8

PERCEPÇÃO AUDITIVA E COMPRESSÃO DE ÁUDIO

PTC2547 - PRINCÍPIOS DE TELEVISÃO DIGITAL

Guido Stolfi – EPUSP – 10 / 2016

1. Transmissão de Áudio na TV Analógica e Digital

Na televisão analógica, a informação de áudio é transmitida por uma portadora independente, modulada em FM. Para esta portadora estão reservadas cerca de 1/12 da banda e 1/30 da potência total do canal (fig. 1.1). Com este processo, a resposta em frequência do canal de áudio é limitada em cerca de 13 kHz, e a relação Sinal/Ruído não passa de 60 dB. A transmissão de som estereofônico (2 canais) é feita por um processo de multiplexação, descrito no item 1.1 a seguir.

Já no caso da televisão digital, a simples amostragem e quantização do sinal de áudio resultará em uma taxa de bits que não pode ser considerada desprezível em relação à taxa do vídeo comprimido. Por exemplo: para um áudio "qualidade CD", os requisitos são:

- 2 canais de áudio, com resposta de 20 kHz
- Frequência de Amostragem de 44,1 kHz
- Quantização com 16 bits (relação sinal/ruído da ordem de 100 dB)

Isso resulta em 2 x 16 x 44,1 = 1,41 Mbits por segundo. Lembramos que uma imagem de TV convencional ("*standard definition*") pode ser comprimida satisfatoriamente para uma taxa de 2 Mbits/s, pouco maior do que a taxa bruta do áudio.

Além disso, para a TV de alta definição, é razoável supor que o áudio deve ser transportado com um padrão de maior qualidade. Para o sistema "5.1 canais", utilizado em salas de cinema, a situação envolve:

- 5 canais principais (Direito, esquerdo, central, traseiro esquerdo e traseiro direito)
- 1 canal de baixa frequência para sub-graves
- Frequência de amostragem de 48 kHz para os canais principais
- Quantização nativa com 24 bits

A taxa total será 5,1 x 48 x 24 = 5,88 Mb/s, que representa mais que ¼ da capacidade de um canal de transmissão adequado para TV digital.

Percebemos então a necessidade de desenvolver um processo de compressão para o áudio. Da mesma forma que procedemos para a imagem, devemos então estudar as características da audição humana, procurando por situações em que seja possível remover informações redundantes de forma imperceptível.

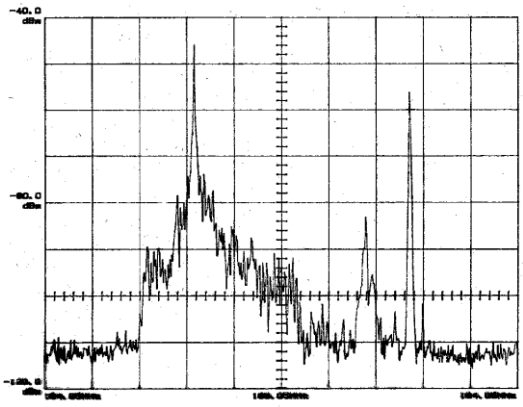

Fig. 1.1- Canal de TV Analógica (portadora de áudio à direita)

1.1 Sistema BTSC

A transmissão de som estereofônico na TV analógica é feita por um processo semelhante ao de radiodifusão de FM estéreo (Sistema BTSC, *Broadcast Television Systems Commitee,* figura 1.2).

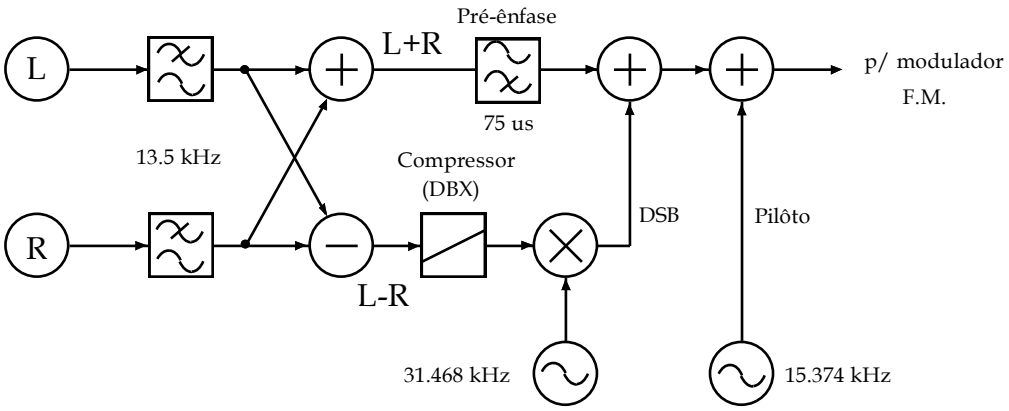

Fig. 1.2 - Áudio estereofônico em TV analógica - Sistema BTSC

Neste sistema, os canais L e R, limitados em banda até cerca de 13 kHz, são somados, sofrem uma pré-ênfase em altas frequências e compõem o sinal de banda base (L+R), que será modulado em FM sobre a portadora de áudio, com desvio máximo de ±25 kHz. Este sinal é compatível com a transmissão monofônica. Já a diferença (L-R) é multiplicada por uma sub-portadora, cuja frequência é o dobro da frequência de varredura horizontal da TV ($2f_h$), gerando um sinal em DSB-SC (banda lateral dupla com portadora suprimida). Antes disso, o sinal (L-R) é comprimido por um sistema denominado DBX (equivalente a um controle automático de ganho com compressão de 2:1), de modo a restringir a faixa dinâmica em 50%, garantindo assim uma melhor relação Sinal/Ruído deste sinal. Para facilitar a demodulação e indicar ao receptor a existência de um sinal estereofônico, é inserido um piloto na frequência de 15.374 kHz (f_H). Lembramos que na radiodifusão FM estéreo, não é usado o compressor, e as frequências da sub-portadora e do piloto são respectivamente 38 e 19 kHz.

O sistema BTSC admite ainda um canal secundário, ou SAP - *Secondary Audio Program* - transportado por uma sub-portadora em $5f_h$, modulada em frequência com desvio de ±10 kHz. O espectro resultante do sinal BTSC está esquematizado na figura 1.3.

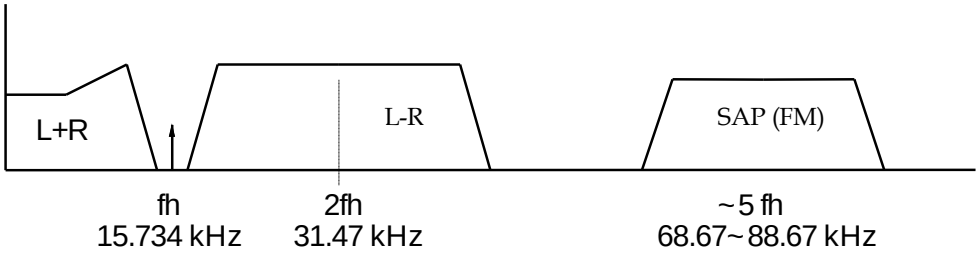

Fig. 1.3 - Espectro de banda base do sinal de áudio estereofônico BTSC

2. Percepção Auditiva: o Ouvido Humano

A figura 2.1 mostra um corte do ouvido humano, cuja estrutura pode ser classificada em 3 partes:

O **Ouvido Externo** é composto pela orelha e pelo canal auditivo, e possui funções de filtragem que aumentam a sensibilidade e auxiliam na localização espacial da fonte sonora. O canal auditivo comporta-se como um tubo sintonizado, com picos de resposta em torno de 4 kHz e 12 kHz, correspondentes às ressonâncias de 1/4 de onda e 3/4 de onda. Já a orelha proporciona uma resposta em frequência diferente, dependendo do ângulo de incidência vertical (elevação) do som; a "assinatura espectral" proporcionada sobre o som recebido permite ao cérebro identificar a origem do som no sentido vertical.

- No **Ouvido Médio** é feita uma adaptação de impedância acústica (ou seja, da relação entre pressão e deslocamento) entre o ar e o meio aquoso que preenche o Ouvido Interno. O tímpano consiste de uma membrana elástica, com área aproximada de 60 a 80 mm2. Os ossículos intermediários (martelo, bigorna e estribo) formam uma alavanca que acopla as vibrações do tímpano para a janela oval, cuja área é de 2 mm2. A diferença das áreas, juntamente com o fator de alavanca dos ossos transmissores, proporciona uma variação de impedância da ordem de 4000 vezes, bastante próxima da relação entre as impedâncias acústicas do ar e da água. Aos ossículos intermediários estão acoplados músculos tensores, que podem atenuar a transmissão de vibrações (especialmente em baixas frequências, onde a atenuação pode atingir 30 dB). Estes músculos proporcionam uma compressão da faixa dinâmica da percepção sonora, além de agirem como proteção reflexa contra sons de alta intensidade.
- O **Ouvido Interno** compõe-se basicamente da cóclea, que consiste de uma cavidade cônica, enrolada na forma de um caracol, preenchida por um meio aquoso (fig. 2.2). Distendida, a cóclea teria comprimento aproximado de 2 a 3 cm. Apresenta-se dividida longitudinalmente em duas câmaras através da **membrana basilar**, sobre a qual estão distribuídos os **Órgãos de Corti**, que são, efetivamente, sensores de vibração ligados a terminações nervosas. A figura 2.3 mostra um corte transversal da cóclea; na fig. 2.4 são apresentados em detalhe os órgãos de Corti.

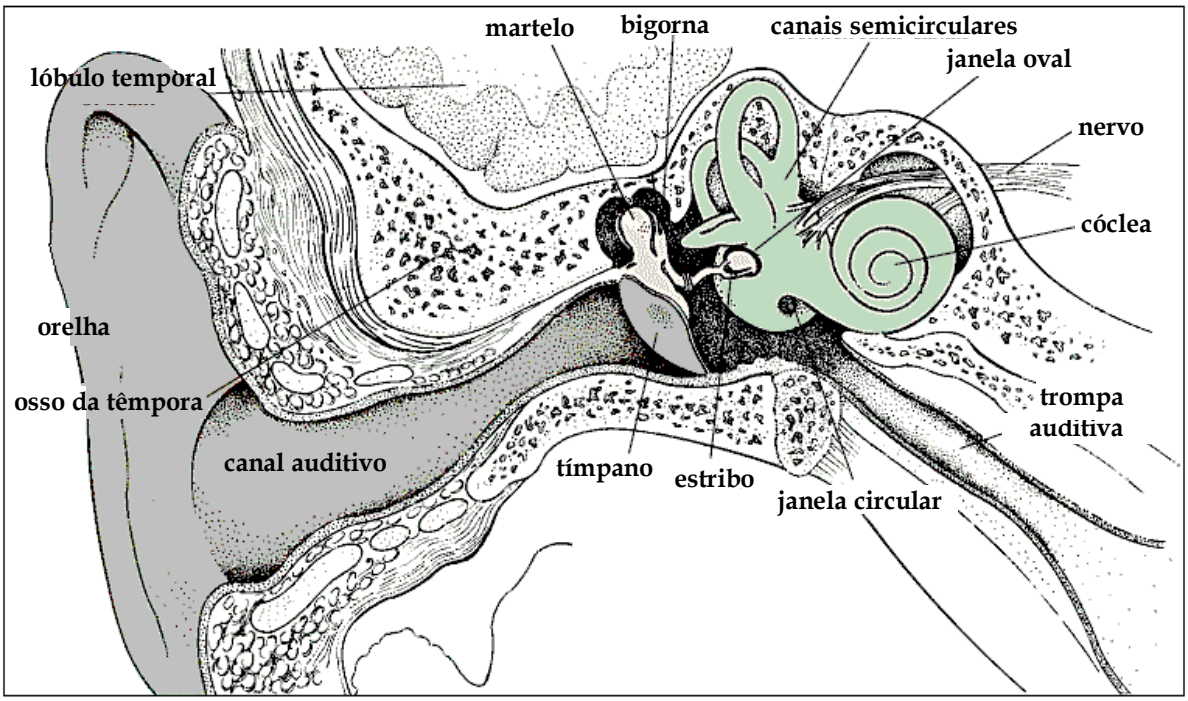

Fig. 2.1 - Estrutura do Ouvido Humano

A figura 2.5 esquematiza a estrutura do ouvido. As vibrações sonoras incidentes no tímpano são acopladas à cóclea pela janela oval, e propagam-se pelo meio aquoso. A janela circular atua como uma terminação, permitindo que o líquido no interior da cóclea vibre livremente.

Devido ao formato cônico da cavidade, e à variação de espessura, largura e elasticidade da membrana basilar, a cóclea apresenta pontos de ressonância distribuídos ao longo do seu comprimento; assim sendo, sons com frequências da ordem de 15 kHz provocam vibração da membrana basilar apenas na região próxima às janelas, enquanto sons de menos de 100 Hz ressoam na extremidade oposta.

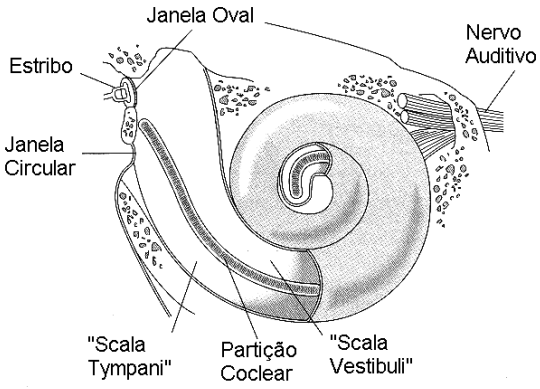

Fig. 2.2 – Estrutura da Cóclea

A cóclea opera então como um analisador de espectro, ou um banco de filtros seletivos distribuídos. O ouvido normal é assim capaz de distinguir sons com frequências entre 20 Hz e 20 kHz, embora a resposta para altas frequências diminua acentuadamente com a idade. Cerca de 50% das pessoas não ouvem acima de 15 kHz.

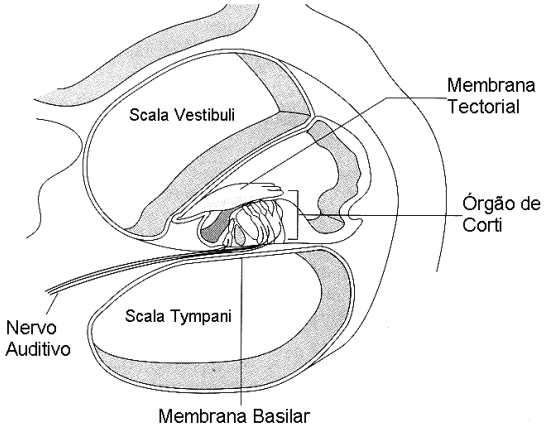

Fig. 2.3 - Seção da Cóclea

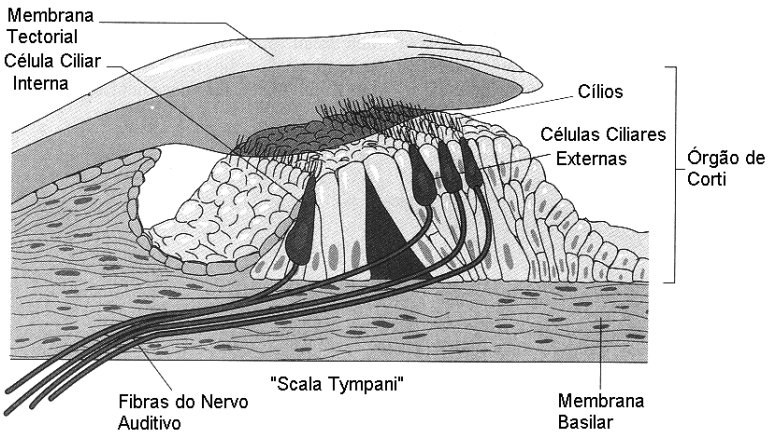

Fig. 2.4 – Órgão de Corti

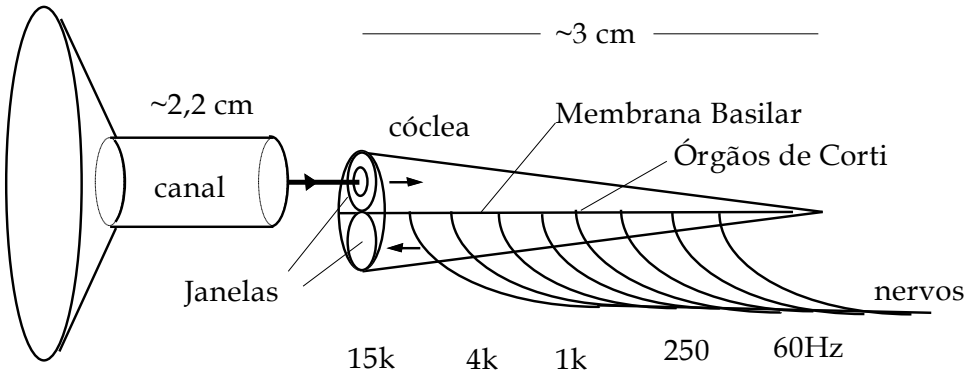

Fig. 2.5 - Representação Esquemática do Ouvido Humano

2.1 Órgãos de Corti

Os Órgãos de Corti contêm células transdutoras de vibração (células ciliares). Estas células estão assentadas na membrana basilar e possuem prolongamentos (cílios), que se apóiam na Membrana Tectorial. A vibração desses cílios, provocada pelo deslocamento relativo entre as membranas basilar e tectorial (fig. 2.6), provoca por sua vez a entrada, na célula, de íons presentes na solução aquosa externa. Ocorre então uma alteração do potencial elétrico na célula, que provoca disparos nas terminações nervosas. A diferença de potencial entre o interior da célula de Corti e o meio aquoso externo é da ordem de 150mV, em repouso. Um deslocamento dos cílios da ordem de 0,1 nanômetro pode ser suficiente para disparar um impulso nervoso.

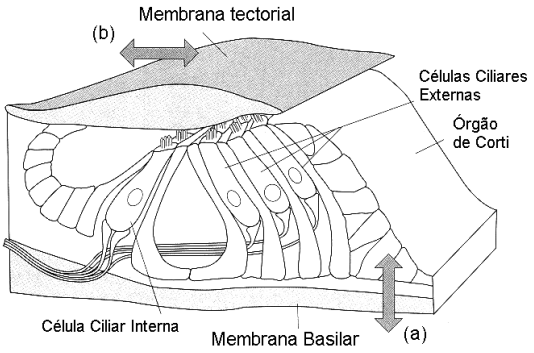

Fig. 2.6 – Transdução de Movimento nas Células Ciliares

A figura 2.7 mostra a disposição das células de Corti, após a retirada da membrana tectorial. As células Internas, em número de 3500 aproximadamente, são os sensores propriamente ditos. Cada uma das células liga-se, em média, a cerca de 8 fibras nervosas aferentes (que levam estímulos nervosos ao córtex cerebral). Apenas 10% das cerca de 30.000 fibras aferentes provêm das células Externas; estas, no entanto, recebem a quase totalidade das fibras eferentes (que recebem sinais provenientes do cérebro). Aparentemente, as células Externas são transdutores ou atuadores que modulam, de alguma forma ainda não totalmente conhecida, a resposta e a sensibilidade das células Internas. Danos às células externas estão associados à perda de seletividade e de faixa dinâmica na audição.

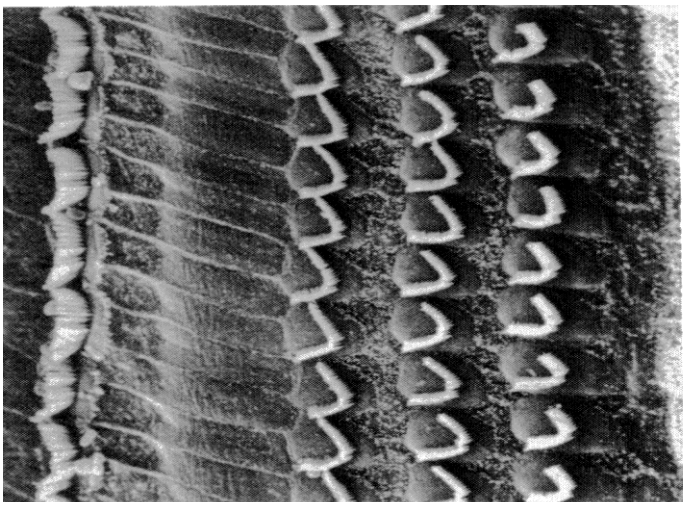

Fig. 2.7 - Células de Corti Internas (esq.) e Externas (3 fileiras à dir.)

2.2 Seletividade e Bandas Críticas

A seletividade, ou seja, a resposta em frequência de cada célula sensorial e de suas respectivas terminações nervosas, varia com a posição da mesma em relação à cóclea. Na figura 2.8 (a) temos as respostas em frequência devidas à ressonância mecânica da membrana basilar. As células ciliares externas influem na seletividade total da cóclea, através de um processo de realimentação ativa, aumentando a seletividade da resposta neural. Na figura 2.8 (b) temos a resposta em frequência medida em fibras nervosas avulsas do nervo auditivo.

A seletividade aparente do nosso ouvido é da ordem de 100 Hz para frequências até 500 Hz, e entre 1/6 e 1/3 de oitava para frequências superiores. Essas faixas de seletividade são denominadas de Bandas Críticas, e estão relacionadas indiretamente com a capacidade do ouvido de discernir sons de amplitudes e frequências diferentes.

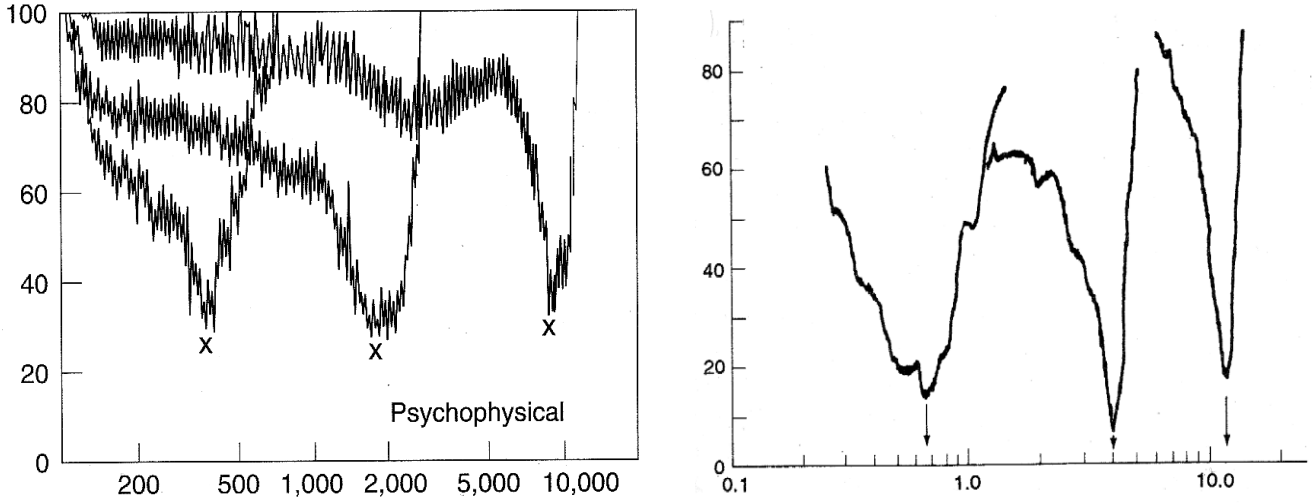

Fig. 2.8 - Seletividade Típica da Membrana Basilar (a) e das Células Ciliares (b)

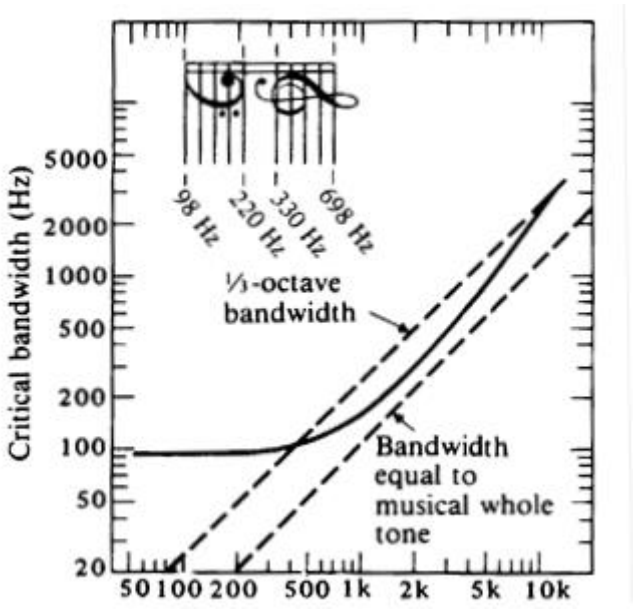

Fig. 2.9 - Larguras das Bandas Críticas do Ouvido Humano, Comparadas com a Escala Musical

A Fig. 2.9 mostra as larguras das bandas críticas para o ouvido humano em função da frequência. Nota-se que a seletividade é praticamente constante (correspondendo a um atraso de grupo constante) dentro das frequências fundamentais da maioria dos instrumentos musicais.

A compreensão do significado das bandas críticas pode ser facilitada analisando um método utilizado para determinação experimental das suas larguras. Este método é esquematizado na figura 2.10, onde se utiliza um estímulo sonoro na forma de um ruído de banda estreita, com frequência central e potência constantes, porém com largura de banda variável. À medida que a largura de banda é aumentada, a sensação subjetiva de intensidade sonora permanece constante até que seja excedida a largura da banda crítica.

A partir desse ponto, ocorre uma sensação de aumento de volume sonoro, apesar da potência permanecer constante. A ideia é que a banda crítica define uma vizinhança espectral, dentro da qual as informações são combinadas em uma única sensação de volume. Sons fora dessa banda contribuem de forma independente, resultando em uma sensação maior de volume.

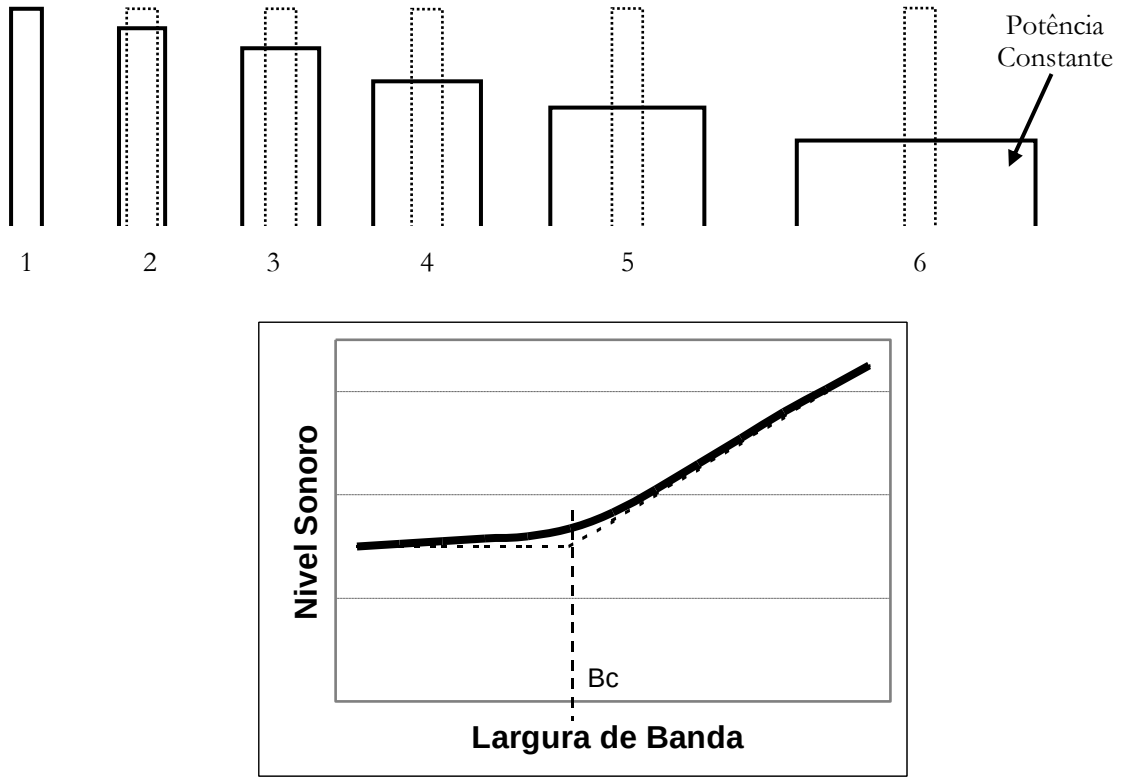

Fig. 2.10 – Determinação das Bandas Críticas

3. Sensibilidade da Audição

O ouvido apresenta resposta perceptual aproximadamente logarítmica, tanto na distinção de frequências como na de intensidade sonora. O limiar inferior de audibilidade para um ouvido normal corresponde a uma amplitude de pressão sonora de 20 μPa r.m.s. (2×10-5 N/m2) na frequência de 1 kHz (aproximadamente 1 pW/m2). Já o limiar de desconforto é atingido com pressões sonoras 1.000.000 de vezes maiores (Tabela 3.1).

É conveniente então medir o nível de pressão sonora em escala logarítmica, sendo então definida a unidade "**dB SPL**" ("*Sound Pressure Level*") como

$$\text{Pressão Sonora} = 20\log\frac{p}{p_o} \quad (\text{dbSPL}) \text{ onde } p_o = 20\mu\text{Pa r.m.s.}$$

	dB SPL		dB SPL
Foguete	195	Conversação	65
Avião a jato	155	Restaurante	60
Limiar de dor	140	Residência urbana	40
Limiar de desconforto	120 = 1 W/m^2	Casa de campo	30
Orquestra fortíssimo	110	Orquestra pianíssimo	30
Rebitadeira	100	Estúdio de gravação	20
Fábrica	78	Folhagens na brisa	10
Tráfego pesado	68	Limiar de audição	0 = 1pW/m^2
Escritório ruidoso	65	Ruído térmico do ar	–10 dB

Tabela 3.1 - Exemplos de Situações Sonoras

Por outro lado, a resposta em frequência do ouvido não é plana, nem linear (varia com a intensidade sonora). Define-se então Volume Sonoro ("Loudness") como representando a

grandeza **subjetiva** associada à percepção sensorial da intensidade de um som. O Volume Sonoro é expresso em Phons, sendo que o nível em Phons é numericamente igual à pressão sonora em dB SPL na frequência de 1 kHz.

A resposta em frequência do ouvido, para vários níveis de Volume Sonoro, foi estudada por Fletcher e Munson em 1933, e posteriormente por Robinson e Dadson, sendo as curvas médias da figura 3.1 adotadas como recomendação pelo ISO.

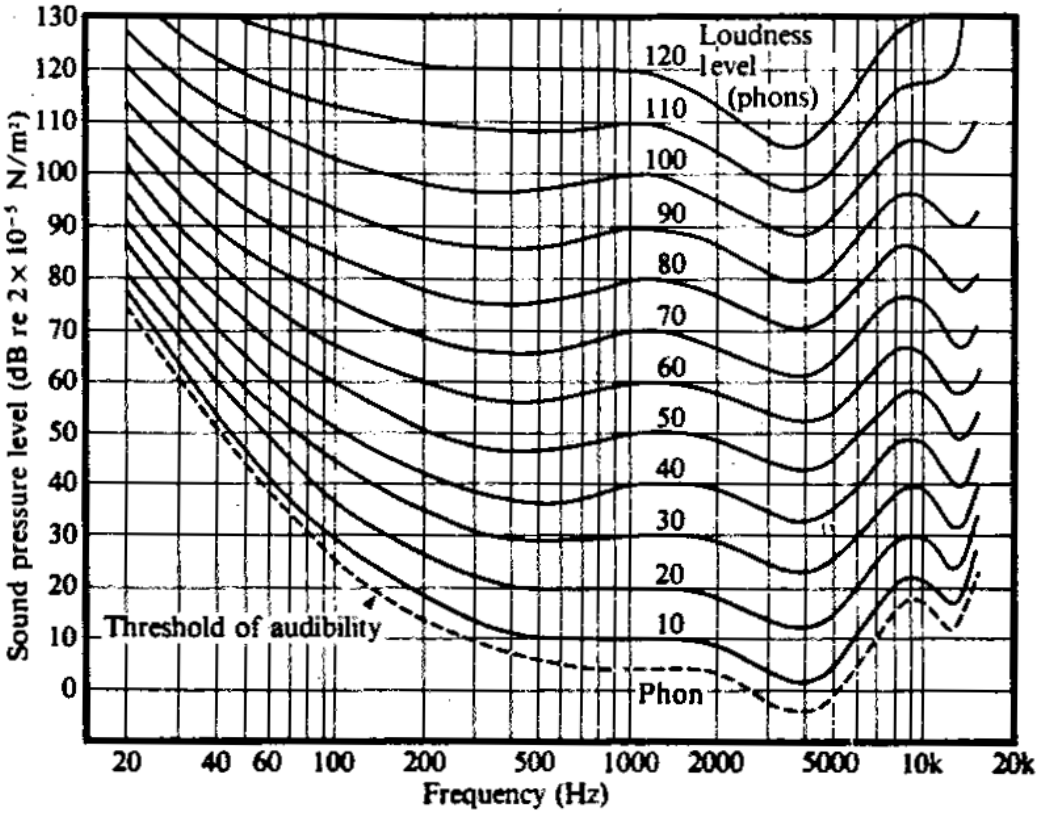

Fig. 3.1 - Curvas de Igual Volume Sonoro para Tons Senoidais

(Rossing - 1982)

Pelas curvas, nota-se que a sensibilidade é bastante reduzida para baixas frequências com baixas amplitudes, rejeitando assim sons produzidos pelo próprio corpo humano e, em particular, pela circulação sanguínea no ouvido interno. A sensibilidade atinge um pico em 4 kHz (frequência de ressonância do canal auditivo), decrescendo novamente até um segundo pico em 12 kHz (também devido ao canal auditivo). Acima de 15 kHz a resposta é variável de indivíduo para indivíduo, e bastante atenuada.

As curvas da fig. 3.1 parecem indicar que um sistema de reprodução sonora deveria responder de 20 a 20.000 Hz (o que é verdade), e proporcionar uma faixa dinâmica ou relação sinal/ruído maior que 120 dB. No entanto, este requisito não corresponde à realidade por duas razões:

- O nível de ruído ambiente em uma residência é de aproximadamente 40 dB SPL; mesmo em um estúdio de gravação é de pelo menos 20 dB SPL (Tabela 3.1).
- O nível de 120 dB SPL corresponde ao limiar de desconforto; experimentalmente determinou-se que as regiões confortáveis para audição de música ou voz abrangem faixas dinâmicas respectivamente de 70 dB e 30 dB (Fig. 3.2).

Considerando um nível de ruído de fundo de 10 dB SPL (inaudível mesmo em condições de estúdio), percebe-se que uma relação S/R de 90 dB é satisfatória para um sistema de

transmissão de áudio de alta qualidade. No caso do CD, por exemplo, a quantização com 16 bits proporciona uma relação S/R da ordem de 100 dB.

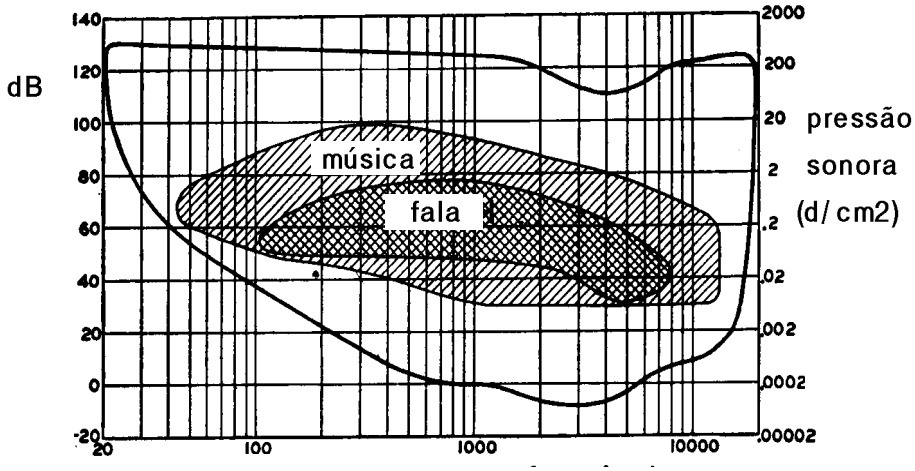

Fig. 3.2 - Faixa dinâmica para Audição de Música e Voz

3.1 Percepção de Direção

Dois mecanismos básicos atuam na percepção da direção de uma fonte sonora. Para frequências baixas, até cerca de 1 kHz, a diferença de fase entre as frentes de onda recebidas nos dois ouvidos fornece indicação de direção (fig. 3.3-a). Para frequências altas, é a diferença de amplitude nos dois ouvidos, provocada pela sombra acústica projetada pela cabeça, que fornece indicação de direção (fig. 3.3-b). Ambiguidades entre sons frontais e posteriores são resolvidos por movimentos da cabeça; portanto, a localização espacial do som é prejudicada quando são usados fones de ouvido.

A percepção de direção no sentido vertical é em parte proporcionada pelo pavilhão auricular, cuja estrutura imprime uma "assinatura espectral" ao som recebido, conforme sua elevação.

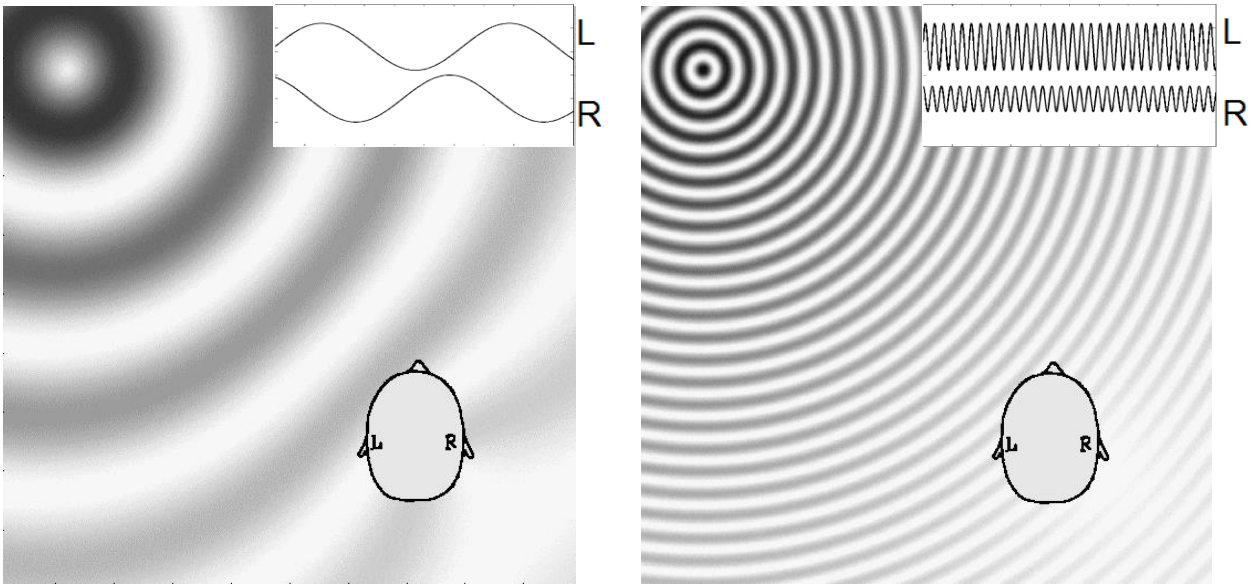

Fig. 3.3 – (a) diferenças de fase **(b) diferenças de intensidade**

3.2 Percepção de Ambiente

Reflexões múltiplas nas paredes, piso e outros elementos de um ambiente trazem ao ouvido pistas que permitem identificar algumas características desse ambiente. Por exemplo, o atraso temporal entre o som direto e a primeira reflexão subsequente (fig. 3.4) fornece à audição uma idéia do tamanho do ambiente. O tempo de decaimento, a relação entre o decaimento em altas e baixas frequências, etc., trazem informações sobre o tipo de ambiente e dos materiais e revestimentos envolvidos.

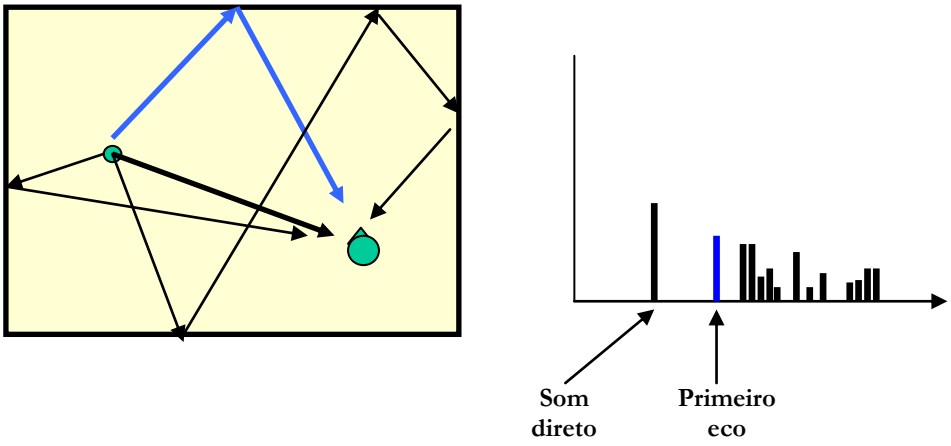

Fig. 3.4 – Percepção de Ambiente

4. Mascaramento

A faixa dinâmica de 120 dB do ouvido humano representa a relação entre os níveis de audibilidade e de desconforto para sons individuais. No entanto, quando dois sons de intensidades e frequências diferentes são combinados simultaneamente, ocorre o fenômeno de mascaramento, que pode ser descrito como um deslocamento relativo do limiar de Audibilidade, provocado pela presença de tons de maior intensidade.

Vários experimentos foram feitos para determinar e modelar as características do mascaramento; algumas conclusões obtidas são as seguintes:

- Tons senoidais com frequências próximas são mascarados mais do que tons em frequências distantes entre si;
- Um tom senoidal mascara mais fortemente outros tons de frequências mais altas do que tons de frequências inferiores;
- Quanto maior a amplitude do tom, maior a banda de frequências mascaradas por ele;
- Ruído de banda estreita apresenta características de mascaramento similares às dos tons senoidais;
- Ruído branco mascara tons senoidais de todas as frequências igualmente e de forma linear;
- O mascaramento provocado por um tom mais intenso permanece por um tempo de 20 a 30 ms após a interrupção do tom (mascaramento temporal);
- Um tom pode ser mascarado por um outro som de intensidade maior emitido até 10 ms após o primeiro (mascaramento temporal retroativo).

A figura 4.1 mostra o limiar de mascaramento para um tom senoidal de 1200 Hz a 80 Phons. Nota-se a presença de "picos" no nível de mascaramento nas frequências próximas a 2400 e 3600 Hz; estes picos sugerem que o ouvido humano gera internamente distorções harmônicas que interferem com os sons captados.

Dentro de uma faixa comparável à banda crítica, o mascaramento atinge níveis até 20 dB abaixo do sinal interferente. Em frequências muito próximas à do sinal interferente (e de suas harmônicas), a sensibilidade melhora um pouco devido à percepção de batimentos, na forma de variações de amplitude do sinal percebido.

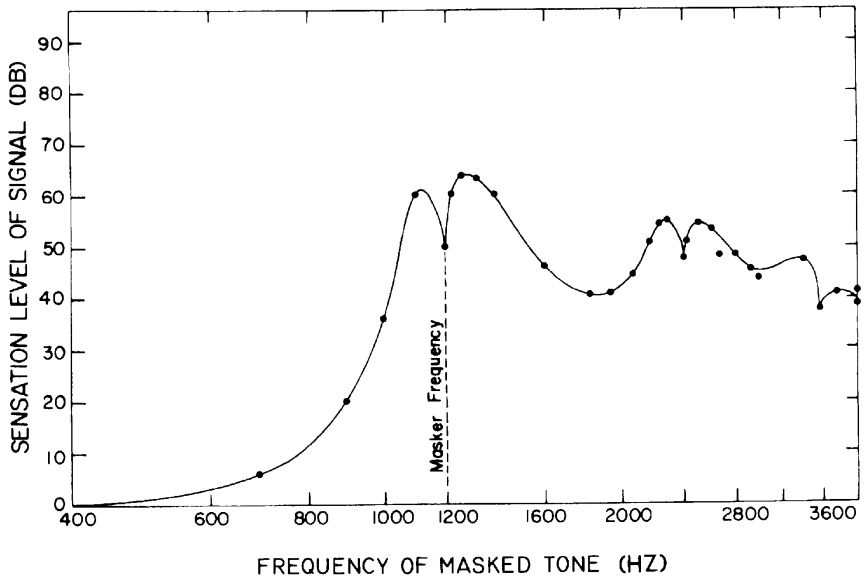

Fig. 4.1 - Mascaramento para Tom Senoidal

A assimetria do mascaramento em função da frequência pode ser explicada pelo padrão de vibração da membrana basilar (figura 4.2); devido à sua assimetria, tons de baixa frequência (f1) mascaram mais fortemente os de alta frequência (f2) do que o inverso. No caso (b), o tom f1 de maior amplitude mascara completamente o tom f2; já no caso (c), um som de frequência f2, mesmo de maior amplitude, não consegue encobrir o som f1.

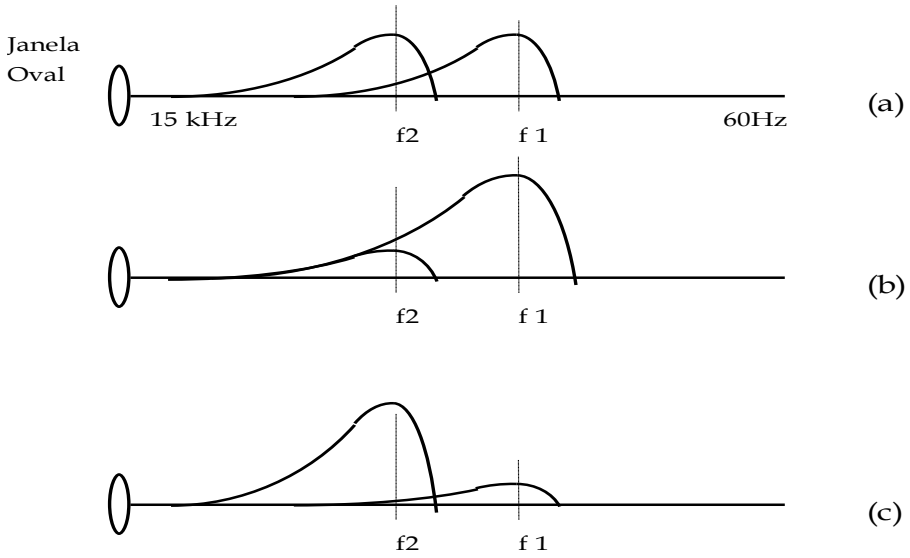

Fig. 4.2 - Resposta da Membrana Basilar para 2 Tons Senoidais

Para sinais de mascaramento constituídos de ruído de banda estreita (Fig. 4.3), os efeitos de batimento desaparecem, mantendo-se semelhantes, no entanto, os efeitos da distorção no ouvido interno (para níveis mais elevados) e da assimetria em frequência do mascaramento.

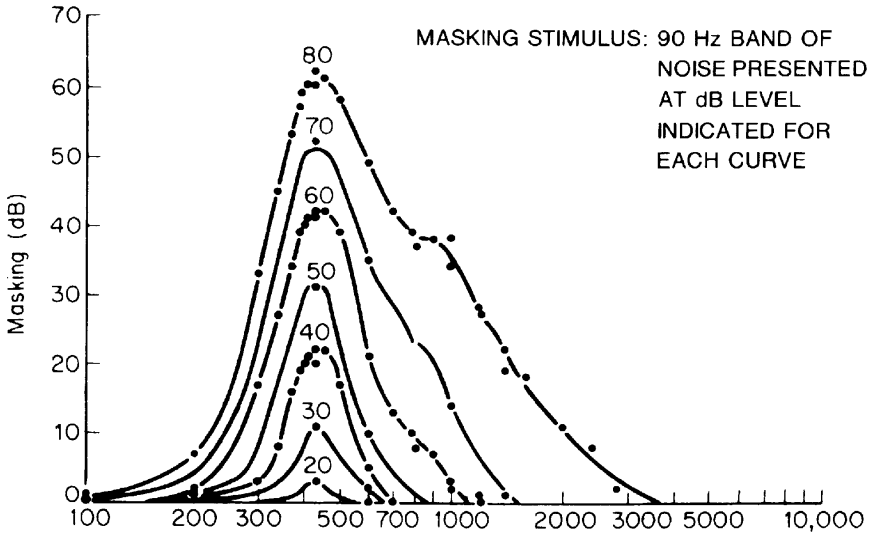

Fig. 4.3 - Mascaramento para Ruído de Banda Estreita

A figura 4.4 mostra famílias de curvas de mascaramento para tons senoidais, ruído de banda estreita e banda larga, para várias frequências e amplitudes. Por exemplo, no gráfico referente a um tom senoidal de 200 Hz, a curva correspondente a um tom de volume sonoro igual a 99dB (Phons) indica que um tom de 1 kHz só será percebido se estiver acima de 62 dB.

Em outras palavras, a presença de um tom de 200 Hz, com amplitude de 99 dB, faz com que a faixa dinâmica da audição seja restrita a (99-62) = 37 dB na banda de 200 a 1000 Hz aproximadamente. Esta característica da audição é bastante explorada pelos processos de compressão de áudio.

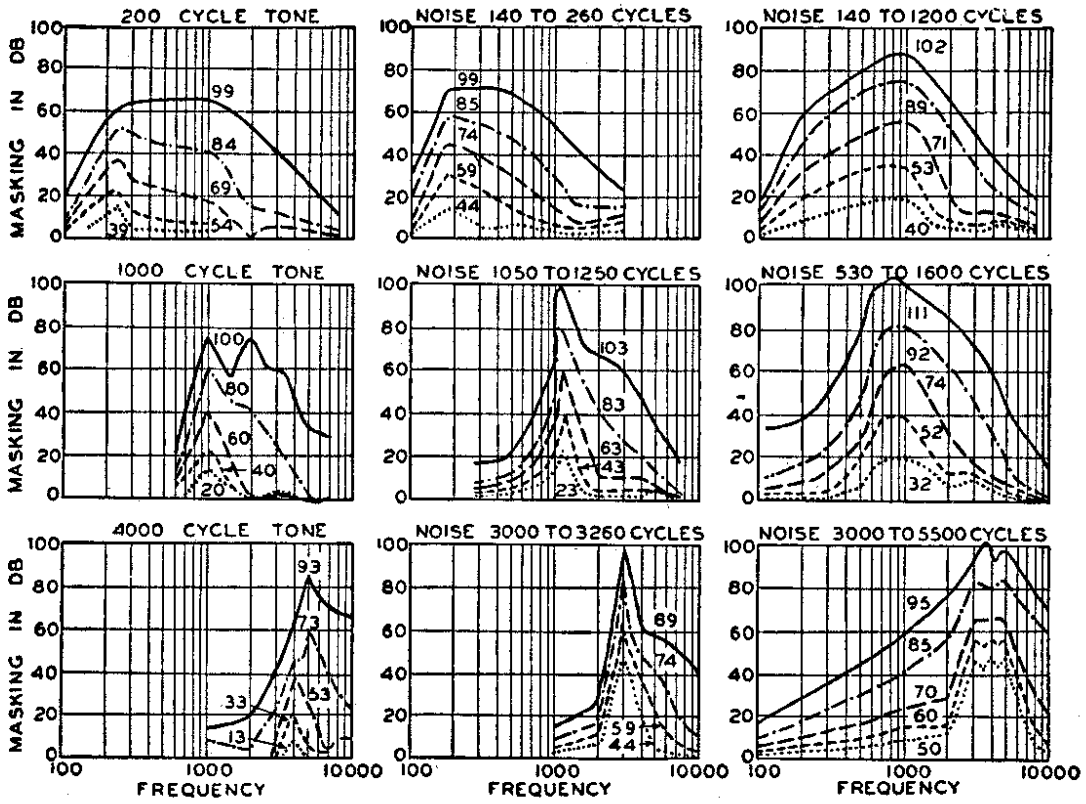

Fig. 4.4 - Curvas de Mascaramento (Fletcher & Munson)

5. Codificação por Sub-Bandas

A maioria dos métodos eficientes de compressão de áudio explora o mascaramento através da subdivisão do espectro em sub-bandas. A idéia geral é decompor o sinal por um banco de filtros passa-banda (fig. 5.1), possivelmente relacionados às bandas críticas de audibilidade, de modo que na saída de cada filtro tenhamos um sinal de banda estreita, cuja amplitude será via de regra menor que a do sinal original.

A seguir, é feita uma estimativa de mascaramento, baseada no espectro do sinal original, de modo que para cada sub-banda é determinada a relação Sinal/Ruído necessária para reproduzir o respectivo sinal filtrado. Então, o sinal de cada sub-banda é quantizado com um número de bits suficiente, de modo que o ruído de quantização esteja abaixo da relação S/R exigida para aquela banda.

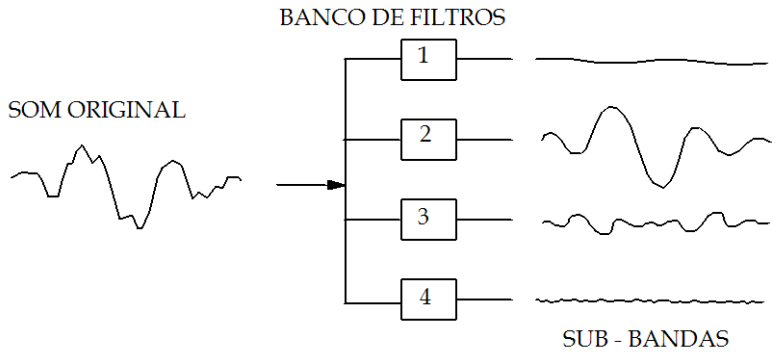

Fig. 5.1 - Análise de um Sinal por Sub-Bandas

A figura 5.3 mostra um diagrama de blocos do codificador de áudio descrito no padrão MPEG; neste caso, o sinal original é analisado por um banco de 32 filtros (fig. 5.2). Supondo por exemplo uma digitalização a 48 kHz, a banda de cada filtro será de 1/32 x 24 kHz = 750 Hz.

As saídas dos filtros são sub-amostradas (decimadas), de modo que a taxa de amostras original (48 kHz) é decomposta em 32 sinais amostrados em 1500 Hz. Uma vez que as respostas dos filtros se superpõem parcialmente, ocorrerá rebatimento espectral ("aliasing"), e um tom dentro de uma sub-banda provocará aparecimento de espúrios nas sub-bandas adjacentes. No entanto, os filtros usados são projetados com simetria da resposta em fase, de modo que os espúrios gerados por uma banda teoricamente se cancelam com os espúrios gerados pelas bandas adjacentes no decodificador. Estes filtros são conhecidos por QMF ("Quadrature Mirror Filters").

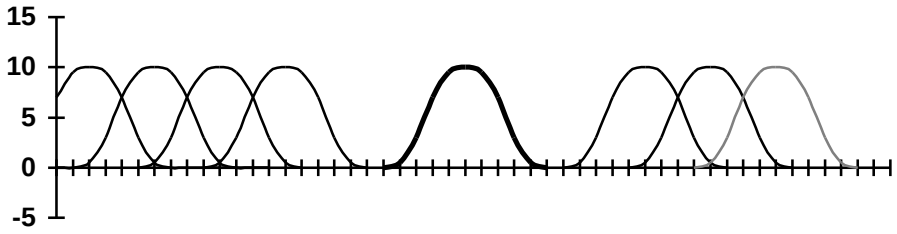

Fig. 5.2 - Respostas do banco de Filtros MPEG

A determinação do nível de mascaramento é feita por um Modelo Psico-acústico, a partir do espectro obtido por uma FFT de 512 ou 1024 pontos. Embora o detalhamento do modelo não esteja especificado no padrão MPEG, sua operação segue o esquema geral da figura 5.4.

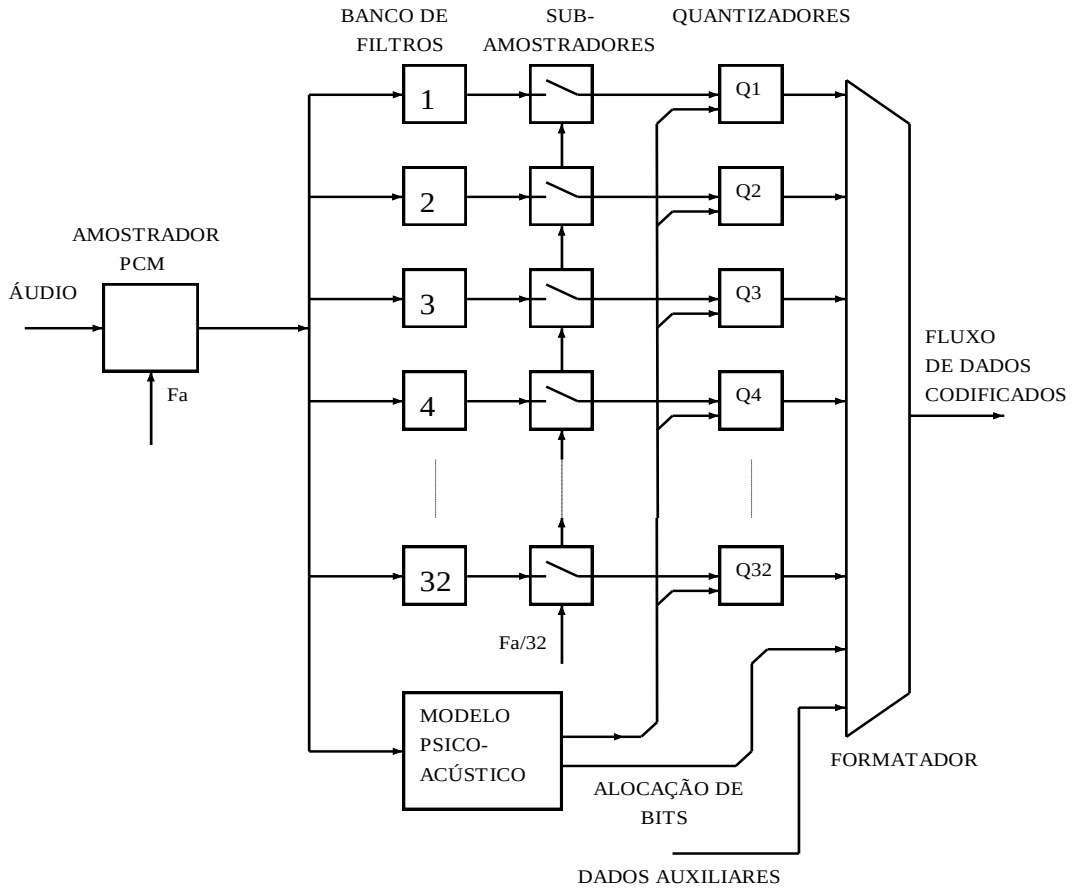

Fig. 5.3 - Estrutura do Codificador MPEG-Áudio

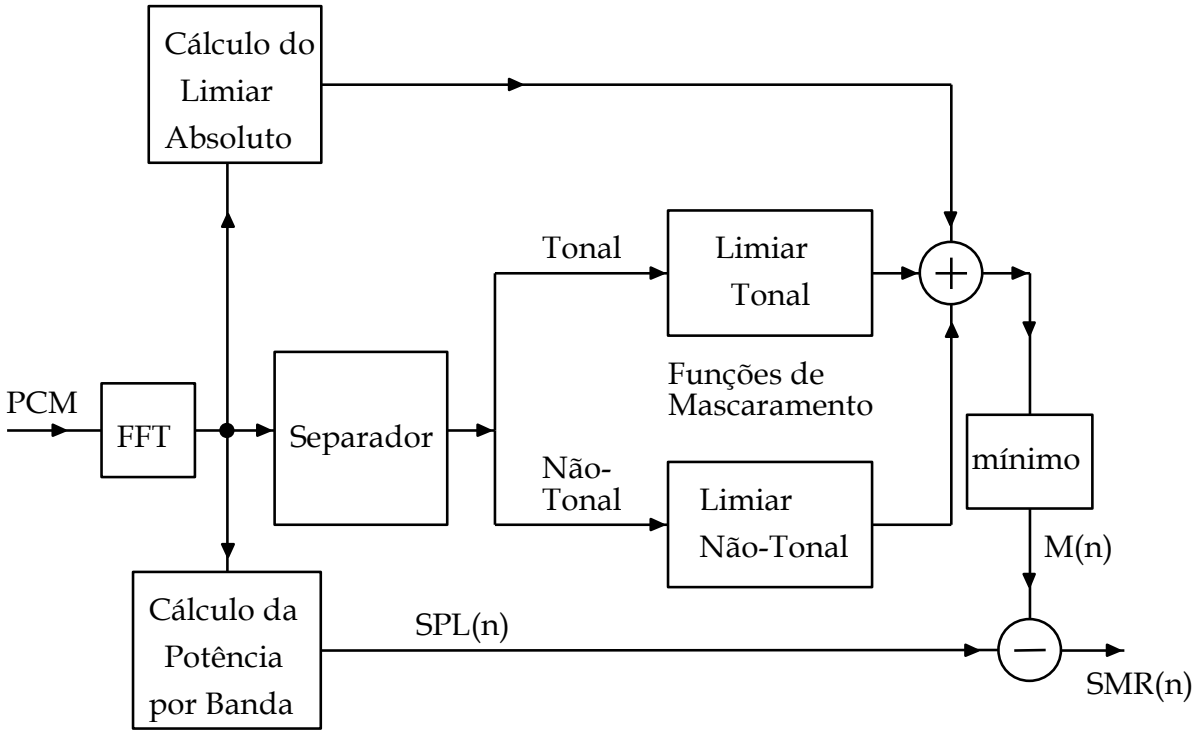

Fig. 5.4 - Exemplo de Estrutura de Modelo Psico-Acústico MPEG

As tarefas envolvidas na codificação seriam as seguintes:

1. Medida do Espectro (por FFT) do sinal completo, dentro de um certo intervalo de tempo (quadro de amostras);
2. Determinação, para cada sub-banda ***n***, do nível de potência sonora ***SPL(n)***;
3. Cálculo do limiar absoluto por banda, correspondente ao limiar de audibilidade;
4. Separação entre sinais tonais e não-tonais para escolha do padrão de mascaramento (esta separação pode ser feita pelo fator de crista, ou seja, pela relação entre valor de pico e valor r.m.s. por sub-banda. Sinais tonais apresentam fator de crista menor que sinais não-tonais ou ruído);
5. Cálculo do efeito de mascaramento devido às componentes tonais, dentro e fora da sub-banda, considerando ainda mascaramento temporal (podem ser usadas tabelas baseadas nas curvas de mascaramento);
6. Idem para as componentes não-tonais (usando tabelas diferentes);
7. Cálculo do limiar de audibilidade total, considerando as contribuições do limiar absoluto e dos efeitos de mascaramento tonais e não-tonais (figura 5.5);
8. Cálculo do valor mínimo (em função da frequência) do limiar total dentro de cada sub-banda, obtendo o nível de mascaramento ***M(n)*** ;
9. Cálculo da Relação Sinal / Mascaramento ***SMR(n)*** (Signal-to-Mask Ratio), como apresentado na Fig. 5.6;
10. Alocação de Bits para quantização de cada sub-banda, em função da ***SMR(n)*** e da porcentagem de ocupação do canal de saída;
11. Formatação do fluxo de dados, incluindo conjuntos de amostras, fator de escala e quantidade de bits alocados para cada sub-banda.

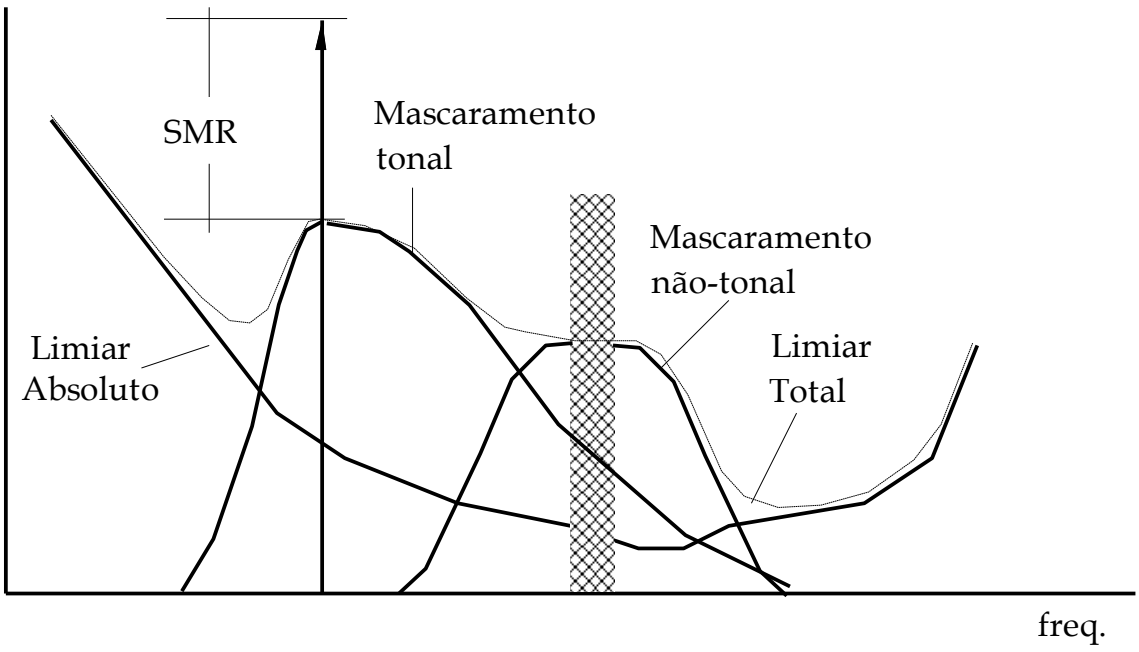

Fig. 5.5 - Cálculo do Limiar de Mascaramento Total

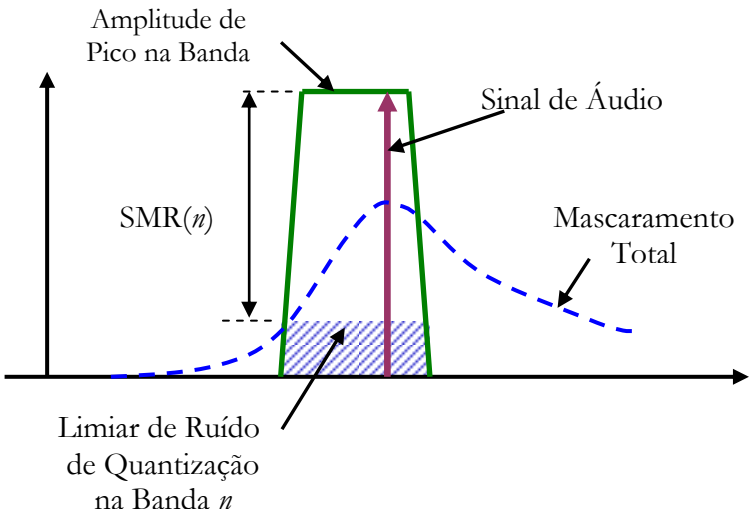

Fig. 5.6 – Relação Sinal / Mascaramento Dentro de Uma Banda

6. Padrões de Compressão de Áudio

6.1 Padrões MPEG

O padrão MPEG-1 de compressão de áudio (ISO / IEC 11172 - Parte 3) especifica 3 níveis de desempenho, associados basicamente à complexidade do modelo psico-acústico e dos bancos de filtros. Já o padrão MPEG-2 (ISO 13818 - Parte 3) estende a codificação para sistemas com 5.1 canais, adequados para "Home Theater" com "Qualidade Cinema".

A maioria dos sistemas trabalha com taxa de bits constante na saída, ao contrário do que ocorre com a compressão de vídeo.

Algumas características de cada sistema são apresentadas a seguir:

6.1.1 MPEG-1 Nível I:

- Codificação feita em Quadros de 12 amostras por sub-banda (384 amostras e 32 sub-bandas);
- Para cada quadro permanecem constantes os fatores de escala e alocações de bits por amostra por banda;
- Aceita taxas de amostragem de 32kHz, 44.1 kHz ou 48 kHz para um ou dois canais independentes ou associados;

O padrão MPEG-1 Nível I é usado no DCC (Digital Compact Cassette) com o nome de PASC (Precision Audio Sub-band Coding).

6.1.2 MPEG-1 Nível II (MUSICAM):

- Codificação em Quadros de 3 × 12 amostras por sub-banda (1152 amostras) proporcionando menor "overhead" na transmissão;
- A análise de espectro para o modelo psico-acústico é feita com FFT de 1024 pontos;
- Admite codificação estatística agrupando 3 amostras consecutivas em um único código;

6.1.3 MPEG-1 Nível III (MP3):

- Codificação em Quadros de 3 × 12 amostras por sub-banda (1152 amostras);
- Usa um modelo psico-acústico mais elaborado;

- Utiliza transformada de cossenos (DCT) para subdividir as Sub-bandas em 6 ou 18 sub-sub-bandas, proporcionando redução de "Aliasing" entre sub-bandas;
- Efetua aplicação de janelas temporais curtas (6 amostras) ou longas (18 amostras) antes da DCT;
- Pode usar Quantizador não-linear : Y=Q(X0.75);
- Adota fator de escala válido por bandas;
- Aplica codificação de Huffman nas amostras quantizadas;
- Utiliza Memória Elástica ("Buffer") para regularizar a taxa média de bits;
- Implementa um controle de taxa de compressão baseado na ocupação do "Buffer";

6.1.4 MPEG-2 (Níveis I, II, III):

- Suporte para 5.1 canais (HDTV), com predição adaptativa dos canais adicionais baseada nos canais L e R;
- Prevê suporte para até 7 canais adicionais de voz (multilingual);
- Codificação e decodificação compatível com MPEG-1 (sintaxe inclui um fluxo de dados compatível com MPEG-1 carregando os canais L e R);
- Aceita outras taxas de amostragem (16, 22.05 e 24 kHz).

6.1.5 Estrutura de Pacotes de Dados MPEG-1

A figura 6.1 apresenta a estrutura dos pacotes de transporte de dados para o padrão MPEG-1, bem como o detalhamento do cabeçalho.

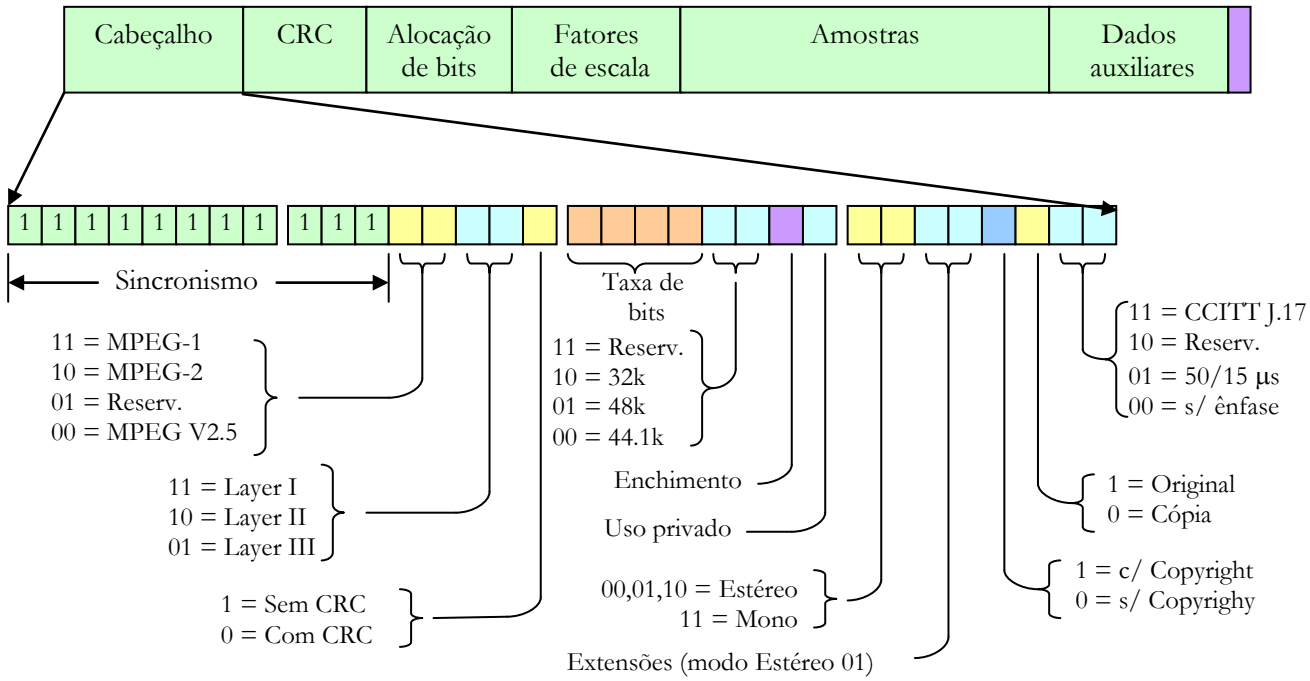

Fig. 6.1 – Pacote de Dados MPEG-1 Áudio

O sincronismo do pacote é feito por uma sequência de 11 bits consecutivos iguais a '1' . Em seguida, 4 bits definem o padrão e o nível da implementação.

A existência de CRC, após o cabeçalho, é sinalizada por mais um bit. Em seguida é informada a taxa de bits, através de índice em para uma tabela, que depende do padrão e do nível. Um

campo de 2 bits seleciona a taxa de amostragem. Finalmente, o último byte traz informações de número de canais, registro de "Copyright", identificação de original ou cópia e uso de pré-ênfase em altas frequências.

Após o cabeçalho (e CRC opcional), há um campo que define a alocação de bits para cada banda. As sub-bandas são separadas em 3 grupos (*Low, Mid* e *High*), sendo que para cada grupo os campos correspondentes selecionam os seguintes quantizadores:

- Bandas '*Low*': 4 bits por banda selecionam 1 entre 15 quantizadores possíveis, variando de 3 a 65535 níveis de quantização;
- Bandas '*Mid*': 3 bits por banda, selecionam um de 7 quantizadores;
- Bandas '*High*': 2 bits por banda, selecionam um de 3 quantizadores possíveis.

Em seguida, há um campo com os fatores de escala (ganhos), ocupando 6 bits para cada banda. Os ajustes de ganho cobrem uma faixa dinâmica de 120 dB.

6.2 Sony ATRAC (Adaptive Transform Acoustic Coder):

A gravação magneto-óptica "Mini-Disc" utiliza o algoritmo denominado ATRAC , que aplica uma combinação de codificação por sub-bandas com transformada DCT (Fig. 6.1).

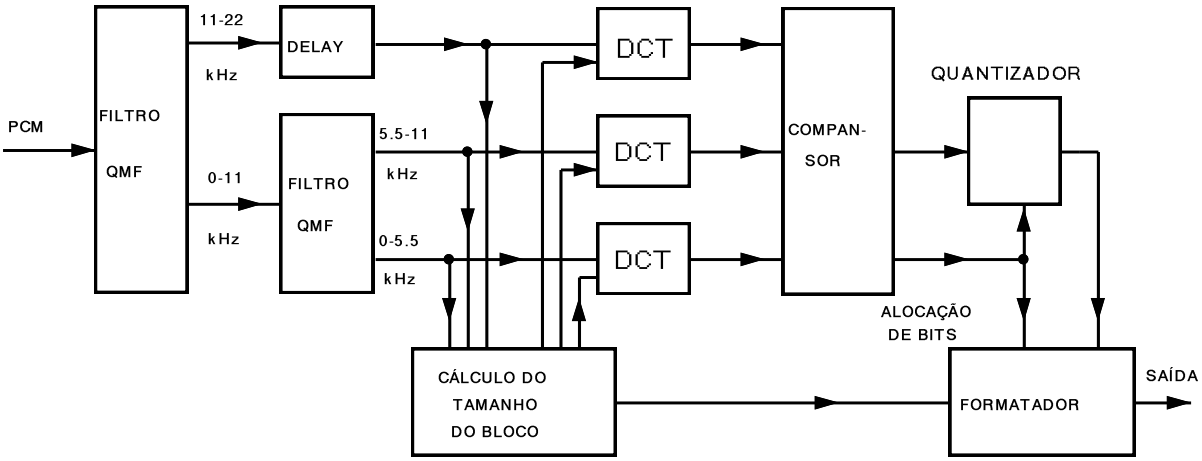

Fig. 6.1 - Codificador ATRAC (Sony)

O sinal de áudio, quantizado em 16 bits a 44.1 kHz, é separado em 3 bandas (0 a 5.5, 5.5 a 11 e 11 a 22 kHz) por dois bancos de filtros QMF. O sinal de cada banda é dividido em blocos com duração variável de 1.45 a 11.6 ms, os quais são decompostos em coeficientes através de uma transformada DCT modificada. Os coeficientes são então reagrupados em 52 bandas de frequência com larguras variáveis, aproximadamente de acordo com as larguras das bandas críticas de audibilidade. Os coeficientes são compandidos e quantizados de acordo com um modelo de mascaramento auditivo. O tamanho do bloco é independente para cada uma das três bandas, sendo escolhido em função da existência de transientes no sinal de áudio.

6.3 Codificador Dolby AC-3:

O padrão AC-3 (adotado para o sistema de radiodifusão HDTV e para o sistema DVD nos EUA) baseia-se na codificação de uma representação do sinal de áudio no domínio da frequência (Fig. 6.2). O sinal de áudio é amostrado a 48, 44.1 ou 32 kHz, sendo que a taxa de amostragem é sincronizada com um sinal de 27 MHz que é usado também na amostragem do sinal de vídeo.

O sinal de áudio é codificado em quadros de 1536 amostras, separadas em 6 blocos de 512 amostras com superposição de 50%; cada um destes blocos é multiplicado por uma janela temporal e transformado em coeficientes no domínio da frequência através de uma transformada DCT (Fig. 6.3). Devido ao formato da janela temporal, os coeficientes são correlacionados entre si e podem ser decimados na razão de 2:1, restando assim 256 coeficientes para cada bloco. Ainda devido à janela, os blocos se superpõem no receptor, cancelando no domínio do tempo os espúrios causados pela DCT.

Os coeficientes X(k) são obtidos a partir das amostras x(n) por:

$$X(k) = \frac{-1}{256} \sum_{n=0}^{511} x(n) J(n) \cos\left(\frac{2\pi}{2048} (2n+1)(2k+1) + \frac{\pi}{4} (2k+1) \right)$$

onde J(n) é a janela temporal (obtida de uma tabela).

Os coeficientes são representados em ponto flutuante, obtendo-se um expoente e uma mantissa para cada um. Os expoentes definem uma representação aproximada do espectro do sinal, em escala logarítmica, denominado de Envoltória Espectral. Antes da codificação de um quadro completo, os expoentes dos seis blocos são examinados quanto às variações de um bloco para outro. Havendo variações significativas entre blocos sucessivos, são formadas envoltórias espectrais válidas para grupos de blocos dentro dos quais as diferenças são menores. Além disso, como os expoentes de coeficientes consecutivos geralmente diferem entre si por valores até +/- 2, os expoentes são codificados de forma diferencial. A codificação da envoltória espectral ocupa assim cerca de 0,3 a 0,4 bits por amostra.

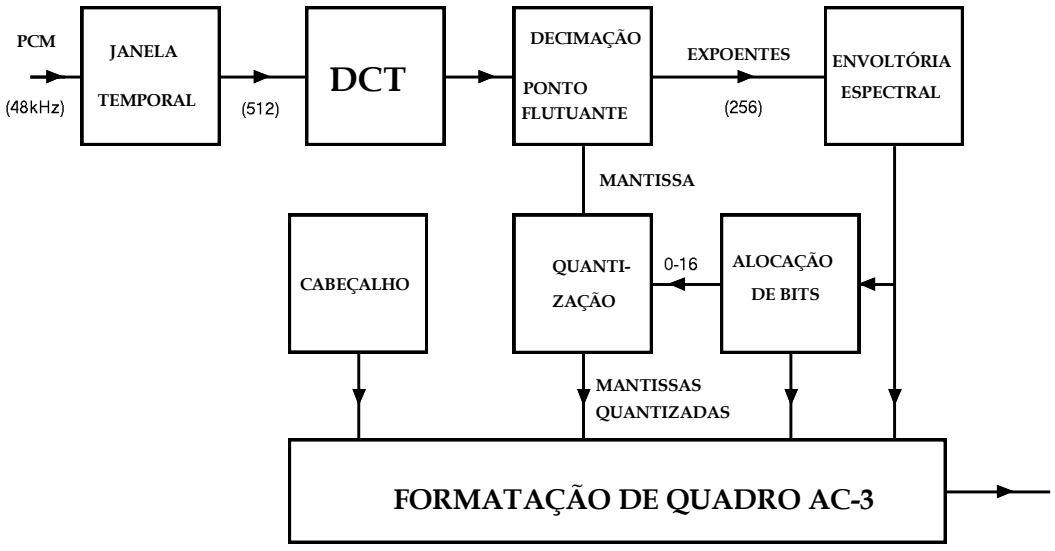

Fig. 6.2 - Codificação de Áudio Dolby AC-3

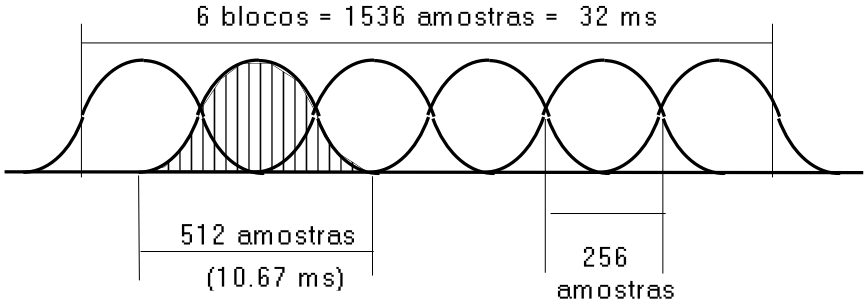

Fig. 6.3 - Janelas Temporais Superpostas para DCT (AC-3)

Baseando-se na envoltória espectral, o algoritmo de alocação de bits determina a precisão necessária para a quantização de cada mantissa. Um modelo psico-acústico determina uma curva de mascaramento para cada bloco; subtraindo esta curva da envoltória espectral, obtém-se a relação S/R necessária para cada coeficiente em função da frequência.

Quando o sistema encontra transientes impulsivos, pode ocorrer um fenômeno de espalhamento do ruído de quantização: o modelo psico-acústico calcula a relação S/R baseando-se na amplitude máxima do transiente, alocando então parâmetros de quantização que podem resultar em ruído elevado. Como esse ruído vale para todo o bloco (5,33ms), pode ser audível em certas condições, especialmente quando o transiente ocorre no final do bloco. Nesses casos, pode ser vantajoso subdividir o bloco de 512 amostras em 2 blocos de 256 amostras; desta forma, o espalhamento de ruído é reduzido.

Além disso, o codificador pode utilizar o recurso de Acoplamento, que consiste em combinar as altas frequências de dois ou mais canais em um único canal conjunto. O codificador transmite então coordenadas de acoplamento, que indicam as atenuações que devem ser aplicadas ao canal conjunto para reconstruir os canais independentes no receptor. Este recurso baseia-se no fato de que a percepção de direção de uma fonte sonora, em altas frequências, deve-se mais à diferença de amplitude do que à forma de onda específica recebida por cada ouvido.

Os canais codificados são combinados em Quadros AC-3 (Fig. 6.4), correspondentes a 1536 amostras. Os quadros são compostos de:

- Palavra de sincronismo: 16 bits de Sinc (0B77h) , 16 bits de CRC (válido para 5/8 do quadro) , mais 8 bits definindo taxa de amostragem (48,44.1 ou 32 kHz) e tamanho do quadro e taxa de transmissão (32 a 640 kb/s)

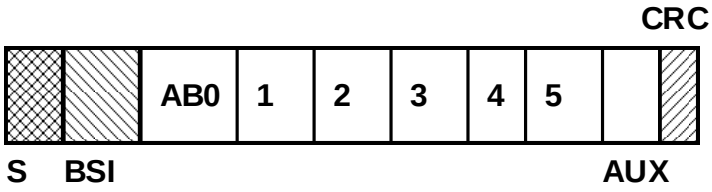

Fig. 6.4 - Formatação do Quadro AC-3

- Parâmetros descritivos do serviço fornecido (BSI - Bit Stram Information): Identificador de padrão AC-3 (5 bits), tipo de serviço (ver detalhamento do tipo de serviço adiante), número de canais (1 a 5), presença de canal de baixa-frequência, informações de língua, compressão de faixa dinâmica, nível de referência de volume (nível de diálogos), proteção contra cópias, time code (28 bits);
- Codificação dos 6 blocos de áudio: informações de acoplamento, subdivisão de blocos, faixa dinâmica; envoltórias espectrais (expoentes); parâmetros para o algoritmo de alocação de bits no decodificador; mantissas (quantizadas com 0 a 16 bits);

- Dados auxiliares (normalmente não utilizados);
- CRC2 para detecção de erros (16 bits), válido para o quadro inteiro.

6.4 Tipos de Serviços de Áudio Suportados pelo Padrão AC-3:

Este padrão admite vários serviços de áudio para diferentes aplicações, sendo que cada serviço é codificado em quadros AC-3 independentes, porém fornecidos sequencialmente no mesmo bit stream. Há 2 tipos de Serviços Principais e 6 tipos de Serviços Associados:

Complete Main Audio Service (CM) : Diálogo, Música e Efeitos ; 1 a 5.1 Canais, podendo ser codificado em múltiplas línguas;

Main Audio Service, Music and Effects (ME) : Música e Efeitos sem Diálogo, podendo ser combinado com áudio proveniente do Serviço D;

Visually Impaired (VI) : Descrição narrativa da informação visual, destinada a atender deficientes visuais; 1 Canal de áudio que pode ser combinado com o Serviço CM;

Hearing Impaired (HI) : Apenas canal de Diálogo, processado p/ maior inteligibilidade, destinado a deficientes auditivos;

Dialogue (D) : Diálogo a ser combinado com áudio proveniente do Serviço ME; pode ser fornecido em várias línguas e codificado em 1 canal (Central) ou 2 canais;

Commentary (C) : Comentários opcionais ao Programa Visual; pode ser fornecido como 1 canal a ser combinado ao serviço ME;

Emergency (E) : Canal de maior prioridade, admite inserção de mensagens de emergência, sobrepondo-se aos outros serviços;

Voice-Over (VO) : 1 canal de áudio, que permite inserções de voz a serem superpostas a um serviço já codificado, sem a necessidade de que o áudio original seja decodificado , misturado com a inserção e re-codificado novamente.

7. MPEG-2 AAC (Advanced Audio Coding)

O padrão AAC, desenvolvido como extensão para compressão de áudio em MPEG-2 e MPEG-4, apresenta ferramentas adicionais que permitem melhora na qualidade da codificação e redução de até 50% na taxa de bits. A figura 7.1 apresenta um diagrama de blocos do processo de codificação MPEG-2 AAC.

Entre os recursos introduzidos neste processo, temos:

- Codificação para 1 a 48 canais de áudio;
- Taxas de amostragem de 8 a 96 kHz;
- Banco de filtros implementado por Transformada Discreta de Cossenos, similar ao Dolby AC-3, com 1024 pontos;
- Utilização de preditor linear (LPC) e equalização para "Noise Shaping" adaptativa;
- Possibilidade de utilização de "replicação espectral".

O conceito de "replicação espectral" envolve a síntese, no receptor, de componentes espectrais (harmônicas ou ruído) que foram suprimidas no codificador pelo filtro *anti-aliasing*, por ocasião da amostragem.

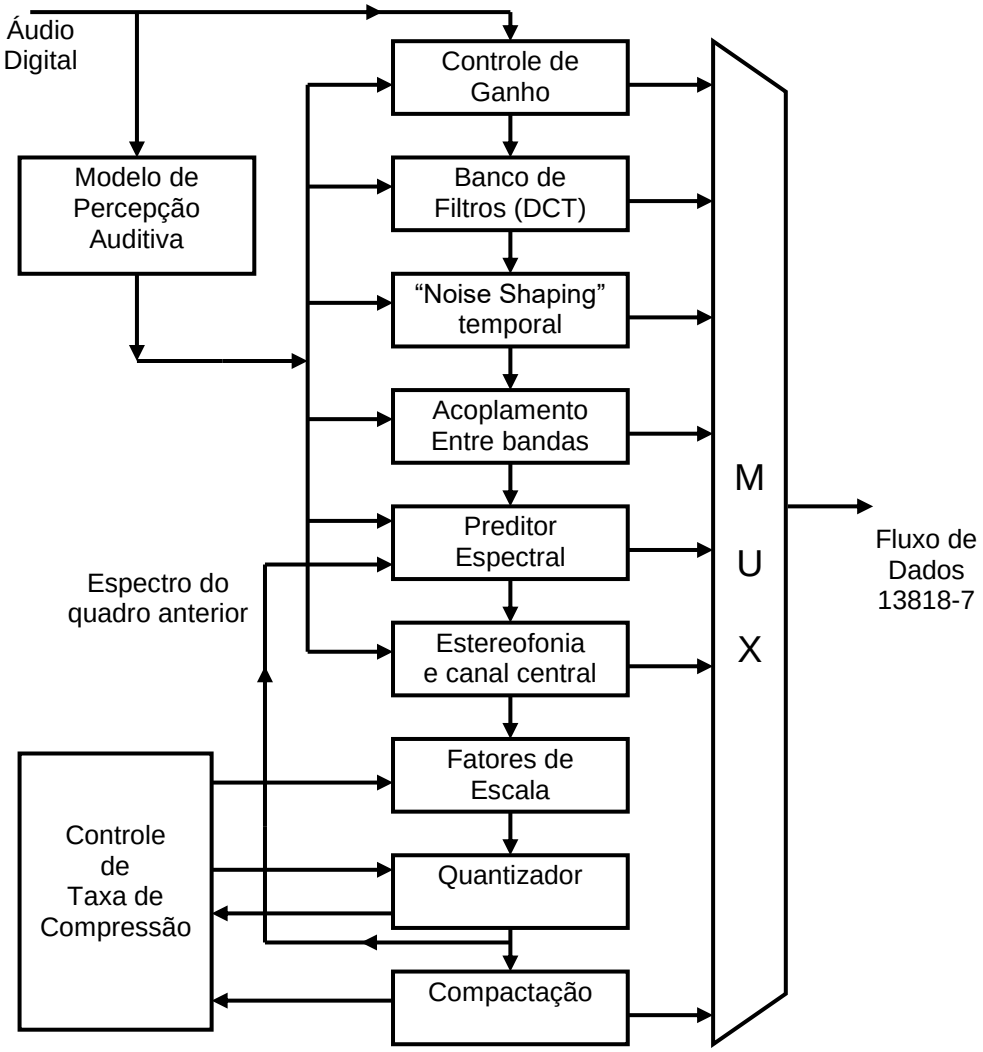

Fig. 7.1 – Diagrama de Blocos do Codificador AAC

Estes recursos são utilizados pelo codificador destinado à compressão de áudio genérico (música, trilha sonora etc.) denominado "Natural Audio Coding". O padrão MPEG-4 admite outros codificadores com ferramentas adequadas para aplicações específicas, como compressão de voz humana ("Vocoder", HVXC, HINL, CELP), síntese de instrumentos musicais (MIDI), etc.

De acordo com o diagrama apresentado, o sinal de áudio digitalizado é analisado segundo um modelo perceptual, o qual controla os parâmetros dos demais blocos. A sequência de processamento das amostras de áudio é a seguinte:

- Controle de Ganho: nesta etapa, o sinal é separado em 4 bandas de frequência, sendo aplicado um ganho individual para as 3 bandas superiores, de modo a equalizar a energia nessas regiões do espectro (fig. 7.2);
- Banco de Filtros: Aplica-se uma janela temporal longa ou curta (dependendo da existência de transientes), e a seguir é feita uma decomposição espectral usando Transformada Discreta de Cossenos, com 2048 ou 256 pontos (de acordo com o comprimento da janela temporal). As amostras são decimadas, resultando em 1024 ou 128 coeficientes espectrais. A figura 7.3 apresenta exemplos de janelas longas e curtas intercaladas.

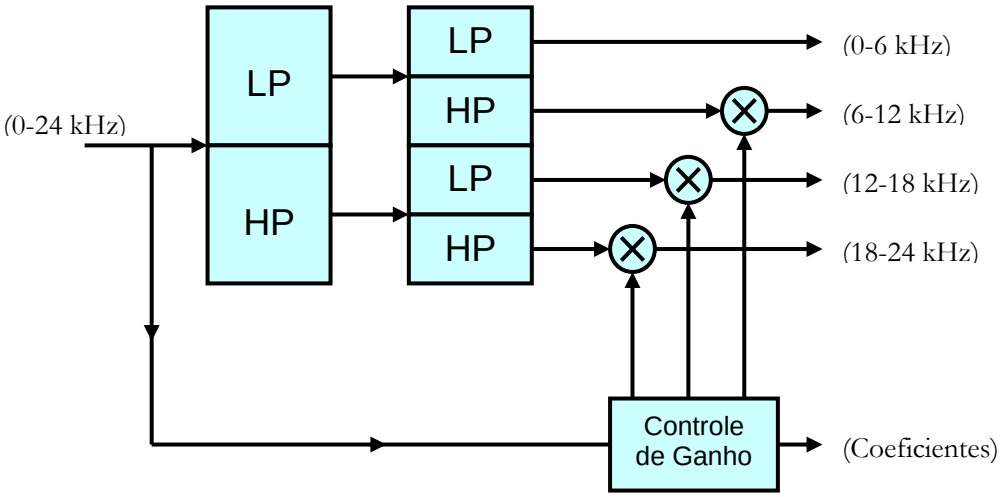

Fig. 7.2 – Controle de Ganho por Bandas

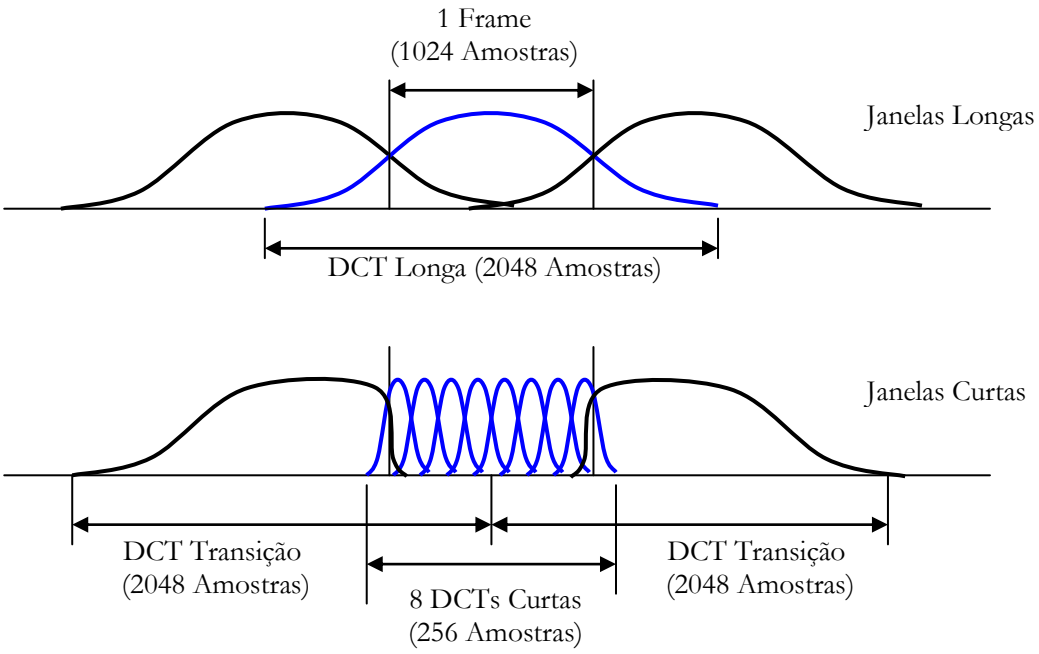

Fig. 7.3 – Janelas Longas e Curtas

- "*Noise Shaping*" Temporal (**TNS**): Este bloco aplica uma pré-ênfase antes da quantização, que deverá ser compensada com uma de-ênfase no receptor. Desta forma, o erro de quantização nas componentes de menor amplitude será reduzido em relação ao das componentes de maior amplitude (fig. 7.4). A forma do filtro de pré-ênfase é obtida a partir do espectro de um "frame", utilizando-se um filtro "passa-baixas" no domínio do espectro para gerar uma curva que representa a envoltória espectral média.
- Preditor Espectral: este bloco efetua a codificação das diferenças espectrais entre o quadro atual e o espectro quantizado do quadro anterior. Cada coeficiente espectral é tratado por um filtro de predição de segunda ordem (fig. 7.5).
- Fatores de Escala: As amostras obtidas do preditor são ajustadas por fatores de escala, válidos para cada frame e grupo de frequências, com resolução de 1,5 dB, de modo a ajustar o fundo de escala do quantizador. Os fatores de escala são codificados de forma diferencial em relação aos fatores usados no quadro anterior.

- Quantizador Não-Linear: Após ajuste de escala, os coeficientes são processados por uma função não-linear da forma $y = x^{0,75}$ e então quantizados com um número variável de níveis. A figura 7.6 esquematiza os fatores de escala e o quantizador não-linear.
- Compactação: A associação de um Fator de Escala e do respectivo Quantizador constitui um símbolo, que é codificado conjuntamente usando um dicionário de Huffman. Os coeficientes quantizados são por sua vez agrupados em 2 ou 4 valores para cada símbolo, e codificados utilizando um dicionário escolhido entre 11 tabelas possíveis.

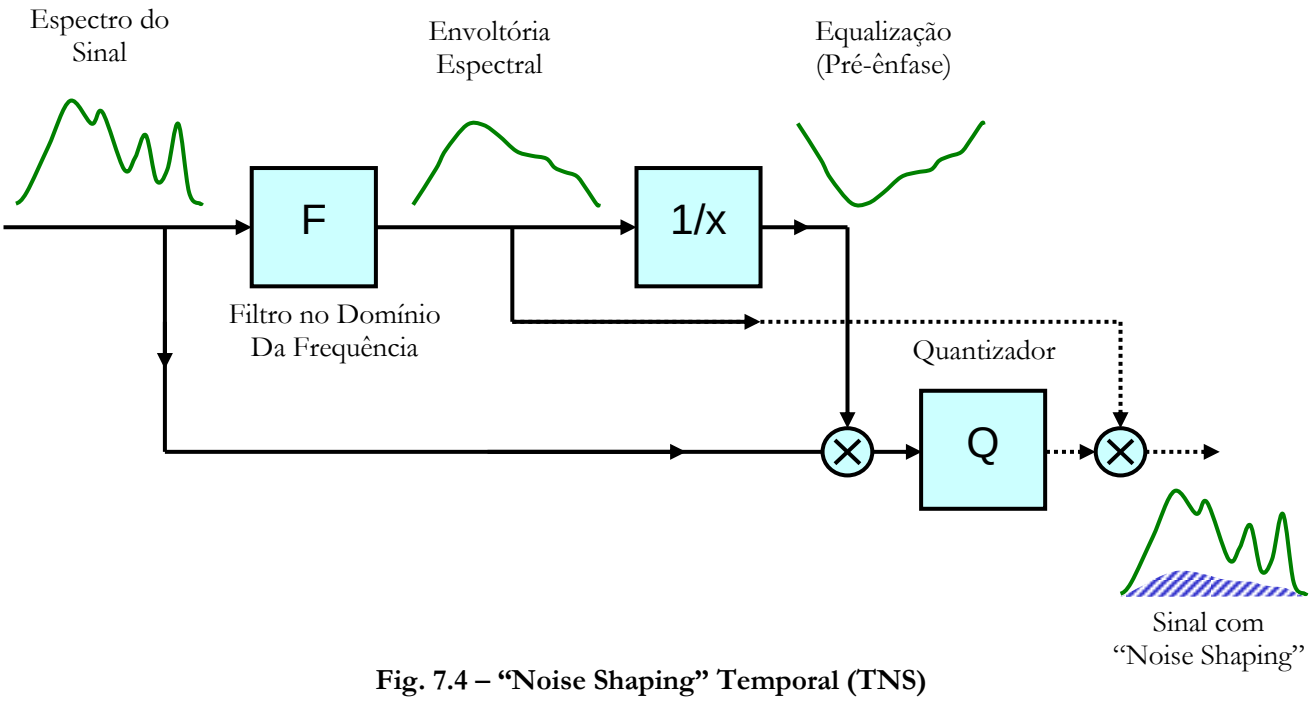

Fig. 7.4 – "Noise Shaping" Temporal (TNS)

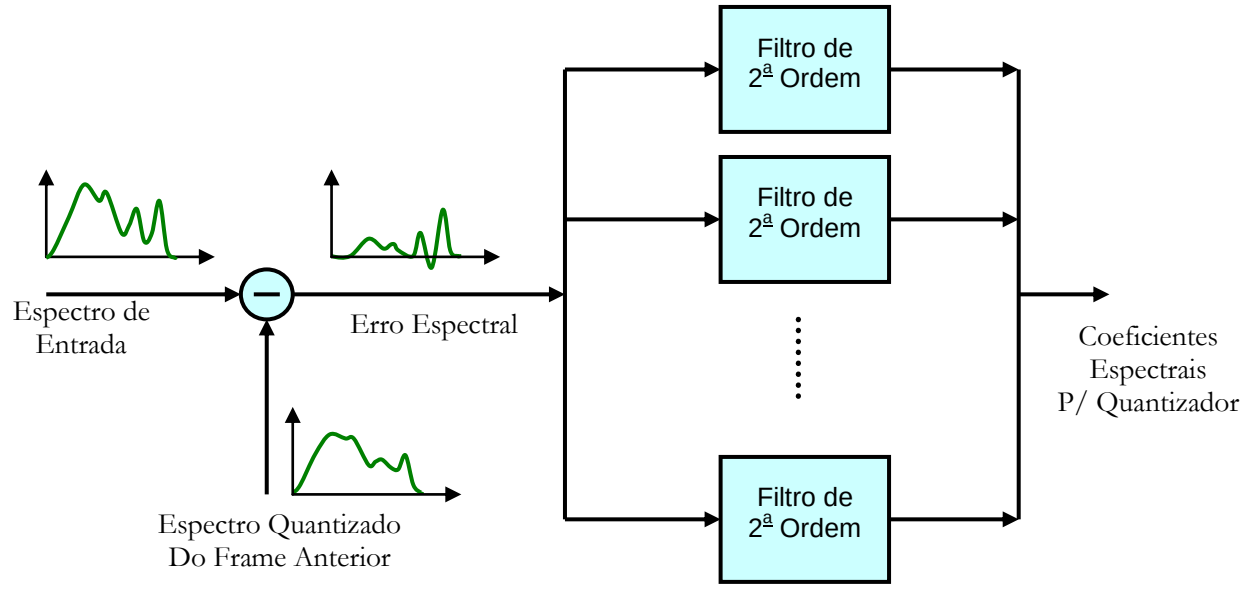

Fig. 7.5 – Preditor Espectral

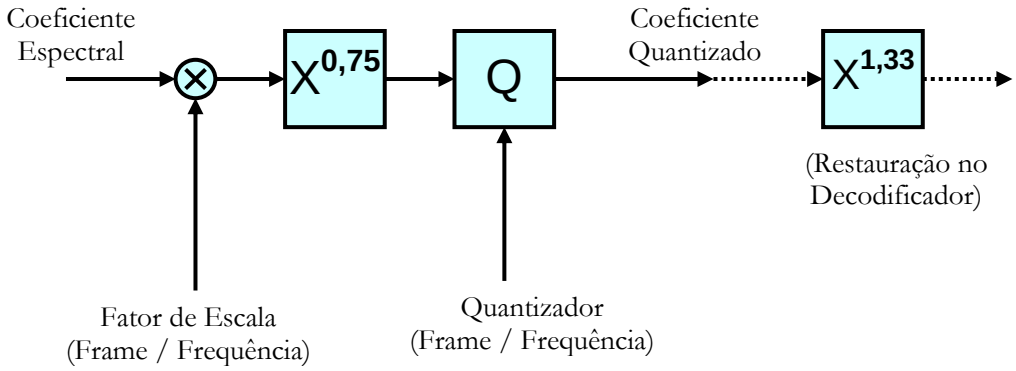

Fig. 7.6 – Fatores de Escala e Quantizador Não Linear

8. Desempenho dos Processos de Compressão de Áudio:

A tabela abaixo apresenta o desempenho de alguns processos de compressão, com parâmetros ajustados para qualidade próxima à de CD.

Padrão de Compressão	Canais	Taxa Média de Compressão	Taxa Original	Taxa Comprimida	Atraso A/D/A
MPEG-1 / I	2	4:1	1,41 Mb/s	384 kb/s	19 ms
ATRAC	2	5:1	1,41 Mb/s	292 kb/s	
MPEG-1 / II (MUSICAM)	2	6:1	1,54 Mb/s	256 kb/s	> 40 ms
MPEG-1 / III	2	8:1	1,41 Mb/s	128 kb/s	> 80 ms
MPEG-2 / III	5.1	10:1	3,53 Mb/s	384 kb/s	
Dolby AC-3	5.1	12:1	3,53 Mb/s	290 kb/s	45 ms
MPEG-2 AAC	5.1	18:1	3.53 Mb/s	196 kb/s	> 80 ms

9. Referências:

A. J. M. Hotsma, T. D. Rossing, W. M. Wagenaars: *Auditory Demonstrations (CD)*; IPO/Philips, 1987

Advanced Television Systems Commitee: *Digital Audio Compression Standard*; 1995

F. Alton Everest: *The Master Handbook of Acoustics*; TAB Books, 1994

Harry F. Olson: *Elements of Acoustical Engineering*; D. Van Nostrand Co., 1947 (figs. 3.2, 4.4)

Jerry Whitaker: *DTV - The Revolution in Electronic Imaging*; McGraw-Hill, 1998

John Watkinson: *Compression in Video & Audio*; Focal Press, 1995

Ken C. Pohlmann: *Principles of Digital Audio*; Sams/MacMillan,1990

Richard M. Warren: *Auditory Perception - A New Analysis And Synthesis* ; Cambridge,1999 (fig. 2.7, 4.3)

Thomas D. Rossing: *The Science of Sound* ; Addison-Wesley, 1990 (figs. 2.9, 3.1)

NAB Engineering Handbook; National Association of Broadcasters, 1999

E. Bruce Goldstein: *Sensation & Perception*; Brooks/Cole, 1999 (figs. 2.2, 2.3, 2.4, 2.6)

ISO/IEC FCD 0.1 14496-3 Subpart 2: *Information Technology - Very Low Bitrate Audio-Visual Coding;* ISO - 1996