UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

AVALIAÇÃO OBJETIVA DE QUALIDADE DE SINAIS DE ÁUDIO E VOZ

Autor: **Jayme Garcia Arnal Barbedo**

Orientador: **Amauri Lopes**

Tese submetida à Faculdade de Engenharia Elétrica e Computação, Departamento de Comunicações, como parte dos requisitos para obtenção do título de doutor.

Banca Examinadora
Prof. Dr. Amauri Lopes (presidente)
Prof. Dr. Fernando Runstein
Prof. Dr. Josué Vieira Filho
Prof. Dr. Dalton Soares Arantes
Prof. Dr. Fábio Violaro
Prof. Dr. João Batista Tadanobu Yabu-uti
Prof. Dr. Luís Geraldo Pedroso Meloni

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

B232a

Barbedo, Jayme Garcia Arnal
Avaliação objetiva de qualidade de sinais de áudio e voz / Jayme Garcia Arnal Barbedo.--Campinas, SP: [s.n.], 2004.

Orientador: Amauri Lopes.
Tese (Doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e da Computação.

1. Áudio. 2. Codificador de Voz. 3. Processamento de sinais. I. Lopes, Amauri. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e da Computação. III. Título.

Este projeto foi financiado pela FAPESP

Processo nº 01/04144-0

AGRADECIMENTOS

Agradeço e dedico este trabalho:

A meus pais, que sempre me incentivaram a ir adiante com meus sonhos e que nunca mediram esforços para que eles se realizassem.

À minha esposa, que tornou os desafios muito mais fáceis de serem vencidos, estando sempre a meu lado, nos bons e maus momentos.

A meus colegas, os quais forneceram um apoio inestimável nos momentos de maior necessidade.

A meu orientador, cuja atuação superou a melhor das expectativas, e ao qual devo boa parte dos créditos pelo sucesso deste trabalho.

A todos aqueles que, de uma maneira ou de outra, contribuíram para a realização deste trabalho.

RESUMO

Este trabalho aborda o problema da avaliação objetiva de qualidade de sinais de áudio e voz. São apresentados os fundamentos teóricos envolvidos na modelagem matemática do ouvido humano, na codificação de sinais de áudio e na realização de avaliações subjetivas e objetivas de qualidade. Dois novos métodos de avaliação objetiva, denominados Medida Objetiva da Qualidade de Áudio (MOQA) e Avaliação Objetiva de Sinais de Voz (AOSV), são propostos. A validação desses métodos é realizada através de uma análise estatística detalhada e de sua comparação com as principais estratégias em uso na atualidade. Por fim, apresenta-se uma análise crítica dos avanços alcançados e da importância da pesquisa aqui realizada, além de se propor algumas diretrizes para futuros estudos.

ABSTRACT

This work deals with the problem of objective quality assessment of audio and speech signals. The theoretic foundation regarding the mathematical modeling of human ear, audio signal codification and subjective and objective assessment, is presented. Two new methods for objective assessment, named MOQA and AOSV, are proposed. The validation of such methods is performed using a detailed statistical analysis and comparing the results with the most important strategies currently in use. Finally, a critical analysis of the advances achieved during the research is presented, along with the proposal of some directives for future studies.

ÍNDICE

CAPÍTULO 1
INTRODUÇÃO

A transmissão e o armazenamento digital de sinais de áudio vêm sendo, cada vez mais, baseados em algoritmos para compressão de dados, os quais são adaptados a diversas propriedades do sistema auditivo humano, destacando-se os efeitos de mascaramento. Tais algoritmos não buscam necessariamente a minimização de distorções, e sim sua manipulação adequada, de maneira que elas sejam minimamente percebidas pelo usuário do sistema. Assim, a qualidade desses assim chamados codificadores perceptuais não pode mais ser medida por métodos tradicionais baseados no valor global de distorção, como a relação sinal-ruído (SNR) e a distorção harmônica total (THD), os quais não são capazes de modelar essas novas características. Em certos casos, as estruturas ruidosas são tão eficientemente mascaradas pelo sinal que se tornam praticamente inaudíveis, ainda que o sinal apresente uma relação sinal-ruído tão baixa quanto 13 dB.

Dessa forma, faz-se necessário o uso de testes subjetivos de audição para a realização de avaliações confiáveis da qualidade de codecs perceptuais. No entanto, tais testes são dispendiosos, seja em termos de tempo ou de custos. Portanto, é altamente desejável o desenvolvimento de medidas objetivas capazes de substituir, de maneira eficiente, os testes subjetivos. Desde o final dos anos 70, alguns métodos foram propostos, mas, com o surgimento dos primeiros codecs perceptuais (MPEG – *Moving Picture Expert Group* e Dolby) no final dos anos 80, tais medidas se tornaram obsoletas. Então, em 1994, a ITU-R (*International Telecommunication Union - Radiocommunication*) fez uma chamada aberta de propostas, a fim de estabelecer um padrão para a medição objetiva da qualidade de áudio. Seis métodos foram apresentados, nenhum deles alcançando o mínimo desempenho desejável. Por esse motivo, todos os esforços se concentraram no desenvolvimento de um método conjunto que pudesse resolver os problemas inerentes às propostas iniciais, surgindo assim o método PEAQ (*Perceptual Evaluation of Audio Quality*). Tal método apresentou um desempenho muito superior aos demais e, apesar de ainda não satisfazer todos os tipos de condições encontrados na prática, deu origem à recomendação ITU-R BS.1387-7. Os conceitos utilizados nesse método serviram de ponto de partida para o desenvolvimento do presente trabalho.

A pesquisa aqui realizada teve como principal objetivo desenvolver novas técnicas e estratégias capazes de superar algumas das principais limitações encontradas nas estratégias tradicionalmente adotadas. A combinação desses avanços resultou em dois novos métodos: Medida Objetiva da Qualidade de Áudio (MOQA) e Avaliação Objetiva de Sinais de Voz (AOSV). A palavra “áudio”, no contexto deste trabalho, é usada para designar sinais de música e voz em banda larga (20 Hz a 20 kHz). O termo “voz”, por sua vez, é usado para designar sinais de voz na faixa de telefonia (300 a 3400 Hz).

O Capítulo 2 apresenta uma breve descrição da fisiologia do ouvido humano e os principais conceitos e modelos provenientes de suas características anatômicas. A

fundamentação teórica contida nesse capítulo é essencial para o entendimento adequado dos capítulos subseqüentes.

O Capítulo 3 apresenta os principais codificadores de áudio, cujas características de implementação são um fator preponderante no desenvolvimento de qualquer medida objetiva de avaliação de áudio. As características perceptuais de tais dispositivos são ali destacadas.

O Capítulo 4 apresenta uma breve descrição dos procedimentos e tipos de testes adotados na realização das medidas subjetivas. As medidas objetivas visam a estimação confiável dos resultados obtidos em tais testes.

O Capítulo 5 faz uma descrição sucinta de alguns dos principais métodos objetivos para avaliação de áudio propostos nos últimos anos, com o objetivo de fornecer uma breve introdução histórica para o trabalho aqui realizado.

O Capítulo 6 descreve as novas estratégias e abordagens desenvolvidas em diferentes estágios do processamento por que passam os sinais de áudio ao serem objetivamente avaliados. A reunião de tais inovações em uma nova estratégia de avaliação é aqui denominada MOQA.

O Capítulo 7 apresenta a validação do método MOQA, onde os testes e respectivos resultados são discutidos e comparados com o desempenho de outros métodos de avaliação de áudio. As conclusões sintetizam as principais características do método e sua aplicabilidade, além de apontarem aspectos passíveis de aperfeiçoamento em pesquisas futuras.

Por fim, o Capítulo 8 apresenta uma série de novas técnicas desenvolvidas com vistas ao aperfeiçoamento do método MOQV (Medida Objetiva de Qualidade de Voz), desenvolvido em pesquisas anteriores. Tais inovações resultaram num novo método, melhor e mais robusto, denominado AOSV.

Lista de Artigos Publicados e Submetidos

Barbedo, J. G. A.; Lopes, A. *Strategies to Increase the Applicability of Methods for Objective Assessment of Audio Quality*, 116th AES Convention, preprint 6080, Berlin, May 2004.

Barbedo, J. G. A.; Lopes, A.; Simões, F. O.; Runstein, F. *Objective Measure of Speech Quality in Channels with Variable Delay*, Revista Telecomunicações, vol. 6, n. 2, pp.19 - 24, December 2003.

Barbedo, J. G. A.; Lopes, A. *A New Method for Objective Assessment of Áudio Quality*, Anais do XX Simpósio Brasileiro de Telecomunicações, Rio de Janeiro, Outubro de 2003.

Barbedo, J. G. A.; Lopes, A. *Innovations on the Objective Assessment of Audio Quality*, Anais da VII Convenção Nacional da AES, São Paulo, Maio de 2003.

Barbedo, J. G. A.; Lopes, A. *Proposal and Validation of an Objective Method for Quality Assessment of Speech Codecs and Communication Systems*, Revista Tecnologia, Fortaleza, Vol. 23, No. 1, pp. 96-112, dezembro de 2002.

Barbedo, J. G. A.; Ribeiro, M.V.; Von Zuben, F.J.; Lopes, A.; Romano, J.M.T. *Application of Kohonen Self-Organizing Maps to Improve the Performance of Objective Methods for Speech Quality Assessment*, Proceedings of the XI European Signal Processing Conference (EUSIPCO2002), Vol. I, pp. 519-522, Toulouse, France, September 2002.

Barbedo, J. G. A.; Ribeiro, M. V.; Lopes, A.; Romano, J. M. T. *Estimation of the Subjective Quality of Speech Signals using the Kohonen Self-Organizing Maps*, Proceedings of the IV International Telecommunication Symposium (ITS), Natal, Brazil, pp. 834-839, September 2002.

Lopes, A.; Romano, J. M. T.; Ribeiro, M. V.; Barbedo, J. G. A.; Lima, C. *Método FL-PMC (Fourier Lapped - Perceptron Multicamadas) para a Estimação de Qualidade de Voz*, patente: privilégio de inovação n. 0204932-5, Método FL-PMC, depósito 06 de novembro de 2002.

Lopes, A.; Barbedo, J. G. A. *Medida Objetiva de Avaliação de Áudio*, pedido de patente em andamento.

Barbedo, J. G. A.; Lopes, A. *A New Cognitive Model for Objective Assessment of Audio Quality*, submitted to the Journal of AES.

Ribeiro, M. V.; Barbedo, J. G. A.; Romano, J. M. T.; Lopes, A. *Fourier-Lapped-Multilayer Perceptron (FLMLP) Method for Speech Quality Assessment*, submitted to the Special Issue on Anthropomorphic Processing of Audio and Speech.

Barbedo, J. G. A.; Lopes, A. *Uma Nova Estratégia para a Estimação Objetiva da Qualidade de Sinais de Áudio*, submetido à revista IEEE Latino.

Barbedo, J. G. A.; Lopes, A. *A New Method for Objective Assessment of Speech Quality*, aceito para publicação na Revista da Sociedade Brasileira de Telecomunicações.

Barbedo, J. G. A.; Lopes, A. *On the Vectorization of Decimation Filterbanks*, to be submitted to a Journal.

CAPÍTULO 2
PRINCÍPIOS FUNDAMENTAIS DA AUDIÇÃO HUMANA

O pré-processamento realizado pelo ouvido sobre o sinal acústico é uma atividade objetiva, pois envolve a transformação do sinal acústico que chega ao ouvido externo em impulsos elétricos nos feixes de neurônios distribuídos ao longo da cóclea. O processamento subjetivo será realizado pelas funções superiores do córtex cerebral, baseado neste sinal condensado gerado pelo ouvido [1]. Esses processos ocorrem de maneira bastante homogênea de uma pessoa para outra. Por isso, um bom modelo dos processos envolvidos na percepção auditiva pode ser aplicado para um vasto número de pessoas. Este Capítulo fornece uma visão geral a respeito das estruturas anatômicas envolvidas na audição humana, os principais conceitos provenientes do estudo do comportamento auditivo e as principais abordagens sugeridas para a modelagem matemática do ouvido.

2.1. FISIOLOGIA DO OUVIDO HUMANO

A seguir, será apresentada uma breve descrição dos elementos fisiológicos envolvidos no processo da audição humana. Algumas dessas estruturas podem ser visualizadas na Figura 2.1, onde é mostrado o corte longitudinal do ouvido.

2.1.1. Ouvido Externo

O ouvido externo e a cabeça são componentes de um complexo sistema de recepção acústica, o qual faz a ligação entre o tímpano e o campo sonoro externo. O ouvido externo protege o tímpano de danos mecânicos e melhora o acoplamento entre este e o campo sonoro, além de contribuir substancialmente para a direcionalidade do sistema, especialmente para altas freqüências. As funções acústicas de vários componentes do ouvido externo têm sido elucidadas através do desenvolvimento de modelos físicos, cujas características são projetadas de modo a reproduzir fielmente seus correspondentes fisiológicos [2].

Assim, as propriedades acústicas do pavilhão auricular podem afetar a propagação do som no espaço e, conseqüentemente, afetar o sinal acústico que chega ao conduto auditivo, agindo como atenuador ou amplificador de sons de determinada freqüência (efeito sombra e efeito ilusório). A concha tem uma ressonância de 5 kHz, enquanto que o restante irregular da orelha produz ressonâncias e anti-ressonâncias. A propriedade de ressonância faz com que os sons externos com freqüência entre 2 e 5 kHz sofram um ganho de 10 a 15 dB. Este aumento permite a detecção e reconhecimento de sons de pequena energia e alta freqüência, como os fricativos (ss, sch) [3,4].

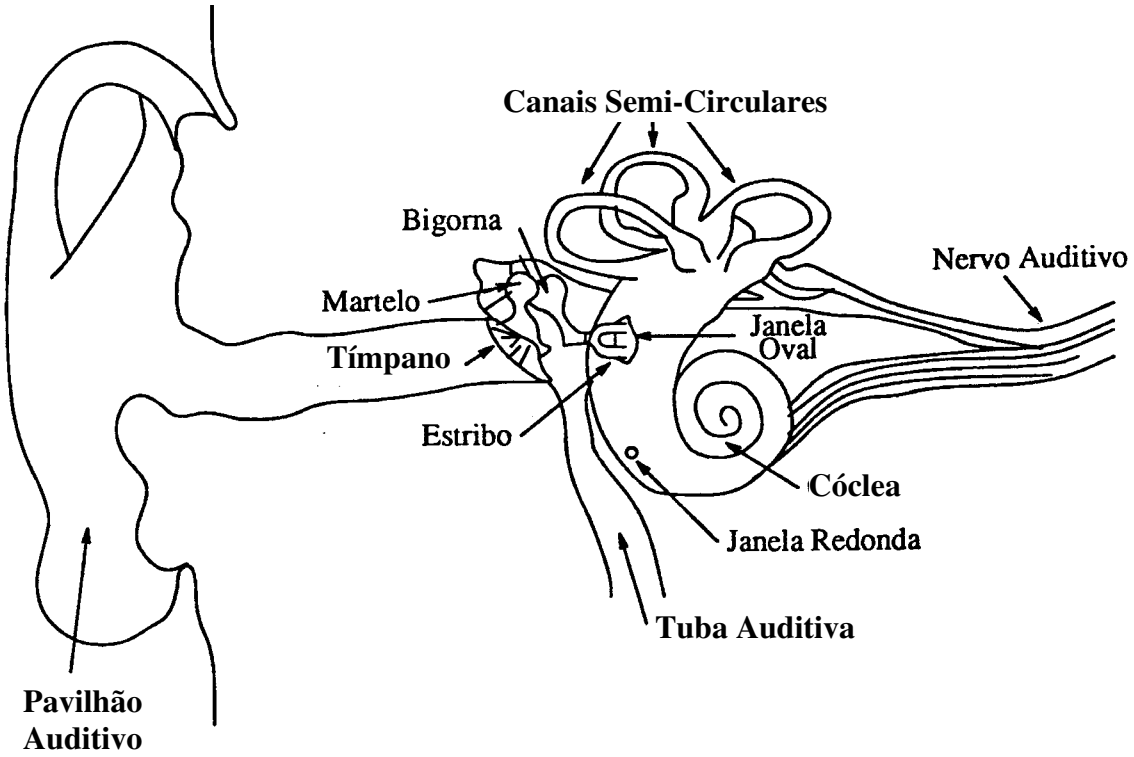


Figura 2.1 - Corte longitudinal do ouvido, com suas principais estruturas.

2.1.2. Ouvido Médio

A principal função do ouvido médio é melhorar a transmissão sonora entre o ouvido externo e o ouvido interno, pois ele tem a capacidade de reduzir a reflexão que ocorre quando uma onda sonora incide em uma superfície fluida. Ou seja, pode-se considerá-lo um transformador de impedância que reduz a alta impedância do fluido coclear (ver seção 2.1.3) para um valor semelhante ao do ar.

As principais estruturas que compõem o ouvido médio são a membrana timpânica, a cadeia ossicular com os respectivos ligamentos e músculos e a cavidade preenchida com ar na qual estão localizados os ossículos. O tímpano é o limite entre o ouvido externo e o médio; as janelas oval e redonda da cóclea são os limites entre o ouvido médio e o interno. Os ossículos (*martelo*, *bigorna* e *estribo*) fazem a transmissão do som que é recebido pelo tímpano diretamente à janela oval, além de protegerem a janela redonda do som, o qual chega a ela com menor amplitude. Se isto não ocorresse, o som chegaria ao mesmo tempo às duas janelas com a mesma magnitude e na mesma fase sonora, pois a distância entre elas é muito pequena, e isto não promoveria a movimentação do fluido do ouvido interno, tornando a audição impossível. A tuba auditiva permite que o ar penetre no ouvido médio através da faringe, igualando a pressão em ambos os lados do tímpano.

O processo de transformação do sinal acústico nas ondas do líquido coclear é chamado de função de transferência do ouvido médio. Ele é equivalente a uma filtragem passa-baixas com corte em 5 kHz, com uma sobre-elevação na faixa entre 2.000 e 5.000 Hz e um pico em torno de 3.500 Hz [5,6]. Como essa filtragem não altera o espectro de forma significativa, ela é, em geral, desconsiderada para sinais com faixa até 5.000 Hz.

2.1.3. Ouvido Interno

Como visto, a onda acústica que chega ao pavilhão auditivo é transformada em movimento das estruturas ósseas que compõem o ouvido médio (*martelo*, *bigorna* e *estribo*). Os ossos do ouvido médio estimulam a *cóclea* através da *janela oval*, fazendo com que seu líquido interno se movimente. A cóclea pode ser modelada como um tubo de aproximadamente 30 mm com duas câmaras separadas por uma estrutura chamada *membrana basilar*, como pode ser visto na Figura 2.2. Na extremidade oposta à janela oval, existe um orifício sobre a membrana basilar que comunica essas duas câmaras, chamado de *helicotrema*. A membrana basilar apresenta uma resistência (mecânica) que varia ao longo de sua extensão: próximo à janela oval ela é mais fina e tensa, ressoando em freqüências mais altas, enquanto no seu final (ápice), ela é espessa e flácida, ressoando então para freqüências mais baixas. As ondas geradas pelo estribo, em resposta a um sinal senoidal, viajam ao longo da cóclea, fazendo vibrar a membrana basilar na mesma freqüência do sinal de entrada [7].

Sobre a membrana basilar existem ainda duas estruturas: as *fibras basilares* e o *órgão de Corti*. As fibras basilares são cerca de 20.000 pequenas estruturas delgadas com comprimentos que variam ao longo da membrana, sendo mais curtas junto à janela oval e mais longas no ápice da cóclea [8]. Sua vibração estimula as *células ciliadas*, que compõem o órgão de Corti, o qual é responsável pelo sensoriamento dos estímulos sonoros recebidos pelo ouvido. As células ciliadas, por sua vez, transformam o movimento das fibras basilares em impulsos nervosos, os quais são então transmitidos pelo nervo coclear para a região específica do córtex cerebral.

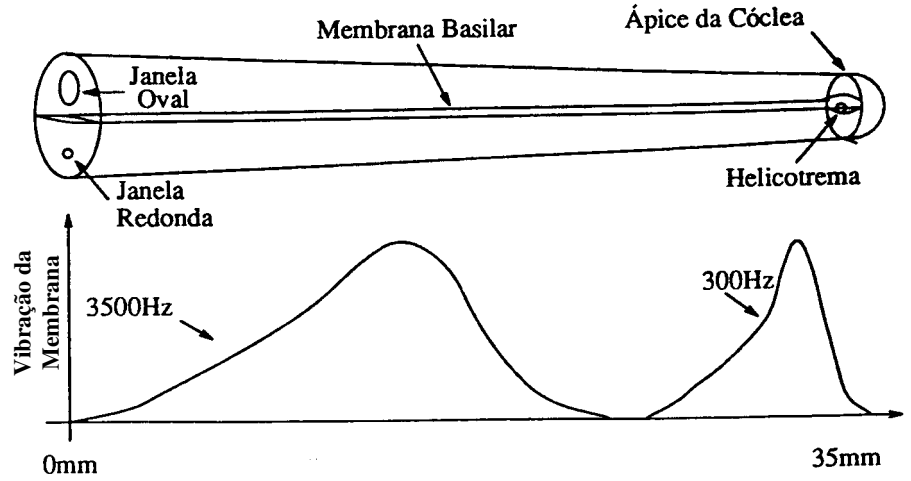


Figura 2.2 - Cóclea, membrana basilar e vibração da membrana em duas diferentes freqüências.

Cada ponto da membrana basilar é mais sensível a uma determinada freqüência, chamada de *freqüência característica*. Para um ponto específico da membrana basilar, a curva de resposta à freqüência de vibração presente na janela oval é equivalente à de um filtro passa-faixa com fator de qualidade aproximadamente constante, resultando numa melhor resolução nas baixas freqüências. Assim, as fibras basilares localizadas na região de altas freqüências características respondem em uma maior faixa de freqüências do que as fibras na região de baixas freqüências características.

Um comportamento similar é obtido ao se traçar a curva de resposta ao longo da membrana basilar para um tom numa freqüência específica, como se pode ver na Figura 2.2. Para cada freqüência, há um ponto da membrana basilar em que a vibração é máxima. A posição desse ponto, medida a partir do helicotrema, é aproximadamente proporcional ao logaritmo da freqüência do som. Ao redor desse ponto haverá uma faixa, de cerca de 1,5 mm, onde a vibração estará presente, atenuando-se conforme se afasta do ponto. Tal faixa determina o conceito das bandas críticas, como será visto mais adiante.

2.2. FENÔMENOS AUDITIVOS

A seguir, serão apresentados alguns dos principais conceitos provenientes do estudo do comportamento auditivo humano.

2.2.1. Limiar Absoluto de Audibilidade em Silêncio

O limiar absoluto de audibilidade em silêncio é o menor nível, em função da freqüência, para o qual um tom se torna audível [9]. Este limiar pode ser aproximado pela expressão analítica dada por

$$\lim = 3{,}64 \cdot f^{-0,8} - 6{,}5 \cdot e^{-0,6\cdot(f-3,3)^2} + 10^{-3} \cdot f^4, \qquad (2.1)$$

onde f é a freqüência em kHz. O limiar é dado em dB_{SPL} (ver Apêndice A). Esta aproximação é usada em quase todos os métodos perceptuais. Consiste de três termos: o primeiro descreve a freqüência de corte para as baixas freqüências; o segundo descreve o aumento de sensibilidade do ouvido para a faixa de freqüências em torno de 3 kHz; o último descreve a freqüência de corte para as altas freqüências. O primeiro termo, ou pelo menos parte dele, é interpretado como um resultado do *ruído interno* (causado por atividade muscular, fluxo de sangue etc.), ao passo que os dois últimos termos são interpretados como a característica de transferência de ouvido médio para o interno. Conseqüentemente, em modelos perceptuais, esta equação é freqüentemente dividida em duas partes: uma chamada *função de ruído interno* e outra chamada *função de transferência do ouvido médio*.

2.2.2. Bandas Críticas

Alguns dos fenômenos de mascaramento, como aqueles que serão apresentados na seção 2.2.3, podem ser explicados em termos de faixas de freqüências conhecidas como *bandas críticas*, as quais foram determinadas através de experimentos psico-acústicos [10]. Uma banda crítica define uma faixa em torno de uma freqüência central, a qual está associada a um ponto da membrana basilar, de modo que a cada ponto é possível definir uma banda crítica. Quando dois sinais se situam dentro de uma banda crítica, o de maior energia poderá dominar a percepção e mascarar o outro estímulo sonoro. Portanto, dependendo dos níveis, dois tons distintos só serão distinguidos um do outro quando estiverem em bandas críticas diferentes. Este é o fenômeno responsável pelo mascaramento simultâneo, como será visto mais adiante. A resolução para a distinção entre uma freqüência e outra varia de 100 Hz, nas freqüências mais baixas, a mais de 6.000 Hz, nas freqüências mais altas. Além disso, sinais com uma largura de banda suficiente para extrapolar os limites de uma banda crítica sempre proporcionarão uma intensidade

perceptual maior que aqueles cujas componentes espectrais estejam limitadas a uma única banda crítica, ainda que o nível de pressão sonora e a freqüência central sejam equivalentes.

É importante ressaltar que as bandas críticas podem ser definidas em torno de qualquer freqüência central. A largura de faixa das bandas críticas corresponde a um espaçamento uniforme de 1,5 mm ao longo da membrana basilar, o que corresponde a aproximadamente 100 Hz para freqüências abaixo de 500 Hz e de aproximadamente 20% da freqüência central da banda para freqüências acima de 1000 Hz (em direção à janela oval) [11]. Portanto, a resposta de amplitude em freqüência, para cada banda crítica, pode ser modelada como a de um filtro passa-faixas com largura de faixa crescente com a freqüência. Tais filtros possuem cortes acentuados: 65 dB/oitava para as bandas críticas em torno de 500 Hz e 100 dB/oitava em torno de 8 kHz.

Embora exista uma banda crítica ao redor de cada freqüência, convencionou-se (com algumas pequenas variações) a adoção dos valores mostrados na Tabela 2.1 [10]. Os valores apresentados na primeira coluna da tabela correspondem à escala Bark.

Tabela 2.1 - Bandas Críticas.

Banda Crítica	**Freqüência (Hz)**		
	Inferior	**Superior**	**Faixa**
0	0	100	100
1	100	200	100
2	200	300	100
3	300	400	100
4	400	510	110
5	510	630	120
6	630	770	140
7	770	920	150
8	920	1080	160
9	1080	1270	190
10	1270	1480	210
11	1480	1720	240
12	1720	2000	280

Banda Crítica	**Freqüência (Hz)**		
	Inferior	**Superior**	**Faixa**
13	2000	2320	320
14	2320	2700	380
15	2700	3150	450
16	3150	3700	550
17	3700	4400	700
18	4400	5300	900
19	5300	6400	1100
20	6400	7700	1300
21	7700	9500	1800
22	9500	12000	2500
23	12000	15500	3500
24	15500	22050	6550

2.2.2.1. Escalas Perceptuais de Freqüência

Devido ao fenômeno das bandas críticas, a resolução espectral da audição não é linear. Assim, escalas de freqüência lineares não modelam adequadamente a percepção de *pitch* de um ouvinte humano, nem é apropriada para explicar os efeitos auditivos no domínio da freqüência como, por exemplo, o mascaramento simultâneo. O termo *pitch* tem sido usado com dois sentidos diferentes: na área de processamento de voz, o termo é freqüentemente utilizado para designar a freqüência de oscilação da glote (vibração das cordas vocais); em psico-acústica, é usado como um atributo da sensação auditiva, segundo a definição encontrada na ANSI (*American National Standards Institute*), a qual estabelece que *pitch* é o atributo auditivo de acordo com o qual os sons podem ser ordenados, em uma escala de freqüência, de baixo a alto. Este é o sentido adotado neste trabalho. Os estudos da percepção humana do *pitch* são complexos. Mais informações podem ser encontradas em [12,13].

Uma representação logarítmica da freqüência é ligeiramente melhor que a linear, porém não é ainda satisfatória. Assim, faz-se necessário a derivação de uma escala que represente adequadamente a audição humana. Várias abordagens foram propostas:

- medição da localização da máxima deflexão da membrana basilar para tons puros a diferentes freqüências [14];

- medição da largura das bandas críticas observadas na percepção de sonoridade (escala Bark, unidade: Bark). Uma distância de 1 Bark corresponde à largura de uma banda crítica. A faixa de freqüência audível corresponde a, aproximadamente, 24 Bark [10].

- medição da relação do *pitch* subjetivamente percebido entre tons puros. Na verdade, esta seria a única escala de *pitch* propriamente dita, ainda que outras escalas auditivas de freqüência sejam freqüentemente referidas como escalas de *pitch*. A escala aqui utilizada é a *mel* [10]. A faixa de freqüência audível corresponde aproximadamente à faixa de zero a 2400 na escala *mel*.

- mínima diferença de freqüência perceptível, resultante da *escala de incremento espectral* (unidade: SPINC) proposta por Terhardt [15]; assim, 1 SPINC corresponde ao menor incremento de freqüência perceptível por um ouvinte humano médio. A faixa de freqüência audível corresponde aproximadamente à faixa entre 0 e 2000 na escala SPINC.

- área coberta pela curva de mascaramento produzida por um sinal de faixa estreita. Esta área, dividida pelo máximo da curva de mascaramento, resulta na largura de um filtro auditivo no caso de este ter um formato retangular. Este valor é chamado de *largura de banda retangular equivalente* e a escala de freqüência correspondente é chamada de *escala ERB* [16-20]. A faixa de freqüência audível corresponde aproximadamente à faixa entre 0 e 38 na escala ERB.

- função de Incremento Espectral: esta escala deriva do limiar de discriminação de freqüência [21]. Esta função raramente é usada em medidas perceptuais.

Segundo Zwicker [9], as escalas Bark e Mel são idênticas, exceto por um fator de normalização (1 Bark = 100 mel). Patterson e Moore postularam uma relação similar entre a escala ERB e a localização da máxima deflexão da membrana basilar [22]. A escala ERB usada por Moore [16] é ligeiramente diferente da escala Bark definida por Zwicker [10], especialmente nas baixas freqüências.

A principal vantagem de se usar uma escala de freqüência auditiva ao invés de uma simples escala de freqüência linear ou logarítmica é a facilidade que ela confere à modelagem dos efeitos no domínio da freqüência. A escala Bark tem sido a mais utilizada nos métodos objetivos de avaliação da qualidade de áudio por fornecer, para este tipo de aplicação, os resultados mais consistentes. Uma aproximação muito simples para a relação entre uma freqüência f e seu valor correspondente z na escala Bark [23] é dada por

$$f \approx 650 \cdot \sinh\left(\frac{z}{7}\right), \tag{2.2}$$

onde f é dado em kHz.

Esta aproximação foi projetada para uma faixa de freqüência relevante para a codificação de voz, isto é, ela só é válida para freqüências abaixo de 5 kHz. A fórmula pode ser facilmente invertida a fim de se ter a transformação inversa, que resulta em

$$z \approx 7 \cdot \operatorname{arcsinh}\left(\frac{f}{650}\right). \tag{2.3}$$

Outra expressão proposta, que relaciona f e z para toda a faixa audível de freqüências, é dada por

$$z = 13 \cdot \arctan(0{,}76 \cdot f) + 3{,}5 \cdot \arctan\left[\left(\frac{f}{7{,}5}\right)^2\right]. \qquad (2.4)$$

2.2.3. Mascaramento

As limitações do ouvido em termos das resoluções temporal, espectral e de amplitude, em combinação com uma faixa dinâmica também limitada, produzem um fenômeno chamado mascaramento. Quando dois tons estão suficientemente próximos um do outro, seja no domínio do tempo ou da freqüência, o tom mais fraco pode se tornar inaudível devido à presença do tom mais forte. Embora o fenômeno do mascaramento deva ser analisado no plano tempo-freqüência, é muito usual considerá-lo como dois efeitos separados, dependendo do domínio que se está considerando. Quando o mascaramento depende unicamente da localização no domínio da freqüência, isto é, os sinais mascarado e mascarador são apresentados no mesmo instante de tempo, tem-se o assim denominado mascaramento simultâneo. Se o mascaramento depende primariamente da localização no domínio do tempo, então ele é chamado de mascaramento temporal. Este último pode ser dividido em dois diferentes efeitos: mascaramento progressivo (ou pós-mascaramento) e mascaramento retrógrado (ou pré-mascaramento). No caso do mascaramento progressivo, os componentes do sinal são mascarados após o término do mascarador, e no caso do mascaramento retrógrado, os componentes são mascarados antes do início da execução do mascarador.

O nível de energia abaixo do qual um componente do sinal é mascarado por outros componentes é chamado de limiar de mascaramento. Além de depender da localização dos sinais mascarador e mascarado no plano tempo-freqüência, o limiar de mascaramento progressivo também depende da duração do mascarador [25].

O mascaramento simultâneo se deve basicamente à existência das bandas críticas. Quando dois tons se encontram em uma mesma banda crítica, o de maior amplitude dominará a percepção sonora. A Figura 2.3 mostra o padrão de mascaramento causado por tons em quatro freqüências distintas (0,25, 1, 4 e 8 kHz). As curvas mostram o nível mínimo que um sinal deve apresentar para se tornar audível (limiar de audibilidade), em função da freqüência. A curva tracejada representa o nível mínimo de audição de tons na ausência de um sinal mascarador. As curvas contínuas mostram o nível mínimo que um sinal deve apresentar para se tornar audível na presença de um sinal mascarador. Assim, ao se colocar um tom mascarador em 4 kHz, por exemplo, os tons nas freqüências próximas têm que apresentar o nível da curva contínua correspondente para que possam ser ouvidos simultaneamente com o mascarador. Note-se que o mascaramento abrange uma largura de faixa menor para baixas freqüências do que para altas freqüências, o que é uma conseqüência direta da definição das bandas críticas.

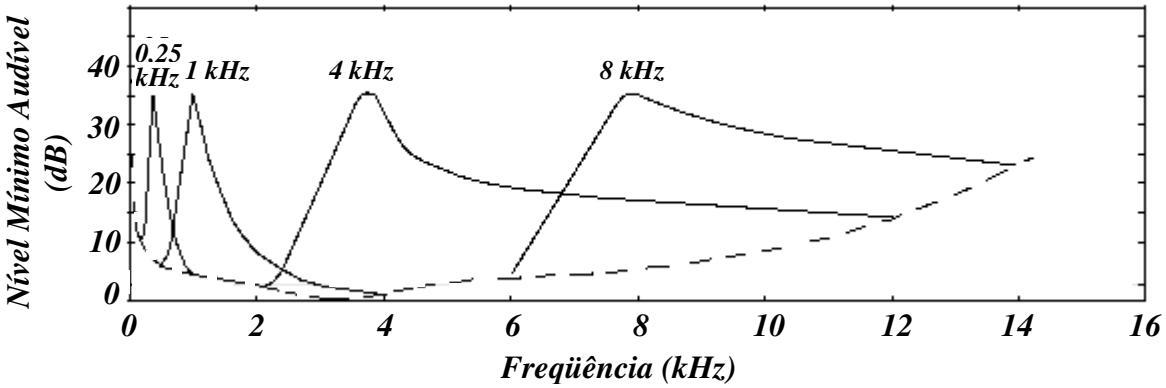


Figura 2.3 - Efeito de mascaramento simultâneo para 4 freqüências distintas.

Já o mascaramento progressivo ocorre devido ao fato de o ouvido humano necessitar de um certo tempo para se recuperar após um sinal de grande amplitude, uma vez que, nessas condições, os neurônios disparados ficam em um estado refratário que pode durar mais de 100 ms. A Figura 2.4 apresenta os resultados de um experimento onde executa-se um sinal de 60 dB e, logo após seu término, no instante zero, traça-se uma curva representando o nível necessário para um sinal de teste se tornar audível, em função do tempo de recuperação do ouvido. Uma aproximação para as curvas de mascaramento progressivo [26], descritas em [15] e exemplificadas através da Figura 2.4, é dada por

$$D(t,T_m) = 1{,}0 - \frac{1}{1{,}5708} \cdot \arctan\left[\frac{t}{13{,}2 \cdot T_m^{\,0{,}25}}\right], \tag{2.5}$$

onde T_m é a duração do mascarador em ms, e t é o tempo após o fim do mascarador em ms.

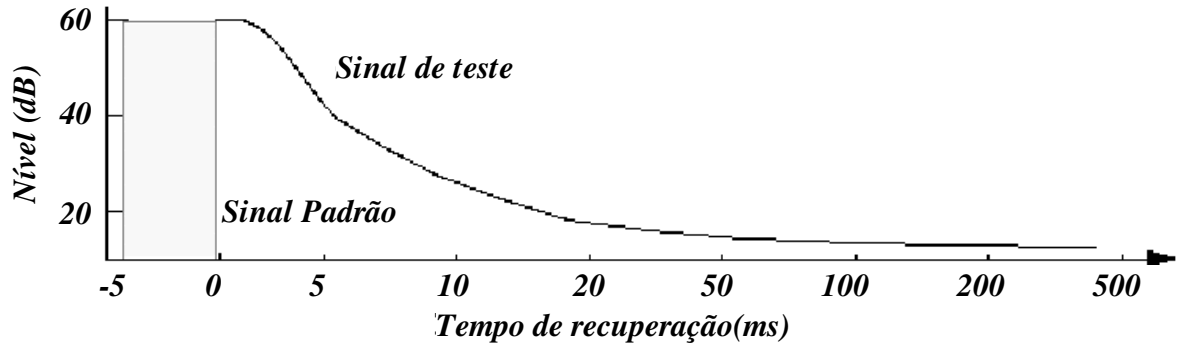


Figura 2.4 - Efeito do mascaramento progressivo [25].

Enquanto o mascaramento simultâneo e o mascaramento progressivo são de fácil entendimento, o mascaramento retrógrado é um fenômeno mais complicado, porque implica em um sinal de grande amplitude mascarar outro sinal antes de o primeiro estar realmente presente. Tal fenômeno é normalmente explicado pela suposição de que o sinal forte é processado mais rapidamente do que o sinal fraco, podendo, portanto, ultrapassar o sinal mascarado durante o processamento dos sinais, ou no nervo auditivo, ou posteriormente, nos níveis mais elevados do sistema auditivo [25].

Os fenômenos de mascaramento podem geralmente ser explicados modelando-se o sistema auditivo como um analisador de sinais de resolução espectral e temporal finita, determinada pelos filtros utilizados nesse analisador, e uma acuidade finita na distinção entre diferentes níveis, determinada pela resolução na representação dos níveis dos sinais. Um componente do sinal só pode ser detectado quando a diferença entre o sinal conjunto (mascarador mais mascarado) e o mascarador, em qualquer posição no plano tempo-freqüência, é maior do que a resolução de amplitude do sistema.

Utilizando esta abordagem na análise do mascaramento retrógrado em particular, tem-se que, como a resposta do filtro ao sinal mascarado levará um certo tempo até alcançar o limiar de audibilidade, o mascaramento já poderá ocorrer se a execução do mascarador se inicia durante esse período, conforme ilustrado na Figura 2.5. Como se pode ver na Figura 2.5, embora o sinal mascarado comece antes do mascarador, este já apresenta uma amplitude maior no instante em que o sinal mascarado excede o limiar absoluto de audibilidade. Esta diferença de amplitudes ocorre porque as respostas de tais filtros são dependentes da amplitude.

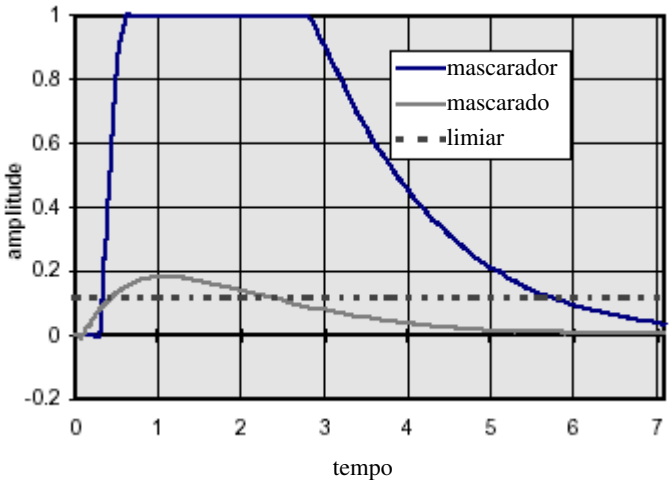


Fig. 2.5 - O mascaramento retrógrado sob o ponto de vista da abordagem utilizando filtros [25]

Mesmo que a diferença entre dois sinais nunca exceda o limiar de detecção, ela ainda pode ser detectada quando integrada sobre uma certa região do plano tempo-freqüência, ou seja, uma certa diferença, que não seria detectável isoladamente, poderá tornar-se perceptível se persistir por um certo intervalo de tempo, o que de fato ocorre no processamento auditivo humano. A inclusão da análise e integração no domínio do tempo pode reduzir a resolução do sistema, porém diminui consideravelmente o limiar de detecção, o que também ocorre na prática no processamento realizado pelo sistema auditivo. Quando se incorpora uma integração temporal, a dependência das funções de mascaramento em relação à resolução tempo-freqüência e ao tempo de integração torna-se mais complexa. Ainda que este modelo seja mais complexo que o anterior, ele explica a ocorrência de todos os tipos de mascaramento temporal e espectral, e funciona até mesmo sem quaisquer suposições sobre o limiar absoluto de audibilidade e a forma da curva dos filtros temporais. O uso de um ou outro modelo depende basicamente do desempenho alcançado em testes comparativos realizados no contexto em que se deseja utilizá-los.

Dentre as três categorias de mascaramento, o simultâneo tem sido mais freqüente e detalhadamente analisado. A medição dos mascaramentos temporais é mais difícil do que a

medição do mascaramento simultâneo, uma vez que estes são fortemente dependentes dos valores das freqüências dos sinais. Além disso, a determinação do mascaramento temporal requer uma boa resolução tanto no domínio do tempo quanto da freqüência. Porém, isto só é possível até certo ponto. No caso do mascaramento progressivo, tal limitação não é grave porque as constantes de tempo observadas são suficientemente grandes (cerca de 100 ms) para permitir sinais de teste com um espectro suficientemente compacto, sem introduzir demasiada incerteza na estrutura temporal. No caso do mascaramento retrógrado, as constantes de tempo observadas são tão pequenas (entre 1 e 20 ms) que este não pode ser medido de uma maneira confiável para sinais de banda estreita.

A seguir, serão apresentadas algumas das características inerentes aos diversos tipos de mascaramento.

2.2.3.1. Limiares de Mascaramento de Tons, Ruídos e Pulsos

Limiares de mascaramento são normalmente medidos como uma função do tempo ou da freqüência central do sinal mascarado, enquanto o nível e a freqüência central do sinal mascarador são mantidos constantes e são tomados como parâmetros. Tais medidas resultam em curvas de mascaramento como aquelas mostradas na Figura 2.6, onde o mascaramento é considerado em ambos os domínios, ou como aquelas mostradas na Figura 2.7, onde apenas o mascaramento simultâneo é analisado.

Na Figura 2.6, o paralelepípedo em destaque representa um ruído de faixa estreita, o qual faz o papel do mascarador. Tal ruído tem uma duração de 500 ms e uma largura de faixa de cerca de 1 Bark. A curva acima do ruído descreve a função de mascaramento relacionada a esse ruído, representando o nível necessário para que um sinal tonal se torne audível na presença do ruído mascarador. O eixo L_T apresenta este nível de forma relativa ao nível do mascarador em uma escala em dB. O eixo t permite observar a variação deste nível com o tempo. Aqui, observam-se os fenômenos do mascaramento retrógrado, de menor duração, e do mascaramento progressivo, mais longo. No eixo z, representando o domínio espectral perceptual, percebe-se o comportamento assimétrico da curva de mascaramento, já que a inclinação para as baixas freqüências é mais íngreme que a inclinação para as altas freqüências. É importante ter em conta que o formato da curva para as freqüências superiores é fortemente dependente do nível dos sinais. Para baixos níveis, ela é quase tão inclinada quanto nas baixas freqüências, enquanto se torna quase plana para níveis muito elevados do mascarador. Essas curvas de mascaramento podem ser aproximadas por exponenciais de dois lados quando representadas em função de uma escala de freqüência perceptual, como no caso da Figura 2.6.

A Figura 2.7 descreve o comportamento do limiar de audibilidade de um sinal na ausência de um mascarador (curva inferior) e na presença de tons mascaradores a 1 kHz e diferentes níveis.

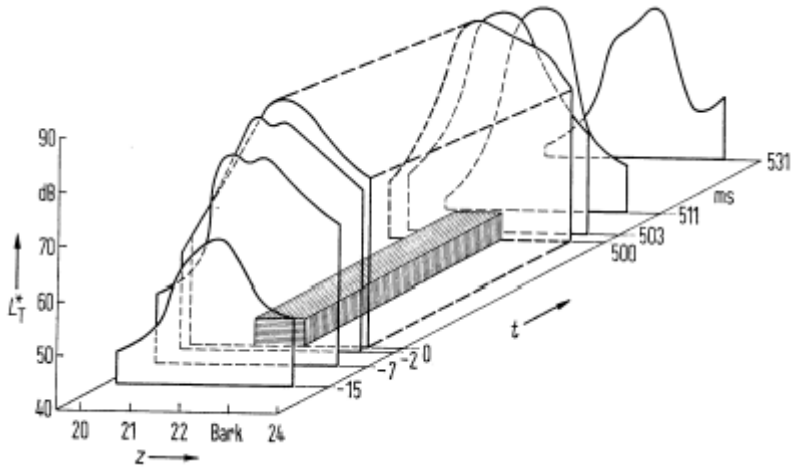


Fig. 2.6 - Padrão de mascaramento causado por ruído de banda estreita [27]

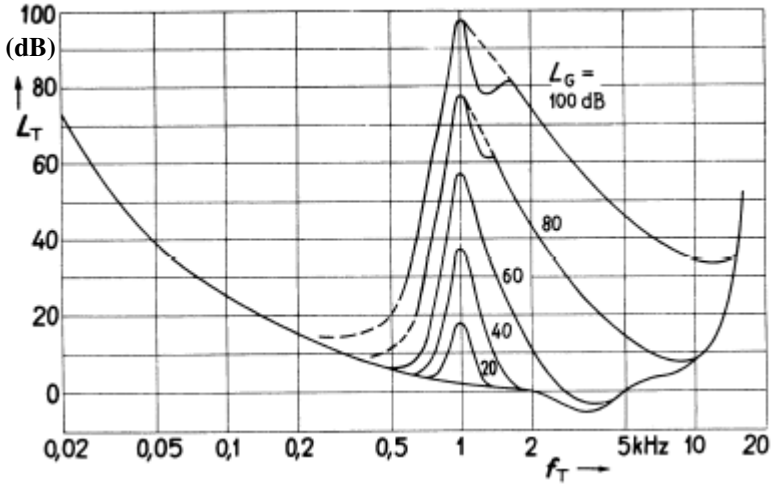


Fig. 2.7 - Representação do mascaramento simultâneo causado por tons a 1 kHz [28]

A diferença de nível entre um sinal e o máximo mascaramento por ele produzido (maior nível que outro sinal pode assumir e continuar sendo mascarado) é chamado *índice de mascaramento* [29,30]. Ele depende da freqüência central do sinal mascarador, mas assume-se que seja independente do seu nível. A representação linear do índice de mascaramento em função da freqüência é chamada fator de limiar [29].

2.2.4. Não-Linearidades da Audição

2.2.4.1. Escala Fônon e Contornos de Sonoridade Equivalente

A sonoridade subjetivamente percebida de um sinal de áudio depende não somente de seu nível de pressão sonora, mas também de outras características como, por exemplo, o fato de que a sensibilidade auditiva humana não é constante com a freqüência. A sonoridade de um sinal é dada na unidade fônon, e corresponde ao nível de pressão sonora em decibéis de um tom puro a 1 kHz que produz a mesma sonoridade percebida para o sinal medido.

A Figura 2.8 mostra uma série de curvas denominadas *contornos de sonoridade equivalente* [15]. Tais curvas são traçadas fixando-se uma certa sonoridade, em fônons,

para um tom, variando-se então a freqüência e verificando-se o nível de pressão sonora necessária para manter a sonoridade percebida constante. Os contornos de sonoridade equivalente, para os níveis de pressão sonora de 0 dB e 60 dB a 1 kHz, são de particular interesse: o primeiro define o limiar absoluto de audibilidade e o último representa um nível de audição intermediário, geralmente usado como referência. Na Figura 2.8, *L* representa o nível relativo dos sinais e L_N representa o valor em fônons das curvas.

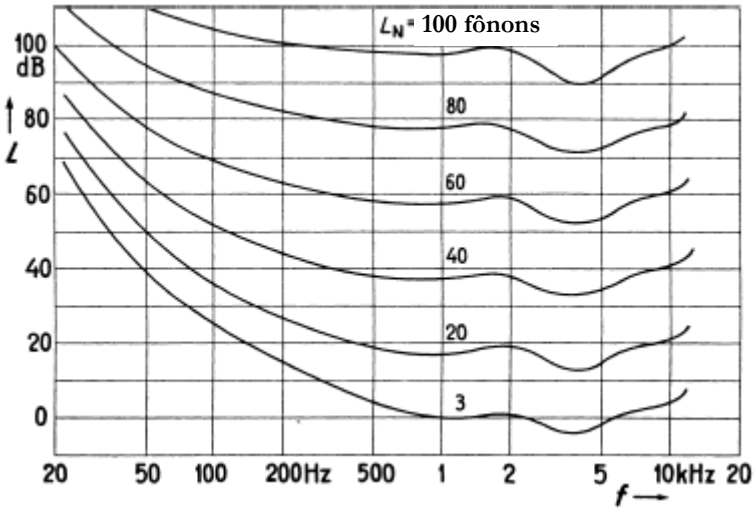


Figura 2.8 - Contornos de Sonoridade Equivalente.

2.2.4.2. Escala Sônon

Apesar de a sonoridade percebida aumentar naturalmente com o aumento do nível de pressão sonora, a relação entre estas duas grandezas não é linear. Por exemplo, se um sinal tem sonoridade em torno de 40 fônons, ao se adicionar 10 fônons ao sinal, a sonoridade percebida dobrará; porém, se o sinal estiver próximo do limiar de audição, a sonoridade percebida decuplicará. Fez-se então necessário transformar esta escala não-linear (fônons) para uma escala linear (sônons) [31]. Por definição, 1 *sônon* é o aumento de potência que faz dobrar a sonoridade percebida. A relação entre a sonoridade *N* e o nível de pressão sonora é então dada por uma lei de potência [15]:

$$N = 2^{\frac{1}{10}(L-40)} = 2^{-4} \cdot 2^{\log_{10}\left(\frac{I_{1kHz}}{I_0}\right)} = \frac{1}{16} \cdot \left(\frac{I_{1kHz}}{I_0}\right)^{0,3} \tag{2.6}$$

onde *N* é sonoridade em sônons, *L* é a sonoridade em fônons, I_{1kHz} é a intensidade correspondente a uma senóide a 1 kHz de igual sonoridade e I_0 é a intensidade de um tom a 1 kHz no limiar absoluto de audibilidade.

Zwicker [15] mediu a mesma relação com ruído de faixa estreita ao invés de tons puros e encontrou um expoente ligeiramente diferente (0,23). A diferença foi interpretada como um resultado da dependência de nível dos filtros auditivos usados para modelar a audição humana, como será visto mais adiante nas próximas seções e também no Capítulo 5. Uma outra expressão para a relação entre as escalas fônon e sônon foi proposta em [32] e é dada pela equação

$$L(b) = \left[P(b)\right]^{0,33}, \tag{2.7}$$

onde $P(b)$ é o sinal em fônons e $L(b)$, em sônons, em relação à banda crítica *b*.

2.2.5. Outros Fenômenos Auditivos

2.2.5.1. Limiar de Diferença

Outro aspecto da percepção de sonoridade é a mínima diferença de nível perceptível entre dois sinais que estejam sendo comparados (no caso deste trabalho, os sinais original e degradado, como será visto no Capítulo 5). Ela é referida ou como *limiar de diferença* ou, como será aqui adotado, *mínima diferença perceptível* (MDP). A MDP apresenta uma dependência com respeito à freqüência e também uma pequena dependência com relação ao nível, mas geralmente assume-se a MDP constante e igual a 1 decibel [33]. Contudo, para alguns sinais a MDP pode ser significativamente menor, podendo chegar a 0,25 dB para sinais de grande amplitude.

2.2.5.2. Sonoridade Parcial

A sonoridade percebida para uma distorção é um dos mais importantes indicadores de qualidade para um sinal de áudio processado. Por esse motivo, a percepção de sonoridade para distorções que estão próximas do limiar de mascaramento produzido pelas componentes do sinal original é de particular interesse para a confecção de modelos perceptuais. Uma componente, submetida à influência de outras componentes, será percebida com uma sonoridade menor do que se esta fosse apresentada isoladamente, a não ser que seu nível seja muito maior do que o limiar de mascaramento produzido pelas outras componentes, não importando se tal componente pertencente ao sinal original ou se é resultado de uma distorção. Quando o nível dessa componente é gradualmente diminuído para um valor próximo ao limiar de mascaramento, a sonoridade também é gradualmente reduzida, até se tornar zero ao alcançar o limiar. Nessa faixa onde a sonoridade de uma componente é reduzida pelas outras componentes, mas ainda não foi ainda completamente mascarada, a sonoridade percebida é chamada *sonoridade parcial*, e o efeito da diminuição da sonoridade percebida para o sinal mascarado é chamado de *mascaramento parcial*. A aproximação normalmente proposta para o mascaramento parcial é bastante simples. Em um teste de audição informal, a intensidade de um sinal parcialmente mascarado foi igualada à sua intensidade na ausência de um mascarador. O resultado é uma intensidade corrigida para o sinal mascarado [23], a qual pode ser aproximada pela equação

$$I'_n = \frac{I_n}{1+\left(\frac{I_s}{I_n}\right)^2}, \tag{2.8}$$

onde I_n é a intensidade do sinal mascarado sem a presença do mascarador e I_s é a intensidade do mascarador, ambos em dB.

Um modelo mais complexo é apresentado na Seção 2.3.2 [18].

Existem apenas alguns poucos modelos para a estimação da sonoridade parcial. Um deles foi introduzido na medida de qualidade de voz descrita em [29], enquanto outro foi proposto por Moore et al. [18].

2.2.5.3. Distorções Não-Lineares

Alguns fenômenos auditivos levam à conclusão de que o sistema auditivo humano produz distorções não-lineares. Exemplos de tais fenômenos são a ocorrência de pulsos (batidas), combinação de tons e a percepção de fundamentais ausentes. Estes fenômenos,

apesar de importantes para se ter um bom entendimento do mecanismo de audição, possuem pouca influência numa avaliação de um sinal de áudio, sendo por isso ignorados na maior parte dos modelos auditivos sugeridos.

2.2.5.4. Reflexo do Estapédio

A grandes pressões sonoras, ocorre um reflexo acústico que ativa dois pequenos músculos no ouvido médio, o estapédio e o tensor do tímpano. A ativação desses músculos afeta a transmissão da pressão sonora para o ouvido interno a fim de protegê-lo contra danos. Como esse reflexo reduz a quantidade de energia que alcança o ouvido interno em aproximadamente 20 dB [34], ele tem considerável influência na percepção de sinais de grande amplitude. Como este é um reflexo com o objetivo de proteger o ouvido, pode-se esperar que ele aconteça apenas quando a pressão sonora atinge um nível potencialmente perigoso para o ouvido. É claro que, em testes de audição com material de áudio de alta qualidade, tais níveis jamais serão atingidos. Portanto, a modelagem do reflexo do estapédio não é considerada importante para medidas perceptuais. De qualquer forma, ele tem sido ocasionalmente modelado em certas aplicações.

2.2.5.5. Reconhecimento de Padrões

Os efeitos de reconhecimento de padrões ocorrem especialmente no contexto de audição binaural (quando ambos os ouvidos estão envolvidos). O fenômeno deste tipo mais conhecido é a habilidade de detectar um sinal que, na verdade, está abaixo do limiar de mascaramento, se este vem de uma direção que não a do mascarador. Em tais situações, o limiar de mascaramento pode, obviamente, ser diminuído através de uma análise binaural. Tal efeito é chamado *diferença de nível de mascaramento binaural*.

Há também outros efeitos de reconhecimento de padrões que não requerem audição binaural. Normalmente, o limiar de mascaramento produzido por um mascarador de múltiplas componentes é claramente superior à soma dos limiares de mascaramento causados pelas componentes individuais do mascarador. No caso de mascaradores de amplitude modulada, o oposto pode ocorrer. A adição de uma nova componente ao mascarador pode diminuir o limiar de mascaramento produzido pelo mascarador total. Este efeito é chamado de mascaramento co-modulado. A maioria das hipóteses para a origem deste efeito pode, em geral, explicar a ocorrência deste fenômeno, mas não predizer inteiramente a quantidade do efeito. A explicação mais simples é a suposição de que o sistema auditivo compara as modulações de envoltória entre diferentes filtros auditivos. Se a mesma estrutura de modulação é encontrada para diferentes filtros, o sistema auditivo "percebe" que um sinal de amplitude modulada está presente, e pode tentar detectar outras componentes do sinal designando um peso maior nos instantes onde o mínimo temporal do mascarador ocorre (*dip listening)*.

2.3. MODELOS PERCEPTUAIS

Diversos modelos acústicos foram criados a partir dos fenômenos descritos na seção 2.2. Esses modelos simulam certas propriedades da audição humana. Como a sonoridade é a principal propriedade de um som, modelos da sonoridade percebida têm sido de particular interesse na pesquisa psicoacústica. Muitos métodos perceptuais usam idéias originárias de tais modelos, especialmente do modelo de Zwicker para a computação da sonoridade

percebida [15]. Outro modelo da sonoridade percebida foi recentemente introduzido por Moore, Glasberg e Baer [18]. Ambos os modelos serão brevemente descritos a seguir.

2.3.1. Modelo de Zwicker para o Cálculo da Sonoridade Percebida

Zwicker descreve um modelo que já inclui a maior parte dos passos de processamento que são usados nas medidas perceptuais. Uma versão simplificada deste modelo, que modela a resposta em freqüência do ouvido usando filtros com um formato determinado pelas terças-oitavas (ver Apêndice A), tem sido amplamente usada para a estimação da sonoridade no campo da prevenção de ruído, tornando-se parte de um padrão internacional [28]. No primeiro passo, o sinal de entrada é transformado para o domínio da freqüência e agrupado em bandas críticas. Esta operação pode ser expressa como uma função da densidade de energia na escala Bark, dada por

$$A(z) = \int_{z-0,5}^{z+0,5} \frac{dI}{dz'} \cdot dz' \,, \tag{2.9}$$

onde z é o valor da banda crítica, em Bark, dI/dz' é a densidade de energia da banda crítica e $A(z)$ denota a energia total na banda crítica z. A operação pode também ser expressa como uma função da densidade espectral de energia, na forma

$$A(z) = \int_{f(z-0,5)}^{f(z+0,5)} \frac{dI}{df} \cdot df \,, \tag{2.10}$$

onde dI/df é a densidade espectral de energia para a faixa de freqüência em Hz correspondente à banda crítica em questão.

Todos os dados utilizados por Zwicker foram coletados através de experimentos fisiológicos, os quais consistiram basicamente na observação da reação da membrana basilar a determinadas excitações e na contagem dos neurônios envolvidos em cada situação. O próximo parágrafo apresenta um breve resumo do estudo realizado. A matemática envolvida nesse processo é complexa, sendo por isso aqui omitida. Para maiores informações, consultar [28].

Em [28], o efeito do mascaramento simultâneo (domínio da freqüência) é interpretado como o resultado de um espalhamento das excitações dos neurônios presentes na área basilar, que corresponde à faixa de freqüência do estímulo sonoro, para áreas que na verdade não respondem ao estímulo sonoro; em outras palavras, a excitação de uma determinada região da membrana basilar irá disparar não apenas os neurônios ligados a tal região, mas também alguns neurônios das áreas adjacentes. Esse fenômeno é modelado através da determinação de como essa energia excita a banda crítica onde ela está concentrada e também as bandas críticas adjacentes. Assim, um sinal de grande amplitude irá disparar um grande número de neurônios na região da membrana basilar correspondente à sua freqüência, e também um certo número de neurônios nas regiões adjacentes; se um outro sinal for executado simultaneamente ao primeiro, este só será audível se for capaz de disparar um maior número de neurônios que o primeiro na região correspondente à sua freqüência na membrana basilar. Logicamente, quanto mais distante do primeiro sinal ele estiver (no domínio da freqüência), mais fácil será alcançar o número mínimo de neurônios necessários para se tornar audível, uma vez que a influência do primeiro sinal já não será muito expressiva. Portanto, o número de neurônios excitados é uma outra maneira de se explicar as curvas de mascaramento geradas por determinado tom.

Assim, as excitações designadas para as bandas críticas adjacentes são determinadas pela forma das curvas de mascaramento; onde o mascaramento é mais intenso, mais neurônios são envolvidos. Tal procedimento resulta em diversas excitações para cada banda crítica (uma originada pela energia da componente presente naquela banda crítica e outras provenientes das energias de componentes presentes nas bandas críticas adjacentes). A maneira como essas excitações parciais são adicionadas umas às outras não é definida em [15]. Em [35], a excitação é determinada pelo maior valor entre as excitações parciais devidas a cada componente, ou seja, as excitações são espalhadas de maneira apropriada e, a cada banda, apenas a maior componente presente irá determinar a excitação resultante. Ainda que tal procedimento não respeite a característica de aditividade do mascaramento (ver seção 2.2.3), sua implementação é consideravelmente mais simples, motivo pelo qual esta estratégia é mais largamente utilizada. Os *padrões de excitação* resultantes do procedimento descrito em [35] são transformados em uma função densidade *N'(z)*, também chamada de sonoridade específica, a qual é definida pela postulação de que a área entre esta função e o eixo da freqüência resulta na sonoridade *N* do estímulo sonoro. A transformação do padrão de excitação *E*(*z*) para o *padrão de excitação específico N'(z)*, dado em sônons/Bark, representa uma função de compressão não-linear dada por

$$N'(z) = k \cdot \left(\frac{1}{s} \cdot \frac{E_l(z)}{E_0} \right)^{\gamma} \cdot \left[\left(1 - s + s \cdot \frac{E(z)}{E_l(z)} \right)^{\gamma} - 1 \right], \tag{2.11}$$

onde:
k : fator de escala (em [15], k = 0,068);
γ : em [15], γ = 0,23;
E : excitação
E_l : excitação correspondente ao limiar absoluto de audibilidade
E_0 : excitação correspondente a um nível de pressão sonora de 40 dB (fator de escala)
s : fator de limiar.

A sonoridade geral do sinal é calculada a partir da equação

$$N = \int_0^{24} N'(z)\,dz\,. \tag{2.12}$$

O intervalo de integração de 0 a 24 corresponde à faixa total da escala Bark.

Os padrões de excitação *E*(*z*) que compõem a Equação 2.13 correspondem à excitação proveniente de determinado componente ou faixa do sinal. Seus valores são dados pela equação

$$E(z) = A(z) \cdot B(z)\,. \tag{2.13}$$

O equacionamento de *B*(*z*) não é trivial, e normalmente depende do modelo e do contexto no qual se está trabalhando. A seguir, são apresentadas três diferentes formulações com grau crescente complexidade e sofisticação, e uma quarta abordagem ligeiramente diferente das demais. É importante observar que o uso dos termos *padrões de excitação* e *curvas de mascaramento* é indiferente, uma vez que estes são conceitos equivalentes.

2.3.1.1. Curvas Independentes do Nível e com Pico Suave

A aproximação aqui adotada usa inclinações fixas para os filtros (curvas de mascaramento): 25 dB/Bark para a inclinação inferior e 10 dB/Bark para a inclinação superior [23], como mostra a equação

$$B(z/Bark) = 15{,}81 + 7{,}5 \cdot (z + 0{,}474) - 17{,}5 \cdot \sqrt{1 + (z + 0{,}474)^2} \ . \qquad (2.14)$$

A variável *B* fornece o formato aproximado, numa escala em dB, para os filtros adotados. Esta expressão modela uma transição suave entre a inclinação inferior e a superior.

2.3.1.2. Curvas Dependentes do Nível e com Pico Agudo

A aproximação aqui adotada leva em conta o mascaramento simultâneo distribuindo a influência de determinada excitação sobre todas as bandas críticas do espectro (termo z-z_c nas Equações 2.15 e 2.16), mas não provê uma transição contínua entre as inclinações inferior e superior [9], como mostram as equações

$$B(z/Bark) = 10^{\frac{1}{10} \cdot S_1 \cdot (z - z_c)} \quad \Big| \quad z \le z_c \,, \qquad (2.15)$$

$$B(z/Bark) = 10^{\frac{1}{10} \cdot S_2 \cdot (z - z_c)} \quad \Big| \quad z > z_c \,, \qquad (2.16)$$

onde

$$S_1 = 27, \ \ S_2 = -\left(24 + \frac{230}{f_c} - 0{,}2 \cdot L\right) \qquad (2.17)$$

e z_c corresponde à localização da máxima excitação na escala das bandas críticas, f_c é a freqüência central das bandas críticas e L é o nível relativo dos sinais em dB. S_1 e S_2 são dados em dB/Bark.

2.3.1.3. Curvas Dependentes do Nível e com Pico Suave

Uma generalização da Equação 2.13 para a curva de mascaramento com transição suave entre as inclinações inferior e superior é dada por

$$B(z/Bark) = A_0 + \left(\frac{S_1 - S_2}{2}\right) \cdot (z + c_1) - \left(\frac{S_1 + S_2}{2}\right) \cdot \sqrt{c_2 + (z + c_1)^2} \qquad (2.18)$$

onde S_1 e S_2 são, respectivamente, os valores das inclinações inferior e superior, A_0 é a excitação de pico, c_1 é o deslocamento de freqüência e c_2 é a largura da excitação de pico.

Esta expressão foi desenvolvida com um grande número de variáveis auxiliares (graus de liberdade), de modo a gerar uma curva com uma transição suave entre as inclinações inferior e superior, com as dependências do mascaramento simultâneo propostas em [9] e descritas na seção 2.2.3. Ele foi usado em [36] e no método perceptual descrito em [32].

2.3.1.4. Aproximação pelo Pior Caso

Se não é possível modelar a dependência de nível de curvas de mascaramento (por exemplo, quando o nível de execução do sinal não é conhecido a priori), o desvio entre a curva de mascaramento correta (dependente do nível) e a curva de mascaramento modelada

(fixa), pode ser reduzido pelo uso de uma curva de mascaramento modificada. Esta curva está na mesma posição ou abaixo da curva de mascaramento dependente do nível, em qualquer ponto acima do limiar absoluto de audibilidade. Originalmente, esta função era tomada a partir de uma tabela e levava à forma do limiar absoluto de audibilidade no silêncio. Com a simplificação em que o limiar absoluto é substituído por um valor constante de zero decibel, uma expressão analítica para a inclinação superior desta *curva de mascaramento para o pior caso* pode ser derivada a partir da dependência de nível dada por

$$B_{pc}(\Delta z) = 10^{-\frac{S_2'}{10}\cdot\Delta z\cdot\left(\frac{5}{5+\Delta z}\right)}, \tag{2.19}$$

onde

$$S_2' = 24\frac{dB}{Bark}.$$

Como a inclinação inferior não é dependente do nível, a curva de mascaramento para o pior caso é idêntica à aproximação dada na seção 2.3.1.2.

2.3.2. Modelo de Moore para o Cálculo da Sonoridade Parcial

Baseados nas medições psico-acústicas usando o método “notched-noise” [37], Moore e Glasberg [17] propuseram modelos para formatos e largura de faixa dos filtros auditivos descritos na seção 2.2.3.1, os quais diferem razoavelmente do modelo de Zwicker. Moore, Glasberg e Baer [18] sugeriram, baseados em tais modelos, um método completo para a predição de limiares e sonoridade total e parcial. De uma maneira geral, este modelo faz uso dos mesmos princípios utilizados no modelo de Zwicker, porém com diferentes abordagens. Apesar deste modelo ser o mais apropriado para muitas aplicações, o modelo de Zwicker se mostrou mais consistente no caso das medidas objetivas de avaliação de áudio. Por esse motivo, o modelo de Moore não será aqui descrito.

CAPÍTULO 3
CODIFICAÇÃO DE ÁUDIO

Quando a tecnologia dos discos compactos foi apresentada ao mundo, em meados dos anos 80, não havia a preocupação de se desenvolver sistemas digitais para compressão de dados, uma vez que a capacidade de armazenamento de um "Compact Disc" (CD) supria com facilidade as necessidades que se apresentavam. Além disso, a tecnologia de compressão de dados existente até então era exageradamente complexa em relação aos recursos computacionais disponíveis na época. Porém, o surgimento de novas mídias com conexões de faixa limitada e, muitas vezes, de alto custo, impôs a necessidade do desenvolvimento de métodos capazes de reduzir ao máximo as exigências para a transmissão e o armazenamento dos sinais de áudio. Desta forma, a compressão de dados, além de reduzir custos, permitiu o desenvolvimento de novas tecnologias que seriam impossíveis sem ela. Neste Capítulo serão descritas algumas das características básicas inerentes aos principais sistemas de compressão de dados de áudio, denominados codificadores de áudio. Estas características são relevantes para o estudo e desenvolvimento de métodos de avaliação objetiva para tais codificadores.

3.1. TIPOS DE CODIFICAÇÃO PARA COMPRESSÃO DE SINAIS DE ÁUDIO

A principal meta de um algoritmo de compressão digital é produzir uma representação mínima de um sinal que, quando descomprimido e reproduzido, é indistinguível do original. Em outras palavras, a mesma informação é transmitida ou armazenada usando uma taxa ou quantidade de dados menor. É importante ressaltar que, em áudio, o termo compressão é normalmente usado no contexto de uma diminuição da faixa dinâmica do som. No presente trabalho, contudo, tal termo será usado para designar uma redução na taxa de bits, sem que haja alteração na faixa dinâmica.

Na literatura afim e neste texto, a compressão digital é usualmente referida como codificação para redução de taxa de bits, ou simplesmente codificação. Para tal, existem diferentes técnicas de codificação que, do ponto de vista do sinal decodificado, podem ser divididas em dois grupos principais [38]:

a) Codificação sem perdas: neste caso, o sinal decodificado é idêntico ao original e, conseqüentemente, os dois são indistinguíveis. Este é o tipo de compressão usado por programas de compactação largamente utilizados em computadores pessoais (WinZip, WinRar, etc.), uma vez que, para estes casos, a corrupção de um único bit pode ser catastrófica. Quando aplicada a áudio, porém, esta abordagem apresenta um baixo ganho de codificação, em torno de 2:1 [39]. É importante observar que o codificador sem perdas não pode garantir um fator de compressão particular, uma vez que este está diretamente relacionado com a quantidade de redundância que se pode extrair de um determinado sinal. Assim sendo, o sistema de comunicações ou gravador utilizados deverão ser capazes de funcionar com uma taxa variável na saída. Codecs deste tipo podem ser concatenados sem

qualquer precaução especial. Um exemplo deste tipo de abordagem é a codificação de Huffman [40].

b) Codificação com perdas: neste caso, o sinal decodificado não será, numa comparação bit-a-bit, idêntico ao original, o que acarretará um determinado grau de degradação. Codecs com perdas não são aconselháveis para compressão de dados, mas devido à sua grande capacidade de compressão, são largamente utilizados para sinais de áudio. Os codecs com perdas mais bem sucedidos são aqueles em que os erros devidos à técnica de compressão utilizada são arranjados de maneira que se tornem subjetivamente difíceis de serem detectados. Portanto, este tipo de codec deve ser baseado em modelos psico-acústicos, sendo por isso chamados de códigos perceptuais [38]. Na codificação perceptual, quanto maior for o fator de compressão requerido, mais preciso deve ser o modelo auditivo utilizado. Este tipo de codificador pode ser forçado a operar a um fator de compressão constante, o que é conveniente para aplicações práticas onde uma taxa fixa pode ser mais facilmente tolerada que uma taxa variável. O resultado de uma compressão fixa é que a qualidade da codificação irá variar significativamente de sinal para sinal, uma vez que há sinais com mais redundância que outros. Este tipo de codec não pode ser concatenado indiscriminadamente, especialmente se diferentes algoritmos são utilizados, pois isso acarretará uma introdução de erros em cascata. Como se pode perceber, uma avaliação da qualidade da codificação só faz sentido para este tipo de codec, já que a codificação sem perdas resulta num sinal idêntico ao original. A codificação perceptual pode ser subdividida em três categorias [38]:

- *Codificação por Sub-bandas*: este tipo de codificação imita o mecanismo de análise em freqüência do ouvido, dividindo o espectro do sinal em um grande número de bandas distintas. Os sinais nessas bandas podem então ser quantizados independentemente. Os erros de quantização de cada banda são confinados aos limites de freqüência da banda correspondente, e podem então ser arranjados de maneira a serem mascarados pelo sinal. Esta é a técnica utilizada pelos codecs tipo MPEG, layers I e II (ver seções 3.2. e 3.3).

- *Codificação por Transformada*: aqui, a forma de onda no domínio do tempo é convertida em uma representação em algum outro domínio através de técnicas como a Transformada Rápida de Fourier (FFT), a Transformada Co-seno Discreta Modulada (MDCT) e Wavelets. A codificação por transformada faz uso do fato de que a envoltória de um sinal varia de forma relativamente lenta, e então todos os coeficientes da transformada podem ser transmitidos de maneira intermitente, ou seja, de maneira descontinuada. Este tipo de abordagem tende a falhar na presença de transientes (surtos ou variações muito rápidas no sinal), os quais fazem com que seja necessária uma atualização freqüente nos coeficientes, uma vez que estes mudam rapidamente sob tais condições. Este tipo de abordagem é usado nos codecs tipo MPEG layer III, Dolby AC-3 e Sony ATRAC. Este último faz uso também da codificação por sub-bandas.

- *Codificação Preditiva*: em um codificador preditivo, existem dois preditores idênticos, um no codificador e outro no decodificador. Sua tarefa é examinar uma quantidade de amostras anteriores e extrapolá-las para estimar ou predizer qual será o próximo valor de código. Este é subtraído do valor real correspondente no codificador para produzir um erro de predição, o qual é transmitido. O decodificador adiciona então o erro de predição à sua própria predição para obter o valor de saída. Este tipo de codificador trabalha com um pequeno atraso de codificação e decodificação, e é útil em telefonia, onde grandes atrasos podem ocasionar sérios problemas.

3.2. VISÃO GERAL DOS PRINCIPAIS CODECS DE ÁUDIO

Para se obter uma qualidade compatível com aquela apresentada pelos CDs de áudio em uma transmissão sem compactação, deve-se utilizar uma amostragem mínima de 44,1 kHz e 16 bits por amostra. Os 16 bits são necessários pelo fato de se exigir, para sinais de áudio, uma relação sinal/ruído de quantização da ordem de 96 dB. Como cada bit de quantização aumenta a relação sinal-ruído em cerca de 6 dB, verifica-se que a quantização deve ser feita com no mínimo 16 bits. Portanto, ter-se-á, para um sinal amostrado a 44,1 kHz, uma taxa de 44100*16 ≅ 705 kbits/s por canal, o que corresponde a 1,41 Mbits/s para um sinal estéreo e a 4,23 Mbits/s para um sistema com 6 canais, taxas estas, na prática, proibitivas. Portanto, os codificadores de áudio são essenciais para a transmissão através dos sistemas de telecomunicações digitais e também para o armazenamento, como em CD, DVD, *minidisc*, etc.

A Tabela 1 fornece uma visão geral de alguns dos sistemas de codificação de áudio disponíveis, dentre os quais destacam-se a família MPEG, o Dolby AC3 e o Sony ATRAC [39].

Tabela 1 - Comparação entre alguns dos Principais Codecs de Áudio Disponíveis no Mercado

Codec	Taxa para uma Boa Qualidade (kb/s)	Complexidade	Principais Aplicações
MPEG-1 Layer I	192 por canal de audio estéreo	Baixa p/ cod. e dec.	Cassete Compacto Digital
MPEG-1 Layer II	128 por canal de audio estéreo	Baixa p/ decodificador	DAB, CD-I, DVD
MPEG-1 Layer III	96 por canal de audio estéreo	Baixa p/ decodificador	ISDN, Sistemas de Radio via Satélite, Áudio de Internet
Dolby AC-2	128	Baixa p/ cod. e dec.	Ponto a Ponto, Cabo
Dolby AC-3	384 para os 6 canais de áudio	Baixa p/ decodificador	Ponto a multiponto, HDTV, cabo, DVD, Cinema, LaserDisc
Sony ATRAC	140 por canal	Baixa p/ cod. e dec.	MiniDisc
MPEG-2 AAC	384 para 6 canais de áudio	Baixa p/ decodificador	HDTV, DVD, rádio na Internet, etc.

As principais características inerentes a estes e alguns outros codecs serão descritas nas próximas seções.

3.3. FAMÍLIA MPEG

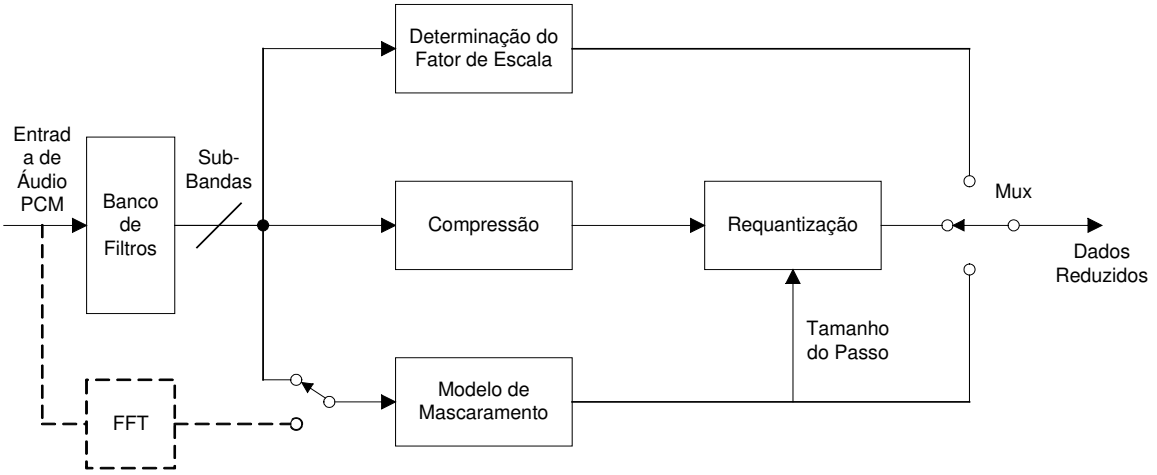
O nome MPEG é na verdade a sigla para um grupo de especialistas em imagens (*Moving Pictures Expert Group*), o qual foi criado pela Organização Internacional de Padrões (ISO) para a determinação de padrões para compressão e transmissão de áudio e vídeo. O primeiro padrão estabelecido foi o MPEG-1 [41,42,43], o qual é composto por três diferentes versões, denominadas *layers*, com crescente complexidade e ganho de codificação. Na seqüência, surgiu o padrão MPEG-2 [44], cuja aplicabilidade se mostrou bem mais ampla. Estes dois padrões são os mais importantes representantes da família MPEG no contexto do presente trabalho. Além deles, outros padrões foram desenvolvidos para diferentes aplicações, como o MPEG-4, o MPEG-7 e o MPEG-21.

3.3.1. Padrão MPEG-1

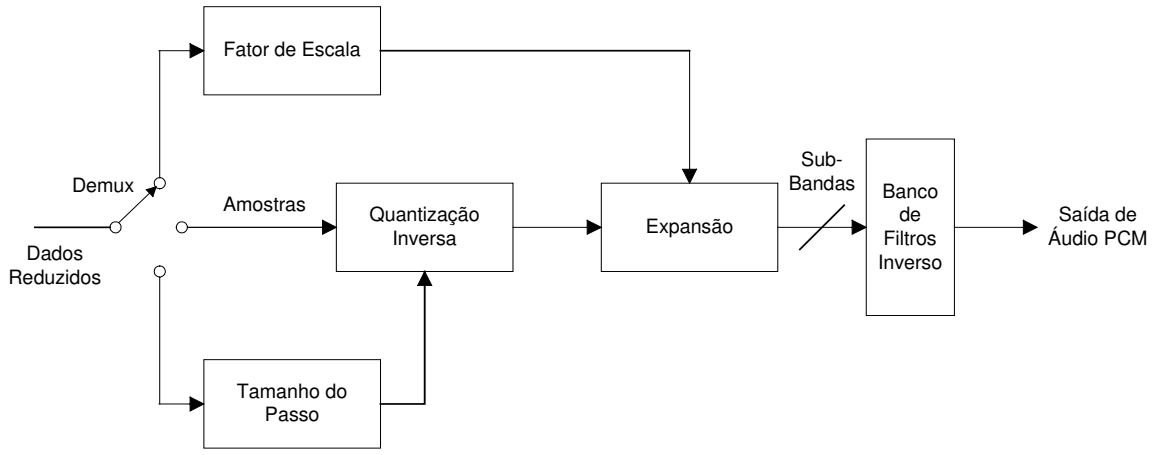
Como já comentado, o padrão MPEG-1 é composto por três versões distintas, cada qual com determinadas características que as tornam mais adequadas a certos tipos de aplicação, como mostrado a seguir.

3.3.1.1. Layer I

O *layer I* do padrão MPEG-1 é uma versão modificada do sistema MUSICAM (*Masking-pattern Universal Subband Integrated Coding and Multiplexing*) [45]. Ele é compatível com o codec usado nos Cassetes Compactos Digitais da Phillips (DCC), também conhecido como Codificador por Sub-Bandas com Precisão Adaptativa (PASC-*Precision Adaptive Subband Coding*). A Figura 3.1 mostra um diagrama de blocos deste codec [38].



CODIFICADOR



DECODIFICADOR

Figura 3.1 - Esquema Geral do Codec MPEG-1 Layer I.

A codificação por sub-bandas faz uso das vantagens oferecidas pelo fato de que sons reais não têm uma energia espectral uniforme. Na sua forma mais simples, a codificação por sub-bandas trabalha dividindo o sinal de áudio em um certo número de bandas de freqüência e comprimindo e expandindo ("companding") o sinal de cada banda de acordo com seu próprio nível e com o nível de mascaramento a que o sinal está submetido, o que não é feito na codificação PCM. Assim, bandas com uma relação sinal-mascarador pequena poderão ser codificadas com palavras-código menores. Por exemplo, uma banda que possua uma relação sinal-mascarador de 30 dB (um valor mediano) poderá ser codificada com uma palavra-código de 5 bits, ao invés dos 16 bits do PCM. Ainda que a codificação em cada banda resulte em palavras-código de comprimento variável, a soma do comprimento de todas as palavras-código será menor que aquela encontrada no PCM, e então um ganho de codificação será obtido (blocos compressão e descompressão na Figura 3.1).

A técnica utilizada para a divisão das bandas é a Filtragem por Quadratura Espelhada (QMF) [46,47]. Este modelo, representado na Figura 3.1 pelo bloco "modelo de mascaramento", pode ser estendido a bandas de qualquer largura para a determinação da quantidade de mascaramento, possibilitando um ganho de codificação. Sub-bandas de largura uniforme não serão capazes de utilizar o mascaramento de maneira tão eficiente quanto bandas "casadas" com as bandas críticas, mas para muitas aplicações o ganho de codificação adicional obtido neste último caso não compensaria o aumento da complexidade do filtro.

Como a codificação MPEG é baseada no princípio do mascaramento auditivo, as sub-bandas devem ser, preferivelmente, mais estreitas que as bandas críticas, sendo, portanto, em maior número. Quanto mais estreita for a sub-banda, maior será o nível de ruído de quantização que pode ser mascarado, uma vez que se poderá fazer uma análise mais refinada do fenômeno de mascaramento, o qual é, desta forma, explorado de maneira mais eficiente. Porém, o uso de um número excessivo de sub-bandas irá aumentar a complexidade e, conseqüentemente, o atraso de codificação. O padrão MPEG-1 layer I utiliza 32 sub-bandas.

Os fatores de escala consistem no maior valor encontrado para as amostras pertencentes a determinada sub-banda (bloco "determinação do fator de escala" da Figura 3.1). Esses valores são usados para normalizar as amostras das sub-bandas.

O tamanho adotado para os blocos de entrada é constante, contendo 384 amostras. A 48 kHz, este número corresponde a um período de 8 ms, tamanho ideal para levar em consideração o fenômeno do pré-mascaramento. Após a filtragem pelas 32 sub-bandas, a dizimação por 32 correspondente da taxa de bits faz com que cada bloco contenha apenas 12 amostras. As amostras em cada bloco filtrado são comprimidas e expandidas de acordo com o valor de pico e o fator de mascaramento medido.

Por simplicidade, nesta versão do algoritmo MPEG, os níveis das 32 sub-bandas são usados como uma análise espectral grosseira da entrada, a fim de se derivar o modelo de mascaramento. Tal modelo utiliza o espectro de entrada para determinar um novo limiar de audibilidade, que por sua vez determina o nível de ruído aceitável para cada uma das sub-bandas. Então, como já visto, onde se detecta mascaramento, o sinal é quantizado de maneira mais grosseira, até que a quantização do ruído se aproxime ao máximo do limiar de mascaramento. Uma quantização mais grosseira requer palavras-código mais curtas, permitindo assim um ganho de codificação.

As palavras-código com diferentes comprimentos são então reunidas no bloco codificado de saída. Se um fator de compressão fixo é empregado, o tamanho do bloco

codificado de saída será fixo. Os tamanhos das palavras-código em cada bloco devem ser tais que a soma de bits de todas as sub-bandas seja igual ao tamanho do bloco codificado. A ação de alocação de bits é que se encarrega de fazer o ajuste fino do comprimento das palavras-código de cada sub-banda, de modo que a soma destas palavras-código seja igual ao tamanho do bloco codificado. A alocação de bits é determinada pela minimização da relação ruído-mascaramento em cada sub-banda, bem como em todo o quadro [41]. Quanto menor esta relação, maior será a qualidade do sinal. Em cada iteração, a sub-banda a ser mais beneficiada (ou seja, aquela com maior relação ruído-mascaramento) recebe mais bits para a quantização, até que todos os bits para a taxa escolhida tenham sido alocados. Esta é a etapa referente ao bloco "requantização" da Figura 3.1.

A fim de que o decodificador possa decompor o bloco em palavras-código de comprimento variável e designar a estas a freqüência apropriada, deve-se informar ao decodificador que alocação de bits foi utilizada. Além disso, algum tipo de sincronização deve ser utilizado a fim de permitir a identificação do início do bloco. Tais procedimentos são realizados pelo multiplexador (MUX), o qual arranja todos os bits de maneira apropriada.

Não é difícil modificar o tamanho do pacote de bits de saída de modo a obter um fator de compressão diferente. Se um pacote maior é especificado, o alocador de bits irá agir iterativamente até que o novo tamanho seja alcançado. Similarmente, o decodificador simplesmente deverá decompor corretamente o novo pacote de saída nas palavras-código codificadas, e então o processo de expansão é idêntico, exceto pelo fato de que palavras-código expandidas contêm menos ruído. Isto possibilita a existência de codecs com diferentes graus de compressão, os quais podem assumir diferentes compromissos entre largura de banda e desempenho usando o mesmo "hardware".

É importante observar que o bloco "FFT" da Figura 3.1 não faz parte do processamento do "layer I"; esta etapa faz parte apenas do "layer II", como será visto na próxima seção.

3.3.1.2. Layer II

A versão "Layer II" do padrão MPEG-1 é idêntica ao MUSICAM. O mesmo banco de filtros de 32 bandas e o mesmo esquema de compressão e expansão encontrado na versão "Layer I" são utilizados. O uso do nível de uma sub-banda para determinação do modelo de mascaramento, tal como adotado no "Layer I", é evitado, uma vez que esta é uma abordagem sub-ótima, pois se conhece o local exato da energia dentro da sub-banda. Ao invés disto, esta versão explora o fato de que as bordas da curva de mascaramento são assimétricas, ou seja, um tom irá mascarar com mais intensidade as freqüências acima da sua. Portanto, se esse mascarador estiver na borda inferior da sub-banda, a intensidade do mascaramento será maior do que no caso deste estar na freqüência superior. Assim, menos bits podem ser alocados para o primeiro caso, resultando num ganho de codificação. A fim de se obter uma resolução espectral melhor que aquela alcançada pelo banco de filtros, uma FFT de 1024 pontos é computada na determinação do modelo de mascaramento, como mostrado na Figura 3.1 Para que a FFT tenha resolução suficiente, o tamanho do bloco (e da respectiva FFT) é aumentado para 1152 amostras (três vezes o comprimento adotado no "Layer I").

As amostras quantizadas em cada sub-banda, os dados de alocação de bits, os fatores de escala e os códigos para seleção do fator de escala apropriado para cada sub-banda são multiplexados em um conjunto de bits de saída.

O esquema de compressão e expansão do “Layer II” é o mesmo usado no “Layer I”, já que o bloco de 1152 amostras é dividido em três de 384 amostras. Porém, nem todos os fatores de escala são transmitidos, devido ao fato de que eles contêm um certo grau de redundância. A diferença entre os fatores de escala de blocos sucessivos excede 2 dB em menos de 10% dos casos. O “Layer II” analisa o conjunto de três fatores de escala sucessivos em cada sub-banda. No caso estacionário, eles serão iguais, e apenas um dos fatores de escala será transmitido. À medida que o conteúdo transiente aumenta em uma dada sub-banda, dois ou três fatores de escala serão transmitidos. Um código de seleção do fator de escala deve ser transmitido para permitir que o decodificador determine quais foram transmitidos em cada sub-banda. Esta técnica reduz em cerca de 50% a taxa de bits para o fator de escala.

O decodificador do “Layer II” não é muito mais complexo que aquele utilizado no “Layer I”, já que o único processamento adicional é a decodificação dos fatores de escala comprimidos a fim de produzir um fator de escala para cada bloco de 384 amostras.

3.3.1.3. Layer III

Esta é a versão mais complexa do padrão MPEG-1, e só é realmente necessária quando há restrições severas quanto à taxa de bits, sem que haja perda significativa da qualidade. Esta é a estrutura de codificação dos arquivos MP3 largamente utilizados na atualidade. Ele é um codificador por transformada baseado no Sistema de Codificação de Entropia Espectral Perceptual Adaptativa (ASPEC), com algumas modificações que asseguram um certo grau de familiaridade com o “Layer II”. O codificador ASPEC original usa uma Transformada Co-seno Discreta Modificada (MDCT) direta nas amostras de entrada. No “Layer III” usa-se uma transformação híbrida incorporando a QMF para geração de 32 bandas utilizada nas outras versões (ver seção 3.3.1.1). Assim, as 32 bandas originadas através da QMF são processadas por uma MDCT de 12 bandas para se obter 384 coeficientes de saída. Dois tamanhos de janela são utilizados para evitar pré-eco em transientes.

Um modelo perceptual bastante preciso é usado para aproveitar a resolução espectral disponível. Uma quantização não-uniforme é usada, juntamente com a codificação de Huffman.

3.3.2. Padrão MPEG-2

O padrão MPEG-1 é capaz de lidar com sinais de áudio de um ou dois canais (mono e estéreo), a freqüências de amostragem comumente usadas em áudio de alta qualidade (32, 44,1 e 48 kHz). A versão MPEG-2 foi criada com o objetivo de ampliar a aplicabilidade do padrão. Tal trabalho foi dividido em três partes [48]:

- extensão para freqüências de amostragem mais baixas (16, 22,05 e 24 kHz), fornecendo uma melhor qualidade a taxas de bits muito baixas (abaixo de 64 kbits/s por canal).
- compatibilidade regressiva com o MPEG-1 para sons multicanal. Esta versão é chamada MPEG-2 BC (Backward-Compatible) [49], e suporta até 5 canais de banda completa e mais 1 canal para enriquecimento das baixas freqüências (este arranjo de canais é referido como “5.1”). Esta extensão multicanal é compatível com o MPEG-1 tanto progressiva quanto regressivamente, ou seja, há uma compatibilidade mútua.

- desenvolvimento de um novo esquema de codificação, mais eficiente, denominado MPEG-2 AAC (*Advanced Audio Coding*) [50,51,52]. Esta versão não pode ser lida ou interpretada pelo decodificador usado no MPEG-1 (inexistência de compatibilidade regressiva).

O MPEG-2 BC é dividido nas mesmas três versões (*layers*) encontradas no MPEG-1. Já o MPEG-2 AAC possui uma única versão [48].

A principal aplicação do MPEG-2 é a televisão digital, uma vez que este é também um codificador de vídeo, sendo capaz de produzir a qualidade necessária na televisão de alta definição (HDTV).

A seguir, serão apresentadas algumas das principais características das duas versões do MPEG-2.

3.3.2.1. MPEG-2 BC

O núcleo do arranjo de bits adotado no MPEG-2 BC é semelhante àquele adotado no MPEG-1, o que possibilita uma compatibilidade total com um decodificador MPEG-1. Adicionalmente, a necessidade de se transferir dois arranjos de bits separados (um para os canais estéreo e outro para os 5.1 canais) é evitada, ao custo de uma certa perda na eficiência de codificação no caso de sinais multicanal. Para manter a compatibilidade regressiva com o MPEG-1, os 5.1 canais em um sinal de áudio multicanal normal não são codificados diretamente [40]. Ao invés, utiliza-se uma matriz de compatibilidade, a fim de se fazer mistura desses canais, tornando-os decodificáveis pelo esquema MPEG-1 (ver Figura 3.2 [53]).

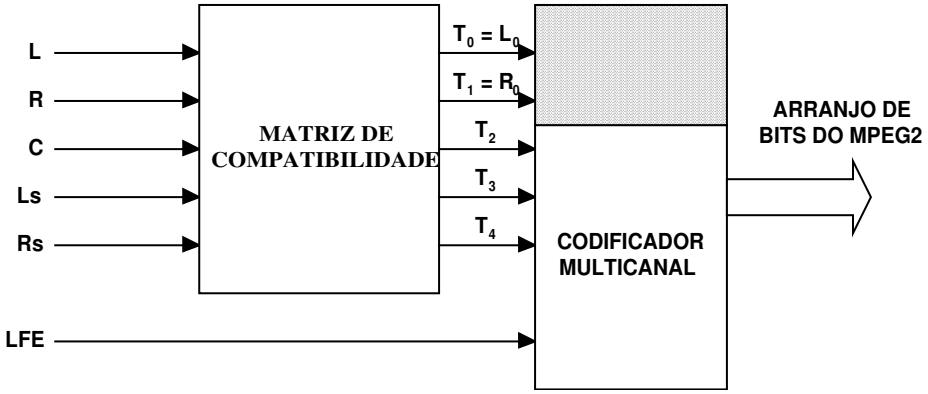


Figura 3.2 - Codificação dos 5.1 canais para o arranjo de bits MPEG-2 com matriz de compatibilidade.

onde,
L (*left*): canal esquerdo;
R (*right*): canal direito;
C (*center*): canal central;
Ls (*left surround*): canal esquerdo envolvente;
Rs (*right surround*): canal direito envolvente;
LFE (*low frequencies enhancement*): extensão para baixas freqüências;
T_i: canais após compatibilização.

O processo de decodificação é mostrado na Figura 3.3 [53]. O decodificador estéreo pode facilmente decodificar os canais necessários, enquanto que o decodificador multicanal deve desfazer a operação realizada pela matriz de compatibilidade no codificador. Esse procedimento tem sido alvo de severas críticas, uma vez que em certos casos há uma perda quase completa do conteúdo de determinado canal, porém o ruído de quantização correspondente a tal canal é preservado, causando sérias degradações [45]. A divisão em *layers* do MPEG-2 é idêntica ao MPEG-1.

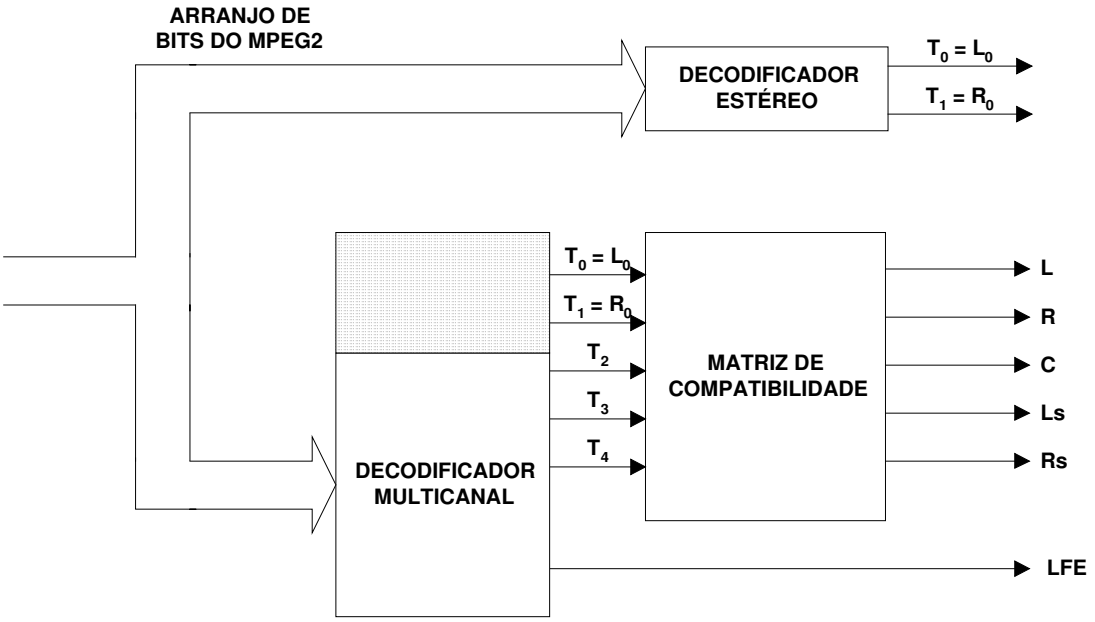


Figura 3.3 - Decodificação do MPEG-2 usando matriz de compatibilidade.

3.3.2.2. MPEG-2 AAC

Este padrão é um aperfeiçoamento dos padrões anteriormente desenvolvidos (MPEG-BC), permitindo que se obtenha uma boa qualidade de áudio a taxas de 64 kb/s por canal para operações multicanal [48]. Ele usa basicamente os mesmos paradigmas de codificação dos padrões MPEG-1 e 2 Layer III, porém com algumas ferramentas adicionais e melhoramento de alguns detalhes, resultando numa eficiência de codificação cerca de 30% melhor que aquela obtida pelo MPEG-2 Layer III.

O MPEG-2 AAC oferece três diferentes perfis, cada qual com certas características particulares:

- *Perfil Principal*: usado quando não há restrições sérias de processamento ou memória.
- *Perfil de Baixa Complexidade*: usado quando há restrições de memória e/ou processamento.
- *Perfil com Taxa de Amostragem Escalonável*: usado em casos em que um decodificador escalonável é necessário.

3.3.3. Outros Padrões

A família MPEG possui ainda outros padrões que fogem ao escopo deste trabalho:

- padrão MPEG-4 [54,55]: voltado a aplicações multimídia, permite a integração dos paradigmas de produção, distribuição e acesso de conteúdo de aplicações como televisão digital, gráficos interativos e multimídia;
- padrão MPEG-7 [56,57,58]: quando comparado aos outros padrões, o MPEG-7 mostra um alto grau de abstração, uma vez que ele foi projetado para representar a informação a respeito da informação, ao passo que os outros buscam representar a própria informação. Em outras palavras, enquanto os MPEGs-1, 2 e 4 permitem que um conteúdo seja disponibilizado, o MPEG-7 permite que tal conteúdo seja descrito e encontrado. Ou seja, ele facilita a busca por conteúdos de mídia, tornando-a mais flexível e eficiente que os sistemas de busca baseados em texto e amplamente utilizados por usuários de todo o mundo na rede mundial de computadores [59].
- MPEG-21 [60]: tem como objetivo a integração de diversas tecnologias, possibilitando um uso mais transparente e amplo dos recursos de multimídia através de uma ampla gama de redes e dispositivos para suportar funções como: criação de conteúdo, produção de conteúdo, gerenciamento e proteção de propriedade intelectual, identificação e descrição de conteúdo, gerenciamento financeiro, privacidade do usuário, abstração de recursos de rede, representação de conteúdo, etc. Tal padrão é apenas o ponto de partida para a integração de tecnologias tão diferentes, e seus autores concordam que seu sucesso dependerá da participação e colaboração de outros grupos e pesquisadores.

3.4. PADRÃO DOLBY AC-3

O AC-3 é um codificador de áudio multicanal desenvolvido pela Dolby Digital [61]. Assim como o padrão MPEG, seu principal objetivo é fornecer a melhor qualidade possível através de um algoritmo de baixa complexidade. Suas principais aplicações incluem a trilha sonora digital de filmes de cinema, HDTV, multimídia, serviços a cabo, etc.

O AC-3 teve como antecessor o padrão AC-2, voltado para a codificação de um único canal [62]. Ambos são fundamentalmente codificadores adaptativos baseados em transformadas, usando um banco de filtros implementado através da técnica de cancelamento de “aliasing” no domínio do tempo (TDAC) [63].

O AC-3 usa uma técnica flexível de alocação de bits para distribuí-los de maneira eficiente através das diferentes freqüências e canais, levando em conta os efeitos de mascaramento intra e inter-canais [64].

A seguir, serão apresentados os principais blocos usados no codificador e no decodificador do padrão AC-3.

3.4.1. Codificador

A Figura 3.4 mostra a estrutura de codificação usada no AC-3 [61]. Tais blocos serão brevemente descritos a seguir.

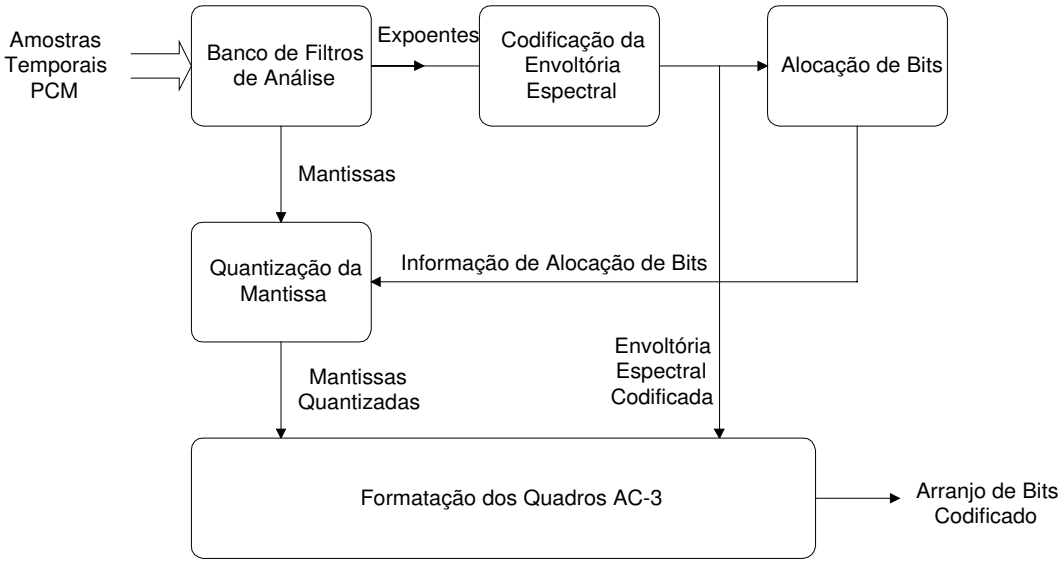


Figura 3.4 - Esquema de Codificação do Padrão AC-3.

3.4.1.1. Banco de Filtros de Análise

Esta etapa é composta por cinco processamentos distintos [64]:

- "Buffer" de entrada: como o AC-3 é um codificador estruturado por blocos, um ou mais blocos (tipicamente de 512 amostras para uma freqüência de amostragem de 48 kHz) dos sinais no domínio do tempo são coletados, a partir de cada canal, em um "buffer" de entrada, antes de se realizar o restante dos processamentos.
- Filtragem de entrada: os sinais de entrada são individualmente submetidos a um filtro passa-altas com freqüência de corte em 3 Hz, a fim de eliminar a componente DC. O sinal a ser enviado para o "subwoofer" (ou canal de extensão para baixas freqüências) é também filtrado por um filtro passa-baixas com corte em torno de 120 Hz.
- Detecção de transientes: os picos de nível de diferentes segmentos dentro de cada bloco dos sinais filtrados são analisados e comparados para a detecção de transientes. Esta informação é usada para ajustar o tamanho dos blocos do banco de filtros, restringindo o ruído de quantização associado ao transiente para uma pequena faixa temporal, garantindo que esse ruído seja mascarado.
- Banco de filtros TDAC: os sinais de entrada de cada canal são então janelados e filtrados individualmente através de um banco de filtros, usando a técnica de cancelamento de "aliasing" no domínio do tempo (TDAC). A transformação TDAC é basicamente uma FFT submetida a algumas manipulações adicionais; ele oferece uma baixa complexidade computacional com uma boa seletividade de freqüência. Os quadros de análise são superpostos em 50%. O ganho de codificação é obtido principalmente a partir da quantização seletiva dos coeficientes da transformada. Como o codificador não adiciona nem elimina informação audível, a saída decodificada deverá ser perceptualmente idêntica à entrada do codificador.
- Conversão para ponto flutuante: mesmo quando implementado em DSPs em ponto fixo, os coeficientes da transformada TDAC são convertidos para ponto flutuante antes da submissão a outros processamentos, com as mantissas assumindo uma

gama de amplitudes entre 0,5 e 1,0. Isto visa assegurar que o processamento intermediário não irá impor limitações práticas na faixa dinâmica. Como resultado, o AC-3 preserva os benefícios da alta resolução dos conversores A-D e D-A (18 a 22 bits). A representação dos dados em ponto flutuante, e particularmente a presença de expoentes, serve também como um auxílio computacional para processos logaritmicamente orientados, como a alocação de bits.

3.4.1.2. Codificação da Envoltória Espectral

Em geral, a demanda média de bits para múltiplos canais é, de maneira aproximada, proporcional à raiz quadrada do número de canais [64]. Se são necessários 128 kb/s para um único canal, então os 5.1 canais requererão, em média, $128 \cdot \sqrt{6} = 289\,kb/s$, o que cabe confortavelmente dentro da taxa mínima de dados de 320 kb/s usada pelo AC-3. Isto implica em que a maioria dos sinais multicanal podem ser apropriadamente codificados usando apenas a flexibilidade da técnica global de alocação de bits.

No entanto, alguns sinais requerem taxas de bits elevadas, e neste caso faz-se necessário o uso de técnicas adicionais. O AC-3 lida com este problema explorando o fenômeno psicoacústico de que, em altas freqüências, o sistema auditivo humano localiza os sons tendo como base as envoltórias dos sinais filtrados pelas bandas críticas, ao invés dos sinais propriamente ditos [65]. Assim, as altas freqüências do sinal são decompostas em envoltória e portadora, codificando-se a informação da envoltória com uma maior precisão do que a informação da portadora. Se necessário, pode-se ainda combinar apropriadamente as componentes da portadora, o que faz com que as redundâncias para a localização das componentes de alta freqüência sejam eliminadas, obtendo-se desta forma um ganho de codificação adicional. Tal procedimento tem um impacto audível mínimo, uma vez que a informação da localização é preservada nos dados da envoltória, e a combinação das portadoras é um fenômeno que ocorre naturalmente no ouvido dos ouvintes, produzindo um resultado equivalente.

A técnica utilizada consegue preservar diversas características acústicas, não somente de sons provenientes de fontes discretas (alto-falantes), mas também de "imagens fantasmas" surgidas entre os alto-falantes.

3.4.1.3. Alocação de Bits

A principal vantagem da codificação multicanal unificada é a habilidade da rotina de alocação de distribuir os bits de quantização através dos canais e freqüências conforme a necessidade, a fim de atender às diferentes demandas de um mesmo sinal. O alocador de bits do AC-3 analisa os coeficientes TDAC com respeito a seus efeitos de mascaramento e sua relação com o limiar absoluto de audibilidade, a fim de computar o número de bits requerido para codificar cada mantissa. O cálculo é realizado globalmente no arranjo dos canais como uma única entidade.

Ainda que os efeitos de mascaramento intercanais sejam levados em consideração, a habilidade de um sinal de mascarar um ruído em um canal diferente é limitada e sua eficácia depende da posição do ouvinte; dessa forma, sua contribuição é mantida pequena.

3.4.1.4. Quantização da Mantissa

Os resultados obtidos no cálculo da alocação de bits são usados para quantizar os dados da mantissa. Ao invés de simplesmente se enviar os *n* bits mais significativos de um

valor, tal valor é escalonado e compensado para prover níveis de quantização centrados em zero, de igual largura e simétricos (simetria ímpar), para minimizar as distorções e facilitar outros processamentos.

3.4.1.5. Formatação dos Quadros (empacotamento dos dados)

Os processos descritos até aqui convertem cada bloco de 6 canais em uma série de arranjos derivados e valores escalares, incluindo os expoentes TDAC e as mantissas quantizadas, informação de alocação de bits, coeficientes de acoplamento, etc. No estágio final do processo de codificação, esta informação é arranjada em um único bloco, juntamente com a informação de sincronização, uma legenda e informação de correção de erros.

3.4.2. Decodificador

A Figura 3.5 mostra a estrutura de decodificação usada no AC-3 [61]. A seguir, cada bloco será brevemente descrito.

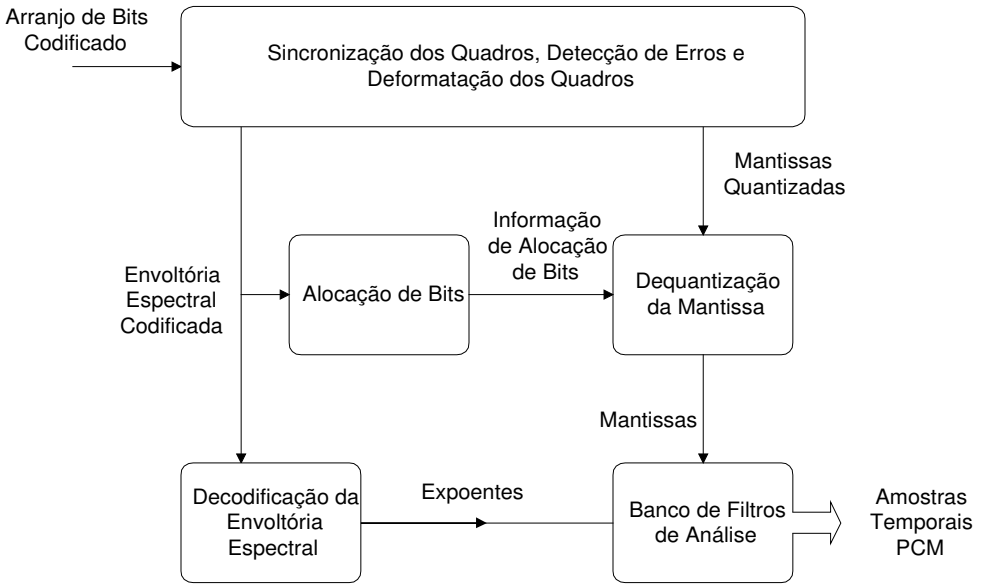


Figura 3.5 - Esquema de decodificação do padrão AC-3.

3.4.2.1. Sincronização dos Dados, Detecção de Erros e Deformatação dos Quadros

Esta etapa é composta por três processamentos, como mostrado a seguir.

- Buffer de Entrada: o decodificador, como o codificador, é estruturado em blocos, e, então, estabelece e mantém sincronismo com o arranjo de dados de entrada, coletando um bloco inteiro em um "buffer" de entrada antes da realização da decodificação.
- Encobrimento de Erro: cada bloco de dados de entrada é checado para verificação da consistência interna, bem como a informação para correção de erros. Se uma condição de erro incorrigível é indicada, o decodificador pode reutilizar o último bloco de entrada livre de erros conhecido no lugar daquele danificado para encobrir o erro. A natureza do processo de reconstrução de sinal torna esse processo de encobrimento de erro relativamente benigno, e um bom bloco pode ser repetido

diversas vezes, se necessário, antes que condições de erro duradouras façam com que o decodificador ou se torne mudo, ou reverta para o uso de sistemas analógicos, como no caso de filmes de cinema.
- Extração dos dados em formato fixo: a extração é feita em dois estágios. Primeiro, os dados em formato fixo são extraídos, incluindo os expoentes e os coeficientes de acoplamento. As porções relevantes destes dados são então usadas pelo decodificador para recuperar a alocação de bits, a qual, por sua vez, é usada para extrair os dados de formato variável, em particular os arranjos de mantissas TDAC.

3.4.2.2. Alocação de Bits

A rotina aqui utilizada é idêntica àquela correspondente ao codificador, exceto pelo fato de que ela usa os resultados intermediários transmitidos, para poupar tempo. Este arranjo permite também que o decodificador compute a alocação de bits de um canal de cada vez, reduzindo a memória necessária. A alocação de bits do decodificador deve ser exatamente emparelhada com a do codificador, a fim de que os dados de formato variável sejam corretamente extraídos, caso contrário distorções podem ser introduzidas.

3.4.2.3. Dequantização da Mantissa

Este bloco consiste de dois diferentes processamentos, como mostrado a seguir.
- Extração dos dados de formato variável: a alocação de bits recuperada no decodificador, especificando o valor quantizado de cada mantissa, é usada para extrair os dados de formato variável a partir do arranjo de bits codificado.
- Conversão para ponto fixo: em preparação para a transformada TDAC inversa, os dados das mantissas e expoentes são combinados para reconstruir os coeficientes TDAC em ponto fixo.

3.4.2.4. Decodificação da Envoltória Espectral

Os coeficientes de alta freqüência que foram codificados como portadora e envoltória são reconstruídos pela combinação das portadoras com os coeficientes de acoplamento correspondentes.

3.4.2.5. Banco de Filtros de Análise (Transformada Inversa)

Os coeficientes da transformada TDAC recuperados para cada canal, sofrem uma transformação inversa para voltar ao domínio do tempo. Os coeficientes do “subwoofer” são feitos iguais a zero para as médias e altas freqüências antes da transformação inversa, de maneira que a saída no domínio do tempo do “subwoofer” esteja à taxa de amostragem completa.

3.4.3. Outras Características

Ao contrário do padrão MPEG-2, o AC-3 não faz uso de uma matriz de compatibilidade (adição e subtração de canais) para torná-lo funcional para sistemas estéreo de dois canais. Seu uso foi evitado devido à sua característica indesejável de, em alguns casos, fazer com que o ruído de quantização associado a determinado canal seja decodificado num canal diferente daquele do próprio sinal, fazendo com que o ruído não seja mascarado. O processo de codificação usado no AC-3 inerentemente preserva a codirecionalidade dos sinais e ruídos de quantização correspondentes, a fim de manter as

características de mascaramento. Para situações em que não há alto-falantes suficientes para representar todos os 5.1 canais, o decodificador pode fazer uma redução dos 5.1 canais para o número requerido de canais de saída sem o uso de uma matriz de compatibilidade.

3.5. SONY ATRAC

Este padrão foi introduzido como o sistema de codificação usado nos MiniDiscs. Existem diversas gerações do codec. Ele opera exclusivamente à taxa de 44,1 kamostras/s [40].

3.5.1. Codificador

Como mostrado na Figura 3.6, o sistema ATRAC usa tanto a codificação por sub-bandas como por transformada. A fim de facilitar o processamento dos sinais, estes são divididos em três bandas (0 Hz a 5.512,5 Hz; 5.512,5 Hz a 11.025 Hz e 11.025 Hz a 22.050 Hz) [66,67]. A decomposição em sub-bandas é feita usando a Filtragem por Quadratura Espelhada (QMF), a qual garante que o "aliasing" causado pela decomposição será cancelado durante a reconstrução [47,68].

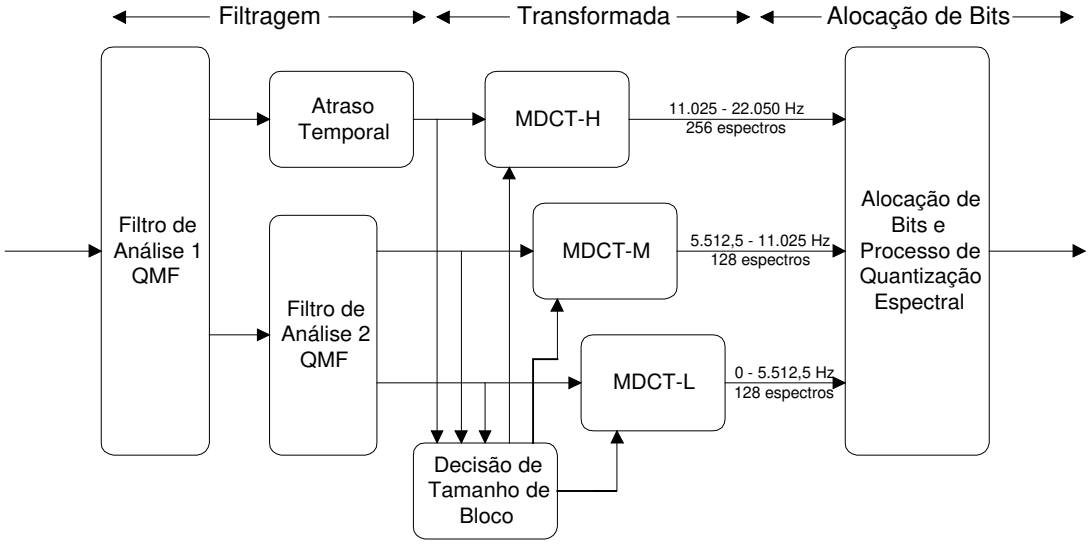


Figura 3.6 - Diagrama de blocos do codificador ATRAC [69].

A seguir, as três bandas são transformadas para o domínio da freqüência através de uma transformada discreta co-seno modulada (MDCT), em substituição à FFT comum. A MDCT permite até 50% de superposição entre os blocos, levando a uma melhor resolução espectral ao mesmo tempo em que mantém a mesma freqüência de amostragem [68].

O tamanho do bloco é alterado de acordo com a natureza do sinal, sendo esta a parte adaptativa do algoritmo ATRAC. Para explicar a motivação para esta característica, imagine-se um transiente de curta duração em relação ao tamanho do bloco. Devido à quantização das componentes espectrais, os erros de quantização serão espalhados por todo o domínio do tempo. Como o mascaramento temporal retrógrado é bastante limitado, este ruído não será mascarado pelo sinal. Este problema é geralmente chamado de pré-eco.

Para evitar os efeitos de pré-eco, o tamanho do bloco pode ser reduzido de maneira que, ainda que o ruído de quantização seja distribuído ao longo de todo o bloco, este é

suficientemente pequeno para garantir o mascaramento. Assim, o ATRAC permite uma codificação eficiente em regiões estacionárias enquanto responde rapidamente a passagens transientes. O tamanho do bloco pode ser selecionado individualmente para cada banda.

Na segunda geração do algoritmo ATRAC, o problema do pré-eco é abordado através da amplificação adaptativa do sinal que precede um ataque (variação brusca do sinal), antes que se realize a MDCT e, após a realização da transformada inversa, o nível original é restaurado [70].

O ATRAC divide o espectro em 52 bandas de diferentes larguras, as quais são divididas de acordo com a definição das bandas críticas. Este número elevado de bandas foi adotado para compensar o fato de que no ATRAC a largura das bandas é fixa, enquanto que as bandas críticas possuem uma alocação dinâmica de acordo com a freqüência central. Esta estratégia é também adotada no Dolby AC-3 [71].

Os valores espectrais resultantes da MDCT são quantizados usando uma notação em ponto flutuante. Os valores são agrupados em Unidades Flutuantes de Blocos (BFU). Cada BFU usa o mesmo fator de escala e o mesmo comprimento de palavra-código. O fator de escala usa 6 bits e é determinado a partir de uma lista fixa de possibilidades, fornecendo uma faixa dinâmica entre –120 dB e +6 dB [66]. O tamanho da palavra-código expressa o número de bits usados pelas mantissas dos valores, e é determinado pelo algoritmo de alocação de bits.

O ATRAC não especifica nem um algoritmo para alocação de bits, nem um modelo psico-acústico. O comprimento das palavras-código de cada BFU é armazenado no arranjo de bits juntamente com os valores espectrais quantizados, de maneira que o decodificador é totalmente independente do algoritmo de alocação de bits. Isto abre a possibilidade de se melhorar o algoritmo de alocação de bits sem que se necessite mudar toda a base de decodificadores existente. Ainda assim, um algoritmo simples é sugerido na referência [68].

3.5.2. Decodificador

O processo de decodificação do ATRAC (Figura 3.7) é bem mais simples que o processo de codificação. O decodificador começa obtendo os BFUs a partir do arranjo de bits, a fim de reconstruir os valores espectrais usando os fatores de escala e as palavras-código. Esses valores espectrais são inicialmente transformados para o domínio do tempo por uma MDCT inversa usando o tamanho de bloco apropriado. Por fim, os três sinais no domínio do tempo são sintetizados em um sinal de saída pelos filtros de síntese QMF.

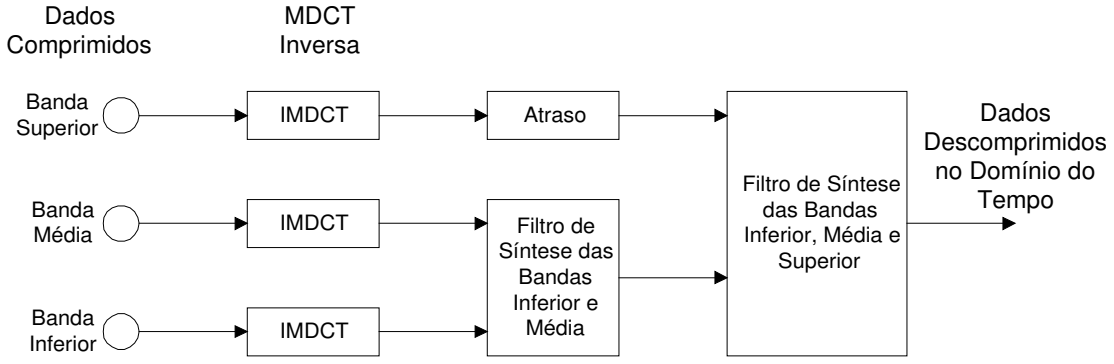


Figura 3.7 - Esquema de decodificação do algoritmo ATRAC.

3.6. DISCUSSÃO

O codec Sony ATRAC tem-se mostrado uma boa solução para sistemas MiniDisc. Isso se deve ao fato de que, tendo-se em mente a portabilidade, necessita-se de decodificadores simples de se implementar e com uma baixa complexidade, o que é alcançado com sucesso por este codec. Por essa razão, o ATRAC não tem competidores neste tipo de aplicação.

Existe hoje uma forte disputa entre os padrões MPEG-2 e AC-3. Não existem muitas referências que apresentem uma comparação imparcial entre estes dois padrões. No entanto, através da coleta de diversas opiniões e depoimentos, chega-se à conclusão de que o AC-3 possui uma certa superioridade em relação ao MPEG-2 BC, o que pode ter motivado o desenvolvimento do padrão MPEG-2 AAC, o qual não apresenta os inconvenientes da matriz de compatibilidade. Segundo seus autores, esta versão deve apresentar um melhor desempenho que o AC-3 por apresentar uma melhor resolução espectral, porém ainda não há nenhuma referência que corrobore esta afirmação para o caso multicanal. Para o caso estéreo (dois canais), tal alegação é confirmada através de alguns estudos comparativos realizados [72].

CAPÍTULO 4
AVALIAÇÃO SUBJETIVA DE SISTEMAS DE ÁUDIO

Apesar dos grandes avanços experimentados nos últimos anos, o uso de medidas objetivas na avaliação da qualidade de sistemas de áudio ainda se mostra restrito a poucas situações. Por esse motivo, a utilização de medidas subjetivas ainda é a mais indicada para a maior parte dos casos, apesar das grandes dificuldades e custos envolvidos na sua realização.

Com o desenvolvimento de novos sistemas de áudio digital, os quais exploram as propriedades psicoacústicas da audição humana, e o surgimento de sistemas de som de múltiplos canais, depara-se com a necessidade de se desenvolver novos métodos subjetivos de avaliação que sejam capazes de lidar com estas novas tecnologias. Portanto, apesar da larga experiência na utilização de medidas subjetivas de avaliação de áudio, ainda há a busca contínua por aperfeiçoamentos que atendam às novas necessidades.

Neste capítulo serão apresentados, resumidamente, os principais procedimentos a serem adotados na realização das medidas subjetivas e os principais tipos de testes recomendados para cada tipo de sistema, de acordo com a Rec. ITU-R BS.1116-1 [73].

4.1. ESQUEMA EXPERIMENTAL

Na avaliação subjetiva de sistemas de áudio com pequenas degradações, deve-se fazer uso de métodos padronizados. Experimentos subjetivos são caracterizados pelo controle e manipulação das condições experimentais e pelos dados quantitativos fornecidos por observadores humanos.

Um planejamento e esquematização cuidadosos são necessários para evitar que fatores não controlados contaminem o teste de audição, evitando desta maneira ambigüidades. Por exemplo, se a seqüência de trechos de áudio em um teste de audição é idêntica para todos os ouvintes, então pode não ser possível determinar se os julgamentos feitos pelos avaliadores foram devidos àquela seqüência, ao invés de serem devidos às degradações que foram apresentadas. Então, as condições de teste devem ser arranjadas de maneira a revelar os efeitos dos fatores independentes, e somente estes.

Em situações em que se espera que as degradações potenciais e outras características sejam distribuídas homogeneamente através do teste de audição, um caráter verdadeiramente aleatório pode ser aplicado à apresentação das condições de teste sem a necessidade de cuidados especiais. Já em situações onde uma não-homogeneidade é esperada, isto deve ser levado em conta na apresentação das condições de teste. Por exemplo, nos casos em que o material a ser avaliado varia no nível de dificuldade de audição, a ordem de apresentação do estímulo deve ser distribuída aleatoriamente, tanto dentro de cada seção, quanto entre elas, evitando-se, tanto quanto possível, que sinais de um mesmo tipo estejam de alguma maneira agrupados.

Similarmente, testes de audição devem ser projetados de maneira a evitar que os ouvintes sejam saturados até o ponto em que haja a perda da precisão de julgamento. Exceto em casos que a relação entre audição e visão é importante, é preferível que a avaliação de sistemas de áudio seja feita sem o acompanhamento de figuras.

Uma consideração importante é a inclusão de condições de controle apropriadas. Tipicamente, tais condições incluem a apresentação de materiais de áudio não-degradados, introduzidos de maneira imprevisível para os ouvintes. São as diferenças entre o julgamento destes estímulos de controle e aqueles potencialmente degradados que garantem que as avaliações estão de fato relacionadas às degradações.

4.2. SELEÇÃO DOS AVALIADORES

4.2.1. Ouvintes Especializados

É importante que os julgamentos dos testes de audição, usados na avaliação de pequenas degradações em sistemas de áudio, sejam fornecidos exclusivamente por ouvintes que tenham experiência na detecção dessas degradações mais tênues. Quanto mais alta a qualidade desejada para o sistema, mais importante é esta condição.

4.2.2. Critérios para Seleção dos Avaliadores

A realização de testes subjetivos de sistemas de som com pequenas degradações, utilizando um grupo selecionado de avaliadores, tem como objetivo primário investigar se um grupo de ouvintes especializados, sob certas condições, é capaz de perceber degradações relativamente súbitas e também de produzir uma estimativa quantitativa das degradações introduzidas.

Às vezes, há a necessidade de se introduzir uma técnica de rejeição antes (*pre-screening*) ou depois (*post-screening*) do teste real. Em alguns casos, ambos os tipos de rejeição devem ser usados. Aqui, a eliminação refere-se a um processo onde todos os julgamentos de um ouvinte em particular são omitidos. Qualquer tipo de técnica de rejeição que não seja cuidadosamente analisada e aplicada pode levar a um resultado polarizado. Então, é extremamente importante que, sempre que a eliminação de dados tenha sido realizada, haja uma descrição clara do critério aplicado, a fim de que o leitor possa fazer seu próprio julgamento.

4.2.3. Número de Avaliadores

O número adequado de avaliadores pode ser determinado se a variância das avaliações puder ser estimada e a resolução requerida para o experimento for conhecida.

Quando as condições de um teste de audição são firmemente controladas, tanto comportamental quanto tecnicamente, tem-se observado que 20 avaliadores são, na maior parte das vezes, suficientes para se alcançar conclusões seguras. É aconselhável que a análise dos dados seja feita à medida em que o teste é realizado, de maneira que, ao se alcançar um nível de significância estatística suficiente para o teste em questão, o processo possa ser interrompido. Tal procedimento visa reduzir custos e tempo demandado.

Se houver interesse em se ter um alto grau de consistência entre os dados fornecidos pelos avaliadores, então se deve utilizar um grande número de ouvintes para garantir uma quantidade adequada de rejeição.

O número de avaliadores não é importante apenas para a resolução desejada. O resultado para um tipo de experimento é, em princípio, válido somente para o grupo de avaliadores especificamente indicados para tal teste. Então, o aumento do número de ouvintes pode resultar em um grupo de avaliação mais geral, e com isto os resultados podem ser considerados mais convincentes. O número de avaliadores pode também ser aumentado para levar em conta a variação, entre os ouvintes, da sensibilidade a diferentes distorções.

4.3. MÉTODO DE TESTE

Na avaliação subjetiva da qualidade de sistemas de áudio, todos os testes são realizados através da comparação entre um par de sinais, diferentemente do que se observa para a avaliação subjetiva da qualidade de voz, onde os testes mais utilizados são os do tipo absoluto [74]. Isto se deve ao fato das degradações encontradas nos sinais de áudio processados serem muito mais tênues e, por essa razão, mais difíceis de serem percebidas pelos avaliadores. A comparação entre os sinais original e decodificado aumenta consideravelmente a sensibilidade do ouvinte a qualquer distorção que venha a corromper o sinal.

Para conduzir avaliações subjetivas em sistemas que geram pequenas degradações, é necessário selecionar um método apropriado. O método de "triplo estímulo, duplamente cego e com referência escondida" tem se mostrado especialmente sensível e estável, além de permitir uma detecção apurada de pequenas degradações, sendo então apropriado para este tipo de teste. O termo "triplo estímulo, duplamente cego e com referência escondida" significa que, além do sinal referência conhecido, o teste é composto ainda por dois sinais a serem avaliados, sendo um deles idêntico à referência. Este método envolve um ouvinte de cada vez, e a seleção de um de três estímulos ("A", "B", "C") fica a cargo deste ouvinte. A referência conhecida (sinal original) está sempre disponível como estímulo "A". A referência oculta (novamente correspondendo ao sinal original) e o objeto (sinal processado) estão disponíveis simultaneamente, mas aleatoriamente distribuídos entre "B" e "C", dependendo do processo.

Pede-se ao ouvinte para avaliar as degradações de "B" e "C" comparados com "A", de acordo com a escala de degradação mostrada na Tabela 4.1. Um dos estímulos, "B" ou "C", deve ser indiscernível do estímulo "A"; o outro pode revelar degradações. Quaisquer diferenças percebidas entre a referência e o outro estímulo devem ser interpretadas como degradações.

A escala de graduação é contínua, com pontos de referência (normalmente valores inteiros entre 1 e 5) derivados da escala encontrada na Recomendação ITU-R BS.1284 [75], como mostrado na Tabela 4.1.

Tabela 4.1 - Escala de Graduação de Degradação

Degradação	**Graduação**
Imperceptível	1
Perceptível, mas não incômoda	2
Ligeiramente incômoda	3
Incômoda	4
Muito incômoda	5

Se pontos de referência intermediários não são usados, é essencial que os resultados para ouvintes individuais sejam normalizados com respeito à média e ao desvio padrão. A equação

$$Z_i = \frac{\left(x_i - x_{si}\right)}{s_{si}} \cdot s_s + x_s \tag{4.1}$$

pode ser usada para obter tal normalização, enquanto se mantém a escala original. Na Equação 4.1, Z_i é o resultado normalizado, x_i é a avaliação do ouvinte i, x_{si} é a média para o ouvinte i na seção s, x_s é a média para todos os ouvintes na seção s, s_s é o desvio padrão para todos os ouvintes na seção s e s_{si} é o desvio padrão para o ouvinte i na seção s.

Recomenda-se que a escala seja utilizada com uma resolução de uma casa decimal.

O método de teste consiste de duas partes: uma fase de familiarização ou treinamento, e uma fase de avaliação.

4.3.1. Fase de familiarização ou treinamento

Antes da avaliação formal, os ouvintes devem estar perfeitamente familiarizados com os equipamentos de teste, com o ambiente de teste, com o processo de avaliação, com as escala de graduação e os métodos para seu uso. Devem também estar familiarizados com os dispositivos sob estudo. Para os testes mais sensíveis, eles devem ser expostos a todo o material que será posteriormente avaliado por eles nas seções de avaliação formal. Durante a familiarização ou treinamento, os ouvintes devem ser agrupados (por exemplo, em grupos de três pessoas), permitindo sua livre interação e discussão a respeito das degradações que eles venham a detectar. Se realizada corretamente, a familiarização pode transformar alguns ouvintes, com pouca habilidade inicial, em especialistas para as propostas do teste.

4.3.2. Fase de avaliação

No início da primeira seção de avaliação formal do dia, deve-se fazer, para cada ouvinte, uma apresentação oral das instruções do teste, preferivelmente suplementada por material escrito. Várias comparações ilustrativas podem ser apresentadas logo antes do início das apresentações para avaliação formal.

Como a memória auricular de média e longa duração não é confiável, o procedimento de teste deve trabalhar exclusivamente com a memória de curta duração. Para isso, recomenda-se o uso de um método de chaveamento quase instantâneo entre os estímulos. O chaveamento instantâneo não é aconselhável, pois este pode introduzir distorções se as formas de onda de estímulos sucessivos não são idênticas. Por isso, deve-se preferir um chaveamento com um tempo total de 40 ms.

Para as avaliações mais críticas, deve-se trabalhar com um ouvinte de cada vez. Somente desta maneira o avaliador pode ter uma completa liberdade individual para escolher entre os estímulos em um método de estímulo triplo. Tal liberdade é essencial para que o ouvinte possa explorar completamente as comparações detalhadas entre os estímulos para cada processo. É desejável que o avaliador seja capaz de escolher entre os estímulos sem orientação visual, podendo então, se ele assim preferir, manter os olhos fechados para uma melhor concentração. Nenhuma distorção deve ser produzida pelo sistema de chaveamento (como “cliques”), já que essas degradações podem interferir seriamente no processo de avaliação.

Uma seção de avaliação não deve durar mais do que 20 ou 30 minutos. A experiência mostra que não mais que 10 a 15 julgamentos devem ser agendados por seção. A fadiga dos avaliadores pode se tornar um fator importante, podendo interferir seriamente na validade dos julgamentos. Para evitar isto, períodos de descanso não inferiores à duração da seção de avaliação devem ser agendados, entre seções sucessivas, para cada ouvinte.

4.4. MATERIAL DE TESTE

Somente material crítico deve ser utilizado na determinação de diferenças entre sistemas sob teste. Material crítico é aquele que exige o maior esforço do sistema sob teste. Não há um material de teste "universal", que possa ser usado para avaliar todos os sistemas sob todas as condições. Assim, o material crítico de teste deve ser especificamente determinado para cada sistema a ser testado em cada experimento. A busca por um material adequado é, em geral, bastante demorada; porém, a menos que um material verdadeiramente crítico seja determinado, os experimentos falharão na tentativa de revelar diferenças entre sistemas, tornando-se dessa maneira inconclusivos.

Antes de se aceitar como sendo válido um caso "nulo" (nenhuma diferença entre os sistemas), deve-se mostrar, empírica ou estatisticamente, que qualquer falha na detecção de diferenças entre sistemas não é devida a uma insensibilidade experimental, causada por uma escolha inadequada do material de áudio, ou quaisquer outros aspectos falhos do experimento. No caso extremo em que vários ou todos os sistemas mostram uma qualidade perfeitamente transparente, pode ser necessário programar testes especiais com "âncoras" (pontos de referência pouco espaçados introduzidos especialmente para o teste em questão), com a proposta explícita de examinar a capacitação dos ouvintes. Estas âncoras devem ser introduzidas de modo a serem detectáveis por ouvintes especializados, mas não por ouvintes comuns. Elas são introduzidas não somente para checagem da capacitação dos avaliadores, mas também para checar a sensibilidade de todos os outros aspectos da situação experimental.

Se estas âncoras são corretamente identificadas por todos os ouvintes em um teste padrão, isto pode ser usado como evidência de que a capacitação dos avaliadores é aceitável e que não há problemas de sensibilidade em outros aspectos da situação experimental.

Por outro lado, se estas âncoras não são corretamente identificadas pelos ouvintes, isto sugere que ou eles não têm a experiência necessária, ou há falhas de sensibilidade na situação, ou ambos.

Na busca pelo material crítico, sinais sintéticos designados especificamente para um determinado sistema devem ser evitados. O conteúdo artístico ou intelectual de uma seqüência utilizada nos testes não deve ser excessivamente atrativo nem muito desagradável, a fim de que os avaliadores não se desviem do objetivo de detectar as diferenças entre os sistemas.

O nível utilizado para a gravação dos sinais deverá ser de -18 dB_{FS} (ver Apêndice A). Os trechos de áudio deverão ter duração entre 10 e 25 segundos.

4.5. DISPOSITIVOS DE REPRODUÇÃO

Os alto-falantes ou fones de ouvido de referência devem ser escolhidos com o objetivo de que todos os sinais de teste possam ser reproduzidos de uma maneira ótima; não devem apresentar nenhum tipo de distorção para qualquer tipo de reprodução e devem ser utilizáveis tanto para uma avaliação monofônica quanto para sistemas de som de dois ou mais canais.

Certas perdas de qualidade são mais claramente perceptíveis no caso de uma reprodução através de fones de ouvido, enquanto outras são mais claramente perceptíveis no caso de uma reprodução através de alto-falantes. Então, é necessário que se determine o tipo dispositivo de reprodução apropriado, através de pré-testes subjetivos.

Especialmente para casos em que as distorções afetarão as características da "imagem" do som estereofônico, deve-se utilizar a reprodução através de alto-falantes. Para a avaliação de sistemas de som estereofônicos de dois canais, o uso tanto de alto-falantes quanto de fones de ouvido pode ser necessário. Para a avaliação de sistemas monofônicos, um alto-falante central e/ou fones de ouvido podem ser utilizados.

Para a avaliação de sistemas de som multicanal, alto-falantes devem ser utilizados, particularmente se há o desejo de se avaliar influências em todos os canais de reprodução simultaneamente.

Em todos os casos, cada alto-falante deve ser acusticamente calibrado nas faixas de freqüência relevantes, a fim de que haja o mínimo de diferenças de timbre entre eles.

4.6. CONDIÇÕES DE AUDIÇÃO

O termo "condições de audição" descreve os atributos acústicos que afetam a percepção sonora de um ouvinte em uma sala acústica, em um determinado ponto de audição de referência, para o som reproduzido por alto-falantes. Isto inclui:

- as características acústicas da sala onde será realizada a audição;
- o arranjo dos alto-falantes na sala de audição;
- a localização da área ou ponto de referência de audição;

Estes três atributos são os responsáveis pela produção das características do campo sonoro resultante naquele ponto ou área.

4.7. ANÁLISE ESTATÍSTICA

O objetivo fundamental da análise estatística dos resultados dos testes é identificar cuidadosamente o desempenho médio de cada um dos sistemas sob teste e a confiabilidade de quaisquer diferenças encontradas entre os desempenhos médios obtidos. Este último aspecto requer a estimação da variância dos resultados (método da análise da variância – ANOVA) [73]. Embora este valor não proscreva ou prescreva qualquer método estatístico particular, se as suposições relativas às estatísticas paramétricas são razoáveis, então esta abordagem pode ser considerada a mais sensível e poderosa, sendo por isso a mais recomendada. Outros métodos estatísticos devem ser utilizados somente em casos particulares, em que a análise da variância não é suficiente. Especificamente, recomenda-se que se aplique o modelo ANOVA como um primeiro estágio. Subseqüentemente, outros métodos, utilizando as estimativas de variância fornecidas pelo ANOVA, podem ser usados para estudar com mais detalhes os efeitos gerais revelados pelo ANOVA.

4.8. OUTROS ASPECTOS IMPORTANTES

A Rec. ITU-R BS.1116-1 aborda ainda outros aspectos que devem ser levados em consideração na realização das medidas subjetivas de qualidade de áudio, como apresentação da análise estatística, conteúdo dos relatórios de teste, detalhes sobre os dispositivos de reprodução utilizados nos testes (alto-falantes e fones de ouvido) e as condições de audição, onde são abordadas as condições requeridas para a sala de audição, as condições do campo sonoro, nível de audição, arranjo dos dispositivos de reprodução em torno do ouvinte, etc. Tais itens não serão aqui explorados, por fugirem ao escopo principal do presente trabalho, porém mais detalhes podem ser obtidos em [73,76,77].

CAPÍTULO 5

MEDIDAS OBJETIVAS DE AVALIAÇÃO DA QUALIDADE DE ÁUDIO

A avaliação da qualidade de áudio é uma das principais etapas no desenvolvimento de um sistema digital de comunicação. As medições subjetivas de qualidade são excessivamente dispendiosas, tanto em termos de custos quanto de tempo consumido. Por esse motivo, é natural que se busque sua substituição por medidas objetivas capazes de modelar satisfatoriamente o comportamento dos ouvintes em avaliações subjetivas. Os métodos objetivos tradicionais, como a relação sinal-ruído (SNR) e a distorção harmônica total (THD), não apresentam resultados satisfatórios, por não levarem em conta as diversas características peculiares à audição humana. Os problemas se tornam ainda mais evidentes quando tais métodos são aplicados aos modernos codecs perceptuais, os quais introduzem, deliberadamente, distorções a serem mascaradas pelas componentes do sinal original.

Dessa forma, em 1994 a ITU-R fez uma chamada aberta de propostas, a fim de estabelecer um padrão para a medição objetiva da qualidade de áudio. Seis métodos foram apresentados: Índice de Distúrbio (*Disturbance Index* - DIX) [78], Taxa Ruído-Mascaramento (*Noise-to-Mask Ratio* - NMR) [79], Medida Perceptual da Qualidade de Áudio (*Perceptual Audio Quality Measure* - PAQM) [80], Avaliação Perceptual (*Perceptual Evaluation* - PERCEVAL) [81], Medida Objetiva Perceptual (*Perceptual Objective Measure* - POM) [82] e Abordagem da Caixa de Ferramentas (*Toolbox Approach*) [83]. Tais métodos serão brevemente descritos mais adiante. Apesar dos esforços realizados pelos diversos grupos de pesquisa, nenhuma das propostas apresentou o desempenho mínimo desejado, levando à conclusão de que bons resultados só poderiam ser alcançados através de um processo colaborativo, onde todos contribuiriam para o desenvolvimento de um único método capaz de atender às exigências impostas. Como resultado, surgiu o método "Avaliação Perceptual da Qualidade de Áudio" (*Perceptual Evaluation of Audio Quality* - PEAQ) [83], o qual será brevemente descrito na seção 5.2.9.

Antes da descrição dos métodos objetivos de avaliação da qualidade de áudio propriamente ditos, serão apresentados alguns conceitos essenciais para a determinação dos modelos básicos usados no desenvolvimento da maioria das medidas objetivas perceptuais propostas.

5.1. CONCEITOS DE MODELOS PERCEPTUAIS

A estrutura básica comum a todas as medidas objetivas de qualidade de sinais de áudio é mostrada na Figura 5.1. O último bloco, representando o mapeamento para as medidas subjetivas desejadas, é opcional.

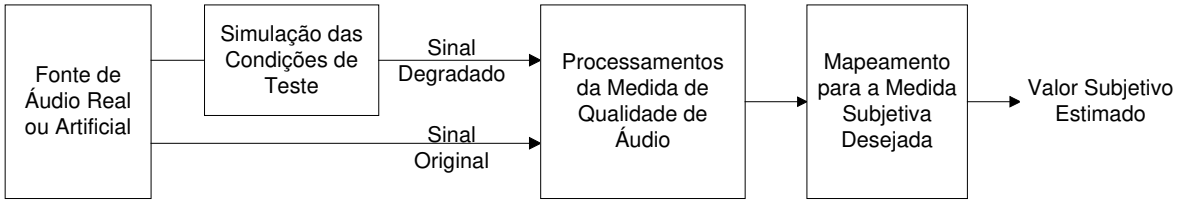


Figura 5.1 - Esquema Básico das Medidas Objetivas de Qualidade de Áudio.

Os métodos de avaliação objetiva de áudio mais bem sucedidos são baseados em conceitos extraídos da psico-acústica, a qual se ocupa do estudo do comportamento da audição. A percepção sonora humana pode ser grosseiramente descrita através de um esquema de cinco estágios, como se pode observar na Figura 5.2. O campo sonoro externo é transmitido para o ouvido interno e decomposto em componentes espectrais. A sensibilidade do ouvido e sua seletividade espectral são melhoradas por processos ativos, os quais normalmente incluem algum tipo de mecanismo de retroalimentação. As excitações neurais no ouvido interno são transmitidas para os centros auditivos do cérebro através do nervo auditivo, sendo então traduzidas em quantidades sensoriais. Os centros auditivos também possuem vários tipos de mecanismos de reconhecimento de padrões que podem influenciar a formação de quantidades sensoriais [51].

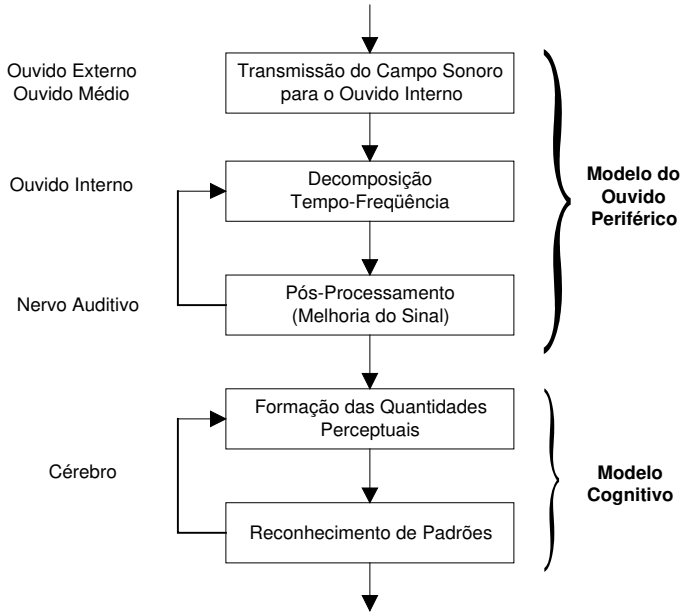


Figura 5.2 - Estágios do Processamento Auditivo.

Os primeiros três estágios na Figura 5.2 descrevem a tradução do campo sonoro externo para as excitações neurais (impulsos elétricos conduzidos pelos neurônios até a região específica do córtex cerebral), e os dois últimos estágios descrevem o processo de transformação destes padrões de excitação em sensações. A translação do campo sonoro externo para as excitações neurais é quase independente de preferências individuais, e representa a parte da percepção sonora primariamente baseada na estrutura fisiológica do sistema auditivo. Em um modelo perceptual, estes passos são chamados de *modelo do ouvido periférico*. Nos últimos estágios do processamento auditivo, as preferências individuais não podem ser claramente separadas das propriedades mais comuns do sistema

auditivo. Esses estágios, que incluem processos de reconhecimento de padrões e fluxo auditivo, são referidos como *modelo cognitivo*.

A percepção auditiva pode ser modelada por diferentes abordagens, as quais são, usualmente, compromissos entre dois conceitos extremos:

- *modelos funcionais* dos processos fisiológicos que ocorrem nos sistemas auditivos, os quais são baseados em medidas de propriedades acústicas e mecânicas do ouvido, bem como em medições das excitações neurais no nervo auditivo e da atividade neural em regiões do cérebro que estão envolvidas no processamento auditivo.

- *modelos heurísticos* de propriedades observadas para o sistema auditivo, os quais são inteiramente baseados em resultados de testes de audição; a estrutura de tais modelos não precisa necessariamente corresponder à estrutura do processamento auditivo humano.

Em teoria, os modelos funcionais podem se aproximar do ideal tanto quanto se deseje para todos os fenômenos auditivos possíveis. Por outro lado, tal modelo normalmente requer um esforço computacional extremamente elevado. Além disso, não se pode esperar um conhecimento suficientemente detalhado de todos os processos fisiológicos incorporados na percepção auditiva.

Os modelos heurísticos são mais práticos e requerem um menor esforço computacional, uma vez que um menor número de fenômenos é modelado. Além disso, eles podem fornecer o máximo de precisão para aspectos particulares da audição, uma vez que ela é diretamente derivada do fenômeno que se deseja reproduzir. Por outro lado, quanto maior o número de fenômenos que se quer modelar, maior a complexidade. Além disso, fenômenos auditivos diferentes podem levar a modelos contraditórios. Então, a maioria dos modelos perceptuais é baseada parte em processos fisiológicos, e parte em fenômenos que podem ser observados através de testes de audição. Dentre as três principais abordagens adotadas na implementação de um modelo do ouvido humano (seções 5.1.1 a 5.1.3), a *comparação de representações internas* está mais próxima a um modelo funcional, enquanto o *limiar de mascaramento* está mais próximo a um modelo heurístico, como será visto a seguir. As três abordagens mencionadas possuem o mesmo objetivo, que é fornecer uma medida da quantidade de distorção audível existente no sinal degradado em comparação com o sinal original; a escolha entre elas no desenvolvimento de uma medida objetiva de avaliação da qualidade de áudio dependerá das características idealizadas para o método e da preferência pessoal do projetista.

5.1.1. Limiar de Mascaramento

O *limiar de mascaramento* tem sido usado em diversos métodos perceptuais [70,79]. Nestas aplicações, o sinal de erro, que é a diferença entre as representações espectrais dos sinais original e degradado, é comparado ao limiar de mascaramento produzido pelo sinal original, conforme a Figura 5.3. Tal procedimento indicará quais componentes desse sinal diferença serão audíveis e quais estarão mascarados. Isto resulta em uma medida da quantidade de distorção audível do sinal degradado, sendo este o valor usado para a determinação da qualidade do sinal sob teste.

A principal vantagem do uso deste conceito é a possibilidade de derivar parâmetros que modelem a diferença entre os sinais diretamente a partir de experimentos de mascaramento. Além disso, tal procedimento pode ser aplicado à codificação de áudio sem modificações importantes no equacionamento original. Por outro lado, a possibilidade de se usar esta abordagem para modelar fenômenos auditivos mais complexos, visando um modelo perceptual mais preciso, é bastante limitada.

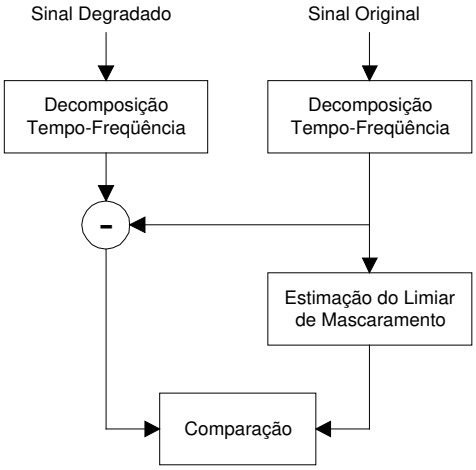


Figura 5.3 - Uso do limiar de mascaramento.

5.1.2. Comparação entre as Representações Internas

O conceito de *comparação das representações internas* foi introduzido em [84], e é usado na maioria dos métodos perceptuais atuais [80,81,82]. É baseado no cálculo dos padrões de excitação dos sinais original e degradado, os quais correspondem às representações internas encontradas na audição humana. As propriedades das distorções (audibilidade, sonoridade, incômodo etc.) são estimadas pela comparação entre estes padrões de excitação. Este conceito é mais próximo de um modelo funcional do sistema auditivo do que o conceito de limiar de mascaramento e, portanto, é um ponto de partida melhor para a modelagem de fenômenos auditivos mais complexos. A Figura 5.4 mostra o esquema básico desta abordagem.

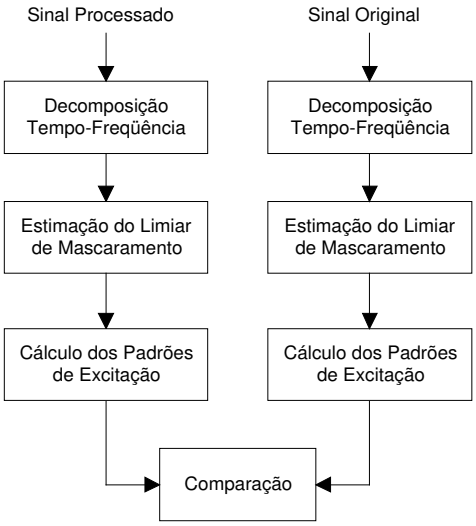


Figura 5.4 - Comparação das representações internas.

A estimação do limiar de mascaramento também é necessária neste modelo, pois o cálculo das representações internas dos sinais deve levar em consideração quais distorções serão ou não mascaradas.

5.1.3 Análise do Espectro Linear do Erro

Alguns efeitos, como, por exemplo, a percepção da freqüência fundamental, são mais facilmente modeláveis quando se utiliza o espectro linear ao invés de um modelo da membrana basilar. Tal abordagem é conhecida como *análise espectral de erros* [25]. De uma maneira geral, este conceito não pode ser usado em modelos baseados em bancos de filtros, mas somente em modelos baseados em transformadas. Ainda assim, pode ser um complemento útil quando implementado em paralelo com os modelos anteriores, uma vez que ele fornece informação adicional a respeito de algumas características difíceis de serem obtidas a partir das outras abordagens. Tal informação adicional pode ser, por exemplo, usada em conjunto com o restante da informação como a entrada para uma rede neural que faça o mapeamento de toda essa informação para um único valor representando a distorção encontrada no sinal degradado.

5.2. MEDIDAS PERCEPTUAIS

Uma característica comum a todos os métodos perceptuais é a modelagem dos efeitos de mascaramento. O mascaramento simultâneo é sempre modelado pela aplicação de uma função de espalhamento, que corresponde ao formato de uma curva de mascaramento média. Efeitos de mascaramento temporal são freqüentemente modelados implicitamente nas expressões utilizadas no modelo, mas de maneira muito grosseira, devido à resolução temporal limitada da decomposição tempo-freqüência normalmente utilizada.

Como se sabe, o aumento da resolução para análise em um dos domínios faz com que haja perda de resolução no outro; como o mascaramento simultâneo tem, em relação ao mascaramento temporal, uma importância maior na codificação de áudio, normalmente prefere-se, para este tipo de aplicação, uma maior resolução no domínio da freqüência, em detrimento da resolução temporal.

As degradações são estimadas pela comparação entre os sinais degradado e original. Os sinais de teste a serem utilizados são usualmente os mesmos trechos musicais usados na avaliação subjetiva de codecs. Porém, em princípio, qualquer tipo de sinal de áudio, incluindo aqueles criados artificialmente, pode ser usado. Os sinais, antes de serem submetidos aos processamentos, são normalmente divididos em quadros. Para sinais de áudio amostrados a 48 kHz, normalmente são utilizados quadros compostos por 2.048 amostras.

O ponto de partida para a maioria dos métodos perceptuais é um filtro ou uma função de ponderação, usados para modelar a função de transferência dos ouvidos externo e médio, seguidos por uma decomposição (ou transformação) tempo-freqüência a curto termo (quadro-a-quadro). A escala de freqüência é então transformada, a fim de levar em conta a resolução espectral não-uniforme do sistema auditivo, como mostra a Figura 5.5, onde a escala em Hz foi transformada para uma escala em Bark. O principal efeito desta transformação, além de simular a resolução perceptual de freqüência (modelando a membrana basilar), é que o formato dos filtros auditivos que realizam a divisão em bandas perceptuais se torna quase uniforme, podendo então ser modelado por uma simples convolução desse espectro transformado com uma função de espalhamento (maiores detalhes podem ser encontrados ao longo da Seção 2.2). Essa função de espalhamento modela a maneira como a excitação causada por uma determinada componente espectral

será distribuída em torno do ponto correspondente na membrana basilar. O resultado de tal procedimento é uma curva que determinará o limiar de mascaramento ao longo de todo o espectro para o quadro que está sendo processado. A Figura 5.6 mostra, como exemplo, o resultado da convolução da função de espalhamento com 4 diferentes componentes espectrais, cada qual com um determinado nível. A linha contínua indica o limiar de mascaramento obtido levando-se em conta a contribuição de cada um dos componentes considerados.

Quando o conceito de limiar de mascaramento descrito na Seção 5.1.1 é usado, a convolução citada é aplicada somente ao sinal original, pois, como visto, será o sinal diferença, e não o sinal degradado em si, que servirá como base para a derivação da quantidade de degradação. Em outras palavras, cada componente do sinal diferença obtido a partir dos sinais sem a aplicação do mascaramento é comparado ao limiar de mascaramento produzido pelo sinal original; se seu nível for maior, significará que a degradação representada por tal componente será audível e, portanto, incômoda para o ouvinte. Quando a comparação das representações internas é usada, esta convolução é aplicada a ambos os sinais, pois neste caso são os próprios sinais (com os respectivos limiares de mascaramento) que serão comparados (Seção 5.1.2).

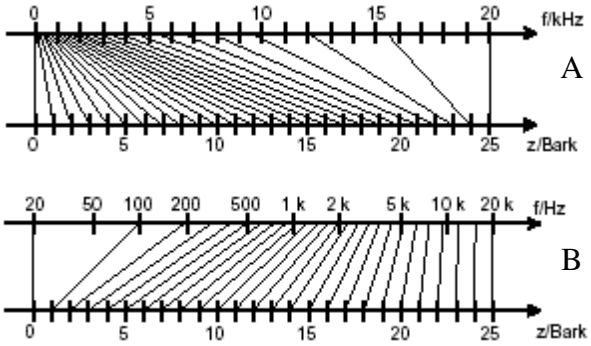


Figura 5.5 - Transformação de freqüência na modelagem de uma escala perceptual de freqüência (figura A: escala de freqüência linear; figura B: escala de freqüência logarítmica).

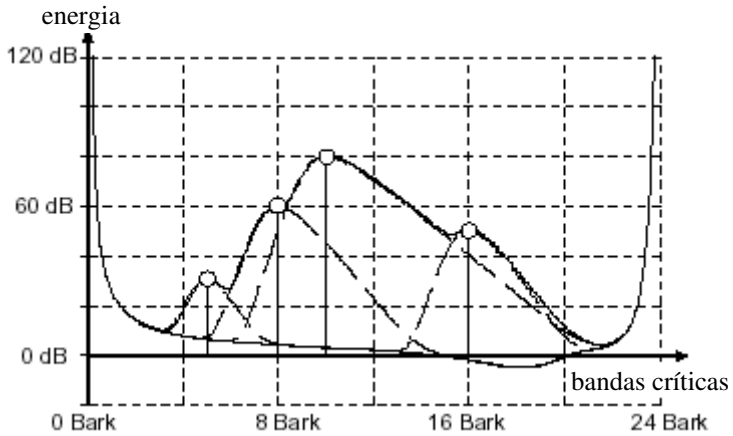


Figura 5.6 - Convolução de um espectro com uma função de espalhamento.

Se o mascaramento temporal é modelado explicitamente, um filtro passa-baixas de primeira ordem é utilizado para realizar uma ponderação exponencial temporal. Tal filtro modela a influência exercida por todos os componentes sobre determinada amostra

ponderando-os, de maneira que as amostras mais distantes tenham sua influência atenuada, e somando-os ao nível da amostra em questão.

As representações perceptualmente adaptadas dos sinais são comparadas umas às outras, a fim de se obter uma estimativa da severidade das degradações do sinal degradado. Os detalhes desta comparação formam as principais diferenças entre os métodos perceptuais existentes.

5.2.1. Primeiras Medidas Perceptuais para Avaliação da Qualidade de Áudio

O predecessor de todos os métodos perceptuais é o procedimento de Zwicker para o cálculo da sonoridade percebida (ver Seção 2.3.1). Implementações deste algoritmo em *hardware* foram feitas no começo dos anos 60, onde até mesmo a dependência do formato dos filtros auditivos em relação ao nível foi modelada. As primeiras aplicações para tais algoritmos foram na avaliação de codecs de voz, e não de áudio [57,84].

5.2.1.1. O Método Degradação de Sinal de Voz (Speech Signal Degradation)

Este método calcula a sonoridade dos erros encontrados no sinal degradado, levando em conta o mascaramento parcial produzido pelo sinal original [57]; tal método faz uso da abordagem descrita na Seção 5.1.1. A decomposição tempo-freqüência é realizada via FFT. Limiares de mascaramento são aproximados por uma função de espalhamento, a qual é derivada de uma curva de mascaramento medida para uma senóide de 1 kHz a um nível de pressão sonora de 80 dB SPL (ver definição de SPL no Apêndice A). Este é o primeiro método conhecido a usar o conceito de limiar de mascaramento.

5.2.1.2. Medida de Distância Espectral Auditiva (*Auditory Spectrum Distance*)

Este método calcula a diferença de sonoridade entre os sinais original e degradado [84]; a principal diferença deste método em relação ao anterior é que este faz uso da técnica descrita na Seção 5.1.2. A decomposição tempo-freqüência é realizada por um banco de filtros FIR, com uma resolução espectral de metade de uma banda crítica. Limiares de mascaramento são modelados diretamente pela resposta em freqüência dos filtros, os quais são projetados para aproximar a mesma curva de mascaramento usada em [57]. O mascaramento temporal é modelado por um filtro assimétrico não-linear, o qual possui uma melhor correspondência com os efeitos de mascaramento temporal do que os filtros FIR de primeira ordem normalmente utilizados [15]. Este foi o primeiro método a usar o conceito de comparação das representações internas.

A partir de 1987, com a introdução do sistema de medidas NMR [79], houve o desenvolvimento de uma grande quantidade de métodos perceptuais para a avaliação de codecs de áudio. Os mais importantes são brevemente descritos nas próximas subseções.

5.2.2. Índice de Distúrbio (DIX)

O método DIX (“Disturbance Index”) é baseado em um banco de filtros auditivos que fornece uma alta resolução temporal, permitindo uma melhor modelagem dos efeitos temporais, como o pré e o pós-mascaramento. A estrutura temporal fina das envoltórias na saída de cada filtro auditivo é preservada e usada para obter informação adicional a respeito dos sinais e distorções introduzidas.

As freqüências centrais dos filtros individuais são igualmente distribuídas sobre a escala de *pitch* perceptual. O topo das curvas dos filtros é ligeiramente arredondado, para garantir que o número escolhido de filtros cubra toda a faixa de freqüência sem a ocorrência de oscilações (*ripple*) na resposta em freqüência total. A fim de modelar os limiares de mascaramento, as inclinações das respostas dos filtros diminuem exponencialmente sobre a escala Bark. As inclinações dependem também do nível dos sinais. A faixa espectral audível é coberta por 40 filtros, representando uma resolução de aproximadamente 0,6 Bark. O algoritmo implementado neste método para o banco de filtros é bastante rápido quando comparado a bancos onde cada filtro é implementado individualmente, porém ainda demanda muito mais tempo que as transformadas baseadas em blocos, como a FFT e a transformada *Wavelet*.

O método DIX adapta dinamicamente os níveis e espectros entre os sinais original e degradado, com o objetivo de separar as distorções lineares das não-lineares. Ele avalia a estrutura das envoltórias temporais nas saídas dos filtros, a fim de modelar a grande quantidade de mascaramento causada pelos mascaradores modulados e do tipo ruído. Vários parâmetros de saída são calculados pela comparação das representações internas de ambos os sinais, incluindo a sonoridade parcial de distorções não-lineares, indicadores para a quantidade de distorções lineares e medidas para efeitos temporais e estéreo. Porém, uma boa estimação da qualidade básica de áudio pode ser alcançada usando-se apenas os dois primeiros parâmetros acima citados, os quais são então mapeados para uma estimativa da qualidade do sinal sob teste.

5.2.3. Relação Ruído-Mascaramento (NMR)

O sistema de medida *NMR* (*Noise-to-mask ratio*) computa explicitamente um sinal de erro, correspondendo à diferença absoluta entre os sinais original e degradado [79]. Este erro e o sinal original são analisados em 27 bandas de freqüência, usando o conceito de limiar de mascaramento. A estrutura geral deste método foi baseada em uma versão simplificada do método descrito em [57]. A curva de mascaramento foi modificada a fim de se obter o formato mais próximo possível das curvas de mascaramento dependentes do nível, determinadas a partir de experimentos psico-acústicos, sem modelar explicitamente essas dependências (“curva de mascaramento para o pior caso”, ver Subseção 2.4.2.4. e Figura 5.7). O pequeno número de bandas de freqüência, em conjunto com um modelo bastante simples do sistema auditivo, facilita a implementação deste método em tempo real, o qual se tornou, por esse motivo, o primeiro sistema de medida perceptual disponível comercialmente. Por outro lado, estas duas características limitam o desempenho deste método na sua utilização para a estimação da qualidade de áudio percebida. Porém, o sistema NMR foi, por muitos anos, usado com sucesso como uma ferramenta no desenvolvimento de codecs.

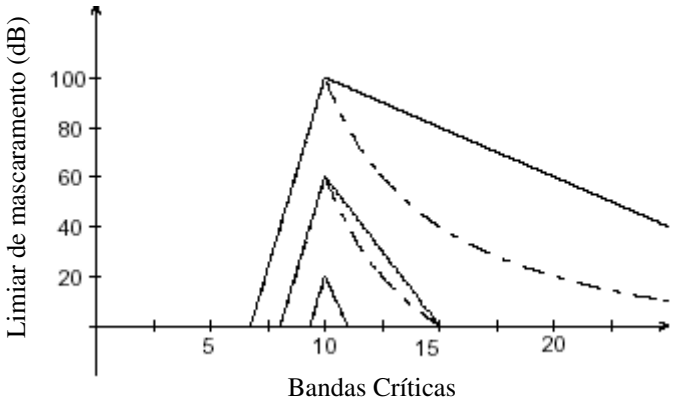


Figura 5.7 - Relação entre a curva de mascaramento para o pior caso (linhas tracejadas) usada no NMR e a as curvas de mascaramento dependentes do nível (linhas sólidas).

5.2.4. Medida Perceptual da Qualidade de Áudio (PAQM)

O método PAQM (*Perceptual Audio Quality Measure*) incorpora um modelo auditivo muito mais complexo que o NMR [80]. Ele modela a dependência de nível do mascaramento, bem como a aditividade não-linear entre diferentes componentes mascaradores e a assimetria entre mascaradores tonais e tipo ruído. Estes efeitos são modelados pela aplicação de uma função de potência às densidades de energia locais, antes das funções de mascaramento espectral e temporal serem aplicadas. O principal valor de saída do PAQM é o logaritmo do *distúrbio ruidoso*, que é a diferença de sonoridade entre os sinais original e processado, baseada no cálculo da sonoridade específica proposta por Zwicker [53].

Os expoentes usados na modelagem da assimetria de mascaramento e o expoente usado no cálculo da sonoridade específica foram ajustados experimentalmente, a fim de se obter a maior correlação possível entre as predições do modelo e os valores subjetivos dos sinais de áudio processados. Conseqüentemente, o PAQM mostrou melhores valores de correlação do que a maioria dos outros métodos perceptuais, mas com isso o modelo deixou de ser inteiramente relacionado com a psico-acústica.

O desempenho do PAQM como um estimador da qualidade subjetiva foi posteriormente melhorado pela modificação das estratégias de ponderação temporal e espectral. Estas melhorias, referidas como correção cognitiva [85] e fluxo perceptual [86], novamente são principalmente determinadas por otimização experimental. Para a medida de qualidade de codecs de voz, uma modificação do PAQM, o PSQM (*Perceptual Speech Quality Measure*) [87] tornou-se um padrão internacional [88] e serviu como base para o desenvolvimento do método MOQV [89,90,91]. Esta modificação, contudo, desvia-se ainda mais da psico-acústica tradicional, não modelando nem mesmo o mascaramento simultâneo.

5.2.5. Avaliação Perceptual (PERCEVAL)

O método PERCEVAL ("*PERCeptual EVALuation*") é caracterizado pelo uso de um número muito elevado de bandas espectrais e a computação de uma probabilidade de detecção como o principal parâmetro de saída [81]; este parâmetro consiste na probabilidade de um ouvinte humano detectar uma distorção quando da comparação entre

os sinais original e degradado. A não ser pelo elevado número de bandas de filtragem (2.000), que é determinado pelo número de células ciliadas ao longo da membrana basilar, o modelo do ouvido utilizado no PERCEVAL é muito simples. Ele não modela nem a dependência de nível da forma dos filtros auditivos, nem o mascaramento temporal. Para evitar as restrições para distorções próximas ao limiar, conseqüência da abordagem baseada na probabilidade de detecção, versões posteriores do PERCEVAL usaram também a distância média entre os padrões de excitação da membrana basilar dos sinais original e degradado como parâmetros de saída. Quando o modelo foi submetido aos testes competitivos, vários parâmetros adicionais foram incluídos, os quais foram calculados separadamente para diferentes regiões espectrais, e mapeados para uma medida de qualidade global usando uma rede neural. Porém, essa versão estendida nunca foi publicada, com exceção das partes incorporadas no método PEAQ.

5.2.6. Modelo Objetivo Perceptual (POM)

O método *POM* ("*Perceptual Objective Model*") combina a abordagem estatística do cálculo de uma probabilidade de detecção com o modelo auditivo detalhado usado no PAQM, incluindo a aditividade não-linear do mascaramento e funções de espalhamento dependentes do nível [82]. Ele modela também as curvas de mascaramento de maneira mais suave. O número de bandas é determinado pelo limiar de discriminação de freqüências. O número de bandas resultante é menor que no PERCEVAL, mas ainda é bastante elevado. Como no PERCEVAL, a probabilidade de detecção foi posteriormente complementada pela adição de uma medida de diferença de excitação, e vários novos parâmetros foram posteriormente gerados e mapeados para um único valor através do uso de redes neurais.

5.2.7. Avaliação Objetiva de Sinais de Áudio (OASE)

O método *OASE* (*Objective Audio Signal Evaluation*) foi introduzido em 1996 [92]. Ele é baseado em um banco de filtros FIR com 241 bandas de freqüência, o que corresponde a uma alta resolução de freqüência, de cerca de um décimo da largura de uma banda crítica. O alto esforço computacional do banco de filtros é reduzido pelo uso de um algoritmo rápido de convolução e pela pré-filtragem dos sinais de entrada por um filtro passa-baixas, implementado na forma de cascata de três estruturas, o que permite reduzir as taxas de amostragem nas baixas freqüências. O formato dos filtros aproxima as curvas de mascaramento para o pior caso, como no NMR. O mascaramento temporal foi originalmente modelado por um filtro passa-baixas não-linear, mas mais tarde foi substituído por um filtro IIR. A principal saída deste modelo é a probabilidade de detecção das diferenças de excitação (da membrana basilar) entre os sinais original e processado.

5.2.8. A Abordagem da Caixa de Ferramentas (Toolbox Approach)

Este método usa uma abordagem de três passos para medir a distância percebida entre o sinal original e o sinal degradado, fornecendo então uma indicação do nível de qualidade subjetiva de áudio. O método é baseado em modelos perceptuais bem conhecidos, os quais são usados para descrever a representação perceptual das diferenças entre os dois sinais de áudio. Além disso, ele inclui um procedimento de ponderação para a qualidade de áudio percebida para um sinal de teste estéreo, levando em conta os resultados de ambos os canais. O método não requer uma rígida correlação amostra-a-amostra entre os sinais.

O primeiro passo do método diz respeito ao cálculo da sonoridade específica. O segundo passo consiste basicamente de uma série de procedimentos de ponderação. Por fim, o terceiro passo cuida da geração dos parâmetros de saída, os quais são mapeados para um único valor que, por sua vez, é mapeado para o valor subjetivo correspondente através de uma função polinomial.

Como já comentado, nenhum destes métodos alcançou um desempenho satisfatório. No entanto, suas características foram intensamente comparadas e criticadas, pontos fortes e fracos foram identificados e informação foi coletada. Tal estudo culminou na proposta de um novo método, a *Avaliação Perceptual da Qualidade de Áudio* (*Perceptual Evaluation of Audio* Quality - PEAQ) [51,83].

5.2.9. O Método PEAQ

A Figura 5.8 mostra o esquema geral do método PEAQ [83].

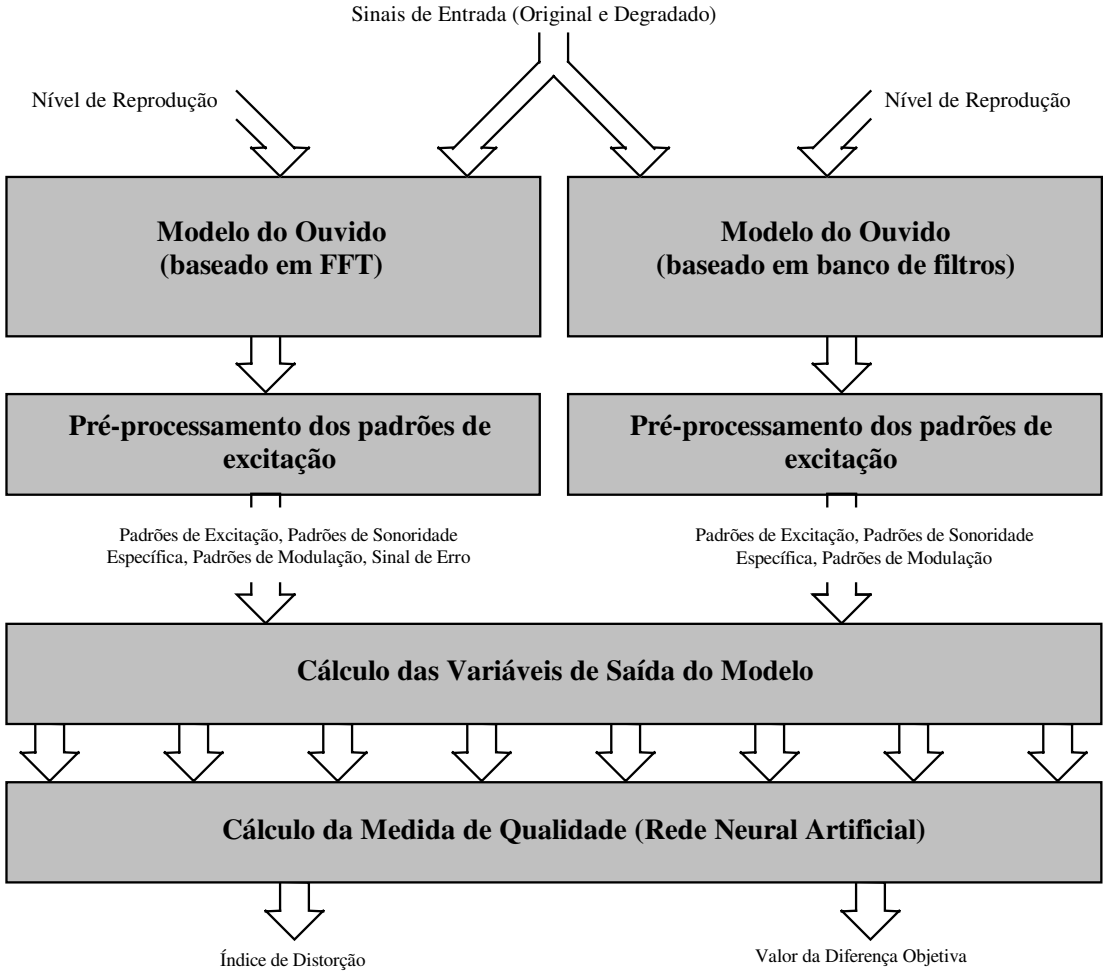


Figura 5.8 - Diagrama de blocos básico do método PEAQ

Este método consiste de um modelo de simulação do ouvido, vários processamentos intermediários (aqui referidos como "pré-processamento dos padrões de excitação") e o

cálculo das variáveis de saída do modelo (MOVs), as quais são, basicamente, uma compilação das medidas apresentadas na Seção 5.2. Essas variáveis são então submetidas a uma rede neural, a partir da qual obtém-se um único valor representando a qualidade de áudio básica do sinal sob teste.

O método PEAQ é composto por duas versões: a primeira, chamada "versão básica", é baseada na FFT (*Fast Fourier Transform*), e tem como principal característica uma baixa complexidade computacional, permitindo a implementação em tempo real; a segunda, chamada "versão avançada", é baseada tanto na FFT quanto em um banco de filtros, e tem como principal característica um melhor desempenho, a fim de se adequar às aplicações mais exigentes, porém com uma complexidade computacional mais elevada. Como se pode notar, ambas as versões usam parâmetros obtidos através do modelo baseado na FFT, porém tais parâmetros são diferentes para cada versão, e até mesmo o processamento realizado após o cálculo da FFT é diferente. Em outras palavras, o modelo baseado na FFT possui duas ramificações que resultarão nos parâmetros usados em cada uma das versões. A estrutura geral, à exceção de alguns poucos processamentos, é semelhante para as duas versões.

As entradas para o cálculo das MOVs, obtidas tanto para o sinal original quanto para o degradado, são:

- Os *padrões de excitação*, obtidos pela distribuição da energia dos sinais em diferentes regiões de *pitch* (o equivalente psicoacústico para a freqüência).
- Os *padrões de excitação espectralmente adaptados*, os quais são computados apenas no modelo baseado no banco de filtros, permitindo a obtenção de resultados mais precisos.
- Os *padrões de sonoridade específica*, simulando a sonoridade percebida pelo ouvinte.
- Os *padrões de modulação*, calculados a partir dos padrões de excitação.
- O *sinal de erro*, calculado como a diferença espectral entre os sinais (somente para o modelo do ouvido baseado na FFT).

Todas as computações para sinais em estéreo são feitas independentemente e da mesma maneira para os canais direito e esquerdo. Ao final, realiza-se uma média aritmética simples entre os valores obtidos para cada canal.

Este método representou um grande avanço em relação aos métodos propostos até então. Ainda assim, ele não foi capaz de alcançar o desempenho desejado para todos os itens de teste. Dessa forma, ainda não é possível abrir mão das avaliações subjetivas, especialmente nos casos que requerem dados realmente confiáveis.

CAPÍTULO 6
O MÉTODO MOQA

Todas as implementações apresentadas neste Capítulo foram realizadas em Matlab® (versão 6.5), a fim de se ter maior flexibilidade e versatilidade para a realização de modificações e testes. A primeira versão implementada baseou-se nas instruções contidas na recomendação BS.1387-1 [83]. Gradativamente, diversos elementos da implementação original foram substituídos por procedimentos mais eficientes e novas estratégias foram incorporadas. Tais mudanças resultaram em um novo método, denominado Medida Objetiva da Qualidade de Áudio (MOQA). Este Capítulo tem como objetivo detalhar cada etapa da implementação, as dificuldades encontradas, as falhas observadas e as soluções adotadas. Devido à complexidade do método, diversas figuras serão apresentadas, no intuito de fornecer ao leitor uma ilustração do que ocorre em cada etapa do processamento.

6.1. CONSIDERAÇÕES GERAIS

A exemplo do que ocorre no método PEAQ, dois diferentes modelos do ouvido foram implementados para o método MOQA, um baseado na FFT e outro baseado em um banco de filtros. No método PEAQ, o resultado deste procedimento foi a geração de duas versões distintas para o programa: uma rápida, porém com resultados mais pobres, e outra mais precisa, porém com elevada complexidade computacional. No MOQA, não há esta divisão em duas versões. A motivação para o uso de dois diferentes modelos foi explorar adequadamente as peculiaridades e vantagens de cada um dos tipos de decomposição tempo-freqüência, a fim de gerar parâmetros de melhor qualidade para alimentação da rede neural artificial usada para estimar a qualidade subjetiva dos sinais. A Figura 6.1 ilustra a estrutura geral adotada para o método MOQA. A descrição detalhada de cada um dos blocos será feita ao longo das próximas seções.

Diversos cuidados foram tomados a fim de tornar a execução do programa mais rápida: utilização de rotinas mais rápidas para o cálculo das FFTs, redução da necessidade de armazenamento de dados e a substituição da maior parte dos "laços" por operações vetoriais e matriciais.

É importante destacar ainda que, para sinais em estéreo, todos os processamentos são realizados independentemente para cada canal, e os resultados correspondentes a cada um dos canais são combinados apenas após a extração dos parâmetros de saída. Assim, ao longo deste Capítulo, quando se coloca que o sinal é submetido a determinado procedimento, deve-se considerar que tal procedimento é aplicado separadamente a cada canal do sinal.

Por fim, deve-se ressaltar que os procedimentos apresentados ao longo deste capítulo foram determinados considerando-se sinais de áudio amostrados a 48 kHz e quantizados com 16 bits por amostra.

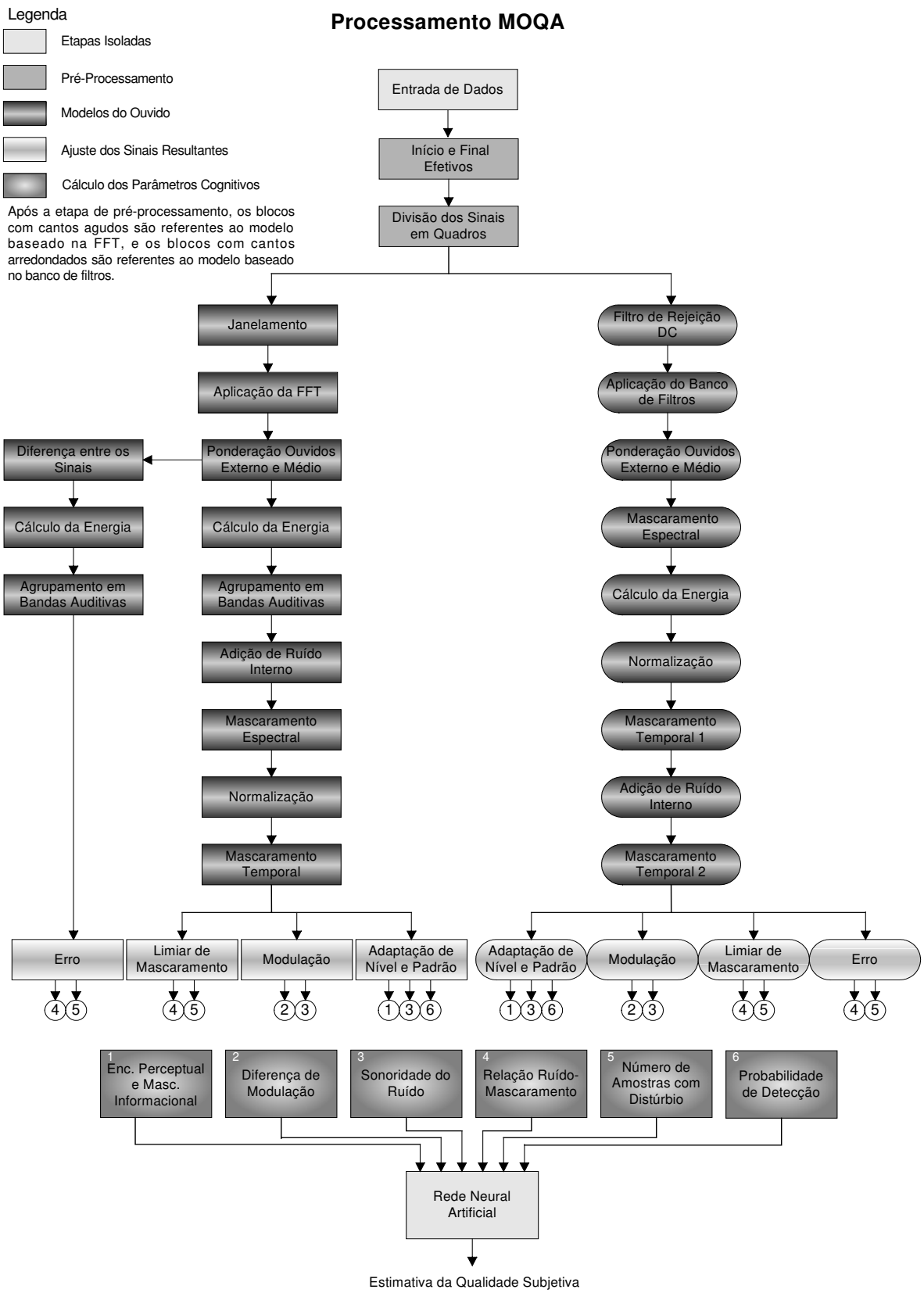


Figura 6.1 - Esquema geral do método MOQA.

6.2. ENTRADA DE DADOS

A entrada do programa que implementa o método MOQA requer o fornecimento de 2 parâmetros obrigatórios, além de permitir a escolha de outros 4 opcionais:

- *Nome dos arquivos de referência e de teste*: estes são os únicos parâmetros que necessariamente deverão ser fornecidos pelo usuário.
- *Número de amostras*: este parâmetro deve ser fornecido pelo usuário no caso de não se desejar que todo o sinal seja analisado, tornando o programa, desta maneira, mais rápido; como padrão, o programa analisa os sinais em todo o seu comprimento.
- *Nível sonoro do sinal*: o nível do sinal, se conhecido, deverá ser fornecido em dB; o valor padrão adotado pelo programa é de 92 dB_{SPL}; apesar de conveniente, o fornecimento do valor correto para o nível do sinal não é crítico para o desempenho do programa, dispensando desta maneira a necessidade do cálculo do nível exato do sinal ou da aplicação de algum tipo de ajuste.
- *Tipo do sinal*: aqui, o usuário deverá escolher entre as opções mono e estéreo; este último é adotado como padrão.
- *Detalhamento e execução do arquivo*: caso esta opção seja selecionada, o programa executará o conteúdo dos sinais de áudio e fornecerá as formas de onda de cada canal separadamente; se esta opção não for explicitamente selecionada, o programa não executará tais instruções.

6.3. PRÉ-PROCESSAMENTO

A etapa de pré-processamento é composta por dois procedimentos, como descrito nas subseções a seguir. Os sinais deverão estar alinhados temporalmente antes de serem submetidos ao restante da rotina.

6.3.1. Identificação do Início e Final Efetivos

Se o arquivo processado contém ruído antes ou após os dados do arquivo original em si, o erro relativo pode ser muito grande, uma vez que o nível do sinal original tende a zero nesses pontos. Quando esse erro é considerado um componente estranho, ele pode ser ignorado pela aplicação de um critério de rejeição de borda, ou limite, de dados. Para isso, é feita a identificação do início e final efetivos do arquivo, definidos, respectivamente, como o primeiro e o último local onde a soma do valor absoluto de cinco amostras sucessivas excede 200, em pelo menos um dos canais. Componentes que se encontrem fora desses limites são ignorados.

6.3.2. Divisão dos Sinais em Quadros

Tradicionalmente, os métodos de avaliação de áudio processam o sinal de áudio como um todo, ou seja, toda extensão do sinal é submetida de uma só vez ao programa em questão. Tal procedimento freqüentemente exige que uma grande quantidade de dados seja temporariamente armazenada na memória de acesso aleatório (RAM) do computador enquanto os cálculos são realizados, tornando-o mais lento. Para minimizar este problema, a solução adotada foi dividir os sinais em trechos de 1,963 segundo (94.240 amostras para uma freqüência de amostragem de 48 kHz), com uma superposição de 41,7 ms (2.000 amostras a 48 kHz). Estes valores foram escolhidos de maneira a se adequarem perfeitamente aos processamentos subseqüentes. Cada trecho é processado separadamente

até a extração dos parâmetros cognitivos. Após, os resultados de cada trecho são combinados através de uma média aritmética simples. A Figura 6.2 ilustra a maneira como os sinais são tratados ao longo da rotina implementada para o MOQA.

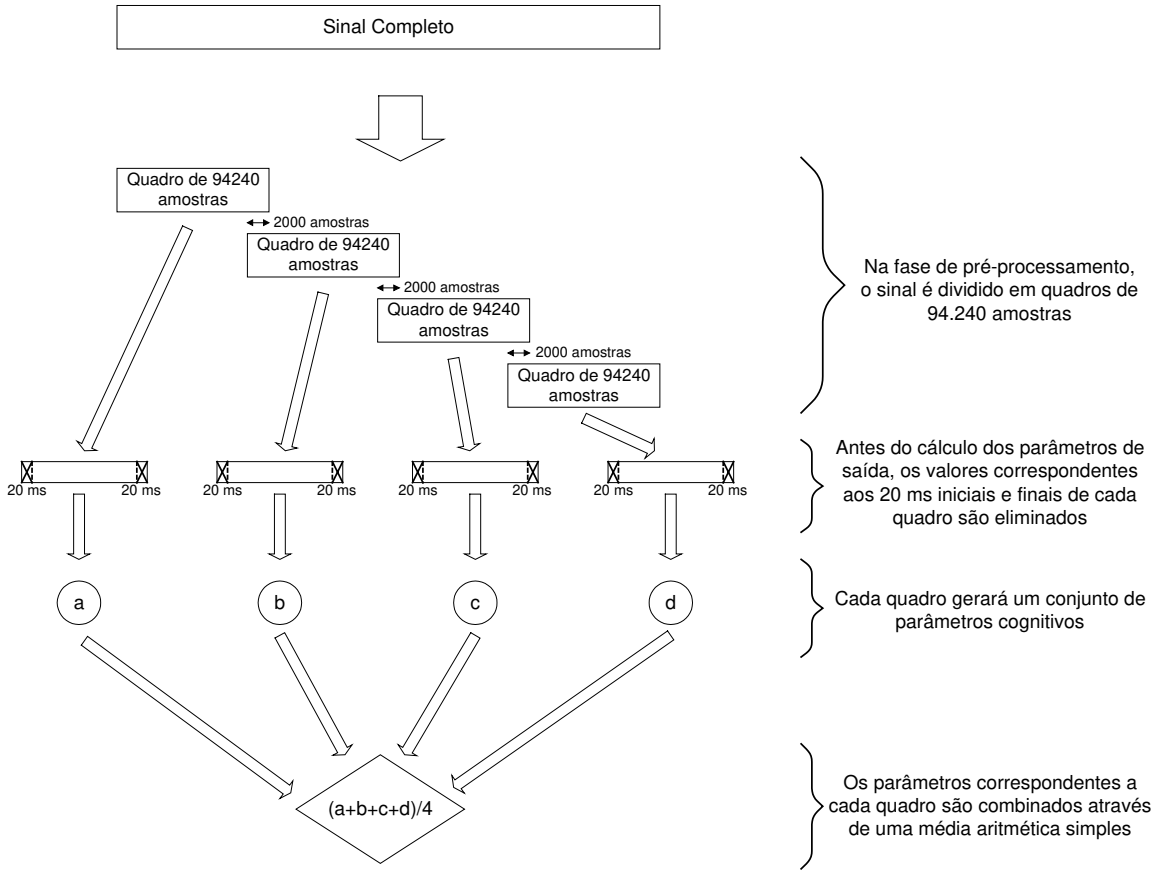


Figura 6.2 - Esquema da divisão do sinal em quadros

Como se pode observar na Figura 6.2, antes de se calcular os parâmetros cognitivos, os padrões resultantes correspondentes aos 20 ms iniciais e finais de cada quadro são eliminados. Tal procedimento é necessário porque alguns componentes cuja influência sobre as amostras das extremidades dos quadros deveria ser modelada estão, por construção do procedimento, localizados em um quadro adjacente, não sendo portanto considerados. Assim, o mascaramento a que tais amostras estão submetidas não é corretamente modelado, e então sua eliminação se faz necessária. Esse é o motivo pelo qual adotou-se uma superposição de 2.000 na divisão de quadros, de modo que nenhuma amostra deixe de ser considerada pelo método. Nas Seções 6.4 e 6.5, o termo “sinal” será referente aos trechos resultantes da divisão descrita.

6.4. MODELOS DO OUVIDO

Como já comentado, dois diferentes modelos para o ouvido foram implementados, de acordo com o tipo de decomposição tempo-freqüência utilizada. Esses modelos, bem como suas principais características, são descritos a seguir.

6.4.1. Modelo Baseado na FFT

6.4.1.1. Divisão em Quadros

Na entrada do modelo do ouvido baseado na FFT, os sinais original e degradado (também denominados de referência e teste) são novamente divididos, desta vez em quadros de 2.048 amostras, com uma superposição de 50% e aplicação de uma janela de Hanning [93]. Um fator de escala, calculado de acordo com a equação

$$fe = \frac{10^{\frac{Lp}{20}}}{Norm}, \tag{6.1}$$

é multiplicado a ambos os sinais. Na Equação 6.1, L_p corresponde ao nível de execução dos sinais e *Norm* representa um fator de normalização de valor igual a $1{,}4258 \times 10^7$. Este valor é obtido tomando-se uma senóide de 1019,5 Hz e 0 dB_{FS} (decibel relativo à escala completa, ver Apêndice A) como sinal de entrada e calculando o máximo valor absoluto dos coeficientes espectrais ao longo de 10 quadros. Se o nível de pressão sonora é desconhecido, recomenda-se a adoção do valor 92 dB_{SPL}.

6.4.1.2. Decomposição Tempo-Freqüência - Aplicação da FFT

O mapeamento para o domínio da freqüência é feito através da aplicação de uma FFT a curto termo (quadro-a-quadro). Os cálculos aqui envolvidos exigiriam o armazenamento de uma quantidade muito elevada de valores, especialmente em se tratando de sinais de áudio. Para resolver tal problema, duas soluções foram adotadas. Na primeira delas, o número total de quadros foi distribuído em 64 grupos separados para o cálculo da FFT; desta maneira, a FFT é calculada para um pequeno número de quadros e as amostras originais correspondentes são imediatamente eliminadas. Este procedimento, por si só, reduz a necessidade de armazenamento pela metade. A segunda solução consiste em se descartar os componentes da FFT que não serão utilizados nos processamentos posteriores. Como a faixa utilizada nos cálculos vai até 18 kHz, a freqüência de amostragem é de 48 kHz e o número de amostras utilizadas no cálculo da FFT é 2.048, tem-se que só é necessário considerar as 768 primeiras amostras; as demais são imediatamente descartadas. Este procedimento reduz a necessidade de armazenamento em mais de 60 %. A Figura 6.3 ilustra este último artifício, onde o primeiro gráfico fornece o espectro de amplitudes completo obtido após a aplicação da FFT a determinado quadro, o segundo mostra apenas os componentes restantes após o descarte das linhas desnecessárias, e o terceiro apresenta um detalhamento do espectro resultante. As duas técnicas combinadas representaram uma redução de cerca de 90% no tempo para o processamento desta etapa da rotina.

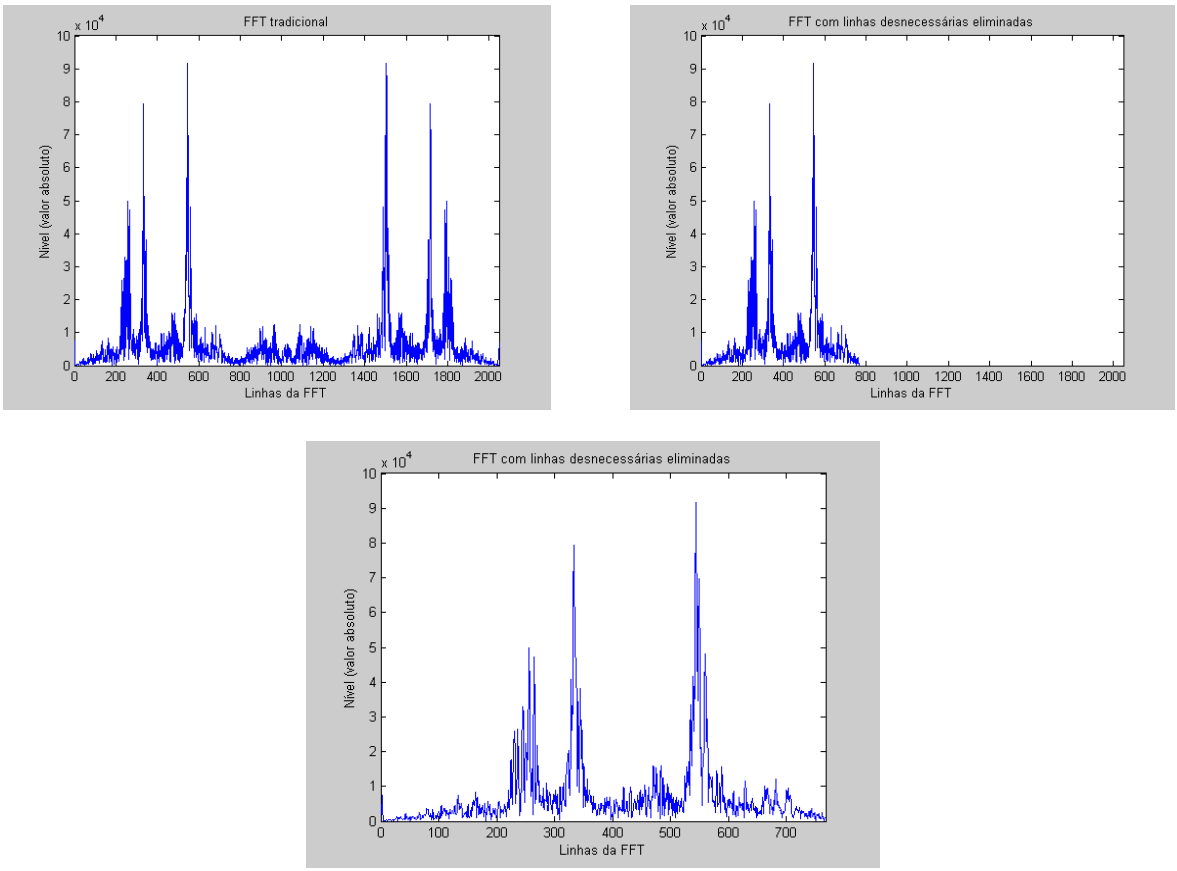


Figura 6.3 - Eliminação das linhas espectrais desnecessárias.

6.4.1.3. Ponderação dos Ouvidos Externo e Médio

A resposta em freqüência dos ouvidos externo e médio é modelada por uma função de ponderação dependente da freqüência, conforme a equação

$$W[k] = -0{,}6 \cdot 3{,}64 \cdot \left(\frac{f[k]}{1000}\right)^{-0{,}8} + 6{,}5 \cdot e^{-0{,}6 \cdot \left(\frac{f[k]}{1000} - 3{,}3\right)^2} - 10^{-3} \cdot \left(\frac{f[k]}{1000}\right)^{3{,}6}, \qquad (6.2)$$

onde $W[k]$ é dado em dB e $f[k]$ é o valor em Hz da freqüência correspondente ao componente k da FFT, conforme a equação

$$f[k] = k \cdot 23{,}4375\,, \qquad (6.3)$$

onde o valor 23,4375 representa o espaçamento, em Hz, entre as linhas espectrais. A Equação 6.2 é baseada nos conceitos apresentados na Seção 2.2.1 e na Equação 2.1. Deste ponto em diante, os índices k e n representarão, respectivamente, os índices das amostras no domínio da freqüência e do tempo.

As saídas da FFT são então ponderadas de acordo com a equação

$$F_w[k,n] = |F[k,n]| \cdot 10^{\frac{W[k]}{20}}. \qquad (6.4)$$

A Figura 6.4 apresenta o formato da função de ponderação aqui utilizada. É interessante notar que, para a faixa de freqüências em torno de 3,3 kHz, o valor da resposta em amplitude fica próximo de 2. Conforme se afasta deste ponto, o ganho vai diminuindo até se transformar em atenuação para as freqüências abaixo de 1.700 e acima de 5.100 Hz.

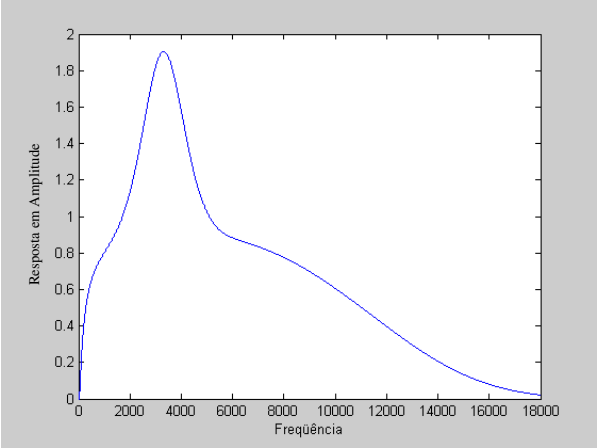


Figura 6.4 - Ponderação simulando as características dos ouvidos externo e médio

6.4.1.4. Cálculo da Energia dos Sinais

Antes do mapeamento da escala de freqüência para a escala perceptual, computa-se o espectro de densidade de energia dos sinais original e degradado através da equação

$$F_p[k,n] = \left| F_w[k,n] \right|^2 . \tag{6.5}$$

6.4.1.5. Agrupamento em Bandas Auditivas

Ao contrário do que ocorre quando se utiliza um banco de filtros, a decomposição tempo-freqüência realizada por intermédio da FFT não é capaz de modelar a resolução espectral não-linear da audição humana. Por esse motivo, é necessário que as linhas espectrais resultantes da aplicação da FFT sejam agrupadas de acordo com as bandas auditivas. Esse agrupamento baseia-se na aproximação dada pela Equação 2.5.

Aqui, adotou-se uma resolução de aproximadamente 0,25 Bark, o que equivale a um total de 109 bandas situadas entre 80 Hz e 18 kHz. Os limites e a freqüência central de cada banda perceptual são aqueles sugeridos em [83] (versão básica). A energia de cada banda, $P_e[k,n]$, é obtida através da soma das energias dos componentes espectrais ($F_p[k,n]$) compreendidos dentro dos limites de cada uma dessas bandas. A energia dos componentes localizados na vizinhança de duas bandas é distribuída entre ambas através de um procedimento que leva em conta a distância entre o componente e o limite entre as bandas. Este procedimento é composto pelos seguintes passos:

- Primeiro, determina-se qual componente está próximo ao limite inferior da banda. Para isso, toma-se a localização do limite e delimita-se uma região determinada por metade da resolução espectral para a esquerda e para a direita. A contribuição desse componente na energia total dessa banda é então dada por

$$ce = \frac{energia \cdot \left[\left(k + 0.5 \right) \cdot res - li \right]}{res}, \tag{6.6}$$

onde *energia* é a energia do componente espectral, *k* é o índice da linha espectral, *res* é a resolução da FFT (23,4375 Hz) e *li* é o valor em Hz do limite inferior da banda correspondente.

- Após, determina-se qual componente está próximo ao limite superior da banda, através do mesmo procedimento descrito no item anterior. A contribuição desse componente na energia total da banda é dada por

$$ce = \frac{energia \cdot [ls - (k - 0.5) \cdot res]}{res}, \tag{6.7}$$

onde *ls* é o limite superior da banda, em Hz.

A Figura 6.5 é apresentada a fim de ilustrar e tornar mais claro o procedimento adotado.

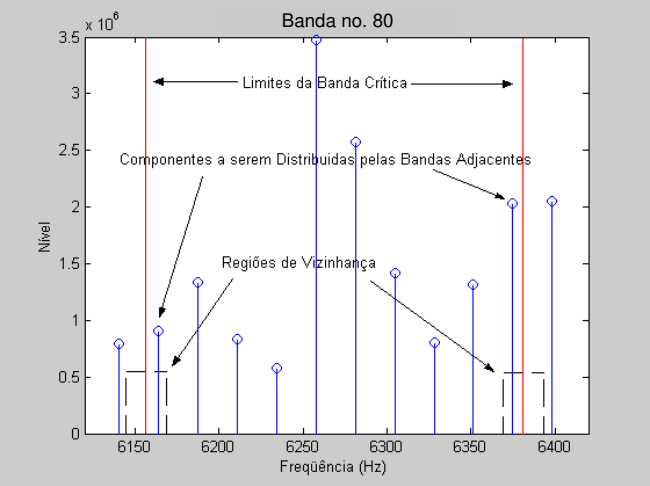


Figura 6.5 - Distribuição da energia de componentes localizadas na vizinhança entre bandas

O limite inferior da banda está localizado em 6.156 Hz e o componente localizado na vizinhança desse limite corresponde à freqüência de 6.164 Hz (linha número 263); sua energia é de 908.230, a qual será assim distribuído:

- para a banda anterior:

$$ce = \frac{908.230 \cdot [6156 - (263 - 0.5) \cdot 23{,}4375]}{23{,}4375} = 141.684;$$

- para a banda mostrada na figura:

$$ce = \frac{908.230 \cdot [(263 + 0.5) \cdot 23{,}4375 - 6156]}{23{,}4375} = 766.546.$$

Da mesma maneira, o limite superior está localizado em 6.382 Hz e o componente localizado nessa vizinhança corresponde à freqüência de 6.375 Hz (linha número 272); sua energia é de 2.028.300, a qual será assim distribuída:

- para a banda mostrada na figura:

$$ce = \frac{2.028.300 \cdot [6382 - (272 - 0.5) \cdot 23{,}4375]}{23{,}4375} = 1.619.936;$$

- para a banda posterior:

$$ce = \frac{2.028.300 \cdot \left[\left(272 + 0.5\right) \cdot 23{,}4375 - 6382\right]}{23{,}4375} = 408.364\,.$$

Este procedimento é repetido para todas as bandas perceptuais.

6.4.1.6. Adição do Ruído Interno

Uma compensação dependente da freqüência, P_l, é adicionada a cada grupo de freqüência, conforme mostrado na equação

$$P_r[k,n] = P[k,n] + P_l[k], \tag{6.8}$$

onde $P[k,n]$ são os padrões resultantes após o agrupamento em bandas auditivas e P_l é dado por

$$P_l[k] = 10^{0{,}4 \cdot 0{,}364 \cdot \left(\frac{f[k]}{1000}\right)^{-0.8}}, \tag{6.9}$$

onde $f[k]$ são as freqüências de cada banda.

6.4.1.7. Modelagem do Mascaramento Espectral

O espalhamento no domínio da freqüência visa modelar a influência que cada componente espectral terá sobre sua vizinhança, conforme descrito na Seção 2.3.1 e ilustrado na Figura 5.6. Os padrões de excitação resultantes após a adição do ruído interno são distribuídos por toda a faixa de freqüência usando uma função de espalhamento, a fim de se levar em consideração as características de mascaramento correspondentes a cada componente espectral. Essa função é uma exponencial de duas caudas. A inclinação inferior (correspondente à influência sobre as freqüências mais baixas) é sempre $S_l = 27$ dB/Bark e a inclinação superior (correspondente à influência sobre freqüências mais altas) é dependente da freqüência e do nível, como mostra a equação

$$S_s[k, L[k,n]] = -24 - \frac{230}{f[k]} + 0{,}2 \cdot L[k,n], \tag{6.10}$$

onde

$$L[k,n] = 10 \cdot \log_{10}\left(P_r[k,n]\right). \tag{6.11}$$

A Figura 6.6 ilustra o formato da função de espalhamento (freqüência no eixo horizontal e amplitude no eixo vertical).

Figura 6.6 - Esboço das inclinações dos filtros.

Portanto, a influência de determinado componente espectral sobre sua vizinhança cai a uma taxa de 27 dB/Bark para as freqüências mais baixas e a uma taxa dada pela Equação

6.10 para as freqüências mais elevadas. O espalhamento é feito independentemente para cada banda perceptual *k*, e é dado por

$$E_e[k,n] = \frac{1}{norm[k]} \left(\sum_{j=0}^{z-1} E_{aux}[j,k,n]^{0,4} \right)^{2,5}, \tag{6.12}$$

onde *z* é igual ao número de bandas (109), E_{aux} é dado por

$$E_{aux}[j,k,n] = \begin{cases} \dfrac{10^{\frac{L[j,n]}{10}} \cdot 10^{\frac{res\cdot(k-j)\cdot S_l[j,L[j,n]]}{10}}}{\sum_{\mu=1}^{j-1} 10^{\frac{res\cdot(\mu-j)\cdot S_l[j,L[j,n]]}{10}} + \sum_{\mu=j}^{z} 10^{\frac{res\cdot(\mu-j)\cdot S_s[j,L[j,n]]}{10}}} & \text{se } k < j \\ \dfrac{10^{\frac{L[j,n]}{10}} \cdot 10^{\frac{res\cdot(k-j)\cdot S_s[j,L[j,n]]}{10}}}{\sum_{\mu=1}^{j-1} 10^{\frac{res\cdot(\mu-j)\cdot S_l[j,L[j,n]]}{10}} + \sum_{\mu=j}^{z} 10^{\frac{res\cdot(\mu-j)\cdot S_s[j,L[j,n]]}{10}}} & \text{se } k \geq j \end{cases} \tag{6.13}$$

e *res* corresponde à resolução da escala perceptual, que é de 0,25 Bark. O fator de normalização dependente da banda perceptual, *norm*, é dado por

$$norm[k] = \left(\sum_{j=0}^{z-1} F[j,k]^{0,4} \right)^{2,5} \quad \text{e} \tag{6.14}$$

$$F[j,k] = \begin{cases} \dfrac{10^{\frac{res\cdot(k-j)\cdot S_l[j,1]}{10}}}{\sum_{\mu=1}^{j-1} 10^{\frac{res\cdot(\mu-j)\cdot S_l[j,1]}{10}} + \sum_{\mu=j}^{z} 10^{\frac{res\cdot(\mu-j)\cdot S_s[j,1]}{10}}} & \text{se } k < j \\ \dfrac{10^{\frac{res\cdot(k-j)\cdot S_s[j,1]}{10}}}{\sum_{\mu=1}^{j-1} 10^{\frac{res\cdot(\mu-j)\cdot S_l[j,1]}{10}} + \sum_{\mu=j}^{z} 10^{\frac{res\cdot(\mu-j)\cdot S_s[j,1]}{10}}} & \text{se } k \geq j \end{cases} . \tag{6.15}$$

O índice *j*, a exemplo de *k*, varia entre 1 e 109, e é apenas uma variável auxiliar para os cálculos, desaparecendo no somatório realizado na Equação 6.12.

A implementação desta etapa foi particularmente complicada, pois muitos dos cálculos são realizados em três dimensões, e não em duas, como ocorre normalmente. Tal característica acarreta o surgimento de alguns problemas; dentre estes, dois se destacam: a impossibilidade de se realizar multiplicações matriciais em três dimensões e o grande aumento da necessidade de armazenamento de dados devido à dimensão adicional.

Para tornar a implementação computacionalmente eficiente, uma estratégia alternativa foi desenvolvida. Os passos descritos a seguir são referentes ao cálculo da Equação 6.15, na

qual a dimensão *n*, relativa às amostras no domínio do tempo, é fixada no índice 1. Para o cálculo da Equação 6.13, utilizou-se exatamente o mesmo procedimento, porém houve a necessidade de se utilizar um laço de iteração para que tais cálculos fossem realizados para cada índice temporal *n*. Em outras palavras, duas dimensões, *j* e *k*, foram abordadas matricialmente, e a terceira, *n*, foi calculada iterativamente.

Inicialmente, gerou-se uma matriz quadrada de dimensão 109, a qual é usada para modelar os fatores (j - k) e (μ - j). Esta matriz tem a seguinte estrutura:

$$\mathrm{C} = \begin{bmatrix} 0 & 1 & \cdots & 107 & 108 \\ -1 & 0 & \cdots & 106 & 107 \\ \vdots & & \ddots & \vdots & \vdots \\ -107 & -106 & \cdots & 0 & 1 \\ -108 & -107 & \cdots & -1 & 0 \end{bmatrix}.$$

Pode-se notar ainda que a única diferença entre os somatórios encontrados nos denominadores das Equações 6.13 e 6.15 é o uso dos valores correspondentes ou à inclinação inferior ou à inclinação superior, dependendo dos valores de μ e de j. Assim, pode-se gerar uma única matriz com os valores a serem utilizados, unificando os somatórios. Essa matriz é construída da seguinte maneira: nos casos em que os elementos da matriz C são maiores que 0, tem-se que $\mu \geq j$, e então, de acordo com a Equação 6.13, utiliza-se a inclinação superior, S_s. Por outro lado, para os elementos da matriz C menores que 0, tem-se que $\mu < j$, e portanto utiliza-se a inclinação inferior, S_l, a qual é sempre igual a 27 dB. A matriz resultante assume a seguinte estrutura:

$$\mathrm{S} = \begin{bmatrix} S_s[1,n] & S_s[1,n] & \cdots & S_s[1,n] & S_s[1,n] \\ 27 & S_s[2,n] & \cdots & S_s[2,n] & S_s[2,n] \\ \vdots & & \ddots & \vdots & \vdots \\ 27 & 27 & \cdots & S_s[108,n] & S_s[108,n] \\ 27 & 27 & \cdots & 27 & S_s[109,n] \end{bmatrix}.$$

Na Equação 6.13, o valor de *n* varia de 1 até o comprimento em número de amostras dos quadros no domínio do tempo; na Equação 6.15, *n* = 1. A matriz *S* é também utilizada no cálculo dos numeradores das equações citadas, já que a situação é exatamente a mesma, com a diferença de que μ é substituído por k. Assim, as Equações 6.13 e 6.14 tornam-se, respectivamente

$$E_{aux}[j,k,n] = \frac{10^{\frac{res\cdot(C_{j,k}[n]*S_{j,k}[n])+L_j[n]\cdot u}{10}}}{\left(\sum_{\mu=0}^{z-1} 10^{\frac{res\cdot(C_{j,\mu}[n]*S_{j,\mu}[n])}{10}}\right)\cdot u}, \tag{6.16}$$

$$norm[j,k] = \frac{10^{\frac{res\cdot(C_{j,k}*S_{j,k})}{10}}}{\left(\sum_{\mu=0}^{z-1} 10^{\frac{res\cdot(C_{j,\mu}*S_{j,\mu})}{10}}\right)\cdot u}, \tag{6.17}$$

onde os subscritos das matrizes indicam as variáveis que determinam suas dimensões e *u* é um vetor linha de dimensão 1 x 109, cujos elementos possuem todos valor unitário. Ele aparece sempre multiplicando vetores coluna de dimensão 109 x 1, e seu efeito é o de reproduzir esses vetores coluna 109 vezes, de modo a gerar uma matriz quadrada de dimensão 109. Isto é feito para que todos os elementos nas equações tenham as mesmas dimensões, permitindo que os cálculos sejam realizados adequadamente para todas as bandas perceptuais ao mesmo tempo. Nas Equações 6.16 e 6.17, as operações de divisão, bem como as multiplicações indicadas pelo símbolo $*$, são todas do tipo escalar, ou seja, são realizadas individualmente entre cada elemento correspondente das matrizes, conforme exemplificado na equação

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a \cdot e & b \cdot f \\ c \cdot g & d \cdot h \end{bmatrix} \tag{6.18}$$

A Figura 6.7 ilustra os efeitos do espalhamento sobre determinado sinal de áudio. O formato da função de espalhamento segue aproximadamente aquele ilustrado na Figura 5.6.

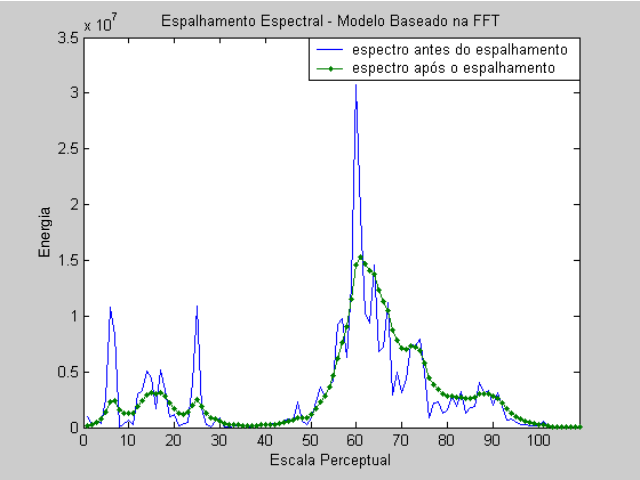


Figura 6.7 - Espalhamento espectral para o modelo baseado na FFT

6.4.1.8. Normalização

Observou-se que o valor para a energia de cada banda perceptual antes e depois da modelagem do mascaramento espectral não é o mesmo, de modo que um novo fator de normalização deve ser aplicado. Esse fator é determinado simplesmente pela relação entre as energias antes e depois da aplicação desta etapa do processamento, dada por

$$norm_2[k] = \frac{\sum_{n=1}^{N} P_r[k,n]}{\sum_{n=1}^{N} E_e[k,n]}. \tag{6.19}$$

Esse fator é então multiplicado ao padrão resultante, segundo a expressão

$$E_c[k,n] = norm_2[k] \cdot E_e[k,n] \tag{6.20}$$

6.4.1.9. Modelagem do Mascaramento no Domínio do Tempo

No modelo do ouvido baseado na FFT, apenas o mascaramento progressivo é modelado no domínio do tempo. Aqui, as energias em cada grupo de freqüência são distribuídas sobre o tempo através de filtros passa-baixas de primeira ordem. Esses filtros são computados de acordo com as equações

$$E_f[k,n] = a[k] \cdot E_f[k,n-1] + (1 - a[k]) \cdot E_n[k,n], \tag{6.21}$$

$$E[k,n] = \max(E_f(k,n), E_n(k,n)), \tag{6.22}$$

onde n é o índice temporal, k é o índice da banda, $E_f[k,0] = 0$ e $a[k]$ é dado por

$$a[k] = e^{-\frac{4}{187,5} \cdot \frac{1}{\tau[k]}}. \tag{6.23}$$

As constantes de tempo $\tau[k]$ dependem da freqüência central de cada grupo, e são calculadas de acordo com

$$\tau[k] = \tau_{\min} + \frac{100}{f[k]} \cdot (\tau_{100} - \tau_{\min}), \tag{6.24}$$

onde $\tau_{min} = 0{,}008$ e $\tau_{100} = 0{,}030$, valores estes sugeridos em [83].

O programa implementado para o método MOQA faz intenso uso de filtragens do tipo IIR de primeira ordem. A implementação recursiva deste tipo de filtro não é eficiente computacionalmente, por isso uma técnica mais apropriada foi desenvolvida, a qual é descrita a seguir.

Primeiro, observa-se que, para cada banda k, a Equação 6.21 pode ser expandida da seguinte maneira:

$$\begin{cases} E_f[k,1] = a \cdot E_f[k,0] + (1-a) \cdot E_n[k,1] \\ E_f[k,2] = a \cdot E_f[k,1] + (1-a) \cdot E_n[k,2] = a^2 \cdot E_f[k,0] + a \cdot (1-a) \cdot E_n[k,1] + (1-a) \cdot E_n[k,2] \\ \quad \vdots \\ E_f[k,N] = a \cdot E_f[k,N-1] + (1-a) \cdot E_n[k,N] = a^N \cdot E_f[k,0] + a^N \cdot (1-a) \cdot E_n[k,1] + \\ \qquad + a^{N-1} \cdot (1-a) \cdot E_n[k,2] + \cdots + a \cdot (1-a) \cdot E_n[k,n-1] + (1-a) \cdot E_n[k,N] \end{cases}, \tag{6.25}$$

onde N é o comprimento de cada quadro no domínio do tempo, em número de amostras. Matricialmente, tem-se:

$$\begin{bmatrix} E_f[k,1] \\ E_f[k,2] \\ \vdots \\ E_f[k,N-1] \\ E_f[k,N] \end{bmatrix} = \begin{bmatrix} a \\ a^2 \\ \vdots \\ a^{N-1} \\ a^N \end{bmatrix} \cdot E_f[k,0] + (1-a) \cdot \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ a^{N-1} & a^{N-2} & a^{N-3} & \cdots & 1 & 0 \\ a^N & a^{N-1} & a^{N-2} & \cdots & a & 1 \end{bmatrix} \cdot \begin{bmatrix} E_n[k,1] \\ E_n[k,2] \\ \vdots \\ E_n[k,N-1] \\ E_n[k,N] \end{bmatrix} \tag{6.26}$$

Mas E_f[k,0] = 0, então a equação fica sendo:

$$\begin{bmatrix} E_f[k,1] \\ E_f[k,2] \\ \vdots \\ E_f[k,N-1] \\ E_f[k,N] \end{bmatrix} = (1-a)\cdot \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a_k & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_k^{N-1} & a_k^{N-2} & a_k^{N-3} & \cdots & 1 & 0 \\ a_k^{N} & a_k^{N-1} & a_k^{N-2} & \cdots & a_k & 1 \end{bmatrix} \cdot \begin{bmatrix} E_n[k,1] \\ E_n[k,2] \\ \vdots \\ E_n[k,N-1] \\ E_n[k,N] \end{bmatrix} \quad (6.27)$$

Como se pode observar, da maneira como o equacionamento está apresentado, os cálculos teriam que ser realizados separadamente para cada banda *k*, através do uso de laços de iteração, o que tornaria o programa lento. Contudo, pode-se verificar que a multiplicação entre a matriz de coeficientes e o vetor no lado direito da equação nada mais é que a realização de uma convolução linear entre a última linha da matriz e o vetor. Assim, pode-se construir uma nova matriz com os coeficientes a serem usados nas convoluções. A Equação 6.27 pode então ser reescrita da seguinte maneira:

$$\begin{bmatrix} E_f[1,1] & E_f[2,1] & \cdots & E_f[109,1] \\ E_f[1,2] & E_f[2,2] & \cdots & E_f[109,2] \\ \vdots & \vdots & \ddots & \vdots \\ E_f[1,N-1] & E_f[2,N-1] & \cdots & E_f[109,N-1] \\ E_f[1,N] & E_f[2,N] & \cdots & E_f[109,N] \end{bmatrix} =$$

$$= (1-a)\cdot \begin{bmatrix} 1 & a_1 & \cdots & a_1^{N-1} & a_1^{N} \\ 1 & a_2 & \cdots & a_2^{N-1} & a_2^{N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & a_{108} & \cdots & a_{108}^{N-1} & a_{108}^{N} \\ 1 & a_{109} & \cdots & a_{109}^{N-1} & a_{109}^{N} \end{bmatrix} \otimes \begin{bmatrix} E_n[1,1] & E_n[2,1] & \cdots & E_n[109,1] \\ E_n[1,2] & E_n[2,2] & \cdots & E_n[109,2] \\ \vdots & \vdots & \ddots & \vdots \\ E_n[1,N-1] & E_n[2,N-1] & \cdots & E_n[109,N-1] \\ E_n[1,N] & E_n[2,N] & \cdots & E_n[109,N] \end{bmatrix}, \quad (6.28)$$

onde o símbolo $\otimes$ indica a convolução entre as linhas da matriz de coeficientes e as colunas correspondentes da matriz $E_n[k,n]$. O cálculo destas convoluções ainda apresenta elevada complexidade computacional. A solução adotada foi transformar as convoluções em multiplicações escalares simples, através da aplicação de uma transformada de Fourier a cada uma das matrizes envolvidas, conforme mostrado nas Equações 6.29 e 6.30. A fim de compactar a notação, deste ponto em diante a matriz do lado esquerdo da Equação 6.28 será indicada por E_f, o fator (1-a) será representado como uma constante *c*, e as matrizes restantes serão denotadas por *A* e E_n, respectivamente.

$$\mathbf{A} = \Im\{A\} \quad e \quad \mathbf{E_n} = \Im\{E_n\} \quad (6.29)$$

$$E_f = c \cdot \Im^{-1}\left\{\hat{\mathbf{A}} * \hat{\mathbf{E}}_\mathbf{n}\right\} \quad (6.30)$$

Nas Equações 6.29 e 6.30, $\Im$ indica a aplicação de uma transformada de Fourier, $\Im^{-1}$ indica a transformada inversa, $*$ indica a multiplicação escalar entre os elementos das matrizes e $\hat{\mathbf{A}}$ e $\hat{\mathbf{E}}_\mathbf{n}$ são as matrizes $\mathbf{A}$ e $\mathbf{E_n}$ modificadas, de modo que a resposta ao impulso do filtro IIR é truncada em 40 amostras. Tal procedimento reduz a necessidade de cálculos, sem prejuízo para a qualidade da filtragem. A fim de reduzir ainda mais a quantidade de operações, a matriz $\hat{\mathbf{A}}$, que é constante, foi calculada apenas uma vez, sendo então

armazenada. Sempre que esta etapa é utilizada na análise de um sinal, o programa busca esta matriz e realiza os cálculos necessários.

A Figura 6.8 mostra um exemplo do efeito da aplicação deste procedimento aos sinais.

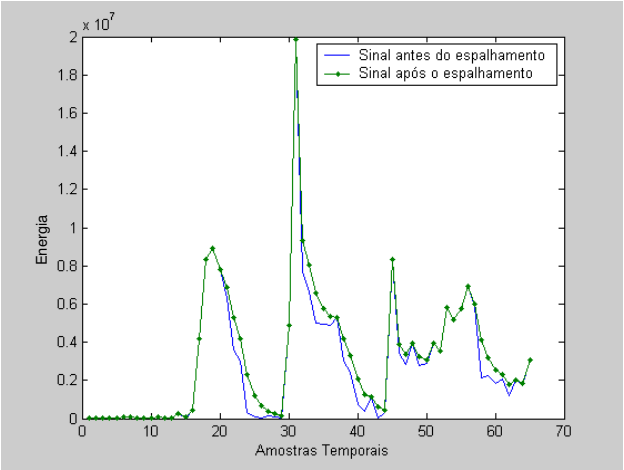


Figura 6.8 - Espalhamento no domínio do tempo para o modelo baseado na FFT

Apenas o mascaramento progressivo é modelado. Pode-se observar também que o espalhamento no domínio do tempo é menos intenso que no domínio da freqüência, uma vez que o mascaramento simultâneo (espectral) é de fato um fenômeno auditivo mais importante que o mascaramento temporal na avaliação de áudio.

6.4.2. Modelo Baseado no Banco de Filtros

6.4.2.1. Filtro de Rejeição DC

Primeiro, os sinais original e degradado são multiplicados por um fator de escala, cujo cálculo é idêntico àquele dado pela Equação 6.1. Contudo, o valor de *Norm* é dado aqui pelo nível de execução assumido para uma senóide em escala completa, que é 32.767.

Como o banco de filtros a ser aplicado é sensível a componentes subsônicas, um filtro de rejeição DC é aplicado a ambos os sinais. Adotou-se um filtro *Butterworth* passa-altas de quarta ordem, com uma freqüência de corte em 20 Hz. Esse filtro é implementado como uma cascata de dois filtros IIR de segunda ordem:

$$y_n = x_n - 2 \cdot x_{n-1} + b_1 \cdot y_{n-1} + b_2 \cdot y_{n-2}, \tag{6.31}$$

onde os coeficientes *b* são, para o primeiro bloco, (1,99517;-0,995174) e, para o segundo bloco, (1,99799;-0,997998).

6.4.2.2. Decomposição Tempo-Freqüência – Aplicação do Banco de Filtros

O banco consiste de 40 pares de filtros, os quais são aplicados a cada um dos sinais. Os filtros são igualmente espaçados e têm uma largura de faixa absoluta constante quando relacionados com uma escala perceptual. As freqüências centrais variam de 50 Hz a 18 kHz. A escala perceptual é calculada através da Equação 2.5. Cada par de filtros possui uma mesma resposta de amplitude, mas com uma diferença de 90° na resposta em fase. Assim, a saída de um filtro representa a transformada de Hilbert do outro (ou a parte imaginária, se for assumido que o primeiro filtro representa a parte real de um sinal

complexo). As envoltórias de sua resposta ao impulso têm um formato $\cos^2$. Tais filtros são definidos por

$$\begin{aligned} h_{re}(k,n) &= \frac{4}{N[k]} \cdot \text{sen}^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \cos\left(2\pi \cdot f[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \\ h_{im}(k,n) &= \frac{4}{N[k]} \cdot \text{sen}^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \text{sen}\left(2\pi \cdot f[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \end{aligned} \quad \Big| \quad 0 \le n < N[k], \qquad (6.32)$$

$$h_{re}(k,n) = h_{im}(k,n) = 0 \quad \Big| \quad \begin{array}{l} n < 0 \\ n \ge N[k] \end{array}$$

onde k é o índice do filtro, n é o índice da amostra no tempo, T é o intervalo de tempo entre duas amostras (1/48000), $N[k]$ é o comprimento da resposta ao impulso de cada filtro e f é a freqüência central de cada banda determinada pelos filtros, segundo tabela encontrada em [83]. Esses filtros são implementados de modo a terem uma resposta ao impulso finita, usando os valores $h_{re}(k,n)$ e $h_{im}(k,n)$ como coeficientes.

A fim de se ter um atraso único para todos os filtros e se alinhar as respostas em fase, a entrada do k-ésimo filtro é atrasada em $D[k]$ amostras, conforme a equação

$$\mathrm{D}[\mathrm{k}] = 1 + 0{,}5 \cdot \left(N[0] - N[k]\right). \qquad (6.33)$$

onde N(0) é o comprimento da resposta impulsiva do primeiro filtro, a qual é a mais longa, segundo valores encontrados em [83]. Na saída do banco de filtros, os sinais são dizimados por um fator de 32.

A Figura 6.9 mostra alguns exemplos dos formatos das respostas dos filtros no domínio do tempo (curvas de cima, correspondentes à resposta ao impulso) e da freqüência (curvas de baixo, correspondentes à resposta em amplitude).

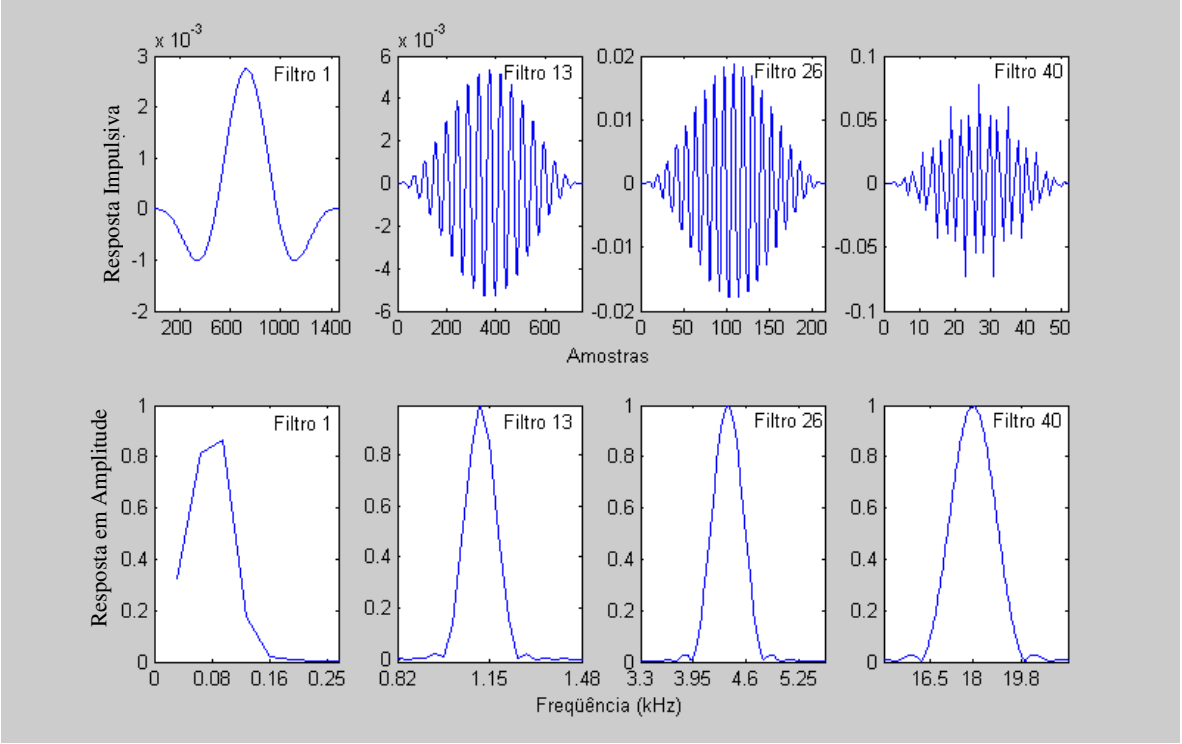


Figura 6.9 - Curvas obtidas para o banco de filtros

Como se pode notar, a largura da banda passante aumenta à medida que sua freqüência central aumenta. Este comportamento é esperado, uma vez que estes filtros modelam o comportamento da membrana basilar e das respectivas bandas críticas (ver Seções 2.1.3 e 2.2.2).

A implementação do banco de filtros demandou uma atenção especial, devido à enorme quantidade de cálculos envolvidos, especialmente em se tratando de sinais que muitas vezes ultrapassam 5.000.000 de amostras. Assim, sete diferentes versões para o banco de filtros foram implementadas:

- *Implementação direta dos filtros FIR amostra-a-amostra*: esta versão consistiu na implementação direta dos filtros e posterior dizimação dos sinais resultantes por um fator de 32; supondo um sinal de 3 milhões de amostras, o número total de multiplicações neste caso ultrapassaria 50 bilhões, além de ocupar uma quantidade muito grande de memória.

- *Implementação direta dos filtros FIR com cálculo a cada 32 amostras*: aqui, a dizimação dos sinais é feita durante a implementação dos filtros, fazendo com que os cálculos sejam realizados apenas a cada 32 amostras. Ainda que a quantidade de cálculos seja 32 vezes menor e o ganho de tempo tenha sido expressivo, seu desempenho em termos de eficiência computacional deixou muito a desejar.

- *Filtro FIR quantizado*: nesta tentativa, os coeficientes dos filtros foram quantizados em um certo número de níveis, de maneira que as amostras a serem submetidas a coeficientes de mesmo valor pudessem ser agrupadas e somadas, reduzindo drasticamente a quantidade de multiplicações. Esperava-se uma melhora significativa no desempenho; o que se observou na prática, entretanto, foi que esta abordagem só apresentou melhores resultados que a primeira tentativa. Tal comportamento pode ser explicado pela complexidade envolvida na implementação desta técnica, uma vez que cada agrupamento de amostras deveria ser feito individualmente, para só então calcular as multiplicações; o Matlab demonstrou não estar otimizado para este tipo de situação, ocasionando um desempenho sofrível.

- *Multiplicação no domínio da freqüência*: uma convolução no domínio do tempo torna-se uma multiplicação no domínio da freqüência. Assim, tomou-se as transformadas rápidas de Fourier dos sinais e dos coeficientes dos filtros, para então multiplicá-los no domínio da freqüência, com o posterior cálculo da transformada inversa. Devido ao grande número de amostras contidas nos sinais, tal abordagem revelou-se desastrosa, pois acarreta o armazenamento de grande quantidade de informação, tornando o programa extremamente lento. Dentre todas as tentativas, esta apresentou o pior desempenho.

- *Método overlap-and-save* [93]: usa o mesmo princípio da versão anterior, porém com uma redução da quantidade de memória requerida através da divisão do sinal em quadros e combinação dos resultados obtidos para cada um dos trechos, segundo determinada metodologia. Esta técnica apresentou resultados ligeiramente superiores àqueles obtidos através da segunda abordagem, porém ainda distantes do desejável. Um dos fatores determinantes para que isto ocorresse foi a impossibilidade de se fazer a dizimação diretamente no cálculo da FFT inversa, o que acarretou uma grande quantidade de cálculos desnecessários. Contudo, ainda que este problema fosse solucionado, esta abordagem ainda teria um desempenho inferior à solução apresentada a seguir.

- *Filtro FIR recursivo*: proposta em [25], esta abordagem insere no equacionamento dos filtros FIR um pólo que, com a alocação adequada dos zeros, deverá ser cancelado; a possível perda de estabilidade ocasionada pelos limites de precisão numérica foi

investigada [25], chegando-se à conclusão que, especialmente para sinais de áudio de menos de 40 segundos de duração, tal ocorrência é muito improvável. Como na saída dos filtros os sinais deverão estar dizimados por um fator de 32, esta abordagem só requer o valor das 32 amostras anteriores e do resultado da componente filtrada anterior, tornando-a eficiente. Esta abordagem apresentou um desempenho pelo menos cinco vezes superior às demais. A descrição detalhada dos procedimentos adotados pode ser encontrada em [94].

- *Divisão do sinal em quadros e implementação matricial dos filtros*: esta abordagem apresentou um desempenho muito próximo àquele apresentado pela técnica anterior [94]. A decisão de se adotar este procedimento na implementação final deveu-se ao fato da abordagem usando o filtro FIR recursivo envolver algumas aproximações que deterioram levemente a qualidade do processamento. A implementação desta técnica é composta por sete etapas básicas, as quais são ilustradas na Figura 6.10.

1) Os coeficientes de cada filtro são agrupados em uma matriz de 40 por 1.456 amostras, onde o número de colunas corresponde ao número de coeficientes do filtro cuja resposta ao impulso é mais longa (os filtros com menor número de coeficientes são completados com zeros) e o número de colunas corresponde ao número de filtros. Como a filtragem é realizada aos pares, tem-se como resultado um par de matrizes com as dimensões citadas.
2) Inverte-se a ordem dos coeficientes dos filtros, ou seja, o último coeficiente se torna o primeiro e vice-versa. Tal procedimento visa facilitar a implementação da convolução, a qual é realizada através das próximas etapas.
3) Os sinais são divididos em quadros de 20.000 amostras; tal divisão visa diminuir a necessidade de armazenamento de valores, tornando o programa mais rápido. Uma superposição de 1.440 amostras entre os quadros se fez necessária a fim de se aplicar a convolução às amostras apropriadas, evitando problemas na etapa de concatenação dos resultados. É importante relembrar que a filtragem é realizada apenas a cada 32 amostras, realizando de forma automática a dizimação apropriada dos sinais. A Figura 6.10 fornece maiores detalhes envolvidos nesta divisão.
4) Cada quadro é dividido em subquadros de 1.456 amostras, com uma superposição de 1.424 amostras, os quais são então concatenados em uma matriz de 625 por 1.456 amostras. Tal divisão, como mostrado na Figura 6.10, tem como objetivo permitir que os coeficientes dos filtros sejam aplicados matricialmente a todo o sinal, de maneira que os 40 filtros possam ser aplicados de uma só vez. Em outras palavras, cada linha da matriz resultante da divisão de cada quadro consiste da linha anterior deslocada de 32 amostras, modelando o fato de que só é necessário aplicar a filtragem às amostras múltiplas de 32. A superposição de 1.440 amostras garante que a primeira amostra para a qual a filtragem é aplicada em um quadro está localizada exatamente 32 posições após a última amostra considerada no quadro anterior, como ilustrado na Figura 6.10.
5) A seguir, é realizada a operação matricial de filtragem, onde as matrizes de coeficientes dos filtros, de dimensão 40 por 1.456, são multiplicadas pela matriz transposta de amostras deslocadas do sinal, de dimensão 1.456 por 625, resultando em uma matriz filtrada de dimensão 40 por 625. Tal procedimento é repetido para cada quadro de 20.000 amostras.
6) Os resultados obtidos para cada quadro são concatenados em uma única matriz composta por 40 linhas e um número de colunas 32 vezes menor que o comprimento total dos sinais.

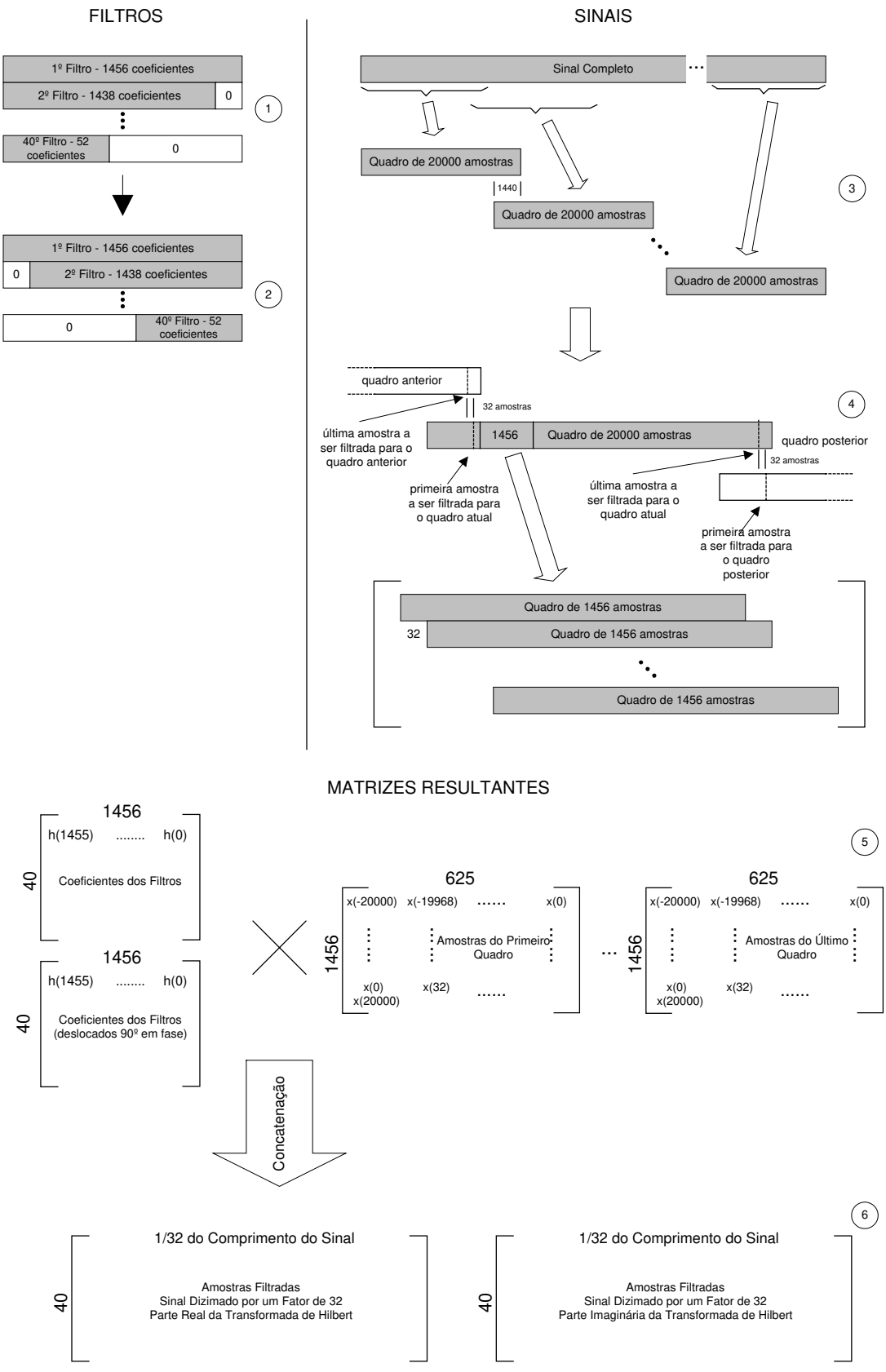


Figura 6.10 - Esquema de filtragem dos sinais

6.4.2.3. Ponderação dos Ouvidos Externo e Médio

Esta etapa é idêntica àquela implementada para o modelo baseado na FFT e descrita na Subseção 6.4.1.3, com exceção das freqüências centrais, as quais são aqui determinadas de acordo com a implementação do banco de filtros.

6.4.2.4. Modelagem do Mascaramento Espectral

As funções de espalhamento usadas no modelo baseado no banco de filtros são essencialmente as mesmas usadas no modelo baseado na FFT, porém a maneira como estas são aplicadas é diferente. O procedimento adotado neste caso é mais simples, devido às próprias características do banco de filtros e ao fato do mascaramento ser aqui modelado antes do cálculo das energias. A estratégia consiste em se tomar determinado componente e adicionar a este a influência de todos os outros componentes correspondentes às demais freqüências. Componentes mais próximos terão maior influência, assim como os de maior nível.

O espalhamento é realizado levando-se em consideração tanto os componentes de freqüências superiores quanto de freqüências inferiores, através de uma função exponencial de dois lados. A inclinação inferior é sempre de 31 dB/Bark, e a inclinação superior varia entre -24 e -4 dB/Bark, como mostra a Equação 6.34. A Figura 6.6 fornece uma visualização da situação. Assim, a influência de determinado componente espectral sobre seus vizinhos cai a uma taxa de 31 dB/Bark conforme se diminui as freqüências e entre 4 e 24 dB/Bark conforme as freqüências são elevadas.

$$S_s[k] = \min\left\{-4, -24 - \frac{230}{f[k]} + 0.2 \cdot L[k]\right\} \tag{6.34}$$

Esta expressão é idêntica àquela correspondente à versão baseada na FFT (Equação 6.10), à exceção do limitante superior em -4 dB, o qual não está presente na Equação 6.10. As freqüências centrais $f[k]$ podem ser encontradas em [83], e o valor de $L[k]$ é dado por

$$L[k] = 10 \cdot \log_{10}\left(x_r^{\,2}[k] + x_i^{\,2}[k]\right) \tag{6.35}$$

onde x_r e x_i representam os sinais na saída dos pares de filtros. O espalhamento espectral é feito inicialmente para a inclinação superior (dependente do nível) e depois para a inferior, e é realizado independentemente para a saída de cada filtro.

A Rec. BS.1387-1 fornece um pseudocódigo para a modelagem do mascaramento espectral, o qual é bastante ineficiente do ponto de vista da complexidade computacional. A implementação eficiente desta etapa tem importância fundamental no desempenho global do método. Se implementada da maneira sugerida, ela responde por 90% do tempo consumido pelo modelo baseado no banco de filtros e por quase 75% do tempo requerido pelo programa como um todo. Isto ocorre porque neste ponto ainda não ocorreu a segunda dizimação dos sinais, portanto a quantidade de amostras envolvidas nos cálculos ainda é muito grande. Os esforços na busca por uma nova estratégia resultaram em um procedimento cerca de 10 vezes mais rápido que aquele sugerido em [83]. Assim, sua participação foi reduzida para 25% do tempo total requerido pelo programa, o qual, por sua vez, se tornou 5 vezes mais rápido. Tal desempenho foi alcançado através do arranjo adequado dos dados de maneira a permitir a substituição dos laços por operações matriciais, conforme mostrado a seguir.

Inicialmente, determina-se o valor de três constantes auxiliares:

$$a = e^{\frac{-32}{4800}} = 0{,}9934\,, \tag{6.36}$$

$$b = 1 - a = 0{,}0066\,, \tag{6.37}$$

$$d = 0{,}1^{\frac{z[40]-z[1]}{39\cdot 20}} = 0{,}9219\,, \tag{6.38}$$

onde $\frac{z[40]-z[1]}{39}$ é a distância em Bark entre duas bandas adjacentes e d é usado na determinação da influência de determinado componente sobre as demais, como será visto mais adiante. Os valores de a e b são usados como coeficientes de um filtro IIR passa-baixas de primeira ordem, usado para suavizar a função de espalhamento. A seguir, modelou-se a influência dos componentes espectrais de menor freqüência sobre as freqüências mais elevadas. Para isso, calcula-se uma nova variável, denominada fração propagada, dada por

$$fp[k,n] = a \cdot d^{S_s[k,n]} + b \cdot fp[k-1,n]\,. \tag{6.39}$$

Como se pode observar, o fator $d^{S_s[k]}$ é suavizado por um filtro passa-baixas com coeficientes a e b. Por esse motivo, a Equação 6.39 pode ser implementada através do procedimento descrito na Subseção 6.4.1.9. Para cada instante de tempo n, os coeficientes fp são aplicados de acordo com o conjunto de equações a seguir:

$$\begin{cases} x_e[1,n] = x[1,n] \\ x_e[2,n] = fp[1,n]\cdot x[1,n] + x[2,n] \\ x_e[3,n] = \left(fp[1,n]\right)^2 \cdot x[1,n] + fp[2,n]\cdot x[2,n] + x[3,n] \\ \quad \vdots \\ x_e[40,n] = \left(fp[1,n]\right)^{39} \cdot x[1,n] + \left(fp[2,n]\right)^{38} \cdot x[2,n] + \cdots + x[40,n] \end{cases}, \tag{6.40}$$

onde $x[k,n]$ são os valores absolutos dos sinais resultantes após a ponderação dos ouvidos externo e médio. Como se pode ver, quanto maior a distância entre as amostras, maior o valor do expoente aplicado aos valores de fp, os quais, por sua vez, são sempre menores que 1. Assim, quanto maior o expoente, menor o valor multiplicado às amostras correspondentes dos sinais. Desta maneira, modela-se o fenômeno em que quanto mais distantes estão as bandas, menor a influência entre elas. A Equação 6.40 pode ser reescrita na forma matricial, resultando em

$$\begin{bmatrix} x_e[1,n] \\ x_e[2,n] \\ \vdots \\ x_e[39,n] \\ x_e[40,n] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ fp[1,n] & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \left(fp[1,n]\right)^{38} & \left(fp[2,n]\right)^{37} & \left(fp[3,n]\right)^{36} & \cdots & 1 & 0 \\ \left(fp[1,n]\right)^{39} & \left(fp[2,n]\right)^{38} & \left(fp[3,n]\right)^{37} & \cdots & fp[39,n] & 1 \end{bmatrix} \cdot \begin{bmatrix} x[1,n] \\ x[2,n] \\ \vdots \\ x[39,n] \\ x[40,n] \end{bmatrix}. \tag{6.41}$$

Como se pode observar, esta equação tem a mesma estrutura da Equação 6.27, portanto o mesmo procedimento pode ser aqui adotado, de acordo com as Equações 6.28 a 6.30. O próximo passo é adicionar a contribuição das freqüências mais elevadas sobre as freqüências mais baixas. Primeiro, calcula-se uma nova fração propagada, dada por

$$fl = d^{31}. \tag{6.42}$$

A seguir, o coeficiente *fl* é aplicado conforme o conjunto de expressões a seguir:

$$\begin{cases} x_f[1,n] = x_e[1,n] + fl \cdot x_e[2,n] + \cdots + fl^{38} \cdot x_e[39,n] + fl^{39} \cdot x_e[40,n] \\ x_f[2,n] = x_e[2,n] + fl \cdot x_e[3,n] + \cdots + fl^{37} \cdot x_e[39,n] + fl^{38} \cdot x_e[40,n] \\ \vdots \\ x_f[39,n] = x_e[39,n] + fl \cdot x_e[40,n] \\ x_f[40,n] = x_e[40,n] \end{cases}. \tag{6.43}$$

A utilização dos expoentes segue os mesmos princípios descritos para a Equação 6.40. Como o fator *fl* não varia com o tempo, ele pode ser aplicado diretamente através de uma operação matricial única:

$$\begin{bmatrix} x_f[1,1] & x_f[1,2] & \cdots & x_f[1,N-1] & x_f[1,N] \\ x_f[2,1] & x_f[2,2] & \cdots & x_f[2,N-1] & x_f[2,N] \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_f[39,1] & x_f[39,2] & \cdots & x_f[39,N-1] & x_f[39,N] \\ x_f[40,1] & x_f[40,2] & \cdots & x_f[40,N-1] & x_f[40,N] \end{bmatrix} = \\ = \begin{bmatrix} 1 & fl & \cdots & fl^{38} & fl^{39} \\ 0 & 1 & \cdots & fl^{37} & fl^{38} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & fl \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_e[1,1] & x_e[1,2] & \cdots & x_e[1,N-1] & x_e[1,N] \\ x_e[2,1] & x_e[2,2] & \cdots & x_e[2,N-1] & x_e[2,N] \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_e[39,1] & x_e[39,2] & \cdots & x_e[39,N-1] & x_e[39,N] \\ x_e[40,1] & x_e[40,2] & \cdots & x_e[40,N-1] & x_e[40,N] \end{bmatrix}, \tag{6.44}$$

onde N é o número de amostras no domínio do tempo.

A Figura 6.11 ilustra a aplicação do espalhamento espectral no modelo baseado no banco de filtros. Como se pode observar, adiciona-se todas as contribuições ao componente em questão. Tal procedimento é repetido para cada linha espectral.

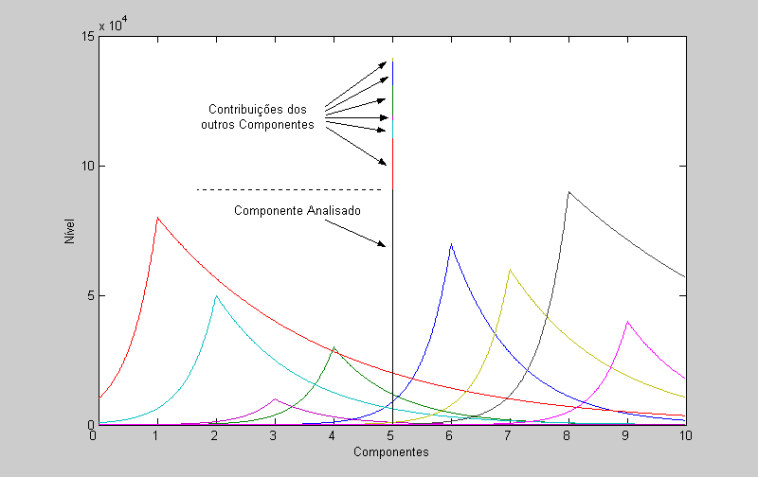


Figura 6.11 - Espalhamento espectral na versão baseada no banco de filtros

6.4.2.5. Cálculo da Energia dos Sinais

As energias são calculadas pela adição dos valores quadráticos dos padrões de saída dos filtros representando as partes real e imaginária do sinal, após seu espalhamento no domínio da freqüência, de acordo com

$$E_0[k,n] = {A_{re}}^2[k,n] + {A_{im}}^2[k,n]. \tag{6.45}$$

Todas as operações apresentadas a seguir são realizadas sobre estas energias.

6.4.2.6. Normalização

Um fator de normalização, consistindo da relação entre as energias dos sinais antes e depois da aplicação do mascaramento espectral, é aplicado aos sinais. O procedimento é similar àquele dado pelas Equações 6.19 e 6.20.

6.4.2.7. Modelagem do Mascaramento Temporal Retrógrado

A modelagem do mascaramento retrógrado é feita através do espalhamento das energias ao longo do tempo por um filtro FIR com uma resposta impulsiva do tipo cos^2, com 12 *taps*. Após a distribuição no tempo, as saídas são dizimadas por um fator de 6, de modo a reduzir a complexidade computacional das etapas subseqüentes. Os valores resultantes são multiplicados pelo fator de calibração de valor 0,9761, a fim de se ajustar a saída para o nível de execução adotado:

$$E_1[k,n] = \frac{0{,}9761}{6} \cdot \sum_{i=0}^{11} E_0[k, 6\cdot n - i] \cdot \cos^2\left(\pi \cdot \frac{(i-5)}{12}\right). \tag{6.46}$$

Aqui, como há uma nova dizimação dos sinais, desta vez por um fator de 6, seria desnecessário realizar a filtragem para todas as amostras. Por essa razão, a Equação 6.46 foi implementada da seguinte maneira:

$$\begin{aligned}\begin{bmatrix} E_1[k,1] & E_1[k,2] & \cdots & E_1[k,M] \end{bmatrix} = 0{,}1627 \cdot \begin{bmatrix} \cos^2\left(\frac{-5\pi}{12}\right) & \cos^2\left(\frac{-4\pi}{12}\right) & \cdots & \cos^2\left(\frac{6\pi}{12}\right) \end{bmatrix} \cdot \\ \cdot \begin{bmatrix} E_0[k,1] & E_0[k,7] & \cdots & E_0[k,N-17] & E_0[k,N-11] \\ E_0[k,2] & E_0[k,8] & \cdots & E_0[k,N-16] & E_0[k,N-10] \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ E_0[k,11] & E_0[k,17] & \cdots & E_0[k,N-15] & E_0[k,N-1] \\ E_0[k,12] & E_0[k,18] & \cdots & E_0[k,N-14] & E_0[k,N] \end{bmatrix}\end{aligned}, \tag{6.47}$$

onde N e M são, respectivamente, o número de amostras temporais por quadro antes e depois da dizimação. Como se pode observar, a dizimação é realizada automaticamente pela construção da matriz E_0. Os cálculos devem ser realizados individualmente para cada banda k, mas como os cálculos aqui realizados não envolvem um número muito grande de operações, este fato não compromete a velocidade de execução da etapa. Adicionalmente, a realização de uma convolução tal como descrita na Subseção 6.4.1.9, com a posterior dizimação dos sinais, teria uma complexidade computacional mais elevada.

6.4.2.8. Adição do Ruído Interno

O procedimento para adição do ruído interno para este modelo é idêntico àquele descrito na Subseção 6.4.1.6, observadas as diferenças quando às freqüências centrais adotadas para cada modelo.

6.4.2.9. Modelagem do Mascaramento Progressivo

Esta etapa é implementada da mesma maneira de sua correspondente no modelo baseado na FFT. As equações são as mesmas, com exceção do parâmetro *a*, o qual aqui é dado por

$$a = e^{-\frac{192}{48000 \cdot \tau}}. \tag{6.48}$$

As constantes de tempo também são diferentes: $\tau_{min} = 0{,}004$ e $\tau_{100} = 0{,}020$.

6.5. AJUSTE DOS SINAIS RESULTANTES

Os cálculos desta etapa têm por objetivo realizar ajustes necessários entre os sinais resultantes da aplicação dos modelos psico-acústicos para que estes possam ser apropriadamente explorados no cálculo dos parâmetros cognitivos. Os procedimentos apresentados valem para os sinais gerados a partir de ambos os modelos apresentados na Seção 6.4. Algumas constantes usadas nas equações apresentadas nesta seção são dependentes do modelo de origem dos sinais. São elas *z* e *ct*, cujos valores são de 109 e 1.024 para os sinais originados a partir do modelo baseado na FFT e 40 e 192 para os sinais gerados a partir do modelo baseado no banco de filtros.

6.5.1. Adaptação de Nível e Padrão

A fim de compensar diferenças de nível e distorções lineares entre os sinais de referência e teste, os níveis médios destes sinais são adaptados um ao outro. No primeiro passo, as energias de cada banda são suavizadas por filtros passa-baixas de primeira ordem. As constantes de tempo são dadas por

$$\tau[k] = \tau_{\min} + \frac{100}{f[k]} \cdot \left(\tau_{100} - \tau_{\min}\right) \Bigg|_{\substack{\tau_{100} = 0{,}050\ s \\ \tau_{\min} = 0{,}008\ s}}. \tag{6.49}$$

Como se pode observar na Equação 6.49, as constantes de tempo dependem das freqüências centrais dos filtros.

As saídas dos filtros passa-baixas de primeira ordem são computadas segundo

$$A_r[k,n] = a \cdot A_r[k,n-1] + (1-a) \cdot E_r[k,n], \tag{6.50}$$

$$A_t[k,n] = a \cdot A_t[k,n-1] + (1-a) \cdot E_t[k,n], \tag{6.51}$$

onde E_r e E_t são os sinais gerados para os sinais de referência e teste após a aplicação dos modelos para o ouvido, e *a* é dado por

$$a = e^{-\frac{ct}{48000 \cdot \tau}}. \tag{6.52}$$

A estratégia utilizada nos cálculos descritos nas Equações 6.50 a 6.52 é a mesma descrita em 6.4.1.9 para o cômputo de filtros IIR de primeira ordem.

6.5.1.1. Adaptação de Nível

O fator de correção momentâneo *CM* é dado por

$$CM[n] = \left(\frac{\sum_{k=1}^{Z} \sqrt{A_t[k,n] \cdot A_r[k,n]}}{\sum_{k=1}^{Z} A_t[k,n]} \right)^2 . \tag{6.53}$$

Esta expressão também foi implementada matricialmente (notação compacta):

$$\mathbf{cm} = \frac{\mathbf{u} \cdot (\mathbf{A_t} * \mathbf{A_r})}{\mathbf{u} \cdot \mathbf{A_t}}, \tag{6.54}$$

onde **u** é um vetor linha de dimensão *Z* cujos elementos são todos iguais a 1, e $*$ representa a multiplicação escalar entre os elementos das matrizes; a divisão também é escalar entre os elementos dos vetores resultantes. O vetor resultante **cm** tem dimensão igual ao número de amostras no domínio do tempo.

Se o fator de correção é maior que 1, o sinal de referência é dividido pelo fator de correção; caso contrário, o sinal de teste é multiplicado pelo fator de correção. Esta estratégia é implementada através das equações a seguir:

$$N_r[k,n] = E_r[k,n] / CM[n] \quad ; \quad CM[n] > 1 \tag{6.55}$$

$$N_t[k,n] = E_t[k,n] \cdot CM[n] \quad ; \quad CM[n] \leq 1. \tag{6.56}$$

As Equações 6.55 e 6.56 foram também aplicadas matricialmente, descartando a necessidade do uso de laços de iteração.

6.5.1.2. Adaptação de Padrão

Aqui são calculados fatores de correção para cada canal, através da comparação das envoltórias temporais dos sinais de referência e teste após a adaptação de nível, conforme a expressão

$$R[k,n] = \frac{\sum_{i=0}^{n-1} a^i \cdot N_t[k,n-i] \cdot N_r[k,n-i]}{\sum_{i=0}^{n-1} a^i \cdot (N_r[k,n-i])^2} . \tag{6.57}$$

A Equação 6.57 pode ser reescrita usando a notação matricial compacta:

$$\mathbf{R} = \frac{(\mathbf{N_t} * \mathbf{N_r}) \cdot \mathbf{A}}{(\mathbf{N_r} * \mathbf{N_r}) \cdot \mathbf{A}} . \tag{6.58}$$

Novamente, $*$ representa a multiplicação escalar entre os elementos das matrizes; a divisão também é do tipo escalar e a matriz **A** é dada por

$$\mathbf{A} = \begin{bmatrix} 1 & a & \cdots & a^{N-2} & a^{N-1} \\ 0 & 1 & \cdots & a^{N-3} & a^{N-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & a \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \tag{6.59}$$

onde N é o número de amostras no domínio do tempo.

Os valores de a são calculados como na Equação 6.52. Se $R[k,n]$ é maior que 1, o fator de correção para o sinal degradado é feito igual a $R[k,n]^{-1}$ e para o sinal original é feito igual a 1. No caso oposto, o fator de correção do sinal original é feito igual a $R[k,n]$ e o do sinal degradado é feito igual a 1. As expressões a seguir resumem o esquema adotado:

$$\begin{aligned} &R_t[k,n] = \frac{1}{R[k,n]}, \; R_r[k,n] = 1 \quad ; \quad R[k,n] \geq 1 \\ &R_t[k,n] = 1, \quad R_r[k,n] = R[k,n] \quad ; \quad R[k,n] < 1 \end{aligned}. \tag{6.60}$$

Se $R[k,n]$ é nulo (e então $R_t[k,n]$ é indefinido), $R_t[k,n]$ é feito igual a zero e $R_r[k,n]$ é feito igual a 1.

Calcula-se então a média temporal dos fatores de correção sobre M quadros adjacentes:

$$\begin{aligned} &CP_t[k,n] = a \cdot CP_t[k,n-1] + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_t[k+i,n] \\ &CP_r[k,n] = a \cdot CP_r[k,n-1] + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_r[k+i,n] \\ &M_1 = M_2 = \frac{M-1}{2} \quad ; \quad M \text{ ímpar} \\ &M_1 = \frac{M}{2} - 1, \quad M_2 = \frac{M}{2} \quad ; \quad M \text{ par} \end{aligned}, \tag{6.61}$$

onde as constantes de tempo são as mesmas vistas anteriormente e M é igual a 3 para o modelo baseado no banco de filtros e igual a 8 para a versão baseada na FFT. Nas bordas da escala de freqüência, a largura da janela de freqüência é reduzida de acordo com

$$M_1 = \min(M_1, k), \qquad M_2 = \max(M_2, z - k - 1), \qquad M = M_1 + M_2 + 1. \tag{6.62}$$

A implementação da Equação 6.61 segue a mesma estratégia adotada para os demais filtros IIR.

Os sinais de entrada adaptados em nível são ponderados com os fatores correspondentes, obtendo-se, dessa forma, os sinais espectralmente adaptados:

$$H_r[k,n] = N_r[k,n] \cdot CP_r[k,n], \tag{6.63}$$

$$H_t[k,n] = N_t[k,n] \cdot CP_t[k,n]. \tag{6.64}$$

6.5.2. Modulação

Esta etapa é aplicada somente aos sinais provenientes do modelo baseado no banco de filtros. Aqui, calcula-se uma sonoridade simplificada, a partir dos sinais anteriores à modelagem do mascaramento temporal progressivo. Este valor é distribuído ao longo do tempo de acordo com

$$E_m[k,n] = a \cdot E_m[k,n-1] + (1-a) \cdot E_2[k,n]^{0,3}. \tag{6.65}$$

A expressão a seguir representa uma modificação da Equação 6.65, denominada "derivação temporal" [83]:

$$E_d[k,n] = a \cdot E_d[k,n-1] + (1-a) \cdot \frac{48000}{ct} \cdot \left| E_2[k,n]^{0,3} - E_2[k,n-1]^{0,3} \right|, \tag{6.66}$$

onde $E_2[k,n]$ representa os sinais antes da aplicação do mascaramento temporal progressivo. Os valores de a são calculados como na Equação 6.48, com as constantes de tempo dadas pela Equação 6.49. O procedimento para aplicação da filtragem é o mesmo descrito em 6.4.1.9.

A partir dos valores resultantes, calcula-se uma medida para a modulação da envoltória em cada saída dos filtros, dada por

$$md[k,n] = \frac{E_d[k,n]}{1 + E_m[k,n]/0,3}. \tag{6.67}$$

6.5.3. Cálculo do Sinal de Erro

O sinal de erro é computado de forma diferente entre os dois modelos.

6.5.3.1. Modelo Baseado na FFT

O sinal de erro é calculado no domínio da freqüência, tomando-se a diferença entre as magnitudes dos espectros filtrados pelas características dos ouvidos externo e médio dos sinais original e degradado, conforme a equação

$$F_{ruído}[k,n] = \left| \left| F_{worig}[k,n] \right| - \left| F_{wdeg}[k,n] \right| \right|. \tag{6.68}$$

A seguir, computa-se a energia do sinal de erro e realiza-se o agrupamento em bandas críticas, de acordo com o procedimento descrito nas Subseções 6.4.1.4 e 6.4.1.5; como resultado, tem-se os padrões de energia do ruído ($P_{ruído}[k,n]$).

6.5.3.2. Modelo Baseado no Banco de Filtros

O sinal de erro para o modelo baseado no banco de filtros é dado simplesmente pelo módulo da diferença entre os sinais de referência e teste após a modelagem do mascaramento temporal progressivo, conforme a equação

$$P_{ruido}[k,n] = \left| E_{forig} - E_{fdeg} \right|. \tag{6.69}$$

6.5.4. Limiar de Mascaramento

Este limiar é obtido pela ponderação dos sinais de referência e teste após a modelagem do mascaramento temporal progressivo com a função dada por

$$m[k] = \begin{cases} 3{,}0 & \mathrm{k} \cdot res \leq 12 \\ 0{,}25 \cdot k \cdot res & \mathrm{k} \cdot res > 12 \end{cases}. \tag{6.70}$$

A ponderação é dada por

$$M[k,n] = \frac{E[k,n]}{10^{\frac{m[k]}{10}}}, \tag{6.71}$$

onde k representa as bandas espectrais e res é a resolução espectral em Bark, e seu valor é 0,25 para o modelo baseado na FFT e 0,6875 para o modelo baseado no banco de filtros. A Figura 6.12 ilustra os sinais resultantes desta etapa.

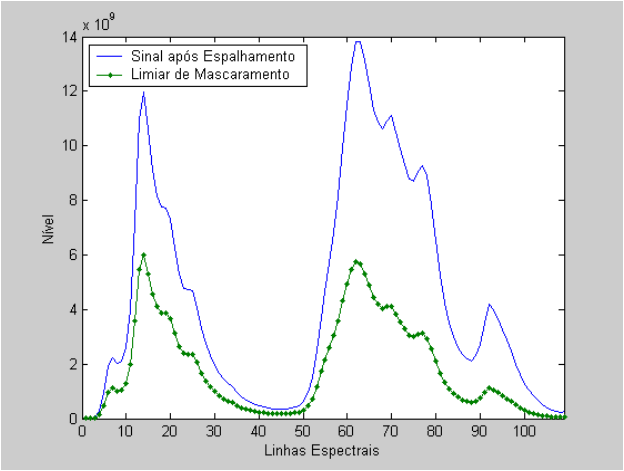


Figura 6.12 - Limiar de mascaramento

6.6. CÁLCULO DOS PARÂMETROS COGNITIVOS

Os parâmetros cognitivos consistem na extração de uma série de variáveis capazes de fornecer o máximo de informação dos sinais que se está analisando. Mais de 20 parâmetros diferentes foram testados para cada modelo; vários deles foram descartados por apresentarem uma contribuição questionável para um bom desempenho do método. Como resultado, adotou-se uma configuração de seis variáveis para cada modelo. As fórmulas apresentadas a seguir são aplicadas separadamente a cada um dos canais dos sinais. Os resultados obtidos para cada um deles são combinados através de uma média aritmética simples, resultando no valor final da variável.

6.6.1. Encadeamento Perceptual e Mascaramento Informacional

Estes dois conceitos, publicados em [86] e não utilizados no método PEAQ, foram combinados para a obtenção de um novo parâmetro cognitivo. A seguir, será feita uma breve descrição de suas principais características, as quais foram usadas na sua implementação.

- *Encadeamento perceptual*: é uma característica cognitiva do sistema auditivo humano que separa diferentes eventos auditivos e os agrupa em diferentes percepções. Se o sinal de entrada é distorcido de maneira que o sinal de saída é separado em duas partes, o

sinal original e a distorção, então o distúrbio provocado por essa distorção será mais intenso do que quando ambas as partes (sinal e distorção) são integradas em uma única percepção.

A modelagem do encadeamento perceptual na fala natural ou em sinais de música é complicada. A abordagem adotada em [86,95], bastante simplificada, assume que quando o sinal é degradado por uma distorção aditiva, é improvável que esse novo componente se integre perfeitamente com o sinal de entrada. Assim, o sinal de saída será decomposto nas duas diferentes percepções anteriormente descritas. Contudo, quando o codec elimina um componente, o sinal de saída não pode ser decomposto da mesma maneira e, portanto, a distorção é menos pronunciada. Este efeito é quantificado no cálculo deste parâmetro através de um fator de correção para o distúrbio ruidoso, que é a diferença perceptual entre os sinais. O fator de correção se baseia na relação entre as potências dos sinais de referência e teste em um certo ponto do plano tempo-freqüência, o que fornece uma medida da "novidade" deste componente, ou seja, indicando quão original, ou diferente, ele é em relação ao sinal de entrada.

Existem duas diferentes formulações para a modelagem do encadeamento perceptual: a primeira, mais simples e largamente difundida, foi usada nas medidas objetivas para avaliação da qualidade de voz PSQM [88] e MOQV [89]; a segunda, melhor e mais sofisticada, foi usada como base para as expressões aqui adotadas. Alguns trechos da formulação original sugerida em [86] foram alterados a fim de adequar as expressões às características presentes no método MOQA.

O primeiro passo no cálculo desta variável é ajustar a resolução temporal em 20 ms. Tal valor, determinado empiricamente em [86], é considerado suficiente para o propósito deste parâmetro. No modelo baseado na FFT, a resolução temporal é de cerca de 21 ms, muito próxima da sugerida, dispensando a necessidade de ajustes. A resolução temporal do modelo baseado no banco de filtros é, após as dizimações, de 4 ms. Assim, para que se alcançasse a resolução desejada, as amostras temporais foram agrupadas de acordo com

$$n_r[k,i] = \sum_{n=i-4}^{i} H_r[k,n], \qquad i = 1,\cdots,\lfloor N/5 \rfloor \tag{6.71}$$

$$n_t[k,i] = \sum_{n=i-4}^{i} H_t[k,n], \qquad i = 1,\cdots,\lfloor N/5 \rfloor. \tag{6.72}$$

Nas Equações 6.71 e 6.72, H é o padrão de excitação espectralmente adaptado, N é o número de amostras no domínio do tempo e os índices r e t representam os sinais de referência e teste, respectivamente. A operação $\lfloor a \rfloor$ indica o maior inteiro menor ou igual a a. Se $n \leq 0$, $H[k,n]$ é feito igual a 0. A seguir, determina-se o valor máximo dentre as amostras temporais para cada k, apenas para o sinal original, dado por

$$p_r[k] = \max(n_r[k,i]), \qquad i = 1,\cdots,\lfloor N/5 \rfloor. \tag{6.73}$$

Então, calcula-se a relação de potências entre os sinais, de acordo com as expressões

$$pp[k,i] = \frac{n_t[k,i]+1}{p_r[k]+1}, \tag{6.74}$$

$$epn[k,i] = 0{,}5 \cdot pp[k,i] + 0{,}5 \cdot epn[k,i-1]. \tag{6.75}$$

Na Equação 6.74 usou-se, no denominador, o valor máximo entre as amostras temporais do sinal de referência, ao invés de seus valores individuais, como na formulação original. Tal estratégia foi adotada como resultado de uma série de testes, os quais revelaram ser esta a melhor configuração. A soma do valor unitário em ambos os termos da divisão visa evitar o surgimento de valores nulos ou indeterminados. Quanto à Equação 6.75, pode-se notar que cada valor da relação é determinado por valores atuais e também por valores anteriores ponderados. Isto é feito a fim de se modelar o fato de que um distúrbio é tanto mais incômodo quanto mais longo é o período em que ele está presente.

Por fim, calcula-se o valor do encadeamento perceptual, dado por

$$ep[k,i] = \left(epn[k,i] \cdot \frac{\left| n_r[k,i] - n_t[k,i] \right|}{n_r[k,i]} \right)^{\alpha}, \tag{6.76}$$

onde $\alpha = 0{,}5$ é um parâmetro determinado empiricamente a fim de melhorar os resultados obtidos para esta variável. É importante ressaltar que os ajustes empíricos, usados aqui no cálculo de alguns dos parâmetros cognitivos, possuem o inconveniente de aumentar o número de graus de liberdade, o que teoricamente diminuiria a confiabilidade dos parâmetros frente a sinais e bases de dados desconhecidas. Contudo, esses ajustes assumem parte do papel que caberia à rede neural, ou seja, eles tentam modelar o processamento cognitivo que ocorre no nível do córtex cerebral humano. Esse procedimento faz com que seja mais simples projetar e treinar a rede neural que fará o mapeamento final entre valores objetivos e subjetivos.

É importante notar que o efeito de encadeamento perceptual leva a uma assimetria entre o distúrbio predito para uma distorção que é causada pela não codificação de um componente em relação ao distúrbio causado pela introdução de um novo componente. Esta propriedade de quebra de simetria faz com que a medida do distúrbio ruidoso não mais seja uma medida de distância, uma vez que $d(x,y) \neq d(y,x)$. Isto está de acordo com a observação de que a troca entre os sinais original e decodificado em um experimento subjetivo não levará a um mesmo valor subjetivo.

Os valores de *ep* são armazenados para posterior combinação com os valores do mascaramento informacional, conforme descrito a seguir.

- *Mascaramento Informacional*: é uma característica cognitiva do sistema auditivo humano em que distorções no sinal que deveriam ser audíveis, por estarem acima do limiar de audibilidade, tornam-se inaudíveis devido ao conteúdo informacional (complexidade) do sinal mascarador. O efeito do encadeamento perceptual no sistema auditivo pode diminuir o efeito do mascaramento informacional [96]. Quando um sinal complexo pode ser decomposto nos termos do encadeamento perceptual, o efeito do mascaramento informacional será menor que no caso de tal decomposição não ser possível. Por esse motivo, ambos os efeitos devem ser modelados em conjunto.

O efeito do mascaramento informacional é aqui implementado tendo como base a variação de potência no tempo para cada faixa de freqüência do sinal original. A variação é levada em conta no cálculo do distúrbio ruidoso para cada quadro temporal, de modo que sinais complexos, possuindo uma maior variação de potência, produzam um maior efeito de mascaramento no distúrbio ruidoso que sinais mais simples. Para cada banda *k* do quadro temporal atual *n*, sua potência é comparada com a potência da mesma banda nos quadros temporais anteriores.

O primeiro passo é o cálculo dos valores de desvio de potência. A densidade de potência do sinal em cada faixa de freqüência, n_r, é comparada com a potência média da mesma faixa de freqüência k calculada ao longo do quadro temporal corrente i e dos quatro quadros precedentes, conforme as expressões a seguir:

$$\begin{cases} p_{med}[k,i] = \dfrac{1}{5}\sum_{j=i-4}^{i} n_r[k,j] & \text{se} \quad i = 5,\cdots,\lfloor N/5 \rfloor \\ p_{med}[k,i] = \dfrac{1}{i}\sum_{j=1}^{i} n_r[k,j] & \text{se} \quad i = 1,\cdots,4 \end{cases}, \tag{6.77}$$

$$p_{xdev}[k,i] = \left| n_r[k,i] - p_{med}[k,i] \right|. \tag{6.78}$$

A seguir, calcula-se a média das potências locais ao longo de cada faixa de freqüência, como mostra a equação

$$p_{dev}[i] = \frac{1}{K}\sum_{k=1}^{K} p_{xdev}[k,i], \tag{6.79}$$

onde K é o número de bandas espectrais e p_{dev} é o valor do mascaramento informacional.

- *Combinação dos Valores*: o último passo no cálculo deste parâmetro é a combinação dos valores obtidos para o encadeamento perceptual e para o mascaramento informacional, dada por

$$cd[k,i] = \frac{ep[k,i]}{p_{dev}[i] + 10}, \tag{6.80}$$

$$em = \frac{1}{K \cdot \lfloor N/5 \rfloor} \cdot \sum_{k=1}^{K} \sum_{i=1}^{\lfloor N/5 \rfloor} cd[k,i]. \tag{6.81}$$

Na Equação 6.80 é realizada a combinação do encadeamento perceptual e do mascaramento informacional. A Equação 6.81 realiza a média dos valores obtidos após a combinação, resultando no valor do parâmetro cognitivo aqui descrito.

6.6.2. Diferença de Modulação

As expressões para o cálculo deste parâmetro, dadas pelas Equações 6.82 e 6.83, derivam de suas correspondentes usadas no PEAQ, com algumas modificações (introdução de ajustes empíricos).

$$dm_p[k,n] = \frac{\left| md_t[k,n] - md_r[k,n] \right|^{\alpha}}{(100 + md_r[k,n])^{\beta}}. \tag{6.82}$$

$$dm = \left(\sum_{k=1}^{K} \sum_{n=1}^{N} dm_p[k,n] \right)^{\delta}. \tag{6.83}$$

Nas equações, md_r e md_t são, respectivamente, as medidas de modulação para os sinais de referência e teste; K e N são, respectivamente, o número de amostras nos domínios da freqüência e do tempo; $\alpha = 2{,}3$, $\beta = 0{,}5$ e $\delta = 0{,}13$ são ajustes empíricos usados para ajustar

o parâmetro às características de projeto do método. O valor de *dm* corresponde à diferença de modulação entre os sinais.

6.6.3. Sonoridade do Ruído

Este parâmetro estima a sonoridade de distorções adicionadas ao sinal de referência. As expressões aqui utilizadas também são baseadas naquelas presentes no PEAQ, porém diversas modificações foram introduzidas. O cálculo da sonoridade da distorção presente em cada componente do sinal é dado por

$$sdi[k,n]=\left(\frac{1}{s_t[k,n]}\cdot P_l[k]\right)^{0,08}\cdot\left[\left(1+\frac{\max\left(s_t[k,n]\cdot H_t[k,n]-s_r[k,n]\cdot H_r[k,n],0\right)}{P_l[k]+s_r[k,n]\cdot H_r[k,n]}\right)^{0,08}-1\right], \quad (6.84)$$

onde P_l é a função de ruído interno como definido na Equação 6.9, H é o padrão de excitação espectralmente adaptado e s é dado por

$$s[k,n]=0{,}1\cdot md[k,n]+eps\,, \quad (6.85)$$

onde *md* são os valores de modulação (ver Seção 6.5.2) e *eps* é um valor arbitrariamente pequeno a fim de evitar que s assuma um valor nulo.

O próximo passo é calcular a sonoridade combinada do ruído em cada instante de tempo, através da expressão

$$sd[n]=\frac{1}{K}\cdot\sum_{k=1}^{K}sdi[k,n]\,. \quad (6.86)$$

Os valores de *sd* menores que 0,2 são feitos iguais a zero, por se considerar que nesses casos o ruído não é percebido. A última etapa no cálculo deste parâmetro é a determinação da média dos distúrbios ruidosos ao longo do tempo, dada por

$$sr=\frac{1}{N}\cdot\sum_{n=1}^{N}sd[n]\,. \quad (6.87)$$

O resultado desta expressão corresponde à sonoridade do ruído.

6.6.4. Relação Ruído-Mascaramento

A relação ruído-mascaramento em cada instante de tempo n é dada por

$$NMR_{local}[n]=10\cdot\log_{10}\left(\frac{1}{K}\sum_{k=0}^{K-1}\frac{\left(P_{ruído}[k,n]\right)^{\alpha}}{\left(M[k,n]\right)^{\beta}}\right), \quad (6.88)$$

onde P_{ruido} é o sinal de erro, cujo cálculo é apresentado na Seção 6.5.3, M é o limiar de mascaramento, calculado de acordo com as Equações 6.70 e 6.71, e $\alpha = 0{,}3$ e $\beta = 0{,}4$ são parâmetros determinados empiricamente. O valor final para a relação ruído-mascaramento é dado por

$$NMR=\frac{1}{N}\cdot\sum_{n-1}^{N}NMR_{local}[n]\,. \quad (6.89)$$

6.6.5. Número Relativo de Amostras com Distúrbios

Este parâmetro foi implementado originalmente no PEAQ, porém calculando o número relativo de quadros, e não amostras, com distúrbios. A abordagem usada no PEAQ mostrou-se excessivamente sensível a diversas situações, tornando seu uso pouco confiável. Por esse motivo, algumas modificações foram introduzidas. As expressões resultantes são muito simples e nenhum parâmetro empírico foi utilizado, conforme se pode observar nas equações a seguir:

$$rr[k,n] = 10 \cdot \log_{10}\left(\frac{P_{ruido}[k,n]}{M[k,n]}\right), \tag{6.90}$$

$$nr = \frac{Z(rr[k,n])}{K \cdot N}. \tag{6.91}$$

O operador Z na Equação 6.91 faz a contagem do número de amostras para as quais rr assume um valor maior ou igual a 0,9 dB. Como se pode observar, o que se faz é simplesmente calcular a relação entre o ruído e o mascaramento e então calcular o número relativo de amostras para as quais essa relação ultrapassa certo nível em dB. Apesar de simples, esta variável apresentou bons resultados, conforme será visto no próximo capítulo.

6.6.6. Probabilidade de Detecção de Distúrbios

Este parâmetro consiste na determinação da probabilidade de determinado distúrbio ser detectado por um ouvinte com audição normal. As probabilidades obtidas para cada componente são ponderadas e combinadas de modo a resultar num único valor para o parâmetro. Quanto maior a probabilidade, pior a qualidade do sinal.

O primeiro passo no cálculo desta variável é a determinação dos valores dos padrões de excitação em dB, de acordo com a equação

$$HD[k,n] = 10 \cdot \log_{10}(H[k,n]). \tag{6.92}$$

A seguir, determina-se a excitação máxima entre os dois sinais, de acordo com a equação

$$L[k,n] = \max(HD_r, HD_t). \tag{6.93}$$

A partir desse valor, calcula-se a mínima diferença detectável, tal como sugerida em [28]:

Se $L[k,n] > 0$:

$$d[k,n] = 5{,}95072 \cdot \left(\frac{6{,}39468}{L[k,n]}\right)^{1{,}71332} + 9{,}01033 \cdot 10^{-11} \cdot (L[k,n])^4 + 5{,}05622 \cdot 10^{-6} \cdot (L[k,n])^3 -$$
$$-0{,}00102438 \cdot (L[k,n])^2 + 0{,}0550197 \cdot L[k,n] - 0{,}198719 \tag{6.94}$$

Senão

$$d[k,n] = 1{,}0 \cdot 10^{30}$$

Na seqüência, calcula-se a diferença entre os padrões de excitação em dB, dada por

$$dif[k,n] = \left|HD_r[k,n] - HD_t[k,n]\right|. \tag{6.95}$$

Essa diferença é então ponderada pela resposta em freqüência dos ouvidos externo e médio (*W*), de acordo com a equação

$$dfp[k,n] = dif[k,n] \cdot W[k], \tag{6.96}$$

onde *W* é calculado de acordo com a Equação 6.2.
O próximo passo consiste no cálculo do fator de escala *a*, dado por

$$a[k,n] = \frac{10^{\frac{\log_{10}[\log_{10}(1,8)]}{b[k,n]}}}{s[k,n]}, \tag{6.97}$$

onde *b* é um fator de inclinação que assume valor 4 ou 6, dependendo se o padrão de excitação do sinal original (HD_r) é, respectivamente, maior ou menor que o padrão de excitação do sinal degradado (HD_t).

A probabilidade de detecção de cada amostra é dada por

$$pa[k,n] = 1\text{-}10^{-(a[k,n] \cdot dfp[k,n])^b}. \tag{6.98}$$

A probabilidade de detecção em cada instante de tempo é dada por

$$pt[n] = 1 - \prod_{k=1}^{K} (1\text{-}pa[k,n]). \tag{6.99}$$

Por fim, a probabilidade de detecção final é dada por

$$pd = \frac{1}{N} \cdot \sum_{n=1}^{N} pt[n]. \tag{6.100}$$

6.7. MAPEAMENTO ENTRE VALORES OBJETIVOS E SUBJETIVOS

Para o mapeamento entre as variáveis de saída do modelo e os valores subjetivos correspondentes foram utilizados dois diferentes tipos de redes neurais:

1- Redes neurais do tipo MLP (*Multi-Layer Perceptron*) com uma camada oculta, usando funções de ativação do tipo tangente hiperbólica na camada oculta e do tipo linear no único neurônio da camada de saída. No treinamento, utilizou-se um método de otimização de segunda ordem do tipo Levenberg-Marquardt [97,98], com um critério de otimização baseado nos mínimos quadrados.

2- Redes de Kohonen com arranjo de neurônios em uma ou duas dimensões e diferentes critérios de vizinhança.

As configurações utilizadas para cada uma das redes, bem como os desempenhos obtidos, serão apresentados em detalhes no Capítulo 7.

CAPÍTULO 7
TESTES E VALIDAÇÃO DO MÉTODO MOQA

Diversos testes foram realizados no intuito de se determinar a configuração mais adequada para uma estimação confiável da impressão subjetiva dos usuários com relação à qualidade dos sinais de áudio. Dezenas de parâmetros foram testados até que se chegasse ao conjunto descrito no Capítulo 6 [99,100]. Cada um desses parâmetros possui informações importantes a respeito dos sinais testados. Para que tais informações fossem adequadamente exploradas, houve a necessidade de se utilizar redes neurais, pois estas possuem uma capacidade inerente para a realização de mapeamentos multidimensionais não-lineares, como é o caso aqui apresentado.

Este capítulo apresenta as configurações mais bem sucedidas e os resultados por elas alcançados. Os resultados são comparados com aqueles obtidos para o método PEAQ [83], a fim de validar o programa MOQA.

7.1. BASES DE DADOS UTILIZADAS

As bases de dados contendo arquivos de áudio e respectivas medidas subjetivas de desempenho não estão disponíveis para o público em geral, sendo por isso muito difíceis de serem obtidos. Um grande esforço foi dispensado na tentativa de se obter as 10 bases de áudio utilizadas na recomendação BS.1387-1 para validação do método PEAQ. Por fim, 3 delas foram disponibilizadas em caráter excepcional para a pesquisa aqui realizada. Essas bases foram usadas no desenvolvimento e calibração dos modelos perceptuais e na configuração dos parâmetros cognitivos. Devido ao número relativamente pequeno de sinais disponíveis, parte da pesquisa foi prejudicada. Em particular, havia a intenção de se aperfeiçoar o processamento temporal dos modelos psico-acústicos utilizados, pois este é um dos pontos mais sensíveis para o bom funcionamento das medidas objetivas de avaliação de áudio, sendo reconhecidamente um ponto fraco nos métodos disponíveis [25]. Contudo, esta seria uma tarefa que exigiria uma quantidade muito maior de sinais do que aquela disponível, de modo que não foi possível colocá-la em prática.

Após a determinação dos modelos psico-acústicos e dos parâmetros cognitivos, o Dr. Thilo Volker Thiede, autor da referência [25], ofereceu-se para extrair os parâmetros para outras 5 bases de áudio. Isto possibilitou que o número de amostras disponíveis para a determinação e treinamento das redes neurais quase triplicasse, tornando os resultados muito mais confiáveis. A seguir, é feita uma breve descrição do conteúdo dessas bases.

7.1.1. Descrição das Bases de Dados

Cada base de dados utilizada possui uma determinada quantidade de pares de sinais (referência e teste), os quais são cuidadosamente escolhidos de maneira a representar as mais diferentes características temporais e espectrais que poderiam ser encontradas na prática. Cada versão degradada é gerada variando o tipo e o número de codecs e usando

diferentes taxas de bits. Os sinais são amostrados a 48 kHz, usando 16 bits para representação de cada amostra. Sua duração varia entre 10 e 40 segundos

Os valores subjetivos correspondentes a cada par de sinais são dados em termos de uma escala de diferença subjetiva (EDS), a qual é determinada pela subtração entre as notas atribuídas pelos ouvintes para os sinais original e degradado, de acordo com a escala apresentada no Capítulo 4; estes valores são, em geral, negativos, mas valores positivos podem ocorrer em casos onde ambos os sinais apresentem uma qualidade muito próxima entre si [83]. A descrição detalhada de algumas das bases de dados pode ser encontrada em [83,101-103].

7.2. CONFIGURAÇÃO DOS TESTES

Como comentado anteriormente, dois tipos de redes neurais foram testados no mapeamento entre os valores objetivos e subjetivos. Diversos testes foram realizados a fim de se determinar as configurações mais adequadas. Esta seção descreve tais testes e a configuração final adotada para cada um dos tipos de rede neural. Adicionalmente, uma subseção é dedicada à descrição dos conjuntos usados no treinamento e no teste das configurações implementadas.

7.2.1. Mapeamento Usando Redes Neurais do Tipo MLP

A rede neural do tipo MLP implementada utilizou funções de ativação do tipo tangente hiperbólica para os neurônios da camada oculta e do tipo linear no neurônio da camada de saída. Maiores detalhes a respeito deste tipo de rede neural podem ser encontrados em [104]. No treinamento, utilizou-se um método de otimização de segunda ordem do tipo Levenberg-Marquardt [97,98], com um critério de otimização baseado nos mínimos quadrados.

O uso deste tipo de rede neural para o mapeamento entre os valores objetivos e subjetivos apresentou algumas peculiaridades, cujos efeitos demandaram a criação de estratégias apropriadas. Uma das particularidades observadas diz respeito ao número de neurônios na camada oculta. Observou-se que este fator não exerce um papel importante no desempenho do método, uma vez que as correlações obtidas não variaram significativamente à medida que este era modificado. Este comportamento, em princípio estranho, pode ser explicado pela própria característica dos parâmetros usados na entrada da rede, bem como pelo comportamento da superfície de mapeamento gerada pela rede neural. Devido à heterogeneidade dos arquivos usados, a rede neural é obrigada a gerar uma superfície de mapeamento bastante complexa e "acidentada". Contudo, como a informação fornecida à rede pelos parâmetros não é perfeita, quanto mais a rede tenta seguir os padrões determinados pelos parâmetros presentes no conjunto de treinamento, mais inadequada se torna a superfície de mapeamento na tarefa de estimar a qualidade subjetiva dos sinais presentes no conjunto de teste, com uma conseqüente queda da correlação entre os parâmetros objetivos e as medidas subjetivas. Este problema poderia ser minimizado através da criação de novos parâmetros capazes de extrair informação de melhor qualidade dos sinais testados.

A estratégia adotada para evitar este tipo de problema foi monitorar o desempenho da rede frente ao conjunto de teste após cada atualização dos pesos; o conjunto de pesos selecionado em cada treinamento foi aquele que resultou no maior valor de correlação. Em

mais de 95% dos casos, o conjunto de pesos escolhido foi obtido até no máximo a décima atualização de pesos. Isto ocorre porque, no início do treinamento, a superfície gerada ainda é suave, ou seja, apesar de a rede ainda não apresentar uma adaptação rigorosa ao conjunto de treinamento, o mapeamento resultante atende de maneira razoável a todos os arquivos presentes tanto no conjunto de treinamento quanto no de teste. À medida que a rede se adapta melhor às peculiaridades do conjunto de treinamento, a superfície de mapeamento tende a assumir um formato inadequado para mapear sinais cujas características não estejam contempladas no conjunto de treinamento.

Assim, ao se interromper o treinamento da rede ainda no seu início, está-se anulando a principal vantagem de se utilizar um número maior de neurônios, que é justamente a de permitir à rede gerar superfícies de mapeamento mais complexas e adaptáveis. Por esse motivo, o desempenho não varia significativamente ao se alterar o número de neurônios. O número de seis neurônios na camada oculta mostrou ser o mais apropriado.

7.2.2. Mapeamento Usando Redes de Kohonen

As redes, ou mapas, de Kohonen (Kohonen Self-Organizing Maps - KSOM) são arranjos de neurônios artificiais que estabelecem e preservam a noção de vizinhança [105]. Se tais mapas possuem a capacidade de auto-organização, então eles podem ser aplicados a problemas de agrupamento e classificação. As topologias mais largamente utilizadas têm os neurônios organizados em estruturas uni ou bidimensionais.

A entrada da rede consiste de um conjunto de parâmetros, selecionados de maneira a fornecer a maior quantidade possível de informação a respeito dos elementos que se deseja classificar. Cada entrada é ponderada por um valor sináptico, adequadamente determinado por um processo de treinamento, o qual é fundamentado na lei de aprendizado competitivo. A competição produzirá apenas um neurônio ativo para cada entrada (*winner-takes-all*). A ativação de tal neurônio terá certa influência, previamente determinada, sobre os demais. A adaptação de pesos é dada por

$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) + \gamma \cdot \left(\mathbf{x}(k) - \mathbf{w}_j(k)\right), \tag{7.1}$$

onde γ é o parâmetro que determina a taxa de aprendizagem. Se cada classe é representada por mais de um neurônio, não somente o neurônio vencedor deve ser ajustado, mas também seus vizinhos, de acordo com algum critério pré-determinado. Após o processo de treinamento, os neurônios devem ser rotulados, de maneira que cada um corresponda a uma classe particular.

Então, quando um conjunto de parâmetros relativos a determinado elemento a ser classificado é fornecido à rede, os neurônios se tornarão ativos, e o maior valor de ativação determinará o neurônio vencedor. Como cada neurônio representa uma classe, o vencedor indicará a classe à qual pertence o elemento analisado. A Figura 7.1 mostra como é estruturado o mapeamento com o uso dos mapas de Kohonen.

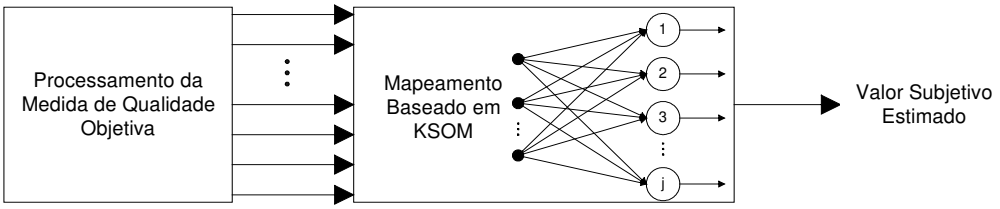


Figura 7.1 - Mapas de Kohonen Aplicados ao Mapeamento entre Medidas Objetivas e Subjetivas.

Como se pode observar na figura, vários valores objetivos podem ser mapeados para um único valor subjetivo. Esta estrutura pode explorar de maneira eficiente a informação contida nos parâmetros objetivos, conforme será visto nas próximas seções.

A estratégia de mapeamento aqui adotada pode ser dividida em 3 estágios:

1 - *Quantização dos Dados*: o princípio de funcionamento das redes de Kohonen é bastante diferente das estratégias de mapeamento clássicas, sejam elas funções polinomiais monotônicas ou redes neurais do tipo MLP. Estas últimas buscam aproximar funções que realizem o mapeamento desejado, capacidade esta não encontrada nas redes de Kohonen. Ao invés disso, elas realizam uma classificação de padrões, ou seja, reúnem elementos com características semelhantes em grupos afins. Os valores subjetivos, por sua vez, podem assumir qualquer valor real dentro de determinado intervalo, podendo, portanto, assumir infinitos valores. Assim, para que a rede de Kohonen seja capaz de estimar a qualidade subjetiva, é necessário quantizar os valores subjetivos de referência, a fim de se obter um número finito de níveis alvo.

Esta divisão em classes causa uma perda na qualidade do mapeamento, mas se a classificação realizada pela rede é confiável, a degradação causada por este tipo de aproximação terá pouco impacto na correlação final. Após uma cuidadosa investigação do comportamento da rede sob diferentes resoluções de quantização, escolheu-se o número de 20 classes, resultando em passos de aproximadamente 0,2 SDG (Subjective Degradation Grade). Para maiores informações sobre as medidas subjetivas, ver Capítulo 4.

2 - *Definição dos Parâmetros de Treinamento*: os fatores de treinamento investigados foram os seguintes:

a) Taxa de aprendizagem: adotou-se uma taxa de aprendizagem variável, variando de 1 no início do treinamento a 0,1 no final; taxas fixas levaram a resultados inferiores.

b) Ordem da vizinhança: foram testadas vizinhanças de primeira a quinta ordem. Uma configuração com ordem de vizinhança variável foi também testada, mas seu desempenho foi equivalente à configuração fixa. Como esta última tem uma complexidade computacional mais baixa, foi adotada como padrão. Os melhores resultados foram obtidos usando uma vizinhança de segunda ordem. Os neurônios localizados nos limites do arranjo são considerados vizinhos entre si.

c) Inicialização dos pesos: os pesos iniciais foram aleatoriamente gerados utilizando-se uma distribuição uniforme compreendida entre 0,1 e 1; outras gamas de valores foram testadas, todas apresentando um desempenho inferior.

3 - *Definição da Arquitetura da Rede*: na definição da topologia final da rede, dois fatores foram investigados:

a) Arranjo dos neurônios: foram testados arranjos unidimensionais e bidimensionais. No primeiro caso, cada neurônio tem apenas dois vizinhos de primeira ordem, dois vizinhos de segunda ordem, e assim por diante, conforme ilustrado na Figura 7.2.

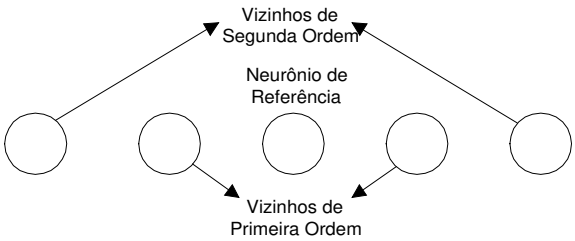


Figura 7.2 - Arranjo unidimensional de neurônios

No segundo caso, o neurônio de referência tem 8 vizinhos de primeira ordem, 16 vizinhos de segunda ordem e *8n* vizinhos de n-ésima ordem, como mostra a Figura 7.3.

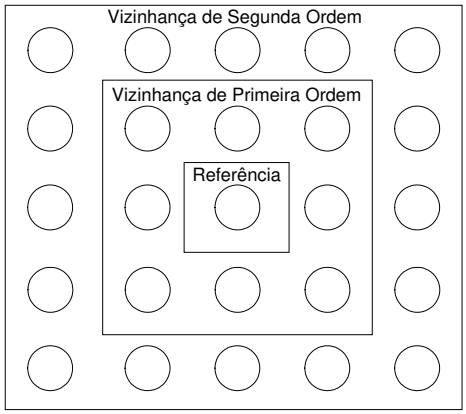


Figura 7.3 - Arranjo bidimensional de neurônios

Testes preliminares revelaram que o arranjo bidimensional não resultou em melhores resultados; como esta estratégia é computacionalmente mais exigente, decidiu-se pelo uso da configuração unidimensional.

b) Número de neurônios: foram realizados testes com 60 a 180 neurônios. Por fim, determinou-se que o número de 5 neurônios por classe (100 no total) é suficiente para uma classificação adequada dos padrões.

O desempenho da rede adotada quando aplicada ao problema em questão será apresentado mais adiante.

7.2.3. Determinação dos Conjuntos de Treinamento e Teste

Antes de se iniciarem os testes com as redes neurais, determinou-se quais conjuntos de parâmetros fariam parte do treinamento ou dos testes. Os conjuntos de parâmetros extraídos a partir dos 593 pares de sinais presentes nos 8 bases de dados foram divididas em grupos, de acordo com a Tabela 7.1.

Tabela 7.1 - Divisão dos conjuntos de parâmetros em grupos

	Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**	**Grupo 6**
Número de Pares de Sinais	98	99	99	99	99	99

Os grupos resultantes dessa divisão foram então usados na determinação de 6 diferentes conjuntos de treinamento e teste, conforme mostrado na Tabela 7.2.

Tabela 7.2 - Determinação dos conjuntos de treinamento e teste

	Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**	**Grupo 6**
Conjunto 1	Teste	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento
Conjunto 2	Treinamento	Teste	Treinamento	Treinamento	Treinamento	Treinamento
Conjunto 3	Treinamento	Treinamento	Teste	Treinamento	Treinamento	Treinamento
Conjunto 4	Treinamento	Treinamento	Treinamento	Teste	Treinamento	Treinamento
Conjunto 5	Treinamento	Treinamento	Treinamento	Treinamento	Teste	Treinamento
Conjunto 6	Treinamento	Treinamento	Treinamento	Treinamento	Treinamento	Teste

Como se pode observar, 5/6 dos sinais foram usados no treinamento em cada conjunto, e o restante nos testes. Todas as redes neurais (MLP e Kohonen) foram testadas frente a cada um dos conjuntos apresentados na Tabela 7.1, como será visto nas próximas seções. Essa estratégia foi adotada para que todos os parâmetros participassem, em algum momento, da etapa de testes, fornecendo assim uma visão muito mais abrangente do desempenho do método. A adoção desta abordagem evita ainda que conjuntos de parâmetros extraídos a partir de sinais potencialmente problemáticos façam parte apenas do treinamento, o que impediria que se pudesse inferir o desempenho do método frente a tais situações no caso destas não estarem representadas no conjunto de teste.

7.3. RESULTADOS

Os resultados apresentados nesta seção estão todos em termos das correlações entre valores objetivos e subjetivos, as quais são calculadas através do coeficiente de correlação de Pearson [106]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(z_i - \bar{z})^2}}, \tag{7.2}$$

onde x_i e z_i representam, respectivamente, os valores subjetivos e os valores objetivos após o mapeamento, enquanto $\bar{x}$ e $\bar{z}$ representam suas médias. Quanto mais próximo de 1 for o módulo de r, melhor é a qualidade do método e do mapeamento.

A seguir serão apresentados os resultados obtidos para cada parâmetro cognitivo individualmente, bem como o desempenho alcançado através do uso das diferentes configurações das redes apresentadas na Seção 7.2.

7.3.1. Resultados Individuais dos Parâmetros Cognitivos

Conforme comentado no Capítulo 6, seis diferentes parâmetros cognitivos foram extraídos para cada modelo. A Tabela 7.3 apresenta as correlações obtidas individualmente para cada um desses parâmetros. Esses valores são obtidos correlacionando-se o valor objetivo representado por um dado parâmetro e o valor subjetivo correspondente. As 3 bases originais citadas na tabela são aquelas que estavam disponíveis desde o início das pesquisas, enquanto que as demais só puderam ser utilizadas na validação do método.

Tabela 7.3 - Correlações obtidas para cada parâmetro

Modelo	Parâmetro	3 Bases Originais	Todas as Bases
FFT	*Enc. Perceptual e Masc. Informacional*	0,583	0,486
	Diferença de Modulação	0,694	0,688
	Sonoridade do Ruído	0,709	0,714
	Relação Ruído-Mascaramento	0,712	0,652
	Número de Amostras com Distúrbio	0,705	0,596
	Probabilidade de Detecção	0,600	0,548
Banco de Filtros	*Enc. Perceptual e Masc. Informacional*	0,681	0,513
	Diferença de Modulação	0,808	0,755
	Sonoridade do Ruído	0,830	0,789
	Relação Ruído-Mascaramento	0,718	0,600
	Número de Amostras com Distúrbio	0,673	0,536
	Probabilidade de Detecção	0,693	0,559

Os resultados apresentados na Tabela 7.3 estão divididos em dois grupos, o primeiro apresentando os resultados obtidos apenas para as bases de dados disponíveis ao longo de toda a pesquisa, e o segundo apresentando os resultados obtidos com todas as oito bases de dados usadas nos testes. Comparando-se os valores, torna-se claro o prejuízo acarretado pela indisponibilidade de todas as bases em todas as etapas de desenvolvimento do método. Ainda assim, os resultados obtidos podem ser considerados muito bons. Tal afirmação se torna evidente ao se constatar que a correlação obtida para a sonoridade do ruído na versão baseada no banco de filtros (0,789) é maior que as correlações obtidas por todos os antecessores do PEAQ. Tal valor é elevado mesmo quando comparado às correlações obtidas pelo PEAQ, as quais não ultrapassaram 0,85 [83], pois este é apenas um parâmetro individual, ainda não submetido a nenhum tipo de mapeamento.

Altas correlações indicam parâmetros de boa qualidade; contudo, se um parâmetro apresenta uma baixa correlação, isto não implica que este seja inadequado, uma vez que ele pode carregar informações importantes para que a rede neural seja capaz de realizar um mapeamento eficiente.

Quanto à diferença entre as versões, pode-se notar que aquela baseada no banco de filtros apresentou um desempenho global ligeiramente superior à versão baseada na FFT. A seguir, serão apresentadas as estratégias mais bem sucedidas na combinação dos parâmetros cognitivos para estimação da qualidade subjetiva.

7.3.2. Resultados Obtidos para as Estratégias Selecionadas

A Tabela 7.4 apresenta os resultados obtidos usando as diferentes estratégias de mapeamento adotadas.

Tabela 7.4 - Resultados gerais para as diferentes estratégias de mapeamento

	MLP	Kohonen (Completo)	Kohonen (Banco de Filtros)
Conjunto 1	0,888	0,885	**0,902**
Conjunto 2	0,871	0,901	**0,902**
Conjunto 3	0,882	0,896	**0,909**
Conjunto 4	0,870	**0,890**	0,886
Conjunto 5	0,874	0,901	**0,914**
Conjunto 6	0,815	0,818	**0,840**

Como se pode observar, os resultados são apresentados para todos os 6 conjuntos determinados na Seção 7.2.3. Os resultados obtidos para o mapeamento usando as redes de Kohonen foram divididos em dois grupos, o primeiro usando todos os parâmetros cognitivos extraídos para ambas as versões do programa, enquanto que o segundo faz uso apenas dos parâmetros provenientes da versão baseada no banco de filtros.

As seguintes conclusões podem ser inferidas a partir da Tabela 7.4:

- O uso das redes de Kohonen levou a resultados superiores àqueles obtidos com o uso das redes neurais do tipo MLP [107-110]. Tal observação pode ser explicada pelo fato deste tipo de rede ser mais robusta à variação da qualidade dos dados a ela submetidos, enquanto que as redes do tipo MLP exigem dados de entrada mais uniformes e representativos. Além disso, as redes de Kohonen modelam de maneira mais eficiente o comportamento dos ouvintes num teste subjetivo [108]. Como o objetivo de se extrair parâmetros de boa qualidade em sinais de áudio nem sempre é alcançado, torna-se essencial a utilização de uma ferramenta de mapeamento capaz de lidar com as mais adversas condições, tarefa esta assumida de maneira mais eficiente pelas redes de Kohonen.

- A rede de Kohonen usando como entrada apenas os parâmetros extraídos a partir da versão baseada no banco de filtros apresentou resultados superiores àquela alimentada com todos os 12 parâmetros [108,109]. Isto ocorreu porque os parâmetros correspondentes extraídos em cada uma das versões carregam, basicamente, o mesmo tipo de informação, ou seja, eles são, em certa medida, redundantes. Como os parâmetros extraídos na versão baseada na FFT carregam informação de qualidade inferior, sua inclusão freqüentemente causa uma deterioração na eficiência da rede. Por esse motivo, optou-se pela exclusão de tais parâmetros.

- Nos testes realizados com a rede de Kohonen alimentada apenas com os parâmetros gerados a partir da versão baseada no banco de filtros, a maior parte dos conjuntos apresentou um desempenho semelhante, com correlações freqüentemente ultrapassando o valor de 0,9. A única exceção foi o conjunto 6, o qual apresentou resultados ligeiramente inferiores, pelo fato de alguns sinais cuja qualidade subjetiva é difícil de ser estimada corretamente terem sido incluídos na etapa de teste da rede.

7.4. CONFIGURAÇÃO FINAL

Tendo como base os resultados apresentados na Tabela 7.4, a configuração final para a estratégia de mapeamento foi definida de acordo com a Tabela 7.5:

Tabela 7.5 - Configuração final para a estratégia de mapeamento.

Rede Neural Utilizada	Kohonen
Arranjo dos Neurônios	Unidimensional
Número de Neurônios	100
Variáveis de Entrada	6, provenientes da versão baseada no banco de filtros
Conjunto de Pesos Utilizado	Resultante do treinamento realizado para o conjunto 5

Como se pode notar, apenas os parâmetros extraídos a partir da versão baseada no banco de filtros são utilizados, o que implica no descarte completo da versão baseada na FFT. Na versão final do programa, contudo, ao invés de sua eliminação total, optou-se apenas por sua desativação, mantendo-se as linhas de programação a ela pertencentes. Tal

procedimento possibilita que novos testes usando a FFT possam ser facilmente realizados no futuro, bastando para isso reativar o trecho correspondente do programa.

7.4.1. Detalhamento do Desempenho da Configuração Escolhida

Como visto na Tabela 7.4, a correlação alcançada para o esquema adotado foi de 0,914. Ao se aplicar esta configuração para os demais conjuntos de teste, observou-se uma correlação média próxima de 0,9, atestando a adequação do esquema escolhido. Além deste, alguns outros critérios para avaliação do desempenho do método foram utilizados, como o erro quadrático médio, a análise dos erros e a determinação dos casos cujo erro de estimação ultrapassou limites previamente estabelecidos.

O erro quadrático médio, determinado de acordo com a equação

$$eqm = \frac{1}{N} \cdot \sum_{n=1}^{N} \left(odg(n) - sdg(n) \right)^2 , \tag{7.3}$$

foi de 0,2441. Este valor pode ser considerado bom, especialmente considerando-se as dificuldades encontradas na avaliação de sinais de áudio. Na Equação 7.3, *odg* e *sdg* representam, respectivamente, os valores subjetivos estimados e reais, e *N* é o número de sinais de teste utilizados no cálculo.

A Tabela 7.6 mostra os erros dos valores subjetivos estimados em relação ao alvo, contendo as porcentagens por intervalo e o erro máximo registrado. A Figura 7.4 sintetiza os resultados na forma de um histograma.

Tabela 7.6 - Percentagens das estimativas para cada intervalo de erro absoluto.

Intervalo de erro absoluto	**<0.1**	**<0.2**	**<0.3**	**<0.4**	**<0.5**	**<0.6**	**<0.7**	**<0.8**	**<0.9**	**<1,0**	**<1,1**	**<1,2**	**<1,3**	**<1,4**	**Máx.**
Percentagem	21,2	38,4	51,5	64,7	71,7	78,8	84,9	87,9	88,9	91,9	97,0	98,0	98,0	100,0	1,39

Como se pode observar na Tabela 7.6 e na Figura 7.4, a maior parte das estimativas se mostrou precisa (mais de 50% dos casos apresentaram erros menores que 0,3, os quais podem ser considerados desprezíveis). Alguns poucos casos apresentaram erros elevados, devido às características presentes nos sinais testados correspondentes, como será comentado mais adiante.

A Rec. BS.1387-1 sugere que se adotem diferentes esquemas de tolerância de acordo com a qualidade do sinal analisado: quanto mais degradado estiver o sinal, maior o erro tolerado. Isto ocorre porque a maioria dos sinais analisados se localiza numa faixa de qualidade próxima da ideal. Portanto, há a necessidade de uma maior precisão para se diferenciar o desempenho de diferentes estratégias de codificação. Assim, uma análise simples dos erros entre os valores esperados e obtidos não é suficiente para inferir a qualidade do método de avaliação implementado. Na validação do método PEAQ, adotou-se um esquema de tolerância variável baseado na acuidade das medidas subjetivas disponíveis [83]. Esse nível de acuidade foi baseado nos valores dos intervalos de confiança obtidos para cada par de sinais, o qual determina o nível de certeza para a medida subjetiva medida. A estratégia adotada foi permitir uma maior tolerância para sinais com intervalos de confiança grandes (alto nível de incerteza). Infelizmente, os valores dos intervalos de confiança não foram disponibilizados para a pesquisa aqui realizada. Por esse motivo, uma estratégia alternativa foi desenvolvida. Os níveis de tolerância adotados são dados por

$$nt = 0{,}5 - \frac{0{,}5 \cdot sdg}{3{,}98}, \tag{7.4}$$

onde *sdg* é a medida subjetiva real para o sinal que se está avaliando e o valor de 3,98 no denominador do segundo termo é o módulo da menor *sdg* encontrada nos bancos de dados utilizados. Assim, são consideradas aceitáveis estimativas que estejam na faixa $sdg \pm nt$.

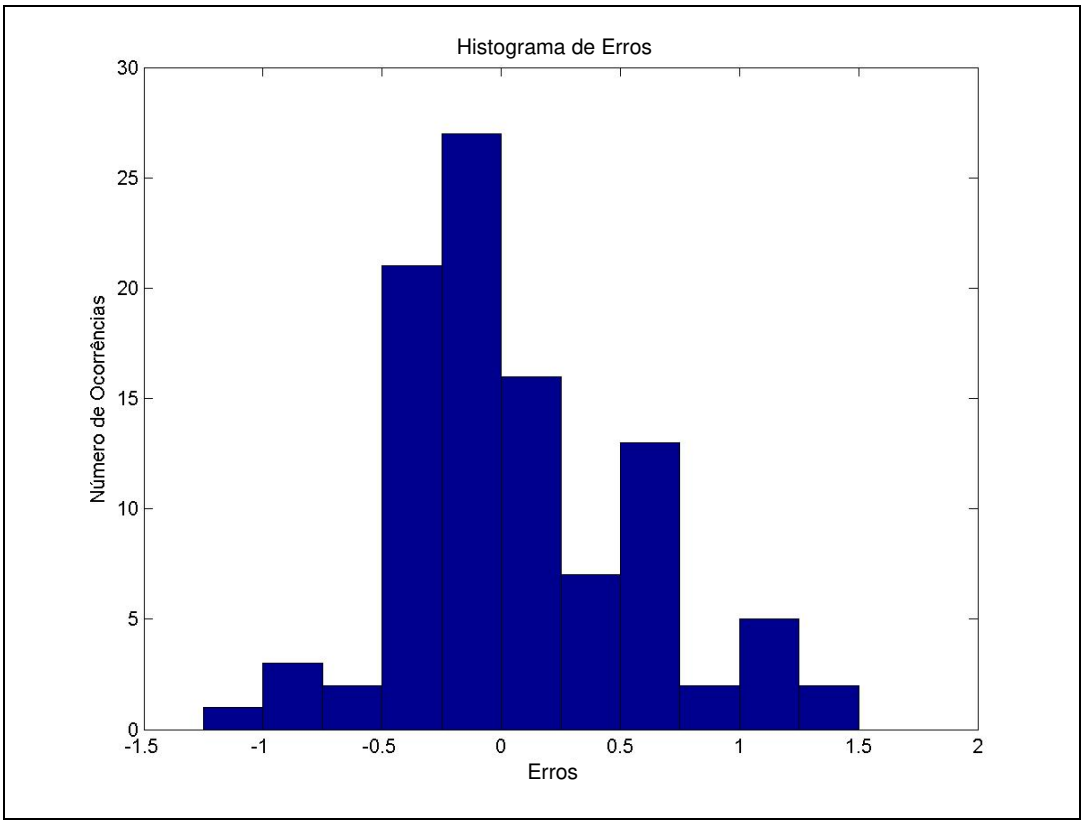


Figura 7.4 - Histograma dos erros.

Como se pode observar na equação, para uma qualidade perfeita (*sdg* = 0), a tolerância é de 0,5 para mais ou para menos. Conforme a *sdg* diminui (aumento da degradação) o nível de tolerância aumenta até um valor máximo de 1 no caso de uma degradação máxima. Para os poucos casos em que a *sdg* é positiva, *nt* será ligeiramente menor que 0,5. Esta estratégia de tolerância é, na média, próxima daquela adotada em [83].

A Figura 7.5 apresenta a curva de mapeamento ideal (linha cheia), o mapeamento efetivamente realizado entre as medidas objetivas e subjetivas (círculos) e os limites de tolerância (linhas pontilhadas).

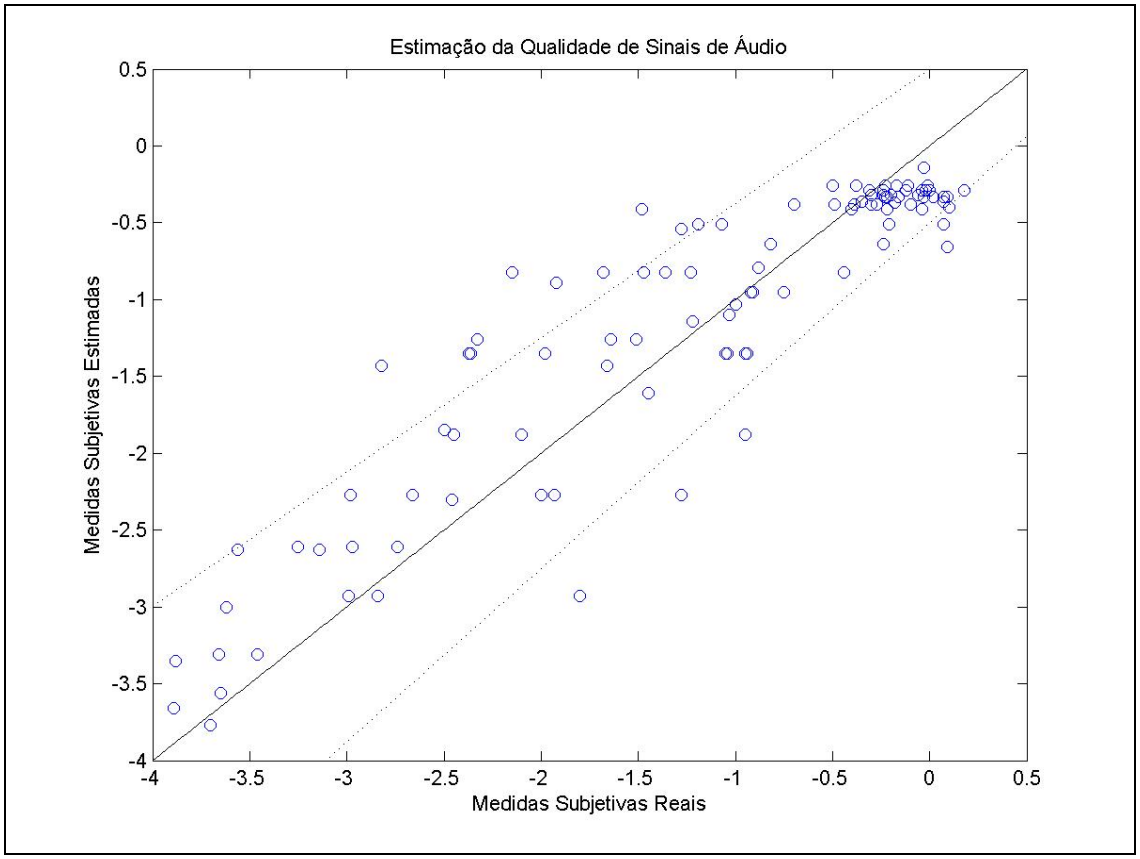


Figura 7.5 - Desempenho do método MOQA na estimação da qualidade subjetiva dos sinais.

Como se pode observar, a maior parte das estimativas se encontra dentro dos limites aceitáveis. A Tabela 7.7 quantifica os casos em que a qualidade da estimativa ficou aquém do desejado. O erro excedente é dado pela diferença entre o erro obtido e a máxima tolerância para o caso em questão.

Tabela 7.7 - Quantidade de estimativas com erros superiores à tolerância máxima.

	Faixas de erro excedente					
	0,0 – 0,1	**0,1 – 0,2**	**0,2 – 0,3**	**0,3 – 0,4**	**0,4 – 0,5**	**0,5 – 0,6**
Quantidade	4	1	5	3	1	2

De acordo com a Figura 7.5 e a Tabela 7.7, 16,2% dos casos analisados ultrapassaram a tolerância permitida. A maior parte desses casos problemáticos é composta por sinais submetidos a múltiplos codecs e por sinais contendo voz masculina.

7.4.2. Tempos de Simulação

A Tabela 7.8 apresenta os tempos de simulação (em segundos) para 3 versões do programa MOQA: 1) versão o uso das técnicas de vetorização, 2) versão usando as técnicas de vetorização e 3) versão com as técnicas de vetorização usando apenas o modelo do ouvido baseado no banco de filtros. Os tempos foram obtidos num microcomputador do tipo PC com processador AMD Athlon® 2000+, 512 MB de RAM e sistema operacional Microsoft Windows XP®. O sinal de áudio utilizado tinha uma duração de 20 segundos.

Tabela 7.8 - Tempos de execução do método MOQA

	Versão 1	Versão 2	Versão 3
Tempo de Execução (s)	69918,3	31,12	20,77

7.4.3. Comparação entre os Métodos MOQA e PEAQ

Os conjuntos de treinamento e teste usados na validação dos métodos PEAQ e MOQA são diferentes e, portanto, as condições enfrentadas podem variar em certo grau. Os autores do método PEAQ tiveram à sua disposição um conjunto de sinais muito mais amplo para o desenvolvimento do modelo psico-acústico e para o ajuste fino de parâmetros. Por outro lado, sua validação foi feita frente a um conjunto de apenas 32 pares de sinais, contra os 99 pares usados na validação do método MOQA. A conseqüência de se utilizar um conjunto de validação muito pequeno é o grande impacto que cada par de sinais tem no cálculo da correlação; assim, um único caso em que o erro tenha sido grande pode causar uma diminuição significativa no valor da correlação. Levando-se em conta essas diferenças, é possível realizar uma comparação crítica entre os dois métodos, através da análise de alguns dos parâmetros de desempenho:

- *Correlação*: as correlações usadas nas comparações são aquelas obtidas para a versão avançada do programa, a qual obteve os melhores resultados. O método PEAQ foi testado frente a dois conjuntos de sinais de áudio. Para o primeiro conjunto, contendo os 32 pares de sinais usados na validação do método, a correlação obtida foi de 0,828; para o segundo conjunto, cuja composição não foi disponibilizada na Rec. BS.1387, a correlação alcançada foi de 0,851. Como se pode notar, os resultados são inferiores não apenas àquele obtido pela configuração adotada para o método MOQA (0,914), como também àqueles alcançados para os diferentes conjuntos de treinamento e teste (ver Tabela 7.4).

- *Erros*: a Rec. BS.1387-1 considera um erro severo quando este é maior que 1, e muito severo quando o mesmo é maior que 1,5. A Tabela 7.9 apresenta a quantidade total e relativa de erros severos e muito severos para ambos os métodos.

Tabela 7.9 - Erros severos e muito severos.

Método	Nº Erros Severos	Nº Erros Muito Severos	Nº Sinais Testados	% Erros Severos	% Erros Muito Severos
PEAQ	4	2	32	12,5	6,25
MOQA	8	0	99	8,1	0

Como se pode observar na Tabela 7.8, o método MOQA também apresentou um melhor desempenho quando os erros absolutos são considerados.

Ao se considerar o número de casos cujo erro ultrapassou o limite de tolerância, não é possível fazer uma comparação direta, uma vez que os métodos utilizados para se determinar as faixas de tolerância são diferentes. Contudo, a diferença entre os limites de tolerância resultante da aplicação desses métodos não é grande, de forma que é possível fazer uma comparação indireta relativamente apurada. Este critério foi o que mostrou a maior diferença entre os métodos: enquanto no PEAQ 47,6% dos testes apresentaram um erro superior ao máximo desejado, no MOQA apenas 16,2% dos testes resultaram em erros excessivos, ou seja, um índice quase três vezes menor de falhas de estimação.

A seguir, serão apresentadas as conclusões a respeito do funcionamento do MOQA.

7.5. CONSIDERAÇÕES FINAIS

Todos os critérios de desempenho adotados revelaram uma superioridade do MOQA em relação ao PEAQ. Isto pode ser creditado, sobretudo, à robustez das redes de Kohonen frente à variação de qualidade dos parâmetros a elas submetidos e à sua capacidade de modelar de maneira eficiente o comportamento dos ouvintes em um teste subjetivo. Isso significa que a estratégia adotada faz um uso eficiente da informação contida nos parâmetros extraídos, mesmo nos casos em que esta informação está diluída devido às dificuldades causadas por determinadas características dos sinais. O ajuste empírico de alguns parâmetros cognitivos também contribuiu para que a rede fosse bem sucedida em estimar a qualidade subjetiva para a maior parte dos sinais testados.

O melhor desempenho foi alcançado sem o uso dos dados provenientes da versão baseada na FFT. Como conseqüência, o custo computacional do método, que já era relativamente baixo devido à implementação vetorial eficiente no Capítulo 6, diminuiu ainda mais.

Apesar dos bons resultados, há situações em que o MOQA tende a falhar, como no caso de múltiplas codificações e para sinais contendo voz. No primeiro caso, os codecs vão, sucessivamente, aplicando regras de mascaramento sobre sinais já modificados por codificações anteriores. Essas codificações sucessivas modificam a estrutura tempo-espectral dos sinais de tal modo que o modelo psico-acústico adotado não mais modela corretamente a real percepção que os ouvintes teriam em relação ao sinal em questão. A tendência do método de falhar para sinais de voz, em particular masculina, pode ser explicada através das diferenças entre os métodos objetivos de avaliação de voz e de áudio; no caso dos primeiros, a inclusão de modelos de mascaramento mais sofisticados nunca resultou em melhores resultados [89]. Este comportamento, apesar de não totalmente explicado, foi repetidamente observado, indicando que modelos psico-acústicos sofisticados não são apropriados na avaliação de sinais de voz.

Nos dois casos citados, a rede neural consegue compensar a falha do modelo psico-acústico apenas de maneira parcial, causando muitas vezes um erro no valor subjetivo estimado. Uma possível solução para este problema seria criar um identificador que detectasse possíveis sinais problemáticos antes de serem submetidos ao método, e então tratá-los de maneira diferenciada. Contudo, esta alternativa não é a mais indicada, pois não se estaria explorando as verdadeiras causas que levam o método a falhar. A solução definitiva para tais problemas só será alcançada com o desenvolvimento de modelos psico-acústicos melhores e mais sofisticados, especialmente em relação ao processamento temporal, o qual ainda deixa a desejar.

É importante ressaltar ainda que a precisão dos testes subjetivos de avaliação de sinais de áudio é limitada pela dificuldade de se ter uma impressão homogênea por parte de todos os ouvintes presentes nos testes subjetivos. Nem mesmo critérios estatísticos para a eliminação de avaliações que se afastem muito da média são capazes de tornar os resultados perfeitamente confiáveis. Assim, há um limite técnico para a máxima precisão possível de ser alcançada por um método objetivo na estimação de uma impressão subjetiva, pois é praticamente impossível modelar matematicamente características como cultura, humor, experiência e outros fatores que certamente exercem influência significativa em um teste subjetivo.

Pode-se concluir que, especialmente se levadas em consideração todas as dificuldades técnicas e circunstanciais envolvidas no desenvolvimento de uma ferramenta deste tipo, o método MOQA atingiu um nível de desempenho muito bom, capaz de situá-lo à frente dos mais bem sucedidos métodos até aqui desenvolvidos. A futura evolução do método envolverá o desenvolvimento de um novo modelo psico-acústico capaz de modelar uma maior gama de fenômenos e peculiaridades auditivas. O sucesso de tal empreendimento dependerá de estudos adicionais a respeito da audição humana e da realização de testes subjetivos com bases de áudio especialmente desenvolvidas para este fim.

CAPÍTULO 8

NOVAS ABORDAGENS PARA AS MEDIDAS OBJETIVAS DE QUALIDADE DE VOZ

Este Capítulo apresenta a criação e o desenvolvimento de novas técnicas capazes de aperfeiçoar e ampliar a aplicabilidade do método "Medida Objetiva de Qualidade de Voz" (MOQV) [89]. Novas estratégias são sugeridas para o cálculo do atraso entre os sinais e para o mapeamento entre as medidas objetivas e subjetivas. Além disso, um novo modelo psico-acústico, baseado nos avanços obtidos nos últimos anos, foi implementado.

A seguir, serão apresentadas as novas abordagens adotadas e seu impacto no desempenho dos métodos desenvolvidos. Os procedimentos descritos ao longo deste capítulo consideram os sinais amostrados a 16 kHz e quantizados com 16 bits por amostra.

8.1. CÁLCULO DO ATRASO: A QUESTÃO DO ATRASO VARIÁVEL

Uma importante limitação encontrada na maioria dos métodos de avaliação objetiva de sinais de voz, incluindo o MOQV, é a incapacidade de lidar com a questão do atraso variável. Este tipo de fenômeno é encontrado cada vez com mais freqüência, especialmente em transmissões baseadas em pacotes, como ocorre na Internet (IP) e em redes ATM, bem como em comunicações móveis. Tal situação motivou o desenvolvimento de métodos capazes de lidar com atrasos variáveis.

Em primeiro lugar, é necessário considerar duas classes de atrasos variáveis:

- atrasos variáveis que causam degradação na percepção subjetiva dos sinais;
- atrasos variáveis que não causam degradação subjetiva, mas afetam o desempenho dos métodos objetivos.

Os atrasos da primeira classe não são eliminados, a fim de evitar erros na estimação da qualidade subjetiva. Por outro lado, aqueles pertencentes à segunda classe devem ser eliminados, uma vez que o desalinhamento temporal entre os sinais causa um aumento artificial na diferença entre os sinais. Como os métodos objetivos usam esta diferença para gerar uma estimativa da qualidade subjetiva, tal desalinhamento poderia causar estimativas de baixa qualidade.

O primeiro método capaz de lidar com o problema foi o PAMS [111], cujo modelo para identificação do atraso foi empregado no método PESQ [112], medida atualmente adotada como padrão pela ITU-T. As estratégias utilizadas nesses métodos inspiraram o desenvolvimento de uma nova técnica. A descrição do algoritmo desenvolvido é apresentada a seguir.

8.1.1. A Rotina

Como já comentado, a técnica para cálculo do atraso aqui desenvolvida é inspirada naquela usada nos métodos PAMS e PESQ, porém a estrutura final de ambas difere em diversos pontos. Essa diferença é em parte devida à pouca informação disponível na literatura disponível, e em parte devida à introdução de estratégias consideradas úteis para o bom funcionamento do algoritmo.

O primeiro passo do algoritmo é estimar e eliminar um atraso médio entre os sinais original e processado. Para isto, calcula-se a correlação cruzada entre as envoltórias dos dois sinais em função do deslocamento relativo entre ambas. A estimativa do atraso é o deslocamento correspondente ao ponto de máximo da correlação. Em seguida, os sinais são alinhados segundo esta estimativa de atraso.

É importante observar que nem todos os atrasos devem ser eliminados. Atrasos muito severos podem comprometer a qualidade da transmissão, fato este que deve ser levado em conta pelo algoritmo. A seleção dos atrasos que devem ou não ser eliminados é feita de maneira automática pelo algoritmo. Por projeto, a estratégia adotada não é capaz de alinhar adequadamente sinais que estejam excessivamente desajustados. Assim, os sinais continuarão desalinhados, modelando de maneira natural o fato de que atrasos severos podem produzir degradação subjetiva.

O segundo passo deste algoritmo tem como objetivo definir os trechos dos sinais onde o atraso é constante. A seguir, esse atraso é estimado e eliminado. Isto é feito através de várias etapas de processamento, as quais serão descritas sucintamente em uma primeira abordagem, e mais detalhadamente nas seções subseqüentes.

Iniciando a descrição simplificada, o algoritmo toma os sinais resultantes da correção de atraso médio e os divide em trechos denominados de locuções (de acordo com critérios que serão detalhados mais adiante). Em seguida, cada locução é submetida a um processo de eliminação de atraso, porém agora utilizando um procedimento distinto, com maior precisão de estimação de atrasos pequenos. Este procedimento emprega um histograma, o qual fornece uma estimativa do atraso e uma medida da confiança desta estimativa.

Esta estimativa é utilizada para o alinhamento fino de uma locução. Após este ajuste, realiza-se um teste. Se a medida de confiança for maior que 98%, o algoritmo passa a tratar da locução seguinte, indicando que a locução anterior apresenta atraso constante e que o mesmo já foi corrigido. Caso contrário, a locução é divida em duas partes segundo um critério que será detalhado mais adiante. Cada uma das duas partes é tratada da mesma maneira que a locução, até que um critério de parada seja satisfeito, como será descrito mais adiante. Após este procedimento, todos os atrasos terão sido compensados, e o algoritmo passa a tratar da próxima locução, até que os sinais completos estejam alinhados.

A rotina proposta será agora descrita de maneira mais detalhada. A Figura 8.1 apresenta o esquema básico de funcionamento do método aqui proposto.

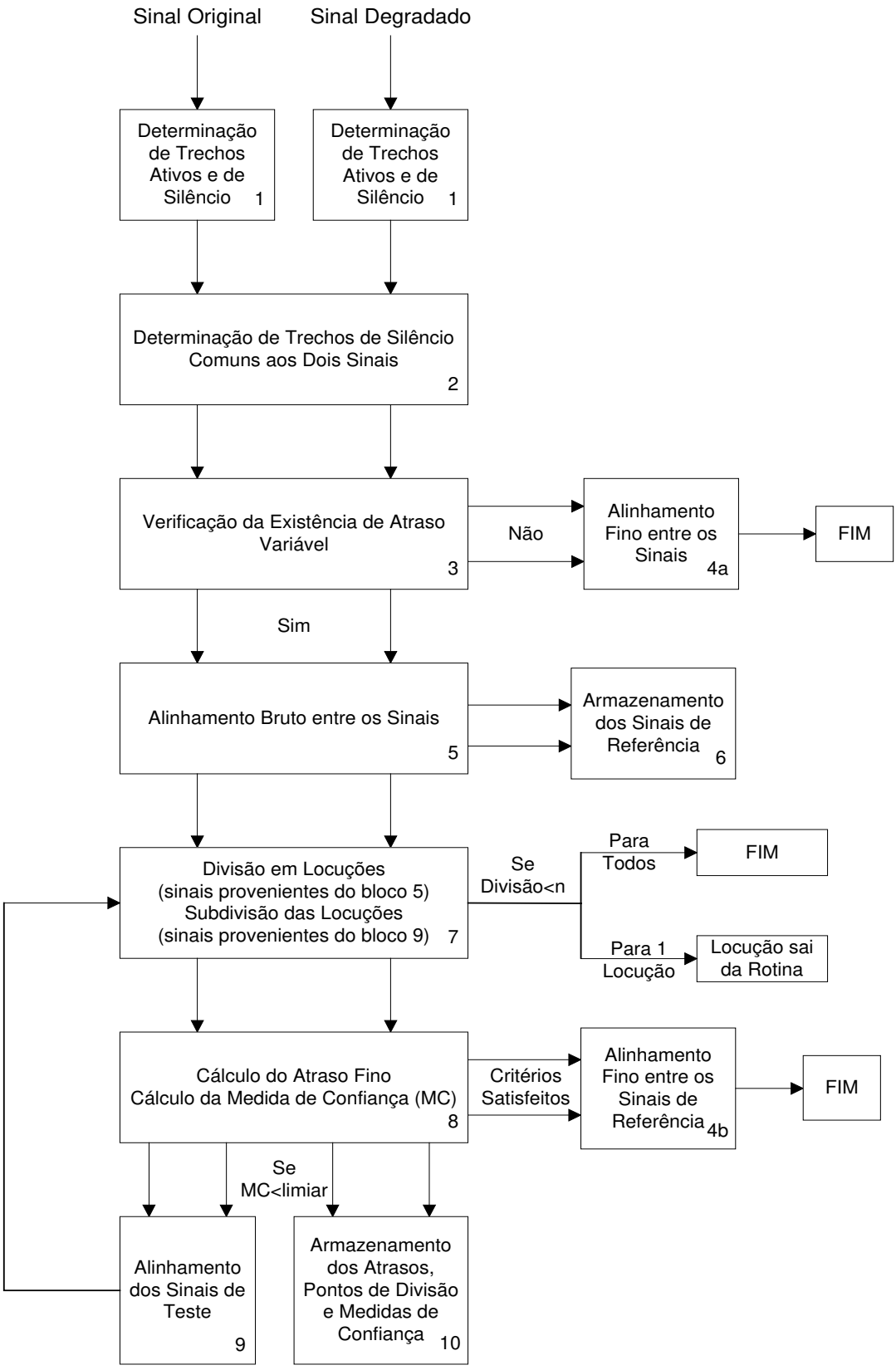


Figura 8.1 - Esquema geral da rotina para compensação de atraso variável.

8.1.1.1. Determinação dos Trechos Ativos e de Silêncio

O primeiro passo desta etapa é a determinação do início e final efetivos dos sinais de voz com base na energia dos componentes, conforme descrito em [89]. Após, o silêncio presente no início e no final de cada sinal de voz é eliminado.

A rotina aqui proposta faz uso intensivo dos trechos dos sinais compostos por voz ativa, bem como dos trechos compostos por silêncio. Assim, é importante que tais trechos sejam corretamente identificados. A determinação dos trechos ativos e de silêncio dos sinais resultantes é feita através de um classificador de trechos de voz baseado em redes neurais [113]. Essa rotina divide o sinal em quadros de 10 ms (160 amostras para uma freqüência de amostragem de 16 kHz), e, para cada quadro, extrai três diferentes parâmetros (energia, número de cruzamentos por zero e autocorrelação). Esses parâmetros são combinados através de uma rede neural, a qual determina se o trecho é silêncio ou voz ativa (na versão original do programa, os trechos de voz ativa são divididos em sonoros e surdos; tal informação, no contexto deste trabalho, não é importante, sendo por isso descartada). A precisão dessa técnica é bastante elevada (cerca de 99% de acerto), podendo ser utilizada sem restrições.

Após a classificação de cada quadro, algumas regras devem ser aplicadas para a determinação dos limites dos trechos de voz e silêncio:

- o primeiro trecho é considerado voz ativa, uma vez que o trecho de silêncio do início do sinal foi eliminado;

- considera-se o início de um trecho de silêncio a primeira amostra de um quadro classificado como silêncio cujo quadro anterior foi classificado como voz ativa e cujos 9 quadros subseqüentes tenham sido classificados como silêncio;

- considera-se o início de um trecho de voz ativa a primeira amostra de um quadro classificado como de voz e cujos 9 quadros anteriores tenham sido classificados como silêncio.

- o último trecho é considerado voz ativa, uma vez que o trecho de silêncio no final do sinal foi eliminado.

As regras acima visam evitar que trechos inativos muito pequenos (menores que 100 ms) sejam de fato classificados como silêncio. Se tal cuidado não é tomado, há a possibilidade que uma grande quantidade de divisões seja gerada, prejudicando o funcionamento da rotina. Além disso, períodos de silêncio menores que 100 ms são naturalmente encontrados durante o pronunciamento de uma sentença, podendo, portanto, serem considerados parte integrante do trecho ativo em si.

As classificações aqui geradas são utilizadas nas etapas representadas pelos quadros 2 e 7 (divisão em locuções), como será descrito a seguir.

8.1.1.2. Determinação dos Trechos de Silêncio Comuns aos Dois Sinais

Esta etapa consiste em se determinar os componentes pertencentes a trechos classificados como silêncio que sejam comuns aos dois sinais. Como se pode observar na Figura 8.2, os sinais são comparados e componentes de mesmo índice pertencentes a intervalos de silêncio em ambos os sinais são identificados e temporariamente eliminados. Tal procedimento visa evitar erros na identificação da existência ou não de atrasos variáveis, como será discutido na próxima subseção. Após esse procedimento, tais trechos são reintroduzidos no exato ponto de onde foram extraídos.

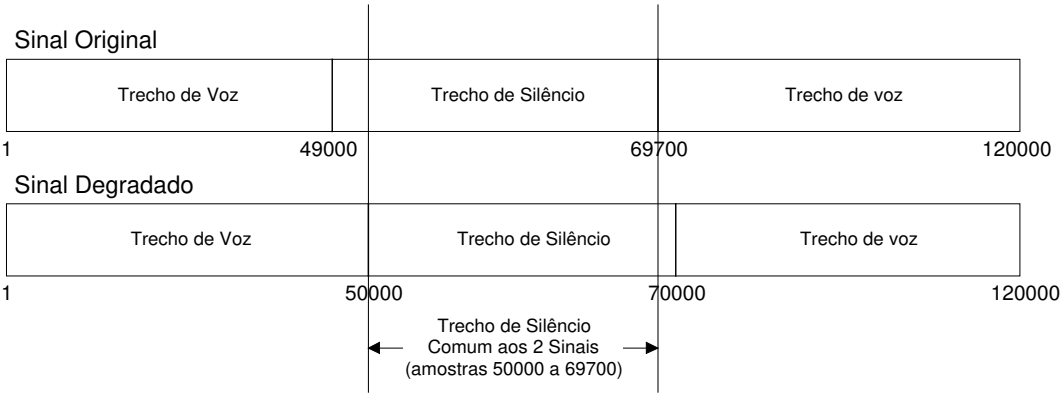


Figura 8.2 - Determinação dos trechos de silêncio comuns aos dois sinais.

8.1.1.3. Verificação da Existência de Atraso Variável

Apesar do crescimento verificado nos últimos anos no número de aplicações capazes de introduzir atrasos variáveis, a maioria dos sistemas e codecs utilizados nos dias de hoje geram apenas atrasos fixos ou de pequena variação. Esta etapa foi introduzida com o objetivo de evitar que sinais com atraso fixo sejam submetidos às etapas subseqüentes, o que representaria um gasto de tempo desnecessário. Por outro lado, como o processamento aqui realizado é muito rápido, o tempo adicional gasto nesta etapa para o caso de testes com sinais que possuam atraso variável (e que, portanto, serão submetidos ao restante da rotina) é desprezível em relação ao tempo total de processamento do método, justificando sua inclusão na rotina.

A estratégia para decisão sobre a presença de atraso fixo ou variável consiste em se calcular o atraso de quadros de 8.000 amostras selecionados em três pontos específicos de cada um dos sinais. O primeiro quadro inicia-se a 1.000 amostras do início do sinal, o segundo é localizado no meio do sinal com os trechos de silêncio eliminados e o último termina 2.000 amostras antes do final do sinal. A identificação de atrasos variáveis baseia-se numa sub-rotina largamente utilizada ao longo da rotina principal, cuja função é calcular o atraso fino entre os sinais ou trechos destes. Esta sub-rotina é descrita em detalhes na Seção 8.1.1.8. Se a diferença entre os atrasos obtidos para os três intervalos não superar 1 ms (16 amostras para uma freqüência de amostragem de 16 kHz ou 8 amostras para 8 kHz) e as medidas de confiança obtidas para os três trechos forem superiores a 0,5 (50 %), então considera-se que o sinal possui um atraso constante e de valor igual à média entre os atrasos obtidos para os três intervalos. É importante destacar que a tolerância de 1 ms na diferença entre os três quadros foi baseada no fato de que, apesar de uma pessoa com audição normal raramente identificar atrasos menores que 4 ms, os métodos de avaliação objetiva, de uma maneira geral, possuem uma sensibilidade maior, e, portanto, um critério mais restritivo se faz necessário.

A eliminação dos trechos de silêncio comuns aos dois sinais visa evitar que algum dos três intervalos escolhidos para cálculo do atraso possua um longo período de silêncio, o que poderia contaminar a estimativa do atraso para aquele quadro. Isto ocorre porque períodos de silêncio são compostos basicamente por ruído, e neste caso qualquer esforço para estimar um atraso falhará, uma vez que tal ruído provavelmente será diferente nos dois sinais, e então o cálculo da correlação cruzada deixa de fazer sentido.

Um problema que se poderia enfrentar com a utilização desta estratégia seria a possibilidade dos trechos selecionados apresentarem um atraso constante, porém com uma

variação importante nos trechos compreendidos entre os mesmos, denotando que, na verdade, o sinal possui atraso variável. Contudo, os critérios e tamanhos de quadros escolhidos tornam esta possibilidade remota. Além disso, tal ocorrência iria de encontro com os padrões de atraso variável observados em condições reais. Por fim, ainda que tal situação ocorra, ela deverá estar limitada a um trecho pequeno (o que é garantido pelo critério de escolha dos intervalos), produzindo um impacto pequeno na estimação da qualidade subjetiva final.

8.1.1.4. Alinhamento Fino entre os Sinais

Esta Subseção descreve as estratégias de alinhamento entre os sinais original e degradado ou entre trechos dos mesmos, estratégias estas usadas ao longo de toda a rotina em questão.

Quando os cálculos dos atrasos e medidas de confiança preenchem certos requisitos (descritos em detalhes nas Seções 8.1.1.3 e 8.1.1.8), os sinais são alinhados da seguinte maneira:

- no bloco 4a da Figura 8.1, os sinais original e degradado, com os trechos de silêncio reintroduzidos, são alinhados utilizando-se os valores de atraso obtidos na etapa descrita na Subseção 8.1.1.3.

- no bloco 4b da Figura 8.1, os sinais de referência armazenados (bloco 6) são alinhados utilizando-se os valores de atraso obtidos para cada trecho, após todos os critérios terem sido satisfeitos (ver Subseções 8.1.1.7 e 8.1.1.8).

O procedimento para o alinhamento é mostrado na Figura 8.3.

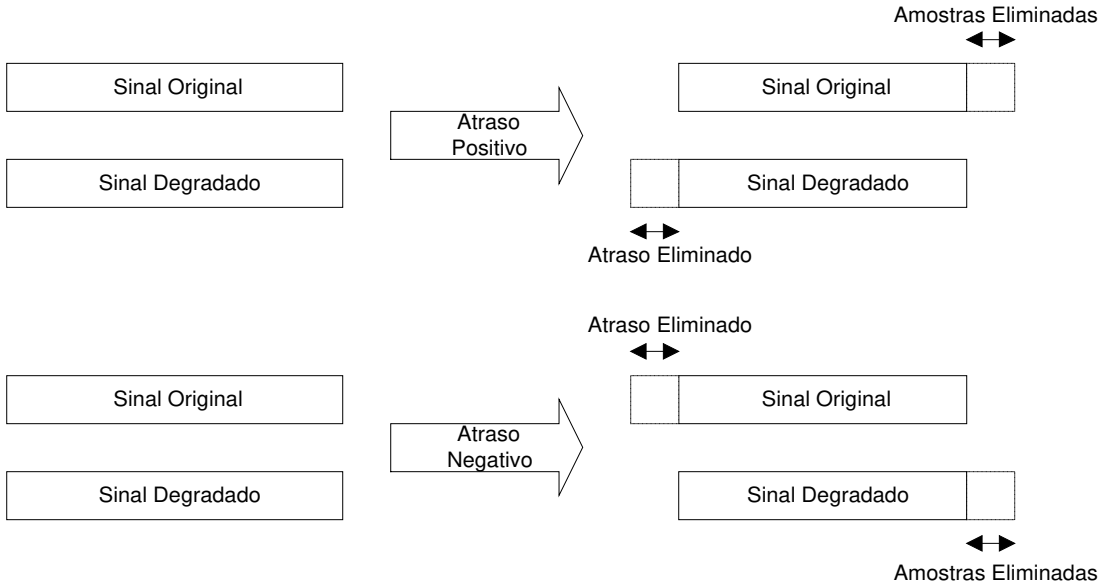


Figura 8.3 - Esquema de realinhamento entre os sinais.

Como se pode observar na figura, se o atraso é positivo e de valor igual a *n*, ou seja, se o sinal degradado está atrasado em relação ao original de *n* amostras, elimina-se os *n* primeiros componentes do sinal degradado, de maneira que os sinais fiquem alinhados; para que os sinais tenham o mesmo comprimento, as *n* amostras finais do sinal original são também descartadas. Se o atraso é negativo, o procedimento é o mesmo, porém invertendo-

se os papéis de cada sinal. Esta técnica pode ser aplicada aos sinais como um todo, se não forem identificados atrasos variáveis, ou a intervalos separados.

8.1.1.5. Alinhamento Bruto entre os Sinais

Caso tenha sido detectada a ausência de atraso variável, segundo os procedimentos da Subseção 8.1.1.3, então executa-se o alinhamento fino descrito na Subseção 8.1.1.4 e a rotina é encerrada. Se determinada a presença de atrasos variáveis, faz-se um primeiro alinhamento entre os sinais, utilizando-se uma sub-rotina denominada *atrgr*, cujo esquema básico é mostrado na Figura 8.4.

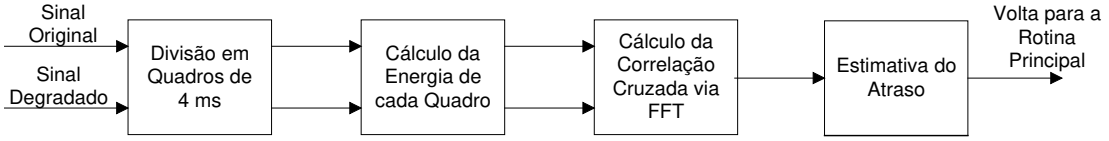


Figura 8.4 - Esquema para cálculo do atraso bruto.

Como se pode observar, os sinais são divididos em quadros de 4 ms, sem superposição. A seguir, calcula-se a energia de cada um deles, a fim de se determinar a envoltória do sinal. A correlação cruzada entre as envoltórias dos sinais é computada no domínio da freqüência através de uma FFT, a fim de reduzir a complexidade computacional. A estimativa do atraso bruto, determinada pela posição do maior valor absoluto da correlação, é então usada para alinhar os sinais. A precisão desta estimativa depende fortemente da quantidade de informação contida na envoltória do sinal. Devido à característica altamente não-estacionária dos sinais de voz, normalmente a envoltória carrega informação suficiente para fornecer resultados com uma resolução de aproximadamente 8 ms (ou 128 amostras para sinais amostrados a 16 kHz), a qual é suficiente para o propósito deste primeiro alinhamento.

Essa estratégia visa fornecer um valor para o atraso médio entre os sinais. O alinhamento baseado neste valor evita que o atraso entre os sinais como um todo seja excessivo. Considerando-se que todas as estimativas de atraso fazem uso, de algum modo, do cálculo da correlação cruzada, e como este cálculo é tanto mais preciso quanto maior for a quantidade de amostras comuns a ambos os sinais (ou trechos), uma diminuição no desalinhamento entre os sinais permitirá que se tenha uma maior precisão nas estimativas de atraso posteriores. Este fato se torna ainda mais importante ao observar-se que as estimativas de atraso fino dividem os sinais em blocos de tamanho limitado, ou seja, há poucas amostras para se trabalhar; se o desalinhamento for excessivo, não será possível extrair informação suficiente para que se tenha uma estimativa confiável para o atraso naquele trecho.

8.1.1.6. Armazenamento dos Sinais de Referência

Após o alinhamento bruto, os sinais resultantes, agora denominados de referência, são armazenados; serão eles que serão alinhados quando todos os intervalos e atrasos correspondentes tiverem sido determinados. Este armazenamento é necessário porque, nas próximas etapas de processamento (Seções 8.1.1.7 a 8.1.1.10), os sinais serão sucessivamente alinhados, a fim de se obter estimativas mais precisas para os atrasos em cada trecho (como foi comentado na subseção anterior, quanto menor o desalinhamento

entre os trechos, maior a precisão dos cálculos). Esses sinais submetidos às diversas etapas de alinhamento, ora compensando atrasos positivos, ora compensando atrasos negativos, perdem uma grande quantidade de amostras no processo, de maneira que usá-los no restante da rotina para estimação da qualidade subjetiva não seria adequado. Assim, o que se faz é armazenar todos os ajustes acumulados para cada trecho e, ao final, após todos os critérios terem sido satisfeitos, os ajustes são aplicados a estes sinais de referência. A Figura 8.5 mostra um exemplo simples das situações encontras na prática.

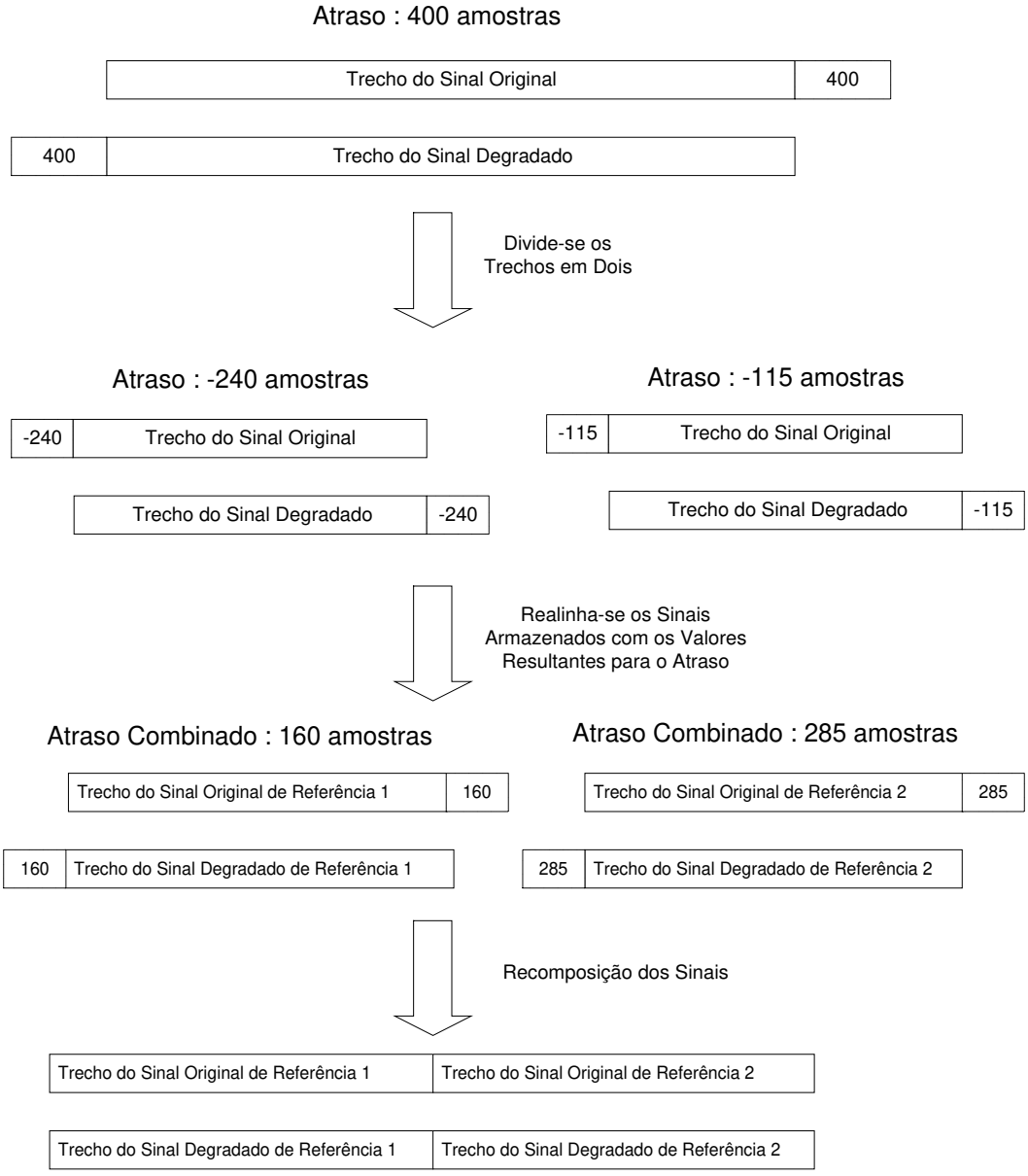


Figura 8.5 - Exemplo do uso dos sinais de referência.

Este exemplo mostra um trecho para o qual determinou-se um atraso de 400 amostras. Supõe-se que os critérios de parada não foram satisfeitos, e então o trecho foi novamente dividido (ver Subseções 8.1.1.7 e 8.1.1.8). Para o primeiro deles, determinou-se um atraso de -240 amostras e, para o segundo, de -115 amostras. Supondo que os critérios de parada

tenham sido satisfeitos, computa-se os atrasos resultantes a serem utilizados para alinhamento dos sinais de referência, cujos valores, neste exemplo, são de 160 e 285 amostras, representando uma perda total de 445 amostras. Por outro lado, os sinais submetidos aos alinhamentos sucessivos perderam um total de 755 amostras, 270 a mais que os sinais de referência. Em muitos casos, os trechos podem ser alinhados 5 ou mais vezes, o que certamente representaria uma perda ainda maior. Assim, o uso destes sinais de referência se mostra essencial para o funcionamento adequado da rotina.

8.1.1.7. Divisão dos Trechos dos Sinais

Esta etapa consiste na divisão dos sinais (ou trechos destes). Após o alinhamento bruto dos sinais (bloco 5 na Figura 8.1), estes são subdivididos em locuções segundo os seguintes critérios:

- a primeira locução inicia-se no começo do sinal, e termina na metade do primeiro intervalo de silêncio;
- as locuções intermediárias iniciam-se na metade de um período de silêncio e terminam na metade do trecho de silêncio seguinte;
- a última locução inicia-se na metade do último período de silêncio e estende-se até o final do sinal.

Se o sinal não possuir trechos de silêncio, haverá uma única locução compreendendo o sinal inteiro.

Cada locução é submetida a um primeiro teste para verificar seu comprimento. Se este é maior que 3.072 amostras, a locução será analisada através dos procedimentos a seguir, a fim de se determinar se há a presença de atrasos. Caso contrário, considera-se que a locução é pequena demais e que o alinhamento bruto corrigiu qualquer eventual atraso; a seguir, o algoritmo começa a analisar a próxima locução. Este teste é o primeiro critério para terminar o processamento de uma locução, e $n = 3.072$ é o comprimento mínimo que será analisado com vistas à determinação do atraso de uma seção individual. Testes práticos revelaram que tal valor é suficiente mesmo em casos com atrasos fortemente variáveis.

Existem outros critérios de parada, como será descrito a seguir. Se nenhum deles é satisfeito, a locução que está sendo testada é dividida ao meio (sinais provenientes do bloco 9 na Figura 8.1). Cada metade é então tratada como uma locução. Este procedimento é repetido até que todas as subdivisões satisfaçam pelo menos um dos critérios de parada. Então, o algoritmo começa a processar a próxima locução.

8.1.1.8. Cálculo do Atraso Fino e da Medida de Confiança

Esta seção descreve a estratégia aplicada às locuções com comprimento maior que 3.072 amostras. Este procedimento visa estimar os atrasos presentes nas locuções, e emprega uma técnica precisa para estimação do atraso, a qual fornece uma medida fina para o atraso uma medida de confiança. A estrutura básica desta técnica é mostrada na Figura 8.6.

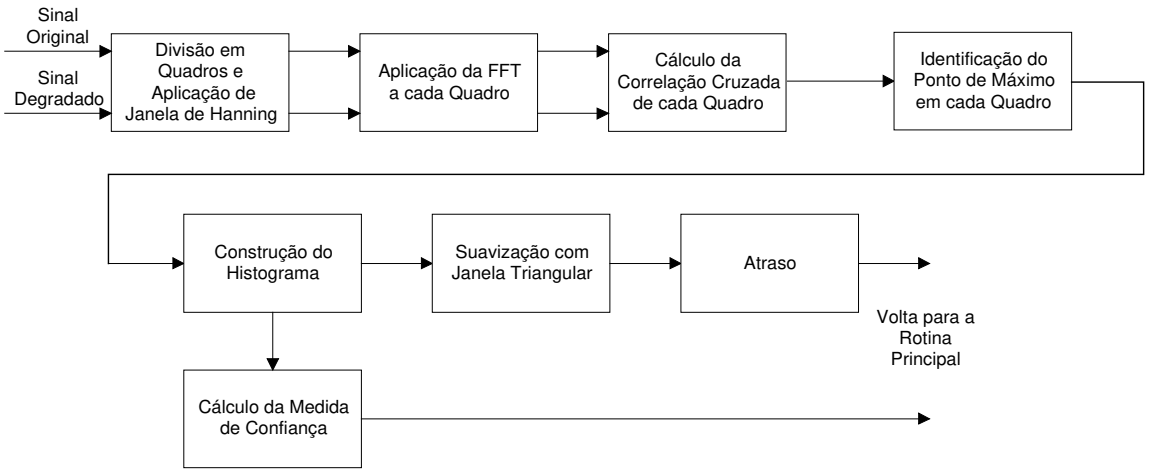


Figura 8.6 - Esquema geral da sub-rotina para cálculo do atraso fino.

Esta sub-rotina foi implementada tendo como base a descrição sucinta encontrada em [114], porém com a inclusão de algumas modificações. Inicialmente os sinais ou trechos que se está considerando são divididos em quadros, utilizando-se uma janela de Hanning. No método PESQ, o tamanho destes quadros é fixo e de valor igual a 64 ms (2.048 amostras), com uma superposição de 75%. Na rotina aqui implementada, utilizou-se um tamanho de quadro dependente do comprimento do trecho para o qual se deseja estimar o atraso: a dimensão do quadro é dada por 20 vezes a raiz quadrada do número de amostras do trecho em questão. Este critério foi adotado em virtude da observação de que, quando os trechos em questão ainda são grandes, seu desalinhamento freqüentemente também o é, uma vez que ainda não se teve a oportunidade de se fazerem ajustes capazes de diminuir o descasamento entre os sinais. Assim, faz-se necessária a utilização de um maior número de amostras no cálculo da correlação cruzada entre os trechos correspondentes a cada sinal. À medida que os trechos são subdivididos, um maior número de alinhamentos já terá sido realizado, e então um menor número de amostras pode ser utilizado na estimação do atraso. A adoção desta solução foi a maior responsável pelos excelentes resultados obtidos nos testes realizados com esta rotina. Outra mudança eficiente foi a alteração da superposição entre os quadros de 75 para 87,5%, permitindo que se tivesse um maior número de pontos na construção do histograma, a qual será descrita mais adiante.

Após a divisão em quadros, calcula-se a correlação cruzada para cada quadro no domínio da freqüência, utilizando-se o método citado anteriormente e descrito em [89]. A seguir, realiza-se a construção de um histograma, segundo o seguinte critério:

- identifica-se, para cada quadro, o valor e o índice da máxima correlação obtida;
- os valores máximos da correlação cruzada de cada quadro, elevados à potência de 0,125 para comprimi-los a valores próximos a 1, são tomados como pesos para cada quadro;
- os pesos são agrupados e somados de acordo com o índice correspondente, gerando as barras do histograma.

O histograma resultante é então normalizado pelo soma de todos os pesos, de maneira que sua área tenha valor unitário.

Neste estágio do processamento, calcula-se uma medida de confiança para os resultados obtidos. Este valor é determinado pela porcentagem da área do histograma

normalizado que está concentrada em torno do índice de seu ponto de máximo (o qual é determinado após a suavização com uma janela triangular, como será visto a seguir). Na implementação aqui adotada, esta área é determinada por todos os valores que se encontram a até 1 ms, para mais ou para menos, desse índice. Por exemplo, para um sinal amostrado a 16 kHz, se o índice do histograma correspondente ao ponto de máximo indica um atraso de 100 amostras, o intervalo utilizado na composição da medida de confiança estará compreendido entre os índices 84 e 116.

Após o cômputo da medida de confiança, o histograma é suavizado através de uma janela triangular de aproximadamente 1 ms e valor de pico igual a 1, de modo a tornar mais confiável a estimativa do atraso, a qual é dada pelo índice de seu ponto de máximo.

Um segmento é retirado do processo de subdivisão da locução quando um dos seguintes critérios é satisfeito: 1) ele apresenta uma medida de confiança maior que 0,98; 2) sua divisão não resulta numa melhoria da medida de confiança; 3) a variação entre o atraso do segmento completo e os atrasos dos trechos resultantes da divisão é menor que 5 amostras. Quando o segmento é retirado do processo, seu atraso e pontos de divisão são armazenados (ver bloco 10 na Figura 8.1 e Subseção 8.1.1.10).

Por outro lado, se um segmento não satisfaz pelo menos uma das condições acima, ele é alinhado (ver bloco 9 na Figura 8.1 e Subseção 8.1.1.9) e novamente subdividido (bloco 7 e Subseção 8.1.1.7). Se todos os segmentos que compõem o sinal satisfazem pelo menos um dos critérios, os sinais de referência são alinhados de acordo com os atrasos e pontos de divisão armazenados no bloco 10 (bloco 4b na Figura 8.1), e o algoritmo é encerrado.

8.1.1.9. Alinhamento dos Sinais de Teste

Quando ainda existem segmentos que não preenchem pelo menos um dos critérios de parada descritos na Subseção 8.1.1.8, eles são alinhados de acordo com a nova estimativa de atraso. A seguir, eles são novamente submetidos aos procedimentos indicados pelos blocos 7 e 8 na Figura 8.1. Esta estratégia é repetida até que todos os segmentos tenham satisfeito alguma das condições de parada.

8.1.1.10. Armazenamento dos Atrasos, Pontos de Divisão e Medidas de Confiança

Esta etapa armazena os pontos de divisão e os atrasos de todos os segmentos que satisfizeram um critério de parada. Ambos os valores são usados no alinhamento final dos sinais de referência.

As medidas de confiança são armazenadas somente para os segmentos que não preencheram os critérios de parada. Essas medidas são comparadas com os valores obtidos após a divisão do trecho, como descrito na Seção 8.1.1.8 (segunda condição de parada).

8.1.2. Testes com a Rotina

8.1.2.1. Base de Dados Utilizada

Nos testes, utilizou-se uma base de dados fornecida pela Fundação CPqD. Esta base de dados consiste de 12 arquivos de voz, cada um dos quais submetidos a 12 diferentes condições, como mostrado na Tabela 8.1, resultando em 144 pares de arquivos.

Tabela 8.1. Base de Dados Utilizada.

N.	Condição	Número de Arquivos Submetidos à Condição
1	Codec G711	6 c/ vozes fem. e 6 c/ vozes masc.
2	Codec G726	6 c/ vozes fem. e 6 c/ vozes masc.
3	MNRU (Q = 25 dB)	6 c/ vozes fem. e 6 c/ vozes masc.
4	MNRU (Q = 15 dB)	6 c/ vozes fem. e 6 c/ vozes masc.
5	Codec com Nível de Entrada de –26 dBov	6 c/ vozes fem. e 6 c/ vozes masc.
6	Codec com Nível de Entrada de –14 dBov	6 c/ vozes fem. e 6 c/ vozes masc.
7	Codec com Nível de Entrada de –38 dBov	6 c/ vozes fem. e 6 c/ vozes masc.
8	Ruído de rua, -26 dBoV, SNR=20dB, sem codec	6 c/ vozes fem. e 6 c/ vozes masc.
9	Ruído de rua, -26 dBoV, SNR=20dB, com codec	6 c/ vozes fem. e 6 c/ vozes masc.
10	R. escritório, -26 dBoV, SNR=20dB, sem codec	6 c/ vozes fem. e 6 c/ vozes masc.
11	R. escritório, -26 dBoV, SNR=20dB, com codec	6 c/ vozes fem. e 6 c/ vozes masc.
12	Cascata de 3 codecs	6 c/ vozes fem. e 6 c/ vozes masc.

Os valores de avaliação subjetiva de qualidade encontrados nesta base correspondem à média das notas obtidas para os doze arquivos constituintes de cada condição. Assim, tem-se somente doze pontos para o cálculo da correlação e para se traçar as curvas. Apesar de limitado, este conjunto de sinais apresenta uma severa variação nos atrasos ao longo dos arquivos. No início dos arquivos, o sinal degradado apresenta um determinado atraso em relação ao original; contudo, o sinal degradado apresenta uma perda contínua de amostras ao longo do tempo, de maneira que, ao final, é o sinal original que apresenta um forte atraso em relação ao degradado. Em certos casos, observou-se uma variação de 1.600 amostras entre o atraso no início e no final dos arquivos. Esta é uma situação crítica para se testar a rotina, pois não há um único trecho com atraso constante e, portanto, este seria o pior caso com o qual se poderia deparar. Assim, um bom desempenho neste tipo de situação necessariamente implicará num bom desempenho para qualquer outro padrão de atraso variável que se possa encontrar na prática. Os resultados obtidos são mostrados na próxima subseção.

8.1.2.2. Resultados Obtidos

Os 144 arquivos do banco de dados foram submetidos ao novo programa MOQV com a rotina de cálculo de atraso variável incorporada. Então, calculou-se as médias das medidas objetivas, de maneira a se ter um único valor para cada condição. A seguir, aplicou-se um mapeamento monotônico usando um polinômio de terceira ordem e calculou-se a correlação entre os valores objetivos e subjetivos. A Figura 8.7 mostra a curva resultante do mapeamento.

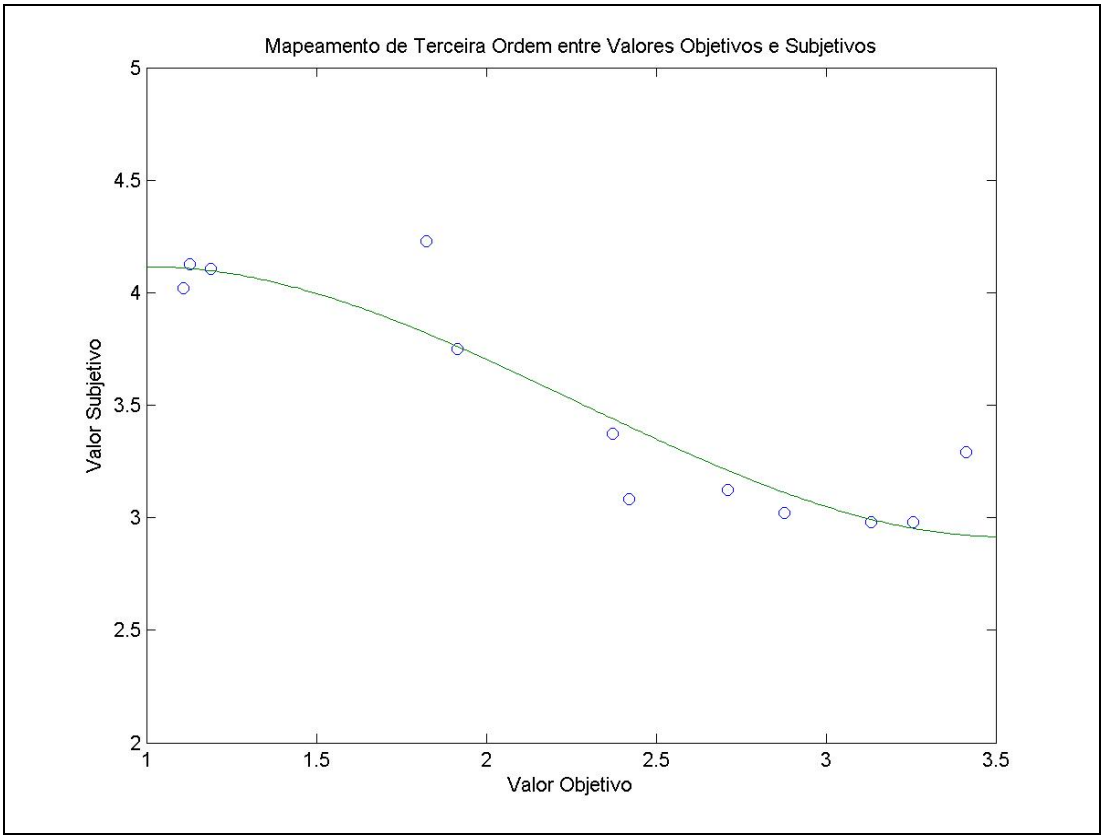


Figura 8.7 - Curva de mapeamento entre os valores objetivos e subjetivos.

A correlação obtida para este caso foi de 0,92, em contraste com o valor em torno de 0,45 obtido sem a aplicação da rotina de atraso variável. É importante destacar que o valor da correlação só não foi maior porque o número de pontos utilizado para o cálculo foi muito pequeno, e então apenas um ponto fora da reta pode ser responsável por um decréscimo importante na correlação. Além disso, o modelo perceptual utilizado no método MOQV não é o melhor disponível ou conhecido atualmente, e então pequenas falhas podem ser observadas.

Além do teste da correlação, foi feita uma inspeção visual detalhada do alinhamento dos arquivos após sua submissão à rotina. Os resultados foram excelentes, com desvios mínimos sendo observados entre os dois arquivos.

Além dos testes realizados com sinais contendo atrasos variáveis, testou-se o novo programa com arquivos de atraso fixo, e os resultados foram idênticos àqueles observados em testes realizados com o programa sem a rotina de cálculo de atraso variável [89].

Como era de se esperar, a nova rotina tornou o programa um pouco mais lento, porém representou uma expressiva melhora na robustez do método MOQV. O uso desta estratégia em conjunto com um modelo psico-acústico aperfeiçoado originou um novo e eficiente método, o qual é descrito em detalhes na Seção 8.3.

8.2. MAPEAMENTO ENTRE AS MEDIDAS OBJETIVAS E SUBJETIVAS

Na busca por métodos de avaliação da qualidade de voz mais eficiente, a maior parte dos esforços tem sido direcionada para alguns poucos fatores, tais como a modelagem do comportamento dos ouvintes em um teste subjetivo e o aperfeiçoamento do modelo do ouvido. Outros fatores foram brevemente investigados, e os resultados têm sido adotados como padrão desde então. Este é o caso do processo de mapeamento, onde o uso de funções monotônicas que minimizam o erro quadrático médio está fortemente estabelecido. Em particular, os mapeamentos polinomiais de terceira ordem têm sido largamente utilizados, principalmente devido à sua capacidade de modelar adequadamente o comportamento dos ouvintes em testes subjetivos [89]. Apesar disso, este tipo de estrutura não é capaz de explorar toda a informação que poderia ser extraída através dos métodos objetivos. Conseqüentemente, ela tende a falhar sob certas condições.

Nesse contexto, duas novas estratégias de mapeamento foram desenvolvidas usando diferentes estruturas de redes neurais, conforme será descrito nas próximas seções.

8.2.1. Abordagem Baseada em Mapas de Kohonen

As informações contidas nesta seção foram compiladas a partir de [115] e [116].

8.2.1.1. Mapeamento Usando Redes de Kohonen

A estratégia de mapeamento aqui adotada pode ser dividida em 3 estágios:

1 - *Quantização dos Dados*. Como já comentado, o problema do mapeamento entre medidas objetivas e subjetivas tem sido tratado por técnicas clássicas usando funções monotônicas. Como foi visto na Seção 7.2.2, o princípio de funcionamento das redes de Kohonen é substancialmente diferente, uma vez que estas não têm a capacidade de aproximar funções. Então, a exemplo do procedimento adotado para sinais de áudio, os valores subjetivos alvo dos sinais foram quantizados. A resolução aqui adotada, de 0,25 MOS e 0,25 CMOS, resultou em 17 classes. A escala objetiva MOS baseia-se na avaliação absoluta da qualidade do material processado, sem que o avaliador disponha de material para comparação. No caso do CMOS, a avaliação é realizada através da comparação entre os sinais original e degradado. Maiores informações a respeito dessas escalas podem ser encontradas em [74] e [89].

2 - *Extração dos Parâmetros de Entrada*. É desejável que o conjunto de parâmetros extraídos dos sinais contenha o máximo de informação a respeito do sinal que se deseja avaliar. Para isso, além da versão original do método MOQV, implementou-se em paralelo uma versão modificada. Nessa versão modificada, a FFT usada na decomposição tempo-freqüência é substituída por uma MLT (*Modulated Lapped Transform*). A MLT é uma ferramenta para a decomposição espectral localizada de sinais [117]. Suas funções básicas são mostradas nas equações a seguir:

$$p_a(n,k) = h_a(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right], \tag{8.1}$$

$$p_s(n,k) = h_s(n)\sqrt{\frac{2}{M}}\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right], \tag{8.2}$$

$$h_a(n) = h_s(n) = -sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{M}\right], \tag{8.3}$$

onde $p_a(n,k)$ e $p_s(n,k)$ são as funções-base para as transformadas de análise e síntese, $h_a(n)$ e $h_s(n)$ são as janelas de análise e síntese e M é o tamanho do bloco. O índice de tempo n varia de 0 a 2M-1 e o índice de freqüência varia de 0 a M-1.

Para cada versão do algoritmo MOQV, cinco parâmetros distintos foram extraídos:

- a diferença entre as energias de cada quadro do sinal, obtida após o mapeamento das freqüências em sub-bandas [89];
- a distância espectral perceptual, dada por

$$PSD = \sqrt{\sum_{b=1}^{B}\left[L_x(b) - L_y(b)\right]^2}\ ; \tag{8.4}$$

onde L_x e L_y representam as funções densidade espectral perceptual dos sinais original e degradado, respectivamente, e b representa a divisão em bandas auditivas [89];

- a distância cepstral perceptual [93], que é uma versão modificada da distância espectral perceptual, dada por

$$PCD = 10\cdot\sqrt{\sum_{b=1}^{B}\left\{\log_{10}\left[L_x(b)\right] - \log_{10}\left[L_y(b)\right]\right\}^2}\ ; \tag{8.5}$$

- os valores MOQV1 e MOQV2 [89], os quais correspondem, aproximadamente, aos valores PSQM e PSQM+ [88].

3 - *Treinamento*. A base de dados utilizada nos testes foi a S-23 [118], a qual é composta por arquivos de voz em inglês, francês, japonês e italiano. Esses arquivos estão associados a um conjunto de codecs e condições de teste. Cada teste tem associado um valor MOS ou CMOS. A estimativa desses valores subjetivos é a meta a ser alcançada a partir dos parâmetros extraídos. Tal material é dividido em três grupos principais:

- primeiro experimento: os arquivos de voz foram submetidos a um conjunto de codecs usados na telefonia fixa e móvel;
- segundo experimento: os arquivos de voz foram submetidos a certos tipos de ruído de ambiente;
- terceiro experimento: os arquivos simulam os efeitos da transmissão de sinais codificados através de canais de comunicações que introduzem erros.

O treinamento foi realizado levando-se em conta todas as linguagens e experimentos encontrados na base de dados mencionada. Os parâmetros são apresentados à rede um a um, e somente uma vez. Em cada apresentação, os pesos relativos ao neurônio vencedor e seus vizinhos são atualizados usando o critério dado pela Equação 7.1. A Figura 8.8 mostra a composição dos dados usados no treinamento.

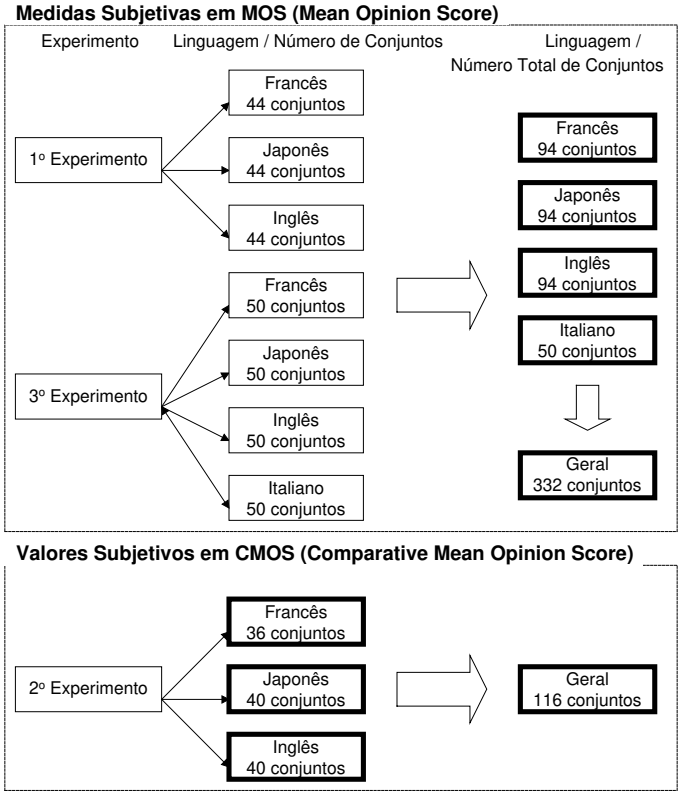


Figura 8.8 - Arranjo dos arquivos de treinamento.

Os quadros em destaque na figura correspondem aos arquivos que foram de fato utilizados no treinamento. Cada conjunto de dados é composto dos parâmetros de entrada e das medidas subjetivas alvo. Para cada um desses conjuntos, e para cada uma das configurações selecionadas, foi testado um total de 10.000 iniciações de pesos distintas, a fim de se determinar os pesos iniciais capazes de produzir a maior correlação entre os valores subjetivos estimados e de referência.

A configuração adotada para os fatores de treinamento (taxa de aprendizagem, ordem da vizinhança e iniciação dos pesos) é a mesma utilizada no caso da avaliação de áudio.

8.2.1.2. Resultados

Nos testes, diversos fatores foram investigados a fim de se encontrar a combinação que resultasse no melhor desempenho, como mostrado a seguir.

a) Arranjo dos neurônios: foram testados arranjos unidimensionais [115] e bidimensionais [116]. Não foram observadas diferenças substanciais no desempenho das duas estruturas testadas; como a estrutura unidimensional é mais simples de ser implementada, esta se tornou a escolha natural para este tipo de aplicação.

b) Número de neurônios: o número de neurônios tem um papel importante no desempenho da estrutura implementada. Foi observado que é necessário um mínimo de 5 neurônios por classe para se obter boas correlações entre os valores subjetivos estimados e de referência. Os melhores resultados foram alcançados utilizando-se de 8 a 15 neurônios por classe. Um número superior de neurônios torna a estrutura excessivamente complexa, sem um ganho correspondente no desempenho.

c) Parâmetros: a rede foi inicialmente testada usando-se os 10 parâmetros de entrada

e, posteriormente, usando-se apenas os valores MOQV1 e 2, levando a um total de 4 parâmetros. Esta última configuração é suficiente para a maior parte das situações encontradas na prática; contudo, no caso dos sinais analisados serem muito complexos, o uso de parâmetros adicionais pode ser útil. Infelizmente, o único modo de se determinar o procedimento correto é testando todas as possibilidades para o problema em questão. Nos testes aqui realizados, o uso de todos os parâmetros se mostrou particularmente útil para alguns sinais com o MOS como valor subjetivo.

d) Quantização: o refinamento da quantização, através do aumento do número de classes, foi também testado, mas os resultados obtidos foram inferiores. Foi observado que a rede tende a perder o foco quando as classes estão muito próximas, de maneira que os erros de classificação crescem rapidamente à medida que se aumenta o número de classes. Por outro lado, classes mais largas acarretam um excessivo descasamento com os valores subjetivos de referência, diminuindo as correlações. Deste modo, o número de 17 classes representa o melhor compromisso entre quantização e classificação.

A Tabela 8.2 fornece a distribuição dos sinais de voz utilizados nos testes.

Tabela 8.2. Sinais de voz usados nos testes.

Língua	**MOS**	**CMOS**
Francês	376	128
Japonês	376	136
Inglês	376	136
Italiano	200	-
Total	1328	400

A Tabela 8.3. apresenta uma comparação entre os resultados obtidos para o algoritmo MOQV original, usando um mapeamento polinomial, e os melhores resultados obtidos para a estratégia proposta usando as estruturas de Kohonen de 1 e 2 dimensões.

Tabela 8.3. Correlações obtidas para cada abordagem.

Língua	**Medida Subjetiva**	**MOQV**		**Kohonen**		
		1	**2**	**1 Dim.**	**2 Dim.**	**Estrutura**
Francês	MOS	0,90	0,88	***0,95***	0,93	5-15-4
	CMOS	0,94	0,94	***0,99***	0,99	2-8-4
Japonês	MOS	0,72	0,77	***0,94***	0,92	10-15-10
	CMOS	0,96	0,96	***0,98***	0,98	5-8-4
Inglês	MOS	0,78	0,80	***0,98***	0,92	10-15-10
	CMOS	***0,96***	0,95	0,93	0,93	1-5-4
Italiano	MOS	0,90	0,90	***0,92***	0,92	10-15-10
	CMOS	-	-	-	-	-

A última coluna da Tabela 8.3 representa a estrutura usada no melhor resultado obtido para a implementação baseada nas redes de Kohonen, onde o primeiro número representa a ordem da vizinhança, o segundo representa o número de neurônios por classe e o terceiro representa o número de parâmetros usados em tal estrutura. Os valores em destaque representam as melhores correlações obtidas para cada caso.

Como se pode observar, a proposta de mapeamento aqui implementada aumenta significativamente os valores das correlações para os casos em que a medida subjetiva MOS foi usada como referência. Nas situações em que a medida CMOS foi tomada como referência para a língua inglesa, o desempenho observado foi ligeiramente inferior, provavelmente porque, em tal caso, as classes não estão bem delimitadas. Conseqüentemente, a rede tende a falhar na classificação de alguns sinais, reduzindo a correlação. É importante frisar que tal observação não se deve à medida subjetiva utilizada, e sim às características dos sinais presentes no segundo experimento, onde a medida CMOS foi tomada como referência. Mas, mesmo nos casos em que não houve melhora no desempenho, os resultados obtidos foram ainda muito bons.

A evolução de desempenho observada pode ser explicada pela capacidade inerente das redes de Kohonen de extrair, a partir dos parâmetros de entrada, a informação que melhor caracteriza as condições testadas. Tal capacidade não é encontrada em mapeamentos polinomiais. Assim, a rede pode identificar, por exemplo, quais sinais foram corrompidos por erros, e então classificá-los apropriadamente, obtendo um bom desempenho para situações em que o método MOQV tende a falhar.

A ampliação da aplicabilidade deste tipo de estrutura está condicionada à disponibilidade de dados associados com uma gama mais abrangente de condições. Sua robustez sob situações para as quais a rede não foi treinada é ainda um ponto a ser investigado, mas uma análise mais profunda do mecanismo de auto-organização permite que se espere um bom desempenho, exceto para sinais com características muito diferentes daquelas encontradas nos sinais usados no treinamento.

O esforço computacional requerido pela rede para treinamento não é importante, uma vez que esta tarefa é executada uma única vez. O esforço requerido pela rede treinada é aproximadamente o mesmo daquele requerido pelo mapeamento polinomial; conseqüentemente, a substituição das técnicas de mapeamento clássicas pelos mapas de Kohonen pode ser feita sem precauções especiais em relação aos recursos computacionais.

8.2.2. Abordagem Baseada no Uso de Redes Neurais do Tipo MLP

A estratégia aqui descrita faz uso de uma rede neural do tipo MLP (*multilayer perceptron*) com uma camada intermediária [102] para fazer o mapeamento entre os parâmetros objetivos e as medidas subjetivas correspondentes. O algoritmo de treinamento utilizado foi o MV-SCGM (*Modified Version of the Scaled Conjugate Gradient Method*), o qual é um dos métodos de segunda ordem mais eficientes para a busca em superfícies multidimensionais não-lineares [119]. O princípio de funcionamento desta abordagem assemelha-se mais aos mapeamentos polinomiais que aos mapeamentos usando redes de Kohonen. Isto se deve ao fato das redes de Kohonen tratarem de problemas de classificação, enquanto que as redes neurais MLP e funções polinomiais tratam de problemas de aproximação. A principal diferença entre estas duas últimas reside no fato das redes neurais do tipo MLP lidarem de maneira muito mais eficiente com a complexidade envolvida na resolução do problema de mapeamento não-linear multidimensional encontrado neste tipo de aplicação.

Os parâmetros usados para alimentar a rede são os mesmos já descritos na seção anterior. A seguir, serão apresentados os testes e resultados obtidos com esta abordagem. Informações mais detalhadas a respeito do algoritmo e estratégia de treinamento podem ser encontradas em [120].

8.2.2.1. Testes Realizados e Resultados Obtidos

Os testes realizados utilizaram a base de dados S-23 [118], a qual é brevemente descrita na Seção 8.2.1.

A rede neural MLP utilizada é composta por 11 entradas, correspondendo aos 10 parâmetros mais a entrada de polarização (*bias*), 12 neurônios na camada intermediária e 1 única saída representando a estimativa da qualidade subjetiva.

Os arquivos de treinamento, a exemplo do que foi feito para a abordagem baseada nos mapas de Kohonen, foram separados de acordo com a língua em que foram gerados, como mostra a Figura 8.11. Os arquivos usados nos testes também foram os mesmos, conforme indicado na Tabela 8.2.

A Tabela 8.4 compara os resultados obtidos. É interessante observar que, mesmo na presença de condições adversas, a estratégia aqui adotada alcançou um excelente desempenho. Como se pode observar, esta proposta superou as outras em todas as situações. O melhor desempenho observado para esta abordagem se deve a dois fatores: a capacidade da rede neural MLP de mapear superfícies complexas e multidimensionais e a informação contida nos parâmetros extraídas a partir das transformadas FFT e MLT, a qual permitiu à rede buscar a melhor superfície de mapeamento para cada caso. Contudo, algumas precauções devem ser tomadas antes que se adote esta abordagem para determinada aplicação, como será comentado na Seção 8.2.3.

Tabela 8.4. Comparação dos Resultados Obtidos Através de Diferentes Estratégias.

Língua	**Medida Subjetiva**	**MOQV**		**Kohonen**		**MLP**
		1	**2**	**1 Dim.**	**2 Dim.**	
Francês	MOS	0,90	0,88	0,95	0,93	***0,97***
	CMOS	0,94	0,94	0,99	0,99	***0,99***
Japonês	MOS	0,72	0,77	0,94	0,92	***0,96***
	CMOS	0,96	0,96	0,98	0,98	***0,99***
Inglês	MOS	0,78	0,80	0,98	0,92	***0,95***
	CMOS	0,96	0,95	0,93	0,93	***0,99***
Italiano	MOS	0,90	0,90	0,92	0,92	***0,95***
	CMOS	-	-	-	-	**-**

8.2.3. Considerações Finais

Ambas as estratégias baseadas em redes neurais apresentaram um desempenho superior àquelas baseadas no mapeamento polinomial. Contudo, sua utilização não pode ser considerada irrestrita, e então algumas considerações se fazem necessárias:

- O uso das redes neurais não elimina a necessidade de se ter um mapeamento particular adaptado às peculiaridades de cada língua, a exemplo do que ocorre nos mapeamentos polinomiais; quando há a necessidade de se utilizar o mapeamento geral, em casos em que a língua dos arquivos de teste em questão não possua um mapeamento individual, as redes de Kohonen, teoricamente, devem ter um desempenho melhor, pois sua característica de tratar o mapeamento como um problema de classificação lhe confere uma robustez não encontrada nas outras abordagens. Em outras palavras, o sinal que se está testando terá que ser designado para uma das possíveis classificações, e a probabilidade maior é que ele seja alocado em um grupo próximo daquele que ele deveria estar de fato. No caso da MLP, há a possibilidade da combinação dos parâmetros se localizar em um

ponto muito ruim da superfície de mapeamento, especialmente se a rede estiver sobretreinada, causando erros substanciais na estimativa desejada. Assim, em situações potencialmente sujeitas à ocorrência de erros de mapeamento, o risco de se ter um desvio importante do valor desejado é muito maior quando do uso das redes MLP.

- A quantidade e qualidade dos sinais usados no treinamento das redes são também fatores preponderantes no desempenho das propostas apresentadas. Quanto maior a base de dados utilizada, melhor deve ser o treinamento. Ainda mais importante que o tamanho do conjunto de treinamento, é a representatividade dos componentes desse conjunto. Quanto maior o número de condições contempladas na base de dados, maior será a gama de condições para as quais a estratégia de mapeamento poderá ser utilizada. É importante observar que as redes de Kohonen são menos sensíveis ao tamanho do conjunto de treinamento utilizado. Assim, a estratégia usando as redes neurais MLP só será superior se ela tiver à sua disposição um banco de dados suficientemente amplo para que ela possa criar uma superfície de mapeamento que realmente se adapte às condições encontradas na prática.

- Finalmente, é importante observar que os tipos de sinais utilizados no treinamento e nos testes são os mesmos e, portanto, não foi possível testar a robustez das estratégias frente a condições desconhecidas. Novamente, observa-se a mesma situação descrita no primeiro item, ou seja, a estratégia baseada na rede de Kohonen terá mais chance de ter um bom desempenho que aquela baseada na MLP.

É importante frisar que a base de dados utilizada, apesar de ser reconhecidamente restrita, é a mais ampla já desenvolvida. Devido aos altos custos envolvidos e às dificuldades inerentes ao processo, a criação de uma base de arquivos de voz realmente representativa é muito improvável. Além disso, a todo o momento surgem novos equipamentos e meios de transmissão que introduzem novas características aos sinais a eles submetidos, de maneira que uma base de dados considerada representativa pode, rapidamente, ser superada.

A única maneira de se ter um método verdadeiramente competente é através do desenvolvimento de um modelo perceptual realmente fiel às características auditivas humanas, pois desta maneira o método, ao se deparar com uma nova condição, será capaz de processá-la da mesma forma que uma pessoa faria, eliminando a necessidade de estratégias de mapeamento ou outros processamentos que pouco têm em comum com o processo auditivo.

Esta última conclusão motivou o direcionamento da pesquisa para um modelo perceptual mais robusto. Os resultados obtidos permitiram a proposta de um novo método, o qual será descrito na Seção 8.3 a seguir. É importante destacar que a implementação deste método é similar àquela adotada para o método PESQ. Contudo, ela não é exatamente a mesma por dois motivos: 1) não se tem toda a informação sobre o PESQ na literatura disponível; 2) foram introduzidas soluções próprias, seja por falta de informação ou por tentativa de aperfeiçoamento.

8.3. UM NOVO MODELO PSICO-ACÚSTICO – O MÉTODO AOSV

Como comentado anteriormente, o modelo psicoacústico utilizado no método MOQV apresenta algumas falhas que o tornam inadequado para uma série de aplicações, limitando assim sua utilização. Dentre esses problemas, dois se destacam pela sua importância e por serem comuns à maioria dos métodos objetivos de avaliação de voz:

- *Modelagem grosseira do mascaramento*: a inclusão explícita do mascaramento nos métodos existentes vem sendo sistematicamente testada, sempre com resultados decepcionantes [121,122]. Nem mesmo o PESQ [113], o mais bem sucedido método objetivo para avaliação de voz existente, foi capaz de lidar adequadamente com este problema. Tal observação é contra-intuitiva, uma vez que o mascaramento nunca deixa de ocorrer, qualquer que seja o evento auditivo. Assim, seria de se esperar que sua inclusão melhorasse as correlações obtidas, a exemplo do que ocorre com métodos de avaliação de áudio. Este problema foi aqui abordado, porém os resultados foram também decepcionantes. Uma solução final para este problema ainda está em estudo por diversos pesquisadores [121].

- *Estrutura excessivamente empírica para o modelo psico-acústico*: é desejável que a modelagem psico-acústica de um método de avaliação objetiva seja o mais fiel possível aos modelos provenientes de estudos fisiológicos da audição humana. Muitos métodos, incluindo o MOQV, são excessivamente dependentes de uma otimização empírica de determinados parâmetros. O primeiro método bem sucedido na tentativa de implementar um modelo psico-acústico com tal característica foi o PESQ. Por esse motivo, ele foi usado neste trabalho como base no desenvolvimento de um novo modelo perceptual que, em conjunto com um processamento cognitivo mais eficiente, deu origem a um novo método, denominado "Avaliação Objetiva de Sinais de Voz" (AOSV) [123].

A Figura 8.9 mostra a estrutura geral do AOSV. O processamento psico-acústico de fato inicia-se somente após o bloco 7, porém uma descrição dos blocos anteriores se faz necessária para um melhor entendimento do método como um todo.

8.3.1. Descrição da Rotina AOSV

8.3.1.1. Determinação do Início e Final Efetivos

O procedimento para detecção do início e final efetivos dos sinais, indicado pelo bloco 1 na Figura 8.9, é idêntico àquele utilizado no método MOQV. A primeira amostra será aquela cuja magnitude, somada às magnitudes das quatro amostras anteriores, supera determinado valor (200 no caso de sinais amostrados a 16 bits). Da mesma maneira, a amostra final será aquela cuja magnitude, somada à magnitude das 4 amostras subseqüentes, é superior ao valor considerado para a primeira amostra. As amostras que não estão contidas no intervalo determinado pelas amostras inicial e final são descartadas.

Figura 8.9 - Estrutura geral do método AOSV.

8.3.1.2. Compensação do Ganho do Sistema

Após a leitura dos sinais e a determinação de seu início e final efetivos, realiza-se a compensação do ganho global do sistema sob teste. Ambos os sinais são escalonados para um mesmo nível de energia. Este nível é de 79 dB SPL no ponto de referência do ouvido [74]. O alinhamento de nível é baseado na potência das versões filtradas dos sinais para a faixa de telefonia (300-3400 Hz). A Figura 8.10 mostra a resposta em amplitude do filtro utilizado.

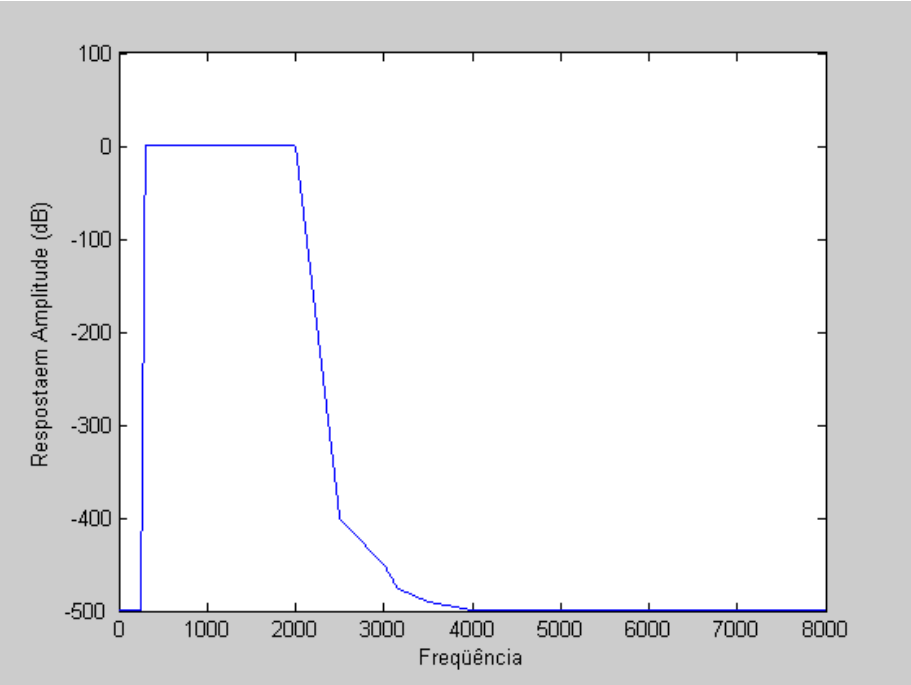


Figura 8.10 - Resposta em amplitude do filtro usado na calibração.

É importante ressaltar que a filtragem aqui realizada não é, a rigor, uma filtragem convencional. Na verdade, ao invés de se utilizar equações a diferenças ou qualquer outro método tradicional de filtragem, simplesmente aplicou-se uma função de ponderação no domínio da freqüência, conforme ilustrado na Figura 8.10. Esta função foi projetada para que, ao se retornar ao domínio do tempo, o resíduo imaginário presente nos sinais seja mínimo. De fato, ao se extrair a transformada inversa, os sinais se tornam complexos. Contudo, a parte imaginária é tão pequena que pode ser eliminada do restante do processo. Esta observação é válida também para a filtragem descrita na Subseção 8.3.1.3.

8.3.1.3. Modelagem das Características do Aparelho Telefônico

Este estágio é indicado pelos blocos 3 e 5 na Figura 8.9. Este procedimento possui duas diferenças básicas em relação à estratégia utilizada no MOQV: 1) aqui ela é aplicada antes do início da modelagem psico-acústica e 2) é usada duas vezes. Assume-se que os testes subjetivos são realizados utilizando-se aparelhos telefônicos com uma resposta no domínio da freqüência que segue a característica de recepção IRS [124]. Um modelo perceptual fiel à avaliação realizada pelos seres humanos deve levar isto em conta, a fim de modelar os sinais que os ouvintes de fato ouvem.

Antes da aplicação da rotina para compensação do atraso variável (deste ponto em diante denominada simplesmente RAV), a qual é descrita na Seção 8.1 e representada no bloco 4 da Figura 8.9, os sinais são filtrados pela primeira vez, de acordo com a resposta em amplitude apresentada na Figura 8.11. Esta primeira filtragem foi omitida nas primeiras versões do programa, mas testes revelaram que a RAV funciona melhor se os sinais estão filtrados. Este fenômeno é devido ao fato de que quanto mais "acidentada" é a forma de onda dos sinais no domínio do tempo, mais eles se parecem com ruído, tornando mais difícil para a rotina identificar corretamente os atrasos. A filtragem elimina os componentes de alta freqüência, os quais na verdade não seriam ouvidos pelos usuários de um sistema telefônico, suavizando a forma de onda temporal dos sinais. Desta maneira, há uma maior probabilidade da RAV realizar alinhamentos corretos.

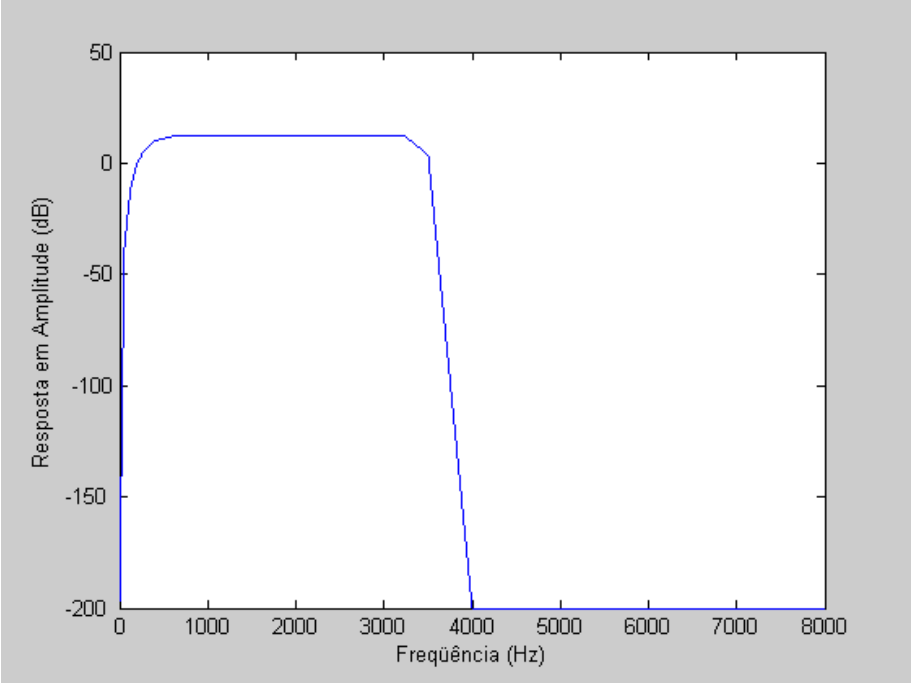


Figura 8.11 - Resposta em amplitude do filtro simulando as características do aparelho telefônico.

Após a filtragem, os sinais são transformados de volta para o domínio do tempo e submetidos à RAV. Após o alinhamento, os sinais são novamente submetidos à filtragem com as características do aparelho telefônico, de modo a eliminar possíveis componentes espúrios surgidos após o alinhamento.

8.3.1.4. Realinhamento de Quadros Ruins

Como comentado anteriormente, os sinais devem ser divididos em quadros antes de serem submetidos aos processamentos subseqüentes. É possível que alguns desses quadros ainda apresentem desalinhamento, mesmo após serem submetidos à RAV. Tais trechos, chamados "quadros ruins", devem ser identificados e tratados. Assim, o propósito desta etapa, representada pelo bloco 6 na Figura 8.9, é dividir os sinais e identificar possíveis quadros desajustados. Aqui, a rotina é dividida em dois possíveis caminhos:

1) Se os sinais possuíam atraso constante antes do alinhamento, isto implica que os sinais como um todo foram alinhados de uma só vez; neste caso, não haverá a ocorrência de

quadros ruins, e então os sinais são simplesmente divididos em quadros de 32 ms com uma janela de Hanning. Os quadros apresentam uma superposição de 50 %.

2) Se os sinais possuíam atrasos variáveis, uma nova estratégia de alinhamento é iniciada, a fim de compensar possíveis desalinhamentos remanescentes da RAV. Esta estratégia é implementada fora da RAV a fim de usar a divisão dos sinais diretamente no restante da rotina AOSV. Se este teste fosse incorporado à RAV, os sinais deveriam ser reagrupados, filtrados com as características do aparelho telefônico, e novamente divididos. Tal estratégia adicionaria esforço computacional desnecessário.

Como se pode observar na Figura 8.12, a estratégia adotada para este último realinhamento, no caso dos sinais apresentarem atraso variável, é composta das seguintes etapas:

1. Os sinais são divididos em quadros de 32 ms mais 50 amostras adicionais (562 amostras para sinais amostrados a 16 kHz), com uma superposição de 16 ms e sem janelamento.
2. Calcula-se a correlação cruzada para cada quadro e o índice do seu ponto de máximo determina o atraso remanescente.
3. Se o atraso remanescente for menor ou igual a 50 amostras, realinha-se os quadros. Isto implicará numa perda de amostras igual ao atraso determinado, justificando a adoção de 50 amostras adicionais na divisão em quadros.
4. As amostras em excesso no final dos quadros são eliminadas, de maneira que cada quadro tenha uma duração exata de 32 ms. Por exemplo, se o atraso obtido para determinado quadro foi de 20 amostras, o total de amostras em excesso a serem eliminadas será de 50 – 20 = 30.
5. Se o atraso remanescente for maior que 50 amostras, determina-se se os quadros em questão são compostos de ruído ou de voz, tendo como base seu nível: um quadro será considerado voz se a soma dos valores absolutos de suas amostras for maior que 100.000 ou se a correlação entre os quadros correspondentes em ambos os sinais for maior que 0,5. Se os quadros forem de voz, seu desajuste é considerado excessivo e eles são eliminados do restante da rotina. Se os quadros forem compostos de ruído, não será possível obter um alinhamento adequado e os quadros são deixados como estão. O classificador neural usado na RAV não foi aqui utilizado porque, neste caso, a precisão requerida para a classificação não é elevada. Desta maneira, seu uso representaria um aumento desnecessário do esforço computacional.
6. Por fim, aplica-se uma janela de Hanning aos quadros de 32 ms e aplica-se o restante da rotina.

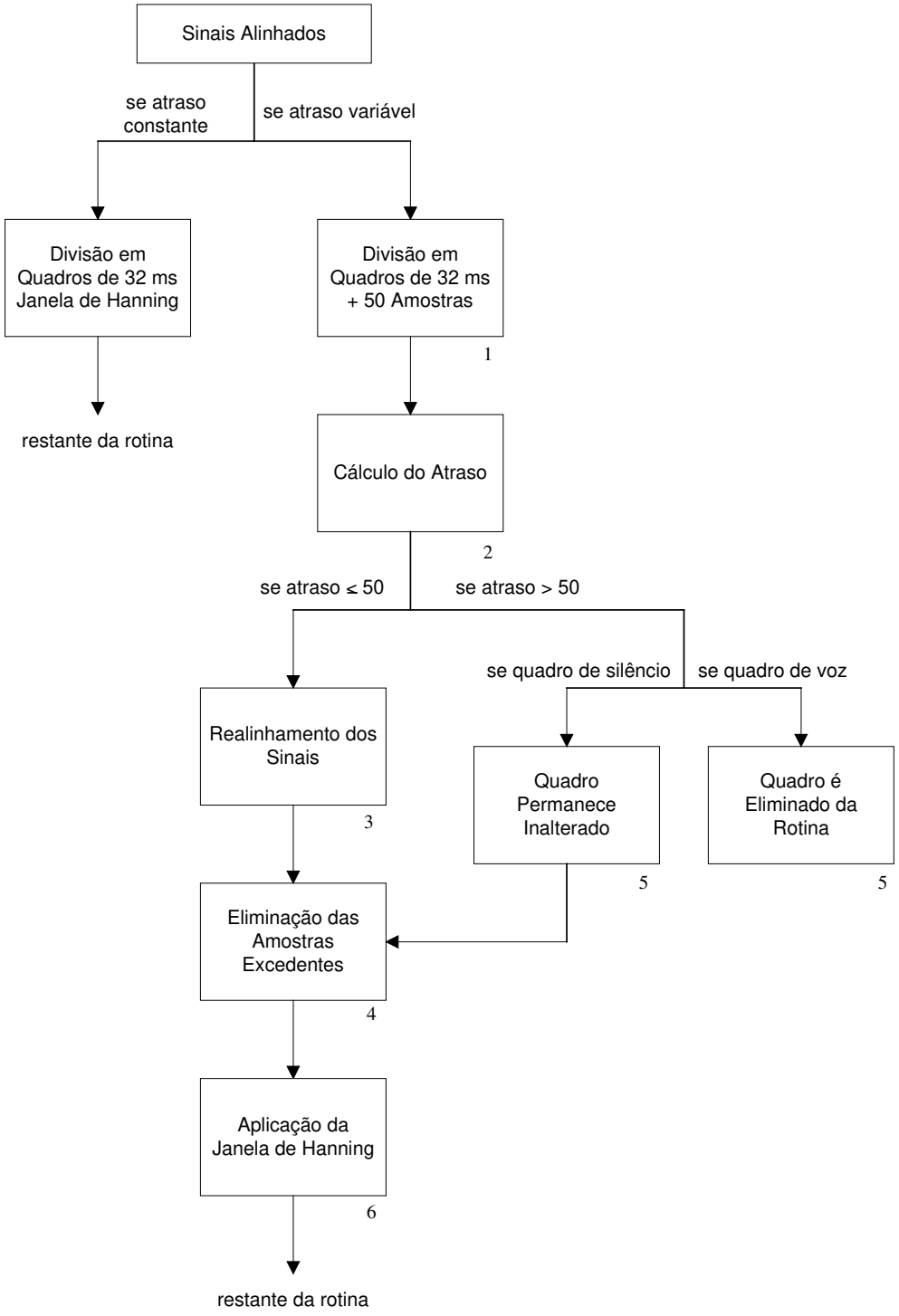


Figura 8.12 - Estratégia para realização do último alinhamento entre os sinais.

8.3.1.5. Energia dos Quadros e Agrupamento em Bandas Auditivas

Esta etapa é similar à correspondente no método MOQV [89], e é representada pelos blocos 7, 8 e 9 na Figura 8.12. Aqui, os sinais são submetidos a uma FFT e a densidade espectral de energia é calculada. Após, essas energias são agrupadas em um número de sub-bandas de freqüência determinadas pelo conceito de bandas críticas [28]. Em outras

palavras, a energia de uma dada sub-banda é determinada pela soma das energias de todos os componentes localizado dentro de seus limites.

Existem algumas diferenças entre os procedimentos adotados para os métodos MOQV e AOSV, como os valores de alguns parâmetros e fatores de correção, e uma importante modificação no número de sub-bandas, que é de 67 para o MOQV e 49 para o AOSV. O número de 49 sub-bandas foi escolhido por modelar melhor a percepção sonora humana. Os padrões resultantes são chamados de *Densidade de Energia de Pitch*.

8.3.1.6. Compensação da Resposta Espectral Linear

Sob certas condições, o sinal degradado pode ser filtrado pelo sistema sob teste, modificando seu espectro. Na maioria das vezes, tal filtragem não é percebida pelo usuário, mas pode causar uma severa degradação na qualidade subjetiva estimada pela rotina. Assim, os espectros de ambos os sinais devem ser novamente ajustados, a fim de minimizar esse efeito (bloco 10 na Figura 8.9). O primeiro passo é calcular a média temporal das densidades de energia de *pitch* de ambos os sinais. Esta média é calculada considerando-se somente componentes cuja potência é pelo menos 20 dB maior que o limiar absoluto de audibilidade [28]. A seguir, calcula-se a relação entre as médias obtidas para os sinais original e degradado. Após o valor máximo da relação ser limitado em 20 dB, ela é usada como uma compensação parcial para a densidade de energia de *pitch* do sinal original, a fim de equalizar os sinais.

É importante notar que o limite de 20 dB imposto para a relação entre os espectros de energia visa modelar o fato de que filtragens severas podem ser detectadas pelos ouvintes como uma degradação. Esse limite garante que diferenças maiores que 20 dB entre os espectros serão apenas parcialmente compensadas. Em outras palavras, se a relação é menor que 20 dB, assume-se que os ouvintes não perceberão qualquer degradação, e então tal diferença é inteiramente compensada. Por outro lado, valores maiores que 20 dB são considerados incômodos, e então a rotina deve considerar tal degradação.

8.3.1.7. Compensação do Ganho Variante com o Tempo

Em certos casos, o ganho pode flutuar ao longo do tempo, causando diferenças entre os sinais. Da mesma forma que para a compensação descrita na Subseção 8.3.1.6, tais variações deverão ser parcialmente ou inteiramente compensadas, dependendo de sua intensidade. Para cada quadro, os valores de todas as amostras que excedem o limiar absoluto de audibilidade são usados para computar as energias dos sinais original e degradado. Então, a relação entre as energias de cada quadro é calculada e limitada entre 0,0003 e 5. As relações localizadas dentre desses limites são totalmente compensadas. Após, um filtro passa-baixas de primeira ordem é aplicado a tais relações, a fim de suavizar possíveis picos agudos. A densidade de energia de *pitch* de cada quadro é então multiplicada pelas relações suavizadas correspondentes.

8.3.1.8. Cálculo das Densidades de Sonoridade

Esta etapa é também muito parecida com sua correspondente no método MOQV. Este procedimento transforma as densidades de potência temporalmente compensadas em uma escala de sonoridade em Sônons, conforme a equação

$$L[i] = E_q \cdot \left(\frac{S_0}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{S_V[i]}{S_0[i]}\right)^{\gamma} - 1\right], \tag{8.6}$$

onde $E_q = 0.1866055$ é um fator de calibração, S_v é a energia do componente espectral que se está considerando, γ é a potência de Zwicker [28], a qual varia de 2 para as freqüências mais baixas, até 1 para as freqüências mais elevadas, i é o índice espectral das amostras e S_0 é o limiar absoluto de audibilidade dado por

$$S_0 = 3{,}64 \cdot f^{-0{,}8} - 6{,}5 \cdot e^{-0{,}6\cdot(f-3{,}3)^2} + 10^{-3} \cdot f^4. \tag{8.7}$$

Os padrões resultantes desta etapa são denominados densidades de sonoridade.

8.3.1.9. Cálculo das Densidades de Distúrbio

Esta etapa calcula a diferença entre as densidades de sonoridade ($L[i]$) dos sinais degradado e original, correspondente ao distúrbio percebido entre os sinais. Se a diferença é positiva, componentes do tipo ruído terão sido adicionados. Quando a diferença é negativa, componentes foram suprimidos. Como resultado, tem-se a "densidade de distúrbio preliminar".

O efeito de mascaramento é modelado através da aplicação de algumas regras a cada componente do plano tempo-freqüência. Primeiro, determina-se o valor mínimo entre cada componente do sinal original e seu correspondente no sinal degradado, resultando nos *valores de mascaramento*. A seguir, as seguintes regras são aplicadas a cada componente, onde *vm* é o valor de mascaramento e *ddp* é a densidade de distúrbio preliminar:

- Se $ddp \geq vm$, o valor de mascaramento é subtraído do distúrbio preliminar.
- Se $-vm < ddp < vm$, o valor do distúrbio é feito igual a zero.
- Se $ddp \leq -vm$, então o valor de mascaramento é adicionado ao distúrbio preliminar.

O efeito deste conjunto de instruções é aproximar as densidades de distúrbio do valor nulo. Além disso, esse procedimento determina uma região para a qual não se percebe uma distorção, o que ocorre quando a densidade de distúrbio preliminar é menor que o valor absoluto do mascaramento. Desta forma, modela-se o fenômeno em que pequenas distorções são inaudíveis na presença de um componente de mesma freqüência, porém de maior amplitude. O resultado é denominado "densidade de distúrbio".

É importante ressaltar que esta estratégia faz uso de artifícios que aproximam, de maneira relativamente grosseira, os efeitos do mascaramento. Como já comentado anteriormente, a modelagem explícita do mascaramento ainda requer estudos adicionais.

8.3.1.10. Modelagem do Efeito de Assimetria

O efeito de assimetria consiste do fenômeno em que a adição de componentes estranhos ao sinal é mais incômoda que sua subtração. Este efeito é modelado calculando-se uma densidade de distúrbio assimétrica para cada quadro, através da multiplicação da densidade de distúrbio convencional por um fator de assimetria. Esse fator consiste da relação entre as densidades de potência de *pitch* dos sinais degradado e original. Se o efeito de assimetria obtido após a multiplicação é menor que 1, ele é feito igual a zero. Se ele excede 12, é limitado a esse valor. Assim, os valores serão diferentes de zero apenas em casos em que o valor de um componente do sinal degradado supera o valor de seu correspondente no sinal original. O valor adotado para o limite superior (12) visa evitar o surgimento de valores muito grandes para o índice de distúrbio assimétrico quando

componentes do sinal original têm uma densidade de energia de *pitch* muito pequena em relação à sua correspondente no sinal degradado. Os valores aqui obtidos são ponderados e somados à densidade de distúrbio convencional após a aplicação dos processamentos descritos nas Subseções 8.3.1.11 e 8.3.1.12. Este efeito será novamente abordado nas Subseções 8.3.1.13 e 8.3.2.

8.3.1.11. Integração das Densidades de Distúrbio ao Longo do Eixo da Freqüência e Processamento dos Intervalos de Silêncio

As densidades de distúrbio convencional (descrita na Subseção 8.3.1.9) e assimétrica (descrita na Subseção 8.3.1.10) são integradas (somadas) ao longo do eixo da freqüência usando-se duas diferentes normas modificadas e uma ponderação para os quadros de baixa energia, conforme as equações

$$D(t) = M_n \cdot \sqrt[3]{\sum_{f=1}^{b} \left(\left| D(f,t) \right| \cdot W_f \right)^3}\ , \tag{8.8}$$

$$DA(t) = M_n \cdot \sum_{f=1}^{b} \left(\left| DA(f,t) \right| \cdot W_f \right), \tag{8.9}$$

onde M_n é um fator de multiplicação de valor $\left[\left(pot.sinal\,orig. + 10^5 \right) / 10^7 \right]^{-0.04}$, resultando em uma ênfase nos distúrbios que ocorrem durante intervalos de silêncio; W_f é uma série de constantes proporcionais à largura das sub-bandas; b é o número de sub-bandas, D e DA são as densidades de distúrbio convencional e assimétrica, respectivamente, e t é o índice das amostras no tempo.

8.3.1.12. Integração do Distúrbio ao Longo do Tempo

As densidades de distúrbio resultantes da integração no domínio da freqüência são divididas em intervalos de 20 quadros, ou 320 ms. Os intervalos também apresentam uma superposição de 50%. As densidades de distúrbio são agregadas em cada intervalo usando-se uma norma L_6. A seguir, os distúrbios são integrados ao longo de todo o comprimento do sinal usando-se uma norma L_2. Essas normas são aplicadas de acordo com a equação

$$L_p = \left(\frac{1}{N} \cdot \sum_{n=1}^{N} distúrbio(n)^p \right)^{\frac{1}{p}}, \tag{8.10}$$

onde p é a ordem da norma (2 ou 6) e N é o número de quadros. O valor $p = 6$ adotado para a integração em cada intervalo acarreta uma forte ênfase nas distorções mais intensas, devido ao fato de que quando pequenos trechos do intervalo estão severamente distorcidos, todo o intervalo pode perder o sentido. No caso do sinal como um todo, a presença de uma sentença distorcida não implica na perda de sentido de todo o sinal. Assim, uma norma de menor ordem deve ser aplicada, a fim de conferir uma menor ênfase às distorções de grande intensidade.

8.3.1.13. Determinação do Valor AOSV

O valor AOSV resulta de uma combinação linear entre os valores resultantes para os distúrbios convencional e assimétrico (bloco 17 na Figura 8.9). No método AOSV, há duas possíveis combinações, enquanto que no método PESQ há apenas uma [112]. Isso se deve

ao fato de se ter observado diferenças importantes no padrão de comportamento dos avaliadores dependendo do tipo de medida subjetiva que se está estimando. Esta discussão será retomada mais adiante, durante a apresentação dos resultados.

8.3.2. Resultados Obtidos

Os testes foram realizados utilizando os experimentos contidos na base de dados S-23 [118], brevemente descrita na Seção 8.2, e os arquivos de voz fornecidos pelo CPqD e descritos na Seção 8.1. A seguir, são apresentados os resultados obtidos e algumas curvas que ilustram o desempenho do método para cada situação. A comparação entre os métodos MOQV e AOSV será apresentada na Subseção 8.3.2.5.

8.3.2.1. Experimento 1 da base de dados S-23

A correlação média obtida para este experimento foi elevada (0,963). Testou-se o desempenho do método suprimindo-se a modelagem do efeito de assimetria, resultando em correlações abaixo de 0,85; portanto, o efeito de assimetria exerce um papel fundamental neste tipo de avaliação. A Figura 8.13 ilustra o padrão de resultados obtidos para este experimento; a figura em questão foi gerada para a língua inglesa utilizando-se um mapeamento polinomial de terceira ordem.

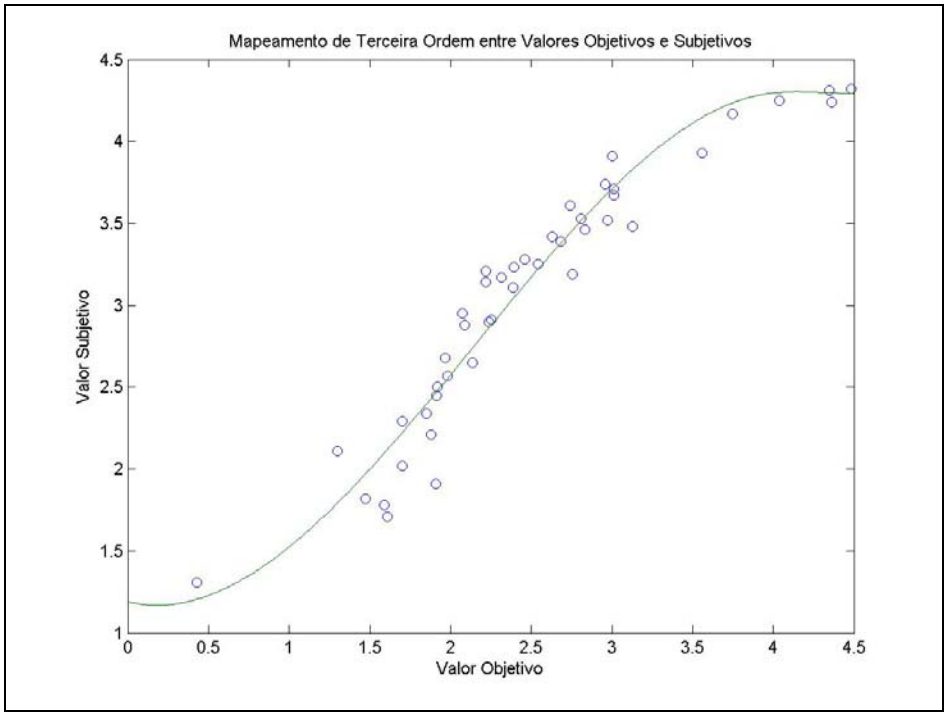


Figura 8.13 - Exemplo de resultado obtido para o primeiro experimento

8.3.2.2. Experimento 2 da Base de Dados S-23

Os resultados aqui observados foram também muito bons, com uma correlação média de 0,965, praticamente a mesma obtida para o primeiro experimento. Durante os testes realizados para este experimento, observou-se que os resultados pioravam ao se considerar o efeito de assimetria, com a correlação caindo para uma média de cerca de 0,91. Este fenômeno pode ser explicado pelo fato deste experimento apresentar como medida

subjetiva de referência o CMOS, onde o sinal original também é apresentado ao ouvinte para efeito de comparação [74]. Desta forma, o ouvinte será capaz de detectar com mais precisão a supressão de determinados componentes, o que não ocorre em testes do tipo ACR (*Absolute Category Rating*), como o MOS. Assim, a supressão e a adição de componentes serão detectadas aproximadamente com a mesma precisão, e o efeito de assimetria deixa de existir. Na versão final da rotina AOSV, o usuário pode optar por mapear os valores objetivos para as medidas subjetivas MOS ou CMOS. Se escolher a primeira opção, o efeito de assimetria será considerado; se optar pelo segundo, o efeito de assimetria é ignorado.

A Figura 8.14 ilustra os resultados obtidos para o segundo experimento. A figura foi gerada a partir de sinais na língua francesa, e a curva foi determinada através de um mapeamento polinomial de terceira ordem.

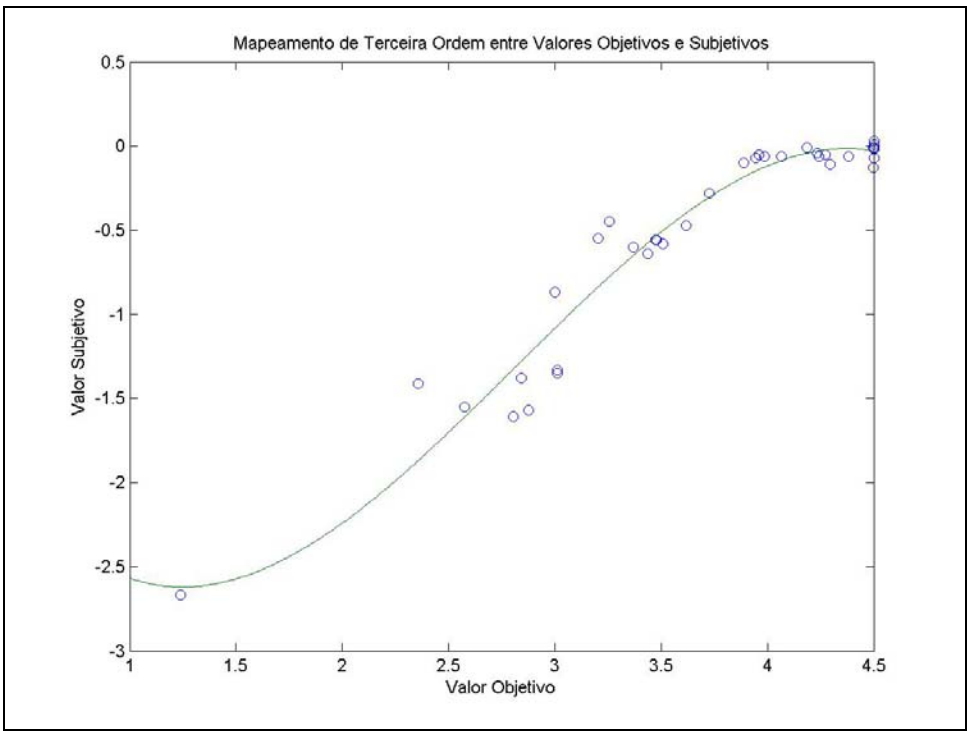


Figura 8.14 - Exemplo de resultado obtido para o segundo experimento

Note-se que há uma pequena perda de monotonicidade nos extremos da curva; por ser muito pequena, ela pode ser considerada desprezível e a curva pode ser tomada sem restrições. Os resultados apresentados na figura foram obtidos sem a modelagem do efeito de assimetria.

8.3.2.3. Experimento 3 da Base de Dados S-23

A correlação média aqui obtida foi de 0,9232, também superior àquela obtida para o método MOQV, como será visto mais adiante. A Figura 8.15 mostra o resultado obtido para este experimento, usando os arquivos de língua japonesa e mapeamento polinomial de terceira ordem.

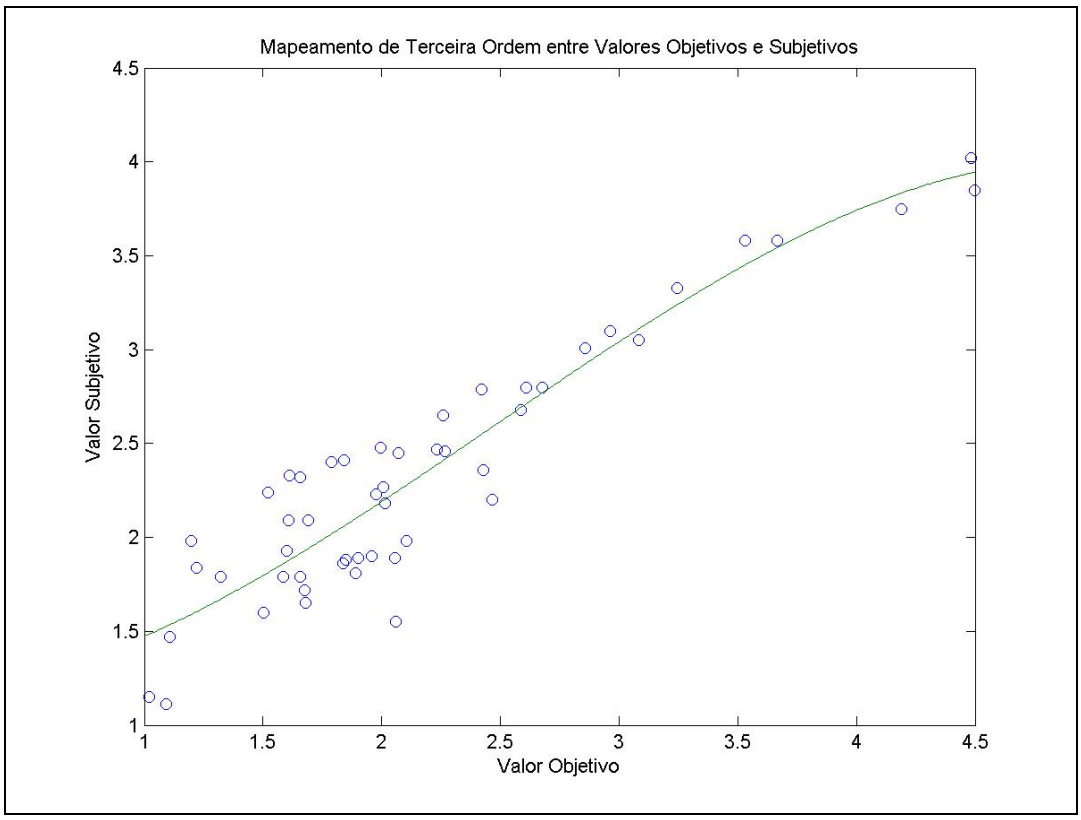


Figura 8.15 - Exemplo de resultado obtido para o terceiro experimento

8.3.2.4. Base de Dados em Português do CPqD

Esta foi a única base de dados com atraso variável testada. A correlação obtida foi mais elevada que aquela obtida nos testes da rotina de atraso variável aplicada ao método MOQV (0,9440 contra 0,9173). Esta correlação pode ser considerada muito boa, especialmente para sinais apresentando atrasos fortemente variáveis; em outras palavras, o método foi validado para as situações mais extremas, em termos do atraso, que se poderia encontrar na prática. Infelizmente, o número de arquivos contidos nesta base de dados é reduzido, o que não permite que se determine uma curva de mapeamento realmente confiável para a língua portuguesa falada no Brasil. No entanto, ela é suficiente para que se possa afirmar que o método funciona adequadamente em situações de atraso variável, atestando sua ótima robustez. Como os valores subjetivos de referência aqui utilizados são do tipo MOS, a modelagem do efeito de assimetria foi incluída. A Figura 8.16 mostra os resultados obtidos para esta base de dados.

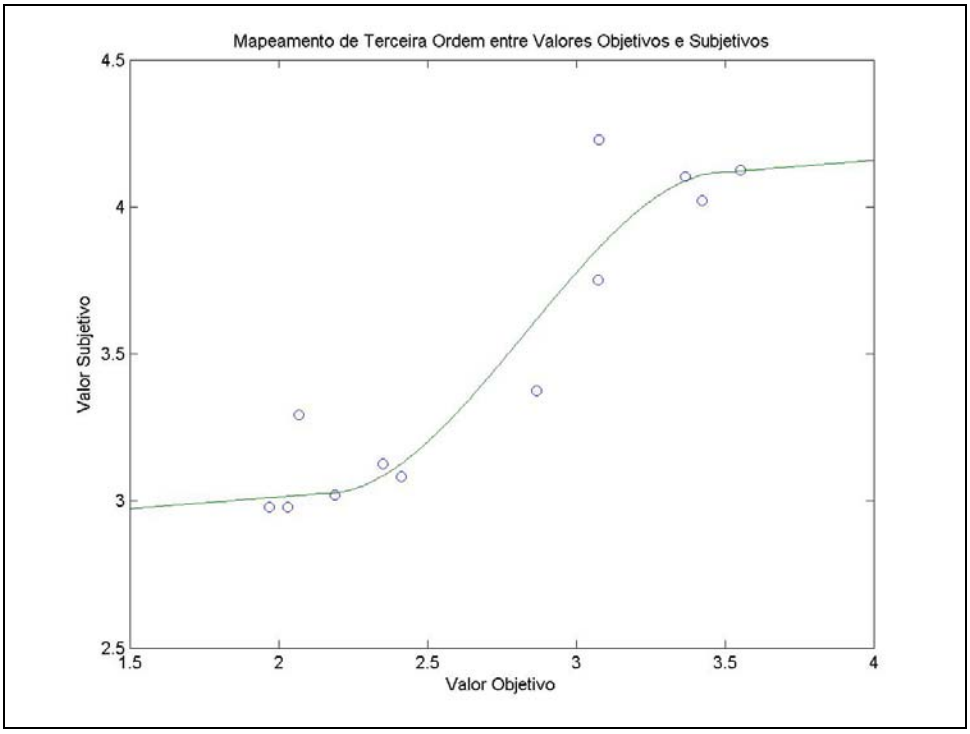


Figura 8.16 - Resultado obtido para a base de dados em português

8.3.2.5. Comparação entre os Métodos

A Tabela 8.5 apresenta uma comparação entre os resultados obtidos para os métodos MOQV, AOSV e PESQ original.

Tabela 8.5 - Comparação entre os Resultados Obtidos para os Métodos MOQV, PESQ e AOSV

Experimento	**Língua**	**MOQV1**	**MOQV2**	**PESQ**	**AOSV**
1º	Francês	0,947	0,943	0,920	***0,956***
	Japonês	0,946	0,939	0,939	***0,968***
	Inglês	0,959	0,962	0,943	***0,967***
2º	Francês	0,937	0,936	0,942	***0,974***
	Japonês	***0,957***	0,956	0,925	0,954
	Inglês	0,959	0,955	0,929	***0,966***
3º	Francês	0,896	0,898	0,871	***0,908***
	Italiano	0,899	0,899	0,929	***0,932***
	Japonês	0,900	0,901	***0,942***	0,929
	Inglês	0,887	0,886	0,916	***0,925***
CPqD	Português	0,452	0,476	***0,964***	0,944
Média		0,885	0,886	0,929	***0,948***

Como se pode observar, os resultados obtidos com o novo método foram os melhores em 8 das 11 situações, contra 2 do método PESQ e 1 do método MOQV1. Mesmo nos casos em que o AOSV foi superado, ele apresentou um desempenho aceitável. Sua superioridade em relação ao método MOQV demonstra a evolução do modelo psico-acústico e cognitivo utilizado.

8.4. CONCLUSÕES

Este capítulo apresentou diversos avanços obtidos na avaliação objetiva de sinais de voz, os quais podem ser divididos em três grupos:

1. *Compensação do atraso*: desenvolveu-se uma rotina funcional para compensação de atrasos variáveis no contexto da avaliação da qualidade de voz. Esta rotina representa uma opção vantajosa frente à estratégia utilizada no método PESQ, a qual, além de não contar com uma descrição detalhada na literatura disponível, é protegida por direitos autorais. Os testes realizados demonstraram que o desempenho da estratégia aqui adotada é muito próximo daquele obtido para o PESQ.

2. *Estratégias de mapeamento*: novas estratégias de mapeamento entre as medidas objetivas e subjetivas foram desenvolvidas, tendo como base a teoria de redes neurais. Dois tipos de redes neurais foram utilizados para este fim: as redes de Kohonen e as redes do tipo MLP. O desempenho de cada possível estratégia foi determinado levando-se em conta dois fatores fundamentais: capacidade de aproximação e robustez. Tais estudos levaram às seguintes conclusões:

- As redes neurais são uma opção vantajosa nos casos em que as técnicas clássicas tendem a falhar, bem como quando se dispõe de uma base de dados para treinamento ampla e representativa. Contudo, as redes neurais possuem uma robustez comparativamente pequena, de modo que, se o método apresenta um desempenho satisfatório utilizando apenas técnicas clássicas, seu uso deve ser evitado.

- No caso de se optar pelo uso de redes neurais, deve-se levar em consideração as características peculiares a cada um dos tipos de rede testados: as redes MLP são muito eficientes quando se tem certeza que a base de dados usada para treiná-la é realmente representativa, e que raramente ocorrerão casos em que o sinal testado é muito diferente daqueles presentes no banco; por outro lado, as redes de Kohonen são mais robustas e fazem um uso mais eficiente da informação contida na base de dados disponível, especialmente se esta é limitada em tamanho e/ou representatividade. Para as bases de dados aqui utilizadas, o uso das redes MLP se mostrou mais eficiente.

3. *Modelo psico-acústico*: um novo modelo psico-acústico, mais fiel às reais características da audição humana do que aquele usado no método MOQV, foi desenvolvido. Como resultado, surgiu um novo método, denominado AOSV.

O método AOSV incorpora alguns dos principais avanços obtidos na avaliação objetiva de voz, incluindo a rotina para cálculo e eliminação de atrasos variáveis. Portanto, ele representa uma compilação do que de melhor se produziu ao longo deste estudo. As principais conclusões relativas a esse novo método são sintetizadas a seguir:

- Comparando a versão final do AOSV com as versões em que o mapeamento neural foi empregado, percebe-se que o modelo aperfeiçoado do ouvido utilizado no método AOSV confere a este uma robustez que não é encontrada ao se utilizar estratégias alternativas de mapeamento, conforme discutido na Seção 8.2.3. Além disso, caso desejado,

o uso de redes neurais para realização do mapeamento pode ser estendido a este método sem grandes dificuldades; contudo, seu uso só será apropriado para possíveis situações em que o método AOSV, da maneira como está implementado, tenda a falhar.

- É importante observar ainda que a aparente superioridade do PESQ frente ao AOSV para o caso de sinais com atraso variável não é conclusiva. Em primeiro lugar, deve-se notar que a rotina para cálculo dos atrasos apresentou um desempenho muito bom, o que é atestado tanto por testes computacionais quanto por inspeções visuais. Além disso, o AOSV foi superior ao PESQ na maioria das condições com atraso fixo, comprovando que seu modelo psico-acústico é, pelo menos, tão bom quanto aquele usado no PESQ. A explicação para esta diferença reside no fato de que o banco de dados utilizado para testar o método frente a atrasos variáveis é bastante reduzido. Assim, mesmo um pequeno erro na estimativa da qualidade subjetiva para determinado sinal pode representar uma diferença significativa no valor da correlação. Neste caso, houve uma pequena diferença em favor do PESQ, mas o contrário poderia facilmente ter ocorrido. No caso de um banco de dados mais amplo, estes pequenos erros de estimação estariam diluídos no cálculo final da correlação, o que provavelmente acarretaria um desempenho similar entre os dois métodos.

- Pode-se concluir então que a nova rotina apresenta diversas vantagens em relação às suas antecessoras, e os testes realizados revelam uma boa robustez frente a uma série de diferentes condições. Algumas condições não encontradas nas bases de dados utilizadas foram testadas durante a validação do método PESQ por seus autores, sempre resultando em bons desempenhos [121]. Como os princípios utilizados no AOSV são baseados naqueles usados no PESQ, pode-se esperar que este também tenha sucesso frente a tais condições. A eficiência do modelo auditivo utilizado permite ainda que se espere um bom desempenho e robustez frente a condições nunca testadas. Novos testes deverão ser futuramente realizados a fim de que se possam confirmar tais suposições.

CAPÍTULO 9
CONCLUSÕES FINAIS

Apresentou-se neste trabalho um novo método para a avaliação objetiva de sinais de áudio, denominado Medida Objetiva da Qualidade de Áudio (MOQA). Abordou-se também a questão da avaliação objetiva de sinais de voz, onde diversos aspectos do método MOQV, desenvolvido durante o projeto de mestrado, foram aperfeiçoados, dando origem ao método AOSV (Avaliação Objetiva de Sinais de Voz). Ambos os métodos desenvolvidos incorporam recursos que visam facilitar e agilizar seu uso.

A fundamentação teórica envolvida no desenvolvimento deste projeto foi apresentada em quatro capítulos (Princípios Fundamentais da Audição Humana, Codificação de Áudio, Avaliação Subjetiva de Sistemas de Áudio e Medidas Objetivas de Avaliação da Qualidade de Áudio). Tais capítulos foram redigidos de maneira a fornecer ao leitor os subsídios necessários para um melhor entendimento dos procedimentos adotados no desenvolvimento dos métodos citados, procurando, contudo, evitar um aprofundamento excessivo dos conceitos, o que poderia tornar a leitura difícil e cansativa. A lista de referências bibliográficas fornece uma ampla base de informações para os leitores interessados em se aprofundar em algum dos tópicos abordados.

Dentre as diversas dificuldades encontradas ao longo do desenvolvimento deste trabalho, destacou-se a inexistência de uma base de dados de áudio publicamente disponível. Este problema foi parcialmente contornado graças à boa vontade de pessoas como o Dr. Thilo V. Thiede, fator este essencial para o sucesso alcançado por este projeto. Os desafios envolvidos no desenvolvimento das estratégias apresentadas ao longo desta tese foram também consideráveis, exigindo o estudo de assuntos tão diversos como análise espectral, modelagem de filtros, redes neurais, vetorização de códigos, anatomia do ouvido, codificação de voz e áudio, entre outros. Tal fato resultou no domínio de uma ampla gama de técnicas potencialmente úteis não apenas na área de avaliação objetiva, mas também para muitas outras aplicações em processamento digital de sinais e engenharia biomédica.

Como comentado anteriormente, os esforços culminaram no desenvolvimento de dois novos métodos de avaliação objetiva, o MOQA para áudio, e o AOSV para voz. Como discutido ao longo deste trabalho, ambos apresentam características inovadoras, não encontradas em nenhum de seus predecessores. Tais aperfeiçoamentos resultaram em ganhos expressivos, tanto em termos da qualidade das estimativas, quanto em termos da complexidade computacional. Testes demonstraram que seus desempenhos não deixam a desejar nem mesmo quando comparados às mais avançadas técnicas atualmente em uso. Mais que isso, eles representam o domínio de uma tecnologia que está fortemente concentrada em alguns poucos centros de excelência na área (com destaque para a Holanda, a Grã-Bretanha e a Alemanha).

Quanto às perspectivas de pesquisas futuras, as avaliações objetivas de voz e áudio vivem situações opostas. No primeiro caso, as medidas objetivas de voz estão muito próximas do limite teórico de desempenho, e abrange uma ampla gama de situações onde

pode ser aplicada. Assim, os avanços que ainda podem ser alcançados não representarão ganhos expressivos. No caso das medidas objetivas de áudio, por outro lado, há ainda um longo caminho a ser trilhado até que estas possam substituir as medidas subjetivas de maneira confiável. É importante destacar que a continuidade das pesquisas nesta área esbarra em três grandes desafios:

1) Em primeiro lugar, os avanços a serem alcançados no futuro estarão fortemente condicionados à evolução das pesquisas relacionadas ao processamento dos sinais acústicos no córtex auditivo, pois, como se sabe, tais mecanismos ainda são pouco conhecidos. Assim, um pesquisador que deseje obter avanços significativos deverá se concentrar no estudo dos processos cerebrais e, adicionalmente, ser capaz de transferir as descobertas realizadas para o domínio do processamento digital de sinais.

2) Para que as inovações que venham a ser introduzidas possam ser adequadamente testadas, haverá a necessidade de se gerar uma base de dados de áudio ampla e representativa. Esta não é uma tarefa simples, uma vez que o processo para gerar esse tipo de material é muito dispendioso, tanto em termos de custos quanto de tempo. Além disso, é importante que as bases de dados que venham a ser desenvolvidas possam ser amplamente disponibilizadas para todos os pesquisadores interessados. Uma possível maneira de minimizar os problemas envolvidos neste tipo de empreitada seria a formação de redes de laboratórios e instituições unidas num esforço conjunto para o desenvolvimento de bases de áudio de referência.

3) O terceiro grande desafio é determinar qual o limite de precisão que uma técnica objetiva de avaliação de áudio é capaz de atingir. Como comentado anteriormente, o principal objetivo a ser alcançado por uma técnica deste tipo é estimar, da maneira mais precisa possível, a impressão que ouvintes humanos teriam de determinado sinal. Contudo, as impressões subjetivas estão sujeitas a variações imprevisíveis, que envolvem fatores tão diversos quanto humor e preferências pessoais dos ouvintes, ambiente dos testes subjetivos, entre outros. Até mesmo pequenos acontecimentos do cotidiano podem alterar as impressões dos ouvintes. Como comentado no Capítulo 4, critérios estatísticos devem ser aplicados para eliminar avaliadores que destoem da tendência revelada pelos demais. Ainda que tal estratégia minimize o problema, um certo grau de incerteza ainda permanece. Tais variações são muito difíceis de serem modeladas, de modo que a precisão das medidas objetivas ainda é limitada por essa incerteza. A avaliação de áudio é muito mais suscetível a esse problema que a avaliação de voz. Não se sabe ao certo qual a correlação máxima que um método de avaliação de áudio pode atingir, mas é muito improvável que valores superiores a 0,94 possam ser atingidos de maneira consistente no curto prazo.

Apesar das dificuldades apresentadas para a continuidade do trabalho, é importante ter em mente que os estudos que porventura venham a ser realizados, ainda que não resultem em melhorias expressivas de desempenho nos métodos de avaliação objetiva de áudio, poderão ter importantes aplicações em outras áreas de pesquisa.

BIBLIOGRAFIA

[1] Noll, P. *Adaptative Quantizing in Speech Coding Systems*, In Int. Zurich Seminar on Digital Comm., pp. B3.1-B3.6, IEEE, 1974.

[2] *Handbook of Acoustics*, John Wiley & Sons, New York, 1998.

[3] Gulick, W.L. *Hearing - Physiology and Psychophysics*, Oxford Univ. Press, 1971.

[4] MacKay, I.R.A. *Phonetics: The Science of Speech Production*, College-Hill Publication Little, Brown and Co., 1987.

[5] Bittencourt, R. *MPEG-Audio*, http://www.lsi.usp.br/~ricardo/mpeg/, USP, 1997.

[6] Robertson, G. *Critical Bands*, http://tuba.music.gla.ac.uk/~george/audio/psy/psy.html, 1996.

[7] Fletcher, H. *Speech and Hearing in Communication*, D. Van Nostrand Co., Toronto, 1953.

[8] Guydon, A.C. *Fisiologia Humana*, Ed. Guanabara, 6ª Edição, Rio de Janeiro, 1985.

[9] Terhardt, E. *Calculating Virtual Pitch*, Hearing Research, vol. 1, pp. 155-182, 1979.

[10] Zwicker, E.; Feldkeller, R. *Das Ohr als Nachrichtenempfänger*, Hirzel Verlag, Stuttgart, 1967.

[11] Fourcin, A.J. et al, *Speech Processing by Man and Machine: Group Report*, pp. 307-351, Bullock, T., 1977.

[12] Hartmann, W.M. *Pitch, Periodicity, and Auditory Organization*, J. Acoust. Soc. Am. 100, 3491-3502, 1996.

[13] Idson, W.L.; Massaro, D.W. *A Bidimensional Model of Pitch in the Recognition of Melodies*, Perceptual Psychophysics, vol. 24, pp. 551-565, 1978.

[14] Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech*, J. Acoust. Soc. Am., vol. 87, no. 4, pp. 88-94, April 1990.

[15] Terhardt, E. *The SPINC Function for Scaling of Frequency in Auditory Models*, Acustica, vol. 77, pp. 40-42, 1992.

[16] Moore, B.C.J.; Glasberg, B.R. *Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns*, Journal of the Acoustical Society of America, vol. 74, no. 3, pp. 750-753, September 1983.

[17] Schroeder, M.R.; Atal, B.S.; Hall, J.L. *Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear*, Journal of the Acoustical Society of America, vol. 66, no. 6, pp. 1647-1652, December 1979.

[18] Moore, B.C.J.; Glasberg, B.R.; Baer, T.H. *A Model for the Prediction of Thresholds, Loudness, and Partial Loudness*, Journal of the Audio Engineering Society, vol. 45, no. 4, pp. 224-240, April 1997.

[19] Glasberg, B.R.; Moore, B.J. *Derivation of Auditory Filter Shapes from Notched Noise Data*, Hearing Research, vol. 47, pp. 103-138, 1990.

[20] Stuart, J.R. *Implementation and Measurement with Respect to Human Auditory Capabilities*, AES UK DSP Conference, London, September 1992.

[21] Hellman, R.P. *Asymmetry of Masking Between Noise and Tone*, Perception & Psychophysics, vol. 11, no. 3, pp. 241-246, 1972.

[22] Patterson, R.D.; Moore, B.C.J. *Auditory Filters and Excitation Patterns as Representations of Frequency Resolution*, in Frequency Selectivity in Hearing, Academic Press, New York, 1986.

[23] Schroeder, M. et al. *Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception*, In: Frontiers of Speech Communication Research, Academic Press, New York, 1979.

[24] Zwicker, E.; Terhardt, E. *Analytical Expressions for Critical Bandwidth as a Function of Frequency*, Journal of the Acoustical Society of America, vol. 68, no. 5, pp. 1523-1525, November 1980.

[25] Thiede, T.V. *Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank*, Ph.D. Thesis, Berlin, 1999.

[26] Kapust, R. *Qualitätsbeurteilung Codierter Audiosignale Mittels Einer Bark-Transformation*, Dissertation an der Technischen Fakultät der Universität Erlangen-Nürnberg, Erlangen, 1993.

[27] Fastl, H. *Temporal Masking Effects: Critical Band Noise Masker*, Acustica, Vol. 36, pp. 317-331, 1976.

[28] Zwicker, E.; Fastl, H. *Psychoacoustics, Facts and Models*, Springer Verlag, Berlin, 1990.

[29] Moore, B.C.J. *An Introduction to the Psychology of Hearing*, Academic Press, New York, 1989.

[30] Humes, L.E.; Jesteadt, W. *Models of the Additivity of Masking, Journal of the Acoustical Society of America*, vol. 85, no. 3, pp. 1285-1294, March 1989.

[31] Stevens, S.S. *A Scale for the Measurement of a Psychological Magnitude: Loudness, Psychological Review*, vol. 43, pp. 405-416, 1936.

[32] Colomes, C.; Lever, M.; Rault, J.B.; Dehery, Y.F. *A Perceptual Model Applied to Audio Bit-Rate Reduction*, Contribution to the 95th Convention of the Audio Engineering Society, Preprint 3742, New York, October 1993.

[33] Cremer, L.; Hubert, M. *Vorlesungen über Technische Akustik*, Berlin, Springer Verlag, 1985.

[34] Green, D.M. *An Introduction to Hearing*, Lawrence Erlbaum Assoc., Hillsdale, New Jersey, 1976.

[35] ISO 532: *Acoustics - Method for Calculating Loudness Levels*, 1975.

[36] Deutsch, W.A.; Noll, A.; Eckel, G. *The Perception of Audio Signals Reduced by Overmasking to the Most Prominent Spectral Amplitudes*, Contribution to the 92nd Convention of the Audio Engineering Society, Preprint 3331, Vienna, March 1992.

[37] Patterson, R.D. *Auditory Filter Shapes Derived with Noise Stimuli*, Journal of the Acoustical Society of America, vol. 59, no. 3, pp. 640-654, March 1976.

[38] Watkinson, J. *MPEG-2*, Focal Press, Oxford, 1999.

[39] Smyth, M.; Smyth, S. *DTS Coherent Acoustics, The Future of Audio, Part Two: The Sonics of Bit Rate Reduction*, Widescreen Review, USA, Issue 16, 1995, pp. 76-80.

[40] Gleerup, T. *Low-Bit-Rate Audio Coding: A Comparison*, Lecture Course C5126 Audio Engineering, Copenhagen, May 1997.

[41] Brandenburg, K.; Stoll, G. *The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio*, 92nd AES Convention, Preprint 3336, March 1992.

[42] ISO/IEC-11172, *Coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbits/s*, International Standard, 1992.

[43] Le Gall, D. *MPEG: a Video Compression Standard for Multimedia Applications*, Communications of the ACM, vol. 34, no. 4, pp. 46-58, 1991.

[44] MPEG Video Standard: ISO/IEC 13818-2, *Information Technology: Generic Coding of Moving Pictures and Associated Audio Information: Video*, 1994.

[45] Brandenburg, K.; Bosi, M. *Overview of MPEG Audio: Current and Future Standards for Low-Bit-Rate Audio Coding*, JAES, Vol. 45, Nos. 1/2, pp. 4-21, January/February 1997.

[46] Jayant, N.S.; Noll, P. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice Hall, Englewood Cliffs, 1984.

[47] Proakis, J.G.; Manolakis, D.G. *Digital Signal Processing: Principles, Algorithms and Applications*, Third Edition, Prentice-Hall International, 1996.

[48] Purnhagen, H. *MPEG Audio FAQ, MPEG-2: coded transmission/storage of sampled sound waves*, November 2001.

[49] ISO/IEC 13818-3, *Information Technology: Generic coding of Moving pictures and associated audio - Audio Part*, International Standard, 1994.

[50] ISO/IEC 13818-7, *MPEG-2 advanced audio coding, AAC*, International Standard, 1997.

[51] Bosi, M.; Brandenburg, K.; Quackenbush, S.; Fielder, L.; Akagiri, K.; Fuchs, H.; Dietz, M.; Herre, J.; Davidson, G.; Oikawa, Y. *ISO/IEC MPEG-2 Advanced Audio Coding*, Journal of the AES, Vol. 45, No. 10, pp. 789-814, October 1997.

[52] Brandenburg, K. *MP3 and AAC explained*, Proc. of the AES 17th International Conference on High Quality Audio Coding, Florence, Italy, 1999.

[53] *MPEG: Questions and Answers*, Philips Pamphlet, 199X

[54] ISO/IEC-14496, *Coding of Audiovisual Objects (MPEG-4)*, International Standard, 1997.

[55] Purnhagen, H. *An Overview of MPEG-4 Audio Version 2*, AES 17th International Conference on High-Quality Audio Coding, Firenze, Sep. 1999.

[56] ISO/IEC 15938, *Multimedia Content Description Inter-face (MPEG-7)*, 2001.

[57] Nack, F.; Lindsay, A. *Everything you wanted to know about MPEG-7* (Part 1), IEEE Multimedia, Vol. 6, No. 3, pp.65-77, 1999.

[58] Nack, F.; Lindsay, A. *Everything you wanted to know about MPEG-7* (Part 2), IEEE Multimedia, Vol. 6, No. 4, pp.64-73, 1999.

[59] Quackenbush, S.; Lindsay, A. *Overview of MPEG-7 Audio*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, June 2001.

[60] ISO/IEC TR 21000-1 *Part 1: Vision, Technologies and Strategy,* MPEG, Document: ISO/IEC JTC1/SC29/WG11 N3939, September 2001.

[61] ATSC, *Digital Audio Compression Standard (AC-3)*, December 1995.

[62] Davidson, G.; Fielder, L.; Antill, M. *Low-Complexity Transform Coder for Satellite Link Applications*, 89th Convention of the Audio Engineering Society, preprint 2966, Sept. 1990.

[63] Princen, J.; Bradley, A. *Analysis/synthesis filter band design based on time-domain aliasing cancellation*, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 34, pp. 1153-1161, 1986.

[64] Davis, M.F. *The AC-3 Multichannel Coder*, 95th Convention of the Audio Engineering Society, preprint 3774, October 1993.

[65] Yost, W.A.; Gourevitch, G. *Directional Hearing*, Springer-Verlag, New York, 1987.

[66] Shibazaki, I. *The Signal Compression Technology of Sony's Newest MD Recorder*, Audio Technology Magazine, December 1995.

[67] Yoshida, T. *The Rewritable MiniDisc System*, Proceedings of the IEEE, vol. 82, no. 10, pp. 1492-1500, October 1994.

[68] Tsutsui, K.; Suzuki, H.; Shimoyoshi, O.; Sonohara, M.; Akagiri, K.; Heddle, R.M. *ATRAC: Adaptive Transform Acoustic Coding for MiniDisc*, Reprinted from the 93rd AES Convention, October 1992.

[69] Yoshida, T. *The Rewritable MiniDisc System*, Proceedings of the IEEE, vol. 82, no. 10, pp. 1492-1500, October 1994.

[70] Watanabe, T.; Ohmoto, M.; Abe, M. *Development of ATRAC2 Encoder/Decoder LSI*, IEEE ICCE: 0-7803-3029, March 1996.

[71] Davidson, G.A.; Fielder, L.D.; Link, B.D. *Parametric Bit Allocation in a Perceptual Audio Coder*, presented at 97th AES Convention, November 10-13, 1994.

[72] Soulodre, G.; Grusec, T.; Lavoie, M.; Thibault, L. *Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs*, Journal of the Audio Engineering Society, vol. 46, no. 3, 1998.

[73] ITU-R Recommendation BS-1116-1, *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, 1997.

[74] ITU-T Recommendation P.830, *Subjective performance assessment of telephone-band and wideband digital codecs*, 1996.

[75] ITU-R Recommendation BS-1284, *Methods for the subjective assessment of sound quality - General requirements*, 1997.

[76] ITU-R Recommendation BS-775, *Multi-channel stereophonic sound system with and without accompanying picture*, 1994.

[77] ITU-R Recommendation BS-645, *Test signals and metering to be used on international sound-programme connections*, 1992.

[78] Thiede, T.; Kabot, E. *A New Perceptual Quality Measure for Bit Rate Reduced Audio*, Contribution to the 100th AES Convention, preprint 4280, Copenhagen, 1996.

[79] Brandenburg, K. *Evaluation of Quality for Audio Encoding at Low Bit Rates*, Contribution to the 82nd AES Convention, preprint 2433, London, 1987.

[80] Beerends, J.G.; Stemerdink, J.A. *A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation*, J. Audio Eng. Soc., vol. 40, pp. 963-978, Dec. 1992.

[81] Paillard, B.; Mabilleau, P.; Morisette, S.; Soumagne, J. *Perceval: Perceptual Evaluation of the Quality of Audio Signals*, J. Audio Eng. Soc., vol. 40, pp. 21-31, Jan. 1992.

[82] Colomes, C.; Lever, M.; Rault, J.B.; Dehery, Y.F. *A Perceptual Model Applied to Audio Bit-Rate Reduction,* J. Audio Eng. Soc., vol. 43, pp. 233-240, April 1995.

[83] ITU-R Recommendation BS.1387-1, *Method for Objective Measurements of Perceived Audio Quality*, 1998.

[84] Karjalainen, M. *A New Auditory Model for the Evaluation of Sound Quality of Audio Systems*, IEEE International Conference of Acoustics, pp. 608-611, 1985.

[85] Beerends, J.G.; Stemerdink, J.A. *Modelling a Cognitive Aspect in the Measurement of the Quality of Music Codecs*, Contribution to the 96th Convention of the Audio Engineering Society, Preprint 3800, Amsterdam, February 1994.

[86] Beerends, J.G.; van den Brink, W.A.C. *The Role of Informational Masking and Perceptual Streaming in the Measurement of Music Codec Quality*, Contribution to the 100th Convention of the Audio Engineering Society, Preprint 4176, Copenhagen, May 1996.

[87] Beerends, J.G.; Stemerdink, J.A. *A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation*, J. Audio Eng. Soc., Vol. 42, No. 3, pp. 115-123, March 1994.

[88] ITU-T Recommendation P.861, *Objective quality measurement of telephone-band (300 - 3400) speech codecs*, 1996.

[89] Barbedo, J.G.A. *Avaliação Objetiva de Qualidade de Codecs de Voz na Faixa de Telefonia*, Dissertação de Mestrado, Unicamp, Campinas, julho de 2001.

[90] Barbedo, J.G.A.; Lopes, A. *Proposta e Avaliação de um Método de Medida Objetiva de Qualidade de Codecs de Voz*, Anais do XIX Simpósio Brasileiro de Telecomunicações, SBrT 2001, Fortaleza, artigo n. 001000000002200007, Setembro de 2001.

[91] Barbedo, J.G.A.; Lopes, A. *Proposal and Validation of an Objective Method for Quality Assessment of Speech Codecs and Communication Systems*, Revista Tecnologia, Fortaleza, Vol. 23, No. 1, pp. 96-112, dezembro de 2002.

[92] Sporer, T. *Objective Audio Signal Evaluation-Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio*, Contribution to the 103rd Convention of the Audio Engineering Society, New York, Preprint 4512, September 1997.

[93] Oppenheim, A.V.; Schafer, R.W. *Discrete Time Signal Processing*, Prentice Hall, New Jersey, 1989.

[94] Barbedo, J.G.A.; Lopes, A. *On the Vectorization of Decimation Filterbanks*, to be submitted to a Journal.

[95] Beerends, J.G.; Stemerdink, J.A. *Modelling a Cognitive Aspect in the Measurement of the Quality of Music Codecs*, 96th Convention of the Audio Engineering Society (AES), Preprint 3800, Amsterdam, February 1994.

[96] Kidd, G.; Mason, C.R.; Deliwala, P.S. *Reducing Informational Masking by Sound Segregation*, J. Acoust. Soc. Am., vol. 95, pp. 3475-3480, 1994.

[97] Fletcher, R. *Practical Methods of Optimization, 2nd Ed.*, John Wiley & Sons, New York, 1987.

[98] Bazaraa, M.S.; Sherali, H.D.; Shetty, C.M. *Nonlinear programming*, John Wiley & Sons, New York, 1993.

[99] Barbedo, J.G.A.; Lopes, A. *Innovations on the Objective Assessment of Audio Quality*, Anais da VII Convenção Nacional da AES, São Paulo, Maio de 2003.

[100] Barbedo, J.G.A.; Lopes, A. *A New Method for Objective Assessment of Áudio Quality*, Anais do XX Simpósio Brasileiro de Telecomunicações, Rio de Janeiro, Outubro de 2003.

[101] *ISO/IEC/JTC 1/SC 2/WG11 MPEG/Audio test report*, Document MPEG90/N0030, October 1990.

[102] *ISO/IEC/JTC 1/SC 2/WG 11 MPEG/Audio test report*, Document MPEG91/N0010, June 1991.

[103] ITU-R Task Group 10/4, *Report on the Sixth Meeting of ITU-R Task Group 10/4*, Doc. 10-4/21, Geneva, 1998.

[104] Haykin, S. *Neural Networks – A Comprehensive Foundation*, Prentice Hall, New Jersey, 1999.

[105] Kohonen, T. *Self-Organizing Maps*, 2nd edition, Springer, 1997.

[106] Rix, A., Hollier, M., *Performance Metrics for Objective Quality Assessment Systems in Telephony*, ITU Study Group 12 - Delayed Contribution D.79, November 1998.

[107] Barbedo, J.G.A.; Lopes, A. *Strategies to Increase the Applicability of Methods for Objective Assessment of Audio Quality*, 116th AES Convention, preprint 6080, Berlin, May 2004.

[108] Barbedo, J.G.A.; Lopes, A. *Uma Nova Estratégia para a Estimação Objetiva da Qualidade de Sinais de Áudio*, submetido à revista IEEE Latino.

[109] Barbedo, J.G.A.; Lopes, A. *A New Cognitive Model for Objective Assessment of Audio Quality*, submitted to the Journal of AES.

[110] Lopes, A.; Barbedo, J.G.A. *Medida Objetiva de Avaliação de Áudio*, pedido de patente submetido à Avaliação da Fapesp.

[111] Rix, A.W.; Hollier, M.P. *The Perceptual Analysis Measurement System for Robust End-to-End Speech Quality Assessment*, Proceedings of the IEEE ICASSP, pp. 1515-1518, vol. 3, June 2000.

[112] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[113] Barbedo, J.G.A.; Lopes, A.; Simões, F.O.; Runstein, F. *Objective Measure of Speech Quality in Channels with Variable Delay*, Revista Telecomunicações, vol. 6, n. 2, pp.19 - 24, December 2003.

[114] Ribeiro, M.V. *Técnicas de Reconstrução de Pacotes Baseadas em Transformada Wavelet e Redes Neurais Aplicadas a Codificadores de Forma de Onda em Telefonia IP*, Dissertação de Mestrado, Unicamp, Campinas, outubro de 2001.

[115] Rix, A.W.; Hollier, M.P.; Hekstra, A.P.; Beerends, J.G. *Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality*

Assessment: Part I – Time Alignment, Journal of the AES, Vol. 50, No. 10, pp. 755-764, October 2002.

[116] Barbedo, J.G.A.; Ribeiro, M.V.; Von Zuben, F.J.; Lopes, A.; Romano, J.M.T. *Application of Kohonen Self-Organizing Maps to Improve the Performance of Objective Methods for Speech Quality Assessment*, Proceedings of the XI European Signal Processing Conference (EUSIPCO2002), Vol. I, pp. 519-522, Toulouse, France, September 2002.

[117] Barbedo, J.G.A.; Ribeiro, M.V.; Lopes, A.; Romano, J.M.T. *Estimation of the Subjective Quality of Speech Signals using the Kohonen Self-Organizing Maps*, Proceedings of the IV International Telecommunication Symposium (ITS), Natal, Brazil, pp. 834-839, September 2002.

[118] Malvar, H.S. *Signal Processing with Lapped Transforms*, Norwood, MA: Artech House, 1992.

[119] Series P Supplement 23, *ITU-T Coded Speech Database*, Telecommunication Standardization Sector of International Telecommunication Union, February 1998.

[120] Santos, E.P.; Zuben, F.J.V. *Efficient Second-Order Learning Algorithm for Discrete-Time Recurrent Neural Networks*, In Recurrent Neural Networks: Design and Applications, CRC Press, Boca Raton, pp. 47-75, 2000.

[121] Ribeiro, M.V.; Barbedo, J.G.A.; Romano, J.M.T.; Lopes, A. *Fourier-Lapped-Multilayer Perceptron (FLMLP) Method for Speech Quality Assessment*, submitted to the Special Issue on Anthropomorphic Processing of Audio and Speech.

[122] Lopes, A.; Romano, J.M.T.; Ribeiro, M.V.; Barbedo, J.G.A.; Lima, C. *Método FL-PMC (Fourier Lapped - Perceptron Multicamadas) para a Estimação de Qualidade de Voz*, patente: privilégio de inovação n. 0204932-5, Método FL-PMC, depósito 06 de novembro de 2002.

[123] Beerends, J.G.; Hekstra, A.P.; Rix, A.W.; Hollier, M.P. *Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment: Part II - Psychoacoustic Model*, Journal of the AES, Vol. 50, No. 10, pp. 765-778, October 2002.

[124] Beerends, J.G.; Stemerdink, J.A. *The Optimal Time-Frequency Smearing and Amplitude Compression in Measuring the Quality of Audio Devices*, Proceedings of the 94th AES Convention, J. Audio Eng. Soc. (Abstracts), vol. 41, p. 409, May 1993, preprint 3604.

[125] Barbedo, J.G.A.; Lopes, A. *A New Method for Objective Assessment of Speech Quality*, submetido à Revista da Sociedade Brasileira de Telecomunicações.

[126] ITU-T Recommendation P.48, *Specification for an Intermediate Reference System*, 1989.

APÊNDICE

A.1. TERÇAS-OITAVAS

A divisão do espectro de freqüências audível em oitavas se faz de maneira que a freqüência central de uma faixa corresponda ao dobro da freqüência central da faixa anterior [80]. A expressão geral é dada então por:

$$f_c(n) = c.2^n \ , \ \text{n} = 0, 1, 2, 3 \ ...$$

onde f_c(n) representa a n-ésima freqüência central e c é a primeira freqüência central desejada.

Então, como se pode observar, se a freqüência central de uma banda tem como valor $c \cdot 2^n$, as freqüências centrais anterior e posterior terão os valores $c \cdot 2^{n-1}$ e $c \cdot 2^{n+1}$, respectivamente. Da mesma forma, os limites inferior e superior para a faixa definida por essa freqüência central serão, respectivamente:

$$lib = c \cdot 2^{n-1/2} = 0{,}7071 \cdot fc$$
$$lsb = c \cdot 2^{n+1/2} = 1{,}4142 \cdot fc \ ,$$

onde *lib* é o limite inferior e *lsb* é o limite superior.

Então, por exemplo, para 1 kHz tem-se 500 Hz para uma oitava abaixo e 2 kHz para uma oitava acima, e tem-se 707 Hz para o limite inferior e 1.414 Hz para o limite superior.

A divisão em terços de oitavas segue o mesmo princípio, sendo que a expressão geral, para este caso, é dada por

$$fc(n) = c \cdot 10^{\frac{n}{10}}.$$

Da mesma maneira, se a freqüência central de uma banda tem como valor $c \cdot 10^{n/10}$, as freqüências centrais anterior e posterior terão os valores $c \cdot 10^{(n-1)/10}$ e $c \cdot 10^{(n+1)/10}$, respectivamente. Então, os limites inferior e superior para a faixa definida por essa freqüência central serão, respectivamente:

$$lib = c \cdot 10^{\frac{[n-(1/2)]}{10}} = 0{,}8913 \cdot fc \ ,$$

$$lib = c \cdot 10^{\frac{[n+(1/2)]}{10}} = 1{,}1220 \cdot fc \ .$$

É interessante observar ainda que $10^{\frac{n}{10}} \cong 2^{\frac{n}{3}}$. Conseqüentemente, a expressão para as terças-oitavas pode ser reescrita como

$$fc(n) \cong c \cdot 2^{\frac{n}{3}},$$

ou seja, é a mesma expressão encontrada para as oitavas, porém com o expoente dividido por três, daí a adoção do nome terças-oitavas.

A.2. SPL

Definição de SPL – Sound Pressure Level (Nível de Pressão Sonora) [81]: tem como unidade o decibel SPL (dB_{SPL}), e é dado pela expressão $20\log_{10}\left(P/P_{ref}\right)$, onde P é a pressão sonora do sinal que se está medindo e P_{ref} é a pressão sonora de referência, a qual pode assumir dois valores:

a) $P_{ref} = 0{,}0002\mu B \quad \left(2\times10^{5}\, N/m^{2}\right)$

b) $P_{ref} = 0{,}1\mu B \quad \left(0{,}1 N/m^{2}\right)$

onde μB é a pressão em microbars.

É importante observar que a pressão sonora de referência dada no item (a) é mais utilizada nas medições relacionadas com a audição e nas medições de nível sonoro no ar e nos líquidos, enquanto que aquela dada no item (b) tem maior aplicação na calibração de transdutores e certos tipos de medição de nível sonoro em líquidos. Os dois níveis de medição diferem um do outro em aproximadamente 54 dB. Por essa razão, é necessário indicar explicitamente o nível de referência adotado (neste trabalho, usou-se o primeiro).

A.3. dB_{FS} (Full Scale)

Esta escala é relacionada à máxima amplitude admissível para um sinal antes que haja saturação, o que, em um sistema digital, significa o maior valor absoluto que o sinal pode assumir. Assim, 0 dB_{FS}, no caso de 16 bits, representa o valor de 32.767.