Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Valmir dos Santos Nogueira Junior

Codificação Perceptiva de Áudio através de Decomposições Atômicas em Exponenciais Complexas

Rio de Janeiro

2018

Valmir dos Santos Nogueira Junior

Codificação Perceptiva de Áudio através de Decomposições Atômicas em Exponenciais Complexas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre em Ciências, ao Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Estado do Rio de Janeiro. Área de Concentração: Sinais e Sistemas de Comunicações.

Orientadores: Michel Pompeu Tcheou, D.Sc.

Flávio Rainho Ávila, D.Sc.

Rio de Janeiro

2018

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

N778 Nogueira Junior, Valmir dos Santos.
Codificação Perceptiva de Áudio através de Decomposições Atômicas em Exponenciais Complexas. / Valmir dos Santos Nogueira Junior. – 2018.
85 f.

Orientadores: Michel Pompeu Tcheou, Flávio Rainho Ávila.
Dissertaçção(Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

1. Engenharia eletrônica - Teses. 2. Processamento de sinais - Teses. 3. Algorítmos - Teses. 4. Compressão de dados (Computação) - Teses. I. Tcheou, Michel Pompeu. II. Ávila, Flávio Rainho. III. Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. IV. Título.

CDU 621.391

Assinatura

Data

Valmir dos Santos Nogueira Junior

Codificação Perceptiva de Áudio através de Decomposições Atômicas em Exponenciais Complexas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre em Ciências, ao Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Estado do Rio de Janeiro. Área de Concentração: Sinais e Sistemas de Comunicações.

Aprovado em 16 de Agosto de 2018.

Banca Examinadora:

Michel Pompeu Tcheou, D.Sc. (Orientador)

Faculdade de Engenharia - UERJ

Flávio Rainho Ávila, D.Sc. (Orientador)

Faculdade de Engenharia - UERJ

Diego Barreto Haddad, D.Sc.

CEFET-RJ

João Baptista de Oliveira e Souza Filho, D.Sc.

COPPE - UFRJ

Rio de Janeiro

2018

AGRADECIMENTO

Agradeço a Deus por todas as experiências proporcionadas até hoje. A vida é um aprendizado e, sem elas, eu não conseguiria valorizar as oportunidades que tive. Agradeço à minha família pelo apoio que sempre me deu, em especial, aos meus pais: à minha mãe que sempre faz o máximo que pode por mim; ao meu pai que, infelizmente, não está mais neste plano, mas celebra essa alegria. Agradeço também a minha noiva, companheira e amiga Aline.

Agradeço ao Programa de Pós-Graduação em Engenharia Eletrônica da UERJ a oportunidade. Aos professores do laboratório de Processamento de Sinais, Aplicações Inteligentes e Comunicações, agradeço o apoio. Aos meus orientadores, agradeço os conhecimentos e, principalmente, ao meu orientador Michel Tcheou, que sempre demonstrou o máximo interesse, apoio e paciência, e esta última eu testei bem.

Agradeço aos meus amigos que contribuíram e aos que compreenderam as minhas ausências.

Portanto, não percam a coragem, pois ela traz uma grande recompensa.

Hebreus 10:35

RESUMO

NOGUEIRA JUNIOR, Valmir dos Santos. *Codificação perceptiva de áudio através de decomposições atômicas em exponenciais complexas*. 85 f. Dissertação (Mestrado em Engenharia Eletrônica) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, 2018.

A decomposição atômica de sinais por algoritmo da classe "Matching Pursuit" (MP) vem sendo aplicada à compressão de áudio. De acordo com a literatura, identificamos que se pode utilizar critérios psicoacústicos, o que possibilitaria uma representação mais compacta do sinal, sem perda de qualidade percebida. Este trabalho descreve uma implementação de um esquema de análise por síntese de sinais de áudio utilizando MP associado ao uso de limiar de mascaramento global psicoacústico, inspirado na camada I do MPEG, além de Dicionários de Exponenciais Complexas (DEC). Para a compressão do sinal utiliza-se a otimização taxa-distorção por curvas operacionais ajustando-se o multiplicador de Lagrange. O desempenho da representação para diversas famílias de sinais é avaliado por uma medida objetiva padronizada pelo ITU, o PEAQ, e através de testes em termos do número de coeficientes necessários para representação do sinal com fidelidade.

Palavras-chave: Matching Pursuit; decomposição atômica de sinais; psicoacústica.

ABSTRACT

NOGUEIRA JUNIOR, Valmir dos Santos. *Perceptual Encoding of Audio through Atomic Decompositions in Complex Exponentials*. 85 f. Dissertação (Mestrado em Engenharia Eletrônica) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, 2018.

The atomic decomposition of signals by algorithms of the class "Matching Pursuit"(MP) has been applied in audio compression. According to the literature, the use of psychoacoustic criteria allows a more compact representation of the signal, with minimal loss of perceived quality. This work describes a scheme of analysis by synthesis of audio signals using MP with direct use of the global psychoacoustic masking threshold, inspired by the MPEG layer I, in addition to Dictionaries of Complex Exponentials (DEC). For signal compression, we use the optimization rate-distortion via operating curves by adjusting the Lagrange multiplier. Its performance of representation is evaluated by an objective measure standardized by the ITU, the PEAQ, and by tests in terms of the number of coefficients needed for representation of signals with high-fidelity.

Keywords: Matching Pursuit; signal decomposition; psychoacoustic.

LISTA DE FIGURAS

Figura 1 Esquema de codificação e decodificação de sinais com base em decomposição atômica psicoacústica. 15
Figura 2 - Curva de limiar de audição humana. 19
Figura 3 Fenômeno de mascaramento na frequência. 20
Figura 4 Resultados experimentais de uma curva de mascaramento. 21
Figura 5 Limiares de mascaramento de ruídos de banda estreita mascarando tons. .. 22
Figura 6 Limiares de mascaramento de tons com 1 kHz e diferentes níveis mascarando tons. 23
Figura 7 Função de espalhamento linear adotada no modelo psicoacústico ISO / IEC MPEG l. 25
Figura 8 - Representação do limiar de silêncio. 27
Figura 9 - Representação da densidade espectral de potência do sinal. 28
Figura 10 - Representação do limiar de mascaramento tonal. 31
Figura 11 - Representação do limiar de mascaramento não-tonal. 32
Figura 12 - Representação de um limiar de mascaramento global. 32
Figura 13 - Representação do limiar de mascaramento global. 33
Figura 14 - Diagrama de blocos do esquema de medidas do PEAQ. 34
Figura 15 - Representação gráfica da projeção ortogonal no primeiro passo da decomposição de x. 39
Figura 16 - Esquema da representação da divisão do sinal em blocos. 41
Figura 17 - Densidade espectral do elemento do sinal com máxima correlação entre os elementos do dicionário e o resíduo do bloco do sinal de áudio. 43
Figura 18 - Densidade espectral em escala logarítmica do elemento do sinal com máxima correlação entre os elementos do dicionário e o resíduo do bloco do sinal de áudio. 44
Figura 19 - Densidade espectral do elemento do dicionário com máxima correlação entre o resíduo do bloco do sinal de áudio. 44
Figura 20 - Densidade espectral em escala logarítmica do elemento do dicionário com máxima correlação entre o resíduo do bloco do sinal de áudio. 45

Figura 21 - Remoção dos máximos encontrados no sinal de entrada: (a) em escala linear e (b) em escala logarítmica. 46
Figura 22 - Esquema geral de compressão de sinais. 48
Figura 23 - Compressão de sinais de áudio realizando a decomposição atômica do sinal via DEC e o binômio taxa-distorção através de curvas operacionais... 49
Figura 24 - Interpretação gráfica da otimização da função Lagrangeana..... 52
Figura 25 - Fecho Convexo contendo os pontos ótimos em termos de taxa-distorção. . 53
Figura 26 - Traçando o fecho convexo. Neste caso $\theta_{min} = \theta_1$. 53
Figura 27 - Traçando o fecho convexo. Neste caso $\theta_{min} = \theta_1$. 54
Figura 28 - Representação de um bloco de um sinal de áudio utilizando a janela: (a) Retangular e (b) Hanning..... 57
Figura 29 - Exemplos de diferentes limiares psicoacústicos em dB. 58
Figura 30 - Número de Iterações por Bloco com Redundância de dicionário de 4N para o Piano A3 61
Figura 31 - Número de Iterações por Bloco com Redundância de dicionário de 4N para a Flauta A4 62
Figura 32 - Número de Iterações por Bloco com Redundância de dicionário de 4N para o Fagote A4 62
Figura 33 - Número de Iterações por Bloco com Redundância de dicionário de 4N para o Violoncelo A4 63
Figura 34 - Número de Iterações por Bloco com Redundância de dicionário de 4N para a Bateria A 63
Figura 35 - Número de Iterações por Bloco com Redundância de dicionário de 4N para a Bateria B..... 64
Figura 36 - Densidade espectral, limiar global psicoacústico e resíduo final do sinal para os blocos: (a) 7 e (b) 115 67
Figura 37 - Histogramas dos sinais com margem de 6 dB e redundância de dicionário de 4N para: (a) Piano, (b) Violoncelo, (c) Fagote, (d) Flauta, (e) Bateria A e (f) Bateria 68
Figura 38 - Curvas de otimização do *piano a3* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread* 70

Figura 39 - Curvas de otimização do *violoncelo a4* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread* 71
Figura 40 - Curvas de otimização do *flauta a4* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread* 72
Figura 41 - Curvas de otimização do *fagote a4* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread* 73
Figura 42 - Curvas de otimização do *bateria a* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread* 74
Figura 43 - Curvas de otimização do *bateria b* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread* 75
Figura 44 - Curvas de otimização de todos os áudios da Taxa-Distorção do quantizador *midrase* 76
Figura 45 - Curvas de otimização de todos os áudios da Taxa-PEAQ do quantizador *midrase* 76
Figura 46 - Curvas de otimização de todos os áudios da Taxa-Distorção do quantizador midtread 77
Figura 47 - Curvas de otimização de todos os áudios da Taxa-PEAQ do quantizador midtread 77

LISTA DE TABELAS

Tabela 1 Valores do PEAQ para diferentes sinais decompostos com dicionários que possuem redundâncias de 4 e 8 vezes o número de amostras por bloco..... 60
Tabela 2 Valores do número médio de iterações para os diferentes sinais decompostos com dicionários que possuem redundâncias de 4 e 8 vezes o número de amostras por bloco..... 65
Tabela 3 Valores de taxa de bits por segundo com boa qualidade 78
Tabela 4 Valores do PEAQ dos diferentes sinais decompostos..... 80
Tabela 5 Valores do número médio de iterações dos sinais decompostos..... 80
Tabela 6 Valores de taxa de bits por segundo 81

LISTA DE SIGLAS

MP	Matching Pursuit
PEAQ	Perceptual Evaluation of Audio Quality
DEC	Dicionário de Exponenciais Complexas
MPEG	Moving Pictures Experts Group
MP3	Moving Pictures Experts Group Audio Layer 3
SPL	Sound Pressure Level
ATH	Absolute Threshold of Hearing
DFT	Transformada Discreta de Fourier
FFT	Fast Fourier Transform
SMR	Razão Sinal Máscara
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
ITU	International Telecommunication Union
CD	Compact Disc
MDCT	Modified Discrete Cosine Transform
ERB	Equivalent Rectangular Bandwidth

SUMÁRIO

	INTRODUÇÃO	14
1	**NOÇÕES DE PSICOACÚSTICA**	18
1.1	Fundamentos Teóricos	18
1.1.1	Nível de Pressão Sonora	18
1.1.2	Limiar Absoluto de Audição	19
1.1.2.1	Mascaramento	20
1.1.3	Escala Bark	20
1.2	MPEG-1	24
1.3	Algoritmo para Cálculo do Limiar de Mascaramento Psicoacústico Global	25
1.4	Avaliação Perceptiva da Qualidade de Áudio	33
2	**DECOMPOSIÇÃO ATÔMICA PSICOACÚSTICA**	35
2.1	Representações de Sinais	35
2.2	Decomposições Atômicas	36
2.3	Matching Pursuit	38
2.4	Dicionário de Exponenciais Complexas	40
2.5	Algoritmo de Decomposição	41
3	**ALOCAÇÃO ÓTIMA DE BITS**	47
3.1	Sistemas de Compressão	47
3.2	Quantização	49
3.3	Otimização Taxa-Distorção	50
4	**PROCEDIMENTOS E RESULTADOS EXPERIMENTAIS**	56
4.1	Procedimentos Experimentais	56
4.2	Resultados Experimentais	59
4.2.1	Decomposição Atômica	59
4.2.2	Alocação Ótima de Bits	66
5	**CONCLUSÕES**	79
5.1	Trabalhos Futuros	80

REFERÊNCIAS 82

INTRODUÇÃO

No processo de obtenção de representações compactas de sinais, deve-se buscar expandir esses sinais em funções-base que apresentem uma grande similaridade com suas estruturas complexas [1]. A modelagem de sinais torna possível descrever matematicamente, e de forma suficientemente acurada, seus fenômenos intrínsecos por meio de ferramentas que propiciam a análise e a síntese desses sinais. Uma poderosa técnica de decomposição de sinais, introduzida por [1], é o *Matching Pursuit* (MP). Trata-se de um algoritmo que calcula, iterativamente, a decomposição do sinal em funções (ou átomos), selecionando, a cada iteração, dentre um conjunto de funções que formam um dicionário, aquelas que melhor se correlacionam com o sinal em análise. Em nossa aplicação, relacionada à representação compacta de sinais de áudio, o sinal é decomposto em formas de ondas selecionadas deste dicionário de átomos de características tempo-frequenciais, que por sua vez é gerado, normalmente, a partir de dilatações, translações e modulações de uma função janela simples.

Para a obtenção de uma representação maximamente compacta do sinal de áudio original, é fundamental explorar aspectos perceptivos da audição, os quais informam que componentes tempo-frequenciais são realmente audíveis. Para tanto, é preciso entender conceitos de psicoacústica, a ciência da percepção sonora humana, em especial os conceitos de audibilidade e de mascaramento [2–4].

A Figura 1 apresenta um esquema de codificação e decodificação de sinais utilizando decomposições atômicas com base em um dicionário de funções elementares, incluindo aspectos psicoacústicos para a seleção dos átomos. Como veremos em mais detalhes em capítulos subsequentes, a entrada do modelo psicoacústico consiste na representação temporal do sinal de áudio durante um determinado intervalo de tempo, e a sua saída corresponde à razão sinal/máscara para cada componente de frequência do trecho de sinal analisado.

O codificador analisa o sinal de forma a encontrar uma boa representação com base em um dicionário e no modelo psicoacústico, alcançando assim os coeficientes e os índices da representação atômica da decomposição de cada quadro do sinal.

Para tornar uma representação compacta, é preciso garantir um número de bits baixo e uma alta *performance*. Nesse sentido, a quantização é responsável pela compres-

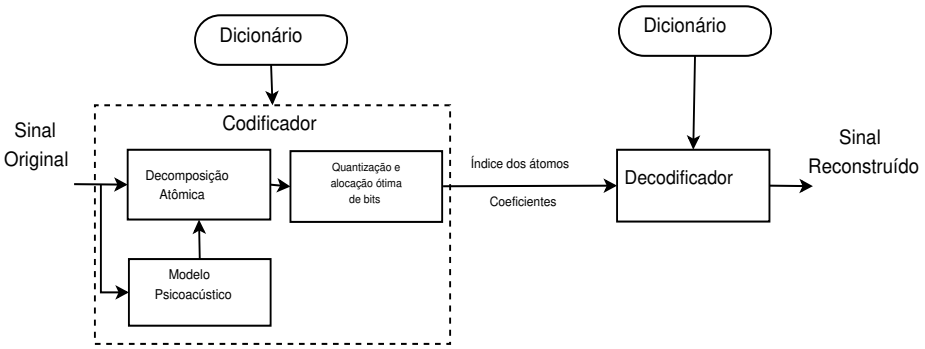


Figura 1 Esquema de codificação e decodificação de sinais com base em decomposição atômica psicoacústica.

são dos coeficientes das representações atômicas dos sinais de áudio. Para se obter tal objetivo é utilizada a otimização taxa-distorção por curvas operacionais, ajustando-se o multiplicador de Lagrange. A teoria por detrás desta metodologia está preocupada com a tarefa de representar uma fonte de dados com o menor número de bits possível, isto é, com um dado compromisso entre a taxa de bits utilizada e a *performance* por ela alcançável. Definir uma taxa-distorção aceitável é o compromisso fundamental no projeto de sistemas de compressão.

Os coeficientes e os índices da representação do sinal são transmitidos ao decodificador, onde é efetuada a reconstrução do sinal com base no mesmo dicionário usado no codificador por quadro e a sobreposição e soma dos sinais de todos os quadros, obtendo-se assim o sinal reconstruído final.

Em [5] o autor utiliza o Iterative Hard Threshold, que é um algoritmo de decomposição esparsa do sinal para decomposição de vários quadros simultaneamente no mesmo dicionário, com o objetivo de permitir impor a dispersão entre os quadros, e a dispersão de cada quadro. Uma abordagem baseada em um algoritmo variacional é demonstrada em [6]. A otimização de uma função de custo leva em conta tanto uma medida de distorção perceptiva derivada de um modelo de audição quanto uma restrição de dispersão. O método baseia-se no algoritmo de MP, realizando uma aproximação esparsa em um dicionário sobre completo, mais precisamente uma união de bases MDCT. Em [7] é proposto uma decomposição atômica tempo-frequancial para codificadores de áudio baseada no MDCT, chamada de ERB-MDCT. Basicamente, o ERB-MDCT é uma MDCT "não-estacionária no domínio da frequência" que segue a escala do ERB [7].

O algoritmo de MP com átomos de Gabor foi utilizado por [8] para investigar quantitativamente as características do Campo Receptivo Espectro-Temporal, medidos

fisiologicamente. O Campo Receptivo Espectro-Temporal representa uma aproximação linear às características de resposta de um neurônio auditivo. Um algoritmo de codificação de áudio paramétrico que utiliza o MP, com dicionário baseado em wavelet packet, um algoritmo de reconfiguração dinâmica da árvore wavelet packet e busca de correspondência com critérios psicoacústicos é caracterizado em [9]. Um método para obter representações esparsas de sinais sonoros é proposto em [10]. No método é utilizado o algoritmo de MP com características psicoacústicas recentes sobre mascaramento de átomos tempo-frequenciais. A representação do sinal é realizada com um dicionário de átomos de Gabor de tamanhos variáveis. Sendo assim, o sinal é decomposto pela primeira vez usando MP e o modelo de mascaramento é aplicado no conjunto resultante de átomos.

No trabalho proposto por [11] apresenta um esquema de extração de características de áudio baseado na decomposição espectral, onde a decomposição é realizada de forma iterativa por busca de correspondência no domínio da freqüência. A classificação geral de áudio é efetuada através da decomposição completa do espectro de áudio em um conjunto de vetores de base harmônica e inarmônica. O trabalho conduz a decomposição no domínio da freqüência, onde a estrutura harmônica mostra um arranjo de forma mais clara [11]. Uma forma de algoritmo para análise de som e re-síntese com adaptação automática local de resolução de tempo-frequência é apresentada em [12]. A fórmula de reconstrução fornece uma boa aproximação do sinal original a partir de análises com diferentes resoluções de variação de tempo dentro de bandas de freqüência complementares. Também é fornecido um limite superior teórico para o erro de reconstrução do método.

O presente trabalho tem como objetivo realizar a decomposição de sinais de áudio, com o auxílio de 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 os autores utilizam o MP com dicionário de exponenciais complexas e ponderação psicoacústica. No lugar da função de ponderação, usamos, nesta dissertação, diretamente a curva psicoacústica obtida pelo modelo MPEG-1 (camada I) [13]. À medida em 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 descartadas nas iterações posteriores. Neste trabalho, a aferição dos resultados obtidos foi feita com uma ferramenta de avaliação perceptiva de qualidade de áudio, o PEAQ *(Perceptual Evaluation of Audio Quality)*.

Os arquivos utilizados no método estão distoníveis em:

$https://github.com/NogueiraJunior/Decomposicoes-Atomicas-em-Exponenciais-Complexas.git.$

A dissertação está organizada da seguinte forma: o capítulo 1 apresenta as características teóricas fundamentais de psicoacústica, incluindo uma descrição do modelo adotado juntamente com o método de obtenção do limiar psicoacústico de mascaramento global. O capítulo 2 explica a decomposição atômica psicoacústica, na qual são definidos o algoritmo de *Matching Pursuit*, o Dicionário de Exponenciais Complexas (DEC), e o funcionamento do algoritmo de decomposição desenvolvido neste trabalho. O capítulo 3 apresenta o sistema de quantização e alocação ótima de bits aplicado a sinais de áudio. Os métodos experimentais, assim como os resultados obtidos, estão demonstrados no capítulo 4. Por fim, no capítulo 5, encontram-se as conclusões desta dissertação e as indicações para desenvolvimento futuro desta pesquisa

1 NOÇÕES DE PSICOACÚSTICA

A psicoacústica, a ciência da percepção sonora, tem como objetivo principal estudar as relações entre as magnitudes dos estímulos físicos e as magnitudes das sensações por eles produzidos. Com significativa importância na caracterização da capacidade de análise tempo-frequência do ouvido interno, ela é de fundamental importância para a detecção de informações irrelevantes do sinal [14–17].

O bom entendimento do fenômeno de mascaramento auditivo fornece a base para o desenvolvimento de modelos psicoacústicos. Na modelagem psicoacústica, utilizam-se formas de mascaramento empiricamente determinadas para se analisar quais componentes de frequência mais contribuem para o limiar de mascaramento global e quando o componente de frequência pode ser descartado.

A psicoacústica apresenta características fundamentais na concepção de um codificador perceptivo de áudio, dentre as quais se podem destacar as unidades de medida de níveis de pressão sonora (SPL), os limiares da audição humana, fenômenos de mascaramento e a escala Bark [14, 17].

1.1 Fundamentos Teóricos

1.1.1 Nível de Pressão Sonora

O som atinge o ouvido humano sob a forma de uma onda mecânica longitudinal de pressão, a qual se propaga em meio material, mais comumente o ar. Denotando por $p(t)$ a pressão sonora em um certo ponto no instante t [15], podemos definir o nível de pressão sonora (*Sound Pressure Level* - SPL) como a grandeza que representa a intensidade de um determinado som, expressa em dB, como [14]:

$$\mathrm{SPL} = 10 \log_{10} \left(\frac{P}{P_0} \right)^2, \tag{1}$$

onde P é a pressão sonora no ponto em questão e, $P_0 = 20$ μPa é aproximadamente igual à pressão sonora no limiar de audição na frequência por volta de 2 kHz, isto é, P_0 representa o menor nível de pressão que pode ser percebido por um ouvinte humano típico.

A sensação de audição que se relaciona com SPL é a sonoridade (*Loudness*), expressa em *phon* [15]. O nível de sonoridade é definido como o nível de um tom sonoro

de 1 kHz, que é percebido tão alto quanto o som em análise para campos planos frontais incidentes [14]. A audição humana é capaz de responder a uma vasta gama de valores SPL em frequências que vão de 20 Hz até 20 kHz.

1.1.2 Limiar Absoluto de Audição

A área da audição é um plano no qual os sons audíveis podem ser exibidos [16]. O limiar absoluto da audição *(Absolute Threshold of Hearing - ATH)* humana, representado na Figura 2, caracteriza a intensidade sonora necessária em um tom puro que pode ser detectado por um ouvinte em um ambiente silencioso [17], isto é, representa o menor nível de pressão sonora em decibéis que se pode ouvir em uma dada frequência. Os componentes de frequência do sinal que estejam abaixo deste nível são irrelevantes para a percepção de sons e, portanto, em um codificador de áudio voltado a ouvintes humanos, não precisam ser armazenados ou transmitidos.

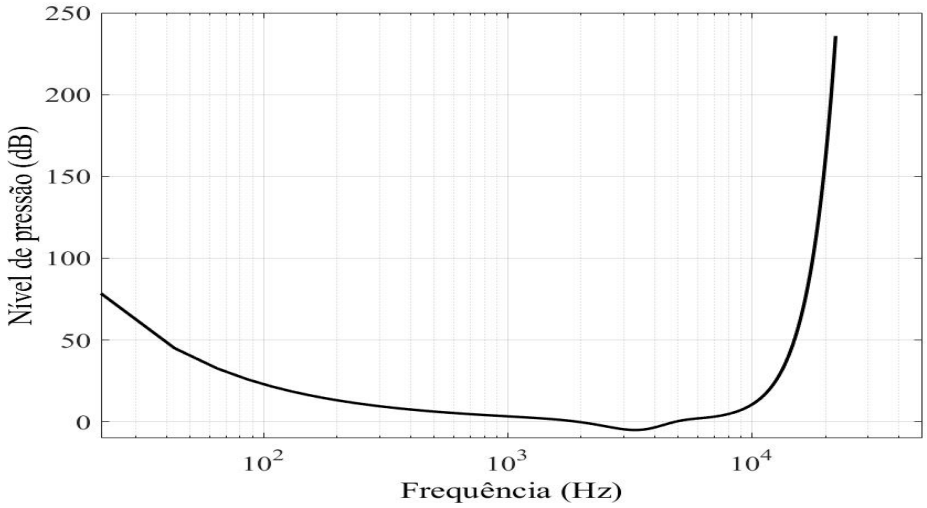


Figura 2 - Curva de limiar de audição humana.

O limiar do silêncio, em dB, dependente da frequência, pode ser numericamente aproximado como [14]:

$$A(f) = 3,64f^{-0,8} - 6,5e^{-0,6(f-3,3)^2} + 10^{-3}f^4, \tag{2}$$

onde f é a frequência em kHz.

1.1.2.1 Mascaramento

O fenômeno de mascaramento é de extrema importância para a codificação de sinais de áudio. Devido a este fenômeno, a percepção de um som está relacionada não apenas com a sua própria frequência e intensidade, mas também com as de seus componentes vizinhos [15]. No mascaramento, um componente de sinal mascarador altera o limiar auditivo e os estímulos físicos apenas produzem sensações auditivas se suas magnitudes físicas se enquadrarem acima desse novo limiar [16]. A Figura 3 apresenta um componente de sinal mascarador alterando o limiar auditivo e impedindo que componentes de frequências vizinhas com menor SPL sejam percebidos.

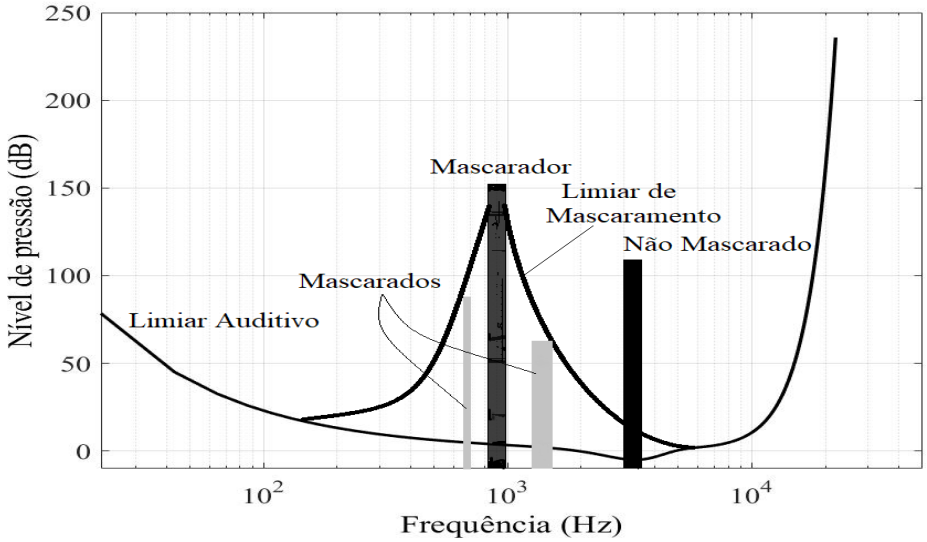


Figura 3 - Fenômeno de mascaramento na frequência. Figura adaptada de [14].

1.1.3 Escala Bark

No ouvido interno, a cóclea realiza uma conversão local-frequência e cada posição da membrana basilar corresponde a uma faixa limitada de frequências [15]. Como resultado da transformação local-frequência, a cóclea pode ser representada sob a ótica de processamento de sinais como um banco de filtros passa-banda altamente sobrepostos [17].

Existe uma faixa de frequências em torno da frequência do sinal mascarador na qual o limiar de mascaramento é plano. Essa faixa de mascaramento plano é conhecida com a banda crítica, e está intimamente relacionada com a escala Bark. A faixa auditiva

principal entre 20 Hz e 16 kHz é dividida em 24 bandas críticas que não se sobrepõem [15]. A equação que relaciona a escala Bark à frequência em Hertz é [14]:

$$z = 13 \arctan \frac{0,76f}{1000} + 3,5 \arctan \frac{f}{7500}, \tag{3}$$

onde f é a frequência em kHz.

O alcance do mascaramento efetivo para os componentes mascarados em diferentes frequências é determinado unicamente pela largura de banda crítica. Se o componente do sinal mascarado estiver na banda crítica do componente do sinal mascarador é mais provável que o sinal mascarado não seja percebido [15]. O mascaramento inter-bandas também ocorre, isto é, uma máscara centrada dentro de uma banda crítica tem efeitos perceptíveis nos limiares de detecção em outras bandas críticas [17].

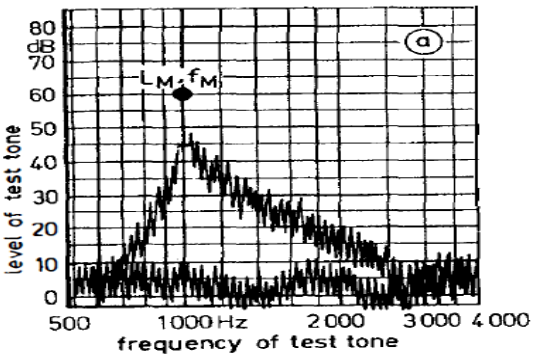


Figura 4 - Resultados experimentais de uma curva de mascaramento. Figura retirada de [16].

A Figura 4 ilustra o resultado de um teste de medida de uma curva de mascaramento. Nesse teste o sinal de mascaramento é um tom puro com 1 kHz e SPL de 60 dB. O nível inferior é o limiar em silêncio detectado pelo indivíduo testado medido na ausência do sinal de mascaramento. O nível superior é o limiar de audibilidade perante o sinal de mascaramento. Note que o mascaramento neste caso é maior nas frequências vizinhas à frequência do sinal mascarador decaindo, rapidamente à medida que o sinal de teste se afasta da frequência de mascaramento em ambas as direções. Outra observação vem do fato que o nível mais alto do limiar de audibilidade está aproximadamente 15 dB abaixo do sinal de mascaramento e, que o retorno é muito mais rápida no sentido das frequências baixas do que se movendo para altas frequências - essas características tendem a ser muito dependentes das especificidades do sinal de mascaramento e o sinal de teste [14].

Os aspectos arbitrários dos componentes do sinal de áudio possuem estruturas complexas de mascaramento simultâneos. É válido destacar alguns tipos:

- Ruído de banda estreita mascarando tom

  A máscara é um ruído com largura de banda menor ou igual ao da banda crítica, mascarando um tom dentro da mesma banda crítica [14].

  Os limiares de mascaramento de tons centralizados em 250 Hz, 1 kHz e 4 kHz e com larguras de banda de 100 Hz, 160 Hz e 700 Hz, mascarados por ruído de banda estreita estão mostrados na Figura 5 [16]. O nível do sinal mascarado é de 60 dB e a linha tracejada horizontal mostra o nível de ruído na figura. As linhas sólidas são os níveis da onda de tom puro para serem apenas audíveis. A curva tracejada na parte inferior representa o limite em silêncio [14].

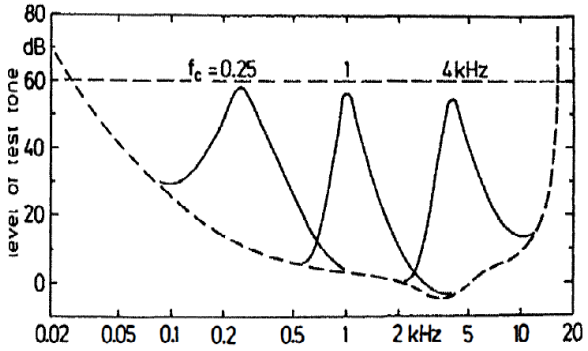


Figura 5 - Limiares de mascaramento de ruídos de banda estreita mascarando tons. Figura retirada de [16].

  As curvas do limiar de mascaramento apresentam características diferentes a depender da frequência do mascador. Os limiares de mascaramento tendem a ser mais amplos para mascaradores de frequências baixas, quando traçados em escala logarítmica [14].

- Tom mascarando ruído

  Um tom puro que ocorre no centro de uma banda crítica mascara um ruído de qualquer largura de banda ou forma, desde que o espectro de ruído esteja abaixo de um limiar previsível diretamente relacionado à intensidade e à frequência central do tom de mascaramento [17].

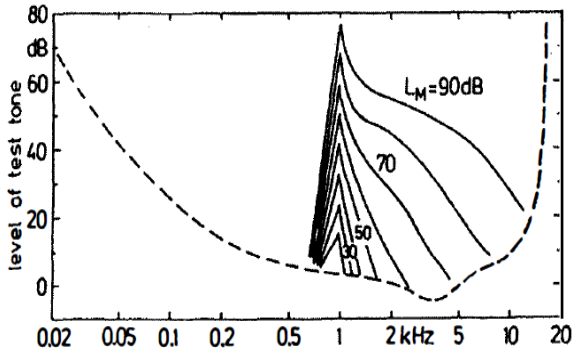


Figura 6 - Limiares de mascaramento de tons com 1 kHz e diferentes níveis mascarando tons. Figura retirada de [16].

No caso onde tons puros mascaram tons puros, isto é, onde um tom é considerado um ruído, o efeito de batimento acarreta maiores dificuldades do que os experimentos de mascaramento de ruído. A Figura 6 mostra os resultados para um mascarador de 1 kHz em diferentes níveis. Com o intuíto de evitar batimentos, a sonda foi ajustada 90 graus fora de fase com o mascarador na frequência de 1 kHz. Note que em baixos níveis de mascaramento, existe uma tendência de dispersão das curvas de mascaramento em direção a frequências mais baixas do que as frequências mais altas. Para altos níveis de mascaramento a situação se inverte, isto é existe uma maior expansão em direção a altas frequências do que frequências mais baixas [14].

- Ruído mascarando ruído

  Trata-se do caso em que um ruído de banda estreita mascara outro ruído de banda estreita. Devido às relações confusas de fase entre mascarador e mascarado, esse caso é mais difícil de se caracterizar [17].

A capacidade de mascaramento de um sinal mascarador é indicada pela razão sinal/máscara (SMR) mínima, isto é, a menor diferença de SPL entre o mascarador e seu limiar de mascaramento [15]. Quanto maior for a SMR, menor é o mascaramento. As SMRs, de mascaramento mínimo são maiores em experimentos de tons mascarando tons do que em experimentos de tons mascarando ruído. A esse fenômeno dá-se o nome de Assimetria de Mascaramento [14].

O modelo psicoacústico adotado no trabalho foi inspirado na camada I do padrão MPEG-1 *(Moving Pictures Experts Group)*, que representa o primeiro padrão internacional que especifica um formato digital para áudio de alta qualidade [13, 14, 16, 17].

1.2 MPEG-1

Projetados, inicialmente, como uma forma de codificação e representação de sinais de áudio com alta-qualidade para o armazenamento em mídias digitais, os algoritmos do MPEG-1 foram recomendados para aplicações de radiodifusão [14].

O padrão MPEG-1, conhecido como *[ISO/IEC 11172-3]*, descreve um algoritmo perceptivo de codificação de áudio projetado para sinais genéricos. O algoritmo considera o sinal de entrada estatisticamente quase-estacionário. O sinal de áudio é então representado por seus componentes espectrais em uma base quadro a quadro, explorando-se modelos perceptuais. O objetivo do algoritmo é proporcionar um esquema de codificação que reduza a taxa de dados enquanto mantém a qualidade de um CD [14].

O MPEG-1 Áudio especifica três camadas. As diferentes camadas oferecem diversos níveis de qualidade de áudio com complexidade variada [14]. O modelo psicoacústico do trabalho foi inspirado no padrão MPEG-1 da Camada I.

As formas da curva de mascaramento são muito mais simples de descrever quando mostradas na escala Bark. Modelar os fenômenos de mascaramento simultâneo é uma das atribuições principais do modelo psicoacústico. A representação da função de espalhamento utilizada para se criar padrões de excitação adotada pelo Modelo Psicoacústico ISO / IEC MPEG l é dada por:

$$10\log_{10}\left(F\left(dz, L_M\right)\right) = \begin{cases} -17\ dz + 0.15\ L_M\ (dz-1)\ \theta(dz-1) & \text{para} \quad dz \geq 0 \\ -(6+0.4L_M)|dz| - (11+0.4L_M)|dz|(|dz|-1)\theta(|dz|-1) & \text{para} \quad dz < 0 \end{cases} \tag{4}$$

onde $F\left(dz, L_M\right)$ é a função de espalhamento, $dz = z(f_{mascarado}) - z(f_{mascarador})$ é a diferença da escala de Bark entre a frequência do mascarado e do mascarador e L_M é o SPL do mascador [14]. A Figura 7 mostra a função de espalhamento do Modelo Psicoacústico do ISO/IEC MPEG 1, onde o eixo das abscissas representa o SPL e o eixo das ordenadas representa a escala Bark.

Podemos dividir o desenvolvimento do modelo psicoacústico em diversas etapas, resumidas abaixo.

1. Inicialmente, o sinal de áudio de entrada é analisado para definir seus componentes de frequências como tonais ou ruído, devido ao fenômeno de assimetria de mascaramento.

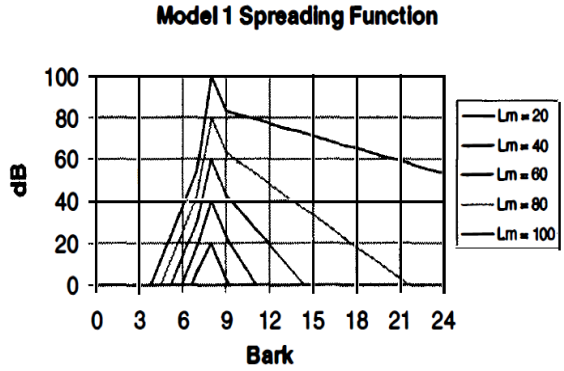


Figura 7 - Função de espalhamento linear adotada no modelo psicoacústico ISO / IEC MPEG l. Figura retirada de [16].

2. As funções de espalhamento são então utilizadas para simular os padrões de excitação dos dois tipos de mascaradores.

3. Em seguida, depois de deslocada para baixo de uma certa quantidade para cada componente mascarador, todos os limiares individuais de mascaramento e o limiar de silêncio são combinados para constituir o limiar global de mascaramento.

Deve-se assumir o nível inferior do limiar de mascaramento global em cada quadro com o intuito de se obter o limiar mínimo de mascaramento [15].

1.3 Algoritmo para Cálculo do Limiar de Mascaramento Psicoacústico Global

O cálculo realizado pelo algoritmo do limiar de mascaramento psicoacústico global utilizado no trabalho é consolidado em nove etapas, detalhadas a seguir.

1. Inicialmente o sinal temporal de entrada x, matematicamente escrito por $x = \{x\,[n]\,;n \in \mathbb{Z}\}$, do algoritmo é dividido em Q quadros com N pontos cada, isto é:

$$x\,[n] = \sum_{i=1}^{Q} x_i[n] \tag{5}$$

2. Cada quadro de sinal no domínio do tempo é multiplicado por uma janela de Hanning $w[n]$ para atenuar os efeitos espectrais causadas por transições abruptas, que é representada da seguinte forma:

$$x_i[n] = w[n]x[n - ip], \tag{6}$$

onde p é o comprimento do salto da janela e i é o quadro em análise.

3. Conversão do sinal do domínio do tempo para o domínio da frequência, através de uma transformada rápida de Fourier *(Fast Fourier Transform - FFT).*

   A conversão do sinal temporal janelado por quadro $x_i[n]$ para um sinal no domínio da frequência se dá pelo uso de uma FFT de M pontos, onde M é o tamanho máximo do dicionário que será definido mais adiante, assim:

$$S_i[f] = \text{FFT}\{x_i[n], M\}, \tag{7}$$

   onde $S_i[f]$ é o trecho do sinal no domínio da frequência. Utiliza-se o processo de inserir zeros no final do sinal para que se possa utilizar a FFT, isso é chamado $zero - padding$.

4. Obtenção das bandas críticas e da curva de limiar de silêncio.

   O espaçamento entre os elementos do dicionário Δf, isto é, o intervalo de frequência entre dois componentes sucessivos do dicionário para a representação de Fourier do sinal, é determinado pela taxa de amostragem do sinal F_s dividida por M, $\Delta f[Hz] = \frac{F_s}{M}$. A banda crítica $z(\Delta fn)$ é determinada com o auxílio da Equação (3), repetida aqui por conveniência [14]:

$$z = 13 \arctan \frac{0,76f}{1000} + 3{,}5 \arctan \frac{f}{7500} \tag{8}$$

   onde f é a frequência em kHz.

   A Figura 8 representa o limiar do silêncio, é calculado para M pontos através da Equação (2), repetida aqui por conveniência [14]:

$$A(f) = 3,64f^{-0,8} - 6,5e^{-0,6(f-3,3)^2} + 10^{-3}f^4, \tag{9}$$

   onde f é a frequência em kHz.

5. Cálculo da densidade espectral de potência;

   O nível do sinal por bloco é então calculado como uma densidade espectral de potência, $X_i[f]$, dada por:

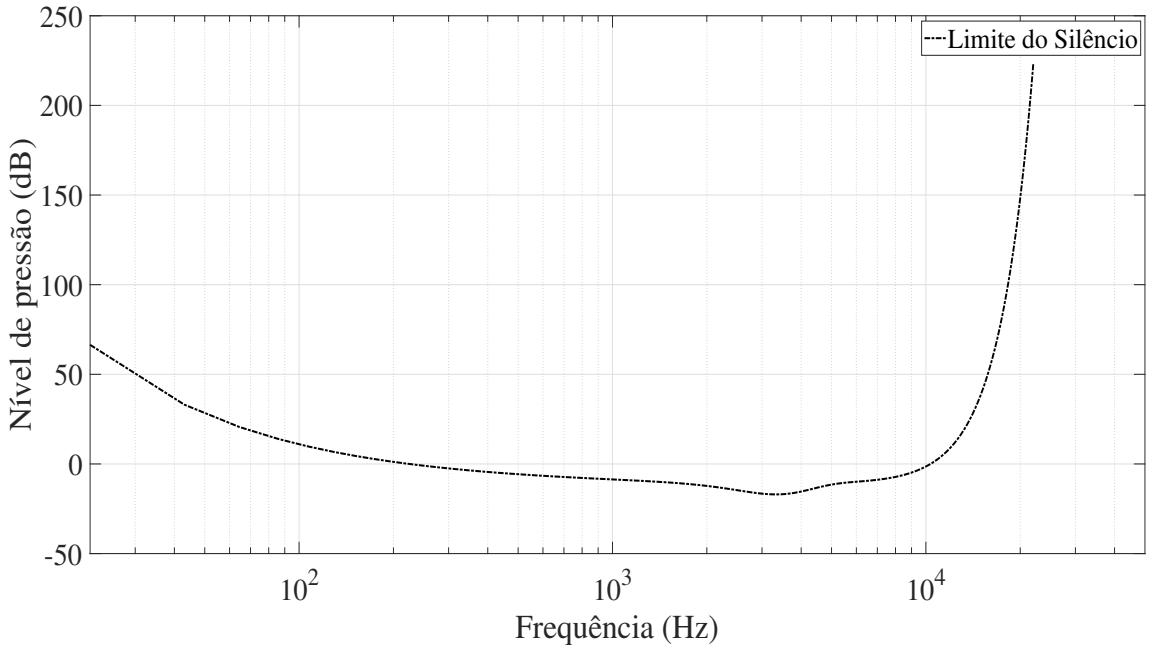


Figura 8 - Representação do limiar de silêncio.

$$X_i[f] = 20\log_{10}(\text{abs}(S_i[f])), \tag{10}$$

onde $S_i[f]$ denota a representação espectral do sinal para cada quadro. Na Figura 9 está ilustrada a densidade espectral de potência de um bloco de um sinal de áudio.

Uma normalização para o nível de referência de 96 dB SLP tem de ser feita, de tal forma que o valor máximo corresponda a exatamente 96 dB. Essa normalização se deve ao fato do desconhecimento prévio sobre os níveis de reprodução real. O nível de pressão em dB de um som só pode ser especificado por comparação com alguma referência dada [15]. O nível do sinal é normalizado de modo que o nível de uma onda senoidal de entrada, definida como $x[n] = \pm 1, 0$, tenha um nível de 96 dB SLP quando integrada pico-a-pico [17].

6. Determinação dos componentes tonais e não-tonais do sinal em análise.

   Existe a necessidade de se diferenciar os componentes de frequência como mascaradores tonais e não-tonais, devido ao fenômeno de assimetria de mascaramento [13]. Essa etapa se inicia com a determinação dos máximos locais, a extração dos componentes tonais e o cálculo da intensidade dos componentes não-tonais dentro de uma

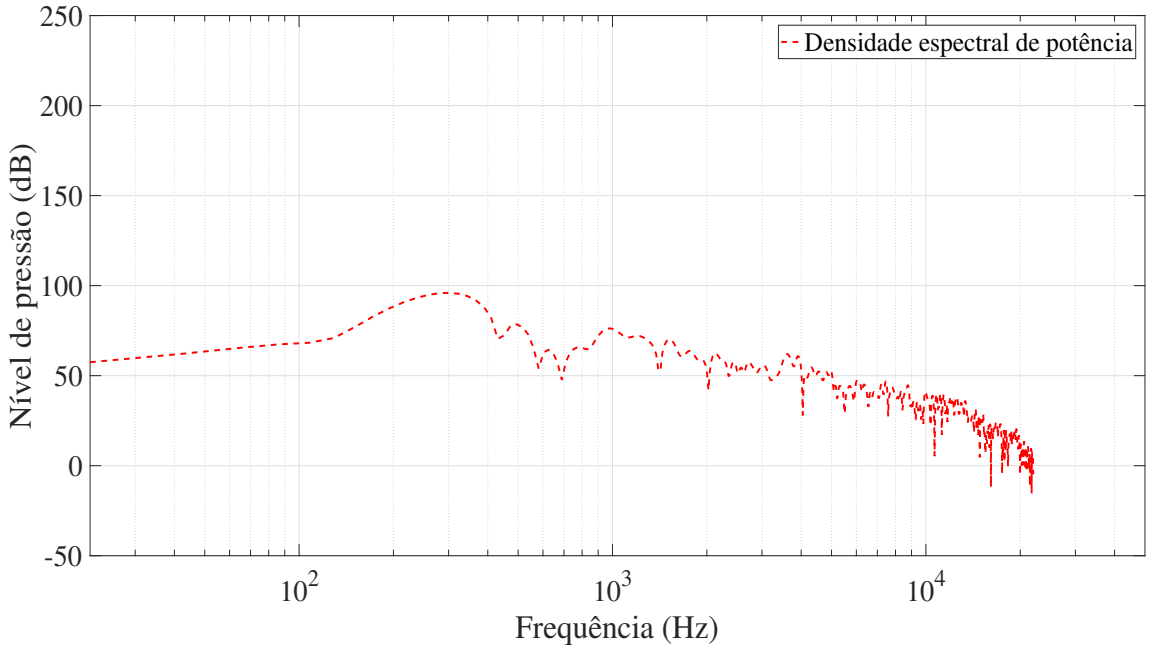


Figura 9 - Representação da densidade espectral de potência do sinal.

largura de banda crítica [13].

As densidades espectrais de potência das linhas espectrais por quadro $X_i[k]$, para cada linha espectral k, são classificadas como tonais e não-tonais através dos seguintes passos:

- Cálculo do Máximo Local. A linha espectral será um máximo local se:

$$\forall\ i\ |\ (X_i[k] > X_i[k-1])\ \&\ (X_i[k] \geq X_i[k+1])\,, \tag{11}$$

  onde k é a linha espectral em análise.

- Arrolar os componentes tonais e calcular o nível de pressão sonora.

  O máximo local na linha espectral k será listado como componente tonal se:

$$(X_{i,dB}[k] - X_{i,dB}[k+j]) \geq 7 \text{ dB}, \tag{12}$$

  onde j é escolhido por:

  $j = [-2, 2],$ para $2 < k < \left(63\frac{M}{N}\right)$

  $j = [-3, -2, 2, 3],$ para $\left(63\frac{M}{N}\right) \leq\ k\ < \left(127\frac{M}{N}\right)$

$j = [-6 : -2, 2 : 6]$, para $\left(127\frac{M}{N}\right) \leq k < \left(255\frac{M}{N}\right)$

$j = [-12 : -2, 2 : 12]$, para $\left(255\frac{M}{N}\right) \leq k < \left(510\frac{M}{N}\right)$,

onde M é a cardinalidade do dicionário e N o tamanho do quadro. Nas expressões acima, o termo $a : b$ indica a sequência de números inteiros entre a e b.

Se $X_i[k]$ é definido como um componente tonal, então os seguintes parâmetros são listados:

- O número do índice k da linha espectral
- Cálculo do nível de pressão sonora

$$X_{TM_i}[k] = 10\log_{10}\{10^{\frac{X_i[k-1]}{10}} + 10^{\frac{X_i[k]}{10}} + 10^{\frac{X_i[k+1]}{10}}\}, \tag{13}$$

onde $X_{TM_i}[k]$ é a densidade espectral de potência por quadro e está em dB.

- Arrolar os componentes não-tonais e calcular o nível de pressão sonora.

  Os componentes não-tonais são calculados a partir das linhas espectrais restantes dentro de cada banda crítica como sendo um único mascarador.

$$X_{NT_i}[k] = 10\sum_{j=1}^{NT}\log_{10}\frac{X[j]}{10}, \tag{14}$$

onde $X_{NT_i}[k]$ é a densidade espectral de potência por quadro e está em dB, NT é o número de componentes não-tonais, e j é o número da linha espectral mais próxima à média geométrica da banda crítica.

7. Determinação dos componentes mascaradores principais.

   A contribuição dos componentes de densidade espectral de potência tonais e não-tonais no limiar de mascaramento global só será considerada se esses estiverem de acordo com os seguintes requisitos:

   - Componentes tonais $X_{TM_i}[k]$ e não-tonais $X_{NT_i}[k]$ são considerados no cálculo do limiar de mascaramento se:

$$X_{TM_i}[k] \geq ATH[k] \tag{15}$$

$$X_{NT_i}[k] \geq ATH[k] \tag{16}$$

onde $ATH[k]$ é o limiar de silêncio.

- Para dois ou mais componentes tonais $X_{TM_i}[k]$ dentro de uma distância menor que 0,5 bark, só será considerado o componente da maior potência.

8. Cálculo do limiar de mascaramento para cada componente mascarador;

Os limiares de mascaramento individuais para os componentes tonais $LT_{TM}[Z(j), Z(k)]$ e não-tonais $LT_{NT}[Z(j), Z(k)]$ são calculados por:

$$LT_{TM}[Z(j), Z(k)] = X_{TM_i}[j] + AV_{TM}[Z(j)] + V_f[Z(j), Z(k)], \text{em dB} \tag{17}$$

$$LT_{NT}[Z(j), Z(k)] = X_{NT_i}[j] + AV_{NT}[Z(j)] + V_f[Z(j), Z(k)], \text{em dB} \tag{18}$$

onde LT_{TM} e LT_{NT} são os limiares de mascaramento individuais na taxa de banda crítica Z, k é o índice da linha espectral e j é o mascarador. Os termos $X_{TM_i}[j]$ e $X_{NT_i}[j]$ são os níveis de pressão sonora do componente de mascaramento j correspondente à banda crítica $Z(j)$, AV_{TM} e AV_{NT} são chamados de índices de mascaramento tonais e não-tonais, respectivamente, e $V_f[Z(j), Z(k)]$ é a função de espalhamento.

Os índices de mascaramento tonais e não tonais são dados respectivamente por:

$$AV_{TM}[Z(j)] = -1,525 - 0,275Z(j) - 4,5 \text{ dB}, \tag{19}$$

$$AV_{NT}[Z(j)] = -1,525 - 0,175Z(j) - 0,5 \text{ dB}, \tag{20}$$

onde $Z(j)$ é a taxa de banda crítica correspondente ao mascarador j.

A função de espalhamento $V_f[Z(j), Z(k)]$ de um mascarador é caracterizada por diferentes inclinações, dependentes da distância em Bark $dZ = Z(k) - Z(j)$ ao mascarador. Nessa expressão k é o índice da linha espectral em que a função de espalhamento é calculada e j é o mascarador. As funções de mascaramento são dadas por:

$$V_f = 17(dZ + 1) - (0,4X_i[j] + 6) \text{ dB}, \quad \text{para} -3 \leq dZ < -1 \,\text{Bark} \tag{21}$$

$$V_f = (0,4X_i[j] + 6) \text{ dB}, \quad \text{para} \quad -1 \leq dZ < 0 \,\text{Bark} \tag{22}$$

$$V_f = 17\ dZ \text{ dB}, \quad \text{para} \ \ 0 \leq dZ < 1 \,\text{Bark} \tag{23}$$

$$V_f = -(dZ - 1)\ (17 - 0,15X_i[j]) - 17 \text{ dB}, \quad \text{para} \ \ 1 \leq dZ < 8 \,\text{Bark}, \tag{24}$$

onde $X_i[j]$ é o nível de pressão sonora, tonal ou não-tonal, do j-ésimo componente de mascaramento em dB. A Figura 10 ilustra o limiar de mascaramento de um mascarador tonal devido a um componente com frequência em 646 Hz e SLP de 68,11 dB. Na Figura 11 está ilustrado o limiar de mascaramento não-tonal consequente de diferentes mascaradores não-tonais encontrados pelo algoritmo para o bloco em análise.

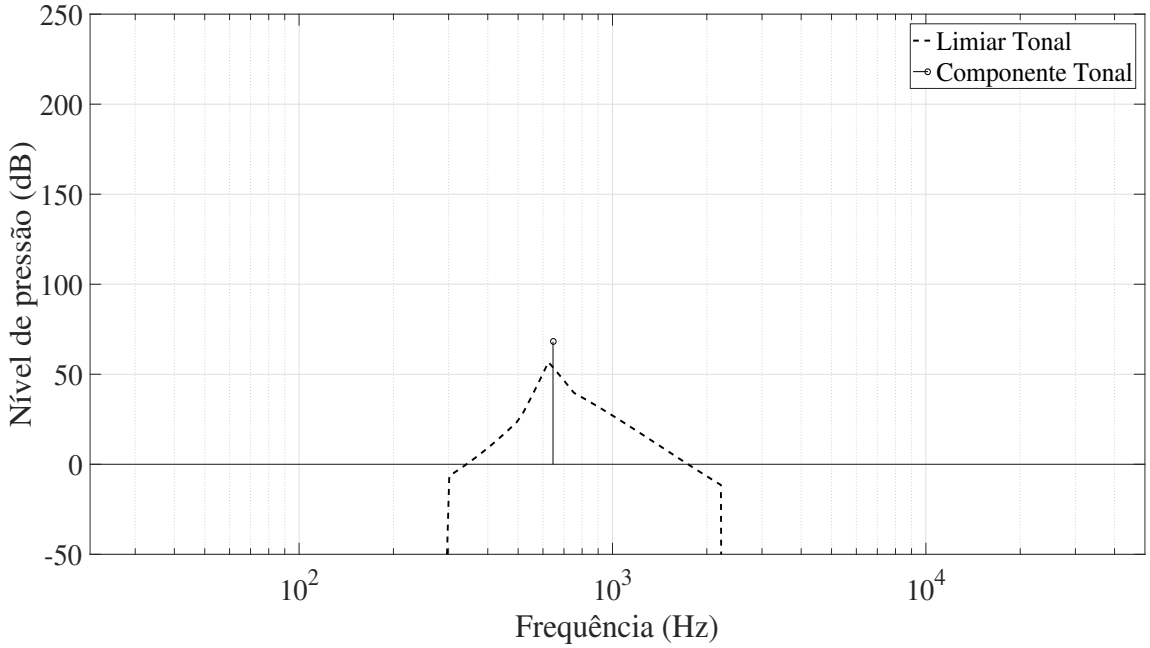


Figura 10 - Representação do limiar de mascaramento tonal.

9. Determinação do limiar de mascaramento global.

A Figura 12 ilustra o limiar global de mascaramento $L_{TG}[k]$, como uma combinação dos limiares individuais e do limiar de silêncio, dada por:

$$L_{TG}[k] = 10\log_{10}\left(10^{\frac{ATH[k]}{10}} + \sum_{j}^{N_{TM}} 10^{\frac{LT_{TM}[Z(j),Z(k)]}{10}} + \sum_{j}^{N_{NT}} 10^{\frac{LT_{NT}[Z(j),Z(k)]}{10}}\right) \tag{25}$$

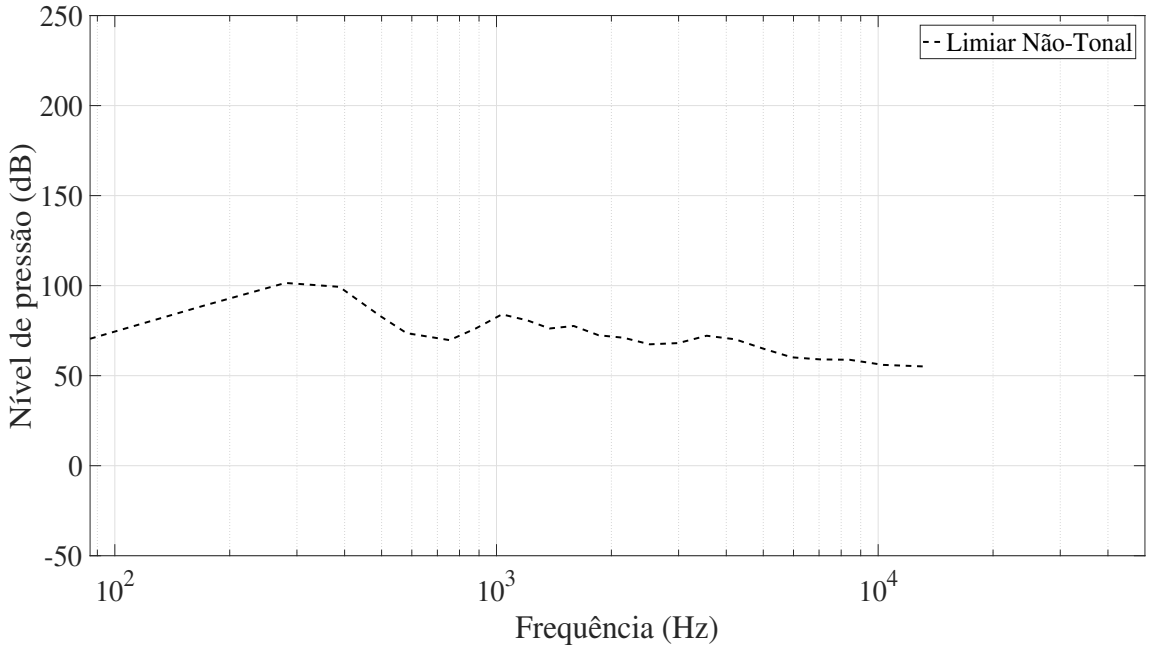


Figura 11 - Representação do limiar de mascaramento não-tonal.

onde $ATH[k]$ é o SPL do limiar de silêncio na linha espectral k, N_{TM} e N_{NT} são os números de mascaradores tonais e não-tonais listados, e LT_{TM} e LT_{NT} são seus limiares de mascaramento individuais correspondentes.

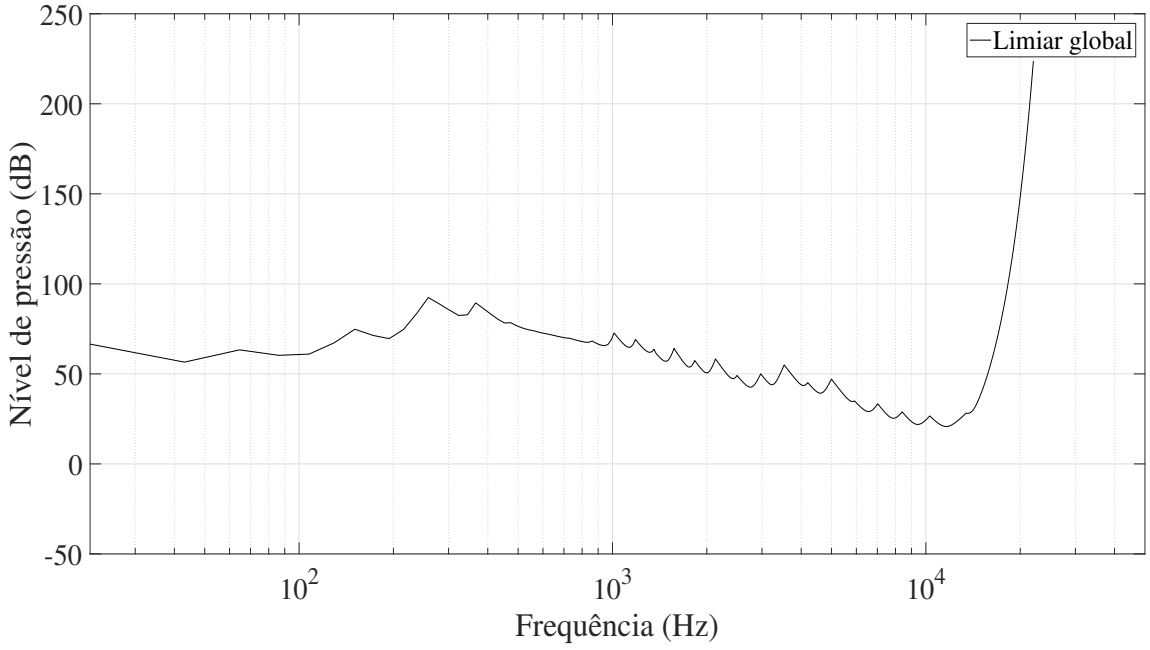


Figura 12 - Representação de um limiar de mascaramento global.

A Figura 15 retrata o resultado final do cálculo do limiar de mascaramento global realizado pelo algoritmo para um determinado bloco, com a densidade espectral de potência de um sinal arbitrário.

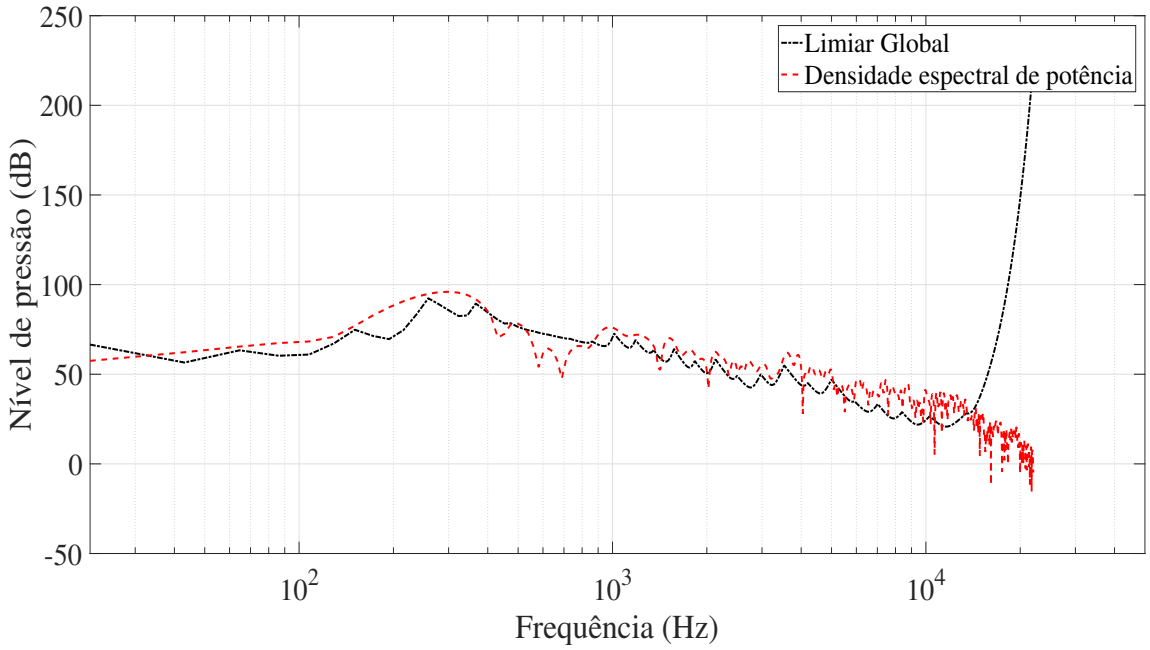


Figura 13 - Representação do limiar de mascaramento global.

1.4 Avaliação Perceptiva da Qualidade de Áudio

O processamento de áudio aproveita as propriedades da percepção auditiva humana para tornar a perda de codificação mais perceptivelmente indistinguível. Assim sendo, o modelo psicoacústico é o principal módulo que determina a qualidade de áudio e a eficiência de compressão de maneira significativa [18, 19].

Testes subjetivos de escuta humana são caros e demorados, pois exigem um grande número de ouvintes humanos treinados e equipamentos de forma adequada. Os algoritmos objetivos baseados em computador são usados para avaliar a qualidade dos sinais de áudio sem a necessidade de qualquer envolvimento humano. Os testes de escuta ainda são necessários para o desenvolvimento e treinamento do algoritmo de avaliação da qualidade objetiva e são freqüentemente usados para verificar a precisão do algoritmo [20].

Algoritmos objetivos de avaliação da qualidade, como PESQ e PEAQ, são geralmente considerados intrusivos, pois requerem um sinal de referência (o sinal original não

distorcido) e o sinal sob analise (o sinal distorcido) [19–21].

O PEAQ é a medida objetiva do Padrão de Recomendação da qualidade de áudio percebida estabelecida pela ITU em 1998, que também é chamada de BS.1387 [18]. Esse padrão pode ser usado em comparações entre dispositivos, podendo combinar outros algoritmos de avaliação de qualidade para fornecer uma avaliação geral eficaz do sistema, especialmente na indústria multimídia, como por exemplo o MPEG-1 Layer 2 e 3 [20].

A versão básica do PEAQ combina a estrutura fisiológica da ouvido humano com o efeito de mascaramento do sinal [18]. O PEAQ inclui modelos baseados na transformada rápida de Fourier (FFT), bem como em um banco de filtros [22]. A Figura 14 mostra um diagrama de blocos representando o esquema geral do modelo.

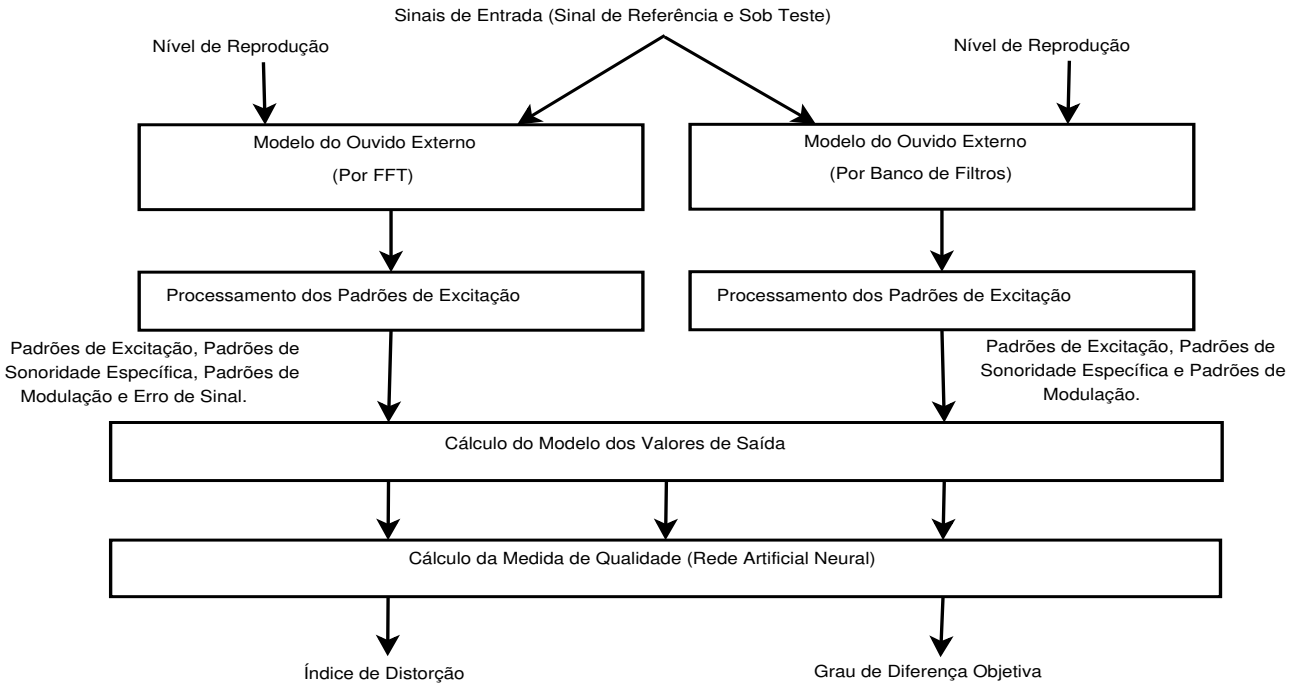


Figura 14 - Diagrama de blocos do esquema de medidas do PEAQ.

A pontuação da avaliação do PEAQ varia de 0 a -4, onde 0 representa um sinal com distorção imperceptível e -4 representam um sinal com distorção muito irritante. Deve-se notar que o PEAQ foi projetado apenas para nivelar sinais com deficiências extremamente pequenas [20].

Este capítulo apresentou de forma resumida os aspectos teóricos da psicoacústica, com o intuito descrever o cálculo do limiar global de mascaramento e a relação sinal/máscara, por quadro, que será utilizada do restante do trabalho.

2 DECOMPOSIÇÃO ATÔMICA PSICOACÚSTICA

O avanço da tecnologia aumentou a demanda pela utilização de sinais digitais em dispositivos móveis. Para que isso possa ocorrer em tais dispositivos, faz-se necessária a incorporação de operações que codifiquem e decodifiquem esses sinais, com o intuito de que eles sejam representados de forma compacta, tornando assim o seu armazenamento e a sua transmissão mais fáceis.

Algoritmos de compressão de sinais são utilizados para produzir representações compactas de sinais de alta qualidade, onde se quer expandir esses sinais por meio de funções que apresentem alto nível de similaridade com suas estruturas complexas [1, 23].

As decomposições atômicas têm 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 um sinal como uma combinação linear desses elementos [1, 23, 24]. Em vez de apenas representar sinais como superposições de senoides (a representação tradicional de Fourier), temos dicionários de coleções de formas de onda parametrizadas alternativos [25]. Tais decomposições atômicas possuem vantagens sobre as decomposições em base, funções ortogonais como as transformadas de Fourier e wavelet, pois apresentam uma maior flexibilidade quanto à representação dos sinais no plano tempo-frequência [1].

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 do dicionário de átomos de tempo-frequência aqueles que melhor se correlacionam com as estruturas presentes no sinal em análise.

2.1 Representações de Sinais

Em muitos casos, um sinal $x[n]$ no tempo discreto consiste de amostras de um sinal $x_a(t)$ no tempo contínuo, isto é, [26]

$$x[n] = x_a(n\Delta T), n \in \mathbb{Z}, \tag{26}$$

onde n é um número inteiro, $\mathbb{Z}$ é o conjunto dos números inteiros, e ΔT representa o período de amostragem do sinal.

O sinal $x[n]$, tratado como um vetor de comprimento N, pode ser representado por um conjunto de coeficientes $X\,[k]$ dado por [27, 28]:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X\,[k]\, e^{\,j\left(\frac{2\pi}{N}\right)nk} \tag{27}$$

$$X\,[k] = \sum_{n=0}^{N-1} x\,[n]\, e^{-j\left(\frac{2\pi}{N}\right)nk} \tag{28}$$

Esta é a Transformada de Fourier Discreta do sinal $x[n]$, a qual é periódica, com período 2π [26]. A Equação (28) pode ser reescrita na forma:

$$X\,[k] = \langle \mathbf{x}, \mathbf{e}_k \rangle, \tag{29}$$

onde $\mathbf{x}$ é o vetor coluna composto pelas amostras de $x[n]$, $n \in [0, ..., N-1]$, e $\mathbf{e}_k$ é um vetor coluna para um k fixo, calculado através da n-ésima componente do vetor k, definida por $e_{kn} = e^{j\left(\frac{2\pi}{N}\right)nk}$ [28].

2.2 Decomposições Atômicas

Em decomposições atômicas é permitido o uso de dicionários compostos por elementos linearmente dependentes, o que as distingue das metodologias clássicas baseadas em transformadas, pelas quais se utiliza uma base ortogonal que provê uma representação única do sinal. A dependência linear do dicionário está associada à existência de mais elementos no dicionário do que os necessários para se gerar um espaço vetorial onde o sinal está contido, e portanto, há mais de uma representação possível do sinal. Nesse caso, o dicionário é denominado redundante ou sobrecompleto [29].

Consideramos que o sinal $\mathbf{x}$ pode ser aproximado por átomos $\mathbf{g}_{m_k}$ que compõem uma família de vetores de um dicionário D, pertencentes ao espaço de Hilbert $\mathbb{H}$, de modo que [30]:

$$\mathbf{x} \approx \sum_{k=0}^{K-1} \alpha_k \mathbf{g}_{m_k}. \tag{30}$$

Os átomos $\mathbf{g}_{m_k}$ possuem $\|\mathbf{g}_{m_k}\| = 1$ e são indexados por m_k, que é definido como $m_k : Z+ \rightarrow \{1, ..., \#D\}$; $\#D$ é o número de elementos do dicionário D, portanto $m_k \in \{1, ..., \#D\}$. O parâmetro α_i é o coeficiente que pondera $\mathbf{g}_{m_k}$ e K corresponde ao número

de átomos selecionados para representar $\mathbf{x}$.

A utilização de dicionários altamente redundantes possibilita a extração direta de uma variedade maior de padrões e fenômenos presentes em sinais, resultando em representações mais compactas e eficientes [1]. Assim, a decomposição referente a Equação (30) não é única, porque alguns elementos no dicionário têm representações em termos de outros elementos [25]. Essa não unicidade nos dá a possibilidade de adaptação, ou seja, de escolher entre muitas representações uma que seja mais adequada aos nossos propósitos [31]. Entretanto, a esparsidade da representação de sinais não depende somente do nível de redundância do dicionário, mas também da acurácia com que o modelo matemático utilizado represente os fenômenos intrinsecos do sinal de forma coerente. Os átomos do dicionário devem apresentar um alto-grau de similaridade com os padrões existentes na classe. Em processos físicos, considera-se o sinal observado como uma mistura de componentes $\mathbf{p}_i$ que representam os fenômenos físicos, dada por

$$\mathbf{x} = \sum_i \beta_i \mathbf{p}_i + \mathbf{u}, \tag{31}$$

onde $\mathbf{u}$ é o ruído inerente à observação. Quanto mais parecidos forem os átomos $\mathbf{g}_{m_k}$ e os seus respectivos coeficientes α_k com as componentes $\mathbf{p}_i$ e os seus coeficientes β_i, utilizados na representação de um sinal $\mathbf{x}$, melhor será a representação obtida para fins de modelagem do sinal e reconhecimento de padrões intrínsecos. O objetivo da decomposição é atingir simultaneamente os seguintes objetivos [25]:

- Esparsidade. Deve-se obter a representação mais esparsa possível do objeto - aquele com o menor número de coeficientes signicativos.

- Super resolução. Deve-se obter uma resolução de objetos esparsos com resolução muito maior do que a possível com abordagens tradicionais não adaptativas.

- Velociade. Deve ser possível obter uma representação em ordem de tempo $O(n)$ ou $O(n \log(n))$

Existem ferramentas capazes de efetuar a decomposição atômica de sinais. Neste trabalho utilizamos o algoritmo Matching Pursuit, o qual consiste em um algoritmo subótimo que nos permite evitar a complexidade proibitiva de se encontrar os membros do dicionário de exponenciais complexas que melhor representam o sinal $\mathbf{x}$ [24].

2.3 Matching Pursuit

Uma música é constituída por notas que possuem durações diferentes e que ocorrem em momentos distintos. Este padrão não-estacionário não é bem representado em uma base senoidal, como a de Fourier. Por outro lado, as notas musicais possuem um comportamento harmônico, sendo assim uma base do tipo wavelet também não é adequada para representá-las [23].

Para obtermos uma decomposição adaptativa do sinal, devemos expandi-lo em uma soma de formas de onda cujas localizações no tempo e na frequência correspondam às verificadas no sinal [24]. O Matching Pursuit é um algoritmo que decompõe um sinal e o representa como uma expansão linear de formas de onda ou funções [1]. A cada etapa, o algoritmo procura em seu dicionário uma função que combina melhor com o sinal atual e a extrai deste uma versão escalada daquela do sinal corrente produzindo o resíduo. O Matching Pursuit utiliza um dicionário redundante, sendo assim não existe uma única configuração para a expansão do sinal. À medida que o número de iterações de uma busca correspondente aumenta, o erro de aproximação converge para a realização de um processo de ruído cuja energia é uniformemente distribuída por todos os vetores do dicionário [1]. O Matching Pursuit continua a ser aplicado nesse sinal residual até que seu critério de parada seja atendido.

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 um grau de liberdade na construção da expansão de funções.

Em nossa aplicação, como veremos em detalhes mais adiante, cada átomo $\mathbf{g_m}$ é caracterizado por parâmetros de amplitude A, frequência f e fase θ, através de aproximações sucessivas de $\mathbf{x}$ por meio de projeções ortogonais envolvendo os elementos do dicionário.

O Matching Pursuit começa projetando um vetor sinal $\mathbf{x}$ em um vetor $\mathbf{g}_{\gamma_0} \in \mathbf{D}$ e calculando o resíduo $\mathbf{r_x}$ [23]:

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{g}_{\gamma_0} \rangle \mathbf{g}_{\gamma_0} + \mathbf{r_x} \tag{32}$$

onde $<,>$ é o operador que representa o produto interno.

Desde que $\mathbf{r_x}$ seja ortogonal a $\mathbf{g}_{\gamma_0}$ e as funções do dicionário tenham energia uni-

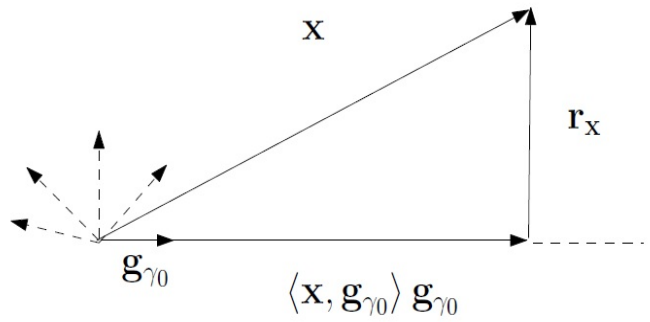


Figura 15 - Representação gráfica da projeção ortogonal no primeiro passo da decomposição de x.

tária, pode-se concluir que:

$$\|\mathbf{x}\|^2 = |\langle \mathbf{x}, \mathbf{g}_{\gamma_0} \rangle|^2 + \|\mathbf{r}_\mathbf{x}\|^2 \tag{33}$$

Logo, para minimizar $\|\mathbf{r}_\mathbf{x}\|$, temos que escolher $\mathbf{g}_{\gamma_0} \in D$, tal que $\langle \mathbf{x}, \mathbf{g}_{\gamma_0} \rangle$ seja máximo. Em alguns casos, é computacionalmente mais eficiente encontrar um vetor $g_{\gamma 0}$ que é quase ótimo:

$$|\langle \mathbf{x}, \mathbf{g}_{\gamma_0} \rangle| \geq \alpha \sup_{\gamma \in \mathbf{M}} |\langle \mathbf{x}, \mathbf{g}_\gamma \rangle|, \tag{34}$$

onde $\alpha \in (0, 1]$ é um fator numérico responsável por maximizar a relação $|\langle \mathbf{x}, \mathbf{g}_\gamma \rangle|$.

A escolha de um vetor $g_{\gamma 0}$ que satisfaça a Equação (34) não é aleatória. É definida por uma função de escolha C, que associa a qualquer subconjunto Λ de $\mathbf{M}$ um índice que pertence a Λ. O conjunto de índices vetoriais dos átomos que satisfazem a Equação 34 é definido por:

$$\Lambda_0 = \{\beta \ \in \mathbf{M} : |\langle \mathbf{x}, \mathbf{g}_\beta \rangle \geq \alpha \sup_{\gamma \in \mathbf{M}} |\langle \mathbf{x}, \mathbf{g}_\gamma \rangle\} \tag{35}$$

A escolha de um vetor $\mathbf{g}_{\gamma 0}$ que satisfaz a Equação 34 é equivalente à escolha do índice γ_0 dentro de Λ_0, formalmente definido por $\gamma_0 = C(\Lambda_0)$. O axioma da escolha garante que exista pelo menos uma função de escolha.

Seja o sinal inicial $\mathbf{x}$, assim o resíduo inicial pode ser definido como $\mathbf{r}_x^0 = \mathbf{x}$. Suponha que o resíduo da k-ésima ordem $\mathbf{r}_\mathbf{x}^k$ já está calculada para $k \geq 0$. A próxima escolha de m_k é tal que

$$\alpha_k = \arg \max_{m \in \mathbf{M}} | < \mathbf{r}_\mathbf{x}^k, \mathbf{g}_{m_k} > | \tag{36}$$

e projetar $\mathbf{r_x^k}$ em $\mathbf{g_{m_k}}$ e subtrair de $\mathbf{r_x^k}$:

$$\mathbf{r_x^k} = \langle \mathbf{r_x^k}, \mathbf{g_{m_k}} \rangle \mathbf{g_{m_k}} + \mathbf{r_x^{k+1}}, \tag{37}$$

onde $\mathbf{r}_x^{k+1}$ é o resíduo da "$k+1$"-ésima iteração, $\mathbf{r}_x^k$ é o resíduo do sinal sendo na k-ésima iteração decomposto e o operador $<,>$ representa o produto interno [1].

A ortogonalidade entre $\mathbf{r_x^{k+1}}$ e $\mathbf{g_{m_k}}$ implica que

$$||\mathbf{r_x^k}||^2 = |\langle \mathbf{r_x^k}, \mathbf{g_{m_k}} \rangle|^2 + ||\mathbf{r_x^{k+1}}||^2 \tag{38}$$

Ao fim de M iterações, tem-se

$$\mathbf{x} = \sum_{k=0}^{M-1} \langle \mathbf{r_x^k}, \mathbf{g_{m_k}} \rangle \mathbf{g_{m_k}} + \mathbf{r_x^M}, \tag{39}$$

Similarmente, tem-se que

$$\|\mathbf{x}\|^2 = \sum_{k=0}^{M-1} |\langle \mathbf{r_x^k}, \mathbf{g_{m_k}} \rangle|^2 + ||\mathbf{r_x^k}||^2 \tag{40}$$

Assim, o sinal original $\mathbf{x}$ é decomposto em uma soma ponderada de elementos de dicionário escolhidos que melhor correspondem aos resíduos obtidos de forma iterativa. Embora essa decomposição não seja linear, o MP é caracterizado pela conservação de energia, característica intrínseca de decomposições lineares e ortogonais.

2.4 Dicionário de Exponenciais Complexas

Definindo o dicionário como um conjunto de M funções redundantes, limitadas no tempo, também conhecidas como átomos, pertencentes a um espaço de Hilbert $\mathbb{H}$, ou seja, $D = (\mathbf{g}_m)_{m \in \mathbf{M}}$, tal que $\|\mathbf{g}_m\| = 1$,em que m é o índice que define o átomo e $\mathbf{M}$ é o conjunto de todos os $m's$ possíveis [32].

Um dicionário de alta-correlação com as formas dos sinais analisados permitirá um baixo erro de aproximação. Neste trabalho, usa-se um dicionário de exponenciais complexas (DEC). Esse dicionário é utilizado por permitir representar a fase do átomo.

Os seus elementos são definidos da seguinte maneira [2]:

$$\mathbf{g}_m = \left\{ g_m[n] = \frac{1}{N} e^{j2\pi \frac{m}{M} n} \right\} \tag{41}$$

onde $n = 0, 1, ..., N-1$ e $m = 0, 1, ..., M-1$. Sendo o sinal real, os coeficientes de correlação aparecem em pares conjugados, logo precisamos pesquisar somente metade dos coeficientes de correlação [2].

Com blocos de 512 amostras, ou seja, de duração de aproximadamente 11,6 ms (assumindo uma frequência de amostragem igual a 44100 Hz), em geral o sinal se comporta aproximadamente de forma estacionária [33], o que torna adequada a utilização da amplitude constante do dicionário.

2.5 Algoritmo de Decomposição

O processamento do sinal é realizado bloco-a-bloco, sendo que cada bloco corresponde a um trecho janelado do sinal, podendo haver sobreposição entre blocos adjacentes, conforme ilustrado na Figura 16. A utilização da janela de Hanning, com sobreposição de metade do tamanho do bloco, tem a finalidade de evitar o surgimento de artefatos audíveis oriundos da separação do sinal em blocos e do próprio janelamento. Neste caso, o sinal é reconstruído através de um procedimento de sobreposição e adição entre os blocos. Seja o i-ésimo bloco do sinal $x[n]$ é descrito como

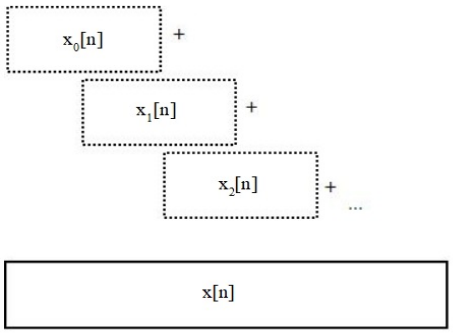


Figura 16 - Esquema da representação da divisão do sinal em blocos.

$$x_l[n] = \omega[n] x[n - lp], \tag{42}$$

onde $i = 0, 1, ..., Q$, Q é número de blocos, l é comprimento a largura da janela, e $w[n]$ possui comprimento N. É importante que $\sum_{i=0}^{Q-1} \omega[n-il] = 1$, de modo que a reconstrução seja perfeita.

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 [26]. Utiliza-se o processo de inserir zeros no final do sinal para que se possa utilizar a FFT, isso é chamado $zero - padding$. Inicialmente, introduz-se uma generalização do produto interno para um produto interno ponderado, $< \mathbf{x}, \mathbf{y} >_{\mathbf{W}} = \mathbf{yWx}$, onde $\mathbf{W}$ é uma matriz positiva definida simétrica, de formato diagonal, em que os elementos são as amostras da janela $w[n]$. Na k-ésima iteração do MP, calcula-se, para $m = 0, 1, ..., M-1$, [2]

$$\frac{| < \mathbf{g}_m, \mathbf{r}_k >_W |}{< \mathbf{g}_m, \mathbf{g}_m >_W} = \frac{R_k^w \left[\frac{m}{M}\right]}{W \left[\frac{0}{M}\right]} \tag{43}$$

onde

$$R_k^w \left[\frac{m}{M}\right] = \sum_{n=0}^{M-1} w[n] r_k[n] e^{-j2\pi \frac{m}{M} n}, \tag{44}$$

e

$$W \left[\frac{0}{M}\right] = \sum_{n=0}^{M-1} w[n]. \tag{45}$$

Em seguida, busca-se o máximo produto interno normalizado

$$m_k = \arg\max_m \frac{| < \mathbf{g}_m, \mathbf{r}_k >_W |}{< \mathbf{g}_m, \mathbf{g}_m >_W} \tag{46}$$

Portanto, o k-ésimo coeficiente é dado por:

$$\alpha_k = \frac{R_k^w \left[\frac{m_k}{M}\right]}{W \left[\frac{0}{M}\right]} = \frac{A_k e^{j\theta_k}}{W \left[\frac{0}{M}\right]}, \tag{47}$$

onde $R_k^w[\frac{m_k}{M}]$ é uma função exponencial complexa que pode ser escrita na forma polar: $A_k e^{j\theta_k}$, onde A_k e θ_k são a magnitude a fase de $R_k^w[\frac{m_k}{M}]$, respectivamente.

Dado que sinais de áudio são reais, $x_l[n]$ também será real. Neste caso é preciso calcular somente metade das correlações a cada iteração de $R_k^w \left[\frac{m}{M}\right]$, ou seja, variando $m = 0, 1, ..., \frac{M}{2}$.

Assim, o algoritmo realiza os seguintes procedimentos [2]:

1. Calcula-se a cada bloco o limiar de mascaramento global;

2. Armazena a FFT (de comprimento M) da janela, $W\left[\frac{m}{M}\right]$;

$$W\left[\frac{m}{M}\right] = \sum_{n=0}^{M-1} w[n]e^{-j2\pi\frac{m}{M}n}, \tag{48}$$

3. Define-se $\mathbf{r}_0 = \mathbf{x}_l$, onde $l = 0, 1, ..., Q$ e Q é número de blocos. Calcula-se a FFT (de comprimento M) de $r_o[n]w[n]$, ou seja, $R_0^w\left[\frac{m}{M}\right]$, onde $w[n]$ é a função janela utilizada.

   No processo do MP, realiza-se o seguinte procedimento a cada iteração k:

   (a) Dado $R_k^w\left[\frac{m}{M}\right]$, é encontrado o valor máximo absoluto dos seus elementos , obtendo assim os parâmetros senoidais de amplitude A_k, frequência f_k e fase θ_k, conforme está ilustrado na Figura 17 até a Figura 20;

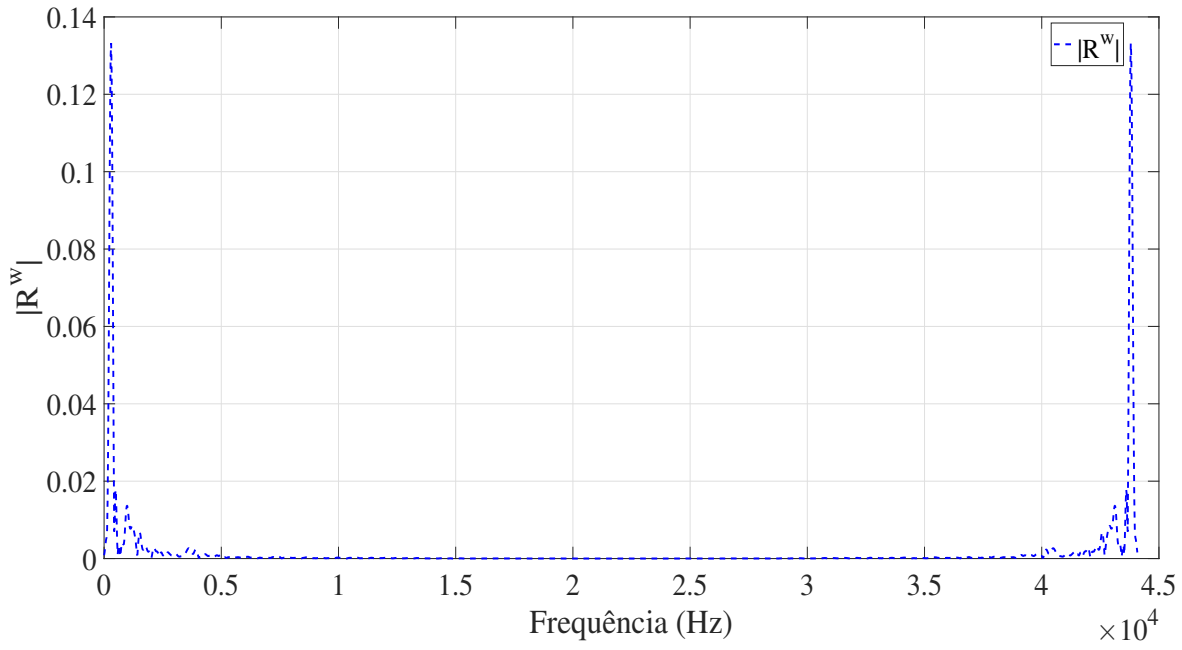


Figura 17 - Densidade espectral do elemento do sinal com máxima correlação entre os elementos do dicionário e o resíduo do bloco do sinal de áudio.

   (b) É calculado o resíduo no domínio da frequência da próxima iteração

$$R_{k+1}^w\left[\frac{m}{M}\right] = R_k^w\left[\frac{m}{M}\right] - \frac{A_k\left(e^{i\theta_k}W\left[\frac{m-m_k}{M}\right]\right)}{2} - \frac{A_k\left(e^{-i\theta_k}W\left[\frac{m+m_k}{M}\right]\right)}{2} \tag{49}$$

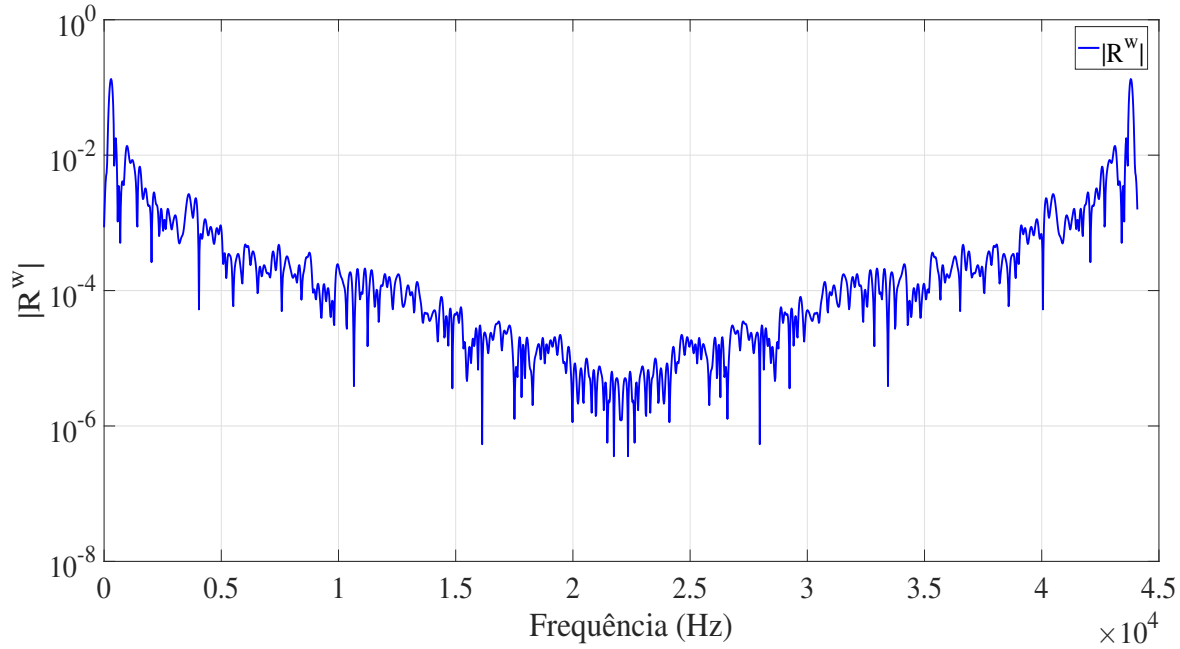


Figura 18 - Densidade espectral em escala logarítmica do elemento do sinal com máxima correlação entre os elementos do dicionário e o resíduo do bloco do sinal de áudio.

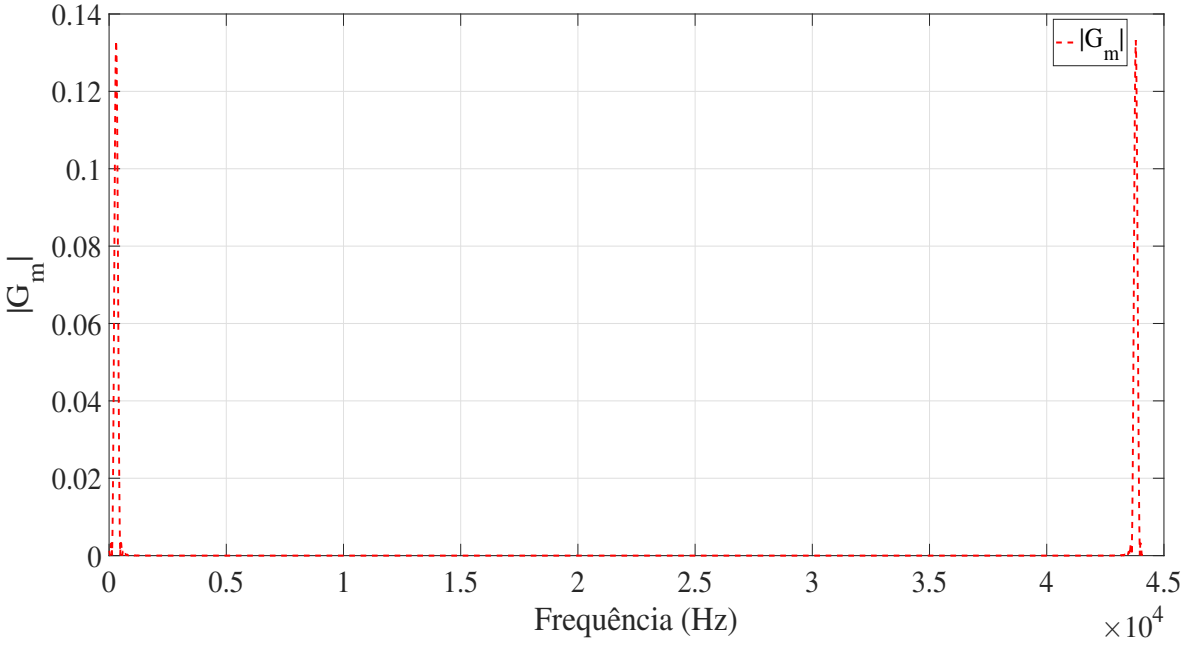


Figura 19 - Densidade espectral do elemento do dicionário com máxima correlação entre o resíduo do bloco do sinal de áudio.

onde $R_k^w\left[\frac{m}{M}\right]$ é o resíduo obtido na k-ésima iteração, $R_{k+1}^w\left[\frac{m}{M}\right]$ é o resíduo obtido na iteração seguinte, A_k e θ_k são a magnitude e fase de $R_k^w[\frac{m_k}{M}]$ e $W\left[\frac{m-m_k}{M}\right]$ e $W\left[\frac{m+m_k}{M}\right]$ são as contribuições dos deslocamentos em frequência dos pontos

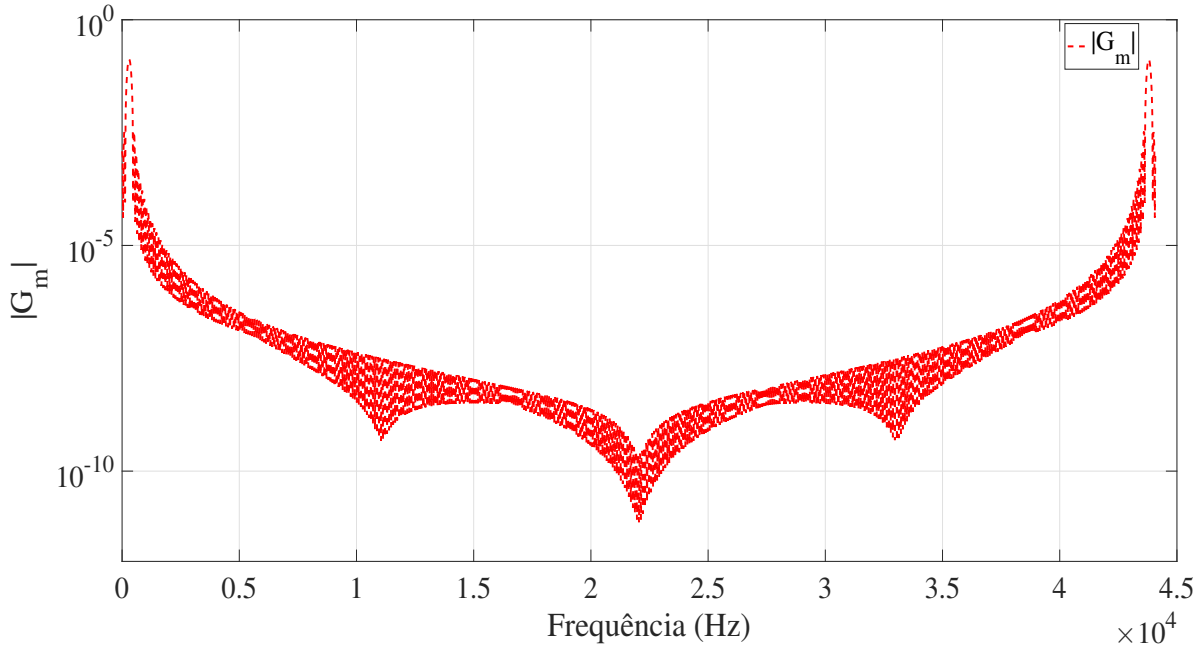


Figura 20 - Densidade espectral em escala logarítmica do elemento do dicionário com máxima correlação entre o resíduo do bloco do sinal de áudio.

selecionados.

Na Figura 21 ilustra a remoção átomo com máxima correlação encontrado no sinal.

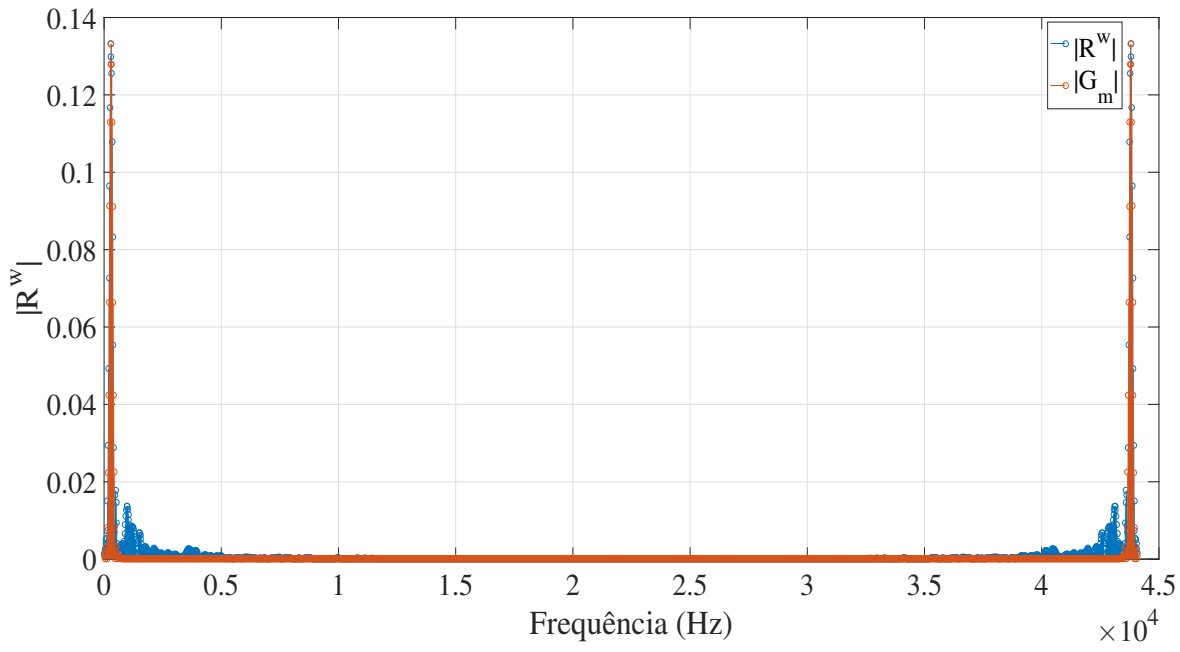
(c) Se $R_{k+1}^{w}\left[\frac{m}{M}\right]$ em dB_{SPL} estiver abaixo do limiar global de mascaramento para todas as frequências, interrompe-se o processo iterativo;

(d) Removem-se do dicionário os elementos correspondentes às frequências com níveis de pressão sonora (SPL) abaixo do limiar global de mascaramento.

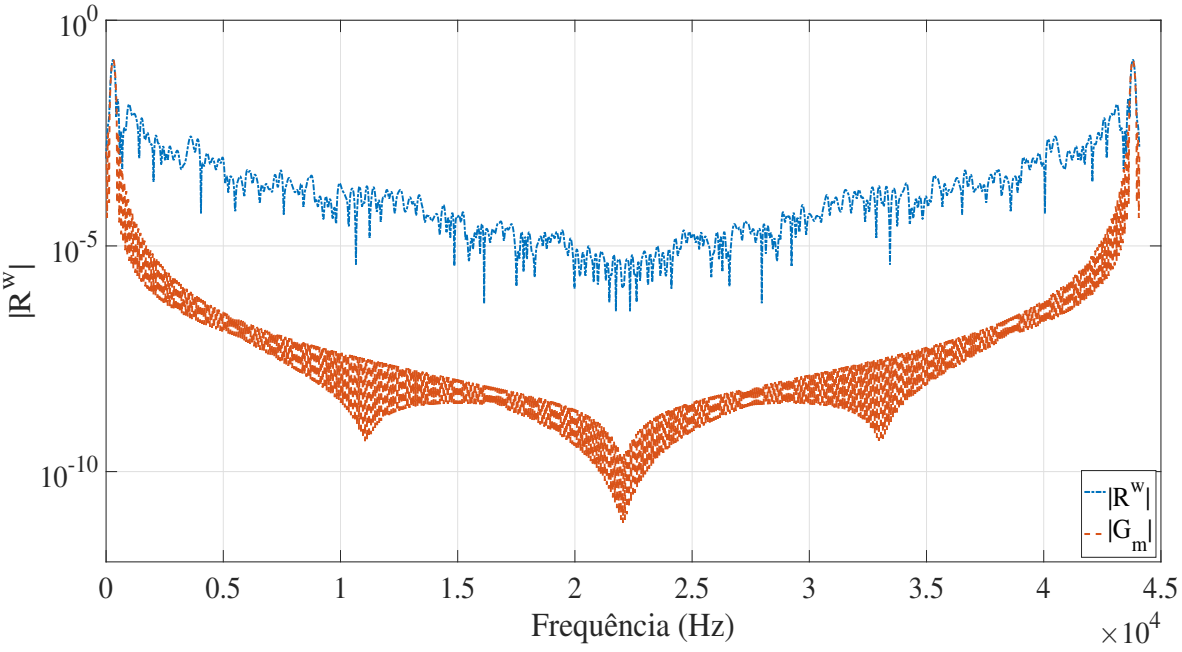
Ao final do processo iterativo, obtém-se uma representação do bloco

$$x_l[n] = \sum_{k=0}^{N_{iter}-1} 2a_k \cos\left[2\pi f_k n + \theta_k\right], \tag{50}$$

onde N_{iter} 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, $O(2M \log_2 M + 2M)$ em função das subtrações da Equação (49).



(a)



(b)

Figura 21 - Remoção dos máximos encontrados no sinal de entrada: (a) em escala linear e (b) em escala logarítmica.

3 ALOCAÇÃO ÓTIMA DE BITS

O sinal sonoro pode ser representado como uma função de onda no tempo contínuo. Para a representação desse sinal sonoro no formato digital, é imprescindível a amostragem do sinal e a conversão de seus valores amostrados em uma representação discreta. Isso se faz necessário, pois os computadores armazenam números utilizando uma quantidade finita de *bits*, de modo que os valores de amplitude são armazenados com precisão limitada [14].

A compressão ou codificação de áudio é utilizada para se obter representações digitais compactas de sinais de áudio com o intuito de se alcançar uma maior eficiência na transmissão ou armazenamento. O objetivo principal neste caso é representar o sinal com o número mínimo de *bits*. No caso de este ser um requisito, deve-se garantir que o codificador seja transparente, ou seja, que o sinal reproduzido não seja perceptivamente distinguível do sinal original. O processo de representar um número infinito de valores com um conjunto finito de símbolos é chamado de quantização [34].

O sinal pode ser compactado utilizando técnicas sem perdas, nas quais os dados codificados são uma cópia exata do sinal original, tornando assim o desempenho de compactação limitado, especialmente se existirem restrições de largura de banda ou de armazenamento, e não for primordial uma reprodução perfeita da fonte. Em tais cenários, a compactação com perdas é necessária. É possível alcançar altas taxas de compressão ao custo da representação imperfeita da fonte. A troca entre a fidelidade da fonte e a taxa de codificação é exatamente o compromisso taxa-distorção. É possível compensar o número de *bits* na representação (a taxa) com a fidelidade da representação (a distorção) [35].

O projeto do quantizador tem um grande impacto no nível de compressão obtido e na perda incorrida em um esquema de compressão com perdas [34]. O ruído de quantização é a principal causa de distorção no processo de codificação de sinais de áudio [14].

3.1 Sistemas de Compressão

Em geral, um método de compressão de sinais pode ser dividido em três partes [34], como ilustrado na Figura 22:

1. **Transformação**: obtém-se uma representação compacta do sinal, resultando em um menor número de coeficientes;

2. **Quantização**: mapeiam-se os coeficientes da transformação em um conjunto finito de símbolos;

3. **Codificação**: mapeiam-se os símbolos em bits.

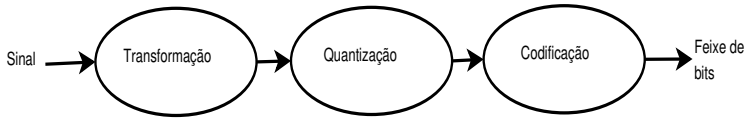


Figura 22 - Esquema geral de compressão de sinais.

O sistema de compressão de áudio proposto neste trabalho é apresentado na Figura 23. Na etapa de transformação é utilizado o método de decomposição atômica psicoacústica, descrito no Capítulo 2, onde é demonstrado que o algoritmo seleciona um dicionário de exponenciais complexas *(DEC)* parametrizadas para um subconjunto de átomos, que são os mais correlacionados com os padrões existentes no sinal. Definindo-se $\mathbf{X}$ como o sinal original, obtém-se o sinal aproximado $\hat{\mathbf{X}}$ com K termos, apresentado novamente por conveniência:

$$\mathbf{x} \approx \hat{\mathbf{x}} = \sum_{k=0}^{K-1} \alpha_{m_k} \mathbf{g}_{m_k} \tag{51}$$

em que cada componente é caracterizado por seu coeficiente α_{m_k} correspondente, e por um conjunto de parâmetros m_k, definido pela exponencial complexa g_{m_k}. Ao final da decomposição, tem-se a sequência dos pares $(\alpha_{m_k}, m_k), k = 1, 2, ..., K$, que formam o livro de estruturas, cujos parâmetros são dados por $m_k = (a_k, f_k, \phi_k)$.

Após a etapa de quantização, realiza-se a otimização da taxa-distorção através de curvas operacionais, que permitem a alocação ótima de bits (essa técnica será apresentada adiante). Definida a alocação ótima de bits entre os coeficientes e parâmetros dos átomos, quantiza-se o livro de estruturas, produzindo assim símbolos que são codificados e transmitidos ao decodificador. No decodificador, o feixe de *bits* é decodificado, gerando os símbolos. Estes, por sua vez, sofrem o processo de quantização inversa, produzindo o livro de estruturas quantizado. Por fim, com base neste livro de estruturas, reconstrói-se o sinal.

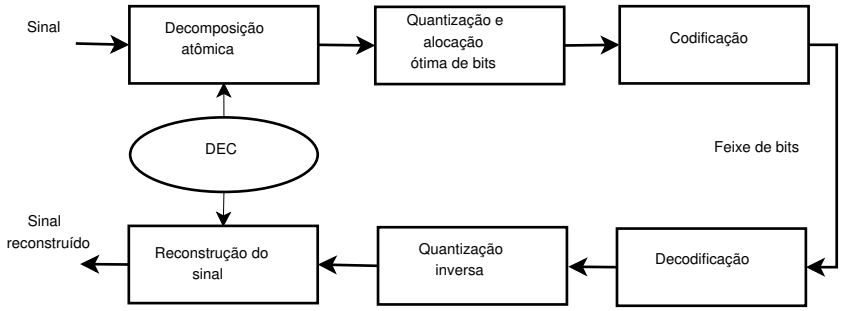


Figura 23 - Compressão de sinais de áudio realizando a decomposição atômica do sinal via DEC e o binômio taxa-distorção através de curvas operacionais.

3.2 Quantização

O tipo de quantizador mais simples é o escalar uniforme, onde todos os intervalos possuem tamanho único, isto é, os limites de decisão são espaçados uniformemente. Os quantizadores podem ser do tipo *midrise*, onde não possuem zero no seu nível de saída, ou do tipo *midtread*, onde possuem um nível de saída zero. Se R for o número de bits, o quantizador *midtread* permite utilizar $2^R - 1$ códigos diferentes em relação aos 2^R códigos permitidos pelo quantizador *midrise*. Apesar do menor número de códigos permitidos, em geral, dada a distribuição das amplitudes dos sinais de áudio, os quantizadores do *midtread* produzem melhores resultados [14], pois é possível representar períodos de silêncio.

O coeficiente α e cada parâmetro de m_k associados aos elementos do livro de estruturas são quantizados utilizando-se um quantizador escalar uniforme definido como [34]:

$$x_q = I_x \ \Delta_{q(x)}, \text{ onde } I_x = \left\lfloor \frac{x + \frac{\Delta_q(x)}{2}}{\Delta_{q(x)}} \right\rfloor, \tag{52}$$

em que x é qualquer parâmetro, x_q representa a sua versão quantizada, $\Delta_{q(x)}$ é o passo de quantização, e I_x corresponde ao símbolo associado a x. Os parâmetros são quantizados de acordo com um intervalo dinâmico definido por seus respectivos valores máximo e mínimo dentre todos os elementos do livro de estruturas [34],

$$\Delta_{q(x)} = \begin{cases} \frac{x_{max} - x_{min}}{2^{b_x} - 1}, \text{se o quantizador for } \textit{mid-rise}, \\ \frac{x_{max} - x_{min}}{2^{b_x} - 2}, \text{se o quantizador for } \textit{mid-tread}, \end{cases} \tag{53}$$

onde b_x é o número de bits alocados a x.

Os parâmetros de amplitude normalizada (a_k) são quantizados de acordo com um intervalo dinâmico definido por seus respectivos valores de amplitude máxima (a_{max}) e mínima (a_{min}), dentre todos os elementos presentes no livro de estruturas, assim:

$$q_{a_k} = \frac{a_{max} - a_{min}}{2^{b_{a_k}} - mid}, \tag{54}$$

em que b_{a_k} é o número de bits alocados à amplitude a_k, *mid* é a especificação do quantizador uniforme utilizado, no caso do quantizador ser *mid-rise* terá valor igual a 1 e, no caso do quantizador ser *mid-tread*, seu valor será igual a 2. A fase ϕ_k é uniformemente quantizada fazendo-se o seu valor máximo igual a $\phi_{max} = 2\pi$ e o seu valor mínimo igual a $\phi_{min} = 0$. A frequência f_k é quantizada de acordo com a discretização da frequência do dicionário de exponenciais complexas. O número de bits alocados para a frequência é $r_f = \log_2(M/2)$ bits, onde $k = 0, 1, \ldots M-1$ e M é a cardinalidade do dicionário.

Assim, pode-se definir o número de bits associados à representação (ou codificação) de um átomo como

$$r = r_a + r_f + r_\phi, \tag{55}$$

em que r_a é a quantidade de bits alocados a amplitude, r_f é a quantidade de bits alocados a fase e, r_ϕ é a quantidade de *bits* alocados para a frequência. Dessa forma, é possível determinar o número total de *bits* gastos por rN_{iter}, onde N_{iter} é o número de iterações necessários para a decomposição do sinal do i-ésimo bloco.

3.3 Otimização Taxa-Distorção

No processo de codificação, deve-se definir o compromisso da taxa de *bits* do codificador com a fidelidade do sinal decodificado—mais *bits* aumentam a carga de taxa de *bits* e, simultaneamente, reduzem o erro de quantização. Usar alguma forma de alocação de *bits* para controlar o nível de erro de quantização é uma característica fundamental dos codificadores de áudio [14]. O objetivo da otimização taxa-distorção é obter a melhor reprodução do sinal para uma dada taxa de compressão alvo [35].

Após a decomposição perceptiva ter sido executada, uma alocação ótima de bits ainda pode ser usada para minimizar uma medida objetiva de distorção [35]. O critério de medida de distorção do trabalho é definido como o erro médio quadrático do sinal, isto é,

a diferença quadrática média entre a entrada e a saída do quantizador. É desejável obter a menor distorção para uma determinada taxa ou a menor taxa para uma determinada distorção [34].

A distorção total pode ser expressa como uma função das taxas de *bits* dos coeficientes e dos parâmetros, resultando em:

$$d_s = f(r_a, r_f, r_\phi). \tag{56}$$

Considere o quantizador uniforme definido pela Equação (52) e os comprimentos de *bits* de cada parâmetro na tripla $b_k = (r_a, r_f, r_\phi) \in \mathcal{B}$, onde $\mathcal{B}$ representa o conjunto de todas as possíveis combinações permitidas de taxas de *bits* dentro do intervalo definido por cada elemento b_k, com $k = [1, 2, ..., K_\mathcal{B}]$ e $K_\mathcal{B}$ o número de elementos em $\mathcal{B}$. A fim de se obter o melhor compromisso taxa-distorção, deve-se buscar o $\mathbf{b}_k$ que minimiza a distorção total inserida no processo de codificação, dada uma quantidade de *bits* disponíveis r_{alvo}. A solução é obtida através da resolução do seguinte problema de otimização [35]:

$$\begin{aligned} &\min_{\mathbf{b}_k \in \mathcal{B}} \; d_S \\ &\text{sujeito a} \quad Mr \geq r_{alvo} \end{aligned} \tag{57}$$

A solução clássica para este problema é baseada na versão discreta da otimização de Lagrange, que tem como ideia básica a introdução de um número real e não negativo chamado de multiplicador de Lagrange $\lambda \geq 0$ que auxilia a minimização da função-custo Lagrangeana [35]

$$J \;=\; d_S + \lambda(\; Mr - r_{alvo} \;), \tag{58}$$

onde M é o número de elementos do livro de estruturas.

Para um dado λ é possível encontrar o par $(d_S^{opt}, r_{alvo}^{opt})$ para o qual J é mínimo, como está ilustrado na Figura 24. Os pontos ótimos relativos a diferentes λ's formam a curva taxa-distorção operacional. O problema se soluciona com a plena minimização da função dada em (58), que é obtida através da resolução do seguinte sistema de equações:

$$\frac{\partial J}{\partial r_a} \;=\; 0 \tag{59}$$

$$\frac{\partial J}{\partial r_f} = 0 \tag{60}$$

$$\frac{\partial J}{\partial r_\phi} = 0 \tag{61}$$

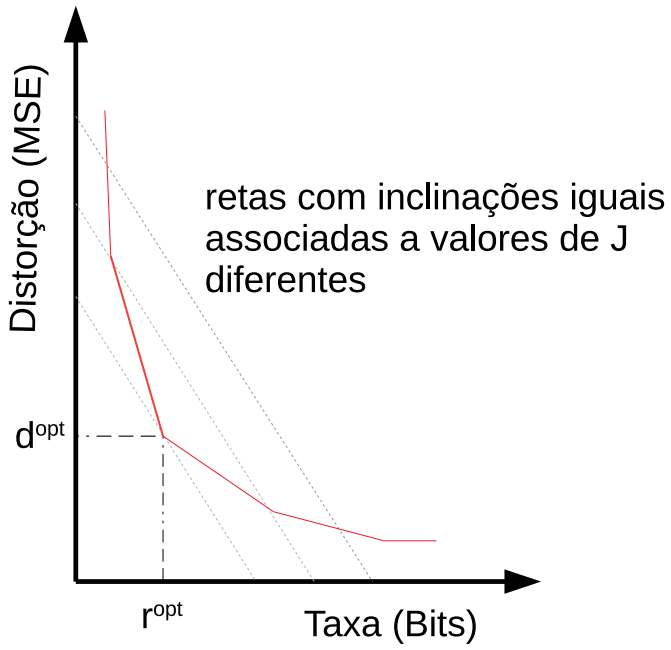


Figura 24 - Interpretação gráfica da otimização da função Lagrangeana.

A fronteira ótima é então definida pelo fecho convexo do conjunto de pontos operacionais representados na Figura 24. Quando não existir forma fechada para d_S em função das taxas (r_a, r_f, r_ϕ), é possível adotar uma abordagem empírica para se obter as curvas operacionais. Para cada $\mathbf{b}_k = (r_a, r_f, r_\phi)$, e para um dado sinal, calcula-se o par taxa-distorção (r_k, d_k), resultando no gráfico taxa-distorção $(T - D)$ apresentado na Figura 25. É nítido que o ponto C possui uma distorção igual ao ponto A, mas com maior taxa de dados; assim, o ponto C não é a melhor escolha. Da mesma forma, o ponto B é pior do que o A, porque apresenta uma maior distorção para uma mesma taxa. Portanto, seleciona-se o ponto que pertence ao fecho convexo, tal que, para uma desejada taxa de compressão, fornece a quantização de coeficientes que leva à distorção mínima.

A curva operacional é obtida quadro a quadro conectando-se os pontos que pertencem ao fecho convexo da região definida pelos pares $T - D$ gerados para cada $\mathbf{b}_k \in \mathcal{B}$, onde $k \in \mathcal{K} = \{1, 2, ..., K_\mathcal{B}\}$, onde $K_\mathcal{B}$ é o número de elementos em $\mathcal{B}$. O procedimento

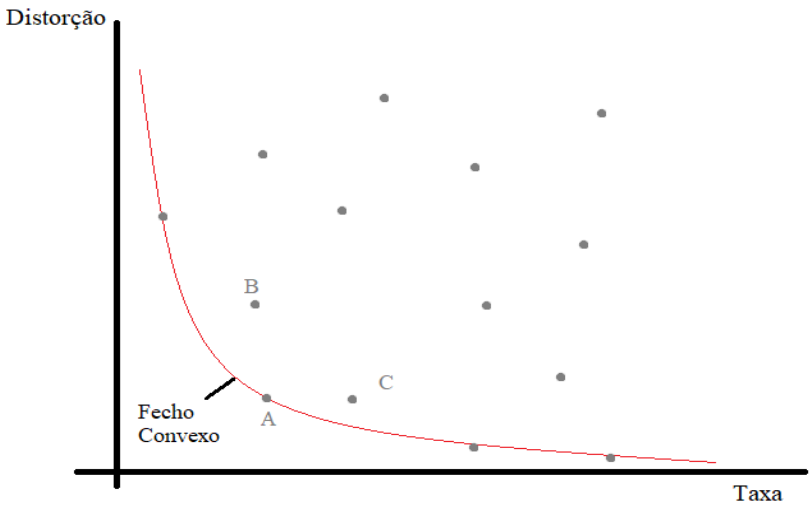


Figura 25 - Fecho Convexo contendo os pontos ótimos em termos de taxa-distorção.

para se obter o fecho convexo é definido da seguinte forma:

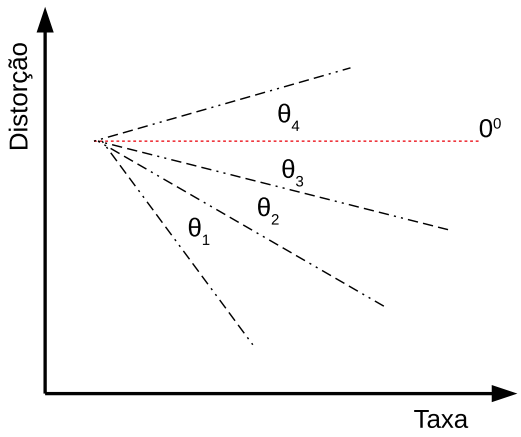


Figura 26 - Traçando o fecho convexo. Neste caso $\theta_{min} = \theta_1$.

1. Busca-se $\mathbf{b_{Ko}} = \arg\min_{\mathbf{b}_k \in \mathcal{B}} \ r_k$;

2. Atribui-se $\mathbf{b_{Katual}} = \mathbf{b_{Ko}}$, portanto $[r_{Katual}; d_{Katual}] = [r_{Ko}; d_{Ko}]$;

3. Traça-se uma reta do ponto $[r_{Katual}; d_{Katual}]$ a todos os outros pontos $[r_k, d_k]$, onde $r_k > r_{Katual}$, como ilustrado na Figura 27. Cada reta possui um ângulo θ_k com

a horizontal (correspondente à inclinação 0^o), calculada da seguinte maneira $\theta_k = \arctan\left(\frac{d_k - d_{katual}}{r_k - R_{katual}}\right)$;

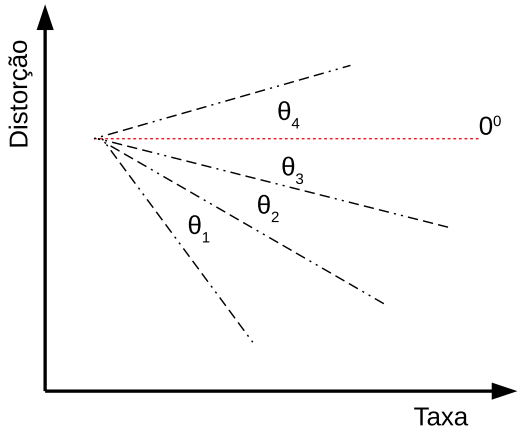


Figura 27 - Traçando o fecho convexo. Neste caso $\theta_{min} = \theta_1$.

4. Obtêm-se $\mathbf{b_{kproximo}}$, cujo par correspondente $[r_{kproximo}; d_{kproximo}]$ possui o menor ângulo θ_k, ou seja, $\theta_{min} = min_k\theta_k$;

5. Se $\theta_{min} \leqslant 0$, inclui-se $\mathbf{b_{Katual}}$ na curva operacional e atualiza-se $\mathbf{b_{Katual}} = \mathbf{b_{Kproximo}}$;

6. Se $\theta_{min} > 0$, interrompe-se o procedimento;

7. Repetem-se os procedimentos de 3 a 6 até alcançar-se o par de r_k máximo;

8. Ao fim, os pontos pertencentes à curva operacional correspondem aos $\mathbf{b_k}$ ótimos do sinal.

A otimização taxa-distorção é realizada quadro a quadro, de forma independente, e considerando a taxa alvo. Assim, a alocação de *bits* é localmente ótima a cada quadro, não sendo globalmente ótima em relação a todo o sinal. Para que a alocação seja globalmente ótima, o quadro de quantização deve possuir o mesmo comprimento do sinal. No entanto, como os sinais de áudio normalmente possuem longa duração, portanto inúmeras amostras, o processo de otimização taxa-distorção nessa situação torna-se impraticável em termos computacionais.

O que se faz é obter as curvas operacionais de taxa-distorção dos blocos e encontrar um multiplicador de Lagrange $\hat{\lambda}$, associado a um ângulo $\hat{\theta}$ que resulte em uma taxa $\hat{r}$ próxima da taxa desejada.

Este capítulo apresentou de forma resumida os aspectos teóricos da quantização e otimização de alocação de bits, com o objetivo de se obter um compromisso otimizado entre taxa e distorção associada que será utilizada no restante do trabalho.

4 PROCEDIMENTOS E RESULTADOS EXPERIMENTAIS

Neste capítulo são apresentados os procedimentos e resultados experimentais envolvendo o uso do codificador perceptivo proposto neste trabalho.

4.1 Procedimentos Experimentais

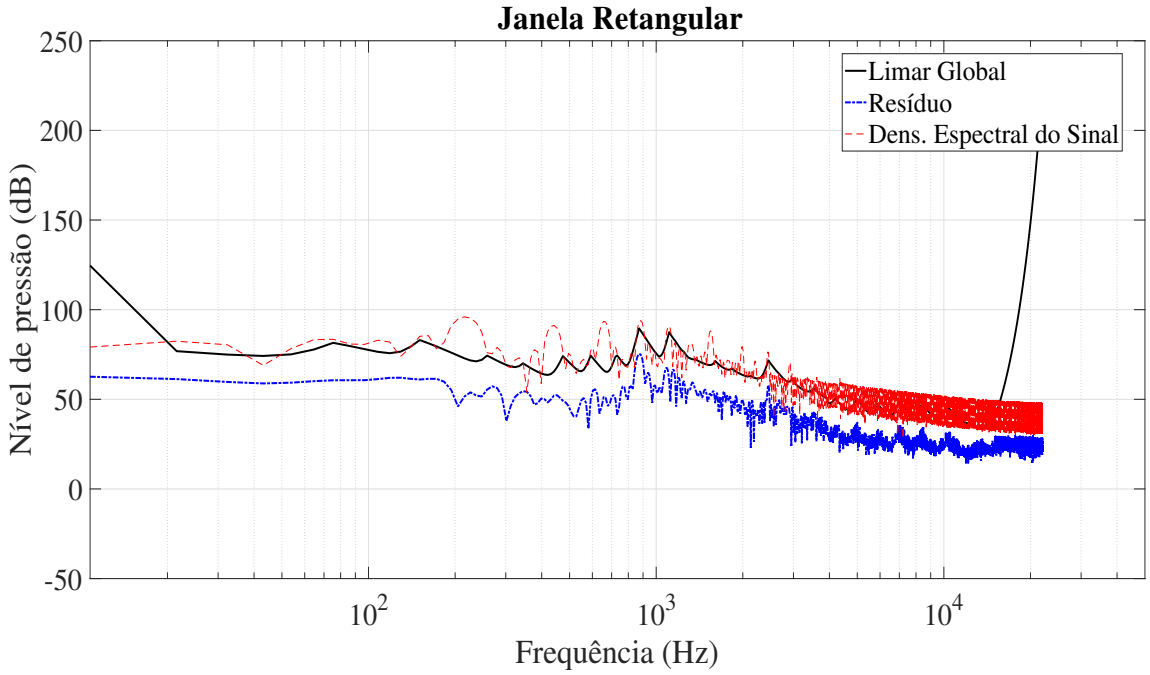
Conforme visto nos capítulos anteriores, o que se deseja alcançar com este trabalho é a representação compacta de sinais de áudio, no qual os componentes não audíveis dos sinais possam ser ignorados. Tal ação torna viável o armazenamento e a transmissão eficiente desses sinais em contextos de banda e espaço em memória reduzidos. Para tanto, é utilizado o Matching Pursuit, uma ferramenta de decomposição de sinais, considerando o limiar global de mascaramento como critério de decisão de quais componentes espectrais são consideradas audíveis.

Inicialmente, o sinal de áudio sob análise foi dividido em Q quadros de tamanho N, ponderados por uma função janela com saltos de l amostras. Neste trabalho, foram testadas duas funções janela: a retangular com saltos de $l = N = 512$ amostras, e a de Hanning, com saltos de $l = N/2 = 256$ amostras. A escolha da janela de Hanning se deve ao fato dela oferecer boa resolução em frequência e dispersão espectral reduzida.

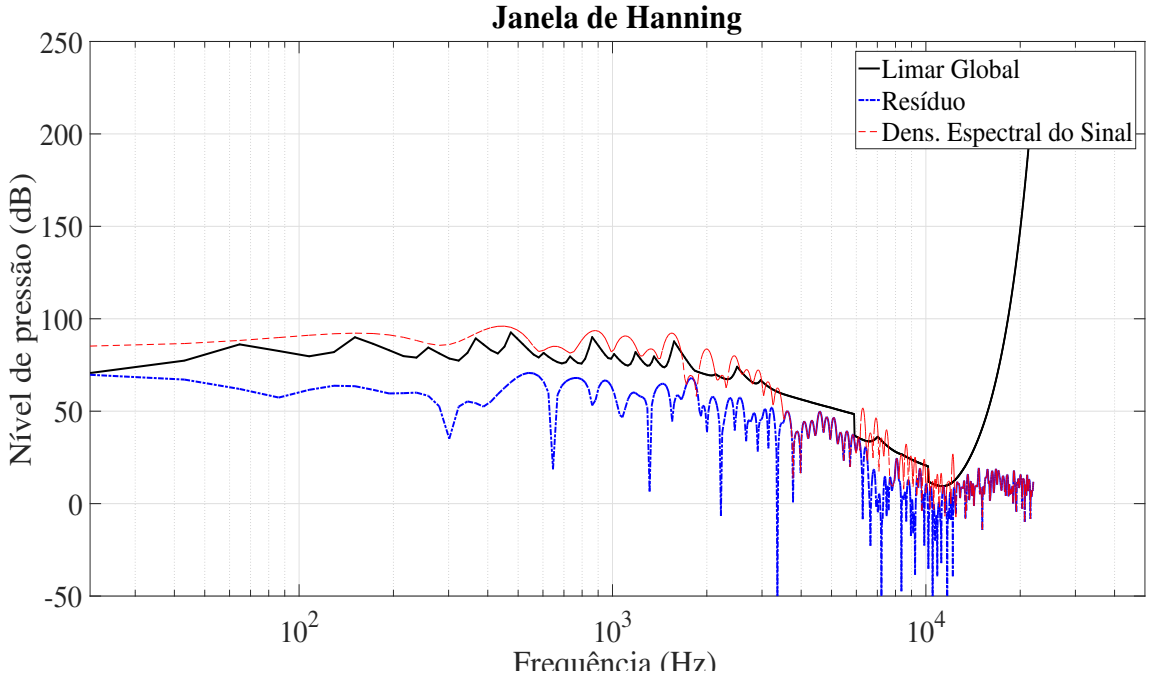
É interessante destacar que a janela de Hanning é capaz de fornecer uma estimação da densidade espectral de potência do sinal mais precisa, permitindo assim melhor identificação dos componentes tonais e não-tonais no processo de obtenção do limiar global e na decomposição por blocos do sinal. Este fato pode ser verificado na Figura 28 (a) e (b), onde estão representadas as densidades espectrais de potência do sinal e o seu limiar global para as janelas retangular e de Hanning, respectivamente.

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. Dessa forma, mais átomos são extraídos para compor a aproximação do sinal. A Figura 29 ilustra alguns exemplos de margens que serão subtraídas no limiar global psicoacústico.

A quantização do livro de estruturas é realizada por dois quantizadores escalares uniformes, um do tipo *midrise* e outro do tipo *midtread*. No quantizador midtread, garante-se a representação de valores nulos, com a contrapartida de ter um nível a menos



(a)



(b)

Figura 28 - Representação de um bloco de um sinal de áudio utilizando a janela: (a) Retangular e (b) Hanning.

de representação que o quantizador midrise, resultando assim, em um maior nível de ruído de quantização. A alocação ótima de bits é realizada através de otimização taxa-distorção apresentada no Capítulo 3, ajustando-se o multiplicador de Lagrange.

A avaliação dos sinais reconstruídos é realizada por meio do algoritmo PEAQ.

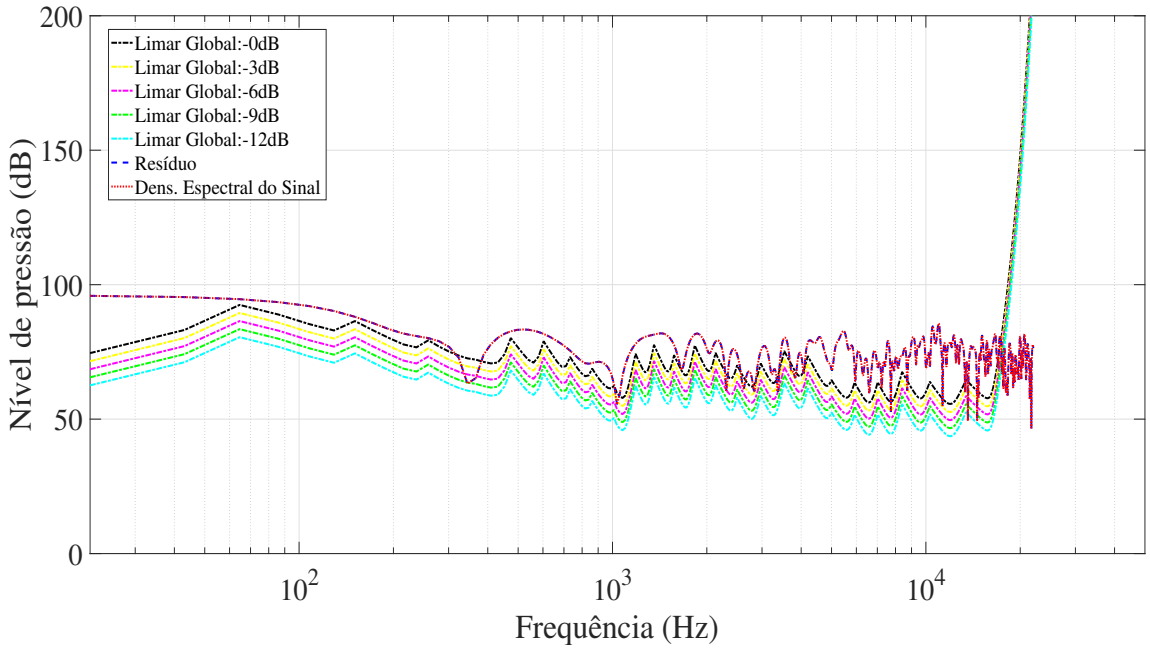


Figura 29 - Exemplos de diferentes limiares psicoacústicos em dB.

Trata-se de um algoritmo que mede objetivamente a qualidade de sinais de áudio, padronizado pelo *International Telecommunications Union* (ITU), na recomendação ITU-R BS.1387 [36]. Essa medida de qualidade é classificada nas seguintes faixas:

- -4 a -3: muito perturbador;
- -3 a -2: perturbador;
- -2 a -1: pouco perturbador;
- -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, nota A4 de um fagote e dois trechos de bateria denominados Bateria A e Bateria B. 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 taxa de amostragem de 44,1 kHz, logo têm 44.100 amostras. Para a decomposição, são divididos em blocos de 512 amostras com sobreposição de 256 amostras.

4.2 Resultados Experimentais

Os resultados experimentais são apresentados em duas partes, a primeira parte estão os resultados referentes a decomposição atômica e a segunda são os referentes a alocação ótima de bits.

4.2.1 Decomposição Atômica

Os resultados da avaliação PEAQ para decomposições realizadas nos sinais de áudio, onde o dicionário possui redundância de 4 e 8 vezes o tamanho dos quadros (N) e as margens a serem subtraídas no limiar global psicoacústico variam de 0 a 10 dB, estão apresentados na Tabela 1 [37].

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

	Piano A3		Violoncelo A4		Fagote A4		Flauta A4		Bateria A		Bateria B	
Redundância / Margem(dB)	4	8	4	8	4	8	4	8	4	8	4	8
0	-0,549	-0,613	-1,576	-1,353	-1,534	-0,743	-2,743	-1,33	-0,385	-0,36	-0,265	0,277
1	-0,62	-0,583	-1,381	-1,203	-1,429	-0,762	-2,618	-1.168	-0,309	0,27	-0,216	-0,226
2	-0,405	-0,552	-1,249	-1,029	-0,991	-0,659	-2,295	-0,876	-0,264	-0,231	-0,158	-0,146
3	-0,33	-0,404	-1,066	-0,839	-1,144	-0,578	-1,933	-0,705	-0,169	-0,162	-0,125	-0,13
4	-0,2	-0,28	-0,854	-0,683	-0,744	-0,465	-1,52	-0,521	-0,147	-0,138	-0,099	-0,104
5	-0,107	-0,191	-0,678	-0,509	-0,662	-0,549	-1,295	-0,4	-0,116	-0,088	-0,089	-0,064
6	-0,092	-0,092	-0,591	-0,385	-0,396	-0,367	-0,955	-0,249	-0,081	-0,061	-0,062	-0,058
7	-0,032	-0,05	-0,474	-0,271	-0,479	-0,236	-0,68	-0,159	-0,03	-0,037	-0,037	-0,037
8	0,002	-0,006	-0,342	-0,174	-0,279	-0,102	-0,558	-0,084	-0,027	-0,021	-0,001	-0,006
9	0,033	0,04	-0,196	-0,051	-0,207	-0,037	-0,418	-0,002	-0,016	-0,019	0,01	0,009
10	0,067	0,082	-0,062	0,02	-0,107	-0,038	-0,223	0,049	-0,001	0,011	0,002	0,017
Média	-0,203	-0,241	-0,769	-0,589	-0,725	-0,412	-1,385	-0,495	-0,141	-0,125	-0,095	-0,093
Desvio Padrão	0,239	0,261	0,499	0,469	0,492	0,274	0,904	0,472	0,129	0,119	0,091	0,095

Como é de se esperar existe uma tendência geral de melhora utilizando um dicionário com maior redundância, pois o aumento da redundância do dicionário possibilita a representação do sinal de maneiras diferentes. Assim, é possível caracterizar melhor as várias formas e padrões presentes no sinal. Outro ponto observado é que, quanto maior for a margem subtraída do limiar global psicoacústico, melhor é o resultado perceptivo da decomposição. Este fato ocorre porque o aumento da margem acarreta o aumento do número de iterações necessárias para se alcançar o critério de parada, aumentando assim o número de elementos que descrevem o sinal.

Na Tabela 1 é possível observar que os sons dos diferentes instrumentos utilizados no trabalho — a saber: cordas, sopro e percussão — exigem diferentes margens para alcançar notas equivalentes na avaliação PEAQ. Uma variação do número de iterações necessárias para cada uma das margens subtraídas do limiar psicoacústico de 0 dB até 10 dB estão representadas da Figura 30 até a Figura 35 para os sinais utilizados no trabalho, com dicionário de redundância correspondendo a 4 vezes o tamanho do quadro N. Na Tabela 2 está ilustrado o número de iterações médio utilizados para cada uma das margens subtraídas do limiar psicoacústico de 0 dB até 10 dB para diferentes instrumentos utilizado no trabalho, com dicionário de redundâncias de 4 e 8 vezes o tamanho do quadro N.

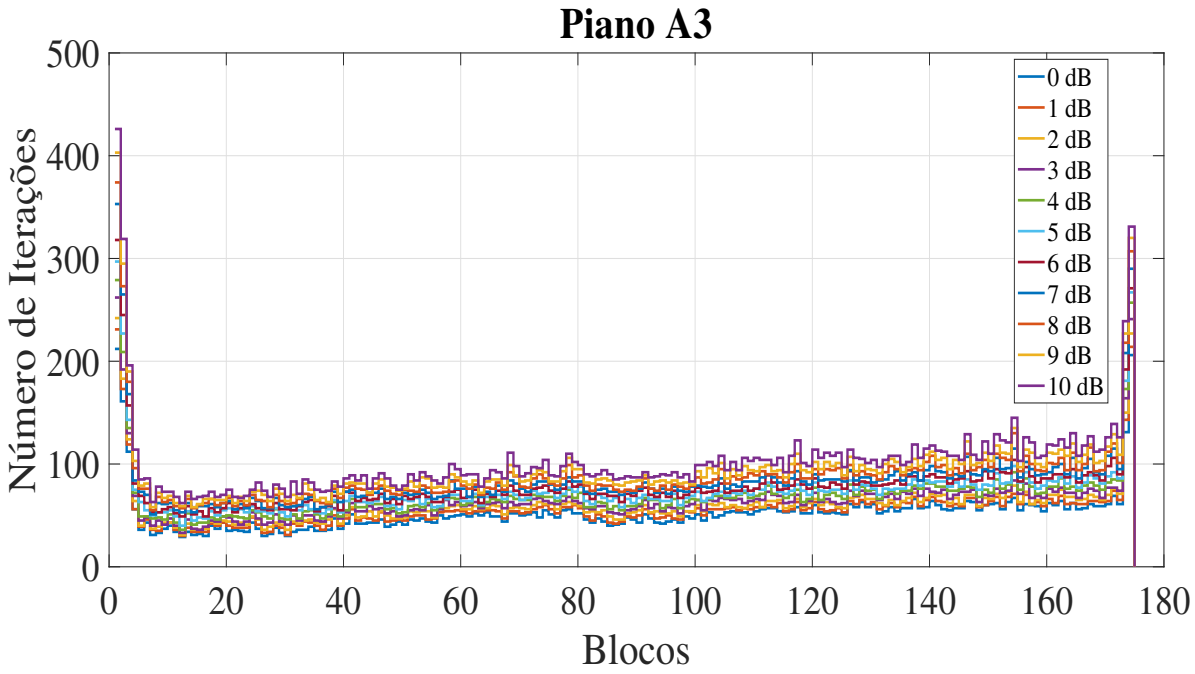


Figura 30 - Número de Iterações por Bloco com Redundância de dicionário de 4N para o Piano A3

Os diferentes tipos de instrumentos apresentam comportamentos sonoros variados.

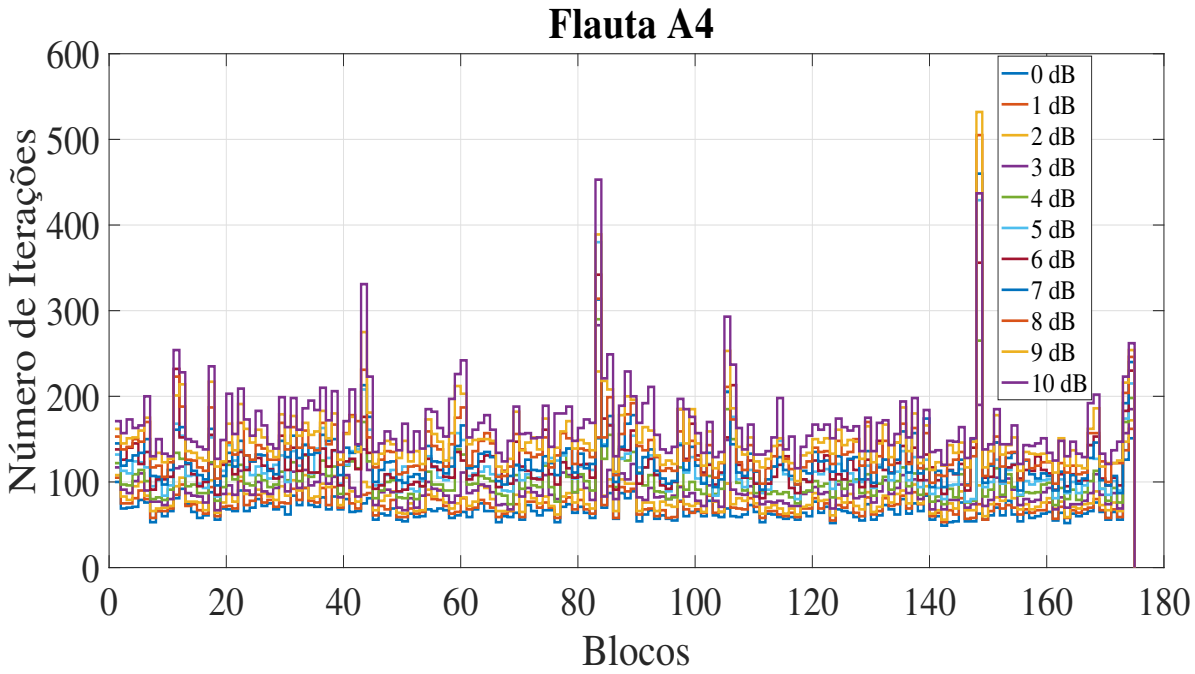


Figura 31 - Número de Iterações por Bloco com Redundância de dicionário de 4N para a Flauta A4

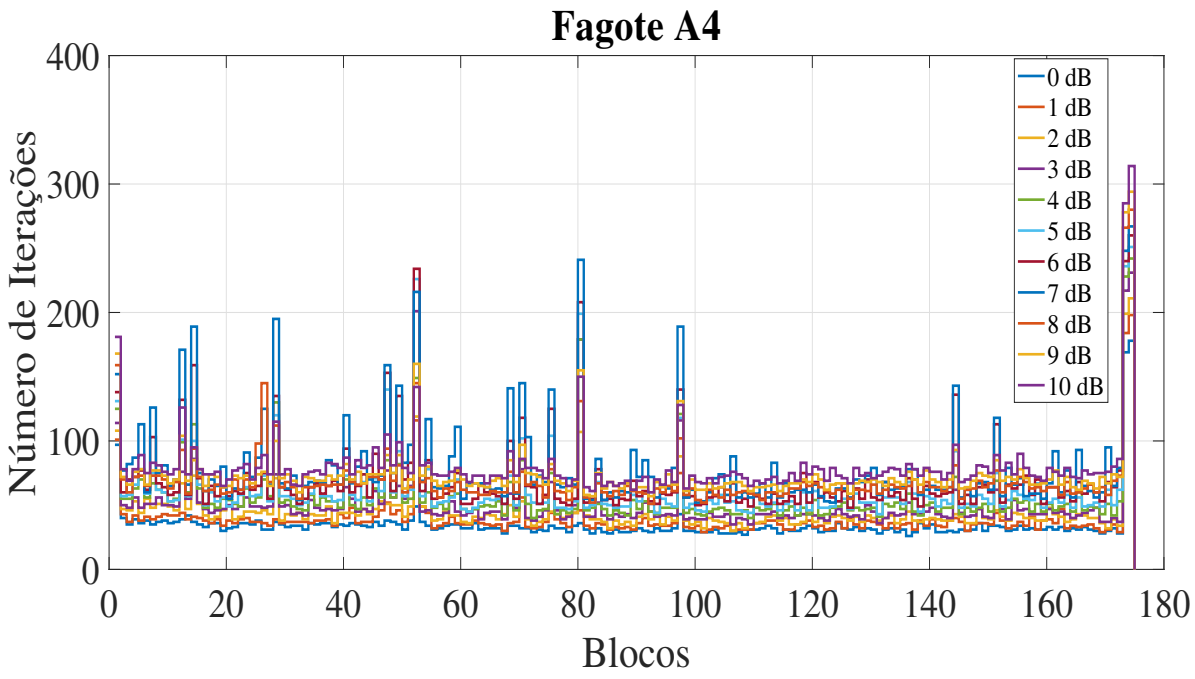


Figura 32 - Número de Iterações por Bloco com Redundância de dicionário de 4N para o Fagote A4

Essa distinção altera o número de iterações necessárias para se obter uma boa avaliação perceptiva da decomposição atômica, conforme o limiar psicoacústico utilizado. Esse comportamento sugere que instrumentos cujos sons têm variações mais acentuadas na am-

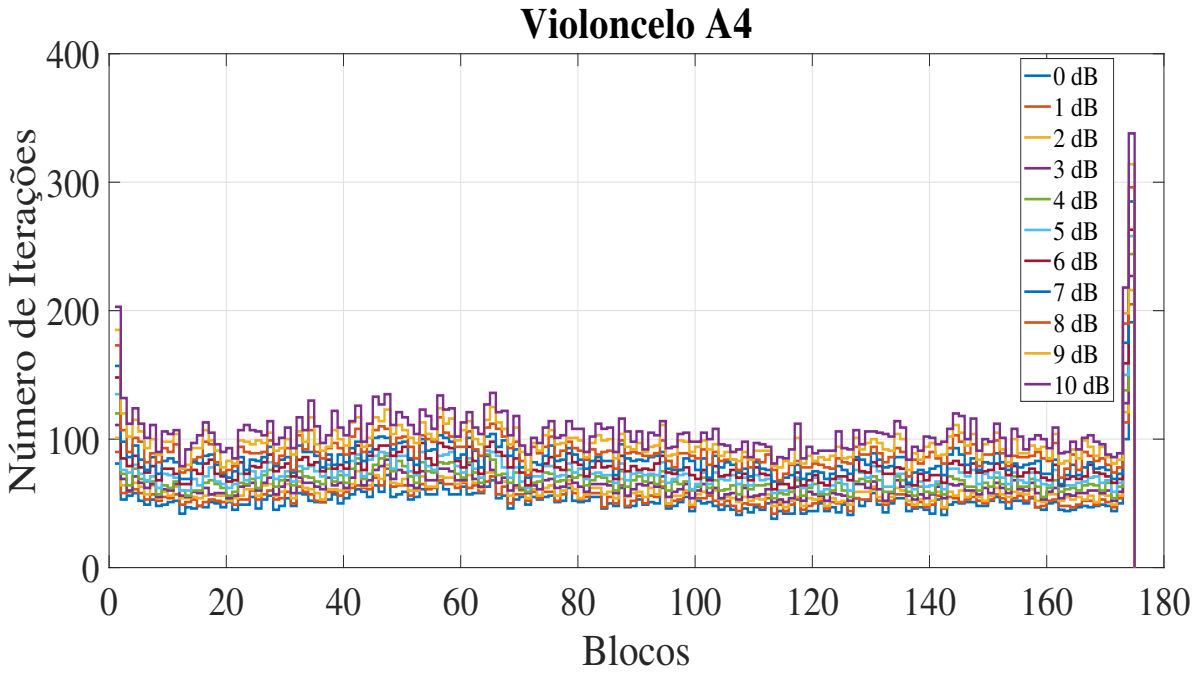


Figura 33 - Número de Iterações por Bloco com Redundância de dicionário de 4N para o Violoncelo A4

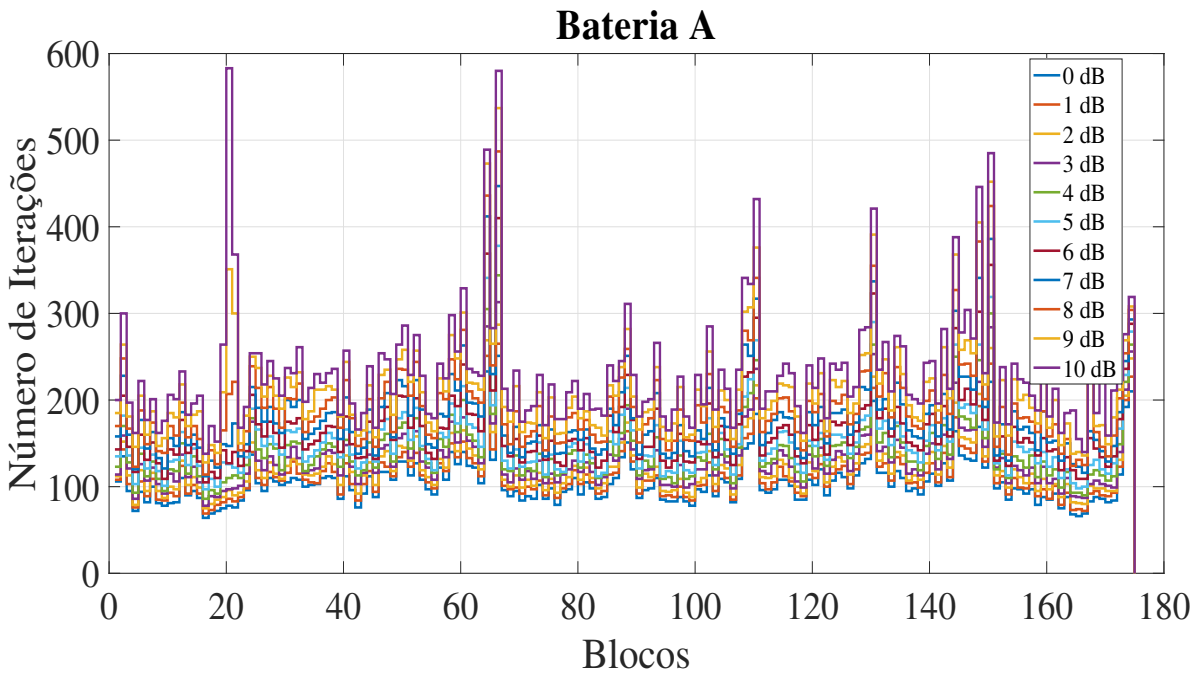


Figura 34 - Número de Iterações por Bloco com Redundância de dicionário de 4N para a Bateria A

plitude e fase da densidade espectral de frequência, como os sinais de *bateria a* e *bateria b* durante o quadro em análise, contribuem para o surgimento de mais fenômenos de mascaramento, tornando a decomposição psicoacústica mais eficaz. O número de fenômenos de mascaramento afeta a quantidade de iterações necessárias para que o Dicionário de

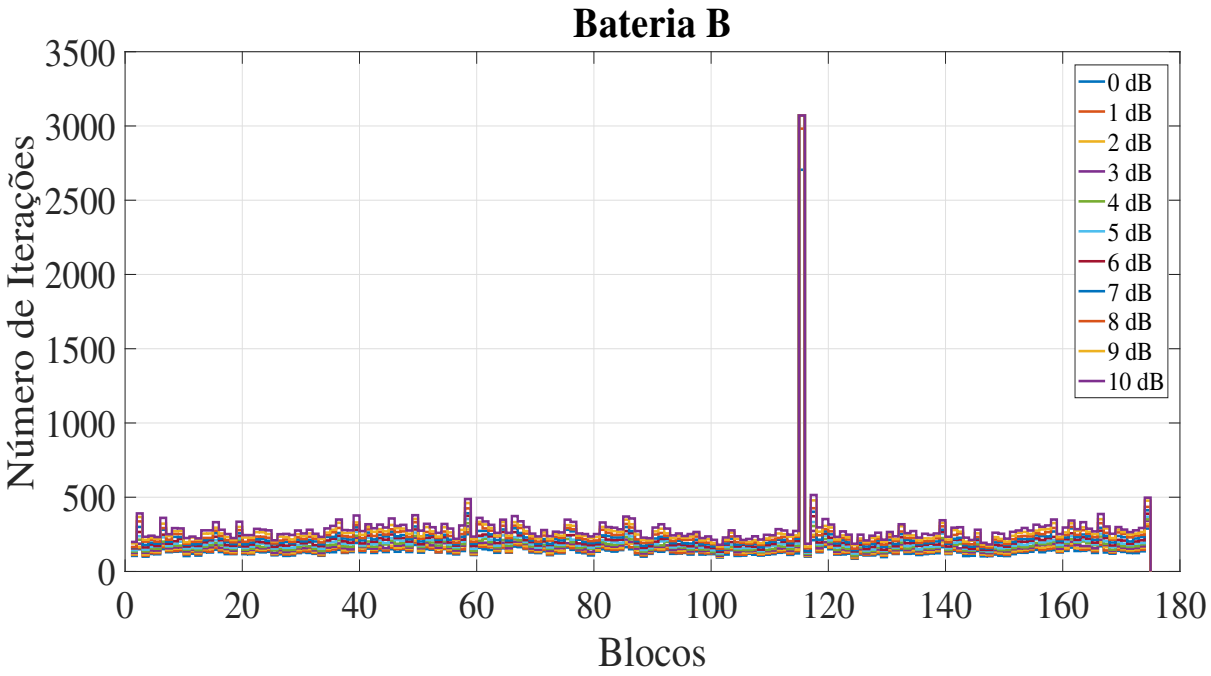


Figura 35 - Número de Iterações por Bloco com Redundância de dicionário de 4N para a Bateria B

Exponenciais Complexas possa representar adequadamente as estruturas que compõem cada quadro do sinal em questão.

Tabela 2 Valores do número médio de iterações para os diferentes sinais decompostos com dicionários que possuem redundâncias de 4 e 8 vezes o número de amostras por bloco.

	Piano A3		Violoncelo A4		Fagote A4		Flauta A4		Bateria A		Bateria B	
Redundância / Margem(dB)	4	8	4	8	4	8	4	8	4	8	4	8
0	52,08	51,95	51,98	52,23	34,37	32,29	66,99	61,10	105,51	104,10	144,42	141,13
1	56,03	55,67	55,70	55,71	39,10	34,38	72,78	66,07	113,48	112,11	155,23	152,03
2	59,81	59,36	59,59	59,35	44,78	36,78	80,50	71,25	122,02	119,85	165,87	162,27
3	64,05	63,39	63,73	63,07	50,63	39,97	90,89	77,02	131,23	128,74	176,54	173,05
4	68,53	67,69	68,16	67,26	55,95	43,96	101,94	82,98	140,88	138,55	188,81	184,65
5	73,15	72,07	73,22	71,83	61,54	48,06	117,15	89,82	151,69	149,22	202,69	197,39
6	77,95	76,53	78,53	76,61	69,30	52,19	120,97	96,83	164,21	161,20	217,27	211,38
7	83,04	81,14	84,29	81,58	78,99	56,70	126,19	104,26	177,93	174,50	233,86	227,63
8	87,95	86,03	90,86	87,25	68,74	62,76	138,87	112,80	192,59	189,03	251,38	245,44
9	93,78	91,39	97,91	93,40	73,91	68,70	151,54	121,69	211,53	206,17	271,43	264,58
10	100,93	97,66	105,71	100,39	79,63	75,60	167,67	132,02	232,09	225,45	292,86	286,09
Média	74,302	72,989	75,425	73,517	59,722	50,126	112,318	92,351	158,468	155,356	209,123	204,149
Desvio Padrão	15,943	14,985	17,703	15,821	15,919	14,456	32,855	23,357	41,311	39,768	48,823	47,445

O número de iterações por bloco do sinal da *bateria b* está representado na Figura 35. Nele é possível observar que existe uma discrepância muito grande no número de iterações utilizados no bloco 115, com margem psicoacústicos com 0 dB, em comparação aos blocos restantes. Na Figura 36, os subgráficos (a) e (b) estão apresentados a densidade espectral de frequência, o limiar global psicoacústico e o resíduo final para os blocos 7 e 115 do sinal *bateria b*, respectivamente. O bloco 7 necessitou de 130 iterações para a sua decomposição, enquanto o bloco 115 demandou 2.705 iterações. Existe um número muito maior de componentes de frequência do sinal no bloco 115 que estão acima do limiar global psicoacústico quando comparado com o bloco 7. Essa diferença de número de componentes de frequência foi a responsável por mais iterações para que toda a energia do sinal que estava acima do limiar global pudesse ser decomposta.

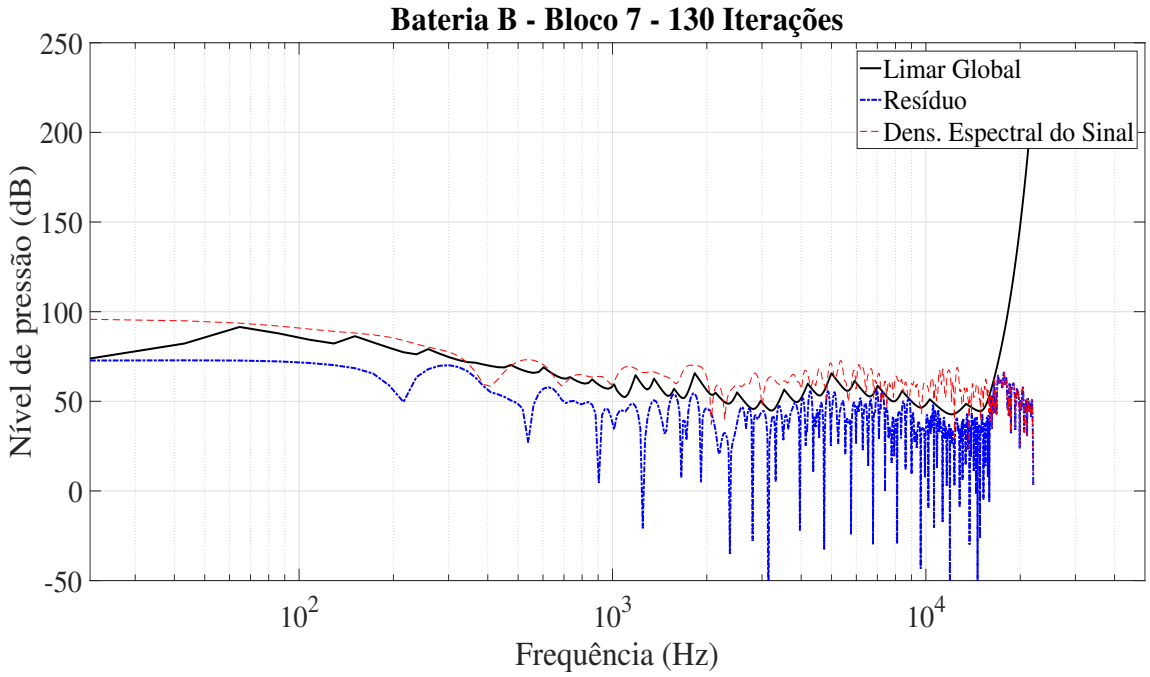
Com o auxílio da Tabela 1 é possível observar que os sinais decompostos pelo algoritmo atingem a nota de PEAQ superior a -1 com a margem de 6 dB. De posse desse critério a Figura 37 apresenta os histogramas dos sinais, com essa figura é possível observar as iterações e as frequências de cada sinal.

4.2.2 Alocação Ótima de Bits

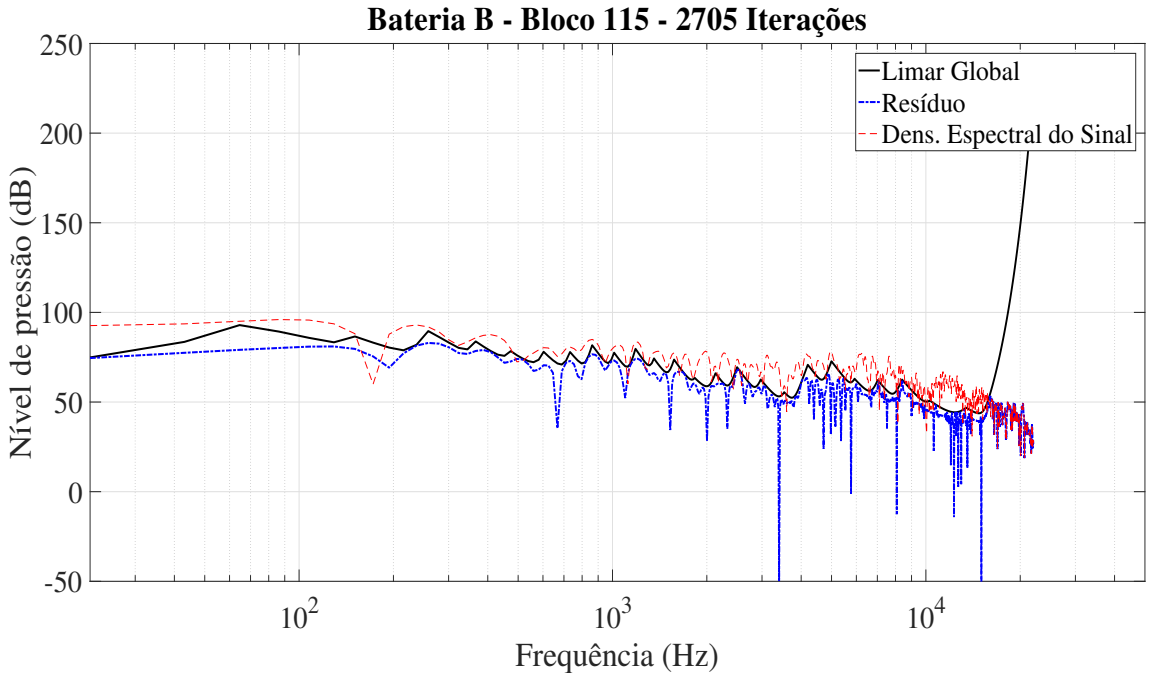
Nesta seção, avaliamos o desempenho da codificação dos sinais realizada com quantizadores escalares uniformes e alocação ótima de *bits* através de taxa-distorção ($T - D$) via multiplicador de Lagrange. O objetivo é conseguir a menor distorção dada uma da taxa de *bits* desejada, lembrando que o número total de *bits* gastos por bloco é determinado por rN_{iter}, onde N_{iter} é o número de iterações e r é o número de *bits* necessários para representar um átomo.

O número de *bits* utilizados para obter a curva de otimização da taxa-distorção para cada áudio é definido da seguinte forma:

- A taxa de amplitude r_a varia de 16 bits até 2 bits;
- A taxa de fase r_ϕ varia de 32 bits até 2 bits;
- Na frequência, o número de bits r_f é dado por $\log_2(M/2)$. Assim é, o número de bits utilizados pelos componentes de frequência é fixo e dependente da cardinalidade do dicionário M, com $M = 2048$ tem-se $r_f = 10$ bits.



(a)



(b)

Figura 36 - Densidade espectral, limiar global psicoacústico e resíduo final do sinal para os blocos: (a) 7 e (b) 115

Dessa forma, o conjunto $\mathcal{B}$ de todas as possíveis combinações permitidas para a alocação ótima de *bits* através de taxa-distorção $(T - D)$ via multiplicador de Lagrange é igual a uma combinação dos bits utilizados, da seguinte forma: $\mathcal{B} = 15 \times 31 \times 10 = 4.650$ bits por quadro.

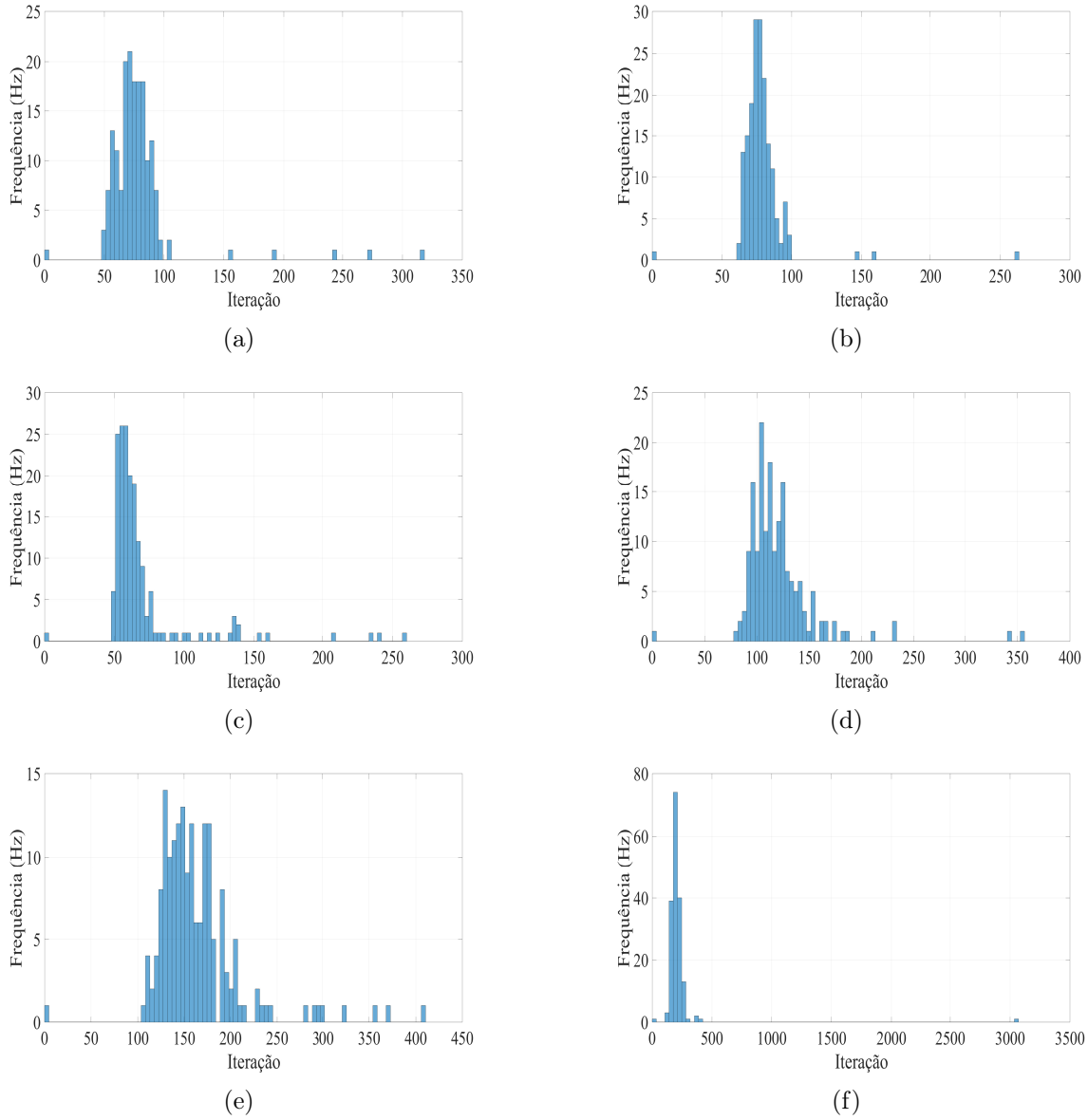


Figura 37 - Histogramas dos sinais com margem de 6 dB e redundância de dicionário de 4N para: (a) Piano, (b) Violoncelo, (c) Fagote, (d) Flauta, (e) Bateria A e (f) Bateria

Os sinais submetidos ao processo de quantização foram decompostos com o DEC, utilizando redundância de $4N$ (para N=512, M=2048) e diferentes margens psicoacústicas. Com o auxílio da Tabela 1, os valores das margens psicoacústicas escolhidas inicialmente foram aqueles que apresentaram seus respectivos valores de PEAQ acima de -0,5, a saber: *piano a3* com margem de 3 dB, *violoncelo a4* com margem de 7 dB, *fagote a4* com margem de 6 dB, *flauta a4* com margem de 9 dB, *bateria a* com margem de 0 dB e *bateria b* com margem de 0 dB.

Durante o processo de codificação, alguns sinais exigiram um maior número de componentes para sua representação. Assim, novas decomposições foram efetuadas para

determinar qual seria a margem psicoacústica a ser utilizada. Esses sinais e suas margens são: *flauta a4*, com margem de 11 dB; e o *fagote a4*, com margem de 9 dB.

A otimização de taxa-distorção através de curvas operacionais foi realizada tanto para quantizadores *midrise* como para quantizadores *midtread*. Para auxiliar a avaliação da qualidade psicoacústica foi elaborada uma curva de qualidade taxa-PEAQ por instrumento, isto é, para cada taxa de *bits* obtida pela curva de otimização da taxa-distorção é gerada uma curva de avaliação de desempenho taxa-PEAQ.

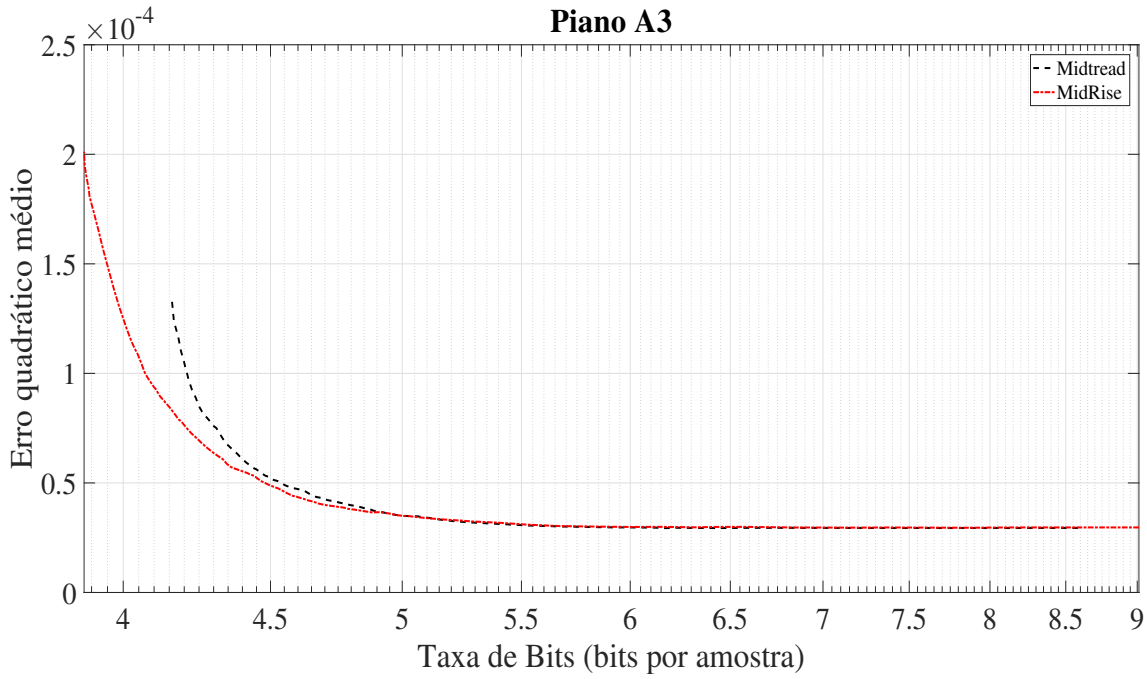
As curvas de taxa-distorção e taxa-PEAQ para os sinais serão apresentadas a seguir. As primeiras curvas estão expostas na Figura 38 e são respectivas ao *piano A3*, com margem psicoacústica de 3 dB e utilizando os quantizadores *midrise* e *midtread*.

A partir de 5,6 bits/amostra, não há uma redução significativa de distorção. A avaliação psicoacústica PEAQ indica tendência de crescimento aproximadamente linear. Conforme aumenta a taxa de bits, ela só é considerada não perturbadora, isto é, o PEAQ é superior a -1, com taxas superiores a 8 bits/amostra. Os dois quantizadores apresentam desempenho equivalentes.

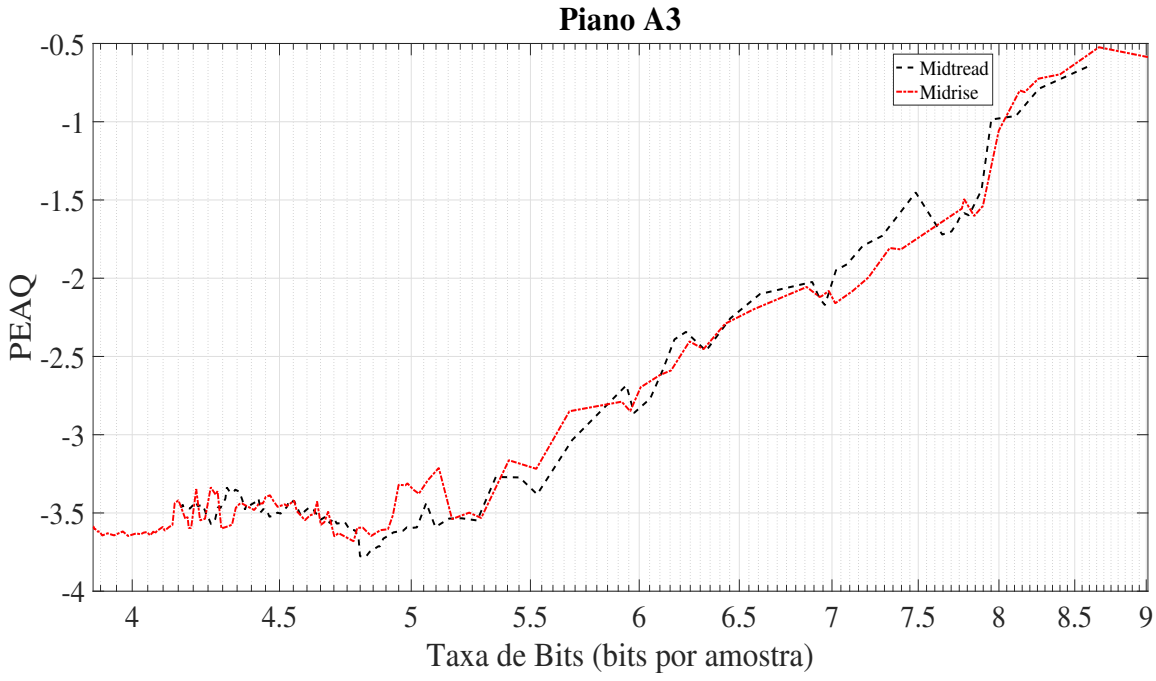
As curvas de taxa-distorção e taxa-PEAQ do *violoncelo a4*, utilizando margem de 7 dB, estão apresentadas na Figura 39. A partir de 7,3 bits/amostra não há redução significativa de distorção. A avaliação psicoacústica por PEAQ apresenta um crescimento acentuado quando a taxa de bits a partir de aproximadamente 7,3 bits/amostra, se tornando estável em valores de -0,6 com taxas superiores a 9 bits/amostra. Mais uma vez, os dois quantizadores possuem desempenho equivalente.

Na Figura 40 estão apresentadas as curvas de taxa-distorção e taxa-PEAQ da *flauta a4*, utilizando margem de 11 dB. A redução de distorção deixa de ser significativa a partir de 14 bits/amostra, para o quantizador *midrise* e para o quantizador *midtread* essa taxa é equivalente a 15 bits/amostra. A avaliação psicoacústica por PEAQ exibe variação nos seus resultados para as taxas de bits entre 13 e 21 bits/amostra, alcançando o valor de -1 com taxas de bits iguais a 21 bits/amostra para *midrise* e 22 bits/amostra para *midtread*. O quantizador *midrise* apresenta um desempenho ligeiramente melhor para a *flauta a4*.

As curvas da Figura 41 correspondem ao *fagote a4*, com margens psicoacústicas de 9 dB. A partir de 6,3 bits/amostra não há redução significativa de distorção para ambos os quantizadores. A avaliação psicoacústica só exibe aumento dos valores das notas PEAQ para taxas a partir de 8 bits/amostra, não atingindo valores acima de -1 nos



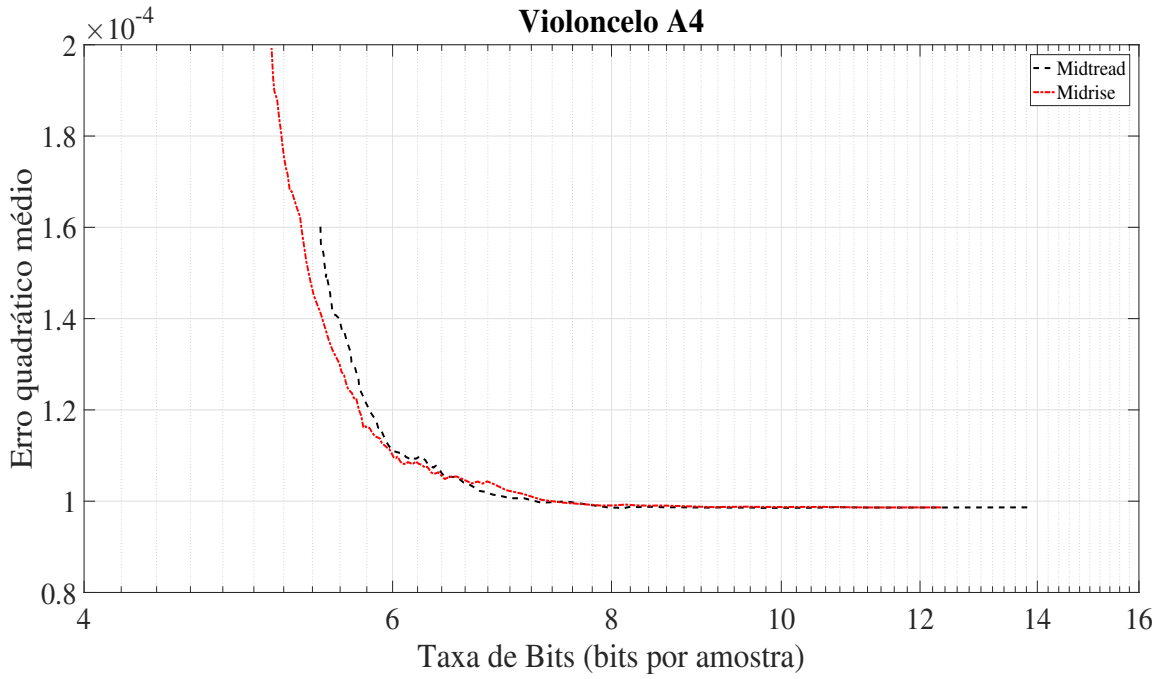
(a)



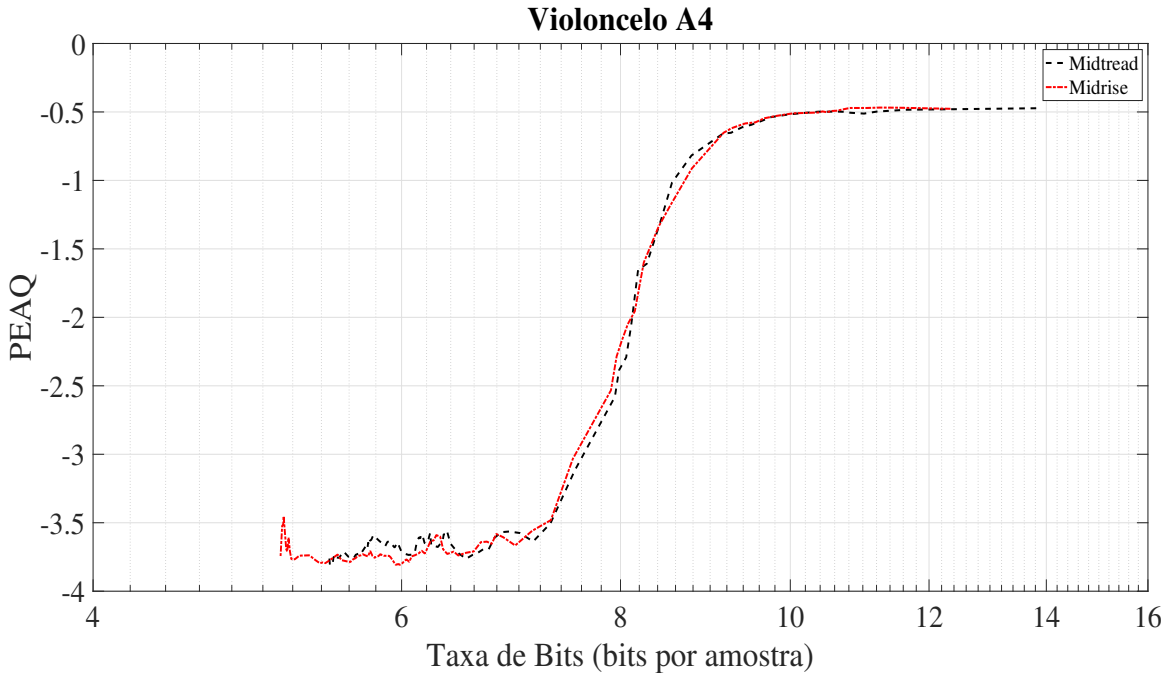
(b)

Figura 38 - Curvas de otimização do *piano a3* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread*

testes realizados. Os valores PEAQ acima de -2, considerados pouco perturbadores, estão relacionados as taxas superiores a 9,8 bits/amostra quando o quantizador é *midrise* e 11 bits/amostra para o quantizador *midthread*. Os quantizadores não apresentaram bons resultados nesse teste com a margem psicoacústica utilizada.



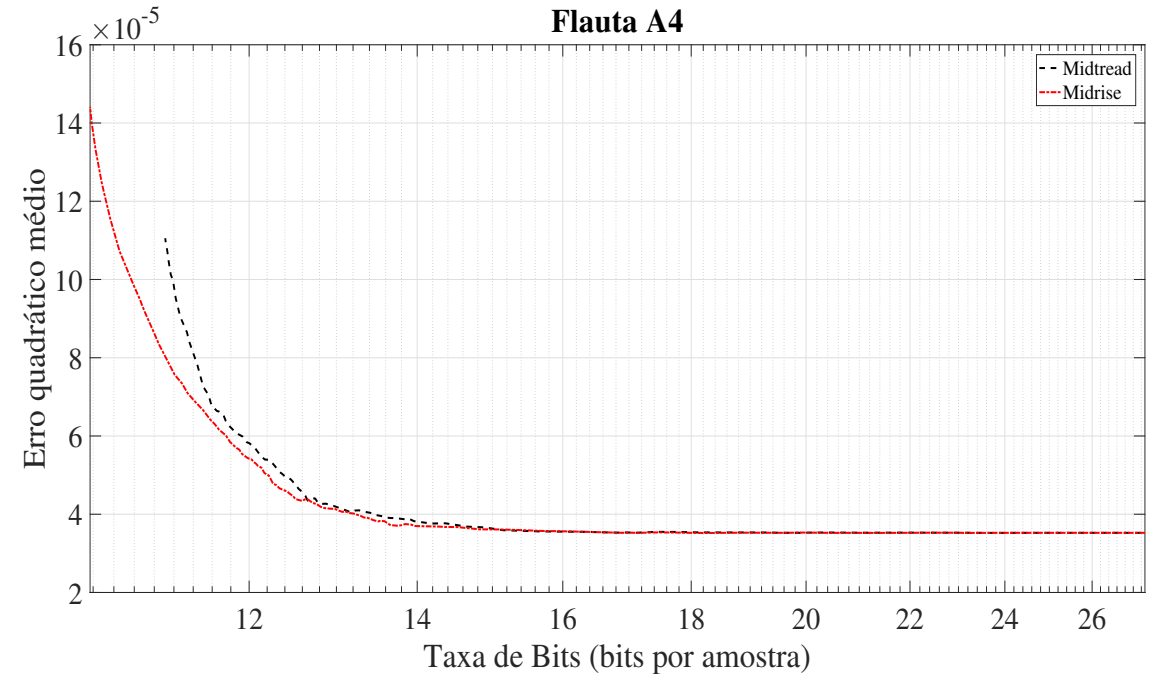
(a)



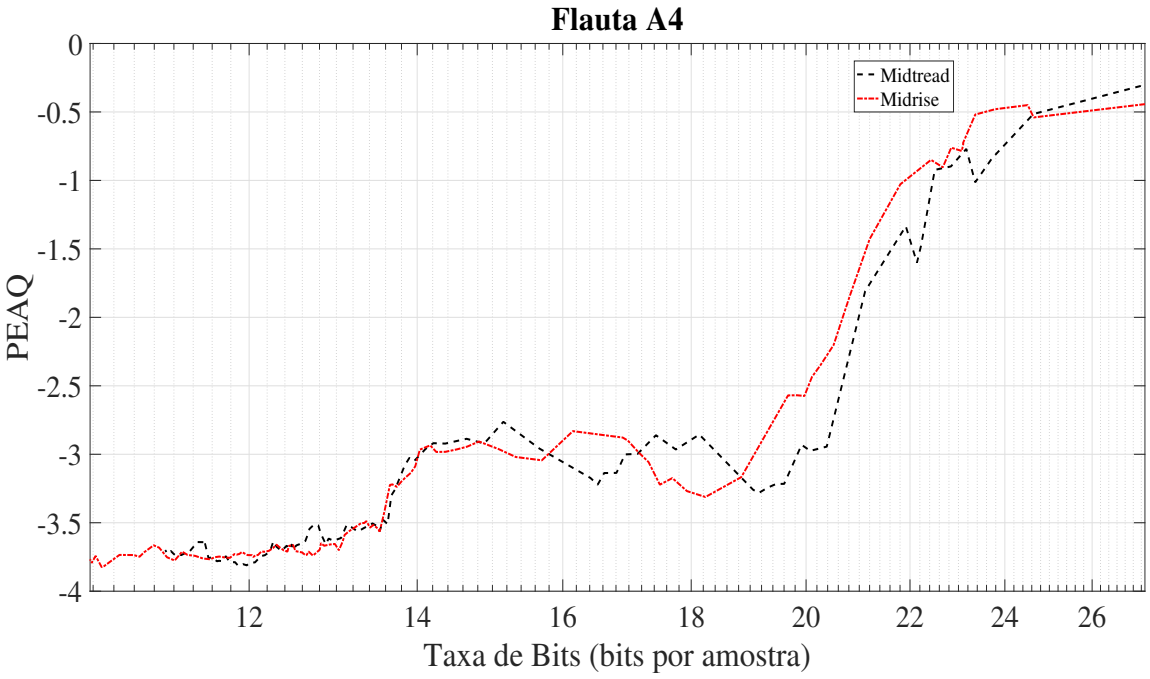
(b)

Figura 39 - Curvas de otimização do *violoncelo a4* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread*

Na Figura 42 estão apresentadas as curvas de taxa-distorção e taxa-PEAQ relativas a *bateria a*, utilizando 0 dB de margem. O erro de quantização não apresenta redução significativa com taxa de bits superiores a 7,9 bits/amostra para ambos os quantizadores. A avaliação psicoacústica por PEAQ indica tendência de crescimento aproximadamente



(a)

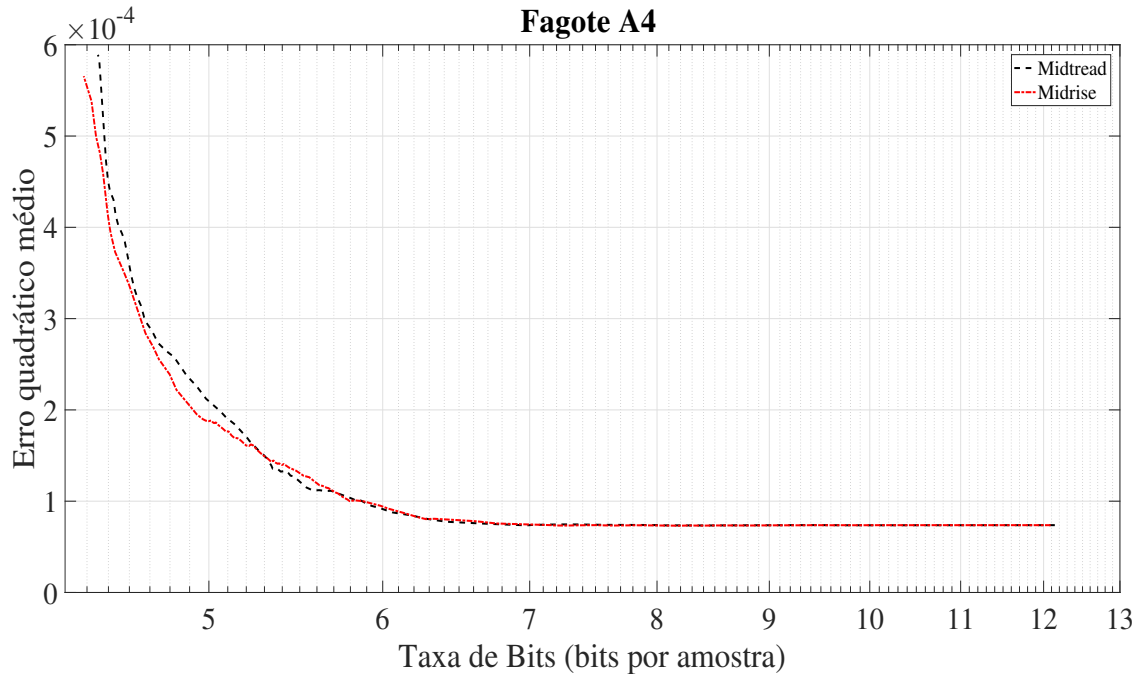


(b)

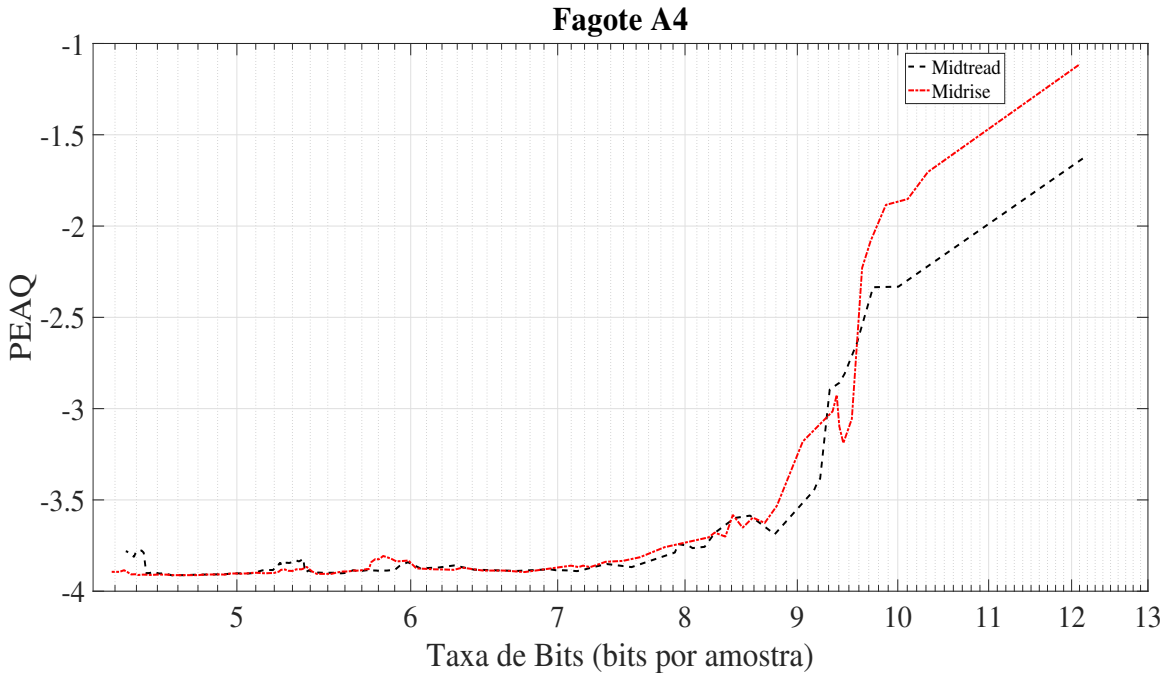
Figura 40 - Curvas de otimização do *flauta a4* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread*

linear. Conforme o aumenta da taxa de bits, o PEAQ só passa a ser -1, com taxas aproximadamente superiores a 11,2 bits/amostra. Os dois quantizadores apresentam desempenho equivalentes.

As curvas de taxa-distorção e taxa-PEAQ do sinal *bateria b*, com margem de 0



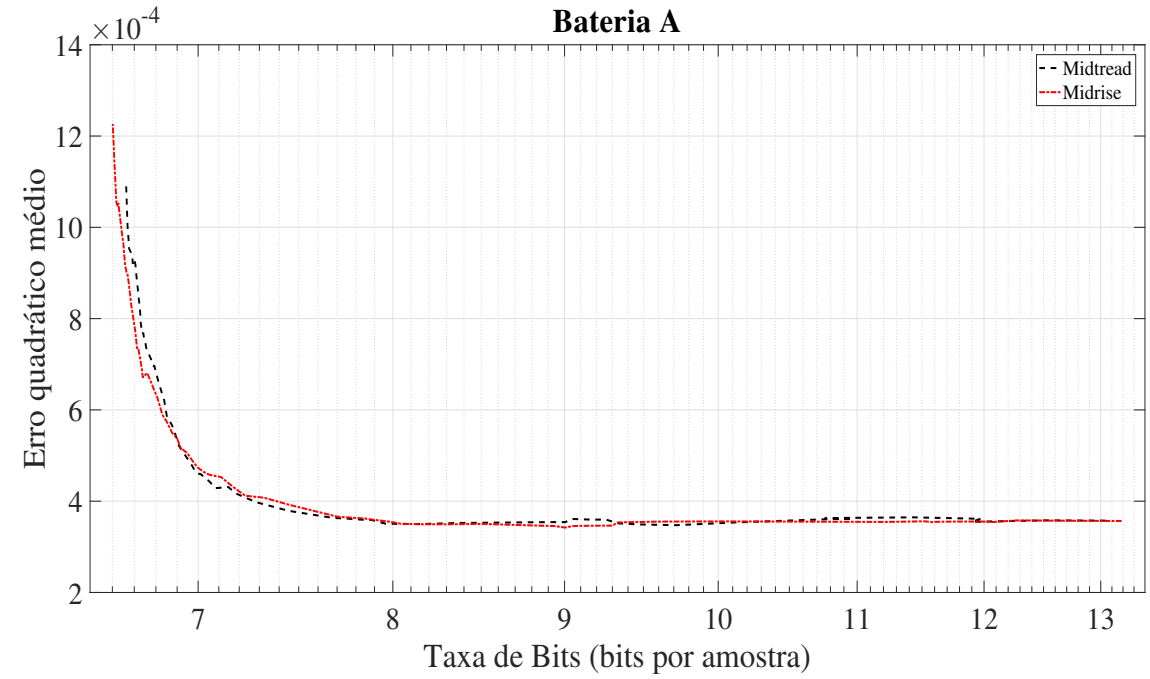
(a)



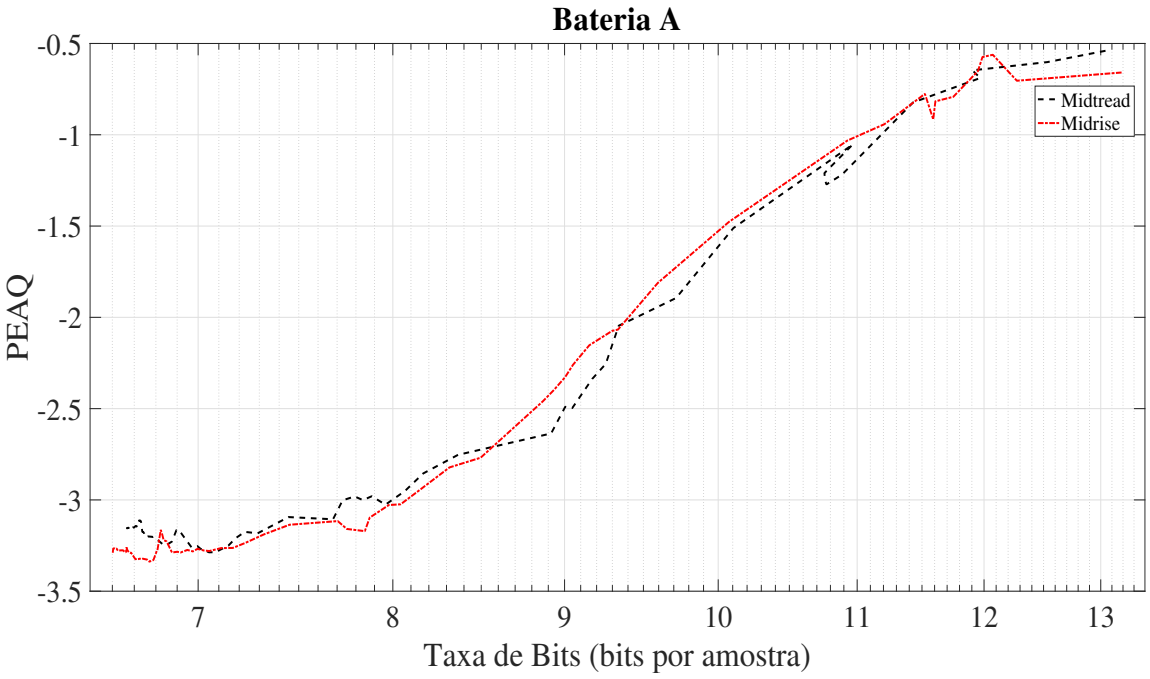
(b)

Figura 41 - Curvas de otimização do *fagote a4* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread*

dB, estão retratadas na Figura 43. A partir de aproximadamente 11 bits/amostra não há redução significativa de distorção. A avaliação psicoacústica por PEAQ apresenta rápido crescimento, alcançando a nota -1 com taxas superiores a 10,3 bits/amostra. Os quantizadores *midrise* e *midtread* são equivalentes, apresentando os melhores desempenhos nos



(a)

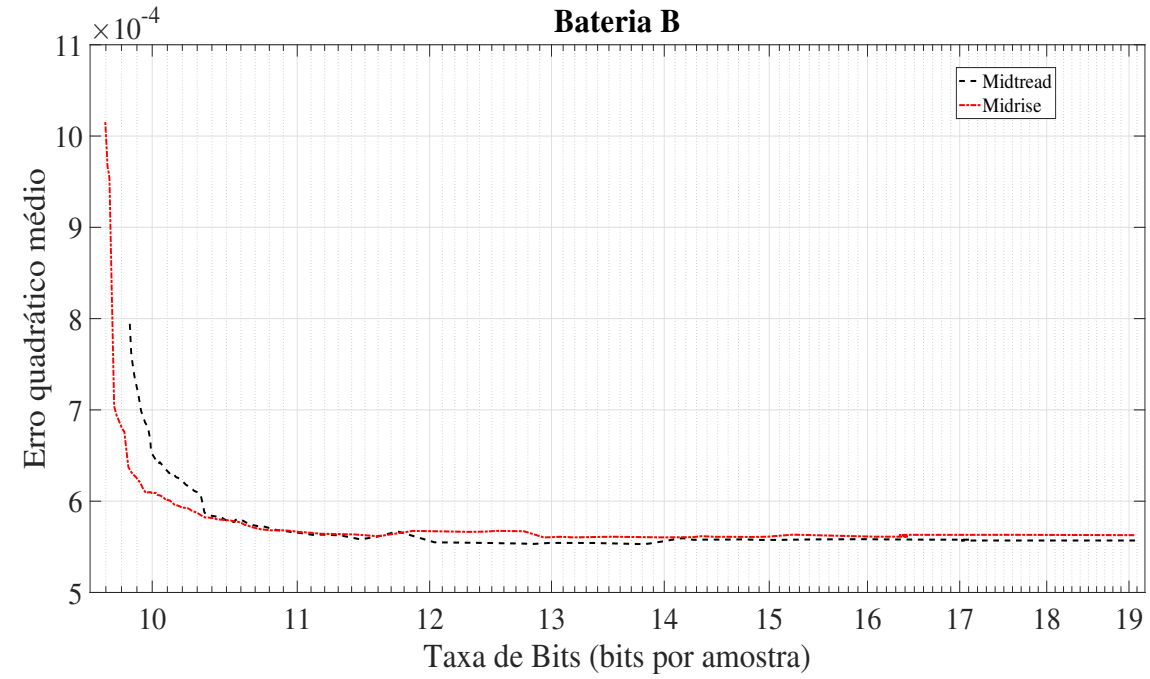


(b)

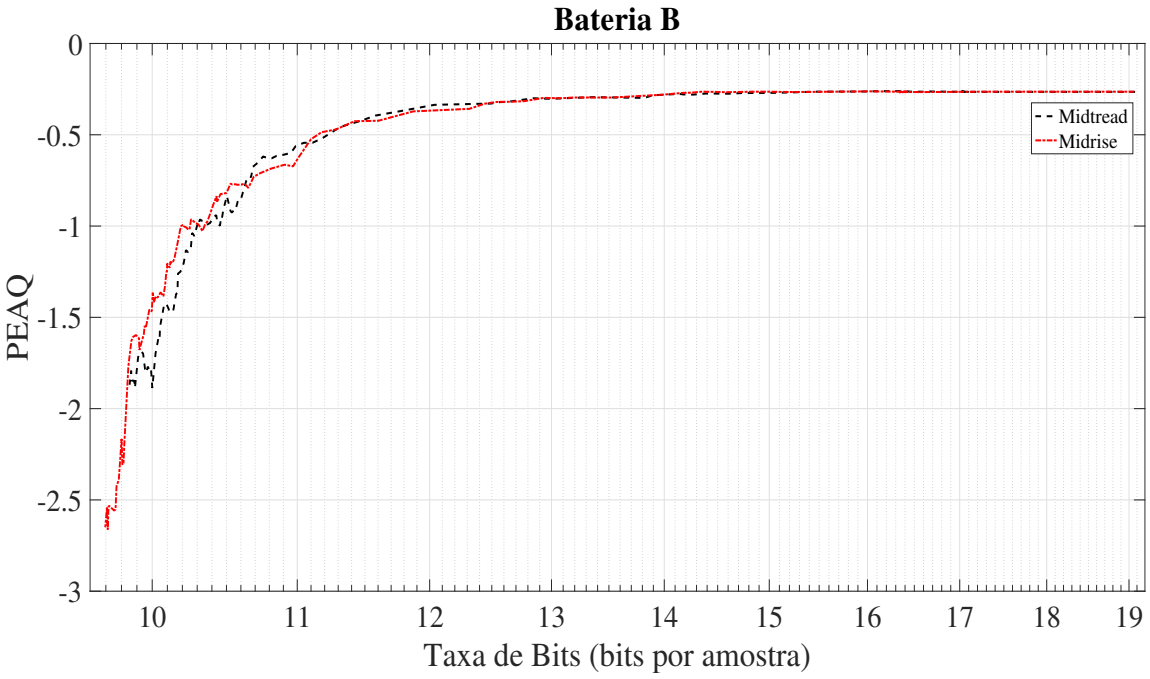
Figura 42 - Curvas de otimização do *bateria a* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread*

testes para o sinal em questão.

Da Figura 44 até a Figura 47 são referentes as curvas de taxa-distorção e taxa-PEAQ de todos os sinais utilizados no trabalho. Por elas é possível ver que as características dos sons de cada instrumento influenciam no desempenho da alocação ótima de



(a)



(b)

Figura 43 - Curvas de otimização do *bateria b* onde: (a) Taxa-Distorção dos quantizadores *midrise* e *midtread*, (b) Taxa-PEAQ dos quantizadores *midrise* e *midtread*

bits. O sinal *piano a3* apresentou a menor distorção, com a melhor avaliação por PEAQ, utilizando para isso a menor taxa de bits para ambos os quantizadores. Os sinais *bateria a*, *bateria b* e *flauta a4* apresentaram bons desempenhos em suas codificações. No sinal *flauta a4* os quantizadores conseguem alcançar boas avaliações psicoacústicas mas a custo

de grandes taxas de bits. Com o sinal *fagote a4*, os quantizadores não conseguem obter bons resultados na avaliação por PEAQ.

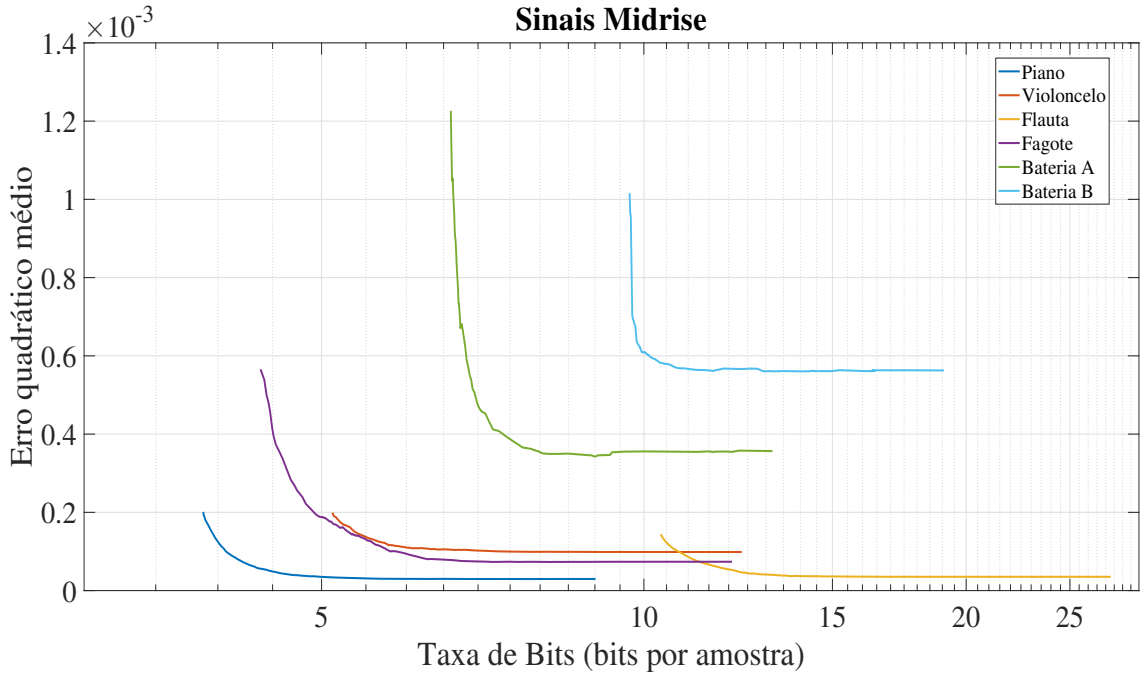


Figura 44 - Curvas de otimização de todos os áudios da Taxa-Distorção do quantizador *midrase*

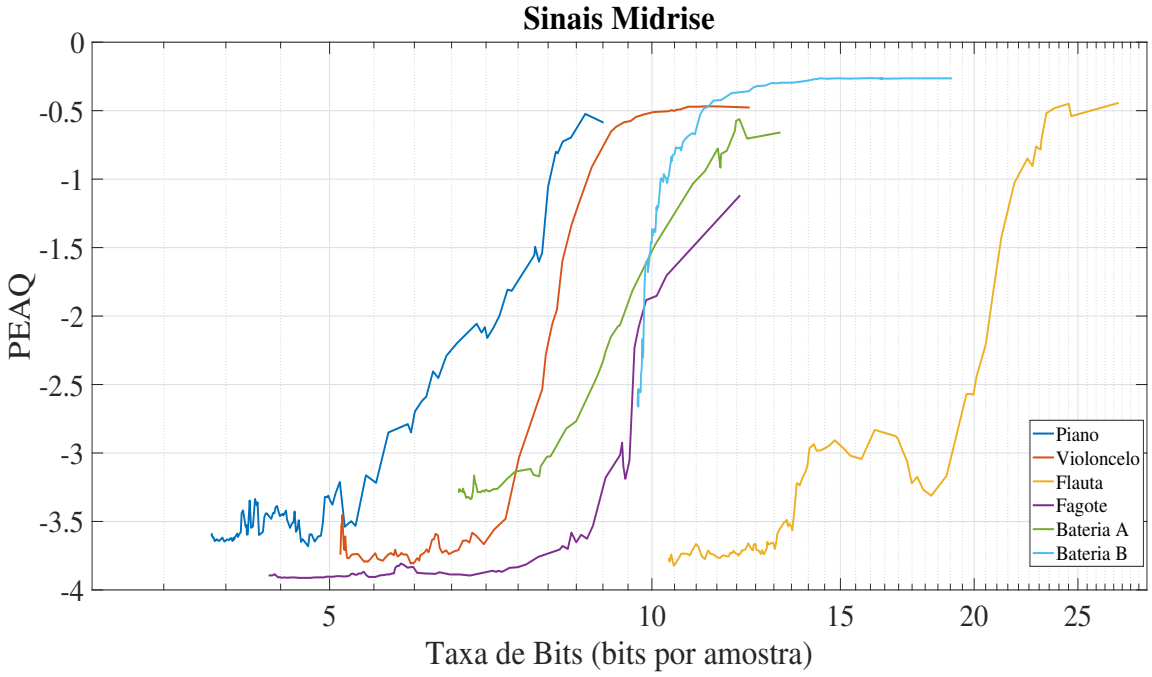


Figura 45 - Curvas de otimização de todos os áudios da Taxa-PEAQ do quantizador *midrase*

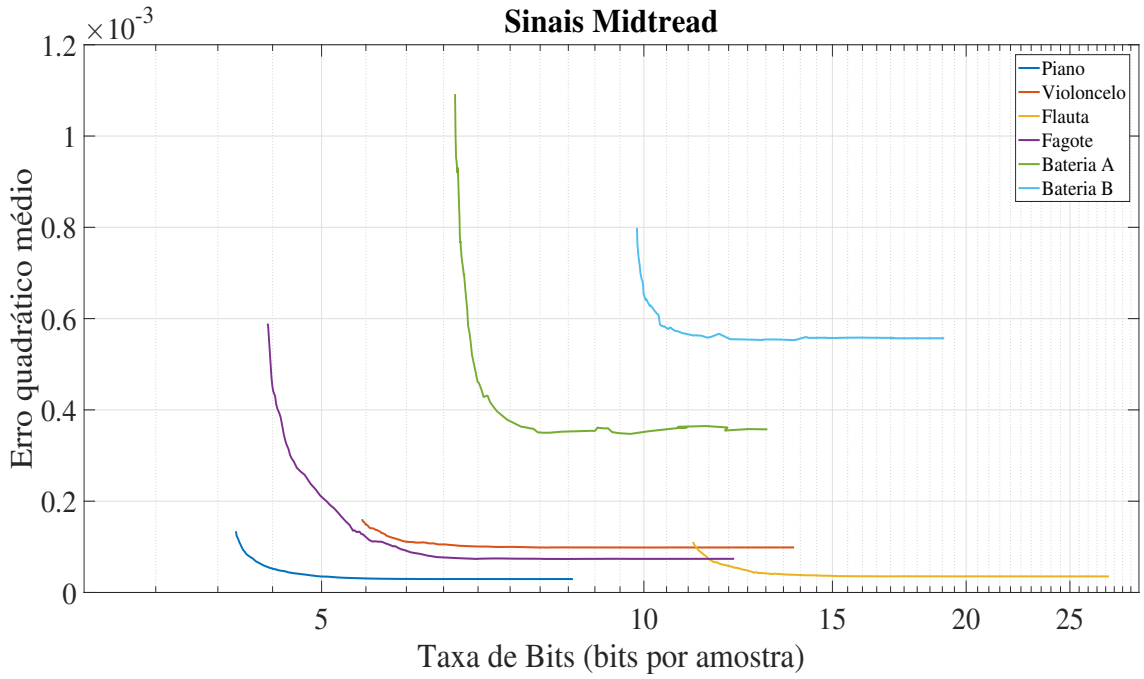


Figura 46 - Curvas de otimização de todos os áudios da Taxa-Distorção do quantizador midtread

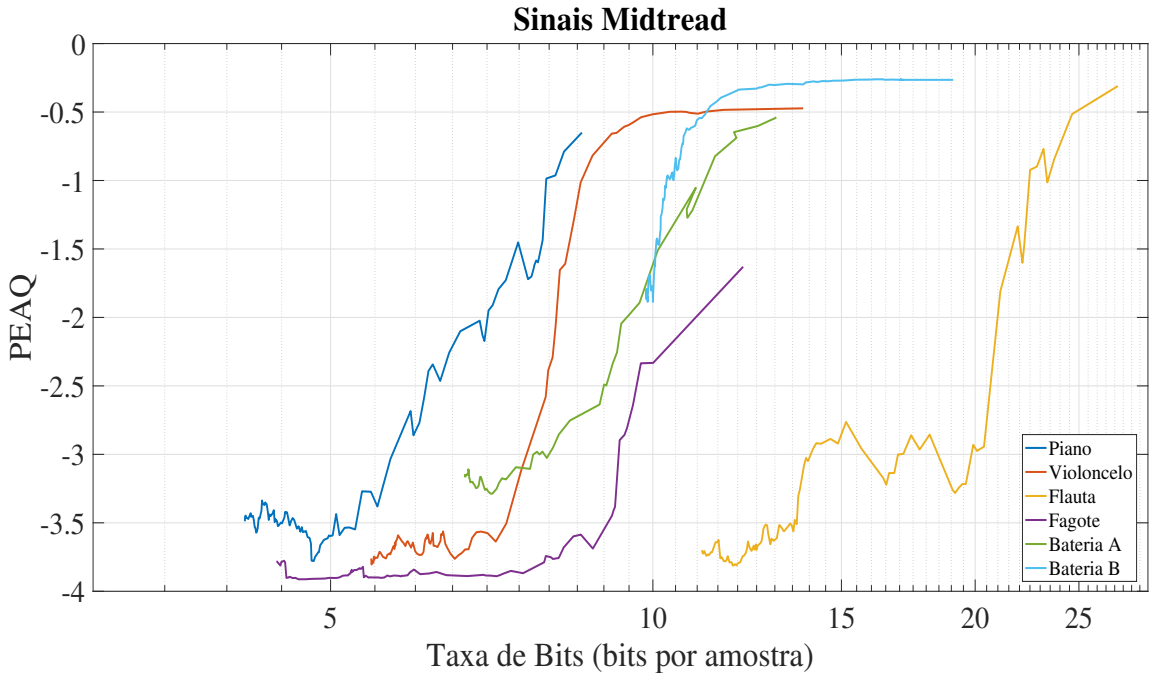


Figura 47 - Curvas de otimização de todos os áudios da Taxa-PEAQ do quantizador midtread

A Tabela 3 estão os valores de taxas de bits por segundo (kbps) com suas faixas de qualidade superior para cada sinal.

O padrão MPEG-1 Layer 3 (*MP3*) alcança qualidade de CD com taxas de bits de

Tabela 3 Valores de taxa de bits por segundo com boa qualidade

	Midrise		Midtread	
	Taxa (kbps)	PEAQ	Taxa (kbps)	PEAQ
Piano A3	352,8	-1	352,8	-1
Violoncelo A4	374,85	-1	374,85	-1
Fagote A4	432,18	-2	485,10	-2
Flauta A4	926,10	-1	970,20	-1
Bateria A	485,10	-1	493,92	-1
Bateria B	454,23	-1	454,23	-1

192 kbps. Observa-se que o codificador proposto apresenta taxas maiores que 192 kbps para que o sinal seja avaliado pelo PEAQ como não perturbador. O MP3 é um padrão desenvolvido e aperfeiçoado por vários pesquisadores no decorrer do tempo. Uma das diferenças entre o padrão proposto e o MP3, é que o último utiliza o código de Huffman e codificação por entropia. Esse tipo de codificação permitiria melhorar o desempenho do codificador proposto.

Outra maneira de aperfeiçoar o método proposto é utilizar alguma forma de alocação ótima de *bits* que leve em consideração aspectos psicoacústicos do sinal em análise.

5 CONCLUSÕES

Essa dissertação propôs um esquema de decomposição de sinais de áudio com o auxílio do algoritmo Matching Pursuit, baseado em Dicionários Redundantes de Exponenciais Complexas e utilizando o princípio de relevância psicoacústica dos componentes do sinal para a obtenção da representação de sinais de forma compacta. O trabalho foi inspirado em [2], mas, ao invés da função de ponderação, utilizamos diretamente a curva psicoacústica obtida pelo modelo MPEG-1 (camada I) [13]. À medida em 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 estas faixas são descartadas.

A codificação de áudio é utilizada para obter representações digitais do sinal com um número mínimo de bits. O quantizador escalar uniforme é o mais simples. Pode ser do tipo *"midrise"*, onde se garante a representação de valores nulos, com a contrapartida de resulta em nível a menos de representação que o quantizador *"midtread"*, levando portanto a um maior nível de ruído de quantização.

A troca entre a fidelidade da fonte e taxa de codificação é exatamente o compromisso da taxa-distorção. A alocação ótima de *bits* é realizada através da otimização da taxa-distorção, realizada pelo método do multiplicador de Lagrange.

A aferição dos resultados obtidos foi realizada por meio de uma ferramenta de avaliação perceptiva de qualidade de áudio: o PEAQ *(Perceptual Evaluation of Audio Quality)*.

Nota-se pela Tabela 4 que existe uma tendência geral de melhora dos resultados perceptivos quando se usa um dicionário com maior redundância, resultado do maior número de possibilidades de representação do sinal. Outro ponto observado é que, quanto maior a margem subtraída do limiar global psicoacústico, melhor é o resultado perceptivo da decomposição. O aumento da margem acarreta no acréscimo do número de iterações necessário para se alcançar o critério de parada, incrementando assim o número de elementos que descrevem o sinal.

Não foi observada uma margem psicoacústica global que possa responder adequadamente a todos os sinais analisados com o DEC. Cada sinal necessita de uma margem correspondente às suas características.

Tabela 4 Valores do PEAQ dos diferentes sinais decompostos.

	Piano A3		Violoncelo A4		Fagote A4		Flauta A4		Bateria A		Bateria B	
Redund. (dB)	4	8	4	8	4	8	4	8	4	8	4	8
0	-0,549	-0,613	-1,576	-1,353	-1,534	-0,743	-2,743	-1,33	-0,385	-0,36	-0,265	0,277
1	-0,62	-0,583	-1,381	-1,203	-1,429	-0,762	-2,618	-1.168	-0,309	0,27	-0,216	-0,226
2	-0,405	-0,552	-1,249	-1,029	-0,991	-0,659	-2,295	-0,876	-0,264	-0,231	-0,158	-0,146
3	-0,33	-0,404	-1,066	-0,839	-1,144	-0,578	-1,933	-0,705	-0,169	-0,162	-0,125	-0,13
4	-0,2	-0,28	-0,854	-0,683	-0,744	-0,465	-1,52	-0,521	-0,147	-0,138	-0,099	-0,104
5	-0,107	-0,191	-0,678	-0,509	-0,662	-0,549	-1,295	-0,4	-0,116	-0,088	-0,089	-0,064
6	-0,092	-0,092	-0,591	-0,385	-0,396	-0,367	-0,955	-0,249	-0,081	-0,061	-0,062	-0,058
7	-0,032	-0,05	-0,474	-0,271	-0,479	-0,236	-0,68	-0,159	-0,03	-0,037	-0,037	-0,037
8	0,002	-0,006	-0,342	-0,174	-0,279	-0,102	-0,558	-0,084	-0,027	-0,021	-0,001	-0,006
9	0,033	0,04	-0,196	-0,051	-0,207	-0,037	-0,418	-0,002	-0,016	-0,019	0,01	0,009
10	0,067	0,082	-0,062	0,02	-0,107	-0,038	-0,223	0,049	-0,001	0,011	0,002	0,017

Com o critério de parada psicoacústico é possível reduzir o número de iterações necessárias para a decomposição de cada sinal, tornando o algoritmo mais rápido e aumentando o grau de compressão obtido, como é possível observar na Tabela 5.

Tabela 5 Valores do número médio de iterações dos sinais decompostos.

	Piano A3		Violoncelo A4		Fagote A4		Flauta A4		Bateria A		Bateria B	
Redun. (dB)	4	8	4	8	4	8	4	8	4	8	4	8
0	52,08	51,95	51,98	52,23	34,37	32,29	66,99	61,10	105,51	104,10	144,42	141,13
1	56,03	55,67	55,70	55,71	39,10	34,38	72,78	66,07	113,48	112,11	155,23	152,03
2	59,81	59,36	59,59	59,35	44,78	36,78	80,50	71,25	122,02	119,85	165,87	162,27
3	64,05	63,39	63,73	63,07	50,63	39,97	90,89	77,02	131,23	128,74	176,54	173,05
4	68,53	67,69	68,16	67,26	55,95	43,96	101,94	82,98	140,88	138,55	188,81	184,65
5	73,15	72,07	73,22	71,83	61,54	48,06	117,15	89,82	151,69	149,22	202,69	197,39
6	77,95	76,53	78,53	76,61	69,30	52,19	120,97	96,83	164,21	161,20	217,27	211,38
7	83,04	81,14	84,29	81,58	78,99	56,70	126,19	104,26	177,93	174,50	233,86	227,63
8	87,95	86,03	90,86	87,25	68,74	62,76	138,87	112,80	192,59	189,03	251,38	245,44
9	93,78	91,39	97,91	93,40	73,91	68,70	151,54	121,69	211,53	206,17	271,43	264,58
10	100,93	97,66	105,71	100,39	79,63	75,60	167,67	132,02	232,09	225,45	292,86	286,09

A quantização escalar uniforme alcançou uma boa qualidade psicoacústica, mas o número de *bits* necessários para tal fim ainda é bastante elevado conforme ilustra a Tabela 6. Melhores desempenhos podem ser atingidos com o desenvolvimento de codificadores que levem em consideração as informações psicoacústicas de cada sinal na etapa de alocação de bits. Nesta dissertação, as informações psicoacústicas foram consideradas na etapa de decomposição, mas não na alocação de bits.

5.1 Trabalhos Futuros

Propõem-se para trabalhos futuros:

- Implementação do princípio de relevância psicoacústica para múltiplos dicionários, com a utilização de um critério de escolha do melhor dicionário;

Tabela 6 Valores de taxa de bits por segundo

	Midrise		Midtread	
	Taxa (kbps)	PEAQ	Taxa (kbps)	PEAQ
Piano A3	352,8	-1	352,8	-1
Violoncelo A4	374,85	-1	374,85	-1
Fagote A4	432,18	-2	432,18	-2
Flauta A4	926,1	-1	970,2	-1
Bateria A	485,10	-1	493,92	-1
Bateria B	454,23	-1	454,23	-1

- Testar o princípio de relevância psicoacústica para outros algoritmos de decomposição, tais como Basis Pursuit, Orthogonal Matching Pursuit e Iterative Hard Threshold .;
- Implementar a codificação considerando o princípio de relevância psicoacústica.

REFERÊNCIAS

[1] MALLAT, S.; ZHANG, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, v. 41, n. 12, p. 3397–3415, Dez. 1993.

[2] VERMA, T. S.; MENG, T. H. Sinusoidal modeling using frame-based perceptually weighted matching pursuits. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 1999.

[3] PETROVSKY, A.; HERASIMOVICH, V.; PETROVSKY, A. Scalable parametric audio coder using sparse approximation with frame-to-frame perceptually optimized wavelet packet based dictionary. In: IEEE. *AES 138th Convention*. [S.l.], 2015.

[4] TOUMI, I.; DERRIEN, O. Sparse decomposition of audio signals using a perceptual measure of distortion. application to lossy audio coding. In: NTNU. *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*. Trondheim, Norway, 2015. p. 1809–1812.

[5] RENCKER, L.; WANG, W.; PLUMBLEY, M. D. Multivariate iterative hard thresholding for sparse decomposition with flexible sparsity patterns. In: IEEE. *Signal Processing Conference (EUSIPCO), 2017 25th European*. [S.l.], 2017. p. 2156–2160.

[6] TOUMI, I.; DERRIEN, O. Sparse decomposition of audio signals using a perceptual measure of distortion. application to lossy audio coding. In: *18th International Conference on Digital Audio Effects*. [S.l.: s.n.], 2015.

[7] DERRIEN, O.; NECCIARF, T.; BALAZS, P. A quasi-orthogonal, invertible, and perceptually relevant time-frequency transform for audio coding. In: IEEE. *Signal Processing Conference (EUSIPCO), 2015 23rd European*. [S.l.], 2015. p. 799–803.

[8] BACH, J.-H.; KOLLMEIER, B.; ANEMÜLLER, J. Matching pursuit analysis of auditory receptive fields' spectro-temporal properties. *Frontiers in systems neuroscience*, Frontiers, v. 11, p. 4, 2017.

[9] PETROVSKY, A.; HERASIMOVICH, V.; PETROVSKY, A. Scalable parametric audio coder using sparse approximation with frame-to-frame perceptually optimized

wavelet packet based dictionary. In: AUDIO ENGINEERING SOCIETY. *Audio Engineering Society Convention 138*. [S.l.], 2015.

[10] CHARDON, G.; NECCIARI, T.; BALAZS, P. Perceptual matching pursuit with gabor dictionaries and time-frequency masking. In: *ICASSP*. [S.l.: s.n.], 2014. p. 3102–3106.

[11] ZHANG, X. et al. An audio feature extraction scheme based on spectral decomposition. In: IEEE. *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*. [S.l.], 2014. p. 730–733.

[12] LIUNI, M. et al. Automatic adaptation of the time-frequency resolution for sound analysis and re-synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 21, n. 5, p. 959–970, 2013.

[13] ISO, M. I. S. 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.

[14] BOSI, M.; GOLDBERG, R. E. *Introduction to digital audio coding and standards*. [S.l.]: Springer Science & Business Media, 2012.

[15] LIN, Y.; ABDULLA, W. H. et al. *Audio Watermark - A Comprehensive Foundation Using MATLAB*. [S.l.]: Springer, 2015.

[16] FASTL, H.; ZWICKER, E. *Psychoacoustics: Facts and Models*. [S.l.]: Springer, 2007.

[17] SPANIAS, A.; PAINTER, T.; ATTI, V. *Audio signal processing and coding*. [S.l.]: John Wiley & Sons, 2006.

[18] HU, X.; HE, G.; ZHOU, X. Peaq-based psychoacoustic model for perceptual audio coder. In: IEEE. *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*. [S.l.], 2006. v. 3, p. 5–pp.

[19] RECOMMENDATION ITU-R BS.1387-1 - Method for objective measurements of perceived audio quality. In: . [S.l.: s.n.], 2002.

[20] CAMPBELL, D.; JONES, E.; GLAVIN, M. Audio quality assessment techniques—a review, and recent developments. *Signal Processing*, Elsevier, v. 89, n. 8, p. 1489–1500, 2009.

[21] BEERENDS, J. G. et al. Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: Psychoacoustic model. *J. Audio Eng. Soc*, v. 50, n. 10, p. 765–778, 2002. Disponível em: <http://www.aes.org/e-lib/browse.cfm?elib=11062>.

[22] THIEDE, T. et al. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, Audio Engineering Society, v. 48, n. 1/2, p. 3–29, 2000.

[23] MALLAT, S. *A Wavelet Tour of Signal Processing*. 2. ed. California, USA: Academic Press, 1998.

[24] DAVIS, G.; MALLAT, S.; ZHANG, Z. Adaptive time-frequency approximations with matching pursuits. In: *Wavelet Analysis and Its Applications*. [S.l.]: Elsevier, 1994. v. 5, p. 271–293.

[25] CHEN, S. S.; DONOHO, D. L.; SAUNDERS, M. A. Atomic decomposition by basis pursuit. *SIAM review*, SIAM, v. 43, n. 1, p. 129–159, 2001.

[26] DINIZ, P. S.; SILVA, E. A. da; NETTO, S. L. *Processamento Digital de Sinais-: Projeto e Análise de Sistemas*. [S.l.]: Bookman Editora, 2014.

[27] OPPENHEIM, A. V.; SCHAFER, R. W.; YUEN, C. Digital signal processing. *IEEE Transactions on Systems, Man, and Cybernetics*, IEEE, v. 8, n. 2, p. 146–146, 1978.

[28] LOVISOLO, L. *Representações Coerentes de Sinais Elétricos*. Dissertação (Dissertação de Mestrado) — PEE/COPPE, UFRJ, Rio de Janeiro, Brasil, 2001.

[29] TCHEOU, M. P. *Compressão de Sinais Usando Decomposições Atômicas*. Tese (Tese de Doutorado) — PEE/COPPE,UFRJ, Rio de Janeiro, Brasil, 2011.

[30] FERRANDO, S. E.; KOLASA, L. A.; KOVAČEVIĆ, N. Algorithm 820: a flexible implementation of matching pursuit for gabor functions on the interval. *ACM Transactions on Mathematical Software (TOMS)*, ACM, v. 28, n. 3, p. 337–353, 2002.

[31] SHAOBING, C.; DONOHO, D. Basis pursuit. In: *28th Asilomar conf. Signals, Systems Computers*. [S.l.: s.n.], 1994.

[32] DAUDET, L. Sparse and structured decompositions of signals with molecular matching pursuit. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 14, n. 5, p. 1890–1902, Set. 2006.

[33] BISCAINHO, L. W. P. Random signals and stochastic processes. In: *Academic Press Library in Signal Processing*. [S.l.]: Elsevier, 2014. v. 1, p. 113–168.

[34] SAYOOD, K. *Introduction to data compression*. [S.l.]: Morgan Kaufmann, 2017.

[35] ORTEGA, A.; RAMCHANDRAN, K. Rate-distortion methods for image and video compression. *IEEE Signal processing magazine*, IEEE, v. 15, n. 6, p. 23–50, 1998.

[36] RECOMMENDATION, B. 1387: Method for objective measurements of perceived audio quality. *International Telecommunication Union, Geneva, Switzerland*, 2001.

[37] NOGUEIRA JUNIOR, V. S.; TCHEOU, M. P.; ÁVILA, F. R. Decomposição psicoacústica de sinais de Áudio com base em dicionários redundantes e exponenciais complexas. *VII Simpósio de Processamento de Sinais, UFABC*, 2017.