UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ELETRÔNICA E DE COMPUTAÇÃO

Técnicas de Análise e Síntese de Sinais Musicais

Autor:
Marcelo Alves Schmalter Soares

Orientador:
Luiz Wagner Pereira Biscainho, D.Sc.

Examinador:
Sergio Lima Netto, Ph.D.

Examinador:
Márcio Nogueira de Souza, D.Sc.

DEL
Novembro de 2004

Agradecimentos

Agradeço, primeiramente, a Deus. Agradeço aos meus pais, que me educaram e tornaram possível esse momento.

Agradeço ao professor e amigo Luiz Wagner, pela orientação em toda a minha vida acadêmica e pelo apoio que me foi dado.

Agradeço aos companheiros de LPS Gustavo Luís, Bruno Carluccio e Fábio Freeland. Ao Cristiano Santos, que iniciou esse trabalho, ao Paulo Esquef, fundamental com seus trabalhos, e ao Iuri Kothe, que me ajudou, e muito, a finalizá-lo e irá continuá-lo.

Agradeço a todos os professores e aos amigos da turma Heróis da Resistência, em especial ao Igor, ao Roberto, ao Fábio Carvalho, ao José Antonio e ao Tiago Hosken, meu fiel escudeiro há mais de 16 anos.

Agradeço também a todos os meus amigos e parentes pelo apoio irrestrito.

Resumo

Descreveremos no presente trabalho ferramentas úteis na análise e na síntese de sinais musicais, com ênfase na separação de fontes sonoras para futuro enriquecimento de gravações musicais.

No capítulo 1, temos a implementação e teste do método de representação tempo-freqüência baseado na DFT[1]. Tratamos da modelagem do sinal, explicitando as diferenças entre esse método e o da DFT, do qual a DFT[1] se origina. São mostrados exemplos práticos da utilização desse método, mostrando sua vantagem sobre o da DFT.

No capítulo 2, um método algébrico para resolver o problema da "mistura" de picos diferentes no domínio da freqüência devido ao janelamento do sinal é descrito com detalhes, implementado e avaliado.

No capítulo 3, temos a apresentação de alguns refinamentos às técnicas de análise vistas nos capítulos anteriores. Posteriormente, temos a apresentação de métodos para realizarmos a separação de sinais com duas vozes instrumentais e, finalmente, apresentamos um método para a ressíntese de sinais. Fechando o capítulo temos a apresentação de resultados para os métodos de separação e ressíntese.

Temos, no capítulo 4, as conclusões acerca desse trabalho.

Palavras-Chave

Análise Espectral

Áudio

Separação de Sinais

Ressíntese de Sinais

Sumário

Capítulo 1

Análise de Sinais Usando DFT^1

1.1 Introdução

A manipulação digital do som requer um bom modelo para este. Para imitar ou transformar sons existentes com precisão, métodos eficazes de análise são de vital importância, para que se possam extrair parâmetros que permitam construir um modelo para os sons. A eficácia desse método de análise determinará a qualidade dos sons ressintetizados resultantes, que chegarão aos nossos ouvidos.

Em se tratando de sinais musicais, a característica predominantemente tonal e não-estacionária dos sinais a analisar leva naturalmente à busca de representações num domínio misto tempo-freqüência. Tipicamente, divide-se o sinal em fatias no tempo e atribui-se a cada fatia (janela) uma representação estacionária em freqüência. Este é o princípio da "short-time Fourier Transform" (STFT). Nas diversas aplicações da STFT ao processamento de sinais, enfrentamos sempre um dilema: Para discriminarmos a ocorrência de eventos num curto espaço de tempo precisamos de janelas curtas; já se desejarmos discriminar informações numa faixa de freqüências muito baixas, precisamos de uma longa janela "de observação".

Em [1], propõe-se uma técnica alternativa que, levando em conta informações da derivada do sinal, objetiva aumentar a precisão do método convencional, no que tange ao balanço tempo-freqüência. Essa técnica é conhecida como DFT^1.

Este capítulo consiste na implementação e teste do método de representação tempo-freqüência baseado na DFT^1. Trataremos da modelagem do sinal, explicitando as diferenças entre esse método e o da DFT, do qual a DFT^1 se origina. São mostrados exemplos práticos da utilização desse método, mostrando sua vantagem sobre o da DFT.

1.2 Modelamento do Sinal

O modelo do sinal deve ser o mais geral possível, para que a maioria dos sons possam ser reproduzidos fielmente e transformados de uma maneira natural e musicalmente expressiva. Estaremos usando o modelo proposto por Serra e Smith [2],baseado numa decomposição do sinal em uma parte determinística e outra estocástica:

$$a(t) = d(t) + s(t), \tag{1.1}$$

onde:

$$d(t) = \text{ parte determinística;}$$

$$s(t) = \text{ parte estocástica.}$$

A parte determinística é uma soma de oscilações senoidais cujas freqüências e amplitudes evoluem, variando lentamente no tempo. Tal oscilação é chamada de parcial. Assim, temos:

$$d(t) = \sum_{p=1}^{P} osc[f_p(t), a_p(t), \phi_p(0)], \tag{1.2}$$

onde P é o número de parciais e

$$osc[f_p(t), a_p(t), \phi_p(0)] = a_p(t)\cos[\phi_p(t)], \tag{1.3}$$

com

$$\frac{d\phi_p}{dt}(t) = 2\pi f_p(t), \tag{1.4}$$

ou seja,

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u)du. \tag{1.5}$$

As funções f_p, a_p e ϕ_p são, respectivamente, a freqüência, amplitude e fase da p-ésima parcial. A fase na origem será ignorada durante a análise.

Não será feita aqui a distinção entre parciais harmônicas e não-harmônicas, isto é, consideraremos terem todas freqüências múltiplas inteiras da fundamental.

A parte estocástica é o que sobra após a parte determinística ter sido subtraída do sinal original e é geralmente reconhecida como uma espécie de ruído.

Como trataremos de sons que, a princípio, têm esse nível de ruído baixo (o que é real para muitos sons naturais, especialmente após seu ataque), iremos ignorar a parte estocástica. Teremos, então:

$$a(t) = \sum_{p=1}^{P} osc[f_p(t), a_p(t), \phi_p(0)]. \tag{1.6}$$

Outra restrição é que os parciais devem ser afastados significativamente na freqüência. Para um som a, existe uma distância mínima $d > 0$ tal que:

$$\text{mín}_{i \neq j, t}\{ \mid f_j(t) - f_i(t) \mid \} > d. \tag{1.7}$$

Essa condição, que previne as freqüências de duas parciais de se cruzarem, é uma hipótese perfeitamente aceitável, verificada ao menos para quase todo som monofônico natural.

1.3 Análise de Fourier de curta duração

De acordo com o teorema de Fourier [3], qualquer função periódica pode ser modelada como uma soma de sinais senoidais com várias amplitudes e com freqüências harmônicas.

A transformada de Fourier [3] converte o sinal temporal (amplitude *versus* tempo) para uma representação espectral (amplitude *versus* freqüência). Ela informa a composição do som nas diversas freqüências e exibe todo o sinal em um só espectro. Porém, esse único espectro corresponde a um sinal estacionário, ou seja, cujas características freqüenciais não evoluem com o tempo. Para analisarmos sinais como a maioria daqueles com que lidamos no nosso cotidiano, cuja composição espectral evolui no tempo, lançamos mão de uma técnica conhecida como "de curta duração", ou seja, que analisa pequenos trechos temporais. Cada trecho desses é tratado como um sinal estacionário, e depois juntam-se todos os trechos compondo um espectro que evolui no tempo.

Essa técnica de divisão do sinal em blocos é utilizada na STFT, definida como

$$X(w, \tau) = \sum_{-\infty}^{\infty} x(t) w(t - \tau) e^{-j\omega t} dt, \tag{1.8}$$

para um sinal $x(t)$ e uma janela $w(t)$. A STFT produz uma série de espectros de curta duração obtidos em pequenos pedaços sucessivos do sinal temporal, geralmente sobrepostos. Na prática, estamos lidando com um sinal discreto resultante da amostragem uniforme do sinal contínuo com freqüência F_s. Então, para cada janela $x[n]$ de N amostras consecutivas, sua transformada discreta de Fourier $X[m]$ é computada:

$$X[m] = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-j(2\pi/N)nm}. \tag{1.9}$$

O que realmente acontece é que uma função de janelamento discreta, w, de comprimento N, está multiplicando a seqüência original. Se $w[n] = 1$ para todo n entre 0 e N,

como na equação (1.9), w é chamada de janela retangular, que é a janela de análise mais simples.

Algumas limitações decorrem do uso de janelas. O formato e o comprimento da janela são fatores-chave nesse aspecto.

1.3.1 Imprecisões de freqüência

Note que a DFT é simplesmente a amostragem em N pontos da $DTFT$ (discrete-time fourier transform) do sinal correspondente, definida como

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega t} \tag{1.10}$$

. Assim, o espectro de magnitude do sinal é amostrado de 0 a $F_s/2$ Hz (freqüência de Nyquist), em passos de F_s/N. A precisão da análise de freqüência é proporcional a N. Para uma boa precisão, necessitamos de um pequeno F_s/N, ou seja, um grande valor de N, o que leva a uma má precisão no tempo (já que N é o tamanho da janela em amostras). Para uma boa resolução no tempo teremos uma má resolução na freqüência. É o já citado dilema freqüência *versus* tempo.

1.3.2 Imprecisões de amplitude

O espectro (DTFT) do produto de w por x, que precisa ser realizado nos processamentos de curta duração, é dado pela convolução $W * X$. Se, por exemplo, x é uma senóide complexa pura com amplitude 1 e freqüência f, X adquire uma forma bastante conveniente para a extração de picos: um impulso unitário na freqüência da senóide. Para que a representação por $W * X$ fosse ideal, W deveria ser o mais perto possível de um impulso, o que é impossível com uma janela finita. Na prática, W consiste de lóbulos, como mostra a figura 1.1, e é desejável que os lóbulos laterais tenham energias desprezíveis se comparados ao lóbulo principal. Porém, a razão entre os lóbulos laterais e o principal é inversamente proporcional à largura do lóbulo principal, o que acaba por determinar a resolução na freqüência da janela utilizada. Para uma boa resolução de freqüências é desejável um lóbulo principal estreito, o que distorce a magnitude do espectro. Tenta-se utilizar uma janela w que minimize a distorção de X na convolução $W * X$.

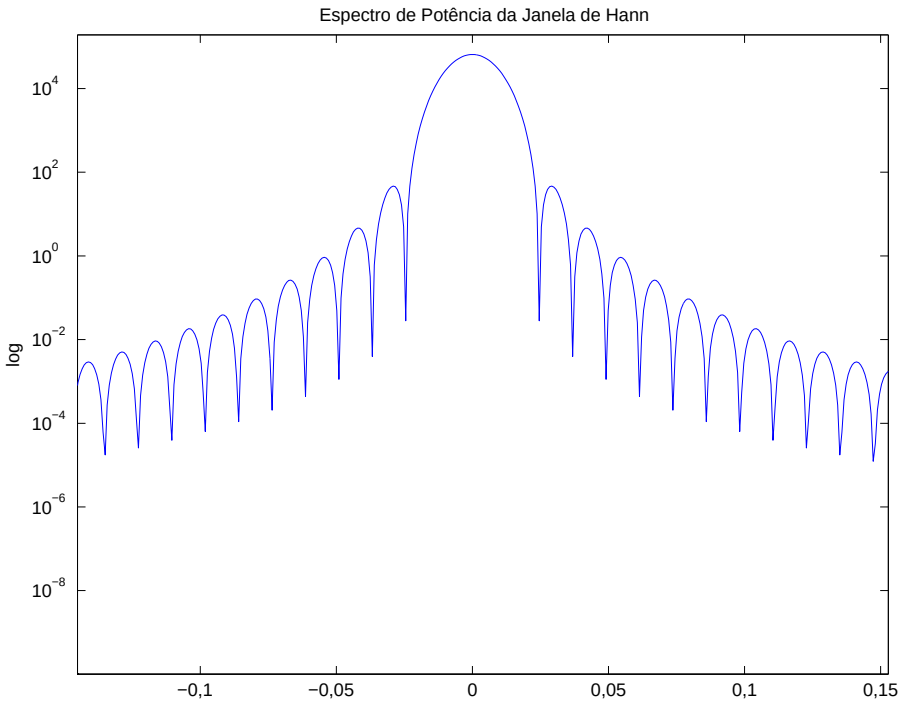

Figura 1.1: Espectro de potência da janela de Hann.

1.4 Transformada de Fourier usando derivadas do sinal

Temos em [1] o método que será explicado a partir de agora.

Para uma dada parcial, temos que a freqüência é a derivada da fase, assim como a derivada da freqüência é a segunda derivada da fase. Ambas, freqüência e fase, serão consideradas como se variassem lentamente no tempo. Mais precisamente, as funções freqüência e amplitude das parciais são supostas limitadas em banda na freqüência. Durante a aplicação de uma janela, suas derivadas são praticamente nulas. Como ao diferenciarmos um seno teremos outro seno, com diferente fase e mesma freqüência, usaremos esse fato para melhor detectar as freqüências, usando as derivadas do sinal.

Considerando:

$$o_p(t) = osc[f_p(t), a_p(t), \phi_p(0)] = a_p(t)\cos[\phi_p(t)], \tag{1.11}$$

temos que

$$\frac{do_p}{dt}(t) = a_p(t)\frac{d}{dt}\{\cos[\phi_p(t)]\} + \frac{da_p}{dt}(t)\cos[\phi_p(t)]. \tag{1.12}$$

Como dito antes: $da_p/dt \approx 0$. Assim,

$$\frac{do_p}{dt}(t) \approx a_p(t)\frac{d}{dt}\{\cos[\phi_p(t)]\} = -a_p(t)\frac{d\phi_p}{dt}(t)sen[\phi_p(t)]. \tag{1.13}$$

Mas $d\phi_p/dt(t) = 2\pi f_p(t)$; então:

$$\frac{do_p}{dt}(t) = -a_p(t)[2\pi f_p(t)]sen[\phi_p(t)] = 2\pi a_p(t) f_p(t)\cos[\phi_p(t) - \frac{\pi}{2}], \tag{1.14}$$

o que nos leva a:

$$\frac{d^k a}{dt^k}(t) = \sum_{p=1}^{P} a_p(t)[2\pi f_p(t)]^k \cos[\phi_p(t) + (-\frac{k\pi}{2})]. \tag{1.15}$$

Vamos chamar de FT^k a transformada de Fourier da k-ésima derivada do sinal $\frac{d^k a}{dt^k}(t)$ ($k \geq 0$), e chamaremos de $FT^k(f)$ sua amplitude na freqüência f. Deve existir um máximo em todo espectro de potência de FT^i para cada parcial p. Usando os espectros de potência de FT^i para diferentes valores de i podemos determinar as freqüências exatas para a gama de parciais.

Para cada parcial p existe um máximo em todo FT^i na freqüência f_p:

$$f_p = \frac{FT^{i+1}(f_p)}{2\pi FT^i(f_p)}. \tag{1.16}$$

Apesar de parecer inútil a definição acima, pois necessita previamente do valor de f_p, a sua versão discreta é de grande interesse. Isto porque a amostragem do espectro de FT^k leva a um valor aproximado f_p^0 da freqüência f_p. Com a versão discreta, podemos chegar a um valor mais correto para a freqüência.

Já foram ditas as conseqüências de se usar determinados tipos de janela, com suas deformações na amplitude e na freqüência. Podemos, porém, obter a amplitude exata a_p de uma parcial p, partindo do valor aproximado $a_p^0 = FT^0(f_p)$, com a equação:

$$a_p = \frac{a_p^0}{W(0)}, \tag{1.17}$$

sendo $W(\omega)$ a transformada de Fourier da janela $w[n]$ empregada na análise do sinal amostrado $a[n]$.

1.5 Transformada discreta

Como a derivada do sinal não é gravada junto com o sinal, ela deve ser calculada através do sinal digital, após a fase de amostragem. Todo o sinal discreto foi amostrado uniformemente com uma taxa de amostragem F_s.

A primeira derivada $\frac{da}{dt}(t)$ ou $a^{'}(t)$ é definida matematicamente como:

$$a^{'}(t) = lim_{\varepsilon \to 0} \frac{a(t+\varepsilon) - a(t)}{\varepsilon}. \tag{1.18}$$

Temos, na verdade, duas derivadas para um sinal, a pela direita e a pela esquerda, para ε positivo ou negativo; porém, no caso do sinal $a(t)$, suposto contínuo, as duas derivadas são equivalentes:

$$a_{-}^{'}(t) = a_{+}^{'}(t) = a^{'}(t). \tag{1.19}$$

Considerando-se o sinal já discretizado, $a[i]$ representa $a(i \times \frac{1}{F_s})$, e o menor ε diferente de zero para a computação da derivada é o período de amostragem $\frac{1}{F_s}$. Faremos as aproximações:

$$a_{-}^{'}[i] = (a[i] - a[i-1])F_s, \tag{1.20}$$

$$a_{+}^{'}[i] = (a[i+1] - a[i])F_s, \tag{1.21}$$

$$a_{+}^{'}[i] = a_{-}^{'}[i+1] \text{ e} \tag{1.22}$$

$$a_{-}^{'}[i] = a_{+}^{'}[i-1], \tag{1.23}$$

ou seja, as duas derivadas são a mesma função, exceto por uma translação de uma amostra no tempo. Vamos usar a aproximação da derivada pela esquerda como uma aproximação para a derivada em geral (considerando $a[i] = 0$ para i negativo).

$$a^{'}[i] = F_s(a[i] - a[i-1]) \tag{1.24}$$

Essa forma é semelhante à forma de um filtro passa-altas de fase linear [4]:

$$y[n] = F_s x[n] - F_s x[n-1], \tag{1.25}$$

cuja função de transferência é

$$H(z) = F_s(1 - z^{-1}), \tag{1.26}$$

que corresponde a um ganho de:

$$|H(e^{j\omega})| = F_s\sqrt{2[1 - \cos\omega]} = 2F_s sen\frac{\omega}{2}, \tag{1.27}$$

onde

$$\omega = \frac{2\pi f}{F_s}. \tag{1.28}$$

A diferenciação é uma operação linear que pode ser considerada uma operação com um ganho de $2\pi f$, de acordo com a equação (1.14), e, pela equação (1.28), igual a $F_s\omega$; este é um ganho bem diferente do ganho prático, $|H(e^{jw})|$, principalmente para grandes valores

de ω. Essa diferença pode ser corrigida multiplicando-se o espectro de magnitude do sinal derivado obtido pela aproximação por um fator F definido como:

$$F(\omega) = \frac{\omega}{2sen\frac{\omega}{2}}. \tag{1.29}$$

Uma versão discreta para a obtenção da freqüência é dada pela equação:

$$f_p = \frac{\mathrm{DFT}^1[m_p]}{2\pi \mathrm{DFT}^0[m_p]}, \tag{1.30}$$

onde m_p é o índice do máximo na DFT^0 correspondente à freqüência f_p. Mais precisamente, m_p é o inteiro mais próximo a $f_p \frac{N}{F_s}$,sendo

$$m_p = \lfloor f_p \frac{N}{F_s} + 1/2 \rfloor \tag{1.31}$$

e

$$F_s/N(m_p - 1/2) \leq f_p < F_s/N(m_p + 1/2). \tag{1.32}$$

Se f_p não satisfizer tais condições, então a análise por DFT^1 falhou para tal freqüência, podendo indicar uma contaminação da raia espectral encontrada, ou seja, não há uma única freqüência nessa raia.

No caso discreto também é necessária uma correção da amplitude, que é feita pela fórmula:

$$a_p = \frac{a_p^0}{W(\mid f_P - f_p^0 \mid)}, \tag{1.33}$$

onde a_p^0 é o valor obtido através da DFT para a amplitude e f_p^0 é o valor obtido para a freqüência pela DFT.

1.6 Algoritmo de análise senoidal de sinal

Resumiremos agora o procedimento para a implementação do método baseado na DFT^1. Para cada bloco do sinal temporal, as seguintes operações devem ser feitas em seqüência:

- Aplicar o janelamento à função a.
- Obter a DFT^0.
- Computar a derivada do sinal original, $a^{'}$.

- Aplicar o mesmo janelamento à função $a^{'}$.
- Obter a DFT[1].
- Corrigir o espectro de magnitude pelo fator F.
- Para cada índice m referente a um máximo na DFT^0:
    1. Computar a freqüência exata, usando a correção.
    2. Computar a amplitude exata, usando a correção.
    3. Adicionar o par (freqüência,amplitude) à lista de resultados do bloco corrente.

1.7 Superposição de janelas

Além do tamanho e do tipo de janela utilizado, um outro parâmetro que deve ser escolhido é o tamanho do salto em janelas. Esse parâmetro determinará o avanço do *frame* no tempo, ou seja, a diferença no tempo (ou em amostras) entre o começo de cada segmento analisado. Sua importância reside em evitar descontinuidades na análise, que seriam geradas nos fins e inícios de cada janela.

Se escolhermos esse parâmetro igual a zero, o avanço será nulo e estaremos sempre analisando o mesmo *frame*. Se o escolhermos igual a um, significa que estamos avançando amostra a amostra, o que geraria um grande custo computacional.

Em geral, o espectro de sinais de áudio e de fala varia lentamente. Então, não se faz necessário o avanço amostra a amostra. Costuma-se associar o valor do avanço ao tamanho da janela. Por exemplo, para uma janela de tamanho N, utilizar avanços de $N/2$ amostras, o que corresponde a 50% de superposição entre amostras de janelas contíguas. De qualquer forma, o efeito das janelas superpostas deve ser tal que não distorça o sinal janelado, como mostra a figura 1.2.

1.8 Exemplos de aplicações - resultados obtidos

Trataremos agora de alguns exemplos do uso da representação freqüencial baseada na DFT[1]. Os exemplos foram implementados utilizando o programa MATLAB, com freqüência de amostragem igual a 44100 Hz e com amostras de 16 bits.

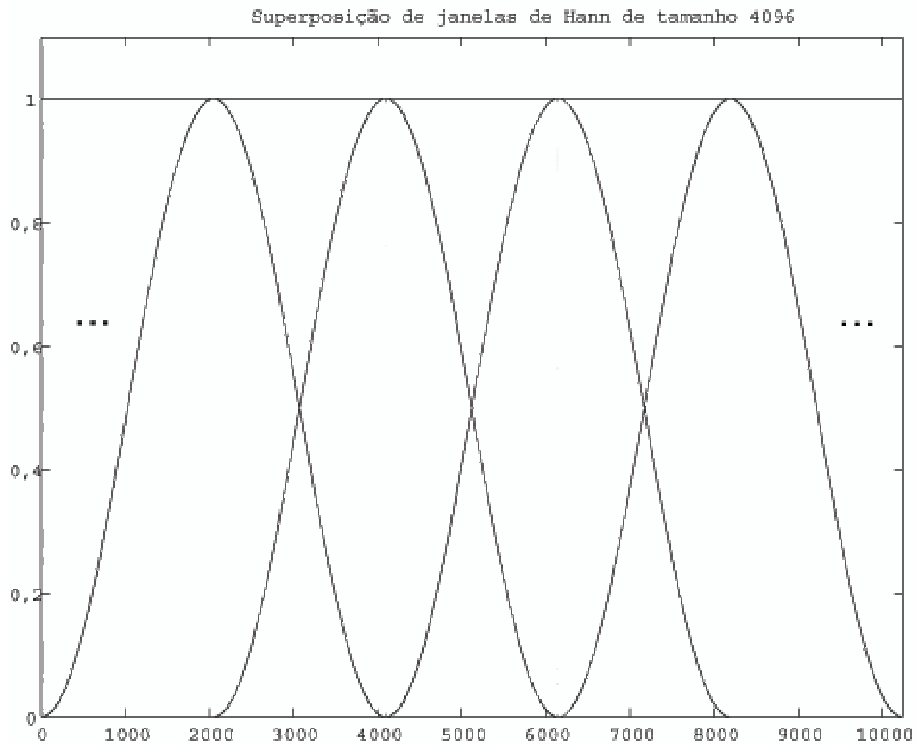

Figura 1.2: Janelas de Hann com superposição de 50% entre amostras de janelas contíguas.

1.8.1 Freqüência variável

O primeiro caso mostrado é o de uma senóide de amplitude 0,8 e de freqüência variável. Durante 5 segundos, a freqüência da senóide varia lentamente de 440Hz a 1660Hz. Iremos analisar a representação dessa senóide com a DFT^1 e a DFT^0, com janelas de 256, 512 e 1024 pontos e superposição de 50% entre blocos adjacentes.

Realizamos os testes com janela retangular (figuras 1.3 e 1.4) e com janelas de Hann (figuras 1.5 e 1.6).

Pudemos notar que, mesmo utilizando janela retangular ou janelas de poucos pontos (256), sempre obtivemos resultados consideravelmente melhores com o método da DFT^1. A DFT^1 conseguiu acompanhar as variações de freqüência do sinal com erro menor ao longo do tempo.

É possível também uma comparação entre os dois tipos de janela utilizados, Hann e retangular. A janela de Hann mostrou-se claramente mais apropriada que a retangular. Podemos verificar isso analisando os resultados, por exemplo, da DFT^1 para 256 e 512 pontos. Quando utilizamos a janela de Hann os resultados para 256, 512 e 1024 pontos foram bem próximos, permitindo, pelo uso de janelas mais curtas, favorecer a resolução no tempo.

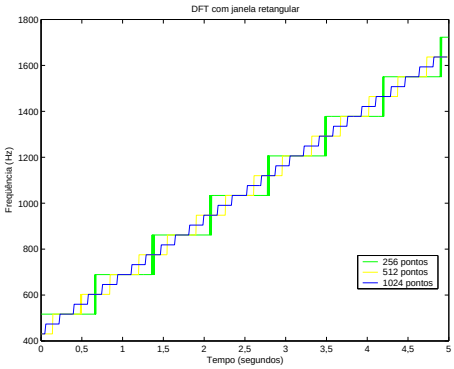

Figura 1.3: DFT com janelas retangulares de diferentes tamanhos.

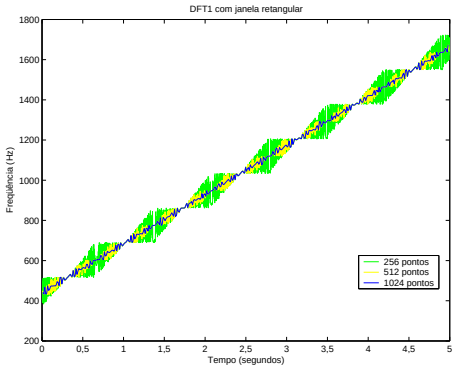

Figura 1.4: DFT[1] com janelas retangulares de diferentes tamanhos.

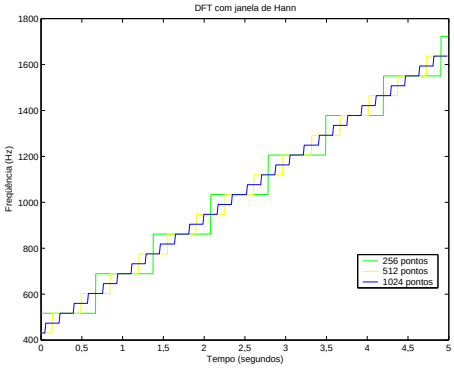

Figura 1.5: DFT com janelas de Hann de diferentes tamanhos.

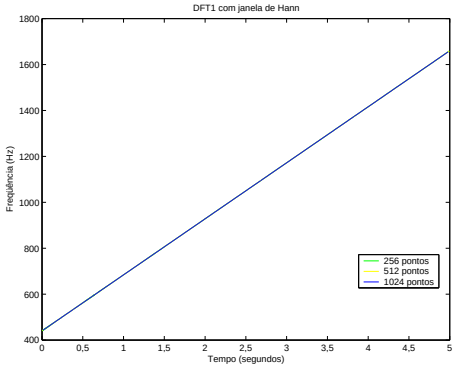

Figura 1.6: DFT[1] com janelas de Hann de diferentes tamanhos.

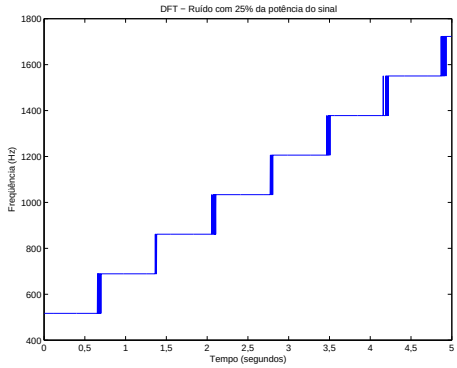

Figura 1.7: DFT com janela de Hann de tamanho 256, sinal contaminado com ruído com 25% da potência do sinal.

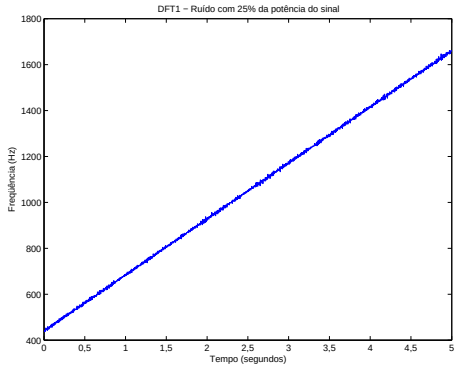

Figura 1.8: DFT[1] com janela de Hann de tamanho 256, sinal contaminado com ruído com 25% da potência do sinal.

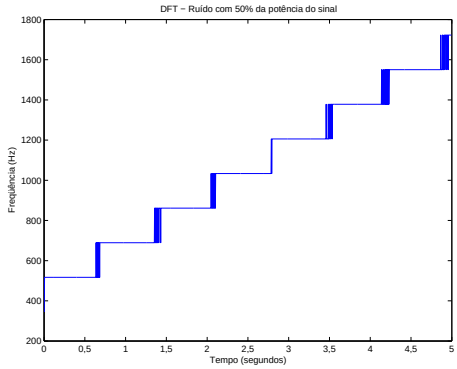

Figura 1.9: DFT com janela de Hann de tamanho 256, sinal contaminado com ruído com 50% da potência do sinal.

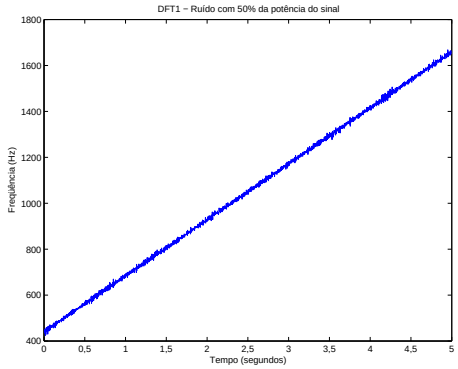

Figura 1.10: DFT[1] com janela de Hann de tamanho 256, sinal contaminado com ruído com 50% da potência do sinal.

1.8.2 Ruído

O segundo caso consta da mesma senóide do exemplo anterior acrescida de ruído branco. Utilizamos ruído branco gaussiano aditivo, variando a potência do ruído em relação à potência fixa do sinal.

Temos nesse caso um exemplo mais gritante de o quão é mais apropriado o uso do método da DFT[1] em relação ao uso simples da DFT. Utilizamos a janela de Hann, que, conforme o exemplo anterior, mostrou-se mais apropriada do que a retangular. Utilizamos ainda a janela com 256 pontos para especificar a diferença entre os resultados e superposição de 50% entre blocos adjacentes. Podemos ver os resultados nas figuras 1.7, 1.8, 1.9, 1.10, 1.11 e 1.12.

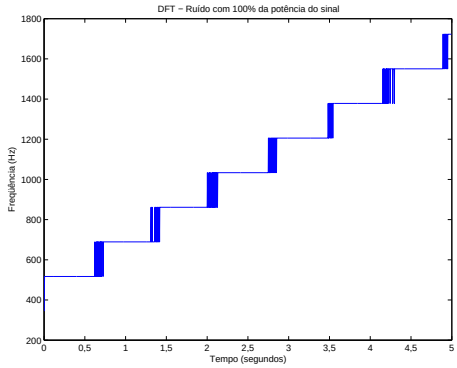

Figura 1.11: DFT com janela de Hann de tamanho 256, sinal contaminado com ruído com 100% da potência do sinal.

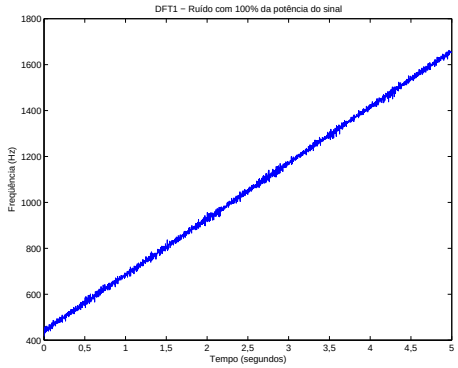

Figura 1.12: DFT^1 com janela de Hann de tamanho 256, sinal contaminado com ruído com 100% da potência do sinal.

1.8.3 Vibrato e trêmolo

Vibrato é uma pequena variação periódica da freqüência de uma nota musical no tempo. Trêmolo é uma pequena variação periódica da amplitude de uma nota musical no tempo.

Passamos então a analisar o comportamento do método de DFT^1 na presença de vibrato, comparando-o com o comportamento do método de DFT. Tanto o vibrato quanto o trêmolo violam as condições de sinal estacionário (o primeiro quanto à freqüência e o segundo quanto à amplitude).

O comportamento de tal método foi estudado usando um oscilador senoidal cuja freqüência é de 200 Hz modulada por vibrato, enquanto sua amplitude permanece constante e igual a 1. Foi usada a janela de Hann, com 512 pontos e superposição de 50% entre blocos adjacentes. Apresentaremos a tabela de erros da freqüência (tabela 1.1), com erros médios percentuais e desvios-padrão entre parênteses, para determinados valores de profundidade e taxa de variação, e gráficos ilustrativos (figuras 1.13, 1.14, 1.15, 1.16, 1.17 e 1.18).

Pudemos notar que o método de DFT^1 apresenta resultados bem superiores ao de DFT enquanto a taxa de variação e a profundidade se mantêm baixos. Conforme esses valores vão aumentando, chega-se a uma paridade. Podemos ver que com valores muito elevados ambos os métodos tornam-se imprecisos, como se tais métodos fossem se corrompendo.

A análise do método frente a um sinal com o efeito de trêmolo será feita utilizando um oscilador senoidal cuja freqüência é constante e igual a 2000 Hz, enquanto que sua amplitude é de 0.5 modulada por um trêmolo. Exibiremos gráficos demonstrativos (figuras 1.19, 1.20,

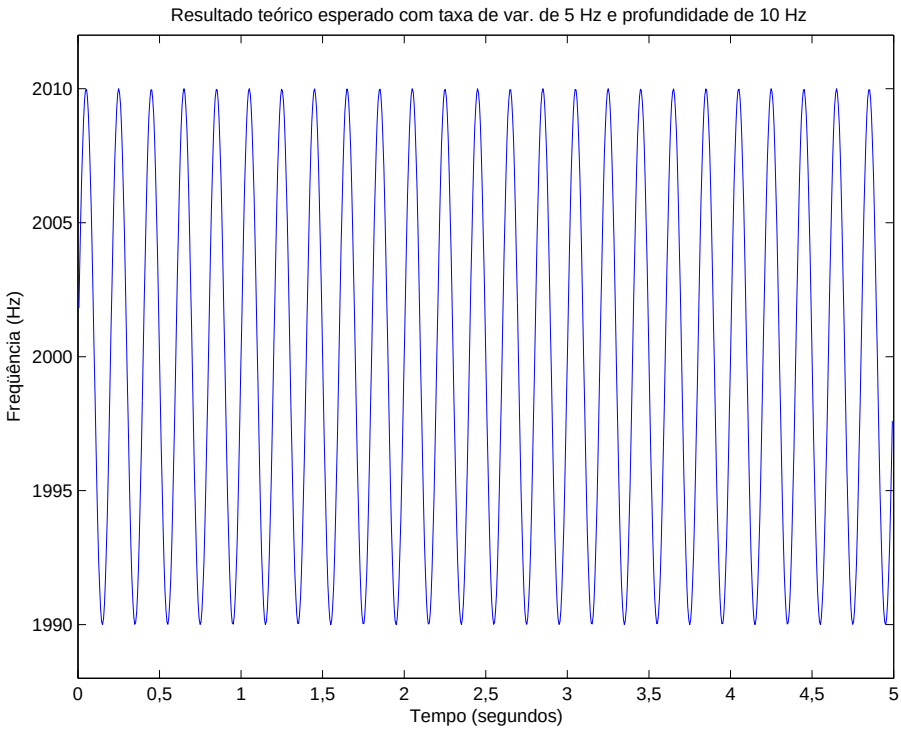

Figura 1.13: Resultado teórico com taxa de variação de 5 Hz e profundidade de 10 Hz.

1.21, 1.22, 1.23 e 1.24) e uma tabela (1.2) com os erros médios e o desvio-padrão percentuais para amplitude, com determinados valores de profundidade e taxa de variação. Foi usada novamente janela de Hann com 512 pontos e superposição de 50% entre blocos adjacentes.

Observamos que a DFT^1 fornece resultados melhores que os da DFT, principalmente para baixos valores de profundidade e taxa de variação. Enquanto a taxa e a profundidade se mantêm com valores razoáveis, o método da DFT^1 prova ser de grande utilidade, por sua maior precisão, corrigindo os valores obtidos pela DFT. Ou seja, para sinais dentro de faixas limitadas de taxa de variação e profundidade, como ocorre com boa parte dos sinais naturais, temos na DFT^1 uma boa ferramenta.

1.8.4 Exemplo com sinal musical

Esta técnica será incorporada a um sistema completo de análise e síntese descrito no capítulo 3. Por ora, apresentamos o resultado de sua aplicação a um sinal musical real tomando como exemplo um trecho de 12 segundos da gravação de "Bachianas Brasileiras número 6", de Villa-Lobos. As figuras 1.25 e 1.26 ilustram o experimento. Temos na figura 1.27 o espectograma do sinal. Podemos identificar na figura 1.26 as linhas freqüenciais (linhas

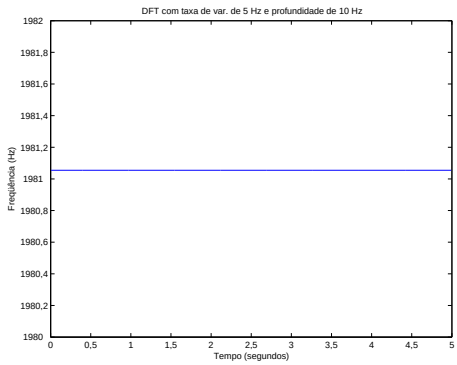

Figura 1.14: DFT com janela de Hann de tamanho 512, taxa de variação de 5 Hz e profundidade de 10 Hz.

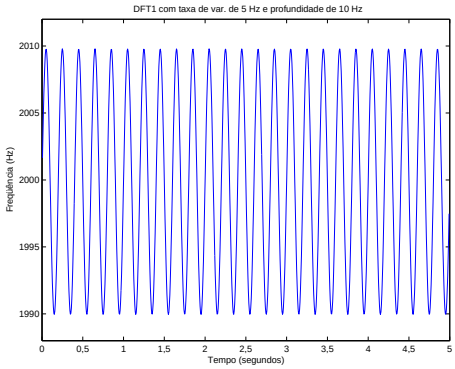

Figura 1.15: DFT[1] com janela de Hann de tamanho 512, taxa de variação de 5 Hz e profundidade de 10 Hz.

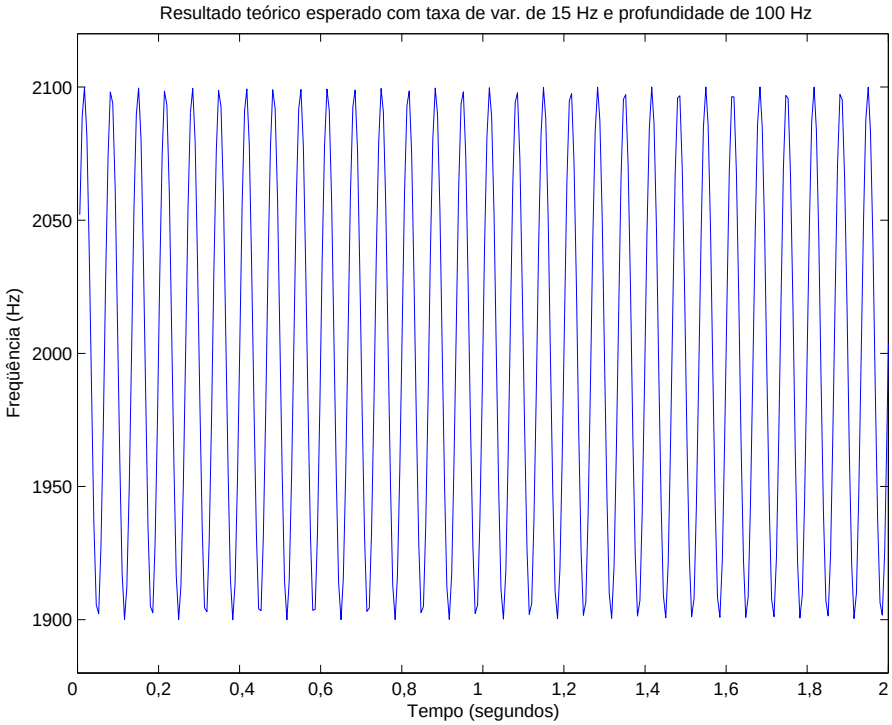

Figura 1.16: Resultado teórico com taxa de variação de 15 Hz e profundidade de 100 Hz.

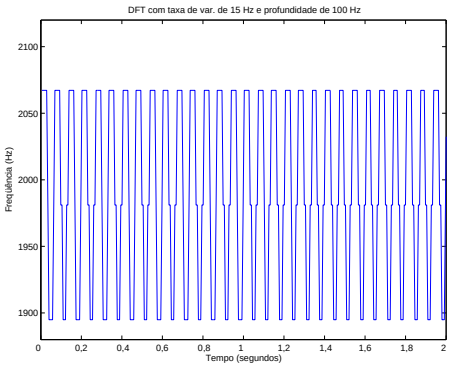

Figura 1.17: DFT com janela de Hann de tamanho 512, taxa de variação de 15 Hz e profundidade de 100 Hz.

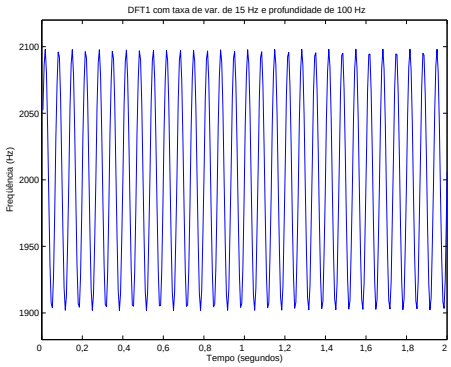

Figura 1.18: DFT^1 com janela de Hann de tamanho 512, taxa de variação de 15 Hz e profundidade de 100 Hz.

Tabela 1.1: Erros percentuais de freqüência para diferentes taxas de variação (primeira coluna) e profundidade (primeira linha): média (desvio-padrão).

	0 Hz	10 Hz	100 Hz	500 Hz	1000 Hz
DFT	Erros em %	Erros em %	Erros em %	Erros em %	Erros em %
5 Hz	-0,94(1×10^{-14})	-0,95(0,35)	-0,37(1,12)	-0,009(1,38)	0,21(1,54)
10 Hz	-0,94(1×10^{-14})	-0,95(0,35)	-0,36(1,12)	-0,57(1,38)	0,20(1,55)
15 Hz	-0,94(1×10^{-14})	-0,95(0,35)	-0,36(1,12)	-0,28(1,44)	0,07(1,64)
20 Hz	-0,94(1×10^{-14})	-0,95(0,35)	-0,34(1,13)	-0,28(1,43)	-0,16(1,97)
25 Hz	-0,94(1×10^{-14})	-0,95(0,35)	-0,33(1,14)	-0,22(1,52)	-0,49(2,68)
DFT^1					
5 Hz	-0,006$(1,9 \times 10^{-4})$	-0,006(0,003)	-0,002(0,0018)	0,003(0,13)	0,028(0,43)
10 Hz	-0,006$(1,9 \times 10^{-4})$	-0,006(0,005)	-0,001(0,04)	0,012(0,41)	0,07(1,04)
15 Hz	-0,006$(1,9 \times 10^{-4})$	-0,006(0,01)	$-2,53 \times 10^{-4}$(0,08)	0,036(0,74)	-0,006(1,40)
20 Hz	-0,006$(1,9 \times 10^{-4})$	-0,006(0,014)	0,0015(0,14)	0,053(1,00)	-0,14(1,90)
25 Hz	-0,006$(1,9 \times 10^{-4})$	-0,006(0,02)	0,004(0,20)	0,03(1,26)	-0,38(2,80)

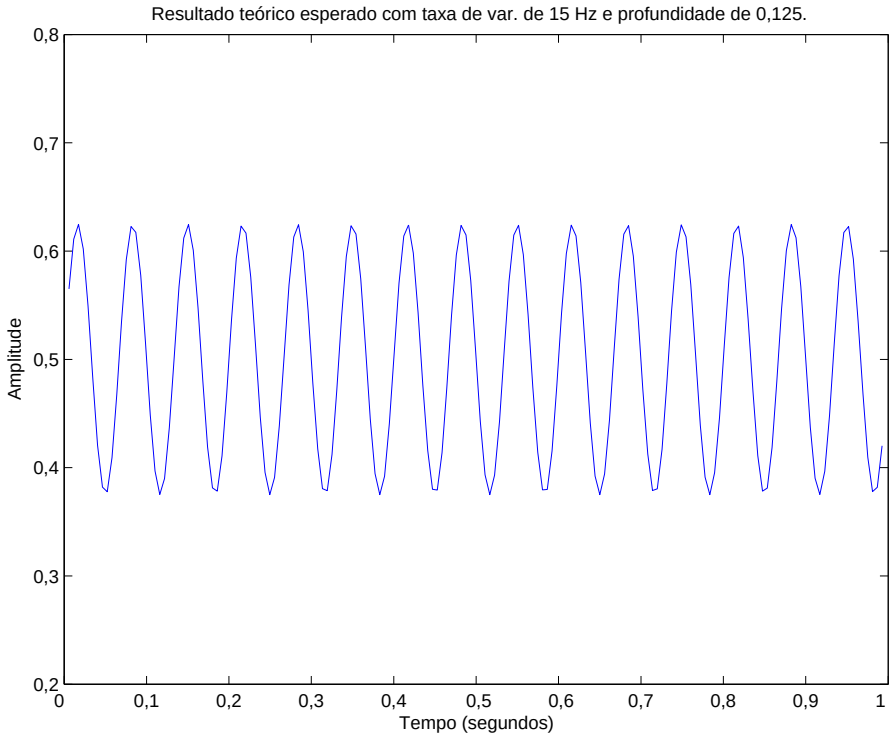

Figura 1.19: Resultado teórico com taxa de variação de 15 Hz e profundidade de 0,125.

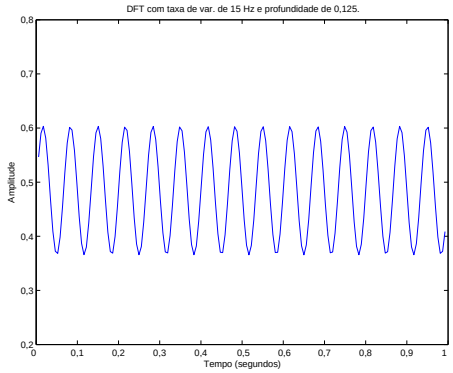

Figura 1.20: DFT com janela de Hann de tamanho 512, taxa de variação de 15 Hz e profundidade de 0,125.

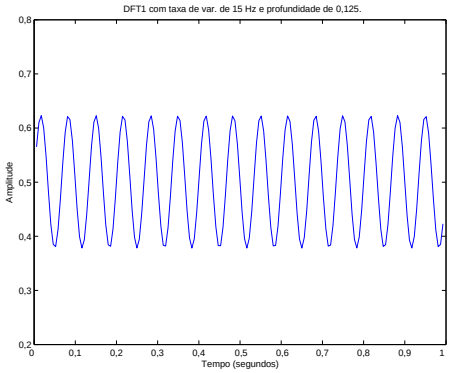

Figura 1.21: DFT[1] com janela de Hann de tamanho 512, taxa de variação de 15 Hz e profundidade de 0,125.

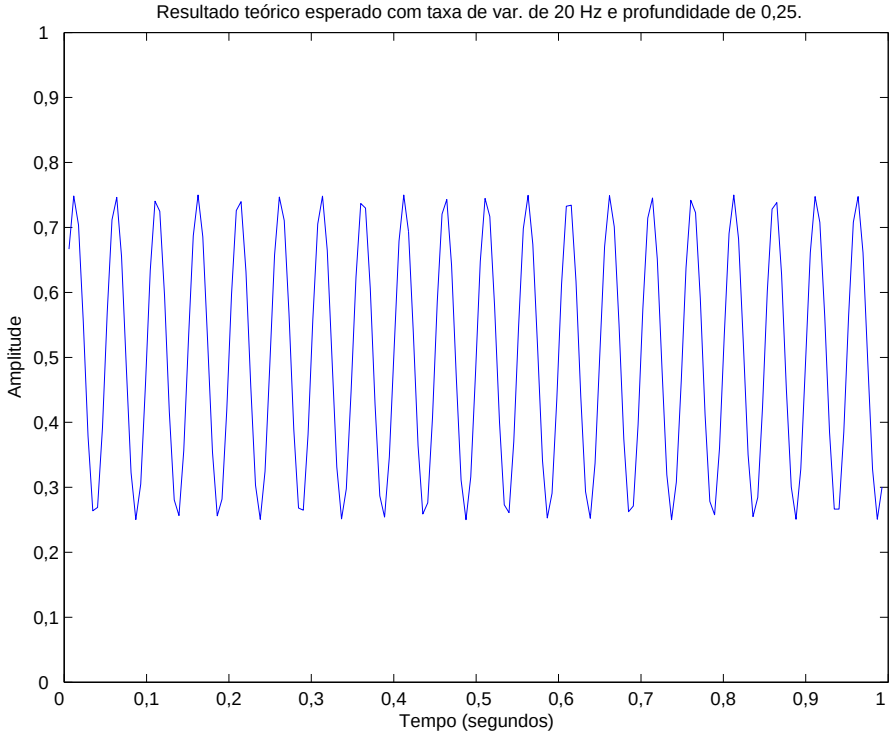

Figura 1.22: Resultado teórico com taxa de variação de 20 Hz e profundidade de 0,25.

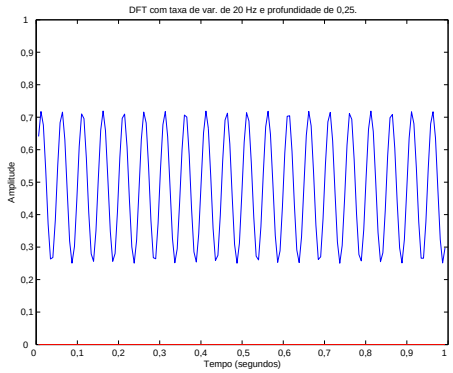

Figura 1.23: DFT com janela de Hann de tamanho 512, taxa de variação de 20 Hz e profundidade de 0,25 Hz.

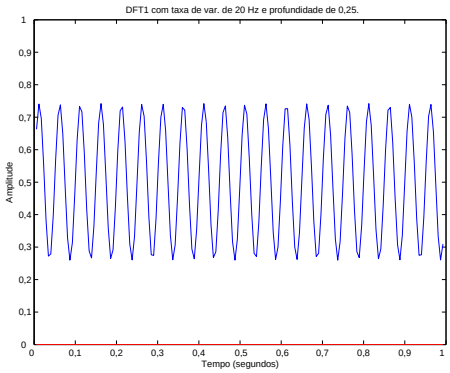

Figura 1.24: DFT[1] com janela de Hann de tamanho 512, taxa de variação de 20 Hz e profundidade de 0,25.

Tabela 1.2: Erros percentuais de amplitude para diferentes taxas de variação (primeira coluna) e profundidade (primeira linha): média (desvio-padrão).

	0	0.125	0.250	0.375
DFT	Erros em %	Erros em %	Erros em %	Erros em %
5 Hz	-3,07($9,1 \times 10^{-5}$)	-3,06(0,038)	-3,03(0,09)	-2,96(0,22)
10 Hz	-3,07($9,1 \times 10^{-5}$)	-3,04(0,15)	-2,94(0,37)	-2,63(0,88)
15 Hz	-3,07($9,1 \times 10^{-5}$)	-3,007(0,34)	-2,77(0,82)	-2,09(1,97)
20 Hz	-3,07($9,1 \times 10^{-5}$)	-2,96(0,59)	-2,55(1,45)	-1,34(3,48)
25 Hz	-3,07($9,1 \times 10^{-5}$)	-2,90(0,92)	-2,26(2,25)	-0,39(5,40)
DFT^1				
5 Hz	0,15(0,001)	0,16(0,057)	0,19(0,13)	0,27(0,29)
10 Hz	0,15(0,001)	0,19(0,19)	0,29(0,45)	0,62(1,05)
15 Hz	0,15(0,001)	0,22(0,41)	0,47(0,98)	1,20(2,31)
20 Hz	0,15(0,001)	0,27(0,71)	0,71(1,71)	2,01(4,06)
25 Hz	0,15(0,001)	0,34(1,09)	1,02(2,64)	3,06(6,32)

horizontais) dos instrumentos, formada pelas fundamentais e seus harmônicos.

1.9 Conclusões

Pudemos verificar, como pretendíamos, que realmente o método da DFT^1 pode ser de grande utilidade na análise espectral de sinais, apresentando uma grande melhora em relação ao método da DFT. Com a DFT^1 obtivemos resultados bem precisos, tanto para freqüência quanto para amplitude. O método da DFT^1 mostrou-se bastante útil ao tentarmos otimizar o compromisso tempo *versus* freqüência, visto que, mesmo com pequenas janelas, fornece bons resultados para freqüência e amplitude, permitindo resolução melhor no domínio do tempo. O comportamento do método da DFT^1 mostrou-se muito bom em presença de ruído e bem satisfatório com vibrato e trêmolo.

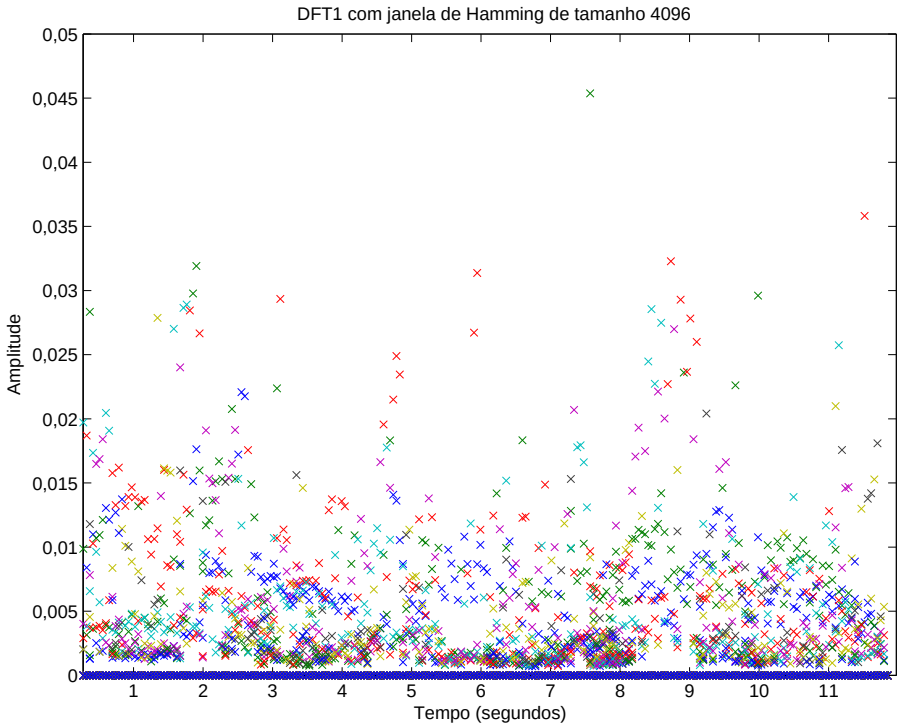

Figura 1.25: Amplitudes obtidas com o método da DFT[1] para um trecho de sinal musical.

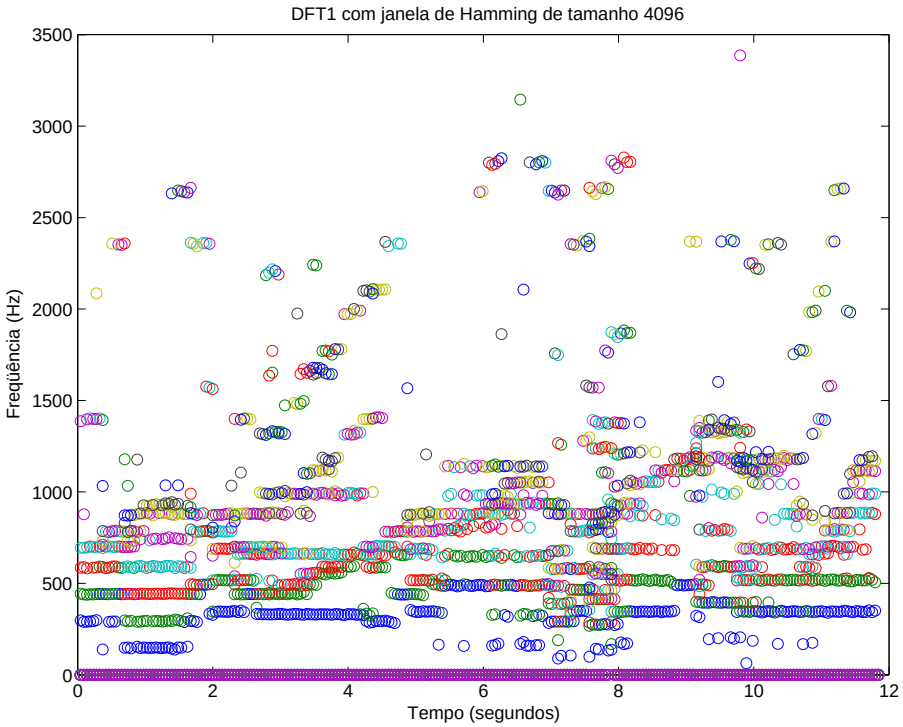

Figura 1.26: Freqüências obtidas com o método da DFT[1] para um trecho de sinal musical.

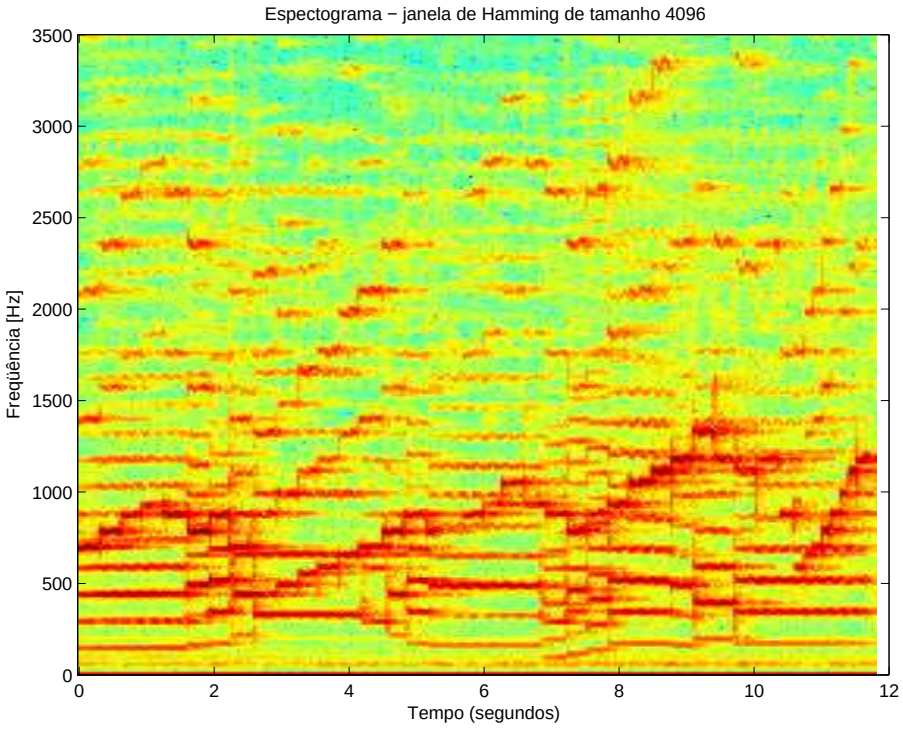

Figura 1.27: Espectograma de um trecho de sinal musical.

Capítulo 2

Distinção de picos espectrais superpostos em um modelo senoidal de duas vozes

2.1 Introdução

A separação de "duas vozes" aborda a separação de sinais provenientes de duas fontes sonoras, como duas pessoas falando simultaneamente ou um dueto musical. Atribuindo-se um modelo senoidal a cada um dos sinais a separar, é possível lançar mão de características como a harmonicidade para tentar agrupar as componentes "coerentes" que pertencerem a cada um deles. Mesmo nessas condições, um problema ocorre freqüentemente: a "mistura" de picos diferentes no domínio da freqüência devido ao janelamento do sinal.

Em [5], aborda-se esse problema. No presente capítulo, o método algébrico lá apresentado é descito com detalhes, implementado e avaliado.

2.2 Modelo de duas vozes

Como visto no capítulo 1, de acordo com o teorema de Fourier, qualquer função periódica pode ser modelada como uma soma de sinais senoidais de freqüências harmônicas, com amplitudes e fases próprias. A transformada de Fourier converte um sinal genérico temporal (amplitude *versus* tempo) em uma representação freqüencial (módulo e fase *versus* freqüência). Ela realiza a decomposição do sinal em infinitas componentes freqüenciais e o

representa na forma de uma densidade espectral, estacionária. Para analisarmos um sinal cuja composição espectral evolui no tempo, lançamos mão de uma técnica conhecida como "de curta duração", ou seja, que analisa pequenos trechos temporais. Cada trecho desses é tratado como um sinal estacionário, e depois juntam-se todos os trechos, compondo um espectro que evolui no tempo. Para cada bloco com amostras consecutivas de $x[n]$, a transformada de Fourier é computada. Isso equivale a ter pré-multiplicado a seqüência original por uma função de janelamento $w[n] = 1$. Essa janela retangular deforma o espectro do sinal original, sendo, em geral, substituída por janelas suavizadoras. O fato de se multiplicar o sinal por uma janela leva a convoluir os espectros do sinal e da janela. Isso tende a piorar a resolução freqüencial da transformada de Fourier. O uso de janelas mais longas poderia minorar esse problema, mas isso requereria uma estacionariedade longa do sinal, o que em geral não é satisfeito pela aplicação em vista. Temos na figura 2.1 o esquema representativo da análise por Fourier.

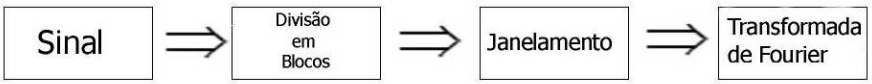

Figura 2.1: Esquema representativo da análise por Fourier.

Podemos facilmente generalizar a modelagem senoidal de uma voz para o caso de duas vozes. É possível representar um sinal gerado por duas vozes simultâneas como uma soma de dois conjuntos de sinais senoidais, com amplitudes, fases e freqüências variantes no tempo:

$$x[n] = x_a[n] + x_b[n], \tag{2.1}$$

onde

$$x_a[n] = \sum_{l=1}^{L_a} a_l[n] \cos\left[\Theta_{a,l}[n]\right] \tag{2.2}$$

e

$$x_b[n] = \sum_{l=1}^{L_b} b_l[n] \cos\left[\Theta_{b,l}[n]\right]. \tag{2.3}$$

Se a excitação é periódica, um modelo harmônico de duas vozes pode ser usado, no qual as freqüências associadas com a voz a e a voz b são múltiplas de duas freqüências fundamentais, $\omega_a[n]$ e $\omega_b[n]$, respectivamente. Para o caso estacionário, no qual a excitação e as características do sistema são considerados fixos durante o intervalo de tempo da análise,

podemos considerar:

$$x_a[n] = \sum_{l=1}^{L_a} a_l \cos\left[\omega_{a,l} n + \phi_{a,l}\right] \tag{2.4}$$

e

$$x_b[n] = \sum_{l=1}^{L_b} b_l \cos\left[\omega_{b,l} n + \phi_{b,l}\right]. \tag{2.5}$$

Para se obter uma representação acurada das duas vozes, devem-se escolher adequadamente as quantidades de senóides, L_a e L_b. O tamanho da janela de análise também é muito importante para resolver picos muito próximos, principalmente em baixa freqüência.

Um sinal de áudio pode ser considerado estacionário em uma janela de até 20ms; isso limita o tamanho da janela e, conseqüentemente, a sua resolução freqüencial. Para aumentar a resolução espectral inserem-se zeros adicionais na janela. E para garantir que a informação espectral pertence ao instante correto, foi utilizado o método de janelamento de fase nula. Maiores informações a respeito dessa técnica podem ser encontradas em [6].

2.3 O problema de dois picos próximos na freqüência

Uma primeira idéia simples e intuitiva para realizar a separação de duas vozes é, sendo válida a hipótese de que um dos sinais componentes tem uma intensidade (volume) maior que o outro, considerar todas as amplitudes maiores pertencendo a uma voz. Escolhendo-se, então, os picos de maior amplitude, obter-se-iam assim as informações necessárias (freqüência, amplitude e fase) para a reconstrução dessa voz separadamente. Finalmente, obter-se-ia a segunda voz pela subtração da primeira voz reconstruída do sinal misturado. Temos o esquema dessa separação na figura 2.2. O grande problema desse algoritmo está na suposição de que os picos de maior amplitude pertencem a uma das vozes separadamente.

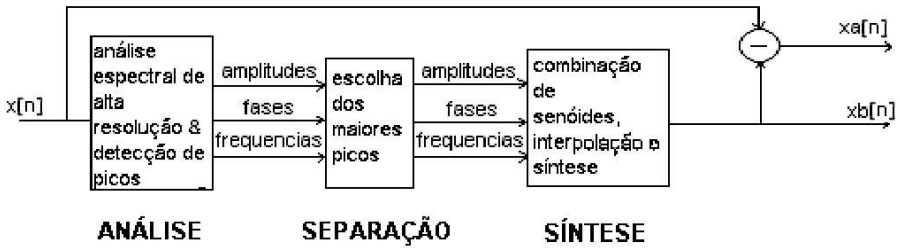

Figura 2.2: Esquema de separação de 2 vozes por diferença de amplitude.

Mas, mesmo antes desse problema, é preciso resolver a ocorrência de outro. Como mencionado anteriormente, a escolha da janela é fundamental para se separar o sinal. Porém, há casos em que dois picos, um pertencente a cada voz, estão muito próximos (na freqüência), fazendo com que a análise de Fourier os apresente como apenas um pico, com freqüência intermediária, ao invés de dois picos separados.

Um exemplo dessa mistura pode ser observado na figura 2.3, onde os picos referentes a duas senóides de freqüências diferentes se misturam devido à janela de análise utilizada, $W(\omega)$. Elas se somam porque as freqüências das senóides A e B, ω_1 e ω_2, respectivamente, estão muito próximas. $S(\omega_1)$ e $S(\omega_2)$ são as transformadas de Fourier do sinal misturado nas freqüências conhecidas. $S_a(\omega_2)$ é a amplitude do sinal A na freqüência do sinal B e $S_b(\omega_1)$ é a transformada de Fourier do sinal B na freqüência do sinal A. Será mostrado um algoritmo que corrige a contribuição de cada componente freqüencial, sabendo-se *a priori* as freqüências presentes no sinal, o que em algumas aplicações é perfeitamente razoável.

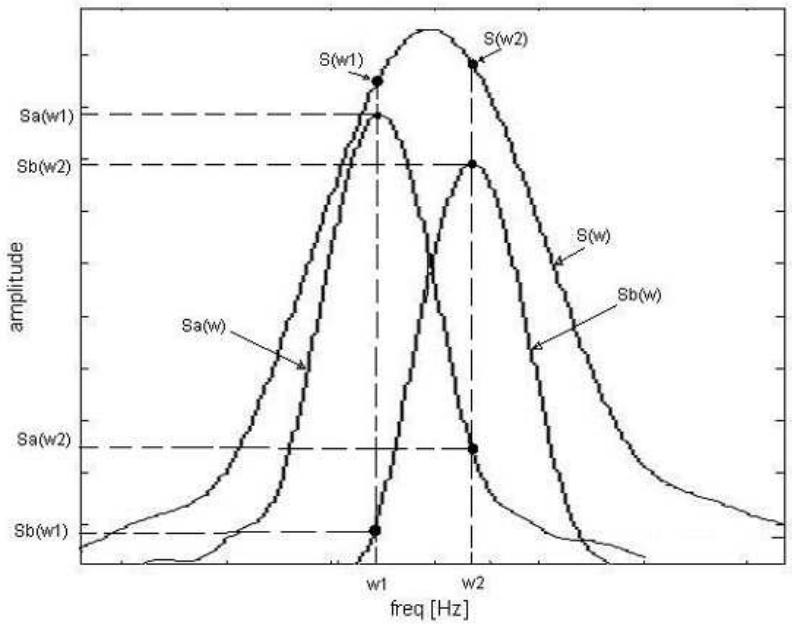

Figura 2.3: Mistura dos picos referentes a duas freqüências em um único pico.

2.4 Possível solução

Pode-se estimar a contribuição de cada componente sabendo-se a priori suas respectivas freqüências, o que em algumas aplicações é possível. Por exemplo, pode-se desejar separar

dois harmônicos cujas duas fundamentais foram pré-determinadas. O método sugerido em [5] será descrito aqui.

Aplicando o janelamento ao sinal composto:

$$x_w[n] = w[n](x_a[n] + x_b[n]) = \tag{2.6}$$

$$= w[n](\sum_{l=1}^{La} a_l \cos[\omega_{a,l} n + \phi_{a,l}] + \sum_{l=1}^{Lb} b_l \cos[\omega_{b,l} n + \phi_{b,l}]) = \tag{2.7}$$

$$= w[n](\sum_{l=1}^{La} a_l \frac{e^{j(\omega_{a,l} n + \phi_{a,l})} + e^{-j(\omega_{a,l} n + \phi_{a,l})}}{2} + \sum_{l=1}^{Lb} b_l \frac{e^{j(\omega_{b,l} n + \phi_{b,l})} + e^{-j(\omega_{b,l} n + \phi_{b,l})}}{2}) = \tag{2.8}$$

$$= \frac{1}{2}\sum_{l=1}^{La} a_l e^{j\phi_{a,l}} w[n] e^{j\omega_{a,l} n} + \frac{1}{2}\sum_{l=1}^{La} a_l e^{-j\phi_{a,l}} w[n] e^{-j\omega_{a,l} n} + \tag{2.9}$$

$$+ \frac{1}{2}\sum_{l=1}^{Lb} b_l e^{j\phi_{b,l}} w[n] e^{j\omega_{b,l} n} + \frac{1}{2}\sum_{l=1}^{Lb} b_l e^{-j\phi_{b,l}} w[n] e^{-j\omega_{b,l} n}. \tag{2.10}$$

Aplicando agora a transformada de Fourier, temos:

$$X_w(\omega) = \frac{1}{2}\sum_{l=1}^{La} a_l [W(\omega - \omega_{a,l}) e^{j\phi_{a,l}} + W(\omega + \omega_{a,l}) e^{-j\phi_{a,l}}] + \tag{2.11}$$

$$+ \sum_{l=1}^{Lb} b_l [W(\omega - \omega_{b,l}) e^{j\phi_{b,l}} + W(\omega + \omega_{b,l}) e^{-j\phi_{b,l}}]. \tag{2.12}$$

Por simplicidade, podemos considerar que cada sinal é composto por apenas uma senóide. Assim, temos:

$$X_w(\omega) = \frac{1}{2}(a[W(\omega - \omega_a) e^{j\phi_a} + W(\omega + \omega_a) e^{-j\phi_a}] + b[W(\omega - \omega_b) e^{j\phi_b} + W(\omega + \omega_b) e^{-j\phi_b}]). \tag{2.13}$$

Fazendo $\omega = \omega_a$ e $\omega = \omega_b$, temos, respectivamente:

$$X_w(\omega_a) = \frac{1}{2}(a[W(0) e^{j\phi_a} + W(2\omega_a) e^{-j\phi_a}] + b[W(\omega_a - \omega_b) e^{j\phi_b} + W(\omega_a + \omega_b) e^{-j\phi_b}]) \text{ e} \tag{2.14}$$

$$X_w(\omega_b) = \frac{1}{2}(a[W(\omega_b - \omega_a) e^{j\phi_a} + W(\omega_b + \omega_a) e^{-j\phi_a}] + b[W(0) e^{j\phi_b} + W(2\omega_b) e^{-j\phi_b}]). \tag{2.15}$$

Agora consideramos que $2\omega_a$, $2\omega_b$ e $(\omega_a + \omega_b)$ estão longe da região de espectro com potência significativa e, portanto, $W(2\omega_a)$, $W(2\omega_b)$ e $W(\omega_a + \omega_b)$ podem ser consideradas nulas. Temos, então:

$$X_w(\omega_a) \simeq \frac{aW(0) e^{j\phi_a} + bW(\omega_a - \omega_b) e^{j\phi_b}}{2} \text{ e} \tag{2.16}$$

$$X_w(\omega_b) \simeq \frac{aW(\omega_b - \omega_a) e^{j\phi_a} + bW(0) e^{j\phi_b}}{2}. \tag{2.17}$$

Chegamos assim ao seguinte sistema de equações, apresentado aqui na forma matricial:

$$\frac{1}{2}\begin{bmatrix} W(0) & W(w_a - w_b) \\ W(w_b - w_a) & W(0) \end{bmatrix}\begin{bmatrix} ae^{j\phi_a} \\ be^{j\phi_b} \end{bmatrix} = \begin{bmatrix} X_w(\omega_a) \\ X_w(\omega_b) \end{bmatrix}. \tag{2.18}$$

A solução

$$\begin{pmatrix} ae^{j\phi_a} \\ be^{j\phi_b} \end{pmatrix} = 2\begin{bmatrix} W(0) & W(w_a - w_b) \\ W(w_b - w_a) & W(0) \end{bmatrix}^{-1}\begin{bmatrix} X_w(\omega_a) \\ X_w(\omega_b) \end{bmatrix} \tag{2.19}$$

determina as amplitudes e fases de cada parcial.

É possível estender esse método para o caso de mais lóbulos misturados [5].

2.5 Exemplo

Examinamos o caso onde tentamos separar de uma mistura o quarto harmônico de Mi_4 (sinal X_a) do quinto harmônico de $Dó_4$ (sinal X_b), na escala temperada 1.318,5 Hz e 1.308,1 Hz, respectivamente. Utilizamos uma freqüência de amostragem de 44,1 kHz, blocos de 1024 amostras, janela de Hanning e superposição de 50% das amostras de blocos adjacentes.

Temos em 2.1 uma tabela com os experimentos realizados e os resultados obtidos. Testamos os casos em que os dois sinais têm a mesma amplitude e a mesma fase (caso I), os dois sinais com mesma fase e amplitudes diferentes (caso II), os dois sinais com mesma amplitude e fases diferentes (caso III) e os dois sinais com amplitudes e fases diferentes (caso IV).

Obtivemos, para o caso I, resultados com erros de 0,0108 e 0,0091 para as amplitudes encontradas e 0,0019 rad e 0,0062 rad para as fases. No caso II, obtivemos resultados com erros de 0,0080 e 0,0074 para as amplitudes e 0,0025 rad e 0,0074 rad para as fases. No caso III, obtivemos resultados com erros de 0,0040 e 0,0046 para as amplitudes e 0,0024 rad e 0,0007 rad para as fases. Finalmente, no caso IV, obtivemos resultados com erros de 0,0046 e 0,0029 para as amplitudes e 0,0003 rad e 0,0065 rad para as fases.

Apresentamos também os resultados nas figuras 2.4 a 2.13 para o Caso I da tabela 2.1.

Tabela 2.1: Resultados encontrados para o método de separação de sinal usando um modelo de duas vozes.

	Caso I	Caso II	Caso III	Caso IV
Sinal x_a	$\cos(\omega_a t + \pi/2)$	$\cos(\omega_a t + \pi/2)$	$\cos(\omega_a t + \pi/2)$	$\cos(\omega_a t + \pi/2)$
Sinal x_b	$\cos(\omega_b t + \pi/2)$	$\frac{1}{2}\cos(\omega_b t + \pi/2)$	$\cos \omega_b t$	$\frac{1}{2}\cos \omega_b t$
Módulo Encontrado para X_a	0,9892	0,9920	0,9960	0,9954
Módulo Encontrado para X_b	1,0091	0,5074	0,9954	0,4971
Fase Encontrada para X_a	1,5727 rad	1,5733 rad	1,5683 rad	1,5711 rad
Fase Encontrada para X_b	1,5770 rad	1,5782 rad	0,0007 rad	0,0065 rad

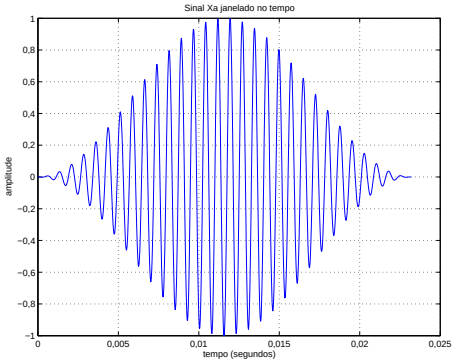

Figura 2.4: Quarto harmônico de Mi janelado no tempo.

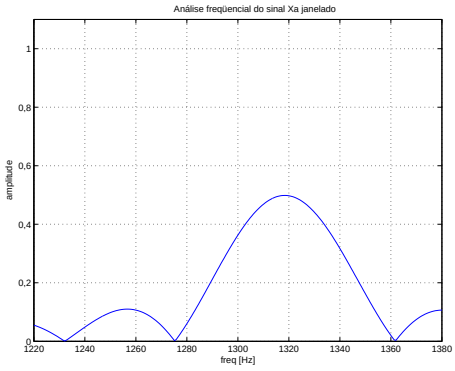

Figura 2.5: Quarto harmônico de Mi janelado.

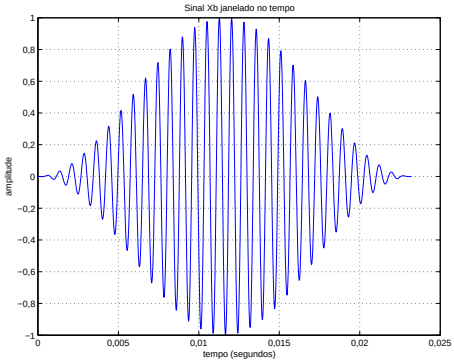

Figura 2.6: Quinto harmônico de Dó janelado no tempo.

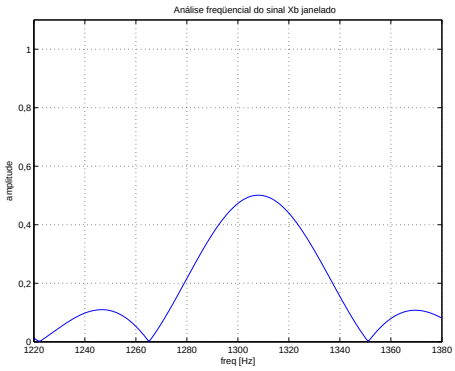

Figura 2.7: Quinto harmônico de Dó janelado.

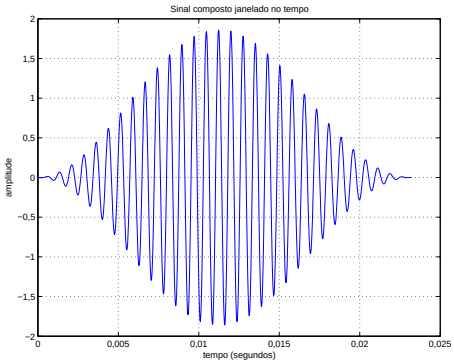

Figura 2.8: Sinal composto janelado no tempo.

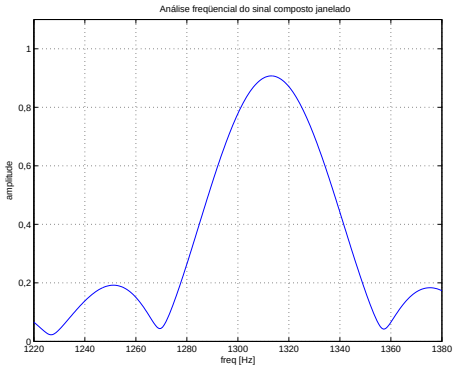

Figura 2.9: Sinal composto janelado.

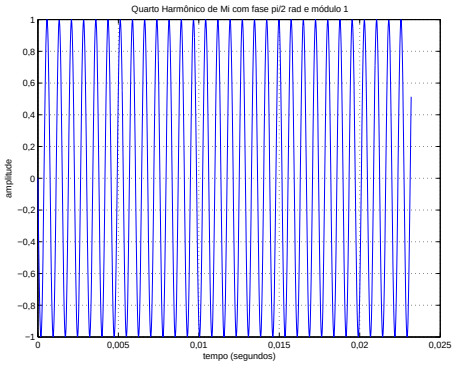

Figura 2.10: Quarto harmônico de Mi.

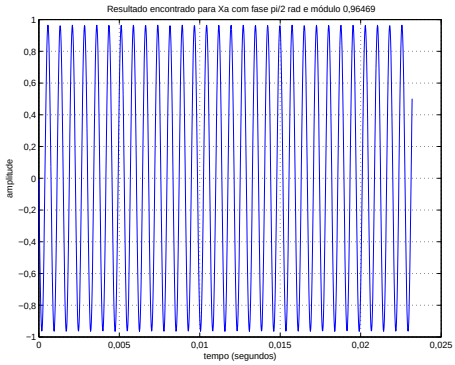

Figura 2.11: Resultado: quarto harmônico de Mi separado do sinal composto.

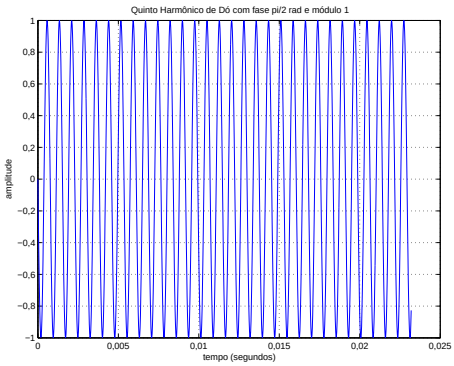

Figura 2.12: Quinto harmônico de Dó.

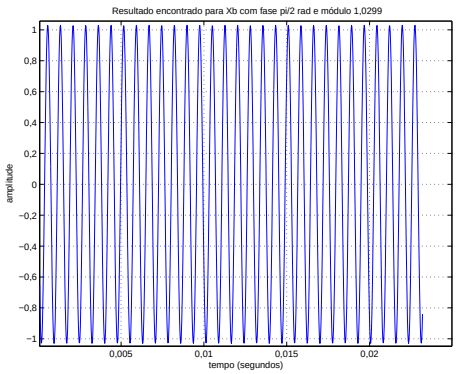

Figura 2.13: Resultado: quinto harmônico de Dó separado do sinal composto.

2.6 Conclusões

Obtivemos excelentes resultados com a implementação do método proposto, que mostrou ser realmente eficiente para separar dois sinais com freqüências próximas, considerando que já se saibam, *a priori*, as freqüências envolvidas.

Capítulo 3

Separação e ressíntese de sinais

3.1 Introdução

Com o advento de novas técnicas e tecnologias em áudio digital e sua utilização em massa por nossa sociedade, aquele que escuta uma gravação tem uma expectativa de grande qualidade, ao menos no aspecto técnico, o que dificulta a aceitação de sons que, na época em que foram gravados, não tinham como receber o tratamento dado a qualquer gravação de áudio de hoje em dia. Os mais diversos métodos de recuperação não se mostraram ainda suficientes para dar a essas gravações uma qualidade compatível com os ouvidos modernos. Um outro caminho a se seguir seria, então, adquirir informações da gravação precária para depois tentar reproduzir o som original artificialmente.

Uma polêmica cerca essa idéia. Alguns resistem a ela alegando que tal técnica tiraria o valor artístico e, por que não, histórico de tais gravações. Há ainda os mais perfeccionistas, que consideram o método insuficiente, pois não recuperaria o som totalmente, apenas utilizaria parâmetros, gerando sempre algo diferente do original. Mas temos também o lado cultural. Seria bem mais fácil fazer esses sons chegarem aos ouvidos modernos e serem aceitos, dando grande contribuição à formação cultural dos indivíduos.

O estudo dessas técnicas de "enriquecimento" consiste genericamente em decompor o som em partes constituintes e parametrizar as características de interesse de forma inambígua a fim de permitir estudar as partes e parâmetros e/ou processá-los independentemente e/ou ressintetizar o som ou sintetizar novo som. As ferramentas mais usuais em análise e síntese são representações em tempo-freqüência, modelos estatísticos, redes neurais etc.

Técnicas de análise e síntese de áudio estão por trás de um sem-número de aplicações.

Podemos utilizá-las na remixagem e na edição de sinais de áudio. Também encontram aplicação na transcrição musical automática, ou seja, através da análise do sinal, identificar as notas que estão sendo tocadas, para podermos traduzir o som em partituras. Pode ser feito uso dessas técnicas de análise e síntese de áudio na composição e na execução musical, permitindo mais facilmente que o compositor utilize novos timbres e alturas musicais genéricas. Uma aplicação próxima seria a síntese de instrumentos, assim como a identificação de instrumentos, temas e estilos. Outro interesse excepcional hoje em dia seria a representação compacta dos sinais de áudio, possibilitando, por exemplo, o transporte mais rápido de um arquivo de som em uma rede de computadores, dado seu tamanho de armazenamento reduzido.

Teremos nesse capítulo, inicialmente, a apresentação de alguns refinamentos às técnicas de análise vistas nos capítulos anteriores. Esses refinamentos incluem uma forma mais aprimorada de se fazer um limiar na detecção de picos, a detecção das fases das parciais (capítulo 1) e a representação do sinal em linhas freqüenciais principais.

Posteriormente, teremos a apresentação de métodos para realizarmos a separação de sinais com duas vozes instrumentais e, finalmente, apresentaremos um método para a ressíntese de sinais. Fechando o capítulo teremos a apresentação de resultados para os métodos de separação e ressíntese.

3.2 Refinamentos na análise

Antes de entrarmos nos detalhes da separação e da ressíntese de sinais, alguns melhoramentos precisam ser feitos nos métodos apresentados no capítulo 1.

3.2.1 Filtragem Two-Pass Split-Window

A filtragem Two-Pass Split-Window (TPSW) [7] é um procedimento não-linear para a estimação de ruído em um espectro de potência.

Uma janela *split* consiste de dois pulsos quadrados separados por um intervalo. Pode ser definida como:

$$h_{SW}[n] = \begin{cases} 0, & |n| < M \\ 1, & M \leq |n| < N \end{cases} \quad (3.1)$$

onde M e N são inteiros não-negativos, satisfazendo $0 \leq M < N$. Assim, o tamanho total

da janela é $L = 2N - 1$ e o tamanho do intervalo central é $G = 2M - 1$.

Considere uma seqüência não-negativa $s[n]$ que, para o propósito de estimação do espectro do ruído, corresponde ao espectro de potência. Uma estimativa $s_{TPSW}[n]$, baseada no método TPSW, para o espectro do ruído pode ser obtida pelos seguintes passos:

1. Compute $s^{'}[n]$ através da filtragem de $s[n]$ por $h_{SW}[n]$, através de:

$$s^{'}[n] = \frac{1}{2(N-M)} \sum_{k=-N}^{N} h_{SW}[k] s[n-k]. \tag{3.2}$$

2. Modifique $s[n]$ de acordo com o seguinte critério de substituição:

$$s^{''}[n] = \begin{cases} s[n], & \text{se } s[n] \leq \alpha s^{'}[n] \\ s^{'}[n], & \text{se } s[n] > \alpha s^{'}[n] \end{cases} \tag{3.3}$$

   onde $\alpha s^{'}[n]$ é um limiar que controla a rejeição de picos e os picos laterais que aparecem na estimação quando há a presença de picos. O parâmetro $\alpha \geq 1$ é chamado de ganho de limiar.

3. Obtenha a estimativa da média local, baseada no método TPSW, $s_{TPSW}[n]$, filtrando $s^{''}[n]$ por $h_{SW}[n]$,

$$s_{TPSW}[n] = \frac{1}{2(N-M)} \sum_{k=-N}^{N} h_{SW}[k] s^{''}[n-k]. \tag{3.4}$$

A convolução entre o espectro de potência e a janela *split* no primeiro passo funciona similarmente a um filtro de média variável. Então, o valor de $s^{'}[n]$ em um dado instante n corresponde à média local das amostras dentro da janela.

É plausível esperar um aumento em $s'[n]$ quando um pico espúrio aparece e permanece na janela de observação. Porém, quando esse pico coincide com o intervalo da janela *split*, ele não afeta a saída $s^{'}[n]$. Então, $s^{'}[n]$ tem dupla funcionalidade. Por um lado, pode formar um limiar variável $\alpha s^{'}[n]$ para a detecção de picos e indica a presença de picos espúrios. Por outro lado, forma a entrada para o segundo passo, $s^{''}[n]$. A idéia é ajustar o ganho de limiar de forma que $s^{''}[n]$ seja igual a $s^{'}[n]$, exceto quando ocorrer a presença dos picos espúrios. Quando isso ocorrer, $s^{''}[n]$ recebe as amostras do espectro original. Assim, a saída final, que dá a estimativa $s_{TPSW}[n]$, fica livre dos picos espúrios.

Aumentando-se o valor de N temos uma média local estimada mais suavemente, porém isso reduz a possibilidade de acompanhamento de possíveis variações rápidas no espectro de potência. O valor de M deve ser escolhido de forma que o tamanho do intervalo

da janela seja aproximadamente tão grande quanto a largura dos picos que se quer rejeitar quando estimamos o ruído. Quanto ao valor do ganho do limiar α, este serve para regularmos os valores de $s'[n]$, possibilitando um controle do equilíbrio entre a rejeição de picos e a redução de picos espúrios na estimativa.

3.2.2 Atribuição de fase pela interpolação parabólica

Para conseguirmos uma posterior ressíntese dos sinais, precisamos da informação de fase, além das freqüências e amplitudes identificadas pelo método da DFT[1], já explicado.

O formato do lóbulo principal do espectro de magnitude de janelas típicas assemelha-se a uma parábola. Se o sinal de entrada tiver uma única ressonância, seu espectro de magnitude corresponde ao espectro de magnitude da janela modulada para a freqüência de ressonância. Aplica-se então o algoritmo de atribuição de fase pela interpolação parabólica [6]:

- Ajusta-se a parábola para três pontos: o pico detectado e seus vizinhos da direita e da esquerda;
- Utilizam-se as coordenadas de máximo da parábola ajustada, tais como amplitude e freqüência.

Sendo $X_{dB}(k) = 20\log_{10}(|X(k)|)$ o espectro de magnitude de $x[n]$ na escala em dB, para o índice k_p de um determinado pico, temos:

$$A_1 = X_{dB}(k_p - 1), A_2 = X_{dB}(k_p), \text{ e } A_3 = X_{dB}(k_p + 1). \tag{3.5}$$

Pode ser mostrado que:

$$f_{k_p} = (k_p + d)\frac{F_s}{N}, \tag{3.6}$$

onde

$$d = \frac{1}{2}\frac{A_1 - A_3}{A_1 - 2A_2 + A_3}.$$

Para a amplitude, temos, em dB:

$$a_{k_{p,dB}} = A_2 - \frac{d}{4}(A_1 - A_3). \tag{3.7}$$

Para obtermos a fase, fazemos o ajuste da parábola separadamente para a parte real e a parte complexa do espectro complexo. Temos a equação para fase:

$$\theta_{k_p} = -arctg(\frac{a_{k_p,imag}}{a_{k_p,real}}), \tag{3.8}$$

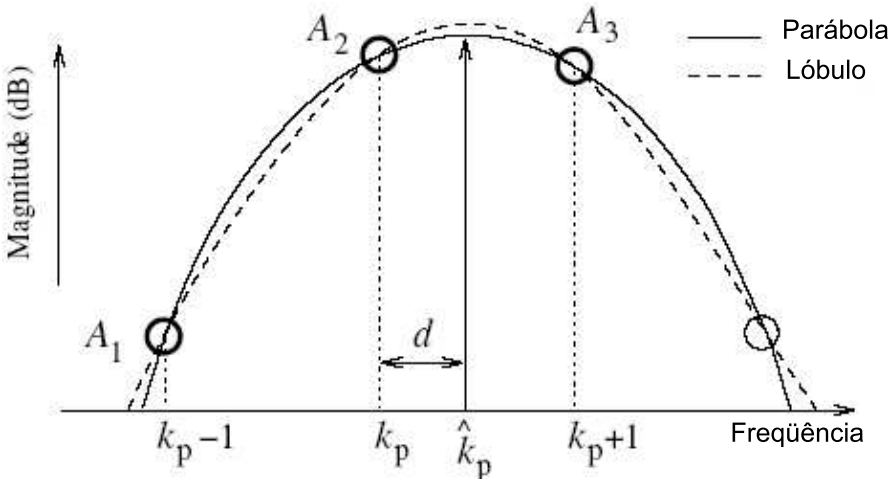

Figura 3.1: Ilustração do processo de interpolação parabólica.

onde $a_{k_p,imag}$ e $a_{k_p,real}$ são computadas assim como a amplitude normal. No entanto, A_i não representam o espectro de magnitude em dB, mas sim as partes real e imaginária do espectro complexo.

Temos uma representação da interpolação parabólica na figura 3.1.

3.2.2.1 Algoritmo de análise senoidal de sinal modificado

Com a interpolação parabólica e a filtragem TPSW, modificamos o algoritmo da seção 1.6, acrescentando a fase:

- Aplicar o janelamento à função a.
- Aplicar a filtragem TPSW.
- Obter a DFT^0.
- Computar a derivada do sinal original, a'.
- Aplicar o mesmo janelamento à função a'.
- Aplicar a filtragem TPSW.
- Obter a DFT^1.
- Corrigir o espectro de magnitude pelo fator F.
- Para cada índice m referente a um máximo na DFT^0:

1. Computar a freqüência exata, usando a correção.
2. Computar a amplitude exata, usando a correção.
3. Através da interpolação parabólica, computar a fase.
4. Adicionar o trio (freqüência,amplitude e fase) à lista de resultados do bloco corrente.

3.2.2.2 Exemplo de aplicação

Para exemplificar a aplicação do método de interpolação parabólica, temos na figura 3.3 o resultado da aplicação do algoritmo para o caso mostrado na subseção 1.8.3, um oscilador senoidal de freqüência 2000 Hz modulada por vibrato, com taxa de variação de 5 Hz e de profundidade de 10 Hz, com amplitude constante e igual a 1. Na figura 3.2 temos o comportamento da freqüência para esse caso.

Podemos notar que a cada mínimo ou máximo na freqüência corresponde um pico ou vale na fase detectada, havendo ambigüidade entre π e $-\pi$. Podemos notar que por uma falha na detecção da freqüência ocorreu o aparecimento de uma outra trilha na fase, distorcendo o valor encontrado nesse ponto.

3.2.3 A formação de trilhas freqüenciais no tempo

A detecção de picos e a estimação de seus parâmetros ocorre *frame* a *frame*, sem que a informação contida em um *frame* afete algum outro *frame*. Para ligar esses *frames* com informações *intraframes*, um esquema de continuação de picos, formando trilhas no espectro, é utilizado.

Deve-se rastrear a história recente do pico, identificando quando o pico apareceu, como evoluiu no tempo e quando desapareceu. Existem métodos heurísticos e métodos baseados em regras, assim como soluções estatísticas para esse problema da continuação do pico. Esses métodos são explicados em [6]. Vamos aqui explicar um método baseado em regras, descrito em [6] e baseado em idéias usadas em [8], que foi utilizado nesse projeto. Métodos baseados em regras foram originalmente propostos em [9] e posteriormente modificados e estendidos em [8] e [5].

Basicamente, na maioria dos métodos baseados em regras, o principal critério para a decisão pela trilha a ser feita é a proximidade das freqüências entre os picos envolvidos.

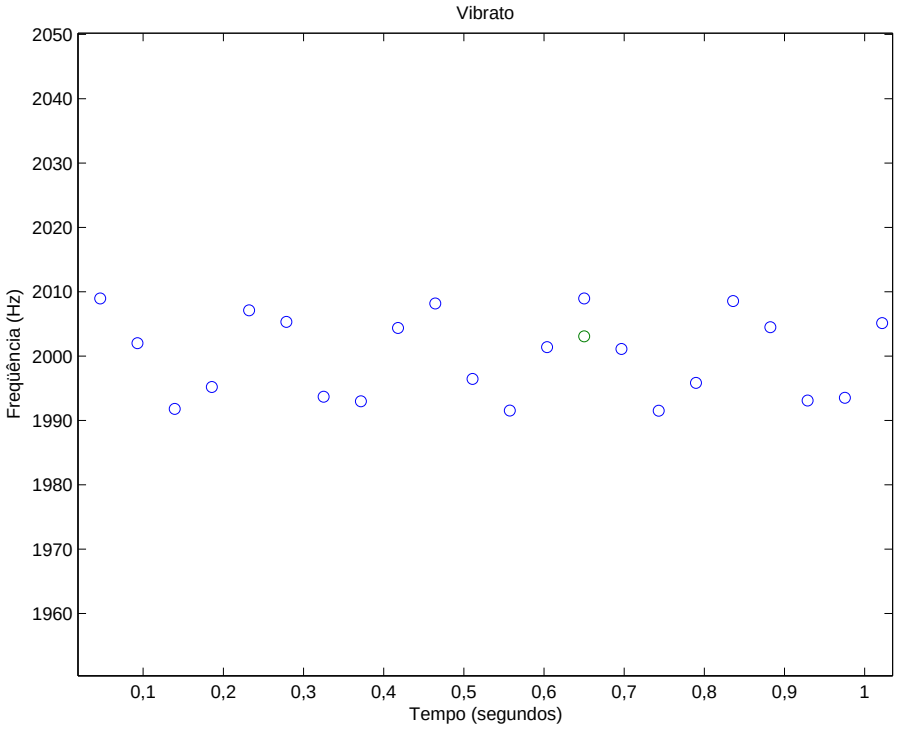

Figura 3.2: Comportamento da freqüência para um oscilador senoidal de freqüência 2000 Hz modulada por vibrato com taxa de variação de 5 Hz e de profundidade de 10 Hz.

Normalmente, o número de picos varia de *frame* para *frame*. A cada pico é atribuído um *status*, que pode ser: trilha emergente, trilha evoluindo e trilha morrendo. Trilhas nos dois primeiros casos são consideradas ativas, sendo consideradas inativas quando no terceiro caso. O procedimento baseia-se sempre na informação do pico em dois *frames* consecutivos. O primeiro *frame* é uma exceção. Nele, todos os picos são inicializados como trilhas emergentes.

Consideremos que estamos no *frame* m, cujos picos devem ser combinados com trilhas ativas do *frame* $m-1$. Existem p picos no frame $m-1$ e suas freqüências são chamadas $f_1, f_2, ..., f_p$. No frame m há r picos cujas freqüências são chamadas $g_1, g_2, ..., g_r$.

Para todas as i trilhas ativas no frame $m-1$, uma busca no frame m é realizada, procurando achar um pico cuja freqüência esteja dentro de um intervalo em torno da freqüência de uma trilha i. Ou seja, a trilha i busca um pico g_j no frame m, tal que a diferença entre f_i e g_j seja menor do que um valor definido. Esse valor pode ser dependente da freqüência, por exemplo, um semitom em torno de f_i. Temos então duas possibilidades:

- Se a trilha i não achar uma continuação, seu *status* muda de ativa para inativa, uma trilha que está morrendo. Na verdade, uma trilha que está morrendo no *frame* $m-1$

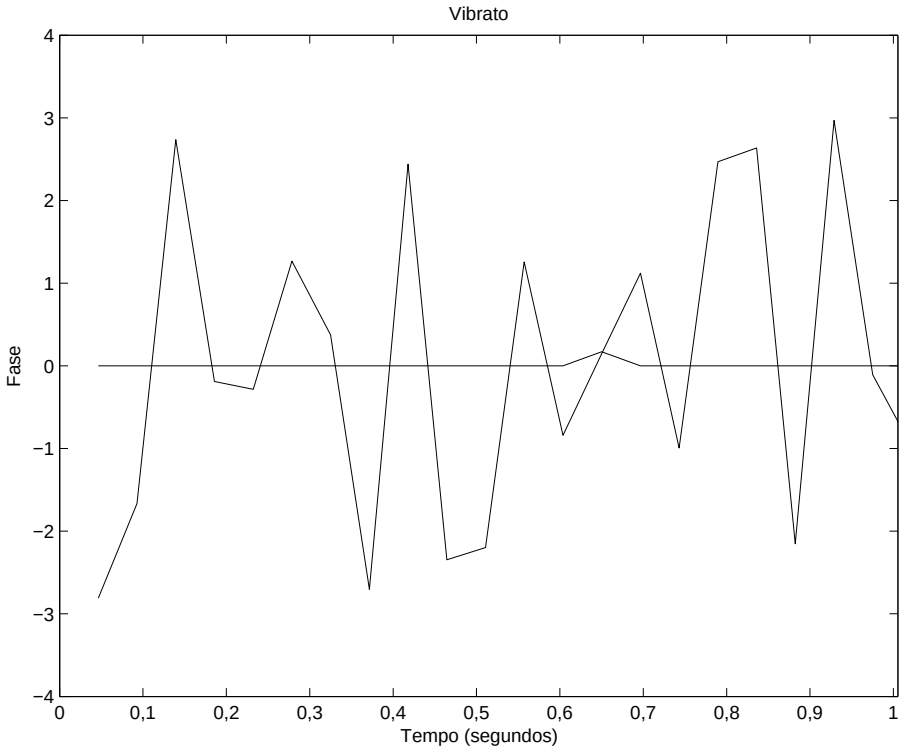

Figura 3.3: Resultado obtido com o método de interpolação parabólica para um oscilador senoidal de freqüência 2000 Hz modulada por vibrato com taxa de variação de 5 Hz e de profundidade de 10 Hz.

morre no *frame* m, pois a trilha é estendida até o *frame* m, onde é criado um pico com a mesma freqüência, porém com amplitude igual a zero.

- Se a trilha i achar uma continuação, seu *status* permanece como ativa (trilha evoluindo) e o pico g_j que é o mais próximo de f_i em freqüência passa a fazer parte da trilha. Mais de um pico no *frame* m pode satisfazer ao critério de proximidade na freqüência utilizado, o que deve ser depois analisado. Temos então mais duas possibilidades:

  - g_j é um pico livre, ou seja, não foi requisitado por nenhuma outra trilha ativa no *frame* $m-1$. Como não há conflito, o pico g_j é imediatamente associado à trilha i.

  - g_j já foi requisitado por outro pico no *frame* $m-1$, diferente de f_i. Para resolver esse conflito, costuma-se medir a diferença entre o pico g_j e os picos requerentes, decidindo-se de acordo com critérios previamente definidos. Suponha que as trilhas u e v requisitem o mesmo pico g_j, e v é a trilha que o está atualmente requisitando. Primeiro, medimos $d_u = |f_u - g_j|$ e $d_v = |f_v - g_j|$. Agora temos duas possíveis situações:

    * Se $d_v > d_u$, a trilha atual, v, perde a disputa e escolhe então o pico mais adequado dentre os que estão disponíveis. Se existe algum pico adequado entre esses picos, a trilha permanece ativa. Se não, troca-se o *status* da trilha para inativa.

    * Se $d_v < d_u$, a trilha atual, v, ganha a disputa e é realizado o procedimento de procura de um pico adequado para a trilha u, que passa a ser a trilha atual. A trilha u irá tentar novamente associar-se ao pico g_j e perderá a disputa. Então, de acordo com o item anterior, associar-se-á ao mais adequado pico dentre os picos disponíveis (se possível) e manter-se-á seu *status* como ativa ou mudar-se-á seu *status* para inativa.

O processo descrito anteriormente é repetido para todas as trilhas ativas no *frame* $m-1$ até que o *status* dessas trilhas tenha sido atualizado. Como já dito, as trilhas cujo *status* mudou para inativa são então estendidas até o *frame* m, onde morrem com a mesma freqüência associada em $m-1$, mas com amplitude igual a zero. Para os picos em m que permanecem não associados, novas trilhas são criadas, com *status* de trilhas emergentes. Similarmente ao que foi feito com as trilhas que estavam morrendo, as novas trilhas que

aparecem no *frame* m são estendidas ao *frame* anterior, $m-1$, onde começam com amplitude zero e as mesmas freqüências às quais foram associadas no *frame* m.

Um refinamento que pode ser feito ao processo descrito consiste em incluir histereses associadas com a decisão de começar uma nova trilha ou terminar uma já existente. Um exemplo onde esse refinamento seria útil é o seguinte: pode acontecer de algumas parciais sofrerem de modulação de amplitude. Ocorrendo tal fato, a amplitude da parcial pode permancer abaixo do limiar de amplitude adotado durante alguns *frames*. Sendo assim, o algoritmo de formação de trilhas irá terminar a trilha sempre que o pico a ela associado desaparecer em um dado *frame*, começando uma nova trilha com o mesmo pico alguns *frames* depois, quando o pico reaparecer. Teremos como resultado diversas trilhas segmentadas, ao invés de uma só trilha contínua, como deveria ser.

A histerese na mudança de *status* consiste em considerar um certo número de chances para que a trilha então termine, antes de mudar seu *status* para inativa. Uma maneira prática de implementar tal esquema seria a seguinte:

1. Aplicar um contador à trilha sempre que esta estiver para ser considerada inativa;
2. Retardar a mudança de *status* até que o contador atinja determinado valor. Isso implica em estender a trilha por *frames* sucessivos através da adição de picos com a mesma amplitude e mesma freqüência;
3. Incrementar o contador a cada *frame* processado;
4. Repetir o procedimento até que o contador atinja o valor determinado. Se nesse tempo a trilha encontrar um pico adequado, o contador deve ser zerado e deve-se proceder normalmente. Se não, confirmar a mudança de *status*, retirando os picos que foram criados artificialmente para estender a trilha, que deve então ser terminada no *frame* no qual deveria inicialmente ser terminada.

Uma estratégia similar pode ser usada para trilhas emergentes, com o objetivo de evitar que picos espúrios iniciem trilhas que serão muito curtas. Assim, uma trilha emergente só seria confirmada se permanecesse ativa durante um certo número de *frames*.

Temos nas figuras 3.4 e 3.5 representações do esquema explicado nessa seção.

Analisamos um trecho de aproximadamente dois segundos de uma gravação de "Bachianas Brasileiras número 6" de Villa-Lobos. Temos na figura 3.6 o resultado encontrado para a formação de trilhas.

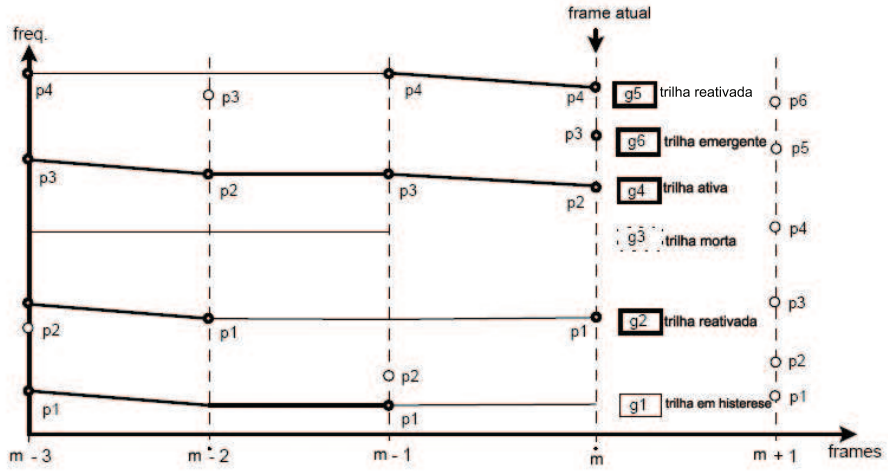

Figura 3.4: Esquema de formação de trilhas, onde g representa as trilhas e p, os picos.

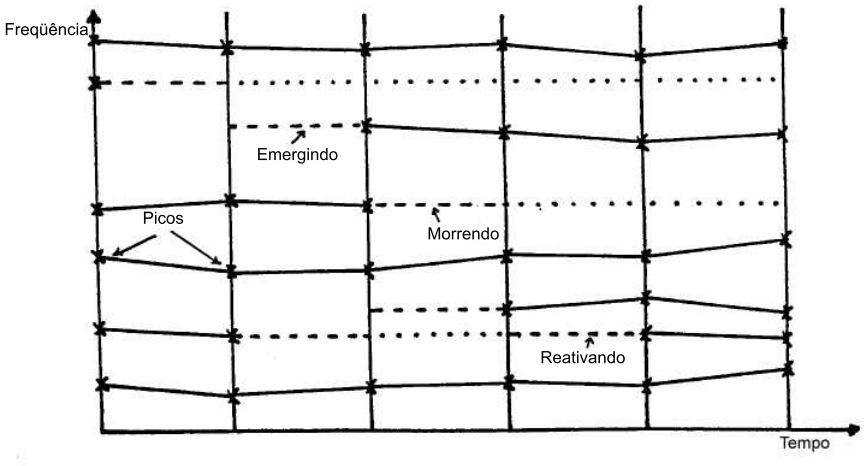

Figura 3.5: Esquema de formação de trilhas, destacando a utilização da estratégia de histerese.

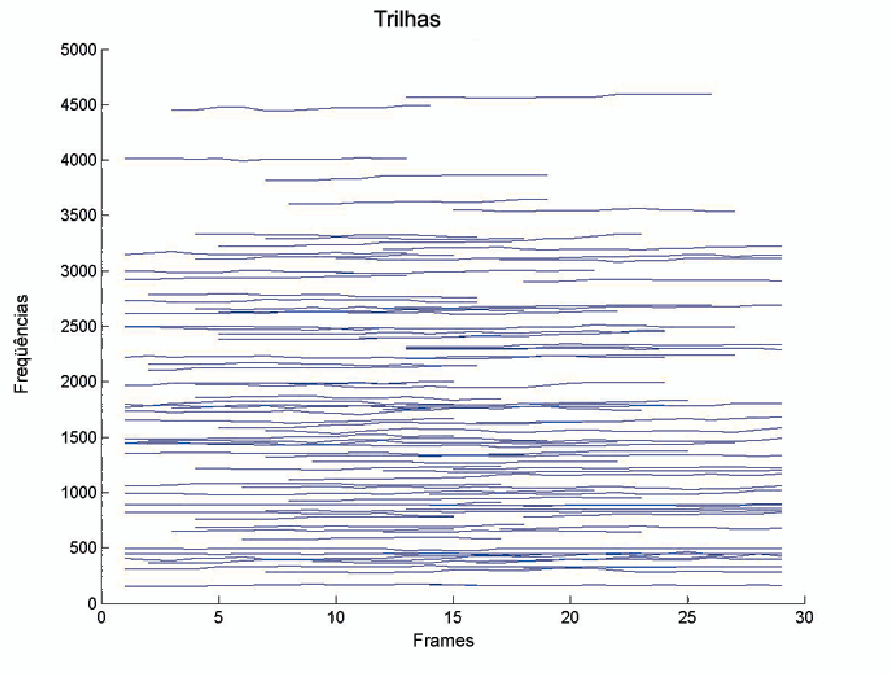

Figura 3.6: Trilhas formadas em um trecho de dois segundos de uma gravação musical.

3.3 Heurística da separação

Uma vez aplicado o método da DFT[1] ao sinal, tendo sido obtidos os respectivos valores de amplitude, freqüência e fase e tendo sido formadas as trilhas, o próximo passo é a separação dos sinais presentes em um determinado sinal, com base nas fontes sonoras.

O estudo de caso neste trabalho restringe-se à separação de um sinal composto de baixa complexidade, gerado por dois instrumentos monofônicos de sopro.

Uma primeira idéia seria separar os sinais da seguinte forma, sabendo que um sinal tem freqüência fundamental f_a e o outro, f_b, sendo $f_b > f_a$: as trilhas cujas freqüências médias fossem inferiores a f_b pertenceriam a uma fonte sonora e as demais pertenceriam à outra fonte sonora. Algo como um filtro passa-baixas para uma fonte e um passa-altas para a outra.

Pode-se prever que o resultado dessa técnica não será satisfatório. Algumas freqüências harmônicas da fonte cuja freqüência fundamental é menor estarão presentes no sinal referente à fonte cuja freqüência fundamental é maior. Por outro lado, o espectro da fonte cuja freqüência fundamental é menor será drasticamente reduzido.

Empregamos esse método no mesmo trecho musical da seção anterior, extraído da gravação de "Bachianas Brasileiras número 6", cujas trilhas estão representadas na figura

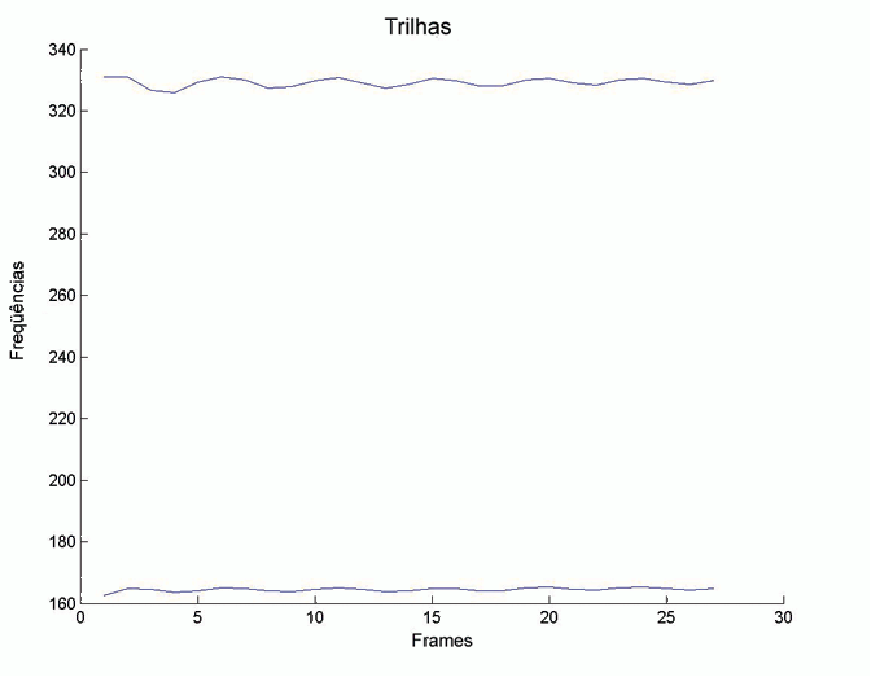

Figura 3.7: Resultado da separação por filtros para o fagote.

3.6. Nessa gravação, temos um fagote cuja freqüência fundamental é 164 Hz e uma flauta cuja freqüência fundamental é 444 Hz.

Temos nas figuras 3.7 e 3.8 os resultados dessa separação.

Uma outra abordagem que pode ser utilizada é separar os sinais de acordo com suas freqüências fundamentais. Sabendo-se *a priori* a freqüência fundamental de cada instrumento, buscam-se seus respectivos harmônicos. Essa foi a abordagem utilizada nesse trabalho. Calcula-se a média de cada trilha e divide-se essa média pela freqüência fundamental. Através da parte inteira desse quociente, calcula-se a harmônica mais próxima da média da trilha. Faz-se a diferença entre a freqüência dessa harmônica e a freqüência média da trilha. Analisa-se então se essa diferença está dentro de um intervalo pré-definido (como já dito antes, esse intervalo é definido de acordo com a freqüência; pode ser, por exemplo, um semitom). Se a resposta for positiva, a trilha atual integra o grupo de trilhas que estão relacionadas ao dado instrumento.

O resultado desse algoritmo pode ser visto nas figuras 3.9 e 3.10. Pode-se observar que, em alguns momentos, temos a superposição de duas trilhas, onde esperávamos ter apenas uma trilha. Esse problema pode ser resolvido diminuindo-se o tempo de histerese utilizado. No nosso caso, diminuímos o tempo de histerese, originalmente $100ms$, para

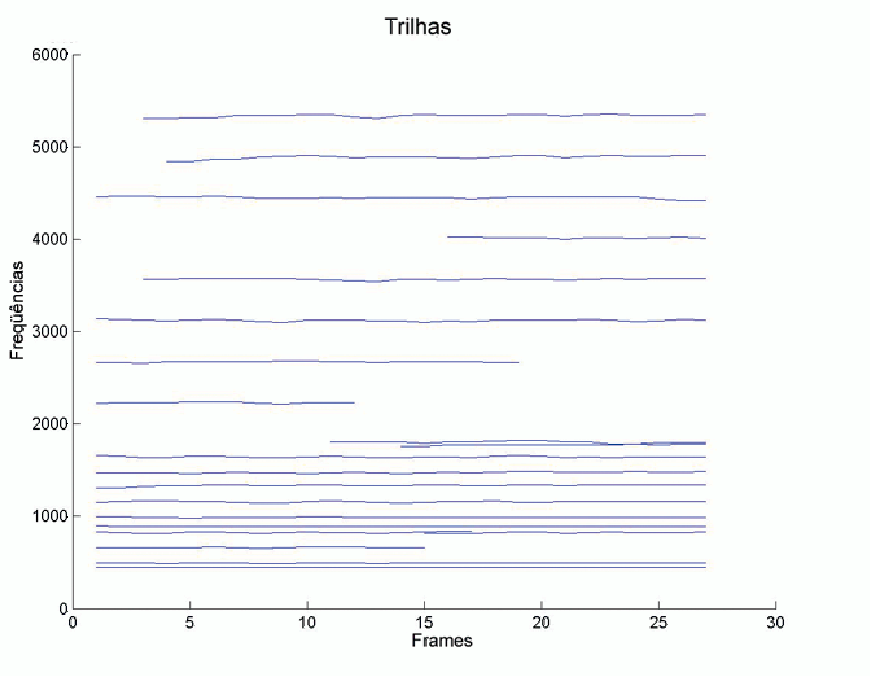

Figura 3.8: Resultado da separação por filtros para a flauta.

$50ms$. Temos os resultados nas figuras 3.11 e 3.12. Obtivemos uma diminuição desse efeito indesejável, porém, obtivemos também um menor número de trilhas para cada fonte sonora, como esperado, prejudicando a qualidade dos sinais quando ressintetizados.

3.4 Ressíntese de sinais

Tendo as informações espectrais dos sinais e os *frames* devidamente ligados, podemos ressintetizar tais sinais a partir de parâmetros de modelamento senoidal obtidos.

A idéia reside em, suavemente, interpolarmos as amplitudes, freqüências e fases das trilhas de um *frame* com os valores do próximo *frame*, para evitarmos descontinuidades no término de um *frame* e no início do *frame* seguinte, o que costuma degradar a qualidade do sinal sintético.

3.4.1 Algoritmo de ressíntese

Como resultado do algoritmo descrito em 3.2.3, todos os parâmetros medidos em um *frame* k são associados com um conjuto de parâmetros no *frame* $k+1$. Considerando $(A_l^k, \omega_l^k, \theta_l^k)$ e $(A_l^{k+1}, \omega_l^{k+1}, \theta_l^{k+1})$ como os parâmetros para a $l-$ *ésima* trilha freqüencial, nos

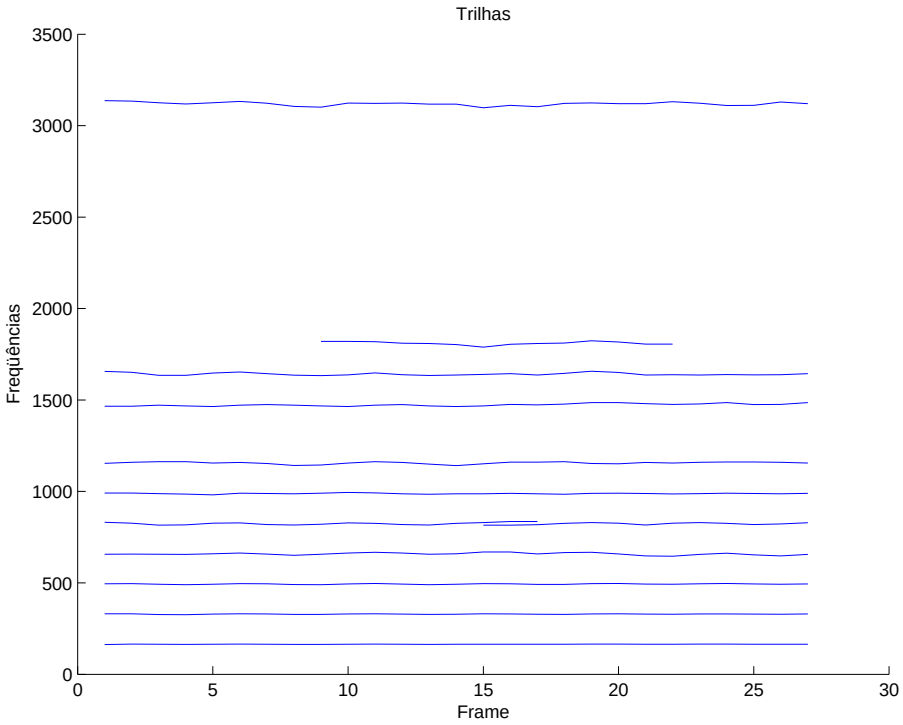

Figura 3.9: Resultado da separação por harmônicos para o fagote.

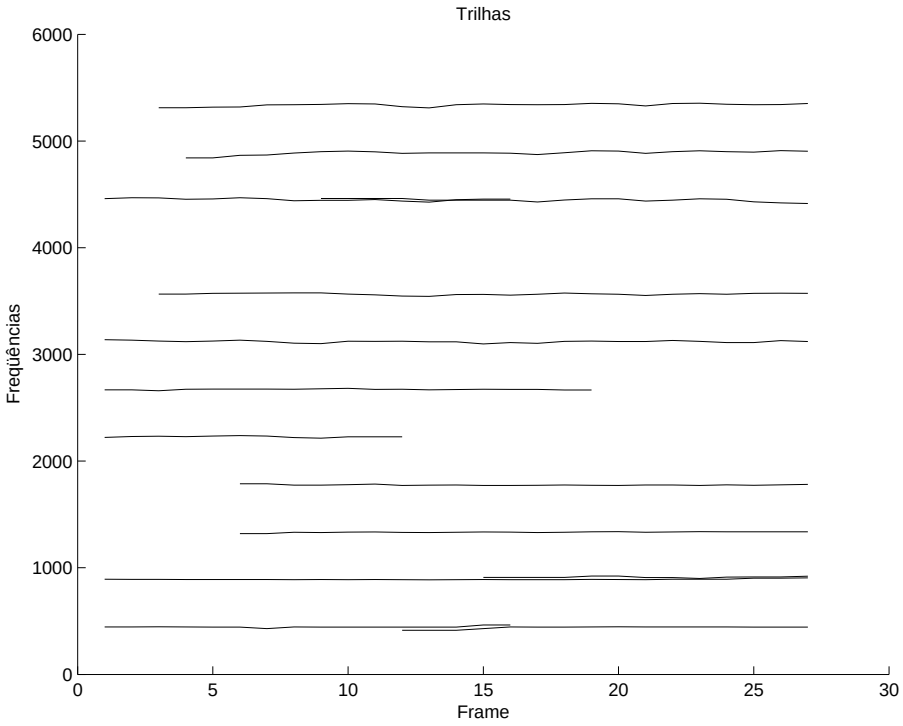

Figura 3.10: Resultado da separação por harmônicos para a flauta.

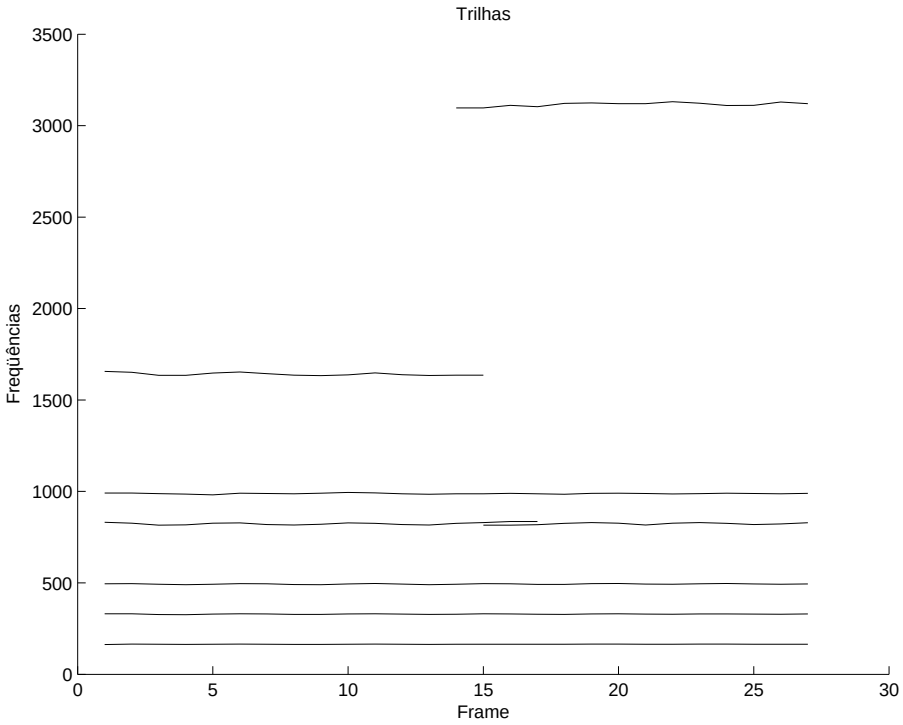

Figura 3.11: Resultado da separação por harmônicos para o fagote com histerese de $50ms$.

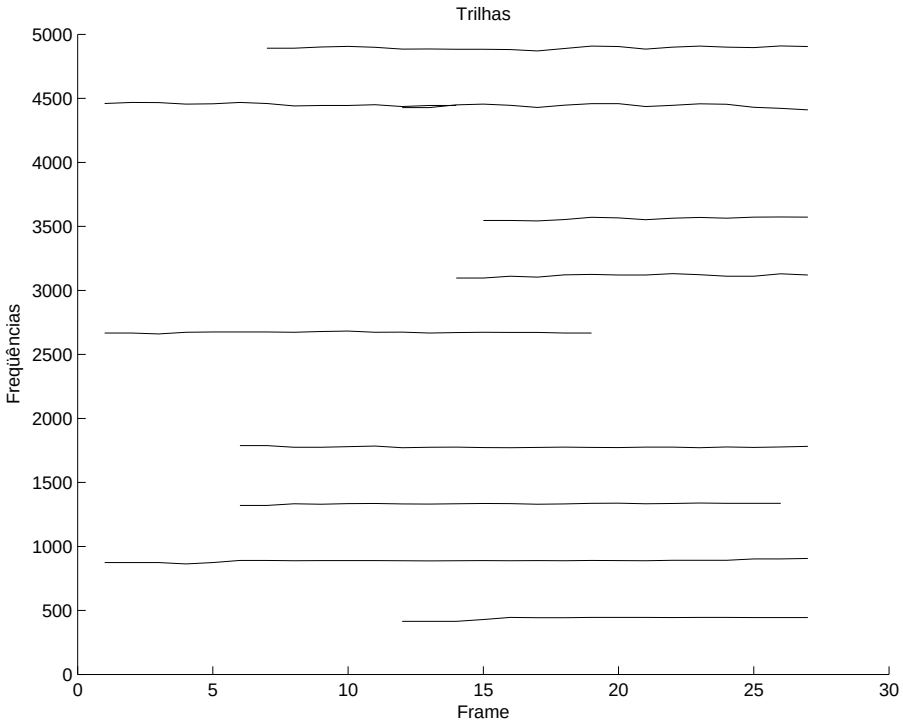

Figura 3.12: Resultado da separação por harmônicos para a flauta com histerese de $50ms$.

frames k e $k+1$, respectivamente, então uma solução óbvia para a interpolação de amplitudes seria:

$$A_l[n] = A_l^k + \frac{(A_l^{k+1} - A_l^k)}{S} n, \tag{3.9}$$

onde $n = 0, 1, ..., S-1$ é o índice temporal no $k-$ *ésimo frame* e S é o número de amostras que será gerado para fazermos a interpolação entre um *frame* e outro..

Infelizmente, uma abordagem tão simples não pode ser usada para interpolarmos a freqüência e a fase, porque a fase medida θ_l^k é obtida em módulo 2π. Então, um desdobramento da fase deve ser feito para garantir que as trilhas freqüenciais tenham uma transição suave entre os *frames*. Dado que quatro variáveis influenciam o valor da fase instantânea (ω_l^k,θ_l^k,ω_l^{k+1} e θ_l^{k+1}), precisamos de pelo menos três graus de liberdade para controlá-la, enquanto que a interpolação linear só nos dá um grau. O primeiro passo na resolução desse problema é definir uma função de interpolação de fases que é polinomial cúbica:

$$\theta_l(t) = \zeta + \gamma t + \alpha t^2 + \beta t^3. \tag{3.10}$$

É conveniente tratar a função de fase como se fosse uma função da variável de tempo contínuo t, com $t = 0$ correspondendo ao início do *frame* k e $t = T$ correspondendo ao início do *frame* $k + 1$. Como a freqüência é a derivada da fase, é necessário que a função de fase cúbica e sua derivada sejam iguais às fases e freqüências medidas no limite do *frame*.

Usando o fato de que a freqüência instantânea é a derivada da fase, temos:

$$f_l(t) = \frac{d}{dt}\theta(t) \tag{3.11}$$

e

$$f_l(t) = \gamma + 2t + 3\beta t^2. \tag{3.12}$$

Em $t = 0$,

$$\theta_l(0) = \zeta = \theta_l^k \tag{3.13}$$

$$f_l(0) = \gamma = \omega^k; \tag{3.14}$$

em $t = T$,

$$\theta_l(T) = \theta_l^k + \omega_l^k T + \alpha T^2 + \beta T^3 = \theta_l^{k+1} + 2\pi M \tag{3.15}$$

$$f_l(T) = \omega_l^k + 2\alpha T + 3\beta T^2 = \omega_l^{k+1}. \tag{3.16}$$

Como a fase θ_l^{k+1} é medida em módulo 2π, é necessário aumentá-la pelo termo $2\pi M$ (M é um inteiro) para que a função da freqüência tenha a transição mais suave possível. A

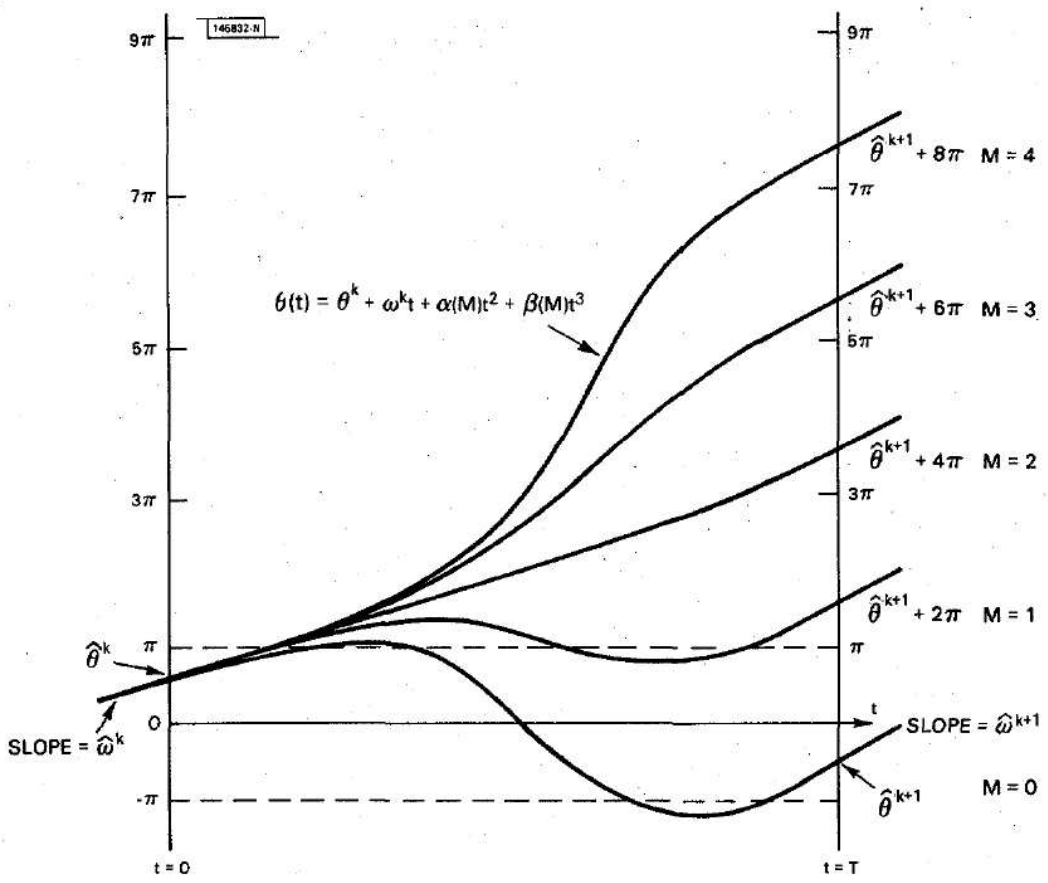

Figura 3.13: Exemplo de funções cúbicas de interpolação de fase para um número de valores de M.

variável M ainda é desconhecida, mas, para cada valor de M, podemos resolver a função para $\alpha(M)$ e $\beta(M)$. A solução seria a seguinte:

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \theta_l^{K+1} - \theta_l^K - \omega_l^K T + 2\pi M \\ \omega_l^{K+1} - \omega_l^K \end{bmatrix}. \tag{3.17}$$

Para determinar M e a solução final para o problema do desdobramento da fase, precisamos especificar o conceito de transição mais suave possível. A figura 3.13 ilustra um conjuto de funções cúbicas de interpolação de fase para um número de valores de M. Parece intuitivo que a melhor função a ser escolhida é aquela com menor variação. Esse é o significado de uma transição mais suave possível. Se as freqüências fossem constantes e o sinal estacionário, a função da fase seria linear.

Temos então que um critério razoável para que a transição seja a mais suave possível seria escolher um M tal que

$$p(M) = \int_0^T [\frac{d^2}{dt^2}\theta_l(t; M)]^2 dt \tag{3.18}$$

seja mínima.

Apesar de M ser um inteiro, o problema pode ser mais facilmente resolvido minimizando-se $p(x)$ com respeito à variável contínua x e então escolhendo M como o inteiro mais próximo de x. Após certa álgebra, podemos chegar à conclusão de que o valor de x que minimiza a função é dado por:

$$x = \frac{1}{2\pi}[(\theta_l^k + \omega_l^k T - \theta_l^{k+1}) + (\omega_l^{k+1} - \omega_l^k)\frac{T}{2}]. \tag{3.19}$$

Através desse valor de x determinamos M, que é usado na equação (3.17) para acharmos $\alpha(M)$ e $\beta(M)$ e, por conseqüência, a função de interpolação de fase

$$\theta_l(t) = \theta_l^k + \omega_l^k t + \alpha(M)t^2 + \beta(M)t^3. \tag{3.20}$$

Como a análise começou com a consideração de que a fase desdobrada θ_l^k corresponde à freqüência ω_l^k no início do *frame* k, é necessário especificar a inicialização do procedimento de interpolação de fase. Quando uma trilha nasce, uma amplitude, uma freqüência e uma fase são medidas no *frame* $k+1$, e os parâmetros no *frame* k da trilha correspondente são definidos como uma amplitude igual a zero ($A_l^k = 0$) e freqüência igual à medida em $k+1$ ($\omega_l^k = \omega_l^{k+1}$). Para garantirmos que as condições de interpolação de fase sejam satisfeitas, a fase desdobrada no início do *frame* k é definida como

$$\theta_l^k = \theta_l^{k+1} - \omega_l^{k+1} S, \tag{3.21}$$

onde S é o número de amostras que será gerado para fazermos a interpolação entre um *frame* e outro.

Com esse esquema de desdobramento da fase, cada trilha freqüencial terá associada uma fase instantânea que estará de acordo com as rápidas mudanças de fase (freqüência) e com as transições mais lentas.

Temos então a seguinte fórmula para a ressíntese de sinais:

$$s[n] = \sum_{l=1}^{L^k} A_l[n] \cos[\theta_l[n]], \tag{3.22}$$

onde $A_l[n]$ é dada pela equação (3.9), $\theta_l[n]$ é a versão discreta da função (3.20) e L^k é o número de senóides estimadas (igual ao número de trilhas) para o *frame* k.

3.4.2 Exemplos de aplicação

Realizamos, primeiramente, a ressíntese de um trecho de 12 segundos de gravação de "Bachianas Brasileiras número 6" de Villa-Lobos já citada na seção 1.8.4, cujas trilhas estão

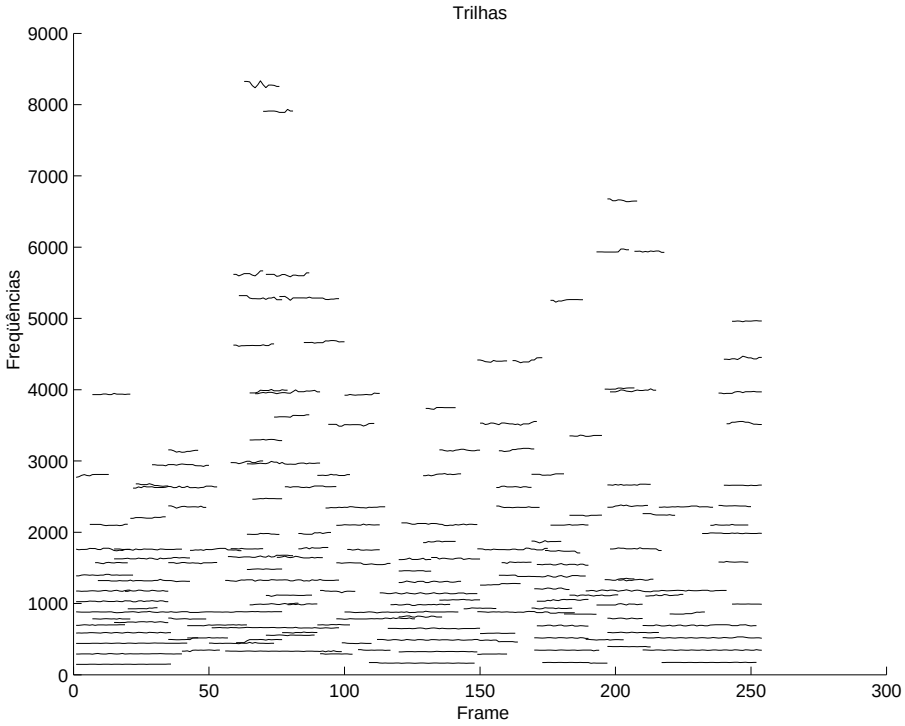

Figura 3.14: Trilhas identificadas para um trecho de 12 segundos de um sinal musical.

representadas na figura 3.14. Utilizamos o método da DFT^1 (N) com 4096 pontos, saltos de 2048 amostras $(N/2)$, janelas de Hann, superposição de 50% dos blocos adjacentes, histerese de 100ms e tamanho mínimo de trilha de 500ms. Os resultados estão nas figuras 3.15 e 3.16. Obtivemos um excelente resultado.

Como forma alternativa de comparar os resultados, as figuras 3.17 e 3.18 apresentam os espectogramas do sinal original e do sinal ressintetizado, respectivamente.

Fizemos então a ressíntese dos sinais separados na seção 3.3 e identificados pelas figuras 3.9 e 3.10. Confirmamos, ao ouvir os sinais ressintetizados, que a separação de acordo com harmônicos é mais eficiente que a separação por filtros, preservando a qualidade de cada uma das fontes sonoras. Temos na figura 3.19 o sinal original, com as duas fontes sonoras. Na figura 3.20 temos o sinal do fagote, após separação, e na figura 3.21 temos o sinal da flauta, também após a separação. Finalmente, temos na figura 3.22 os dois sinais reagrupados.

Fizemos ainda testes com sinais de voz bastante simplificados, algo que a princípio não estava no escopo de nossos testes, alterando o valor do tamanho mínimo de trilha de 500ms para 90ms. Os resultados, que foram muito bons, estão nas figuras 3.23 e 3.24.

Partimos então para um teste mais ousado. Novamente fizemos a ressíntese de sinais

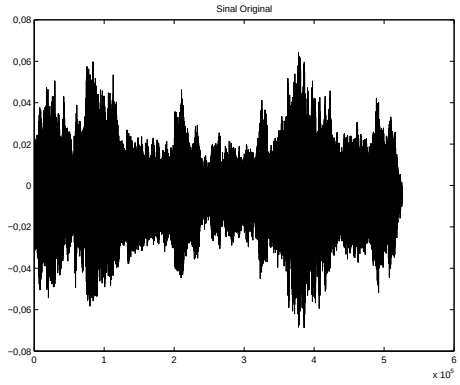

Figura 3.15: Forma de onda do sinal musical original.

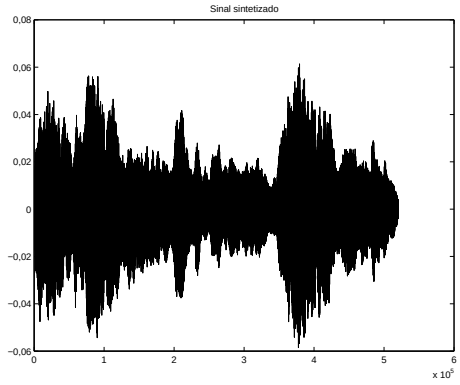

Figura 3.16: Forma de onda do sinal musical ressintetizado.

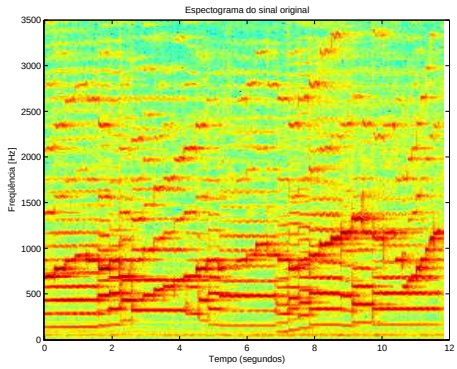

Figura 3.17: Espectograma do sinal musical original.

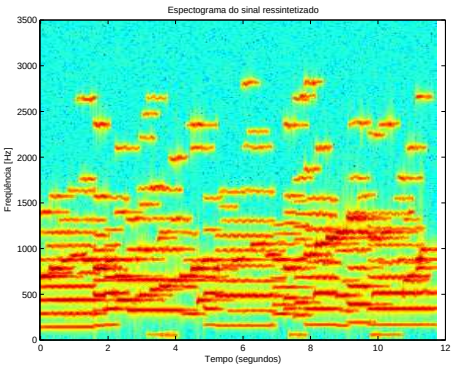

Figura 3.18: Espectograma do sinal musical ressintetizado.

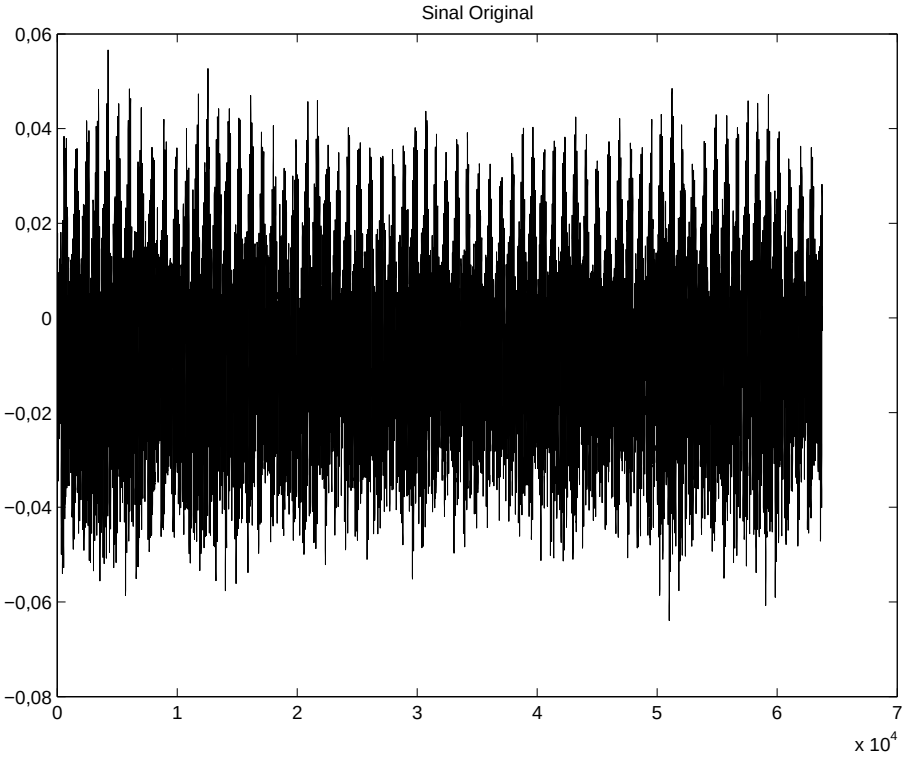

Figura 3.19: Sinal musical composto por flauta e fagote.

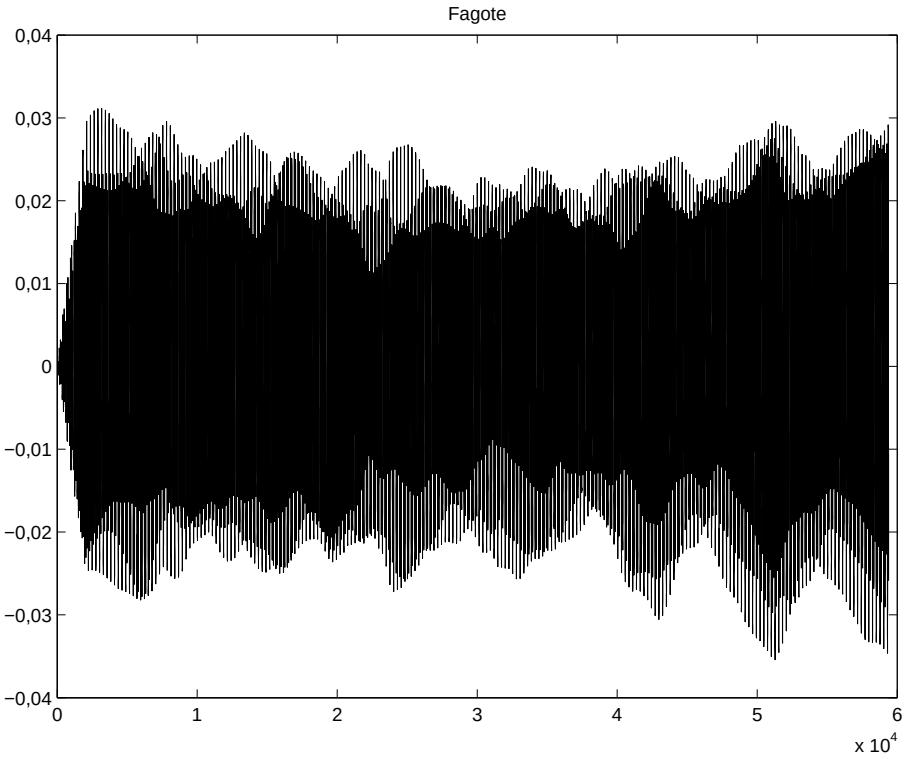

Figura 3.20: Forma de onda do sinal do fagote.

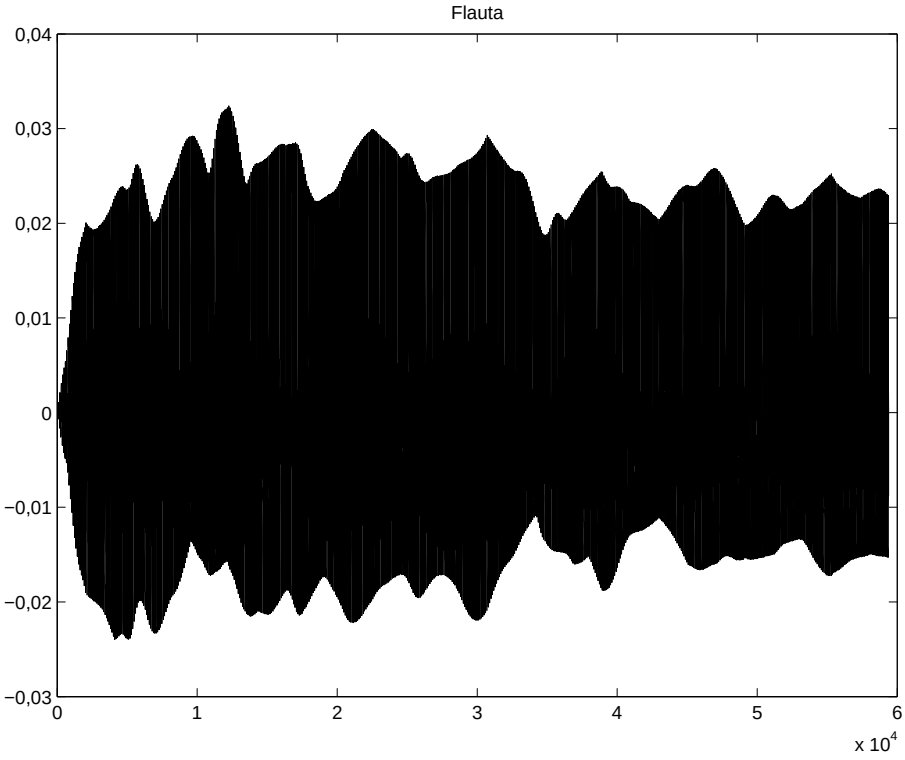

Figura 3.21: Forma de onda do sinal da flauta.

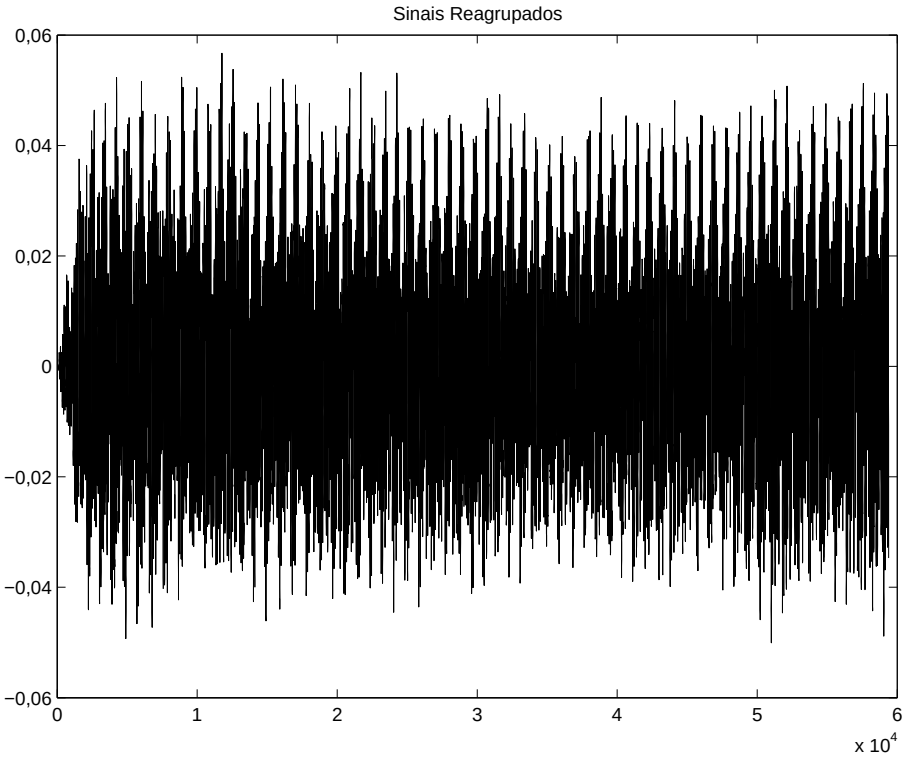

Figura 3.22: Forma de onda do sinal composto pelos sinais da flauta e do fagote reagrupados.

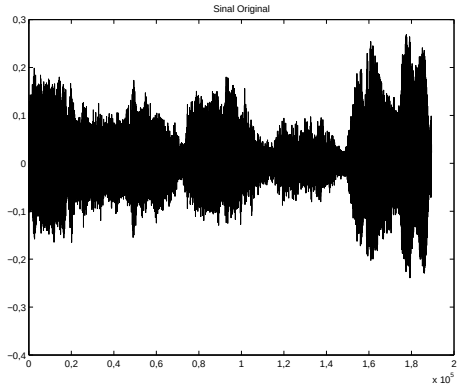

Figura 3.23: Forma de onda de um sinal musical composto por voz, de Naná Vaconcelos.

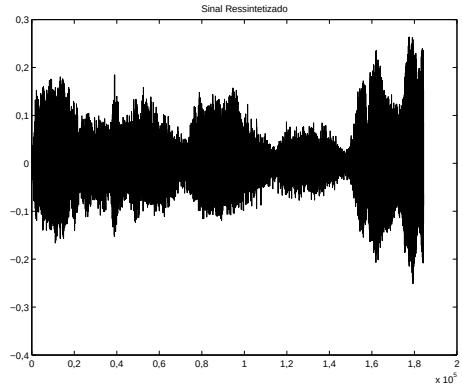

Figura 3.24: Forma de onda do sinal musical composto por voz, de Naná Vaconcelos, ressintetizado.

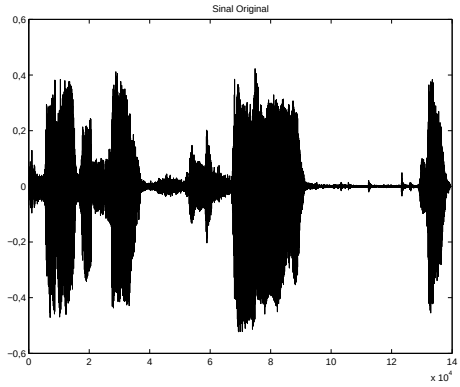

Figura 3.25: Forma de onda de um sinal de voz, do filme "Monty Python: Em busca do cálice sagrado".

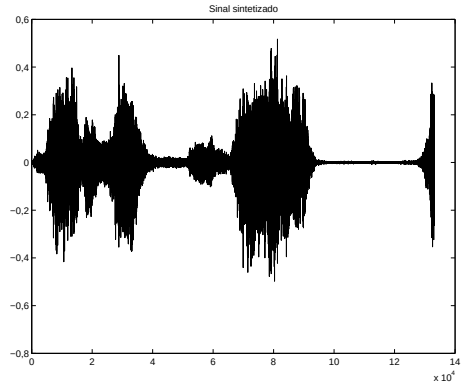

Figura 3.26: Forma de onda do sinal de voz, do filme "Monty Python: Em busca do cálice sagrado", ressintetizado.

de voz, só que agora num trecho de uma fala, não mais um sinal musical. Os resultados estão nas figuras 3.25 e 3.26. Pudemos ver que, onde ocorre as transições de silêncio para fala, no sinal original, estas são suavizadas no sinal ressintetizado, por causa do esquema de interpolação, gerando um efeito não desejável.

Temos nas figuras 3.27 e 3.28 um exemplo de ressíntese de um sinal composto por um instrumento de sopro e outro de cordas. Como esperado, na ressíntese, o instrumento de sopro é reconstituído quase perfeitamente, enquanto o de cordas acaba por ser sintetizado com distorções. Esses problemas ocorrem nos ataques de notas, que envolvam transitórios bruscos, um problema a ser resolvido em uma possível continuação do trabalho.

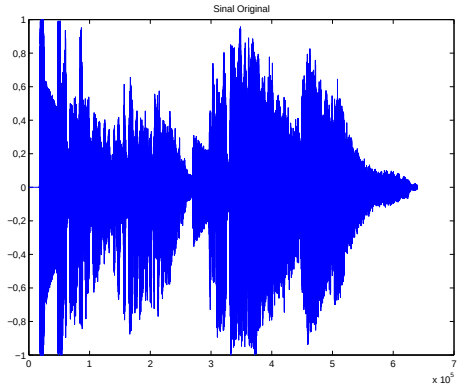

Figura 3.27: Forma de onda de um sinal musical, do início de uma gravação de “Cotidiano”, de Chico Buarque.

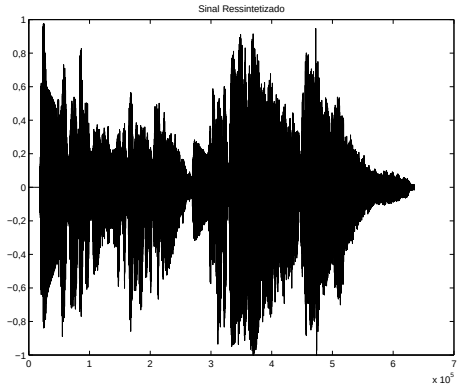

Figura 3.28: Forma de onda do sinal musical, do início de uma gravação de “Cotidiano”, de Chico Buarque, ressintetizado.

3.5 Conclusões

Apesar de o conceito da ressíntese ser bastante simples, na prática tivemos que implementar um sistema muito detalhado, com a identificação de trilhas freqüenciais e um procedimento de interpolação da fase com função cúbica. Os resultados de todos os testes de ressíntese de sinais foram considerados muito bons, mas lembrando que os sinais testados foram sinais considerados simples. Para os sinais mais complexos, obtivemos problemas já esperados nos resultados finais.

Capítulo 4

Conclusões

No capítulo 1, implementamos e testamos com sucesso o método de representação tempo-freqüência baseado na DFT[1], mostrando sua vantagem sobre o método da DFT.

No capítulo 2, o método algébrico para resolver o problema da "mistura" de picos diferentes no domínio da freqüência devido ao janelamento do sinal foi implementado e obtivemos ótimos resultados sobre os sinais de testes.

No capítulo 3, os refinamentos às técnicas de análise vistas nos capítulos anteriores foram implementados e realmente serviram aos seus objetivos. Posteriormente, realizamos a separação de sinais com duas vozes instrumentais, segundo métodos apresentados e, finalmente, fizemos a ressíntese de sinais.

Obtivemos resultados bastante positivos, tanto na parte de análise quanto na de síntese de sinais. Os objetivos do trabalho foram plenamente atingidos.

O trabalho ainda pode ser mais desenvolvido e melhorado, utilizando-se novas abordagens para atacar problemas referentes a sons com outras características mais gerais.

Um problema desafiador na formação das trilhas freqüenciais ocorre quando a trajetória de duas ou mais trilhas se cruzam. O primeiro problema é que, no ponto de interseção, os picos podem ser detectados como um único pico. Caso já se saiba *a priori* as freqüências desses picos, pode-se utilizar a técnica descrita no capítulo 2. Ainda assim, o critério da distância em freqüência para identificar as trilhas torna-se frágil. Pode-se resolver esse problema utilizando informações de um passado mais longo das trilhas em questão, identificando suas trajetórias, aumentando-se a complexidade do algoritmo de formação das trilhas. Pode-se usar também, para ajudar na resolução do problema, um segundo parâmetro na resolução do conflito entre as trilhas, talvez a amplitude dos picos.

Uma outra melhoria no trabalho seria a de identificar-se primeiro a freqüência fundamental de um grupo de trilhas e, *frame* a *frame*, identificar seus harmônicos, possibilitando um melhor controle sobre as trilhas que pertençam à dada fonte sonora.

Pode-se, também, incorporar ao trabalho a identificação automática das freqüências fundamentais.

Por fim, para permitir a utilização dos algoritmos com sinais gerais, é necessário dar um tratamento específico aos ataques de notas (transitórios bruscos).

Referências Bibliográficas

[1] DESAINTE-CATHERINE, M., MARCHAND, S., "High-Precision Fourier Analysis of Sounds Using Signal Derivatives", *Audio Eng. Soc.*, v. 48, n. 7/8, pp. 654 – 667, 2000.

[2] SERRA, X., SMITH, J. O., "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition", *Comput. Music J.*, v. 4, n. 4, pp. 12 – 24, 1990.

[3] HAYKIN, S., VEEN, B. V., *Sinais e Sistemas*. Porto Alegre, RS, Brasil, Bookman, 2001.

[4] MOORE, F. R., *Elements of Computer Music*. Englewood Cliffs, NJ, EUA, Prentice-Hall, 1990.

[5] QUATIERI, T. F., MCAULAY, R. J., "Audio Signal Processing Based On Sinusoidal Analysis/Synthesis". In: Brandenburg, K., Kahrs, M. (eds.), *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 9, Norwell, MA, USA, Kluwer Academic Publisher, 1998.

[6] ESQUEF, P. A. A., *Spectral-Based Sound Synthesis - A Review*, Report, HUT, Laboratory of Acoustics and Audio Signal Processing, Helsink, Finland.

[7] ESQUEF, P. A. A., Biscainho, L. W. P., VÄLIMÄKI, V., "AN Efficient Algorithm for the Restoration of Audio Signals Corrupted with Low-Frequency Pulses", *Audio Eng. Soc.*, v. 51, n. 6, pp. 502 – 517, 2003.

[8] SMITH, III, J. O., SERRA, X., *PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation*, Report, CCRMA, Department of Music, Stanford University, Stanford, California, USA.

[9] QUATIERI, T. F., MCAULAY, R. J., "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Transactions ons Acoustics, Speech, and Signal Processing*, , n. 4, pp. 744 – 754, 1986.