# Explorando envoltórias espectrais em sistemas musicais interativos

*José Henrique Padovani*
*Escola de Música da UFMG*
e-mail: josepadovani@ufmg.br

*Sérgio Freire*
*Escola de Música da UFMG*
e-mail: sfreire@musica.ufmg.br

**Sumário:**
Tradicionalmente, os sistemas musicais interativos dependiam de comandos de controle – que, a partir de meados da década de 1980, passaram a utilizar predominantemente o protocolo MIDI – e hardwares específicos de síntese e processamento. Há pouco mais de uma década tornou-se possível o uso de computadores pessoais para o processamento de sinais de áudio em tempo real. O presente trabalho apresenta duas possibilidades de exploração de envoltórias espectrais obtidas pelas técnicas de deconvolução por filtragem de *cepstrum* e pela redução das informações espectrais de uma FFT a quartos de oitava. Estas abordagens abrem possibilidades de aplicação em sistemas musicais interativos e na obtenção de dados de controle a partir de sinais de áudio.

**Palavras-Chave:** sistemas musicais interativos, transformada de Fourier, cepstrum, deconvolução, processamento digital de sinais.

## Introdução

Recentemente a computação musical ganhou forte impulso com ambientes computacionais voltados à interação musical como Max/MSP1, Pure Data2 e JMax3. Em seus estágios iniciais, esses ambientes lidavam apenas com fluxo de informações discretas para controle – em sua maior parte no protocolo MIDI – utilizando como sintetizadores ou processadores de áudio hardwares específicos4. A implementação do processamento digital de sinais vem ganhando espaço em sistemas interativos desde o início da década de 1990. Desde meados de 1996 estes sistemas permitem procedimentos de manipulação de áudio no domínio do tempo e no domínio das freqüências em computadores pessoais[5].

Nesta comunicação de pesquisa apresentamos duas implementações de algoritmos de processamento que lidam com a extração de envoltórias espectrais de sinais de áudio em tempo real.

## A transformada de Fourier

Em processamento de sinais, a ferramenta mais largamente utilizada para a manipulação de um sinal no domínio das freqüências é a transformada de Fourier (FT). Neste trabalho vamos nos referir à transformada rápida de Fourier (FFT), uma implementação específica da FT em sistemas computacionais que se caracteriza por utilizar análises de períodos curtos (STFT) e pela eficiência

[1] Informações sobre Max/MSP podem ser acessadas a partir do site http://www.cycling74.com.

[2] Informações sobre Pure Data podem ser acessadas em http://www.puredata.info/.

[3] Informações sobre JMax podem ser acessadas em http://freesoftware.ircam.fr/rubrique.php3?id_rubrique=14.

[4] Cf. Rowe, 1993, p.9-38.

[5] Cf. Puckette, 2002, p.6.

de processamento quando comparada à implementação da FT clássica em sistemas discretos (DFT)6.

Pode-se dizer que todo sinal digital de áudio é um conjunto de amostras ordenadas no tempo, formando uma função que representa a variação da amplitude desse sinal. Podemos considerar tal função como uma somatória de $k$ ondas senoidais, cada uma com amplitude $A_k$, fase $\phi_k$ e freqüência $f_k$. O que a FFT permite é justamente a descrição de um sinal qualquer a partir destes parâmetros.

Desta maneira, um sinal descrito no domínio do tempo pela somatória de $k$ senóides a partir da fórmula

$$y(t) = \sum_{k=1}^{5} \frac{A}{k} \cos(2\pi k f t + \phi_k)$$

– onde todas as fases $\phi_k$ nulas sejam consideradas nulas – pode ser representado no domínio do tempo e, a partir dos valores obtidos por seu processamento em uma FFT, no domínio das freqüências:

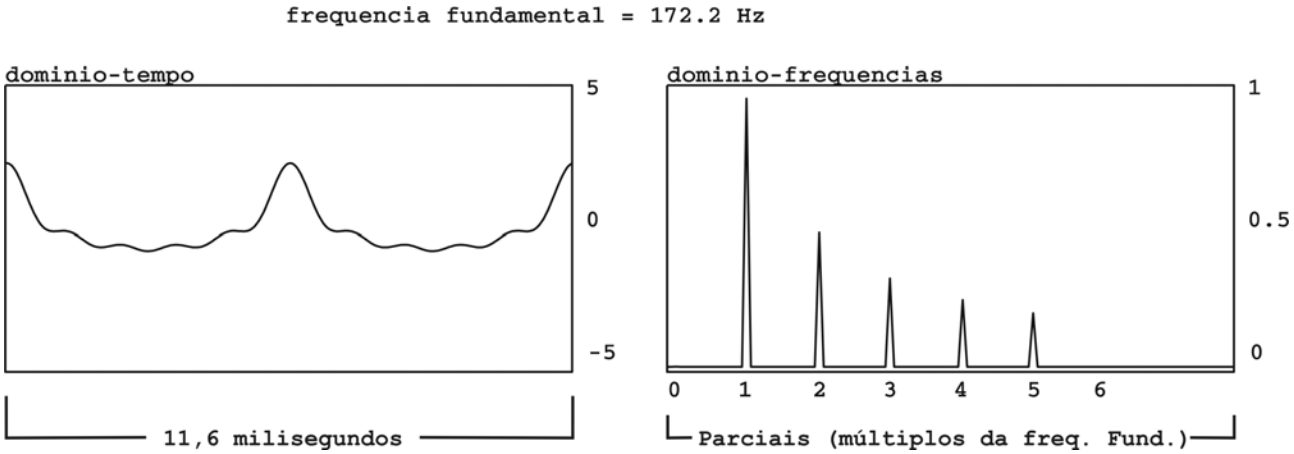


*Figura 1: Representações de um sinal no domínio do tempo e no domínio das freqüências.*

Dentre as inúmeras aplicações da transformada de Fourier para o processamento em tempo real de áudio (como o vocoder de fase ou os redutores de ruído), vale a pena ressaltar duas possibilidades que estão diretamente ligadas ao trabalho aqui exposto. Estas possibilidades são a de se filtrar com extrema precisão faixas de freqüência de um espectro – a partir da redução da amplitude de regiões delimitadas por esse filtro FFT – e a de se realizar convoluções entre sinais diferentes.

Filtros baseados em FFT agem de maneira pontual nas freqüências de um sinal analisado, possibilitando a transformação espectral para a ressíntese. Uma FFT que trabalha em blocos de análise de $N$ amostras de um sinal em freqüência de amostragem $f_{sr}$ é capaz de

[6] Devido ao caráter breve desta comunicação, não entraremos em detalhes matemáticos para explicar minuciosamente a FT, a DFT, a STFT e a FFT – o que se pode pesquisar com facilidade a partir da bibliografia existente. Nesta comunicação nos limitaremos a descrever de maneira geral o que faz e possibilita o algoritmo da FFT em sistemas musicais interativos.

ornecer $N/2$ faixas com largura de banda igual a $f_{sr}/N$ [7]. Estas faixas são cunhadas como os *bins* da FFT, sendo numeradas pela sua ordem no bloco de análise. No caso acima, temos 1024 *bins* (numerados de 0 a 1023), sendo o *bin* de ordem zero correspondente ao *DC offset*, o de ordem 1 à freqüência central de 43,06 Hz, o de ordem 2 à freqüência central de 86,12 Hz, e o de ordem 512 igual à freqüência central de 22050 Hz[8].

A resposta a impulso (*impulse response*) de qualquer sistema linear é o sinal resultante da aplicação de um único impulso em sua entrada. Se considerarmos um sinal como uma seqüência de impulsos, cada um deles gerará uma resposta defasada e modificada em amplitude, que se somará aos demais. A esse processo, dá-se o nome de convolução. No domínio das freqüências, isto equivale à multiplicação das amplitudes do espectro do sinal de entrada pela resposta a impulso do sistema a cada faixa de freqüência. Em uma FFT, a convolução pode ser facilmente implementada multiplicando-se, entre si, as amplitudes dos respectivos *bins* de dois sinais analisados – seguida da FFT inversa para retorno ao domínio do tempo. A reverberação é um bom exemplo de convolução entre uma fonte sonora qualquer e a resposta a impulso de um ambiente acústico real ou fictício. Ao processo que reverte os efeitos de uma convolução se dá o nome *deconvolução*. Conhecendo-se o resultado sonoro da convolução e a resposta a impulso do sistema em questão, pode-se obter o sinal da fonte sonora original através deste processo reverso.

## Obtendo envoltórias espectrais de um sinal

### *Obtendo envoltórias por deconvolução de cepstrum*

Tempelaars[9] define a análise de *cepstrum* como uma técnica que permite inferir informações sobre a resposta a impulso de um sistema linear e sobre o sinal a este aplicado, utilizando como material de análise apenas o sinal resultante. Para isso, é necessário algum conhecimento prévio das características do sinal de entrada e da resposta a impulso. Esta técnica foi introduzida primeiramente em 1963, por Bogert, Healy e Tukey[10] e tinha como finalidade discernir se as alterações em dados sismológicos correspondiam a explosões ou terremotos, tendo sido seu desenvolvimento intimamente ligado às negociações do tratado de proibição de testes nucleares[11]. No ano seguinte, o pesquisador e artista digital Michael Noll publica um trabalho[12] descrevendo a aplicação da análise de *cepstrum* para a detecção de alturas e de presença/ausência de componentes vocais em sinais de áudio.

Em termos computacionais, pode-se considerar que o espectro é uma forma de onda, passível de ser analisada a partir de seus componentes e periodicidades. A técnica de análise de *cepstrum* faz exatamente isto, possibilitando separar o contorno espectral de um sinal analisado em componentes "lentos" – relacionados à resposta de impulso de um sistema – e "rápidos" – relacionados ao sinal excitador.

---

[7] Como exemplo, para $N = 1024$ e $f_{sr} = 44100$ temos uma somatória de 512 faixas de freqüência de largura de banda de 43,06 Hz com valores de amplitude e fase individuais entre 0 e 22050 Hz, além de um valor constante (*DC offset*), que indica o grau de assimetria entre as partes positiva e negativa do sinal no domínio do tempo.

[8] Tratando-se de sinais reais (sem parte imaginária, como é o caso do áudio), os bins acima de $N/2+1$ não apresentam informação espectral adicional por serem simétricos aos valores inferiores a $N/2$.

[9] Cf. Tempelaars, 1996, p. 267.

[10] apud Tempelaars (1996, p. 268). *Bogert*, B.P.; Healy, M.J.R.; Tukey, J.W. (1963) "Quefrency analysis of time series of echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking". In: *Proceedings of the Symposium on Time Series Analysis*. pp. 209-243. New York: Wiley.

[11] Cf. Brillinger, 2002, p. 1606.

[12] Cf. Noll, 1964, pp. 296.

Tomemos como exemplo um sinal vocal $y(t)$. Este sinal pode ser entendido como a convolução do sinal de uma fonte excitadora $x(t)$ (relacionado aos impulsos gerados pelas pregas vocais) e a resposta a impulso $h(t)$ (relacionada à filtragem ressonante do trato vocal):

$y(t) = x(t) * h(t) \rightarrow$ convolução (domínio do tempo).

No domínio das freqüências consideramos este como o resultado de uma multiplicação:

$Y(f) = X(f) \times H(f) \rightarrow$ - multiplicação no domínio das freqüências.

Para separarmos os dois componentes desta função é necessária a utilização de um filtro. Contudo, a operação subtrativa de filtragem realizada pelos filtros não é capaz de separar os componentes do produto $Y(F)$. Isso é resolvido com a utilização de logaritmos, que transformam a relação de multiplicação em uma soma:

$$\log Y(f) = \log X(f) + \log H(f)$$

Para a subtração utiliza-se uma nova FFT a partir da qual se pode multiplicar por zero as amplitudes dos *bins* que não integram a curva de ressonância ou o sinal excitador procurados. Após realizar esta segunda FFT, deve-se retornar ao domínio do tempo realizando procedimentos inversos aos realizados até então. O esquema geral do processamento é o seguinte:

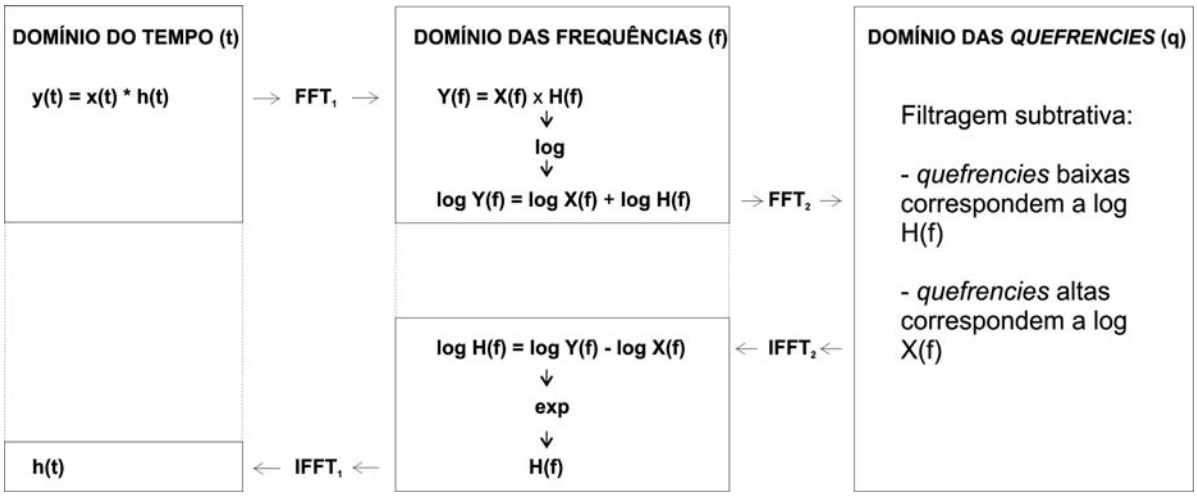


*Figura 2: Esquema geral da deconvolução por filtragem de cepstrum.*

Como o *cepstrum* é obtido a partir do logaritmo das amplitudes de cada faixa de freqüência da FFT anterior (i.e., antes de se realizar a FFT inversa, que converteria os dados freqüenciais em dados temporais), os autores do estudo original sobre esta técnica cunharam novos nomes para os parâmetros e procedimentos aos quais esta segunda FFT está relacionada. Assim, o espectro do logaritmo de um espectro é chamado de *cepstrum*, e neste novo domínio falamos de *quefrency*, *saphe*, *tilfering*, anagramas correspondentes às palavras freqüência, fase, filtragem em inglês. Destes nomes sobrevivem no meio técnico-acadêmico *cepstrum* e *quefrency*. Estas palavras servem para evitar a confusão de se dar o mesmo nome a questões diferentes: uma faixa de *quefrencies*, por exemplo, não está relacionada com as freqüências que ouvimos em um som qualquer, mas com

"parciais" de uma "forma de onda" que é formada pelo logaritmo das amplitudes das faixas de freqüência do espectro de um sinal analisado.

De fato, em termos computacionais, um vetor representando um sinal não está no domínio do tempo ou das freqüências, mas no domínio digital. A FFT de um sinal permite que se represente um sinal complexo a partir de componentes senoidais. No caso de um sinal que representa o comportamento espectral de um som, o que se encontra no *cepstrum* é a descrição de amplitudes de senóides determinadas que, somadas, correspondem ao logaritmo das amplitudes de cada faixa frequencial desta representação.

Vale dizer que esta implementação do *cepstrum* em tempo real abre possibilidades para outras implementações relacionadas à análise da resposta de impulso em sistemas lineares, tal como é o caso de sua aplicação na análise e processamento de reverberações.

### *Obtendo envoltórias por redução de informações espectrais a quartos de oitava*

O sinal gerado pela análise espectral via FFT apresenta uma discriminação de freqüências que cresce do grave para o agudo, dobrando sua definição a cada oitava: um valor para a primeira oitava, dois para a segunda, quatro para a terceira, oito para a quarta, dezesseis para a quinta etc. Com uma janela de análise de 1024 pontos, a última oitava é representada por 512 pontos, um número excessivo tanto para definir envoltórias quanto para servir de modelo de comparação com outros sinais. Procuramos, então, realizar uma filtragem desse sinal, de modo que cada oitava seja representada, no máximo, por quatro valores, por meio da média aritmética de valores vizinhos. Assim a envoltória espectral resultante equivale aos coeficientes de um filtro de quarto de oitava.

A figura abaixo demonstra a diferença de resolução de um espectro sonoro com valores de freqüência reduzidos a quartos de oitava e em sua resolução original a partir da FFT:

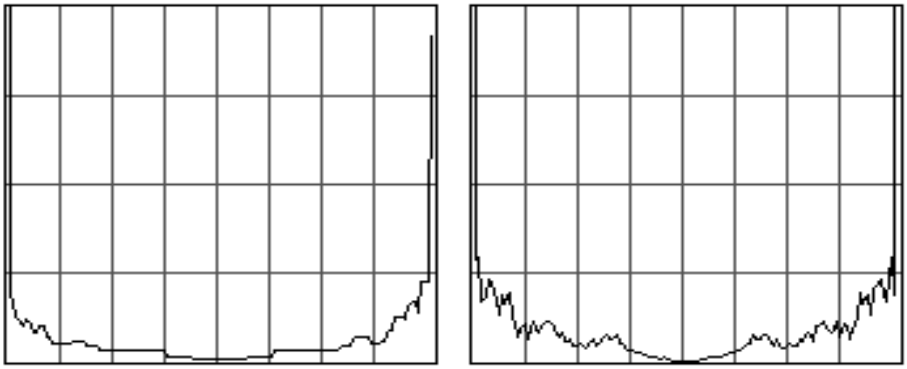

*Figura 3: O gráfico à esquerda mostra o espectro com resolução reduzida a quartos de oitava e o da direita mostra sua plotagem original. Nas figuras acima todos os bins das FFTs estão representados, daí a simetria do espectro.*

## Implementação em tempo real

### *a) cepstrum*

Foram elaborados algoritmos nos ambientes Max/MSP e Pure Data que realizam este processo em tempo real. A *quefrency* de separação é tomada por estimativa. De maneira geral, valores de controle para uma separação das *quefrencies* "graves" variando entre 15 e 35 *bins* – para uma janela de análise de 512 amostras – permitem uma reconstrução arredondada da envoltória, caracterizando um espectro constituído por regiões formânticas. Valores abaixo de 20 tendem a uma sonoridade de filtro passa-baixa enquanto valores superiores acabam por oferecer um resultado sonoro muito semelhante ao de uma convolução comum, isto é, com um delineamento de envoltória mais inconstante e aperiódico.

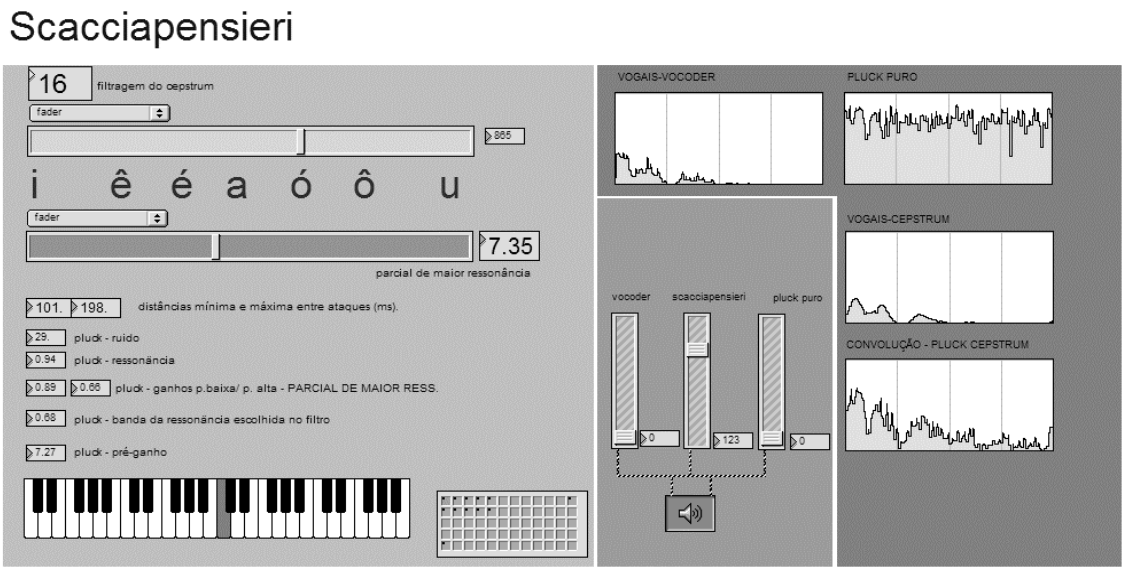


*Figura 4: Tela do algoritmo Scacciapensieri no ambiente Max/MSP. Note-se o espectro mais "arredondado" da envoltória extraída, plotada no segundo gráfico de cima para baixo à direita, quando comparado ao espectro original, representado no gráfico à esquerda.*

O primeiro algoritmo realizado sintetiza um berimbau de boca (*scacciapensieri*) a partir da deconvolução de vogais pré-gravadas, da síntese de sons pinçados a partir do algoritmo de Karplus-Strong e da convolução destes dois sinais. Isto faz com que o som pinçado (*pluck*) do algoritmo de Karplus-Strong ganhe o colorido das formantes que caracterizam as vogais pré-gravadas. Para aperfeiçoar o efeito é usado um filtro ressonante que faz com que certos harmônicos brilhem com mais vigor (esta filtragem ressonante é uma característica marcante no resultado sonoro da performance dos *scacciapensieri*). Para o controle de qual vogal irá soar em determinado momento é utilizado um vocoder de fase que percorre um arquivo que contém informações espectrais das vogais [a], [e], [i], [o] e [u], ou através de vogais capturadas em tempo real por um microfone.

O segundo algoritmo realiza a deconvolução de um sinal vocal de fala. Em seguida utiliza-se este como envoltória na convolução com arquivos pré-gravados ou com ruído branco. No primeiro caso a voz ganha colorido de clarineta, violoncelo, ou do instrumento pré-gravado em questão. No segundo caso, a voz ganha a característica de sussurro ou rouquidão (dependendo da *quefrency* de corte utilizada para a deconvolução do sinal vocal original). Pode-se, também, usar sinais capturados em tempo real tanto para se realizar a deconvolução inicial quanto para realizar a convolução que se dá em seguida com a envoltória obtida pela filtragem de *cepstrum*. Este algoritmo foi implementado na plataforma Max/MSP e tem uma versão preliminar no ambiente Pure Data[13].

[13] Vale dizer que, apesar da semelhança e da origem comum, os ambientes Max/MSP e Pure Data apresentam diferenças consideráveis quanto ao modo de trabalhar com dados de controle e sinais de áudio, além de possuírem objetos e bibliotecas diferentes.

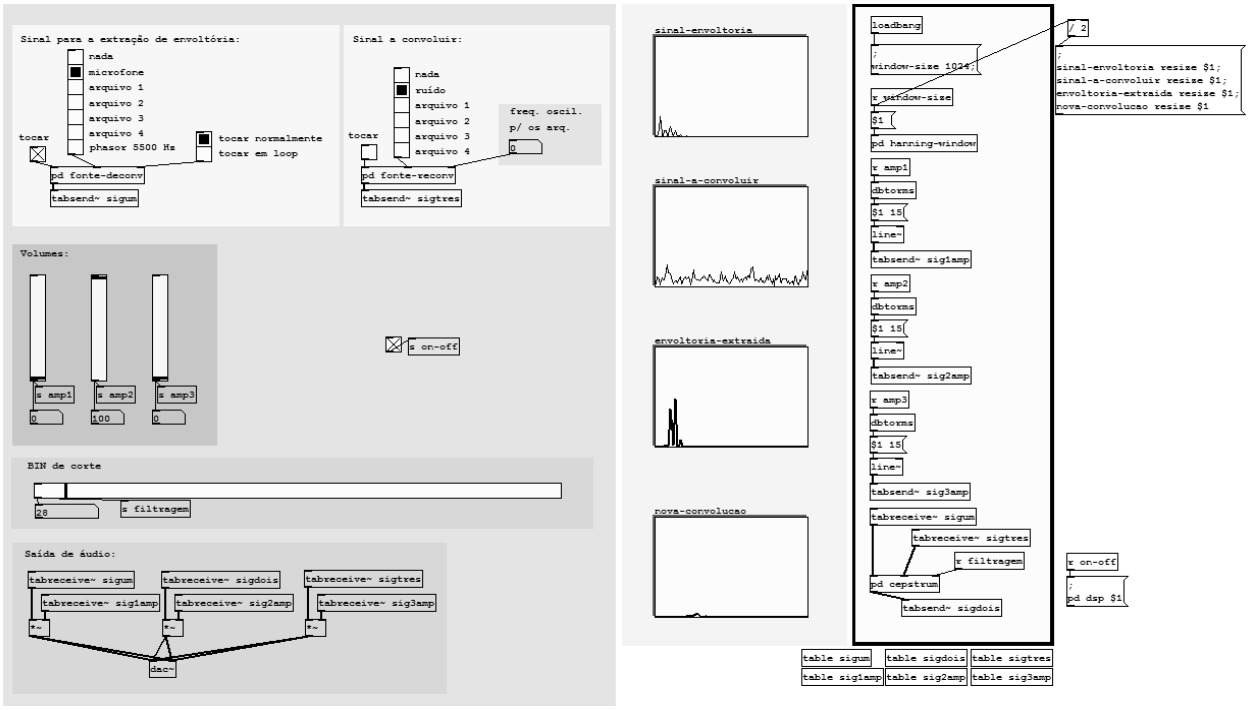


*Figura 5: Tela da versão preliminar do algoritmo de deconvolução e reconvolução por filtragem de cepstrum no ambiente Pure Data.*

### *b) redução espectral a quartos de oitava*

A redução espectral a quartos de oitava não necessita de explicações técnicas adicionais quanto às questões de processamento envolvidas em sua implementação; basta lembrar que o espectro de amplitudes assim "filtrado" pode ser facilmente multiplicado pelo espectro de outro sinal de entrada, imprimindo a este suas características espectrais globais. Essa redução oferece ainda a possibilidade de uma comparação bastante confiável entre a envoltória de um sinal de entrada e envoltórias previamente armazenadas (dentro de uma mesma execução). Esse tipo de reconhecimento – baseado no próprio áudio, sem limitações de complexidade do sinal de entrada – pode ser um elemento valioso na composição e performance de música interativa.

A extração de envoltórias espectrais em tempo real, seja por filtragem baseada em *cepstrum* ou por filtragem a ¼ de oitava, apresenta uma latência mínima devida às análises espectrais envolvidas. No caso do *cepstrum*, uma análise FFT feita com 1024 pontos e superposição de 2 janelas, essa latência é de cerca de 35 ms (com uma freqüência de amostragem de 44.1 KHz). Esse valor se refere à extração da envoltória; a convolução de um sinal de entrada com uma envoltória (preferencialmente de evolução lenta) se dá, nesse mesmo caso, com um atraso de 11,6 ms, valor aceitável para uma grande tipologia de sons musicais.

## Conclusão

As envoltórias extraídas pelos dois processos possibilitam a utilização do sinal obtido como controlador dinâmico de informação pois implicam numa redução de informações espectrais cuja utilização é de utilização mais simples que aquela de um sinal complexo. Diferentemente do que ocorria há pouco mais de uma década – quando o protocolo MIDI e as mensagens discretas eram essenciais para a música interativa –, o processamento digital de sinais e suas técnicas permitem que o sinal de áudio se torne, ele mesmo, um elemento controlador em sistemas interativos.

## Referências Bibliográficas

Brillinger, David R. (2002). “John W. Tukey's Work on Time Series and Spectrum Analysis”. The Annals of Statistics. 30 2002, 1595-1618.

Noll, Michael. (1964). “Short-time Spectrum and ‘Cepstrum’ Techniques for Vocal-Pitch Detection”. The Journal of American Society of Acoustic. V.36 n.2, 296-302.

Puckette, Miller (2002). “Max at seventeen”. Disponível em http://www-crca.ucsd.edu/~msp/Publications/dartmouth-reprint.pdf. Acessado em 17 de maio de 2006. Publicado originalmente em 2002: Computer Music Journal 26/4, pp. 31-43.

Rowe, Robert (1992). Interactive Music Systems: Machine Listening and Composing. Cambridge, London: The MIT Press.

Tempelaars, Stan. (1996). Signal processing, speech, and music. Lisse: Swets & Zeitlinger.