Um Sistema Automático de Transcrição Melódica *

Adriano Mitre e Marcelo Queiroz

Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)
Rua do Matão, 1010 – 05508-090 – São Paulo, SP – Brasil

amitre@linux.ime.usp.br, mqz@ime.usp.br

Abstract. *A system for the automatic transcription of melodies which implements state-of-the-art techniques of sinusoids estimation is presented. A novel technique, which benefits from robust partial estimates, is proposed for the estimation of fundamental frequency. Furthermore, system's flexibility allows it to serve also as a signal analysis tool, aiding in the production of plots such as sonograms and F0-grams.*

Resumo. *Um sistema de transcrição automática de melodias que implementa diversas técnicas de estimação robusta de senóides é apresentado. Partindo de estimativas confiáveis de parciais, uma nova heurística para determinação de freqüência fundamental é proposta. Tendo sido concebido de maneira modular, o sistema pode ainda ser usado como ferramenta de análise de sinais, servindo à produção de gráficos como sonogramas e F0-gramas.*

1. Introdução

O problema da transcrição automática de melodias, embora estudado há vários anos (vide [Roads, 1996, cap.12] para um estudo taxonômico), dificilmente pode ser considerado um problema fechado. Uma variedade de aspectos técnicos e cognitivos fazem 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 na seção 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 única/objetiva, 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 nova, baseada em estimativas confiáveis de parciais, para a determinação de freqüência fundamental.

O sistema tem como entrada um sinal de áudio monofônico (por exemplo uma gravação acústica) e produz uma transcrição de seu conteúdo melódico. No contexto deste artigo, transcrição significará a obtenção de uma representação simbólica, mais formalmente um conjunto de quádruplas da forma $(t_0, n, \Delta t, v)$, onde n denota a altura da

*O projeto teve apoio da FAPESP, projeto temático 02/02678-0.

nota, v sua intensidade, t_0 seu início e Δt sua duração. No restante do texto iremos nos referir a essas quádruplas como notas.

Nosso sistema possui três modos de impressão de notas: descritivo, *piano roll* e "MIDIficável". O primeiro tem como saída um texto de fácil leitura, em que as notas são descritas com letras (notação latina) e oitavas em relação à central, por exemplo $\mathrm{E}\flat(-1)$. Já o segundo modo produz um arquivo que ao ser carregado no *gnuplot*[1] produz uma visualização das notas em estilo *piano roll*.

O terceiro modo, por sua vez, tem como saída um arquivo em formato CSV[2] que, ao ser processado pelo utilitário *MIDICSV*[3], dará origem a um arquivo MIDI. Neste caso, o andamento, a forma de compasso e a pausa inicial serão fornecidos pelo usuário. É possível ainda, a partir do arquivo MIDI, obter partituras das transcrições com o uso de sistemas como o *Lilypond*[4].

2. Arquitetura do Sistema

Uma parcela significativa dos sistemas de transcrição encontrados na literatura mesclam 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 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*.

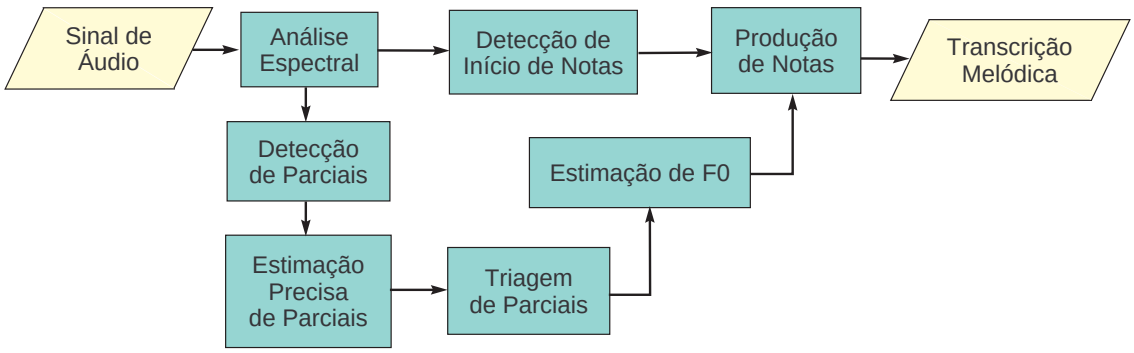

Figura 1: Módulos do sistema e seu fluxo de dados.

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].

[1] disponível gratuitamente em http://www.gnuplot.info
[2] acrônimo, em inglês, para *valores separados por vírgula*.
[3] idem, http://www.fourmilab.ch/webtools/midicsv
[4] idem, http://lilypond.org/web

3. 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 1 apresenta um resumo dessas características.

Tabela 1: Modelo evolutivo de uma nota típica.

Estágio	Energia	Caráter
Ataque	Cresce rapidamente	Transiente
Amortecimento	Decresce	Em estabilização
Sustentação	Peramenece 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

$$s\left(t\right) = d\left(t\right) + e\left(t\right) \tag{1}$$

onde

- $d\left(t\right)$ é a parte determinística (componente de banda-estreita)
- $e\left(t\right)$ 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

$$d\left(t\right) = \sum_{p=1}^{P} osc\left(f_p\left(t\right), a_p\left(t\right)\right) \tag{2}$$

onde P é o número de parciais, dados por

$$osc\left(f_p\left(t\right), a_p\left(t\right)\right) = a_p\left(t\right) \cdot \cos\left(\phi_p\left(t\right)\right) \tag{3}$$

com

$$\frac{d\phi_p}{dt}\left(t\right) = 2\pi f_p\left(t\right) \quad \text{i.e.} \quad \phi_p\left(t\right) = \phi_p\left(0\right) + 2\pi \int_0^t f_p\left(u\right) \cdot du \tag{4}$$

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

$$f_p\left(t\right) = p \cdot f_1\left(t\right) \tag{5}$$

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:

$$\min_{1 \leq i < j \leq P} \left\{\left|f_i\left(t\right) - f_j\left(t\right)\right|\right\} \geq f_1\left(t\right) \tag{6}$$

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 constante, mas variar em função de sua espessura, comprimento e tensão (logo, B é determinado por estes parâmetros). 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 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*). Apesar de o sistema implementar diversas técnicas de detecção de ínicio de nota (por exemplo [Duxbury et al., 2003]), foi necessário suprimir sua apresentação por conta do limite de páginas desta publicação.

4. Análise Espectral

O sistema obtém estimativas espectrais por meio da Transformada Discreta de Fourier de Curto Tempo (abreviada em inglês por STDFT). 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.

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

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

e também

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

6. Estimação Precisa de Parciais

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 $\mathrm{F0}_{\min} \cdot \left(\sqrt[12]{2} - 1\right)$ Hz, onde $\mathrm{F0}_{\min}$ denota a freqüência fundamental da nota mais grave da seqüência melódica de entrada. Para se extrair nuances expressivas como vibratos e glissandos, entretanto, a precisão necessária é 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 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 $\mathrm{F0}_{\min}$ Hz. Existem, para situações como essa, técnicas para obter estimativas mais precisas a partir de quadros com duração $L_w \cdot (\mathrm{F0}_{\min})^{-1}$ s. 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.

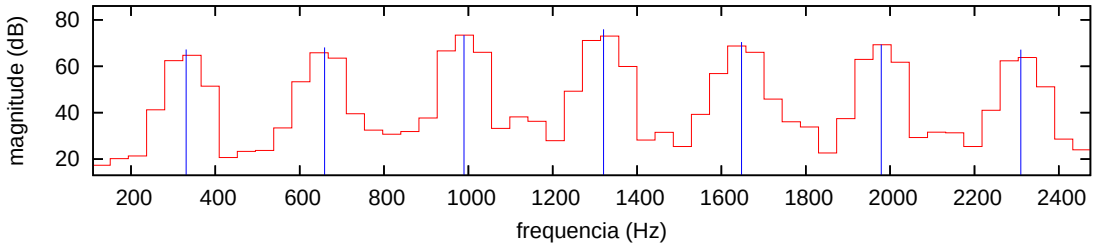

Figura 2: Espectro de magnitude e estimativas precisas dos parciais.

A superamostragem espectral, popularmente referida como *zero-padding*, é uma dessas técnicas. 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. A desvantagem dessa técnica é seu custo computacional: para aumentar a precisão em m vezes é necessário calcular uma STDFT com tamanho mN.

Outros métodos obtêm estimativas mais precisas dos parciais por meio de interpolação com coeficientes espectrais adjacentes. A interpolação parabólica pertence a essa categoria. Essa técnica 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. Essa técnica 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.

Especificamente para a janela de Hann[5], Grandke desenvolveu um algoritmo de interpolação que utiliza o coeficiente vizinho de maior magnitude para obter estimativas mais consistentes do que as obtidas pela interpolação parabólica [Grandke, 1983].

Há técnicas mais sofisticadas de estimação de parciais, as quais consideram, além da magnitude, a informação de fase presente no espectro. Dentre essas, destacam-se o Método da Derivada [Desainte-Catherine e Marchand, 2000], que se baseia na relação entre os espectros do sinal e de sua derivada (aproximada com a aplicação de um filtro), e a Reatribuição Espectral [Kodera et al., 1978, Auger e Flandrin, 1995], que redispõe os pontos do reticulado tempo-freqüência de acordo com os centros de energia de cada célula. Essas técnicas apresentam desempenho superior, mas possuem a desvantagem de exigirem o cômputo de STDFT adicionais. A Figura 2 exemplifica os resultados obtidos com a Reatribuição Espectral. A Tabela 2 lista as principais técnicas de estimação de parciais aplicáveis a sinais musicais e algumas de suas características.

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

Tabela 2: Estimadores de parciais implementados 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
Reatribuição	$O(N \lg N)$	$O(1)$	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

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] e [Hainsworth e Macleod, 2003].

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.

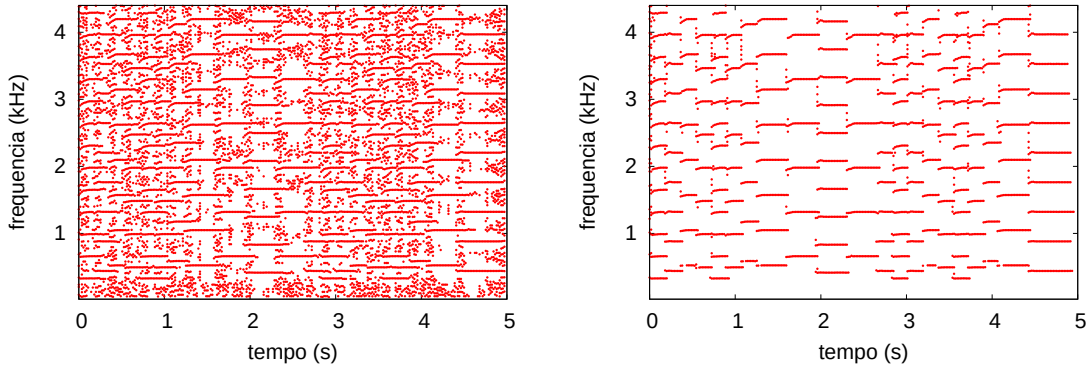

Figura 3: Exemplo do efeito da triagem de parciais.

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.

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

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.

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

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

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

$$C = \left\{ c_n \stackrel{def}{=} \frac{f_\kappa}{n} : 1 \leq n \leq \left\lfloor \frac{f_\kappa}{\text{F0}_{\min}} \right\rfloor \right\} \tag{7}$$

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)
1  for each (f, a) ∈ P
2       do i ← ⌊f / f_cand + 0.5⌋
3          d ← DESVIO-RELATIVO(f_cand · i, H[i]_freq)
4          d̂ ← DESVIO-RELATIVO(f_cand · i, f)
5          if d̂ ≤ min{d, ∜2}
6             then if d̂ < d or a > H[i]_mag
7                     then H[i]_freq ← f
8                          H[i]_mag ← a
9  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, seção 3.4].

As funções Φ e Ψ, definidas a seguir, são adaptações do modelo de soma harmônica apresentado em [Klapuri, 2004, seção 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

$$\Phi(n) = \sum_{i=1}^{I(n)} H[n][i]_{mag} \cdot \Psi(i) \tag{8}$$

$$\text{onde } I(n) = \max\{j \in \{1, \ldots, |H|\} : H[n][j]_{mag} > 0\} \tag{9}$$

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

$$\Psi(i) = \begin{cases} 1, \text{ se } i = 1 \\ \min\left\{\log_{2^{1/3}}\left(i \cdot \sqrt{\frac{i+1}{i}}\right) - \log_{2^{1/3}}\left((i-1) \cdot \sqrt{\frac{i}{(i-1)}}\right), 1\right\}, \text{ c.c.} \end{cases} \tag{10}$$

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 seguinte equações

$$\varphi = \min\left\{n : c_n \in C^\Phi \text{ e } \chi(n) \geq \chi(m), \forall c_m \in C^\Phi\right\} \tag{11}$$

$$\chi(n) = \frac{\sum_{i=1}^{I(n)} H[n][i]_{mag} \cdot \Psi(i)}{\sum_{i=1}^{I(n)} \Psi(i)} \tag{12}$$

$$C^\Phi = \left\{c_n \in C : \Phi(n) \geq \beta \cdot \max_{m|c_m \in C}\{\Phi(m)\}\right\} \tag{13}$$

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:

$$\text{F0}' = \frac{\sum_{i=1}^{I(n)} H[i]_{freq}/i \cdot H[i]_{mag} \cdot \Psi(i)}{\sum_{i=1}^{I(n)} H[i]_{mag} \cdot \Psi(i)} \tag{14}$$

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

9. 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,

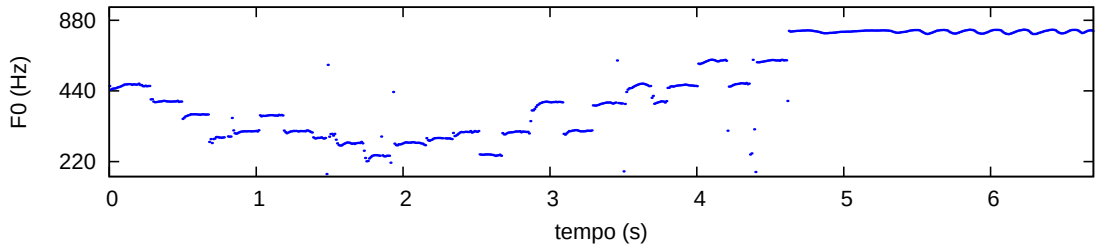

Figura 4: Exemplo de saída do módulo de estimação de F0.

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. Além de uma seqüência de estimativas de F0 (Figura 4), 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.

O módulo de produção de notas tem essencialmente três finalidades. A primeira é evitar que irregularidades na estimação de F0 ocasionem notas espúrias. A segunda é eliminar o atraso decorrente do fato de as notas muitas vezes não possuírem, em seus estágios iniciais, altura (e, portanto, F0) bem definida. A terceira é impedir que uma seqüência de notas repetidas seja interpretada como uma única nota.

PRODUZ-NOTAS$(Fund, T)$

1 **for** $i \leftarrow 1$ **to** $tamanho[Fund]$
2 **do** $n \leftarrow \lfloor \log_{\sqrt[12]{2}} (freq[Fund[i]]\,/440) + 0.5 \rfloor + 69$ ▷ número de nota MIDI
3 **if** $ultimaConstatacao[Nota[n]] \neq$ INDEFINIDA
4 **then if** $inst[Fund[i]] - ultimaConstatacao[Nota[n]] > \gamma_{\max}$
5 **then if** $intens[Nota[n]] > \nu_{\min}$ **and** $duracao[Nota] > \Delta_{\min}$
6 **then** IMPRIME$(Nota[n])$
7 REINICIA$(Nota[n])$
8 **if** $inicio[Nota[n]] =$ INDEFINIDO
9 **then** $t \leftarrow \max\{T[i] \leq inst[Fund[i]]\}$
10 **if** $inst[Fund[i]] - t \leq \sigma_{max}$
11 **then** $inicio[Nota[n]] \leftarrow t$
12 **else** $inicio[Nota[n]] \leftarrow inst[Fund[i]]$
13 $ultimaConstatacao[Nota[n]] \leftarrow inst[Fund[i]]$
14 $intens[Nota[n]] \leftarrow \max\{intesidade[Fund[i]], intens[Nota[n]]\}$
15 $duracao[Nota[n]] \leftarrow ultimaConstatacao[Nota] - inicio[Nota]$
16 **for** $m \leftarrow 1$ **to** $tamanho[Nota]$
17 **do if** $inst[Fund[i]] - inicio[Nota[m]] > \gamma_{\max}$ **or** $i = tamanho[Fund]$
18 **then if** $intens[Nota[m]] > \nu_{\min}$ **and** $duracao[Nota[m]] > \Delta_{\min}$
19 **then** IMPRIME$(Nota[m])$
20 REINICIA$(Nota[m])$

10. Resultados e Discussão

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.

11. Conclusão

Neste trabalho apresentamos um sistema de transcrição automática de melodias com desempenho bastante promissor. 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.

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 desta publicação foram produzidos desta forma.

O sistema está disponível em www.mitre.com.br/asymut/, onde se encontram também exemplos em que se pode comparar a gravação original com a resintetizada a partir da transcrição.

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.

Referências

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.

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.

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.

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

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

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.

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.

Kernighan, B. W. e Pike, R. (2000). *A Prática da Programação*. Editora Campus, primeira edição.

Klapuri, A. (2004). *Signal Processing Methods for the Automatic Transcription of Music*. tese de PhD, Tampere University of Technology.

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.

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.

Roads, C. (1996). *The Computer Music Tutorial*. The MIT Press.

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

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.