2002] Duxbury, C., Sandler, M., e Davies, M. (2002). A Hybrid Approach to Musical Note Onset Detection. In *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany. [cited at p. 43]

[Grandke, 1983] Grandke, T. (1983). Interpolation algorithms for discrete Fourier transforms of weighted signals. *IEEE Trans. Instr. and Meas.*, 32(2):350–355. 1983. [cited at p. 22]

[Hainsworth, 2004] Hainsworth, S. (2004). *Techniques for the Automated Analysis of Musical Audio.* tese de PhD, Cambridge University Engineering Department, Cambridge, UK. [cited at p. 26, 30]

[Hainsworth e Macleod, 2003a] Hainsworth, S. e Macleod, M. (2003a). On Sinusoidal Parameter Estimation. In *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03)*, Londres, Reino Unido. [cited at p. 30]

[Hainsworth e Macleod, 2003b] Hainsworth, S. e Macleod, M. (2003b). Onset Detection in Musical Signals. In *Proceedings of the International Computer Music Conference*, Singapore. [cited at p. 43]

[Harris, 1978] Harris, F. J. (1978). On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceedings of the IEEE*, 66(1). [cited at p. 30]

[Hofbauer, 2004] Hofbauer, K. (2004). Estimating Frequency and Amplitude of Sinusoids in Harmonic Signals - A Survey and the Use of Shifted Fourier Transforms. Graz University of Technology - Graz University of Music and Dramatic Arts. Diploma Thesis. [cited at p. 30]

[Järveläinen et al., 2000] Järveläinen, H., Verma, T., e Välimäki, V. (2000). The effect of inharmonicity on pitch in string instrument sounds. In *Proc. International Computer Music Conference*, Berlin, Germany. [cited at p. 9]

[Keiler e Marchand, 2002] Keiler, F. e Marchand, S. (2002). Survey On Extraction of Sinusoids in Stationary Sounds. In *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburgo, Alemanha. [cited at p. 30]

[Kernighan e Pike, 2000] Kernighan, B. W. e Pike, R. (2000). *A Prática da Programação.* Editora Campus, primeira edição. [cited at p. 13]

[Klapuri, 1999] Klapuri, A. (1999). Sound Onset Detection by Applying Psychoacoustic Knowledge. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 3089–3092, Phoenix, Arizona. [cited at p. 42]

[Klapuri, 2004] Klapuri, A. (2004). *Signal Processing Methods for the Automatic Transcription of Music.* tese de PhD, Tampere University of Technology. [cited at p. 36]

[Kodera et al., 1978] Kodera, K., Gendrin, R., e de Villedary, C. (1978). Analysis of time-varying signals with small BT values. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):64–76. [cited at p. 26]

[Lagrange et al., 2003] Lagrange, M., Marchand, S., Raspaud, M., e Rault, J.-B. (2003). Enhanced Partial Tracking Using Linear Prediction. In *Proc. of the 6th Int. Conference on Digital Audio Effects(DAFx-03)*, Londres, Reino Unido. [cited at p. 17]

[Lagrange et al., 2005] Lagrange, M., Marchand, S., e Rault, J.-B. (2005). Improving Sinusoidal Frequency Estimation Using a Trigonometrical Approach. In *Proceedings of the 8th Int. Conference on Digital Audio Effects (DAFx'05)*, Madrid, Spain. [cited at p. 24, 30]

[López e Rocamora, 2005] López, E. e Rocamora, M. (2005). Tararira: Sistema de búsqueda de música por melodía cantada. In *Anais do X Simpósio Brasileiro de Computação Musical*, páginas 142–153, Belo Horizonte, Brasil. [cited at p. 4]

[Marchand, 2000] Marchand, S. (2000). *Sound Models for Computer Music (analysis, transformation, and synthesis of musical sound)*. tese de PhD, Université Bordeaux 1, Talence Cedex, France. [cited at p. 23]

[Mitre, 2005] Mitre, A. (2005). Um sistema automático de transcrição melódica. In *Anais do I Simpósio de Iniciação Científica e Pós-Graduação do IME-USP*, São Paulo, Brasil. [cited at p. 64]

[Mitre e Queiroz, 2005] Mitre, A. e Queiroz, M. (2005). Um sistema automático de transcrição melódica. In *Anais do X Simpósio Brasileiro de Computação Musical*, páginas 174–185, Belo Horizonte, Brasil. [cited at p. 63]

[Nuttall, 1981] Nuttall, A. H. (1981). Some Windows with Very Good Sidelobe Behavior. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(1):84–91. [cited at p. 28, 30, 31]

[Regalia, 1993] Regalia, P. A. (1993). *Handbook for Digital Signal Processing*, chapter Special Filter Designs. John Wiley & Sons. [cited at p. 47]

[Roads, 1996] Roads, C. (1996). *The Computer Music Tutorial*. The MIT Press. [cited at p. 5]

[Roederer, 2002] Roederer, J. G. (2002). *Introdução à Física e Psicofísica da Música*. EdUSP, São Paulo, primeira edição. Tradução Alberto Luis da Cunha. [cited at p. 36]

[Serra e Smith III, 1990] Serra, X. e Smith III, J. O. (1990). Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal*, 14(4):12–24. [cited at p. 8, 10]

[Smith III, 2003] Smith III, J. O. (2003). *Mathematics of the Discrete Fourier Transform (DFT) - with audio applications*. W3K Publishing. available online at http://ccrma.stanford.edu/ jos/mdft/. [cited at p. 15, 20]

Parte II

Parte Subjetiva

Capítulo 14

Desafios e Frustrações

A realização deste trabalho foi, sem dúvida nenhuma, um grande desafio. Abaixo listamos algumas atividades desafiadores que fizeram parte de sua realização:

- Busca de textos científicos relacionados, direta ou indiretamente, ao problema da transcrição melódica. A ferramenta Google foi extremamente valiosa neste processo, já que muitos artigos foram encontrados por meio de buscas por termos chave. Além disso, outra forma muito importante de descoberta de textos foi a consulta às referências bibliográficas de cada texto já encontrado.

- Alguns artigos puderam ser obtidos a partir de versão eletrônica disponível livremente. Contudo, a obtenção de alguns artigos dependeu de convênios que bibliotecas da USP tem com revistas científicas, por exemplo a IEEE.

- Ler artigos científicos foi uma experiência nova e desafiadora. Muitas vezes os textos não são claros e raramente são autocontidos.

- O supervisor deste trabalho, o Prof. Marcelo Queiroz, sugeriu no início do ano, que apresentássemos o trabalho no X Simpósio Brasileiro de Computação Musical (SBCM 2005). O autor deste trabalho nunca havia escrito um artigo científico até então, mas aceitou o desafio. Um artigo foi escrito e submetido à apreciação da comissão técnica. Felizmente, o texto foi aceito, com avaliações bastante positivas. Apesar disso, alguns revisores sugeriram que o artigo fosse tipografado em LaTeX, uma vez que parte substancial do mesmo empregava linguagem matemática. O processador de texto Microsoft Word havia sido usado na redação da primeira versão do artigo. O autor deste texto não possuía, então, praticamente nenhuma experiência com LaTeX. Aprender rapidamente a usar esta ferramente foi certamente um desafio, que foi vencido e a versão final do artigo ([Mitre e Queiroz, 2005]) foi feita em LaTeX. Viajar para outra cidade a fim de apresentar o trabalho foi, sem dúvida, um outro grande desafio.

- Embalado pelo "sucesso" obtido no X Simpósio Brasileiro de Computação Musical, o autor deste trabalho se inscreveu no I Simpósio de Iniciação Científica e Pós-Graduação do IME-USP, em cujos anais figura uma versão abreviada do artigo enviado para o SBCM 2005 ([Mitre, 2005]).

- Os gráficos deste texto foram produzidos, quase que em sua totalidade, utilizando o *gnuplot**. O autor não possui prévia experiência com este aplicativo e aprender a usá-lo foi um desafio. Mais do que aprender a usá-lo, foi necessário entender a sintaxe esperada dos arquivos de entrada, pois a saído do sistema aqui apresentado deveria conformar-se a ela. Isso viabilizou a produção dos gráficos que ilustram este texto.

- MIDI é um formato binário e, portanto, "chato" de manipular. Uma solução alternativa para a produção de arquivos MIDI foi procurada e encontrada: *MIDICSV*†. Este aplicativo recebe um arquivo em formato CSV‡ e produz o arquivo MIDI correspondente.

- Não há uma base padrão com gravações para testar sistemas de transcrição musical. Coletar gravações monofônicas foi difícil, mas felizmente foi possível reunir uma base representativa e de porte razoável.

Com relação às frustrações, é possível citar a redução no escopo do trabalho, que originalmente ambicionava, por exemplo, estimar o andamento da música e lidar com entradas polifônicas. Contudo, o autor compreendeu que era melhor resolver bem um problema menor a resolver mal um problema maior. Por conta disso, reduziu-se o escopo do trabalho à transcrição de melodias.

* disponível gratuitamente em http://www.gnuplot.info

† idem, http://www.fourmilab.ch/webtools/midicsv

‡ acrônimo, em inglês, para *valores separados por vírgula.*

Capítulo 15

Disciplinas do curso de maior relevância

A seguir, são apresentadas as disciplinas do Bacharelado em Ciência da Computação de maior relevância para a realização do presente trabalho. Os conceitos de cada curso que o tornaram importante são listados sob o título da disciplina.

15.1 Computação

MAC 0110 - Introdução à Computação

- Primeiros contatos com programação em C

MAC 0122 - Princípios de Desenvolvimento de Algoritmos

- Mais experiência com a linguagem de programação C
- Utilização de estruturas de dados mais sofisticadas
- Habilidade de desenvolver algortimos, avaliar sua corretude, bem como analisar (e possivelmente melhorar) sua eficiência

MAC 0211 - Lab.Prog.1

- Modularização de código
- Portabilidade

MAC 0242 - Lab.Prog.2

- Linguagens de programação baseadas em scripts

- Desenvolvimento de sistemas maiores e mais complexos do que os já desenvolvidos até então

15.1.1 MAC0337 - Computação Musical

- Análise de Fourier — "coração" do sistema
- Algoritmo FFT — implementação eficiente da STDFT

Além disso, foi por ocasião do curso de Computação Musical que o autor deste trabalho conheceu seu supervisor, o Prof. Marcelo Queiroz.

15.2 Matemática

MAT 0111 - Cálculo I

- Integral — imprescindível à análise Fourier
- Derivada — fundamental para a compreensão dos métodos de estimação precisa de parciais

MAT 0139 - Álgebra Linear para Computação

- Espaço vetorial — as funções contínuas T-periódicas formam um espaço vetorial
- Base ortogonal — $\left\{e^{i2\pi nT^{-1}}\right\}_{n\in\mathbb{N}}$ é uma base deste espaço

MAC0338 - Análise de Algoritmos

- Análise assintótica de desempenho — implementação eficiente do filtro da mediana (§9.4.2)
- Diversos algoritmos importantes
- Aquisição de maior rigor e clareza no desenvimento e análise de algoritmos — por exemplo, o algoritmo de Produção de Notas

MAE0121 - Introdução à Probabilidade e à Estatística I

- Mediana e sua generalização, o quantil — filtro baseado em ordem estatística, por exemplo o da mediana (§9.4.2)

15.3 Física

FAP 0137 - Física II

- Ondas — instrumentos melódicos produzem sons harmônicos (§2.1)

Capítulo 16

Agradecimentos

Durante o desenvolvimento deste trabalho, contei com a ajuda de muitas pessoas. Algumas merecem destaque.

Tendo realizado a maior parte deste trabalho em minha casa, devo agradecer, em primeiro lugar, à minha família. Sou muito grato ao meu pai, inesitante financiador de meus estudos e cientista exemplar. Agradeço também à minha mãe, com seu fôlego infinito para discutir comigo sobre uma infinidade de assuntos. Por último, mas não menos importante, devo agradecer ao meu irmão, companheiro ímpar em meus momentos de lazer. Devo agradecer a todos os que vivem comigo pela paciência quando eu realizava testes sonoros.

Agradeço aos meus amigos, cujo encorajamento foi indispensável. Eles sabem quem são.

Devo agradecimentos também aos amigos novos, feitos no X Simpósio Brasileiro de Computação Musical. Conhecer gente que partilha dos mesmos interesses é sempre gratificamente. Martín, Ernesto e Gabriel, espero poder continuar aprendendo e compartilhar o que aprendi com vocês.

A meu supervisor, o Prof. Marcelo Queiroz, cujo estímulo e senso crítico foram inestimáveis.

Sou grato ainda à Fundação de Amparo à Pesquisa do Estado de São Paulo, que apoiou financeiramente este trabalho (projeto temático 02/02678-0).