Decomposição Psicoacústica de Sinais de Áudio com Base em Dicionários Redundantes e Exponenciais Complexas

Valmir dos Santos Nogueira Jr, Michel Pompeu Tcheou, Flávio Rainho Ávila
Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil
e-mail: vsnjunior@gmail.com, mtcheou@uerj.br e flavio.avila@uerj.br

Abstract - *The atomic decomposition of signals by algorithms of the class "Matching Pursuit"(MP) has been applied in audio. According to the literature, the use of psychoacoustic criteria allows more compact representation of the signal, with minimal loss of perceived quality. This work describes a scheme of analysis and synthesis of audio signals using MP with psychoacoustic elements inspired by the MPEG layer I, in addition to Complex Exponential Dictionaries. Its performance is evaluated by an objective measure described by the ITU, the PEAQ, and by tests in terms of the number of coefficients of the representation.*
Palavras-chave: *Matching Pursuit, Signal Decomposition, Psychoacoustic.*

Introdução

Uma poderosa ferramenta de decomposição de sinais, introduzida por [1], é o *Matching Pursuit* (MP). Trata-se de um algoritmo que calcula, iterativamente, a expansão do sinal em funções (ou átomos), selecionando em um dicionário, aquelas que melhor se correlacionam com as estruturas do sinal em análise. Neste sistema, o sinal é decomposto em formas de ondas selecionadas deste dicionário de átomos de tempo-frequência, com uma coleção de dilatações e modulações de uma função janela simples. Para a obtenção de uma representação compacta do sinal de áudio original, que utilize somente os componentes realmente audíveis, é preciso entender os aspectos psicoacústicos do sistema auditivo humano [2–4].

O presente trabalho tem como objetivo realizar a decomposição de sinais de áudio, com o auxílio do algoritmo do *Matching Pursuit*, utilizando o princípio de relevância psicoacústica dos componentes do sinal. A elaboração do trabalho foi baseada no artigo [2], no qual o autor utiliza o MP com dicionário de exponenciais complexas e ponderação psicoacústica. No lugar da função de ponderação, usa-se diretamente a curva psicoacústica obtida no modelo MPEG-1 (camada I) [5]. À medida que a energia do resíduo obtido no processo iterativo do MP, passa a estar abaixo da máscara psicoacústica em determinadas faixas espectrais, as exponenciais complexas de frequências referentes a essas faixas são removidas do dicionário. A avaliação dos resultados obtidos foi feita com uma ferramenta de avaliação perceptiva de qualidade de áudio, o PEAQ *(Perceptual Evaluation of Audio Quality)*.

Materiais e métodos

A psicoacústica tem como objetivo principal fazer com que as magnitudes das sensações sejam equivalentes às magnitudes dos estímulos [6]. Ela apresenta características fundamentais na concepção de um codificador perceptivo de áudio, dentre as quais podem-se destacar as unidades de medida de níveis de pressão sonora (SPL), os limiares da audição humana, fenômenos de mascaramento e a escala Bark.

O limiar de audição representa o menor nível de pressão sonora em decibéis que se pode ouvir em uma dada frequência. Em fenômeno de mascaramento um componente de sinal mascarador altera o limiar auditivo, os estímulos físicos apenas produzem sensações auditivas se suas magnitudes físicas se enquadrarem acima desse novo limiar [6]. A Fig. 1 apresenta um componente de sinal mascarador alterando o limiar auditivo e impedindo que componentes de frequências vizinhas com menor SPL sejam percebidos.

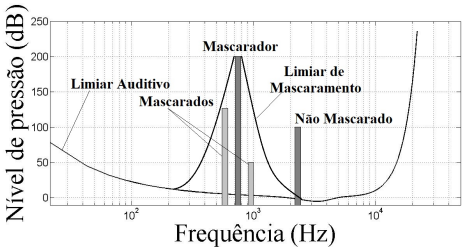

Figura 1 – Ilustração do fenômeno de mascaramento. Figura inspirada a partir de [7].

O modelo psicoacústico do trabalho foi inspirado na camada I do padrão MPEG-1 *(Moving Pictures Experts Group)* [5], onde o cálculo realizado pelo algoritmo do limiar de mascaramento psicoacústico global é consolidado em seis etapas: (1) Conversão do sinal do domínio do tempo para o domínio da frequência, através de uma transformada rápida de Fourier, (2) Obtenção da curva

de limiar de silêncio, (3) Determinação dos componentes tonais e não-tonais do sinal em análise, (4) Determinação dos componentes mascaradores principais, (5) Cálculo do limiar de mascaramento para cada componente mascarador e (6) Determinação do limiar de mascaramento global.

De grande importância, a decomposição atômica de sinais tem como objetivo selecionar um subconjunto de elementos, denominados átomos ou estruturas, a partir de um dicionário de formas de onda pré-definidas, a fim de aproximar o sinal como uma combinação linear desses elementos [1]. Considerando que um sinal x pode ser aproximado por átomos $\mathbf{g}_\gamma$ pertencentes a um dicionário D de modo que:

$$\mathbf{x} \approx \sum_{i=0}^{M-1} \alpha_i \mathbf{g}_{\gamma(i)} \tag{1}$$

Os átomos g_γ são indexados por $\gamma(i)$, que é definido como $\gamma : Z+ \rightarrow \{1, ..., \#D\}$; $\#D$ é o número de elementos do dicionário D, portanto $\gamma(i) \in \{1, ..., \#D\}$. O parâmetro α_i é o coeficiente que pondera $\mathbf{g}_{\gamma(i)}$ e M corresponde ao número de átomos selecionados para representar x.

Na prática, buscam-se representações compactas, isto é, representações que utilizem o menor número de átomos, para um dado nível de acurácia na representação.

O Matching Pursuit é um algoritmo que decompõe um sinal e o representa como uma expansão linear de formas de ondas [1]. A cada etapa, o algoritmo procura em seu dicionário uma forma de onda que combina melhor com o sinal atual e a subtrai do sinal inicial deixando um resíduo de sinal. O Matching Pursuit continua a ser aplicado nesse sinal residual até que seu critério de parada seja encontrado.

Desejamos representar um sinal de dimensão N em um conjunto de M coeficientes, onde $M < N$ [1]. Um dicionário redundante apresenta uma cardinalidade maior que a dimensão N do sinal, propiciando alto grau de liberdade na construção da expansão de funções.

Cada átomo $g_{\gamma i}$ é caracterizado por parâmetros de escala (s), deslocamento (τ) e de frequência (ξ), através de aproximações sucessivas de x com projeções ortogonais sobre os elementos do dicionário. O sinal x é decomposto de forma iterativa, de acordo com a seguinte expressão:

$$\mathbf{r}_x^k = < \mathbf{r}_x^k, \mathbf{g}_{\gamma k} > \mathbf{g}_{\gamma k} + \mathbf{r}_x^{k+1} \tag{2}$$

onde $\mathbf{r}_x^{k+1}$ é o resíduo da "k+1"-ésima iteração, $\mathbf{r}_x^k$ é o sinal a ser decomposto e o operador $<,>$ representa o produto interno [1]. O átomo é selecionado da seguinte forma:

$$\gamma_k = \arg\max_{\gamma \in \Gamma} | < \mathbf{r}_x^k, \mathbf{g}_{\gamma k} > | \tag{3}$$

O vetor original x é aproximado como uma soma ponderada de elementos selecionados no dicionário, a cada iteração, como descrito na equação (1).

O dicionário de exponenciais complexas é utilizado devido ao fato dele permitir encontrar informação de fase do átomo. Seus elementos são definidos da seguinte maneira [2]:

$$\mathbf{g}_{\gamma(i)} = \left\{ g_{\gamma(i)}[n] = \frac{1}{N} e^{j2\pi \frac{m}{M} n} \right\} \tag{4}$$

onde $n = 0, 1, ...N-1$ e $m = 0, 1, ...M-1$. Com blocos de duração de aproximadamente $11,6$ ms (frequêcia de amostragem igual a 44100 Hz) o sinal se comporta de forma estacionária, o que torna adequada a utilização da amplitude constante do dicionário.

O processamento do sinal é realizado bloco a bloco, sendo que cada bloco correspode a um trecho janelado do sinal, podendo haver sobreposição entre blocos adjacentes. Nesse caso, o sinal é reconstruído através de um procedimento de sobreposição e adição entre os blocos.

Em [2], propõe-se uma implementação baseada em Transformada Discreta de Fourier (DFT) que permite utilizar algoritmos rápidos como a FFT *(Fast Fourier Transform)*. Para aproveitar ao máximo a FFT, a cardinalidade M do dicionário deve ser potência de 2. Inicialmente, introduz-se uma generalização do produto interno para um produto interno ponderado, $< \mathbf{x}, \mathbf{y} >_W = \mathbf{yWx}$, onde W é uma matriz positiva definida simétrica, de formato diagonal, em que a diagonal é composta pelas amostras da janela $W[n]$. Na k-ésima iteração do MP, calcula-se, para $m = 0, 1, ...M-1$,

$$\frac{| < \mathbf{g}_m, \mathbf{r}_k >_W |}{< \mathbf{g}_m, \mathbf{g}_m >_W} = \frac{R_k^w \left[\frac{m}{M}\right]}{W \left[\frac{0}{M}\right]} \tag{5}$$

onde

$$R_k^w \left[\frac{m}{M}\right] = \sum_{n=0}^{M-1} w[n] r_k[n] e^{-j2\pi \frac{m}{M} n}, \tag{6}$$

e

$$W \left[\frac{0}{M}\right] = \sum_{n=0}^{M-1} w[n]. \tag{7}$$

e, em seguida, busca-se o máximo produto interno normalizado

$$m_k = \arg\max_m \frac{| < \mathbf{g}_m, \mathbf{r}_k >_W |}{< \mathbf{g}_m, \mathbf{g}_m >_W} \tag{8}$$

Portanto, o k-ésimo coeficiente é dado por

$$\alpha_k = \frac{R_k^w \left[\frac{m}{M}\right]}{W \left[\frac{0}{M}\right]} = \frac{A_k e^{j\theta_k}}{W \left[\frac{0}{M}\right]} \tag{9}$$

onde A_k e θ_k são o módulo e a fase de $R_k^w \left[\frac{m_k}{M}\right]$ respectivamente.

Dado que sinais de áudio são reais, $x_l[k]$ também será real. Neste caso, $R_k^w \left[\frac{m}{M}\right]$ é uma matriz Hermitiana, e é preciso calcular somente metade

das correlações a cada iteração, ou seja, variando $m = 0, 1, ..., \frac{M}{2}$.

O algoritmo calcula a cada bloco o limiar de mascaramento global, armazena-se a FFT (de comprimento M) da janela, $W\left[\frac{m}{M}\right]$; define-se $r_0 = x_l$; e calcula-se a FFT (de comprimento M) de $r_o[n]w[n]$, ou seja, $R_0^w\left[\frac{m}{M}\right]$. No processo do MP, realiza-se o seguinte procedimento a cada k iteração: (1) Dado que $R_k^w\left[\frac{m}{M}\right]$, encontre o valor máximo absoluto do conjunto resultante, obtendo assim os parâmetros senoidais de amplitude A_k, frequência f_k e fase θ_k; (2) calcula-se o resíduo no domínio da frequência da próxima iteração, $R_{k+1}^w\left[\frac{m}{M}\right]$.

$$R_k^w - \frac{A_k\left(e^{i\theta_k}W\left[\frac{m-m_k}{M}\right]\right)}{2} - \frac{A_k\left(e^{-i\theta_k}W\left[\frac{m+m_k}{M}\right]\right)}{2} \quad (10)$$

(3) se $R_{k+1}^w\left[\frac{m}{M}\right]$ em dB_{SPL} estiver abaixo do limiar global de mascaramento para todas as frequências, interrompe-se o processo iterativo e (4) removem-se do dicionário os elementos correspondentes às frequências com niveis de pressão (SPL) abaixo do limiar global de mascaramento.

Ao final do processo iterativo, obtém-se uma representação do bloco

$$x_l[n] = \sum_k^{Niter} 2\frac{A_k}{W\left[\frac{0}{M}\right]}\cos\left[2\pi\frac{m_k}{M}n + \theta_k\right] \quad (11)$$

onde Niter é o número de iterações, $a_k = \frac{A_k}{W\left[\frac{0}{M}\right]}$ e $f_k = \frac{m_k}{M}$. Observe que, a cada bloco, a complexidade de inicialização é $O(2M\log_2 M)$ em função das duas FFT's e, a cada iteração, $O(2M)$ em função das subtrações da equação (10).

O sinal de áudio original sob análise foi dividido em quadros de tamanho N, ponderados por uma função janela com saltos de H amostras. Neste trabalho, foram utilizadas duas funções janela, a retangular e de Hanning, com saltos de H=N para a janela retangular e para a janela de *Hanning* saltos de $H = N/2$ amostras. A escolha da janela de *Hanning* se deve ao fato dela oferecer boa resolução em frequência e dispersão espectral reduzida, melhorando assim a identificação dos componentes tonais e não-tonais no processo de obtenção do limiar global. Observe que a densidade espectral com a janela retangular (Fig. 2) resulta em diversas componentes tonais inexistentes quando estima-se a densidade usando-se a janela de Hanning (Fig. 3), especialmente em altas frequências.

Resultados

O critério de parada da decomposição utilizado se baseia no limiar global de mascaramento psicoacústico. Muitas vezes é necessário fazer uso de uma margem, a ser subtraída do limiar global, de modo a garantir que o resíduo da decomposição seja inaudível. Quando o critério de parada psicoacústico não é considerado, usa-se o número máximo de iterações definido corresponde ao tamanho do bloco, N.

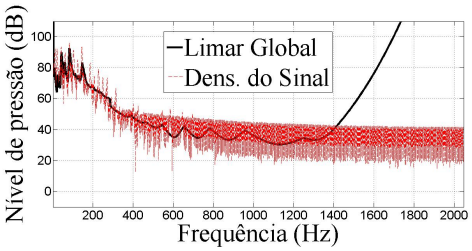

Figura 2 – Densidade espectral de potência do sinal com janela retangular.

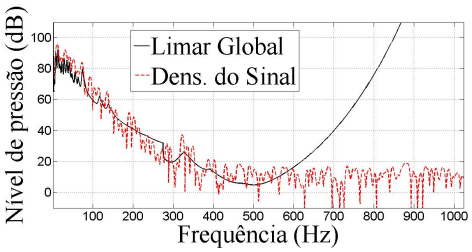

Figura 3 – Densidade espectral de potência do sinal com janela de *Hanning*.

A avaliação dos sinais reconstruídos é realizada através do algoritmo PEAQ, que é um algoritmo para medir objetivamente a qualidade de sinais de áudio, padronizado pelo *International Telecommunications Union* (ITU), na recomendação ITU-R BS.1387 [8]. Essa medida de qualidade é classificada nas seguintes faixas [9]: (1) -4 a -3: muito perturbador; (2) -3 a -2: perturbador; (3) -2 a -1: pouco perturbador e (4) -1 a 0: não perturbador.

Os sinais de áudio utilizados nos experimentos se referem a notas de instrumentos musicais: nota A3 de um piano, nota A4 de flauta, nota A4 de um violoncelo e nota A4 de um fagote. Foram utilizados dicionários de exponenciais complexas com redundância de 4 e 8 vezes o tamanho do bloco. Os sinais possuem 1 segundo de duração e são amostrados a uma taxa de 44,1 kHz, logo têm 44100 amostras. Para a decomposição, são divididos em blocos de 512 amostras com sobreposição de 256 amostras.

Os resultados da avaliação PEAQ para os testes realizados com dicionário com redundância de 4 e 8 vezes o tamanho dos quadros são apresentados na Tabela 1.

Tabela 1 – Valores do PEAQ para diferentes sinais decompostos com redundâncias de 4 e 8 vezes o número de amostras por bloco.

	Piano A3		Violoncelo A4		Fagote A4		Flauta A4	
Redundância / Margem (dB)	4	8	4	8	4	8	4	8
0	-1,3797	-1,4323	-2,9854	-2,0189	-1,9883	-2,4923	-2,8371	-2,0125
1	-1,2049	-1,2825	-2,4789	-1,7482	-2,0956	-1,6445	-2,904	-2,3696
3	-1,1036	-1,1104	-1,7289	-1,2079	-1,4777	-1,0831	-2,4916	-2,1056
6	-0,7239	-0,6662	-1,9303	-1,2899	-0,8779	-0,6548	-1,9057	-1,4405
9	-0,4163	-0,3616	-1,3581	-1,0140	-0,5884	-0,5345	-1,1401	-0,9209
12	-0,1439	-0,1208	-0,9295	-0,6809	-0,5564	-0,5126	-0,6588	-0,5578
Sem critério psicoacústico	0,3807	-0,3059	-0,8078	-0,8098	-1,0016	-0,6792	-1,3214	-1,1261

Discussão e Conclusões

É possível notar uma tendência geral de melhora utilizando um dicionário com maior redundância. Outro ponto observado é que quanto maior a margem subtraída do limiar global psicoacústico melhor é o resultado perceptivo da decomposição. Testes informais de audição também foram realizados. Notou-se que com a margem de 6 dB foram obtidas decomposições com boa qualidade, classificadas como pouco ou não perturbador.

Observe que ao não se adotar o critério de parada psicoacústico, obtém-se ótimas notas do PEAQ. No entanto, ao custo de sempre se utilizar 512 coeficientes na representação dos sinais a cada bloco. Com o critério de parada psicoacústico, é possível reduzir o número de coeficientes usados a cada bloco, obtendo-se avaliações de qualidade equivalentes.

Na continuação imediata deste trabalho serão realizados testes sistemáticos sobre um conjunto mais extenso de sinais de áudio, tais como outros instrumentos, orquestras e sinais de voz. Além disso, um processo de quantização deve ser incorporado ao algoritmo.

Referências

[1] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, Dez. 1993.

[2] T. S. Verma and T. H. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[3] A. Petrovsky, V. Herasimovich, and A. Petrovsky, "Scalable parametric audio coder using sparse approximation with frame-to-frame perceptually optimized wavelet packet based dictionary," in *AES 138th Convention*, IEEE, Mai. 2015.

[4] I. Toumi and O. Derrien, "Sparse decomposition of audio signals using a perceptual measure of distortion. application to lossy audio coding," in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, (Trondheim, Norway), pp. 1809–1812, NTNU, Nov. 2015.

[5] ISO, MPEG International Standard, "IEC 11172: Information technology-coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s," *Part1: Systems, Part2: Video, Part3: Audio*, 1993.

[6] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Springer, 2007.

[7] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*, vol. 721. Springer Science & Business Media, 2012.

[8] ITU-R, "BS.1387: Method for objective measurements of perceived audio quality," *International Telecommunication Union, Geneva, Switzerland*, 2001.

[9] P. Kabal, "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University*, pp. 1–89, 2002.