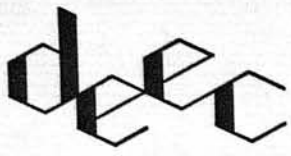

DEPARTAMENTO DE

ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES

CODIFICAÇÃO PERCEPTUAL DE ÁUDIO DIGITAL ESTEREOFÓNICO

Aníbal João de Sousa Ferreira

FACULDADE DE ENGENHARIA

UNIVERSIDADE DO PORTO

Rua dos Bragas, 4099 Porto Codex – PORTUGAL

CODIFICAÇÃO PERCEPTUAL DE ÁUDIO DIGITAL ESTEREOFÓNICO

Aníbal João de Sousa Ferreira

CODIFICAÇÃO PERCEPTUAL DE ÁUDIO DIGITAL ESTEREOFÓNICO

Aníbal João de Sousa Ferreira

Dissertação submetida para satisfação parcial dos requisitos
do Curso de Mestrado em

Engenharia Electrotécnica e Computadores
Perfil de Telecomunicações

Departamento de Engenharia Electrotrotécnica e Computadores
Faculdade de Engenharia da Universidade do Porto

Porto
Janeiro de 1992

Tese realizada sob a supervisão do Doutor

Mário Jorge Moreira Leitão

Professor Auxiliar da Faculdade de Engenharia
da Universidade do Porto

SUMÁRIO

O familiar Disco Compacto permite o registo de sinais áudio que preenchem quase totalmente a região de funcionamento útil do ouvido humano. É por isso comum a sua designação de áudio de alta qualidade. O tema desta dissertação é a compressão deste tipo de sinais, sem afectar ou comprometer a sua qualidade subjectiva original. Com este objectivo, só se admitem duas premissas: os sinais áudio são representados por dois canais (estereofonia) e o receptor chama-se ouvido humano. É sobre este último elemento que se procurará fundamentar o elevado ganho de compressão.

A codificação perceptual de sinais áudio de alta qualidade é uma área de investigação que derivou da compressão de sinais de voz e que se demarcou desta e autonomizou na década de oitenta. Este percurso de desprendimento foi também o permitido pela própria evolução tecnológica e o esboçado pela sedimentação de conhecimentos acerca do sistema auditivo humano. As actividades de normalização no âmbito da ISO/MPEG, ainda em curso, estimularam também a consolidação de resultados.

A psico-acústica fornece importantes medidas de correlação entre estímulo acústico e sensação auditiva, permitindo a identificação da tolerância auditiva com a irrelevância do sinal e da acuidade auditiva com a Entropia Perceptual. Estes aspectos podem aglutinar-se numa entidade matematicamente tratável para postular um limiar de mascaramento. Neste contexto, o ganho de compressão afirma-se como resultado da extracção da redundância e da irrelevância contidas no sinal áudio. Estas componentes exibem uma grande eficiência no plano intracanal. Porém, a dimensão intercanal é também potenciadora de ganhos adicionais. Neste sentido, partindo de sugestões da psico-acústica, o limiar de mascaramento estereofónico associa-se à percepção binauricular, assim como o limiar de mascaramento monofónico retrata a percepção monauricular.

Os limiares de mascaramento são implementados num codificador perceptual e prova-se que é possível codificar transparentemente ou com muito boa qualidade, qualquer sinal estereofónico, usando uma taxa de informação total não superior a 128Kbit/s.

Palavras chave: *codificação perceptual, mascaramento, irrelevância, percepção monauricular, percepção binauricular, psico-acústica.*

PREFÁCIO

Assumindo uma prosa mais liberta, outras chaves poderiam também abrir o conteúdo das palavras. Prazer da descoberta, Beleza, Arte. São também os agentes activos de um trabalho de Engenharia, os catalizadores invisíveis no produto da reacção. Mas reais e legítimos.

A motivação neófila pode traduzir-se na procura de uma coerência ou de um estado de perfeição, acidentalmente, de natureza científica. E o ponto de partida pode ser a Entropia Perceptual de um sinal acústico, incorporando um esboço das propriedades analíticas do sistema audivo. E estas, não raro, são tão belas quanto insondáveis. Então, só o espírito expande novos espaços para além do inequivocamente observável. *André Breton* tinha razão.

As bases multidisciplinares podem não ser necessariamente ortogonais: acústica, fisiologia, psicologia, estatística, processamento de sinal, psico-acústica. Mas a sua aglutinação resulta num espaço revelador de harmonia que não deixa de ter vincado o *gesto* do autor. E o resultado pode ser arte. Como a música.

O trabalho apresentado nesta tese foi totalmente desenvolvido durante um estágio realizado nos Laboratórios Bell da AT&T em Murray Hill, New Jersey, Estados Unidos da América, de Outubro de 1990 a Setembro de 1991. Representou uma experiência particularmente estimulante e recompensadora quer no plano científico, quer no plano humano e só se tornou possível graças ao apoio e dedicação de várias pessoas e instituições.

As minhas primeiras palavras de agradecimento são dirigidas ao Prof. José Manuel Tribolet[†] , ao Prof. Mário Jorge Leitão[†] e ao Dr. Nikil Jayant[‡] , pela oportunidade que me proporcionaram, pela confiança que depositaram em mim e pelo encorajamento que sempre me induziram.

Quero agradecer especialmente a pessoa que mais de perto acompanhou, estimulou e iluminou o meu trabalho: James David Johnston[‡].

Agradeço também ao Nuno Guimarães[†] e ao Fernando Pereira[‡] a Amizade e simpatia que suavizaram, desde logo, os sobressaltos decorrentes da minha entrada nos Estados Unidos.

[†] INESC, Portugal
[‡] AT&T Bell Laboratories, USA

Nos Laboratórios Bell disfrutei sempre de um ambiente de trabalho bastante receptivo e agradável, o que é seguramente um dos vectores da sua contagiante sinergia. Quero agradecer a todos os meus colegas e amigos que sempre se disponibilizaram para ouvir, conversar e ajudar.

Quero também deixar uma palavra de agradecimento ao INESC que me acolheu e me proporcionou condições excelentes de trabalho e também à Junta Nacional de Investigação Científica e Tecnológica que me facultou uma Bolsa Mestrado.

Por último, quero expressar o meu profundo agradecimento aos meus pais, ao meu irmão e à Betsy, pois foi o seu incondicional calor e entusiasmo que me deram alento e me ajudaram a vencer, em particular, as amiudadas depressões gélidas de New Jersey.

ÍNDICE

GLOSSÁRIO

binauricular	*binaural*
bits	*bits*
codificação de fonte	*source coding*
codificação de poço	*sink coding*
combinador	*multiplexer*
disco compacto	*compact disc*
efeito de mascaramento	*masking effect*
estado-da-arte	*state-of-the-art*
intensidade percebida	*loudness*
monauricular	*monaural*
norma	*standard*
planura espectral	*spectral flatness*
psico-acústica	*psychoacoustics*
sistema auditivo	*auditory system*
sobreposição espectral	*spectral aliasing*
sobreposição temporal	*temporal aliasing*
ondas progressivas	*travelling waves*

1
INTRODUÇÃO

1.1 Enquadramento

A introdução, em 1982, do Disco Compacto no universo do grande consumo criou uma nova norma de qualidade e exigência associadas às fases de gravação, armazenagem e reprodução de sinais acústicos, genericamente identificados como música. Esta nova fase na área do áudio está intimamente ligada ao desenvolvimento da tecnologia digital e a uma crescente actividade de investigação na área de Processamento Digital de Sinal.

De facto, o aparecimento de conversores analógico-digitais e digital--analógicos de grande qualidade, com resoluções superiores a 16 *bits* e capazes de operar a taxas de amostragem superiores a 30 000 conversões por segundo, viabilizaram a representação digital de sinais áudio contemplando um conteúdo espectral bastante superior ao tradicionalmente associado a sinais de voz. Por outro lado, o aparecimento de Processadores Digitais de Sinal dedicados e cada vez mais poderosos, permitiram a implementação em tempo real de algoritmos de processamento complexos. A manipulação e processamento de sinais áudio tornou--se assim realizável no plano digital e com precisão matemática definida, evitando os erros e derivas do processamento analógico e permitindo, além disso, a aplicação de técnicas de detecção e correcção de erro não existentes no plano analógico.

No entanto, as vantagens decorrentes da tecnologia digital são possíveis à custa de uma acrescida necessidade de largura de banda para representar o sinal áudio no seu formato digital. O desenvolvimento de suportes admitindo grandes densidades de informação como os Discos Compactos ("CD") e as Bandas de Áudio Digital ("DAT") favorecem a disponibilidade, a preços reduzidos, de grandes larguras de banda necessárias para aplicações de estúdio ou comércio discográfico. Contudo, estas larguras de banda não estão ao alcance de uma gama de aplicações perspectivada para serviços de áudio digital que, nos últimos anos, tem vindo a ampliar-se e diversificar-se, com a consequente necessidade de explorar estratégias específicas de codificação [V5]. De facto, destacam-se, em particular, três áreas relevantes. As Redes Digitais com Integração de Serviços (RDIS/ISDN) contemplam serviços que se apoiam ou involvem áudio digital, como por exemplo, a difusão radiofónica, a teleconferência, a televisão de alta definição ou o acesso a bancos de registos sonoros ou musicais. A difusão directa de áudio digital por satélite tem naturalmente a ganhar com uma melhor utilização do espectro. Mais

recentemente, o conceito de ambiente Hipermédia consagra o áudio digital como um elo crucial da dinâmica interactiva. O interesse comum das diferentes áreas condiciona a orientação da investigação no domíno da codificação de áudio digital, no sentido da utilização de larguras de bandas cada vez mais reduzidas, sem sacrifício subjectivo da qualidade.

É neste último contexto que a contribuição implícita nesta tese procura enquadrar-se. Salvo referência em contrário, a palavra *áudio* referir-se-á ao espectro audível, abarcando frequências desde 20Hz até 20 000Hz.

1.2 Objectivos

Numa pré-publicação de Junho de 1991 do Grupo de Teste do CCIR e referente aos requisitos para sistemas de codificação de áudio digital funcionando a baixas taxas de informação, especificam-se vários objectivos a serem observados por tais sistemas. Os dois primeiros e porventura mais relevantes, aludem à Qualidade Básica de Áudio e à Qualidade da Imagem Estereofónica. O primeiro especifica que << ... *a qualidade do som reproduzido, depois de descodificado, deve ser subjectivamente indistinguível da qualidade do de Disco Compacto, para a maioria dos tipos de material áudio* ... >>. O segundo refere que << ... *a qualidade da imagem sonora deve ser conservada* ... >>.

A pré-publicação referida enquadra-se numa intensa actividade de investigação e normalização desenvolvida no âmbito de três grupos de estudo ligados às organizações ISO, CCIR e CCITT. Interesses comerciais têm também actuado como um catalizador, não totalmente imparcial, na evolução deste processo. O objectivo global é a compressão sem perdas subjectivas ou, por outras palavras, a *codificação perceptual* de um canal aúdio. Pretende-se com este conceito minimizar a taxa de informação, conservando a qualidade subjectiva, *i.e.*, conservando a qualidade percebida pelo ouvido humano, depois do sinal reconstruído.

A taxa de informação de um canal monofónico de Disco Compacto é cerca de 700Kbit/s. Diferentes algoritmos e sistemas de codificação permitem, hoje, a compressão transparente de um canal monofónico em 128Kbit/s. Porém, o desempenho destas soluções degrada-se significativamente quando se tenta reduzir a taxa de informação aquém deste último valor. Este sintoma é comum para todas as soluções conhecidas, sugerindo este facto que o limite de compressão transparente

está a ser assimptoticamente atingido. Por outro lado, ainda não foi publicada qualquer estratégia que apresente ganhos de compressão na codificação multicanal. Neste caso, e considerando em particular o cenário estereofónico, a solução usualmente sugerida consiste em agrupar canais monofónicos, codificados independentemente.

A codificação transparente a taxas mais reduzidas, sobretudo a 64Kbit/s, que corresponde a um canal B de RDIS, é um objectivo muito tentador e certamente na mente das várias entidades envolvidas na área de áudio digital. Contudo, os resultados até agora publicados são, deste ponto de vista, insuficientes. De uma forma geral, tem-se também concluído que, associando dois canais codificados independentemente e a taxas de informação próximas do limite de um canal B de RDIS por canal áudio, revelam-se novos problemas cuja expressão mais evidente é a criação de artefactos graves na imagem estereofónica, para além de distorções mais subtis.

Ao som, *i.e.*, à luz dos dois objectivos referidos no início deste parágrafo, o trabalho e resultados expostos nesta dissertação pretendem contribuir para a obtenção de maiores ganhos na compressão transparente de áudio digital, revendo por um lado, a dimensão *intracanal* do problema e desenvolvendo, por outro lado, conceitos base e um modelo para a dimensão *intercanal* do problema, concretizada na associação familiar de dois canais que é o par estereofónico.

1.3 Estrutura da Tese

O problema da compressão de áudio digital é abordado no capítulo 2. Procura-se fornecer uma panorâmica dos vários conceitos base do problema, fazer uma breve referência a diversas estratégias de codificação vulgarmente usadas e identificar o actual estado-da-arte.

O capítulo 3 ocupa-se da análise pormenorizada dos vários aspectos ligados à percepção auditiva. Para além de uma caracterização descritiva, procura-se também identificar modelos que possam, de alguma forma, caracterizar quantitativamente aspectos mais relevantes. Estes modelos revestem-se de extrema importância pois são directamente incorporados nos algoritmos de codificação e são responsáveis pelos ganhos de codificação.

As contribuições desta dissertação encontram-se expostas no capítulo 4. Os tópicos julgados mais relevantes e producentes, apresentados nos capítulos anteriores, são retomados, desenvolvidos e implementados num sistema codificador/descodificador, na forma de um simulador escrito em linguagem FORTRAN. Este sistema envolve contribuições originais desde um algoritmo optimizado para cálculo de um banco de filtros/transformada, até à organização física da informação espectral que interliga codificador com descodificador. Porém, a maior ênfase é colocada na contribuição para modelos melhorados de codificação monofónica e no desenvolvimento do primeiro modelo para codificação estereofónica. Aspectos de sincronização, atraso de codificação e complexidade de implementação são abordados, mas aspectos de protecção de erro no canal de transmissão, não foram contemplados.

A fim de avaliar a qualidade subjectiva do processo de codificação desenvolvido, foram realizados vários testes de audição. O planeamento, preparação e condução dos testes são descritos no capítulo 5. Apresentam-se também os respectivos resultados e conclusões finais.

O capítulo 6 conclui o texto da dissertação. Resumem-se as conclusões globais do trabalho desenvolvido e dos resultados alcançados e perspectivam-se algumas orientações para a sua evolução num futuro próximo.

2
COMPRESSÃO DE ÁUDIO DIGITAL

2.1 Introdução

A compressão de áudio digital inspirou-se naturalmente na actividade e resultados da investigação na área de compressão de voz [V1-V4][V7]. Por um lado esta precedeu a investigação na área de áudio digital que só começou a ganhar expressão em meados de 1979 [A20]. Por outro lado a abordagem do problema da compressão na área de áudio começou essencialmente por adoptar estratégias previamente desenvolvidas na área de voz [V6]. A evolução das duas áreas diferenciou-se sobretudo na década de oitenta. Contudo, é possível encontrar, hoje, aspectos comuns relativos aos objectivos, às particularidades e às concretizações específicas de cada área.

Em termos gerais, os objectivos coincidem: dada a representação digital de um sinal analógico, referida ao mínimo múltiplo da frequência de *Nyquist*, pretende-se maximizar a redução da taxa de informação, de modo a transmitir e reconstruir o sinal, de acordo com uma dada qualidade pretendida para todo o processo.

A acção de reduzir a taxa de informação de um sinal digital, ou seja, a compressão de um sinal, supõe a exploração de algumas características explícitas do sinal ou implicitamente associadas e que viabilizem tal redução. Estas características, que representam as particularidades de que se tenta tirar partido, são essencialmente a natureza da *fonte* que produz os sinais a codificar, as características estatísticas e espectrais dos próprios sinais e, finalmente, a natureza do *poço*, isto é, o sistema que irá captar e absorver o sinal.

Para a área de compressão de voz a *fonte* é o mecanismo natural de produção de voz, envolvendo sobretudo as cordas vocais e o tracto vocal afectado pela língua, lábios e nariz [P20]. Para a área de compressão de áudio, a *fonte* é qualquer sistema capaz de produzir sinais acústicos dentro dos limites de percepção do ouvido humano. Em cada circunstância, as características estatísticas e espectrais dos sinais são as determinadas pelas respectivas *fontes*. Em ambas as áreas, o *poço* é o sistema auditivo humano.

Para cada área a estratégia específica de actuação sobre as várias particularidades para conseguir a redução da taxa de informação é concretizada num algoritmo de codificação que, no contexto desta tese, se assume sinónimo de

algoritmo de compressão. A sequência deste capítulo pretende fornecer uma panorâmica das estratégias usadas actualmente na compressão de áudio e referir as premissas e os objectivos de qualidade subjacentes. A referência à compressão de voz visa sobretudo afirmar a coerência evolutiva dos conceitos e estratégias.

2.2 Codificação de Fonte e de Poço

No contexto da codificação de voz é usual distinguir entre codificação de *onda* e codificação *paramétrica* [V1][V2]. Esta última, associada a taxas de informação mais reduzidas, baseia-se na parametrização rígida do sinal de voz, adoptando para isso um modelo para o sistema natural de produção de voz, formado pelas cordas vocais e tracto vocal. O modelo procura sintetizar as ocorrências de sons vozeados e não-vozeados, silêncios e padrões de formantes. O critério de qualidade associado a este tipo de codificação é sobretudo a inteligibilidade do sinal produzido. A identificação da forma de onda sintetizada com o sinal original é bastante grosseira pelo que o outro tipo de codificadores, ditos codificadores de *onda*, procura extrair redundância do sinal, mas observando uma medida de fidelidade à forma de onda original. Por vezes aspectos de percepção auditiva são também considerados, no entanto, a maior ênfase é colocada na exploração das características do sinal de fonte.

O espectro médio e típico dos sinais de voz, para além de só conter energia significativa na banda de frequências até 5KHz, revela-se não plano, o que denota redundância estatística [V6]. Por outras palavras, existe correlação entre as amostras temporais do sinal. Este facto revela que há algum grau de predizibilidade do sinal que pode ser convertido em ganho de codificação, ou seja, é possível usar uma taxa de informação inferior à associada à representação Codificação por Modulação de Impulsos (PCM), para transmitir o sinal com a mesma relação sinal--ruído inicial. O ruído inicial é obviamente o ruído de quantificação PCM original, que se supõe ser branco.

A extracção de redundância linear pode ser efectuada no domínio dos tempos, nomeadamente através de PCM diferencial (DPCM), em que se faz uso de um preditor linear, para depois quantificar de forma conveniente o erro de predição. Prova-se [V6] que o ganho máximo desta estratégia, no sentido de minimização do erro quadrático médio do erro de predição, conduz a um erro de reconstrução que

também é branco. Porém, reconhece-se [V2] que o espectro plano do ruído de reconstrução não é perceptualmente o mais adequado.

A estratégia DPCM com malha aberta (D*PCM) dispõe o ruído de reconstrução com um perfil espectral aproximadamente idêntico ao sinal original mas a alguns decibeis abaixo deste. De uma forma geral, o ruído de reconstrução resulta mais concentrado em regiões de maior energia do sinal como por exemplo, as regiões de formantes. Esta estratégia revela alguma melhoria qualitativa, apesar do ganho de codificação ser inferior ao de DPCM. A importante conclusão a retirar é que existe um nítido efeito de *mascaramento* do ruído de reconstrução na vizinhança de zonas espectrais com mais energia de sinal.

Uma terceira estratégia do tipo DPCM, designada por codificação com realimentação de erro (NFC), procura estabelecer um compromisso entre as duas situações referidas anteriormente. O ganho de codificação, no mesmo sentido de mínimo erro quadrático médio, é também aqui inferior ao verificado em DPCM. Contudo, o ruído de reconstrução - ou quantificação - é modulado de tal forma que é mais eficientemente mascarado pelo próprio sinal, o que é equivalente a dizer que é menos audível. Este facto revela uma outra importante conclusão: o ouvido humano não é igualmente sensível em todo o espectro, ao ruído de quantificação.

Qualquer uma das estratégias anteriores pode assumir formas complexas de implementação para incluir adaptatividade a variações acentuadas do sinal, ou seja, a não-estacionaridades do sinal. Em termos gerais, há a necessidade de actualizar o filtro preditor a partir do cálculo dos coeficientes da função autocorrelação, tomando um excerto curto do sinal. O ganho de codificação resulta melhorado; porém, há a necessidade de transmitir adicionalmente informação lateral.

A redundância linear pode ser efectuada também no domínio das frequências, usando a codificação por transformada. Este contexto revela as mesmas conclusões anteriores quanto à importância de uma estratégia de quantificação perceptualmente adequada [V1]. A codificação por transformada é abordada no parágrafo seguinte.

As várias estratégias de codificação apresentadas foram essencialmente optimizadas para se adaptarem, com diferentes graus de fidelidade, às características da fonte de sinal que se assumiu tratar de voz. Podemos portanto designá-los de codificadores de *fonte*. Contudo, os sinais de voz são sinais muito

específicos pois são caracterizáveis, no curto e longo termos, por comportamentos e parâmetros bem definidos. Esta situação não ocorre para os sinais designados genericamente no contexto desta tese por áudio. De facto, áudio ou música pode ser todo o arranjo possível e imaginável de sons naturais e/ou sintetizados. Assim, não se estabelece qualquer premissa restritiva quanto aos padrões estatísticos e espectrais destes sinais.

Considerando eventualmente adaptitividade na estratégia de codificação por predição linear, ou mesmo por transformada, como se referirá em breve, e atendendo a que o espectro áudio, extremamente dinâmico, pode assumir qualquer perfil, os ganhos de extracção de redundância não são suficientes para justificar a complexidade de implementação exigida [A25] que, em certos cenários mais críticos, poderia ser simplesmente impraticável.

De uma forma geral, pode concluir-se que a codificação de fonte exibe sérias limitações se for aplicada a sinais de áudio. De facto, mesmo com ganhos sub-óptimos, no sentido dos mínimos quadrados, é possível conseguir melhores resultados perceptuais adaptando a quantificação às características de percepção do sistema auditivo. As propriedades de mascaramento surgem assim como aspectos promissores em novas filosofias de codificação. É portanto imperioso conhecer os atributos, propriedades e limitações do sistema auditivo, ou seja, do poço do processo de codificação. Dado que esta é a certeza básica na codificação dos sinais áudio, procurar-se-á adaptar todo o processo de codificação às características do poço, pelo que se passará a designar de codificação *perceptual* ou codificação de *poço*.

2.3 Redundância e Irrelevância

O efeito de *mascaramento*, que resumidamente se pode exprimir como a ofuscação ou completa ocultação de um sinal com baixa energia e situado na mesma região espectral de um sinal com maior energia, pode ser desejável, se o primeiro for o ruído de quantificação e o último uma componente do sinal original. Pode também ser indesejável se ocorrer a situação inversa. Surge assim a necessidade de controlar rigorosamente a disposição do ruído de quantificação ao longo do espectro, de forma que a *quantificação perceptual* numa região espectral não vá afectar adversamente regiões vizinhas, comprometendo o desempenho global do codificador.

As técnicas de codificação preditiva, em condições de ganho menos favorável (no sentido dos mínimos quadrados), usam uma realimentação controlada do erro de quantificação e filtros de ênfase espectral, para realizarem a modulação espectral do ruído de quantificação. Porém, o controlo desta modulação por este processo não é muito efectivo, sobretudo a muito baixas taxas de codificação, em que o desempenho deste tipo de codificadores se degrada significativamente [V6][V7].

Melhores perspectivas são oferecidas por um outro processo de codificação que, tal como a codificação preditiva, também fornece informação não correlacionada a partir de um sinal com redundância. Este processo de codificação opera no domínio das frequências e tem nesta tese importância fundamental por diversas razões que irão oportunamente ser apresentadas. Em termos muito genéricos, o sinal é decomposto em várias bandas e cada uma delas tem um tratamento diferente, em função da sua importância objectiva, avaliada a partir da medida de planura espectral [V6], e subjectiva, incluindo ponderações psico-acústicas [P5]. Deste último ponto de vista, é útil dispor da maior resolução espectral possível para analisar o sinal e para realizar uma quantificação mais "fina", o que torna a codificação no domínio das frequências mais adequada e natural.

A codificação no domínio das frequências pode ser realizada através de dois tipos de implementação, designados historicamente por sub-bandas ou transformada. Ambos procedem a uma análise do sinal, decompondo-o em várias componentes espectrais. São, por isso, só diferentes formulações matemáticas do mesmo processo analítico [T7].

Se o objectivo da codificação no domínio das frequências for a atribuição óptima de *bits* pelos vários coeficientes espectrais, minimizando, no sentido dos mínimos quadrados, o erro de reconstrução, ou seja, o ruído de quantificação assumindo um canal de transmissão sem erros, prova-se [V6] que a solução deste problema fornece variâncias dos erros de quantificação idênticas para todos os coeficientes. Considerando quantificadores lineares, isto supõe o uso do mesmo passo de quantificação para todos os coeficientes, embora o comprimento da palavra (em *bits*) varie de coeficiente para coeficiente de acordo com a sua variância. Em consequência, o ganho de codificação relativamente ao formato PCM é limitado superiormente pelo inverso da medida de *planura espectral.* Este ganho é o quociente entre a média aritmética e a média geométrica das variâncias dos

coeficientes espectrais, no caso da codificação por transformada. No caso da codificação por sub-bandas, assumindo iguais larguras de banda, o ganho é a mesma relação das variâncias das amostras de cada sub-banda. Dado que o ganho cresce com o número de sub-bandas, eleva-se este número até ao seu limite máximo mas esta circunstância é precisamente a da codificação por transformada. Esta solução é, portanto, a que se reveste de maior interesse.

Prova-se também que o ganho objectivo teórico, considerando diferentes transformadas com a mesma dimensão dos vectores de base, é mais rapidamente aproximado pela transformada de Karhunen-Loève (KLT) que tem a propriedade de diagonalizar a matriz auto-covariância do sinal. Os coeficientes espectrais são assim completamente descorrelacionados e podem, por isso, ser quantificados de forma independente e efectiva. Estes apresentam também uma variância cuja média geométrica é minimizada, o que significa melhor ganho de codificação. Porém, os vectores de base desta transformada dependem do sinal. É por isso necessário recalculá-los sempre que se codifica um segmento diferente do sinal. Para que a operação de transformação seja tratável, é necessário que a dimensão da transformação seja finita, ou seja, que o comprimento do segmento de sinal transformado seja finito. Contudo, apesar de se obterem coeficientes completamente descorrelacionados para um segmento particular, pode naturalmente haver correlação entre coeficientes homólogos de segmentos adjacentes de sinal. O limite teórico de redundância do sinal não é assim completamente extraído. Tomando os ganhos de compressão da transformada óptima (KLT) como referência, é possível encontrar outras transformadas, independentes do sinal, que forneçam ganhos semelhantes para a mesma dimensão da transformação e que eliminem algumas dificuldades daquela, nomeadamente a necessidade de calcular em cada transformação os seus vectores base e a inexistência de um algorimto rápido de cálculo.

A transformada discreta em cosseno (DCT) tem sido usualmente identificada como a que fornece melhores ganhos de extracção de redundância nas áreas de voz [V7] e imagem. Na área de áudio, uma DCT modificada [T3][T4] tem conseguido uma grande aceitação [A13][A16-A19][A30], sobretudo devido a propriedades interessantes de selectividade em frequência e eliminação dos problemas clássicos de ruído na fronteira entre segmentos transformados. Esta transformada será abordada em detalhe no parágrafo 4.3.

A discussão anterior assumiu o significado da compressão no sentido de extracção de redundância do sinal. Referiu-se que esta era identificada pela disposição não plana do espectro médio do sinal, que se considerou tratar essencialmente de voz. Porém, ainda neste contexto, conclui-se [V3][V4] que, em vez de projectar a atribuição óptima de *bits* pelos vários coeficientes espectrais, definida a partir do espectro médio do sinal, se se redefinir essa atribuição optimizada de acordo com o espectro do sinal avaliado em períodos de tempo curtos, então poder-se-á mehorar o ganho. Esta técnica de adaptividade, idêntica em termos de objectivos à referida no parágrafo 2.2 para o caso da codificação preditiva, permite a obtenção de ganhos substanciais, apesar de ser necessário incluir informação lateral para caracterizar a adaptividade. A utilização e a necessidade de adaptividade é natural pois o sinal de voz é basicamente um sinal não-estacionário.

Foi referido que as condições de maximização do ganho objectivo de compressão implicam um ruído de quantificação branco. Tal como apontado para os codificadores preditivos, também na codificação por transformada se inclui uma coloração do ruído de quantificação para se obter melhores resultados perceptuais [V3][V6]. Isto é conseguido prevendo um parâmetro de ponderação na equação que define a atribuição óptima de *bits* por coeficiente. Este parâmetro pondera a importância perceptual do ruído de quantificação em função da frequência. É possível também aqui modular o ruído de quantificação com uma forma situada entre duas disposições extremas. Uma, corresponde naturalmente à circunstância de ganho máximo objectivo em que a ponderação não é activada. Neste caso o ruído de quantificação é plano (branco) e a relação sinal-ruído calculada ao longo do espectro tem a forma do próprio espectro. A outra corresponde à circunstância em que todos os coeficientes espectrais são codificados com o mesmo número de *bits*. Neste caso o ruído de quantificação tem a forma do espectro do sinal. Consequentemente, a relação sinal-ruído é constante ao longo do espectro.

A vantagem decorrente da codificação por transformada, reside não só no facto de com relativa facilidade poder determinar-se adaptativamente a melhor situação intermédia de ponderação mas, sobretudo no facto de poder definir-se uma variação individualizada da ponderação de cada coeficiente espectral. Este cenário de codificação extremamente flexível e orientado sobretudo para as vantagens de uma codificação perceptual, requer naturalmente um conhecimento profundo das características do sistema auditivo. A configuração espectral arbitrária do ruído de quantificação é viabilizada pela possibilidade de efectuar uma quantificação tão

precisa quanto a resolução espectral da transformada, o que corrobora o interesse e a adequação desta para a codificação perceptual.

Em conclusão, os antecedentes históricos da codificação perceptual revelam que, mesmo para os sinais de voz cuja compressão envolveu sobretudo conceitos de redundância estatística ou codificação sem perdas, é notório o efeito de mascaramento do ouvido humano, o que vem confirmar que, em circunstâncias de ganho objectivo menos favorável, pode-se obter uma codificação perceptualmente preferível. O efeito de mascaramento manifesta-se neste contexto pela reduzida audibilidade ou completa inaudibilidade, em certas zonas espectrais, de maiores quantidades do ruído de quantificação camufladas ou "apagadas" devido à presença de componentes do sinal geralmente com maior energia. Visto de outra perspectiva, isto pode ser interpretado como a existência de componentes do sinal *irrelevantes* do ponto de vista de percepção, devido ao efeito mascarante ou ofuscador de outras componentes do sinal. Como tal, a *extracção de irrelevância* de um sinal surge assim como um factor complementar e importante para a obtenção de ganhos adicionais de compressão. Um corolário desta abordagem é claramente a inadequação do ganho objectivo tratado até agora para identificar o limite do ganho de compressão que é, aliás, um conceito que deixa de ter sentido pois passa a ser limitado só pelo completo conhecimento das tolerâncias e limitações perceptuais do sistema auditivo. Este estudo é genericamente referido como o estudo da psico-acústica. Ó domínio completo das características do sistema auditivo é um desafio bastante arrojado dada a complexidade e dificuldade inerentes. O capítulo 3 aborda esta temática.

O primeiro projecto de um critério de codificação de voz apoiado em parâmetros subjectivos ou perceptuais foi desenvolvido por Schroeder, Atal e Hall [V4]. Em vez do ganho de codificação fundamentado na minimização do erro quadrático médio do ruído de quantificação, criou-se um objectivo de codificação fundamentado na minimização da audibilidade subjectiva do ruído de quantificação relativamente à audibilidade do sinal. Esta medida de audibilidade relativa usa um modelo simplificado do sistema auditivo que procura simular a análise do sinal por este efectuada. O efeito de mascaramento é representado por um modelo psico-acústico que fornece um limiar de (in)audibilidade do ruído de quantificação em toda a largura espectral. Na medida da capacidade oferecida pela taxa de informação disponível, o ruído de quantificação é modulado ao longo do espectro de acordo com o limiar de mascaramento calculado. Os resultados relatados incluem a

codificação transparente de voz sintética, com 3.5KHz de largura de banda, em 7Kbit/s.

O critério de codificação referido marca uma abordagem diferente do problema da compressão e estabelece a base dos algoritmos de compressão de áudio digital.

Os sinais de áudio caracterizam-se por uma gama dinâmica e um espectro que exploram quase toda a capacidade do sistema auditivo. O desenvolvimento de codificadores perceptuais para áudio digital assenta em duas áreas de investigação. Por um lado, é necessário investigar e caracterizar quantitativamente os diversos mecanismos de conversão de informação acústica em sensação e impressão auditivas. Esta área é do domínio da psico-acústica. Por outro lado, é necessário investigar o desenvolvimento, incorporação e validação de modelos auditivos, partindo de informação da psico-acústica, e definir algoritmos de compressão perceptualmente efectivos, isto é, algoritmos que codifiquem sinais de forma perceptualmente inofensiva e usando a taxa de informação mais baixa possível.

Em síntese, no contexto da compressão de áudio, os ganhos potenciais podem exprimir-se em dois vectores não necessáriamente ortogonais. Por um lado, a *extracção de redundância* continua a ser pertinente pois a maior parte dos sinais expectáveis promete alguma correlação estatística no domínio dos tempos. A transformação para o domínio das frequências, usando vectores de base ortogonais e fixos, proporciona sempre alguma descorrelação da informação. Por outro lado, a *extracção de irrelevância* que depende do grau de conhecimento do sistema auditivo, é susceptível de garantir a componente mais importante do ganho total de compressão. Neste espírito, o ganho de compressão será a partir de agora considerado como sendo a razão simples entre a taxa de informação PCM original e a taxa final do sinal comprimido.

2.4 Estratégias de Codificação

2.4.1 Objectivos

De uma forma geral, os algoritmos de codificação que serão referidos nos parágrafos seguintes visam comprimir sinais áudio de acordo com uma dada relação qualidade/custo.

Por sinais áudio entende-se todo o tipo de sinal representável em formato de Disco Compacto. As amostras digitais são sempre expressas em palavras com o comprimento de 16 *bits*. Assim, pode-se associar a estes sinais os valores de 90dB para a sua gama dinâmica e 98dB para a relação sinal-ruído. A frequência de amostragem pode depender de caso para caso; contudo, há três frequências normalizadas. A frequência de 32KHz associa-se ao contexto de difusão radiofónica ou outras aplicações de difusão e contempla a largura de banda de sinal tipicamente associada a "FM", ou seja, 15KHz. A frequência de 44.1KHz identifica-se com a representação digital de Disco Compacto e a frequência de 48KHz está ligada a aplicações de estúdio e áudio profissional. Apesar destas duas últimas admitirem largura de banda de sinal até 20KHz, a frequência de amostragem 48KHz é a que será a doptada para a discussão do desenvolvimento exposto no capítulo 4.

O critério de qualidade preconizado por vários autores pode também variar. De uma forma geral, procura-se uma qualidade não inferior à do familiar "FM" e ao mais baixo custo possível, quer em matéria de taxa de transmissão, quer em matéria de complexidade dos módulos codificador e descodificador.

Dado que as gradações dos conceitos de qualidade e complexidade de implementação podem ser definidos e refinados à posteriori, o único critério (objectivo) de qualidade para o algoritmo de codificação considerado no âmbito do capítulo 4 é o referido no parágrafo 1.2, isto é: qualquer que seja o sinal e para qualquer potencial ouvinte, o sinal áudio depois de codificado e descodificado deve ser indistinguível do sinal original. Em particular, o processo de compressão deve ser transparente mesmo para indivíduos que em resultado de treino específico tenham desenvolvido uma acuidade auditiva acima do normal, ou seja, mesmo para os exigentes "ouvidos de ouro". Por outro lado, a função custo que se quer minimizar, assume-se depender de uma única variável: a taxa de informação do sinal comprimido.

No contexto desta tese, ruído representa o elemento que se pretende dominar e controlar, ou seja, o ruído de quantificação. Por isso, a referência a *relação sinal--ruído* significa, na verdade e implicitamente, relação *sinal-ruído de quantificação*. Por outro lado, como se pretende que este ruído de quantificação seja efectivamente mascarado pelo sinal, aquela relação também é sinónima de *relação sinal-ruído mascarado* (*signal-to-mask-ratio*, SMR).

2.4.2 Domínio dos Tempos

As técnicas de codificação no domínio dos tempos foram as primeiras a surgir e visavam uma redução simples da taxa de informação, sobretudo com vista a aplicações de difusão directa por satélite.

A quantificação uniforme em PCM com resolução de 14 *bits* era considerada suficiente para tornar o ruído de quantificação imperceptível. O espectro de sinal admitido era 15KHz e a taxa de amostragem adoptada era 32KHz.

Com este dados e partindo da constatação simples que os erros de quantificação são muito mais notórios em passagens musicais de muito baixa amplitude do que em passagens de alta amplitude, definiu-se uma estratégia de codificação em que o erro de quantificação é aproximadamente linear com a amplitude das amostras temporais. A Recomendação J.41 do Comité Internacional Consultivo de Telefones e Telégrafos (CCITT) é ilustrativa das técnicas mais usadas: compressão instantânea e compressão quase instantânea (NICAM). A primeira tem associada uma tabela de 11 segmentos que comprime cada amostra temporal de 14 para 11 *bits*, de acordo com uma lei logarítmica. Esta lei mantém a relação sinal-ruído essencialmente constante em toda a gama dinâmica do sinal. A segunda técnica considera blocos de 32 amostras e de acordo com a maior amplitude nesse bloco, é seleccionada uma dentre 5 possíveis escalas de ganho que comprime os 14 *bits* de cada amostra do bloco em 10 *bits*. Considerando a informação de protecção de erro e sincronização, as duas técnicas comprimem a taxa original do sinal de 448Kbit/s em 384Kbit/s, por canal.

Uma outra técnica distinta é um caso particular da técnica DPCM adaptativa (ADPCM), em que o resíduo de predição é expresso num só bit e em que a frequência de amostragem é bastante elevada relativamente à frequência de Nyquist original. Trata-se da codificação Delta adaptativa (ADM). Tipicamente, esta técnica comprime os 448Kbit/s por canal em 330Kbit/s por canal. Comparativamente, a codificação ADM revela um desempenho ligeiramente inferior ao da codificação NICAM e ao da compressão logarítmica, sobretudo para as frequências mais baixas [A14].

As técnicas anteriores caracterizam-se por introduzir um ruído de quantificação com espectro plano, apesar de poder ter energia variável na escala dos tempos. No entanto, este espectro pode ser modulado por filtros de pré-ênfase e

pós-ênfase como, por exemplo, os definidos pela recomendação J.17 do CCITT. A inclusão destes filtros melhora de facto a qualidade subjectiva da codificação dado que a energia do ruído de quantificação é diminuída na zona espectral do sinal onde a energia é tipicamente mais reduzida, isto é, às altas frequências.

Uma outra técnica de predição linear aplicada a áudio [A15] utiliza os conceitos mais recentemente introduzidos na área de voz como o codificador Multipulso (Multipulse LPC) ou o codificador de predição linear excitado por impulsos (CELP) [V5]. Em vez da tradicional distinção entre segmentos vozeados e não vozeados, são aplicados conceitos de correlação de curto termo e correlação de longo termo. Dada a estrutura em malha fechada destes preditores, é efectuada uma análise por síntese com ponderação perceptual, de que resultam os impulsos de excitação que modelizam a forma de onda codificada. Usando adicionalmente codificação por entropia, consegue-se comprimir a taxa de informação original do sinal áudio para 128Kbit/s. Porém, o processo não é transparente e a melhoria de qualidade tem uma evolução saturada com o aumento da complexidade do sistema.

De uma forma geral, estes métodos de compressão que manipulam o sinal em toda a sua largura de banda, não reduzem significativamente a taxa de informação nem exercem uma ponderação perceptual de forma rigorosa e eficiente. A qualidade de codificação destas técnicas não pode assim ser qualificada de transparente.

2.4.3 Domínio das Frequências

2.4.3.1 Limiar de Mascaramento

A codificação no domínio das frequências visa realizar uma quantificação do sinal adaptada ao processamento analítico efectuado pelo sistema auditivo.

Como será detalhadamente exposto no capítulo 3, a análise auditiva recorre a uma decomposição de tipo logarítmica do espectro do sinal. Esta decomposição estrutura-se num conjunto de bandas, designadas por *bandas críticas* [P4], onde se pode identificar um comportamento psico-fisiológico regular e análogo. A particularidade mais relevante é que definem a mínima resolução espectral que permite caracterizar qualitativamente e também quantitativamente, embora de forma grosseira, o efeito de mascaramento. Por exemplo, o mascaramento de um

pequeno sinal devido a um tom puro de amplitude elevada é máximo e constante, até que a separação em frequência destes dois sinais ultrapasse o limite da banda crítica em que ambos se encontram. Por outro lado, a detecção de um sinal acústico ocorre só se a sua energia no interior de uma dada banda crítica exceder um nível definido, independentemente de se tratar de um tom puro, banda de ruído ou uma combinação dos dois [P9].

Com base nestes conceitos, os diversos métodos de codificação no domínio das frequências que serão sumariamente descritos nos dois parágrafos seguintes, dividem e analisam o espectro do sinal em bandas relacionadas com as bandas críticas. Para cada banda e de acordo com regras que variam de caso para caso, é calculado um limiar de mascaramento (*threshold of masking*) que representa a máxima quantidade de ruído, confinado nessa banda, que pode ser introduzido de forma imperceptível ou benigna. Seguidamente, a operação de quantificação em cada banda é realizada de modo que a energia do ruído de quantificação resultante não ultrapasse o limiar de mascaramento previamente calculado.

Todo este processo é extremamente dependente das características espectrais e da dinâmica temporal do sinal. O primeiro aspecto está ligado aos efeitos de mascaramento no domínio das frequências referidos até agora e que supõem condições permanentes do sinal. O segundo está ligado a efeitos de mascaramento oberváveis também no domíno dos tempos [P5]. Designam-se por efeito de *pré--macaramento* e efeito de *pós-mascaramento* e serão pormenorizadamente descritos no capítulo 3. É contudo oportuno referir, desde já, que o efeito de pré-mascaramento refere-se à circunstância em que um sinal de baixa amplitude e duração limitada é mascarado por um sinal de elevada amplitude que surge tipicamente poucos milissegundos depois daquele. À primeira vista, esta é uma situação estranha se não se considerar factores de latência e velocidade de processamento aos níveis mais altos do sistema auditivo. O efeito de pós--mascaramento refere-se à circunstância em que um sinal de baixa amplitude e duração limitada é mascarado porque surge imediatamente após a cessação de um sinal de amplitude muito superior.

A maior ou menor adaptatividade do método de codificação a estes parâmetros, é uma outra variante dos diversos algoritmos que se passam a abordar.

2.4.3.2 Codificação por Sub-bandas

Os codificadores de sub-bandas têm a estrutura genérica representada na Fig. 2.1. O sinal digital é inicialmente decomposto em várias bandas através de um banco de filtros.

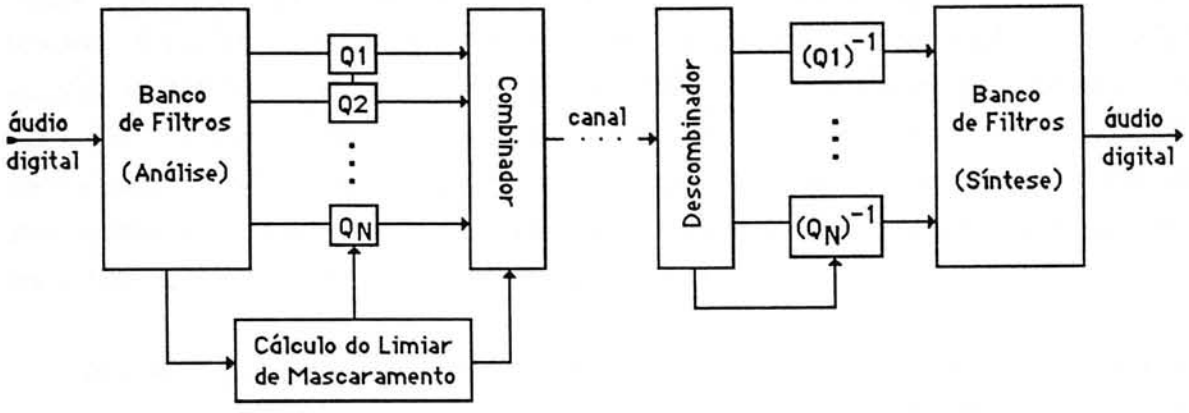

Figura 2.1: Diagrama de blocos de um codificador e descodificador por sub-bandas. Os quantificadores assinalados por Q incluem também a operação de codificação.

O tipo de filtro mais usado é o Filtro Espelho em Quadratura (QMF) que por sua vez combina Filtros de Meia Banda (*Half-Band Filters*). A forma básica dos QMF garante linearidade de fase e cancelamento de sobreposição espectral. Porém, não asseguram reconstrução exacta das bandas do sinal previamente isoladas [T9]. A reconstrução perfeita é necessária pois pretende-se obter o sinal original na ausência de quantificação e de uma perspectiva de controlo, pretende-se que toda a distorção gerada seja devida unicamente à operação de quantificação. Uma versão modificada dos QMF, designada de *Lattice*-QMF, soluciona o problema da reconstrução perfeita. Os Filtros Digitais de Onda (*Wave Digital Filters*) e os Filtros Conjugados em Quadratura (*Conjugate Quadrature Filters*) também proporcionam reconstrução perfeita; porém, estes últimos envolvem uma complexidade computacional dupla da dos *Lattice*-QMF. Um projecto de banco de filtros adaptado aos limites das bandas críticas é proposto em [T5].

Um bloco independente monitora as características espectrais e temporais do sinal para gerar o limiar de mascaramento que governa a quantificação em cada sub-banda. A monitoração do sinal pode usar as próprias amostras de cada sub--banda, ou então pode recorrer a informação obtida através de uma outra estrutura analítica como por exemplo, uma transformada. Em certos casos, pode-se usar

simplesmente um padrão característico médio e fixo de mascaramento. Por exemplo, a Recomendação G722 do CCITT envolve quantificação DPCM com uma atribuição fixa de *bits* nas diversas sub-bandas.

Os diversos blocos de quantificação e codificação utilizam o limiar de mascaramento para introduzir ruído perceptualmente inofensivo, em cada sub--banda. De uma forma geral, além de amostras quantificadas, aqueles blocos podem também fornecer informação lateral devido a estratégias de codificação PCM adaptativa (APCM) [V6], isto é, do tipo NICAM. Esta informação é combinada com outra informação lateral devida ao limiar de mascaramento, informação de sincronização e protecção e é depois enviada para o receptor. Dado que o limiar de mascaramento é determinado no emissor, o receptor terá sempre uma estrutura mais simples pois só usa este limiar para reconstruir o sinal.

Um dos primeiros desenvolvimentos que se consagrou como referência para trabalhos subsequentes foi o trabalho publicado por Krasner [A.20]. O bloco de análise e síntese é implementado com um banco de filtros QMF que dividem o espectro até 15KHz, em 24 bandas. Os QMF estão dispostos numa organização de decomposição espectral sucessiva, de modo que as 24 sub-bandas aproximam por defeito os limites das bandas críticas. A atribuição dos *bits* de quantificação pelas várias bandas é fixa. A compressão em cada banda é do tipo APCM, a adptatividade consiste, por isso, na conservação da relação sinal-ruído ao longo da gama dinâmica das amostras respectivas. A relação sinal-ruído considerada adequada para mascaramento em cada sub-banda, é determinada a partir de uma curva em que tons de diversas frequências mascaram ruído confinado na banda crítica em torno de cada frequência. A adopção desta curva é uma opção pragmática, pois reconhece-se que bandas de ruído mascarando tons puros são muito mais eficientes e como tal, exibem menores valores de mascaramento em décibeis (dB). O comprimento de cada bloco APCM é de 8 amostras - para cada banda - para contornar as condições mais restritivas do efeito de pré-mascaramento. A codificação de um espectro musical com 15KHz de largura de banda é conseguido em 124Kbit/s. A codificação é referida como sendo transparente; porém, é apontado que os trechos musicais de teste não são muito "limpos" devido a algumas limitações técnicas na digitalização directa dos sinais.

Em 1987, uma nova abordagem da compressão por sub-bandas definiu novas referências [A7]. O espectro de 16KHz é subdividido em 24 sub-bandas. Estas são só uma aproximação grosseira aos limites das bandas críticas. A divisão é conseguida

com uma cascata simples de QMF de que resultam 16 sub-bandas de 500Hz de largura cada, na região até 8KHz e 8 sub-bandas de 1KHz de largura cada, na banda de 8KHz até 16KHz. São introduzidos novos modelos de mascaramento que se fundamentam no comportamento do ouvido humano, referido às bandas críticas [P5]. Além de efeitos de mascaramento intrabanda crítica são também considerados efeitos interbanda crítica. Considerando situações de mascaramento mais extremas, a relação sinal-ruído em cada sub-banda é tornada fixa e dependente somente da largura do filtro respectivo. O efeito de pré-mascaramento é o que exige maiores precauções e, por isso, os factores de escala da codificação APCM em cada sub-banda são actualizados mais rapidamente do que o normal, quando são detectadas condições de não-estacionaridade. É afirmado que sinais com 16KHz de largura de banda são codificados transparentemente em 160Kbit/s.

Um sistema similar [A21] é obtido linearizando o modelo psico-acústico usado em [A7]. Este modelo só contempla o mascaramento de ruído por tons puros e assume uma mesma forma para todo espectro. O número de bandas é 26 e é referido que se obtém codificação transparente em 220Kbit/s para sinais com 20KHz de largura de banda.

Os melhores resultados de compressão estão associados a um sistema recente, denominado MUSICAM [A28][A32] e que proporciona uma codificação transparente, à taxa de 128Kbit/s, para sinais áudio amostrados à frequência de 48KHz. O espectro do sinal é dividido em 32 sub-bandas de igual largura. O processo é conseguido através de uma estrutura eficiente pseudo-QMF, em que um filtro passa-baixo protótipo é deslocado para as frequências centrais de diversas sub-bandas [T9]. O limiar de mascaramento é obtido dinamicamente através do cálculo independente de uma Transformada Discreta de Fourier (DFT). A alta resolução espectral da transformada permite identificar a *tonalidade* do sinal, ou seja, os atributos tipo tom e tipo ruído do sinal. Isto é, permite localizar as componentes do som com comportamento sinusoidal (com coerência de amplitude e fase). As restantes componentes que não exibem comportamento coerente associam-se genericamente a ruído. Este tipo de componentes é o produzido, por exemplo, pelos sons não vozeados da fala. Dado que os dois tipos de componentes têm propriedades de mascaramento diferentes [P18], como já havia sido notado anteriormente, usa-se um modelo psico-acústico com duas expressões, uma aplicada a cada tipo de componente. O valor de mascaramento final é calculado por ponderação adequada destas expressões.

Blocos de 36 amostras de cada sub-banda são quantificados, isto é, têm uma nova atribuição de *bits*, de acordo com o limiar de mascaramento calculado. Internamente a cada bloco, sub-blocos de 12 amostras são codificados pela técnica APCM. Os factores de escala têm portanto uma frequência de actualização tripla da do limiar de mascaramento.

O sistema de codificação admite várias configurações em que a uma maior taxa de informação está associada, para a mesma qualidade, uma menor complexidade de implementação, ou vice-versa.

Um sistema particularmente interessante é contituído por uma estrutura híbrida [A18] que procura combinar as vantagens de codificação por transformada, com as da codificação por sub-bandas, de modo a explorar melhor as características analíticas do sistema auditivo. Como foi já esboçado, este exibe uma resolução espectral decrescente com o aumento da frequência e, correspondentemente, uma crescente resolução temporal.

A problemática é exposta da seguinte forma: se a resolução espectral é maximizada no caso da codificação por transformada, o limiar de mascaramento é calculado por esta via com melhor precisão e, adicionalmente, beneficia-se de um melhor ganho de codificação já que existe uma maior descorrelação das componentes espectrais. Porém, melhor resolução espectral é sinónimo de pior resolução temporal [T9], o que contribui para problemas, vulgarmente designados por pré-eco, resultantes da violação das condições de pré-mascaramento [A1]. A sugestão consiste em adoptar uma solução híbrida, envolvendo bancos de filtros (QMF), com a intrínseca melhor resolução temporal, para isolar bandas espectrais e envolver uma análise adicional por transformada em cada sub-banda. As sub--bandas iniciais, em número de 4, são assimétricas segundo potências de dois. Como são também usadas transformadas de diferentes resoluções, resulta globalmente uma menor resolução temporal (maior resolução espectral) às baixas frequências do espectro do sinal codificado. Inversamente, resulta uma maior resolução temporal (menor resolução espectral) às altas frequências do espectro. Deste modo, as resoluções adequam-se a uma conveniente avaliação do limiar de mascaramento. Por outro lado, adequam-se também a um melhor controlo do efeito de pré--mascaramento. De facto, dado que este é consequência de não-estacionaridades do sinal, ou seja, das suas componentes de maior frequência, a maior resolução temporal associada permite uma actuação mais "rápida" do algoritmo.

O limiar de mascaramento é calculado para cada transformada de acordo com métodos descritos em [A11]. As fronteiras das sub-bandas finais são alinhadas de acordo com as das bandas críticas, prefazendo um total de 25 sub-bandas no espectro de 20KHz. A taxa de informação para sinais amostrados a 48KHz é 64Kbit/s. Porém, testes de audição não confirmam ganhos qualitativos destes codificadores e para a mesma taxa, relativamente a outros usando exclusivamente transformada [A17]. Por outro lado, é apontada uma desvantagem significativa, associada com a maior complexidade computacional da estratégia proposta.

2.4.3.3 Codificação por Transformada

Os codificadores perceptuais por transformada têm a estrutura geral representada na Fig. 2.2. Os blocos de análise e síntese realizam a transformação directa do sinal para o domínio das frequências e a transformação inversa, respectivamente. O tipo de tranformada usada varia desde a DFT até à Transformada Discreta em Cosseno (DCT). A escolha depende essencialmente da selectividade espectral pretendida, da existência de algoritmos rápidos e eficientes de cálculo e da maior ou menor "aptidão" para solucionar problemas específicos, como seja a correcção de efeitos de fronteira de bloco ou a conservação de um sistema criticamente amostrado [T9].

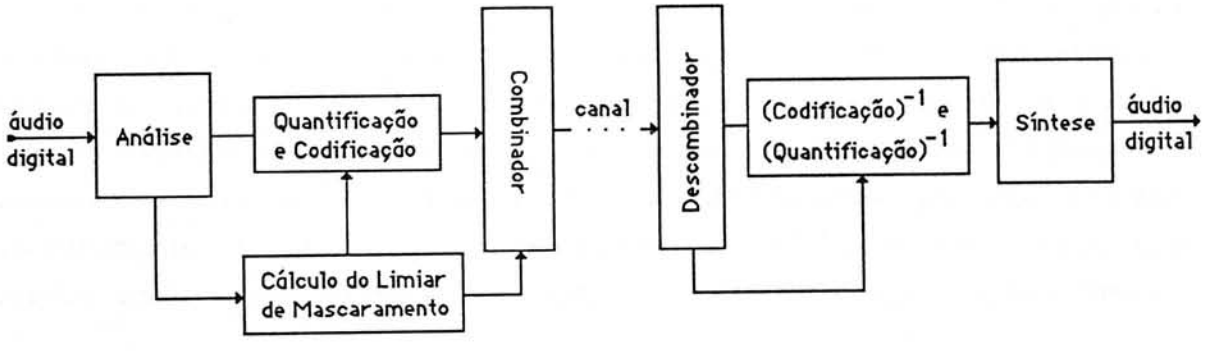

Figura 2.2: Diagrama de blocos de um codificador perceptual por transformada.

De uma forma geral, pretende-se obter uma grande resolução espectral para aplicar convenientemente os modelos psico-acústicos, coeficientes espectrais o mais descorrelacionados possível uns dos outros para conseguir melhores ganhos de codificação e reconstrução perfeita do processo análise-síntese para controlar a injecção e coloração do ruído de quantificação.

O cálculo do limiar de mascaramento é feito, em regra, a partir dos próprios coeficientes espectrais. Contudo, pode-se recorrer também a um bloco analítico distinto [A19][A31].

Os modelos psico-acústicos podem ter uma forma relativamente simplificada ou podem assumir formas complexas, de modo a simular com mais rigor o processamento realizado pelo ouvido e a conseguir, portanto, melhores ganhos de codificação.

O bloco de quantificação e codificação modula o ruído de acordo com o limiar de mascaramento. Este processo pode ser iterativo se se impuser uma taxa de informação constante para a saída do codificador. Pode também haver a necessidade de codificar informação lateral se se optar, por exemplo, por uma forma de codificação APCM.

O combinador reune e formata toda a informação -incluindo informação de sincronização e protecção de erro- para ser transmitida ao descodificador. Tal como o descodificador por sub-bandas, o descodificador por transformada tem também uma estrutura mais simples do que a do codificador já que a sua função é simplesmente descodificar e reconstruir o sinal digital.

A codificação perceptual por transformada trata, na realidade, de equacionar o compromisso entre as exigências impostas pelas características perceptuais, o balanço da resolução espectral/temporal intrínseca à operação de transformada, e algumas variáveis adicionais, de modo a maximizar a qualidade da codificação e a minimizar a taxa de informação. Estas variáveis adicionais, por vezes, actuam contrariamente ao objectivo da compressão. A codificação por transformada implica assim alguns problemas específicos que exigem soluções suplementares.

Um problema particularmente importante relaciona-se com as restrições da propriedade de pré-mascaramento do ouvido humano, inicialmente referidas no parágrafo 2.4.3.1 e abordadas no parágrafo anterior para o caso da codificação por sub-bandas. De facto, na etapa de reconstrução, o ruído de quantificação espraia-se em toda a extensão do segmento de signal transformado. Se o sinal mascarante assumir uma forma de não-estacionaridade, como por exemplo, um transitório no domínio dos tempos, também designado na gíria por "ataque", pode acontecer que na pior das hipóteses, considerando uma transformada com 1024 pontos e um sinal

amostrado à taxa de 48000 amostras por segundo, o ruído de quantificação se inicie 21 milissegundos antes do próprio sinal mascarante. O resultado é comprometedor, quer seja avaliado objectivamente pela relação sinal-ruído, quer subjectivamente pela inevitável detecção de artefactos graves do sinal. Se o cenário for um ataque na sequência de instantes de silêncio, o ruído de quantificação é perfeitamente audível na forma de um desagradável "pré-eco", porque é entendido como um eco invertido nos tempos, isto é, antes do sinal que o provoca e que simultaneamente lhe retira todo o "ímpeto".

Um outro problema prende-se com dificuldades no mascaramento de ruído na fronteira entre segmentos adjacentes. De facto, se os segmentos transformados forem tomados como blocos disjuntos de amostras, embora contínuos no tempo, está-se implicitamente a ignorar a correlação de sinal para além das fronteiras de cada segmento. Na reconstrução, o ruído também se revela não correlacionado e com padrões distintos de bloco para bloco, sobretudo nas suas fronteiras. Este é também um facto audível e designa-se por efeito de fronteira de bloco (*bloc edge effect*). A solução consiste em admitir alguma correlação entre segmentos adjacentes, isto é, admitir a transformação de segmentos com alguma percentagem de sobreposição entre si. Esta operação tem a desvantagem de forçar o sistema de análise (transformada) a deixar de ser criticamente amostrado [T9], o que conduz a um aumento da taxa de informação.

Em geral, as estratégias de codificação procuram evitar que estes problemas comprometam os ganhos proporcionados pela codificação por transformada. Procurar-se-á de seguida dar uma panorâmica geral de várias soluções propostas na literatura recente.

Dado que a codificação por transformada é eficiente para sinais pseudo--estacionários [V6], isto é, aqueles cujos atributos estatísticos variem de forma lenta relativamente à dimensão do segmento transformado, uma possível solução [A1] consiste em alterar a dimensão do segmento transformado, de modo que o sinal se possa considerar sempre quase estacionário. Isto é, isolam-se secções do sinal que exibam características distintas. As dimensões possíveis dos segmentos transformados variam segundo potências de dois. O processo de codificação por transformada é idêntico ao codificador de Zelinski e Noll [V7]. A relação sinal (original)-ruído (de reconstrução) é utilizada para comutar a dimensão do segmento transformado. O algoritmo iterativo consiste em partir da dimensão máxima de transformada, calcular a SNR do segmento reconstruído, subdividir o

bloco em dois, calcular novamente a SNR de cada sub-segmento e repetir a operação para cada sub-segmento. O algoritmo progride até que a relação sinal--ruído média, avaliada na extensão no segmento inicial, seja maximizada. Esta solução é a que dita o comprimento adequado para a transformada dos vários sub--segmentos. A melhoria objectiva deste método tem uma correspondência subjectiva, sobretudo para sinais com fortes ataques, como é o caso dos sons de percussão e castanholas.

Foi definido um conceito importante no contexto da codificação perceptual que traduz de algum modo a capacidade do ouvido humano em absorver informação ou, por outras palavras, a informação perceptualmente útil existente num sinal áudio. O conceito é o da Entropia Perceptual (PE) e encontra-se extensivamente descrito em [A10]. A PE é, na realidade, uma medida e traduz o número mínimo de *bits* necessários para codificar de forma transparente um segmento curto de sinal. A PE é calculada com base num modelo psico-acústico resultante de estudos sobre as propriedades do ouvido humano, sobretudo as de mascaramento de tons puros por bandas de ruído e as de mascaramento de bandas de ruído por tons puros. Este modelo, completado por mais alguns parâmetros psico-fisiológicos, é incorporado num algoritmo de codificação por transformada [A11]. Quando excitado por um sinal áudio, o algoritmo fornece um limiar de mascaramento para cada coeficiente espectral. Idealmente, se for injectado ruído no sinal de acordo com o limiar de mascaramento, através de uma quantificação independente de cada coeficiente, então ter-se-á uma taxa de informação ideal que representa os coeficientes quantificados, ignorando os respectivos factores de escala. Esta taxa de informação ideal dividida pelo número total de coeficientes é a Entropia Perceptual. Representa, portanto, um limite inferior teórico de informação para a codificação transparente de qualquer sinal. É claro que estas considerações supõem a validade e fidelidade do modelo psico-acústico.

As estatísticas apresentadas revelam que sinais com 15KHz de largura de banda apresentam PE entre 0.5 e 2.1 *bit*/amostra, com média aproximada em 1.3 *bit*/amostra.

A PE é utilizada num desenvolvimento posterior [A11] para codificar sinais com 15KHz de largura de banda. A codificação é transparente para a taxa de 128Kbit/s, o que equivale a 4 *bit*/amostra. Este último valor tem obviamente o custo de toda a informação lateral necessária.

O processo de cálculo especificado em [A10] é usado para avaliar a máxima quantidade de energia de ruído, convenientemente disposta ao longo do espectro, que pode ser injectada no sinal de forma inaudível. Tomam-se como base de codificação segmentos curtos de 64ms. Segmentos adjacentes de sinal são sobrepostos em 1/16 do seu comprimento total para obviar artefactos de codificação na fronteira entre blocos. A energia do sinal é inicialmente avaliada em cada banda crítica. Seguidamente, uma função de espraiamento [P12] é convoluída com estas energias para traduzir o efeito de mascaramento entre bandas críticas. As duas expressões do modelo psico-acústico, uma para tons puros mascarando ruído e outra para ruído mascarando tons, são ponderadas pela medida de planura espectral [V6]. Esta medida é usada para avaliar o carácter tonal de secções do espectro. O limiar de mascaramento final é calculado depois de ajustar a energia de ruído anteriormente calculada com o limiar absoluto de audição. Esta operação certifica que o sinal não é quantificado com um ruído inferior à capacidade mínima de audição (capítulo 3).

Dado que este processo de codificação é inerentemente gerador de informação a taxa variável, é usado um procedimento iterativo de ajuste para manter a taxa de informação constante. Em consequência e considerando o limiar de mascaramento, alguns blocos terão uma margem favorável de codificação (são sobrecodificados) e outros blocos poderão ter uma margem desfavorável (são subcodificados), tudo dependendo da disponibilidade média de *bits*.

Uma referência importante na evolução dos codificadores perceptuais é o trabalho desenvolvido por Brandenburg [A16][A17]. O seu primeiro codificador eficiente [A16] comprime sinais com 20KHz de largura de banda em 2.5 *bit*/amostra. O sinal digital é transformado para o domínio das frequências usando uma DCT modificada (MDCT) [T3]. A melhor selectividade da MDCT relativamente a outras transformadas é apontada como uma razão importante para a eficiência global da codificação, além de resolver os problemas de fronteira de segmento. Estes aspectos serão analisados em pormenor no capítulo 4. A maximização objectiva do ganho de codificação (parágrafo 2.3) conduz a uma quantificação inicial de todos os coeficientes usando o mesmo passo de quantificação. Adicionalmente, realiza-se codificação por entropia. Usam-se depois duas malhas de controlo para obter uma taxa de informação constante, respeitando simultaneamente os limites de mascaramento no interior de cada banda crítica.

Uma malha interior aumenta ou diminui o passo de quantificação global para reduzir ou aumentar o número total de *bits* necessários para representar o bloco, tendo em vista o número de *bits* disponíveis. Neste processo iterativo, o limiar de mascaramento numa banda crítica particular pode resultar violado. Por isso, existe uma segunda malha, exterior, que monitora a diferença entre o ruído de quantificação introduzido em cada banda crítica e o respectivo limiar de mascaramento. Sempre que a diferença for positiva, o passo de quantificação é diminuído nesta particular banda crítica. O processo iterativo anterior, assim como a operação descrita a seguir exigem que o sinal seja reconstruído no emissor. Diz-se que o processo de codificação se baseia na análise por síntese.

Para solucionar problemas de pré-eco, é usada uma "pós-filtragem" do sinal reconstruído. Se se detectar que a energia de ruído do sinal reconstruído é superior à energia do próprio sinal, então a ocorrência de pré-eco é declarada. Dado que esta ocorrência resulta essencialmente de ruído introduzido às altas frequências, calcula-se até que frequência se confina 90% da energia total do sinal. Esta frequência é transmitida ao receptor como informação lateral. No receptor, e antes da operação de reconstrução, as componentes espectrais acima daquela frequência são tornadas nulas, isto é, o espectro é "pós-filtrado".

A avaliação da qualidade de codificação é feita de forma preliminar por medidas que pretendem de alguma forma representar, mas não substituir, o resultado dos fastidiosos e sempre longos testes de audição. São a relação ruído de quantificação-limiar de mascaramento (NMR) e indicador de mascaramento *(masking flag)*. A NMR fornece a distância média, para todos os segmentos de sinal, entre ruído de quantificação e o limiar de mascaramento. Por outras palavras, fornece uma medida da margem de segurança do processo de codificação relativamente ao limite teórico de transparência. O indicador de mascaramento traduz a percentagem de segmentos em que pelo menos uma banda crítica viola o limiar de mascaramento.

Apesar do seu interesse prático, estas medidas não têm rigor absoluto pois estão condicionadas à validade do modelo psico-acústico que utilizam.

Um outro codificador [A17], ainda baseado na filosofia anterior [A16], inclui algumas modificações do controlo de pré-eco, que agora está incluído no cálculo do próprio limiar de mascaramento e extrai redundância adicional entre coeficientes adjacentes usando, nomeadamente, DPCM. A codificação de sinais com 20KHz de

largura de banda é transparente à taxa de 128Kbit/s e de aceitável qualidade a 64Kbit/s.

Um codificador particularmente interessante é o codificador denominado ASPEC [A19]. O algoritmo é adaptável a várias taxas de informação, desde 64Kbit/s com qualidade "FM" até 128Kbit/s com qualidade CD. Combina estratégias importadas de vários trabalhos anteriores [A11][A12][A17][A18][A30][T1]. Segmentos do sinal são transformados através da MDCT. São usados segmentos com dois comprimentos básicos: um longo (21ms) e outro curto (5ms). Cada segmento é usado de acordo com uma estratégia de comutação dinâmica de janela de transformada para assegurar por um lado, um ganho de codificação interessante e por outro lado, para controlar efeitos de pré-eco. O cálculo do limiar de mascaramento é semelhante ao já descrito a partir das referências [A11] e [A17]. A operação iterativa de quantificação e codificação insere-se numa malha de análise por síntese que visa respeitar uma taxa de informação fixa [A17]. O sinal é codificado de modo que o limiar de mascaramento não é violado ou é-o de forma inofensiva, por exemplo, reduzindo a largura de banda do sinal.

Existem outras soluções de codificação (e. g. [A5]) que seguem basicamente a mesma estrutura das referências anteriormente citadas, com algumas diferenças de modelos ou implementação.

A referência [A6], por exemplo, não usa um modelo perceptual explícito. De facto, os coeficientes são agrupados em blocos segundo os limites das bandas críticas. Os blocos assumem uma representação em vírgula flutuante. As mantissas têm uma quantificação parcial fixa, bastante grosseira (*coarse*). Os expoentes são enviados como informação lateral para o descodificador e são também usados para realizar adicionalmete uma quantificação "fina" e adaptativa das mantissas. É indicado que a codificação é transparente em 128Kbit/s para sinais com 15KHz de largura de banda.

Uma outra solução de codificação [A29] propõe, adicionalmente, codificação preditiva inter-bloco dos coeficientes espectrais mais importantes, para explorar as estacionaridades de longo termo do sinal. A transformada usada é a DFT com 1/16 de sobreposição entre segmentos adjacentes. O limiar de mascaramento é obtido a partir da energia do sinal em cada banda crítica. Estas energias são convoluídas com uma aproximação linearizada da função de espraiamento, resultando uma curva de excitação. O limiar de mascaramento final é calculado com um

deslocamento fixo (definido por Zwicker [P4]) da curva de excitação. A atribuição de *bits* pelos vários coeficientes é feita minimizando a medida NMR ao longo de todo o espectro. Só é necessária a transmissão de informação lateral quando ocorre uma não estacionaridade do sinal. Neste caso, os preditores devem ser reinicializados.

Uma versão evoluída deste codificador [A30] usa um modelo pragmático para o mascaramento de tons por ruído para avaliar o limiar de mascaramento. Faz-se uso também de uma medida de tonalidade para favorecer a codificação de regiões tonais do espectro. Calculam-se os coeficientes completamente mascarados e os não mascarados. Associando um digito binário a cada um deles, o espectro é codificado por comprimento de série (*run length coding*) e adicionalmente, por entropia. Para os coeficientes não mascarados, é usada uma codificação preditiva em frequência (intrabloco) ou nos tempos (interbloco). A primeira é escolhida sempre que existe uma não-estacionaridade do sinal. Para sinais com 20KHz de largura de banda, a codificação não é transparente a 64Kbit/s e tem mesmo um desempenho sofrível para certos trechos musicais mais críticos.

Foram também ensaiadas algumas soluções baseadas em codificação vectorial [A26][A27]. Os codificadores propostos funcionam exclusivamente a taxa de informação variável entre 64Kbit/s e 100Kbit/s e não estão especialmente preparados para solucionar problemas de pré-eco. Em síntese, os coeficientes obtidos a partir de uma DCT são normalizados por uma envolvente de potência espectral interpolada. Grupos de coeficientes adjacentes são quantificados vectorialmente segundo um padrão de distorção perceptualmente adequado. A sobreposição entre segmentos adjacentes é de 6.7%. Os resultados de codificação revelam que o processo não é transparente para as taxas indicadas.

Embora de uma forma simplificada, as técnicas de codificação vectorial são também retomadas no codificador detalhado no capítulo 4.

2.5 Actividades de Normalização

A par da grande actividade verificada durante a década de 80 na área da codificação perceptual de áudio, surgiu o objectivo, por parte de dois grupos da International Standards Organization (ISO/IEC JTC1/SC2/WG8 e MPEG-Motion Picture Experts Group), de criar uma norma para codificação de sinais áudio a muito baixas taxas de informação, desde 32Kbit/s até 128Kbit/s por canal. Este

objectivo insere-se num projecto mais vasto que visa repartir a taxa de informação típica de um CD (aproximadamente 1.4Mbit/s) por dois serviços: cerca de 1.15 Mbit/s para um sinal de vídeo comprimido e 256Kbit/s para um canal áudio estereofónico comprimido.

Com a perspectiva de marcar um posição activa no espectro de futuras aplicações de áudio, várias empresas situadas em diferentes áreas contribuíram com propostas para a norma. Em função de similaridades básicas, as diferentes propostas foram agrupadas em 4 sistemas distintos.

ASPEC, um codificador por transformada e já descrito no parágrafo 2.4.3.3, reuniu os Laboratórios BELL da AT&T, a CNET francesa (Centre National d'Etudes des Télécommunications) e as Deutsche Thomson Brand e FhG-AIS (Fraunhofer--Gesellschaft), ambas alemãs.

MUSICAM, um codificador por sub-bandas e descrito no parágrafo 2.4.3.2, reuniu o CCETT francês (Centre Commun d'Etudes de Télédiffusion et Télécommunications), o IRT alemão (Institut fur RundFunktechnik), a MATSUSHITA japonesa e a PHILIPS holandesa.

Um outro codificador por transformada, ATAC, reuniu só empresas japonesas: FUJITSU Laboratories, JVC (Japan Victor Company), NEC Corporation e a SONY Corporation.

Finalmente, um codificador por sub-bandas sem nome específico e denominado simplesmente SB/ADPCM, reuniu a BTRL inglesa (British TELECOM, U.K.) e a NTT japonesa (Nippon Telegraph & Telephone Corporation).

Este último codificador é um extensão da estrutura preconizada pela recomendação G722 do CCIR. Resumidamente, o sinal é decomposto em 8 sub-bandas cujas amostras são depois codificadas de acordo com um algoritmo do tipo ADPCM. O codificador não recorre a um modelo psico-acústico explícito.

O codificador ATAC é semelhante ao codificador ASPEC com algumas diferenças no cálculo do limiar de mascaramento, na forma da janela de amostragem temporal para a transformada, além de outras diferenças mais formais.

Em Julho de 1990 os quatro sistemas foram sujeitos a testes completos na SBC (Swedish Broadcasting Corporation), Estocolmo, Suécia. Os sistemas ATAC e SB/ADPCM foram desde logo preteridos dado que exibiam resultados preliminares muito inferiores aos dos seus concorrentes e além disso, não dispunham de suporte (*hardware*) completo para apoiar a sua demonstração.

Os testes de avaliação incluiam diversos parâmetros de desempenho tanto objectivos (*e.g.* complexidade) como subjectivos (*e.g.* qualidade da imagem estereofónica). ASPEC granjeou a melhor pontuação em termos de qualidade de codificação, sobretudo a muito baixas taxas de codificação. Em termos de complexidade e outros parâmetros formais, MUSICAM obteve uma pontuação cuja margem de vantagem relativamente ao ASPEC foi suficiente para superiorizar a sua pontuação geral.

Após este teste foram definidos novos objectivos no âmbito do grupo de trabalho 11 da ISO (WG11). As entidades ligadas ao MUSICAM e ASPEC foram incumbidas de colaborar no projecto de um sistema único, melhorado, estratificado em camadas, e com base nas respectivas propostas. O projecto a desenvolver durante o período de 1990 a 1992 tem como objectivo [A8] obter em 1992, e para a taxa de 2*64Kbit/s, a mesma qualidade de codificação, conseguida em 1990, para a taxa de 2*128Kbit/s (canal estereofónico).

2.6 Conclusões

Dadas as várias estratégias de codificação apresentadas neste capítulo, parece resultar evidente que as técnicas de codificação no domínio das frequências são as que proporcionam melhores ganhos. Dentre estas, a codificação por transformada é a que fornece melhores perspectivas de ganhos adicionais pela facilidade em se adaptar a novas configurações e novos modelos. Basta referir que a redefinição das bandas é extremamente flexível pois resulta só do conveniente agrupamento de coeficientes espectrais. Codificação por transformada será, portanto, a filosofia base de suporte ao codificador estereofónico detalhado no capítulo 4.

Dada a exiguidade de publicações e resultados conclusivos, a compressão explorando correlação intercanal (estereofonia) não foi referida. Sê-lo-á no capítulo 4. Neste capítulo serão também retomados alguns aspectos julgados mais

relevantes, nomeadamente, uso da transformada MDCT e comutação dinâmica de janelas temporais, associadas a transformadas com diferentes resoluções.

3
CODIFICAÇÃO PERCEPTUAL

3.1 Introdução

A psico-acústica é a área de estudos que se ocupa de quantificar a correlação verificável entre estímulos acústicos e impressão ou sensação auditiva. Para tal, são realizados testes de audição em que vários ouvintes (*subjects*) são sujeitos a sinais acústicos especialmente criados para excitar aspectos particulares de comportamento do sistema auditivo. A partir dos resultados obtidos são calculadas estatísticas e definidos modelos matemáticos.

O presente capítulo aborda alguns conceitos relativos à natureza e comportamento do sistema auditivo. Sem pretender ser exaustivo, o que seria aliás uma cândida veleidade, dada a exiguidade de resultados e desconhecimento dos parâmetros psico-fisiológicos que os condicionam, procurar-se-á caracterizar essencialmente aspectos de mascaramento e respectivos modelos que são de primordial importância para o desenvolvimento do capítulo 4.

O emprego frequente de alguns adjectivos susceptíveis de serem interpretados como sinónimos, exige que seja esclarecido o seu significado no contexto da codificação perceptual. Assim, *monauricular* associa-se ao processamento realizado pelo sistema auditivo baseado exclusivamente nas capacidades analíticas de *um* só ouvido. *Binauricular* associa-se ao processamento realizado pelo sistema auditivo, baseado nas capacidades analíticas conjugadas dos *dois* ouvidos. *Monofónico* refere-se ao estímulo acústico expresso num só sinal. *Estereofónico* refere-se ao estímulo acústico expresso em dois sinais, um para cada ouvido. Pode-se deste modo concluir que um estímulo monofónico invoca só propriedades monauriculares quer seja aplicado a um só ouvido ou aos dois ouvidos (em verdade, neste último caso existe um efeito de soma das intensidades apresentadas a cada ouvido). Por outro lado, um estímulo estereofónico invoca propriedades monauriculares e também binauriculares. Neste cenário, o familiar par de canais estereofónico é um estímulo acústico condicionado que invoca, mas não esgota, todas as propriedades binauriculares do sistema auditivo. Por exemplo, a impressão de espaço e localização são bastante restringidas pela disponibilidade de dois únicos focos acústicos (canais).

Importa desde já identificar a sensibilidade máxima e os limites de risco do sistema auditivo, de modo a identificar a área de funcionamento de qualquer

codificador perceptual. O plano de audição traduz a excursão da intensidade acústica admitida pelo ouvido para cada frequência. Definindo a intensidade de pressão sonora (em dB SPL) como sendo o logaritmo da razão entre a intensidade acústica de um sinal e a intensidade acústica de referência igual a 10E(-12) W/m^2 e graduando logaritmicamente o eixo das frequências em três oitavas, a área de funcionamento encontra-se identificada pelas curvas indicadas na Fig. 3.1.

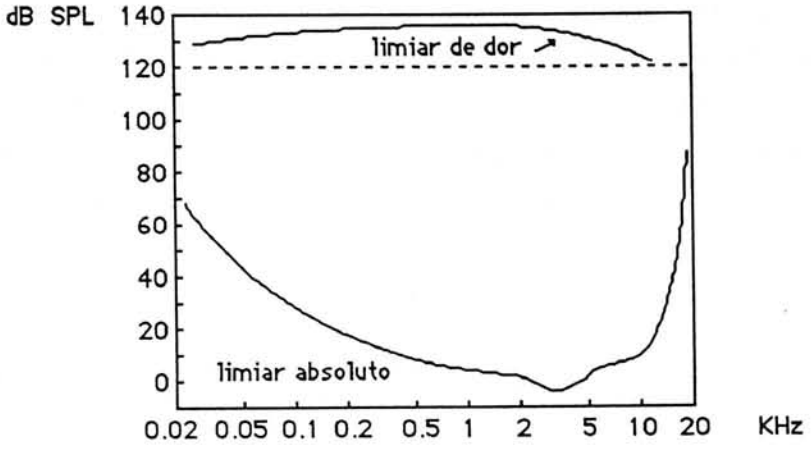

Figura 3.1: A região de audição considerada para a codificação perceptual encontra-se limitada inferiormente pelo Limiar Absoluto de Audição e superiormente pelo segmento a tracejado.

A curva inferior é a do conhecido limiar absoluto de audição (também referida como a curva a 0 *phons*) e traduz a sensibilidade máxima de audição (i.e. a intensidade mínima audível) para cada frequência, em condições de campo aberto (*free-field*). Verifica-se que a acuidade auditiva é máxima para uma frequência aproximada de 4 KHz e diminui assimetricamente para frequências superiores e inferiores. A curva superior está associada às intensidades acústicas que originam dor. Intensidades acústicas superiores às desta curva, mesmo para tempos de exposição bastante curtos, podem provocar destruição literal do sistema auditivo. No entanto, podem ocorrer também perdas permenentes de audição, sobretudo na região de maior sensibilidade (que se reflete numa subida localizada da curva a 0 *phons*), devido a exposições prolongadas a sinais de elevada amplitude. Torna-se portanto necessário respeitar o balanço entre tempo de exposição e intensidade acústica para as regiões próximas do limiar da dor, de modo a permitir a recuperação do sistema auditivo [P5]. No entanto, é comum admitir, para efeitos de codificação perceptual, que a área de funcionamento se limita superiormente a 120dB SPL [P20]. Este valor está associado ao limiar da sensação cutânea (*threshold of feeling*) e assume tempos de exposição curtos.

3.2 Sistema Auditivo

3.2.1 Anatomia do Ouvido

O sentido da audição envolve várias fases de processamento dos estímulos acústicos. A região periférica, constituída pelos elementos imediatamente ligados aos ouvido, capta o sinal acústico e realiza a sua conversão mecânico-electroquímica, de modo a transmitir a informação, através dos nervos auditivos, a regiões centrais localizadas no cérebro. Estas traduzem finalmente a informação acústica em sensação auditiva. Toda a cadeia de processamento contém fortes não--linearidades, observáveis mesmo ao nível do próprio ouvido.

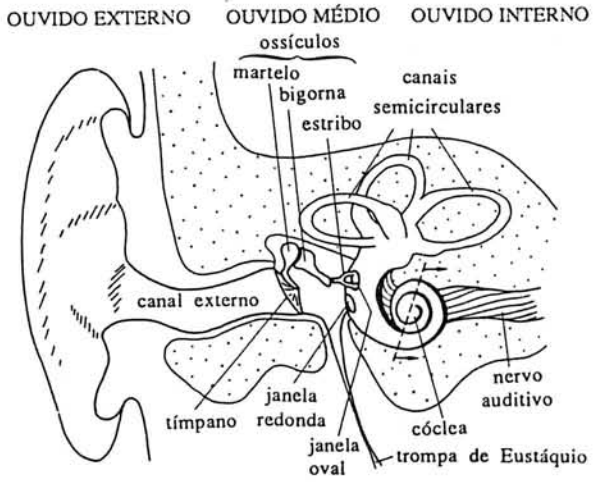

Figura 3.2: Estrutura do ouvido externo, médio e interno.

O ouvido divide-se em três regiões de acordo com a representação da Fig. 3.2. O ouvido externo compreende os elementos que canalizam as vibrações acústicas, isto é, o movimento das partículas de ar. A aurícula faz convergir ondas acústicas para o interior do canal auditivo externo. Estes dois elementos contribuem significativamente para a resposta em frequência do sistema auditivo. Em particular, o comprimento do canal externo (cerca de 2.5cm) relaciona-se com a grande sensibilidade verificada para a frequência aproximada de 4KHz. Refira-se que a cabeça e o tronco humanos também modificam o espectro das ondas acústicas devido a efeitos de reflexão e difracção. Estes efeitos estão associados às importantes funções de localização. De facto, a localização no azimute explica-se por diferenças de intensidade e fase das ondas sonoras que chegam aos dois ouvidos. Porém, para fontes sonoras colocadas simétricamente aos dois ouvidos, isto é, no plano médio, a

capacidade do sistema auditivo em detectar a elevação (desde 0° até 180°) é atribuída à introdução de vales e picos no espectro do sinal acústico, devido aos efeitos referidos anteriormente.

O tímpano situa-se na fronteira entre o ouvido externo e o ouvido médio. As vibrações acústicas são-lhe comunicadas no extremo terminal do canal externo. O tímpano está mecanicamente ligado a uma estrutura de 3 ossículos (martelo, bigorna, estribo) que actuam sobre uma membrana da cóclea, a janela oval, situada no ouvido interno. Esta janela induz um movimento dos fluidos no interior da cóclea, através dos movimentos na base do estribo. O conjunto do tímpano e ossículos actua como um adaptador de impedâncias entre o movimento de partículas de ar, captadas pela área do tímpano, e o movimento dos fluidos no interior da cóclea, comunicados pela área da janela oval. Para além deste papel, o ouvido médio desempenha também funções de controlo de ganho. Este visa proteger o ouvido interno de sobrecargas ou, no limite, de destruição. Para sinais acústicos com intensidades bastante elevadas, a transmissão do ouvido médio torna-se não linear, introduzindo sobretudo termos quadráticos na função de transferência de amplitudes.

O ouvido interno é composto pelos canais semi-circulares, pelo vestíbulo e pela cóclea. Os dois primeiros assistem sobretudo em funções de referencial espacial e equilíbrio motor. É na cóclea que se desenvolvem as funções relevantes de transdução mecânico-electroquímica dos sinais acústicos.

A Fig. 3.3 representa um corte transversal da cóclea. Esta é formada por um conjunto de 3 ductos (rampa vestibular, ducto coclear, rampa timpânica) enrolados em espiral, tal como num caracol, e preenchidas por dois fluidos diferentes.

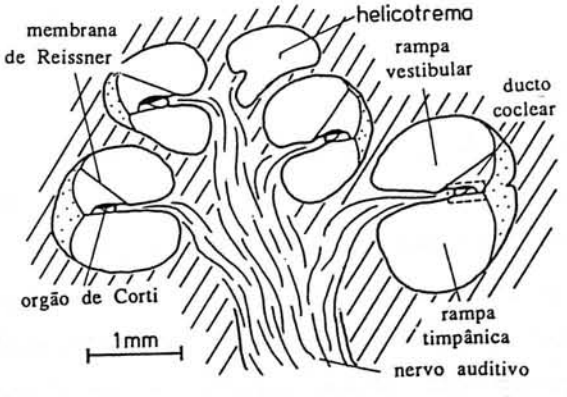

Figura 3.3: Secção transversal da cóclea.

No vértice da cóclea, a rampa vestibular comunica com a rampa timpânica e encerram um fluido com grande concentração de sódio e designado por perilinfo. Este ponto de encontro tem o nome de helicotrema. O ducto coclear encontra-se isolado da rampa vestibular pela membrana de Reissner, e da timpânica pela membrana basilar. Contém um outro tipo de fluido com grande concentração de potássio: o endolinfo. Dado que a membrana de Reissner é extremamente fina, hidromecanicamente, a rampa vestibular e o ducto coclear podem ser vistos como um só canal.

A base do estribo comunica vibrações ao perilinfo no início da rampa vestibular, através da janela oval. Por sua vez, os fluidos transmitem as vibrações à membrana basilar. De facto, como as paredes exteriores aos ductos são rígidas e os fluidos são essencialmente incompressíveis, a igualização do movimento de fluidos verifica-se na janela redonda (situada no início da rampa timpânica) e propaga-se através da membrana basilar.

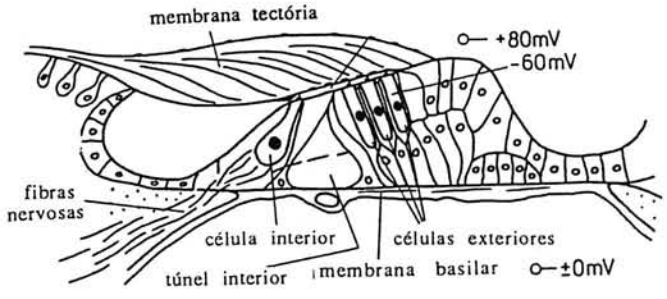

Figura 3.4: Secção transversal do orgão de Corti.

O comprimento da membrana basilar é cerca de 32mm. Esta membrana é estreita e rígida na região da janela oval, enquanto que junto ao hicotrema é mais larga e flexível. Sobreposta e solidária com a membrana basilar encontra-se o orgão de Corti. Este orgão contém células de suporte e células sensitivas (células ciliadas). Estas estão dispostas em várias filas, uma de células interiores e três de células exteriores (Fig. 3.4). As células exteriores e interiores exibem diferenças estruturais e também diferenças funcionais. É admitido [P5] que, para estímulos acústicos com amplitude bastante elevada, os estereocilios das células interiores são directamente estimuladas pelo movimento local da MB. Para estas amplitudes, as células exteriores encontram-se essencialmente em estado de saturação. Para baixas amplitudes, os movimentos da MB são suficientemente suaves para para

estimular adequadamente as células exteriores, não afectando significativamente as interiores. Nestas duas circunstâncias, admite-se existir uma interacção entre os dois tipos de células, de modo que a gama dinâmica resulta aumentada na periferia do sistema auditivo. Esta interacção supõe algum efeito de realimentação activa, por vezes não-linear, que em certas circunstâncias pode originar auto-oscilação ou ainda, o aparecimento de produtos de distorção. Este fenómeno é observável, por exemplo, a partir das emissões espontâneas do ouvido [P5] (*emissões otoacústicas*). Os produtos de distorção aparecem também em consequência de não linearidades hidrodinâmicas para estímulos de grande amplitude [P20].

Uma membrana tectoria cobre ligeiramente as várias filas de células ciliadas. Quando um estímulo acústico atinge a MB, esta é animada de movimento que é comunicado às células e, por acção do seu movimento relativo com a membrana tectoria, o estereocilio situado na sua extremidade é deformado. Dada a diferente polarização do endolinfo, do perilinfo e de outras substâncias complexas que rodeiam as células sensitivas, a inclinação do estereocilio dá origem a um fluxo iónico variável na superfície da célula sensitiva que se reflecte na modulação de impulsos nervosos no nervo auditivo associado. Se o estereocilio é inclinado numa direcção provoca uma excitação de actividade do nervo auditivo, se é inclinado na direccção oposta, origina inibição da sua actividade.

O disparo de impulsos no nervo auditivo ocorre espontaneamente à taxa aproximada de 50 impulsos por segundo. Durante os instantes de subida acentuada de estímulos acústicos, a taxa de disparo pode alcançar os 1000 impulsos por segundo. Este padrão de disparo decai mesmo que o estímulo se mantenha com o mesmo nível de excitação. Uma justificação encontrada [P11] é a necessidade da célula sensitiva associada recuperar as suas características de polarização estacionária. O período de recuperação é cerca de 1ms. O nervo auditivo exibe também outras características importantes como linearidade e controlo automático de ganho [P20].

Von Békésy observou as vibrações da membrana basilar (MB) sob estimulação acústica e notou a existência de ondas progessivas deslocando-se desde a base da cóclea até ao vértice. À medida que a onda vai progredindo, a sua velocidade vai diminuindo. A velocidade é também dependente da frequência do sinal, ou seja, a MB é um meio dispersivo. Békésy identificou o meio de propagação associado à MB como sendo uma espécie de linha de transmissão não uniforme: as altas frequências viajam só distâncias muito curtas na MB antes de serem

acentuadamente atenuadas, enquanto baixas frequências viajam distâncias mais longas antes de se extinguirem. Em termos concretos, considerando grandezas características da MB como a massa, resistência e rigidez, pode-se definir uma impedância complexa [P11]. Assumindo adicionalmente algumas premissas, como por exemplo, incompressibilidade do fluido no interior da cóclea e fronteiras (paredes) rígidas, pode-se definir uma equação de movimento cujas soluções confirmam o comportamento das ondas progressivas.

Como é exemplificado na Fig. 3.5, para cada frequência, a envolvente da onda progressiva atinge um valor máximo para uma dada posição da MB. Frequências mais baixas atingem o seu valor máximo em posições que se situam a distâncias maiores da janela oval. Por outras palavras, a MB está sintonizada em frequência, em função da distância à janela oval. No próximo parágrafo, esta selectividade em frequência será melhor caracterizada recorrendo à organização das *bandas críticas* que, por sua vez, definem uma escala "natural" para o eixo das frequências.

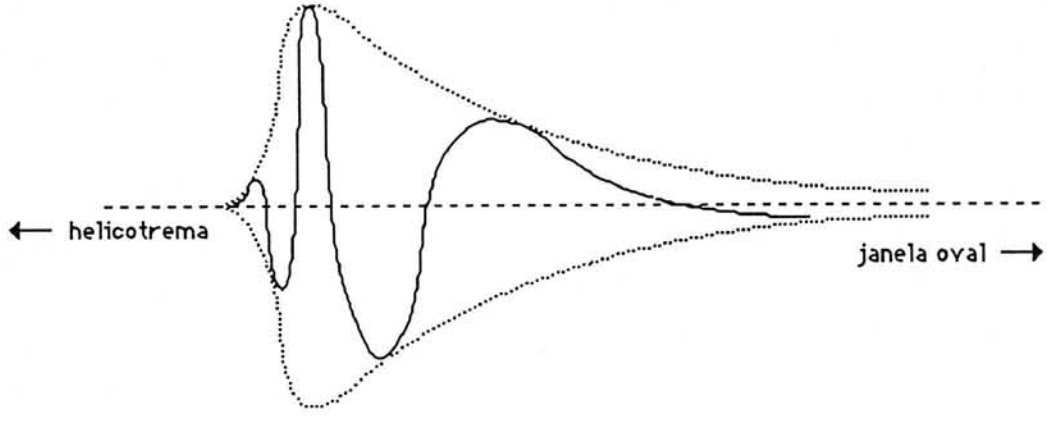

Figura 3.5: Em função da sua frequência, a onda progressiva induz uma vibração máxima na membrana basilar para uma distância específica da janela oval.

Tendo em consideração caracterizações simplificadas do sistema auditivo, há modelos matemáticos que procuram reproduzir vários processos, como por exemplo, o da transdução mecânico-electroquímica [P8][P11][P19]. Estes modelos, além de serem simplificações, exibem por vezes uma complexidade considerável. Contudo, esta não é a razão principal pela qual não são usados directamente no cálculo do limiar de mascaramento. De facto, a impressão ou sensação auditiva resulta também do processamento da informação coligida pelos nervos auditivos, em estágios mais elevados do sistema auditivo, localizados no cérebro. Este processamento envolve, entre outras, funções de memorização, filtragem e

correlação. No actual contexto de conhecimento, estes estágios não estão isoladamente acessíveis ou são desconhecidos, o que dificulta naturalmente a quantificação individualizada das suas propriedades. Por isso, só é possível realizar uma análise conjunta e modelização de todo o processo, desde o estímulo acústico até à consequente impressão auditiva. Os dados necessários são obtidos através de testes psico-acústicos adequados. Neste espírito, será a partir destes que se retirará informação relevante para o objectivo da codificação perceptual.

3.2.2 Bandas Críticas

As bandas críticas são, antes de mais, um conceito empírico pois resultam da coerência dos resultados de diferentes testes relativos às características de selectividade espectral do ouvido humano.

A designação de bandas críticas foi originalmente proposta em 1940 por H. Fletcher, partindo da hipótese de que quando ruído branco mascara um tom, só uma banda estreita de frequências nas vizinhanças do tom, igualando-o em potência, é que contribui efectivamente para o seu mascaramento. Assim, a largura de cada banda crítica seria muito simplesmente dada pela razão entre a potência do tom e a densidade de intensidade espectral de ruído branco. Esta razão passou a designar-se por *razão crítica*. Testes posteriores provaram consistentemente que aquela hipótese estaria incorrecta para frequências inferiores a 200Hz e estaria "correcta" para frequências superiores a 200Hz, desde que as larguras das bandas fossem 2.5 vezes menores do que as iniciais [P9].

Usando procedimentos de teste diferentes, verificou-se haver concordância quanto às bandas em que o sistema auditivo exibe um comportamento peculiar. Verificou-se, por exemplo, (Zwicker 1955, Scharf 1959) que a intensidade percebida (*loudness*) de sons complexos, com amplitude constante e em bandas sub-críticas, era constante e igual à mesma amplitude de um tom puro centrado na banda crítica. Só quando a diferença em frequência das várias componentes do som ultrapassasse os limites da banda crítica é que a intensidade percebida aumentava. Por outras palavras, a intensidade percebida de uma banda crítica de ruído é idêntica à de um tom nela centrado.

Um outro teste (Greenwood 1961) envolveu bandas de ruído mascarante e um tom. Provou-se, por um lado, que o mascaramento do tom era máximo quando se

posicionava no centro da banda crítica. Por outro lado, o mascaramento do tom aumentava com a largura da banda de ruído que o envolve e era máximo quando esta atingia os limites da banda crítica. Para além destes limites, o mascaramento era constante.

Um método mais preciso foi planeado por Zwicker em 1954. Considerou o limiar de mascaramento de uma banda de ruído devido a dois tons, um em cada lado da banda. Aumentando a diferença em frequência dos dois tons, concluíu-se que o limiar de mascaramento mantinha-se inalterado até que esta diferença de frequências atingisse os limites da banda crítica. A partir daí, o limiar de mascaramento diminuía bastante.

Banda Crítica	Limite inferior (Hertz)	Limite Superior (Hertz)
1	0	100
2	100	200
3	200	300
4	300	400
5	400	510
6	510	630
7	630	770
8	770	920
9	920	1080
10	1080	1270
11	1270	1480
12	1480	1720
13	1720	2000
14	2000	2320
15	2320	2700
16	2700	3150
17	3150	3700
18	3700	4400
19	4400	5300
20	5300	6400
21	6400	7700
22	7700	9500
23	9500	12000
24	12000	15500
25	15500	19200
26	19200	21000

Tabela 3.1: Limites das Bandas Críticas.

Outras experiências provaram que a sensibilidade do ouvido humano a diferenças de fase de componentes de um som complexo é mais notória se as componentes se situarem dentro da mesma banda crítica. Provou-se também que a

identificação de harmónicos ou a agradibilidade de um conjunto de tons é maior quando a sua separação é função das bandas críticas.

Com base nestes testes, reuniu-se consenso quanto à definição das bandas críticas de acordo com as frequências indicadas na tabela 3.1. As duas últimas bandas desta tabela são uma extensão adaptada de [A24].

Dado que cada banda crítica se assemelha a um filtro passa-banda natural do sistema auditivo, convencionou-se adoptar uma nova escala de frequências, adaptada às características perceptuais [P4]. Escolheu-se Bark como unidade (em memória a Barhausen que, por sua vez, introduziu a unidade de intensidade percebida: *phon*). Um Bark corresponde à largura de uma qualquer banda crítica. Por exemplo, a frequência de 2000Hz corresponde à frequência perceptual de 13 Bark. Esta escala é sobretudo coerente com a constatação [P9] de que uma banda crítica corresponde a um comprimento fixo, cerca de 1.3mm, na membrana basilar.

É curioso verificar que, por uma via distinta, também se concluíu que a inteligibilidade da voz era consequência da análise auditiva baseada em bandas não uniformes. De facto, French e Steinberg, em 1947, isolaram 20 bandas contíguas em que cada uma contribuía igualmente (em 5%) para a inteligibilidade de sinais de voz [V5][V6][P9]. Esta divisão espectral ficou conhecida como Medida de Índice de Articulação e, apesar de se limitar à frequência máxima de 5600Hz, é interessante observar que tem uma correspondência notável com a divisão do espectro em bandas críticas. Outros trabalhos confirmaram esta aparente tendência do sistema auditivo para "integrar" a energia do sinal em cada banda crítica. Este é um aspecto que também se revela interessante no contexto específico do reconhecimento de voz.

Dado que os diversos testes que confirmaram a divisão do espectro em bandas críticas, consideraram essencialmente condições permenentes de excitação (mais de 100ms), é pertinente questionar a validação das bandas críticas para diferentes condições de excitação. De facto, algumas experiências [P9] (sobretudo baseados em mascaramento e com excepção para as baseadas em intensidade percebida) parecem sugerir que a largura das bandas críticas aumenta para durações curtas de excitação. Neste cenário, admite-se que a selectividade espectral do sistema auditivo está associada a algum processo que necessita de um certo tempo (*e.g.* 10ms) de validação (*build-up time*). Porém, os resultados obtidos até agora não são totalmente conclusivos, encerrando mesmo alguma contradição. Assim, no próximo

capítulo, considerar-se-á que as bandas críticas são fixas, de acordo com a tabela 3.1.

3.2.3 Mascaramento Monauricular

3.2.3.1 Domínio dos Tempos

Como foi já referido em capítulos anteriores, o mascaramento de um som devido a outro, traduz-se pela subida do limiar de mascaramento do primeiro, devido à presença do segundo. Este facto foi apontado para a circunstância em que os dois sons estão simultaneamente presentes. Além de depender fortemente da amplitude de cada sinal, a subida do limiar de mascaramento depende sobretudo da relação de frequências entre ambos.

O presente parágrafo aborda o efeito de mascaramento que depende da relação de amplitudes, de frequências e também da relação temporal entre os dois sons. Isto é, supõe-se que os dois sons ou estímulos estão afastados no tempo. Este é um cenário típico em música pois existe uma frequente alternância de sons com alta e baixa amplitudes, originando efeitos complementares de mascaramento.

Os testes psico-acústicos publicados contemplaram duas ocorrências típicas. Por um lado, analisou-se o mascaramento de um sinal breve e de baixa amplitude que precede um outro sinal de maior duração e amplitude bastante mais elevada. Esta situação é conhecida por pré-mascaramento. Por outro lado, avaliou-se o mascaramento de um sinal breve e de baixa amplitude que se sucede a um sinal de maior duração e amplitude. Esta situação é conhecida por pós-mascaramento.

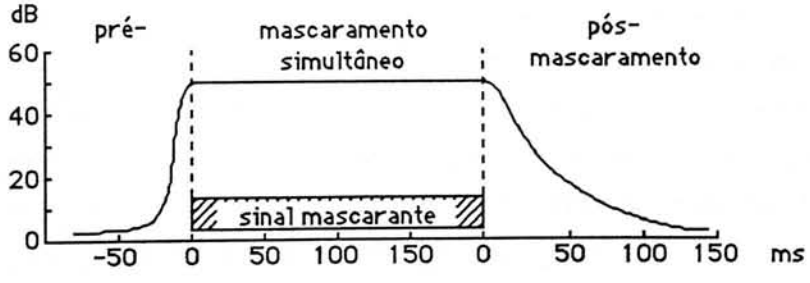

Figura 3.6: Mascaramento no domínio dos tempos. A escala temporal está referida aos efeitos de pré-mascaramento e pós-mascaramento.

Supondo um sinal mascarante com a duração de 200ms e um pequeno sinal (tonal) mascarado, com duração muito inferior, a variação do limiar de mascaramento em função da distância temporal entre os dois sinais é a representada, de forma aproximada, na Fig. 3.6.

Conclui-se facilmente que o mascaramento é máximo e constante quando os dois sinais ocorrem simultaneamente. Este nível de mascaramento decresce exponencialmente quando o sinal mascarado precede o mascarante. Dado que é impossível o sistema auditivo "ouvir" no futuro, este facto explica-se pela possibilidade de existência de diferentes mecanismos de processamento para sinais de elevada e baixa amplitude. A explicação poderá residir mesmo ao nível periférico e ligada às células sensitivas exteriores e interiores, como foi brevemente referido no parágrafo 3.2.1. Não há, porém, resultados conclusivos [P5][P20] e o facto global parece ser que o processamento mais rápido de sinais com elevada amplitude "apaga" o processamento, iniciado alguns instantes atrás, de sinais com baixa amplitude.

O efeito de pré-mascaramento parece revelar-se apenas nos 20ms que antecedem o início do sinal mascarante e parece independer da duração deste [P5]. Contudo, depende ligeiramente da amplitude do sinal mascarante e depende significativamente da relação de frequências entre os dois sinais.

Quando o sinal mascarado sucede ao sinal mascarante, o limiar de mascaramento também decai, mas muito mais lentamente do que na situação de pré-mascaramento. De facto, ainda há um significativo efeito de mascaramento para uma diferença temporal de 100ms. Como no caso anterior, a amplitude do sinal mascarante também não modifica significativamente o padrão de pós--mascaramento. Porém, a duração do primeiro já tem consequências importantes, assim como a relação de frequência dos dois sinais. Concretamente, o nível de mascaramento será tanto maior quanto maior for a duração do sinal mascarante e quanto mais próximas forem as frequências dos dois sinais. O pós-mascaramento é interpretado como o resultado de uma persistência da representação do sinal mascarante nos canais e centros auditivos [P20], com a consequente redução da sensibilidade para processar sinais com baixa amplitude.

Do ponto de vista da codificação perceptual por transformada, as considerações da literatura acerca de pré e pós-mascaramento fornecem só uma

indicação qualitativa. De facto, a informação mais relevante é que o efeito de pré--mascaramento verifica-se para tempos muito inferiores aos de pós-mascaramento.

A circunstância de interesse, no contexto do capítulo 4, envolve sinais mascarados que, em vez de serem extremamente curtos, como assumido acima, se prolongam desde o início (fim) do sinal mascarante, até algum instante antes (depois) deste. Esta reformulação dos conceitos de pré e pós-mascaramento é a que reflecte, com maior fidelidade, o efeito de espraiamento do ruído de quantificação no domínio dos tempos, devido à reconstrução do sinal por transformada. Dado que não existem resultados quantitativos para estas condições de teste, no capítulo 4 procurar-se-á encontrar uma resposta satisfatória para o caso considerado.

3.2.3.2 Domínio das Frequências

No contexto da codificação perceptual, o sinal mascarado é o ruído de quantificação. O sinal mascarante é uma qualquer representação musical. Para realizar uma codificação transparente, é necessário conhecer e dominar o efeito de mascaramento do ruído de quantificação devido a uma qualquer componente musical, quer tenha um perfil tonal (coerente) ou incoerente. Os estudos e resultados da psico-acústica relevantes para este objectivo são os que retratam cenários extremos, isto é, os que envolvem testes de tons puros mascarando bandas de ruído e os que envolvem bandas de ruído mascarando tons puros. Estes últimos representam uma abordagem pragmática do caso que mais se lhe assemelha e que também nos interessa: o mascaramento de bandas de ruído por sinais incoerentes.

Zwicker publicou bastante trabalho relativo a mascaramento, embora não dirigido especialmente à codificação perceptual. O modelo que propôs em 1967 para cálculo do limiar de mascaramento, dada uma excitação arbitrária, está detalhado numa recente publicação [P5] e serviu de base ao desenvolvimento de diversos codificadores perceptuais (*e.g.* [A7][A23]) que introduziram, contudo, algumas adaptações ao modelo original.

Far-se-á, de seguida, uma breve referência à metodologia proposta e seguida Zwicker, relevando as suas virtualidades, usadas por praticamente todos os codificadores perceptuais e apontando as suas insuficiências. Este será o prefácio do parágrafo seguinte que apresenta o modelo base adoptado e discutido no capítulo 4.

O modelo de Zwicker fundamenta-se sobretudo em testes de tons puros mascarados por ruído confinado às bandas críticas. Supondo um sinal de excitação composto por ruído na terceira, nona e décima-oitava bandas críticas (cujas frequências centrais são, respectivamente, 250Hz, 1KHz e 4KHz), de acordo com a tabela 3.1, o limiar de mascaramento de um sinal tonal, em função da frequência, é o representado na Fig. 3.7. O tom manter-se-á inaudível desde que, para cada frequência, não ultrapasse a amplitude definida pelas curvas assinaladas ou pelo limiar absoluto de mascaramento.

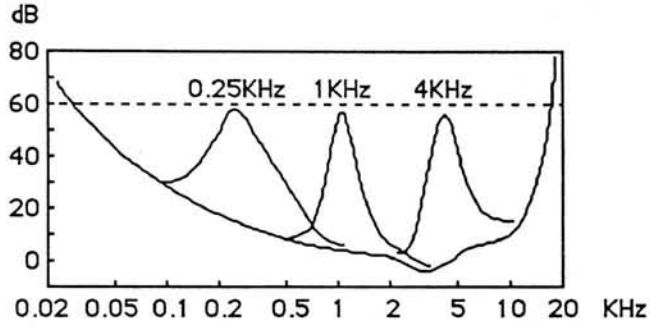

Figura 3.7: Curvas de mascaramento de um sinal tonal devido a três bandas críticas de ruído centradas nas frequências indicadas. A intensidade de ruído em cada banda é 60dB. No sentido de frequências crescentes, o mascaramento máximo é 2dB, 3dB e 5dB, respectivamente.

As várias curvas revelam três aspectos importantes. Para além de ser notório que o mascaramento é máximo quando o sinal tonal se situa no centro de cada banda crítica, o valor máximo de mascaramento é diferente em cada banda. Por outro lado, atentando na inclinação ascendente e descendente de cada curva de mascaramento, pode-se concluir que de uma forma geral, é conseguido maior mascaramento para frequências superiores ao sinal mascarante do que para frequência inferiores. Relacionando este facto com os conceitos introduzidos no parágrafo 3.2.1, pode-se concluir que a maior influência a altas frequências pode estar simplesmente ligada à perturbação causada pela onda progressiva cuja envolvente aumenta ao longo da membrana basilar, desde o seu início na janela oval. Isto é, das mais altas para as baixas frequências. Após atingir o seu valor máximo, a onda progressiva atenua-se rapidamente (Fig. 3.5) causando muito pouca perturbação para pontos mais distantes na membrana basilar, ou seja, para frequências mais baixas.

Finalmente, para a escala adoptada verifica-se que as curvas de mascaramento não são idênticas para as três bandas críticas de ruído mascarante

consideradas. Se a escala de frequências for graduada em Bark, realizando a conversão sugerida no parágrafo anterior, verifica-se (Fig. 3.8) que as curvas de mascaramento exibem um perfil idêntico com inclinações aproximadas de 27dB/Bark na parte ascendente e 10dB/Bark na parte descendente.

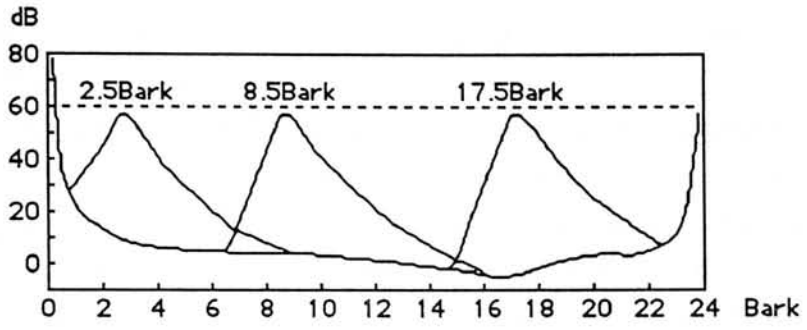

Figura 3.8: Mesma circunstância da Fig. 3.7 mas com o eixo das frequências graduado em unidades *Bark*.

A vantagem decorrente é que a partir do cálculo da excitação para cada banda crítica, o mascaramento correspondente é predizível e obtido a partir da deslocação da curva de mascaramento normalizada para o centro dessa banda. Isto é, a curva de mascaramento independe da frequência central. Esta é uma das vantagens de graduar o eixo das frequências segundo a escala perceptual, em Bark.

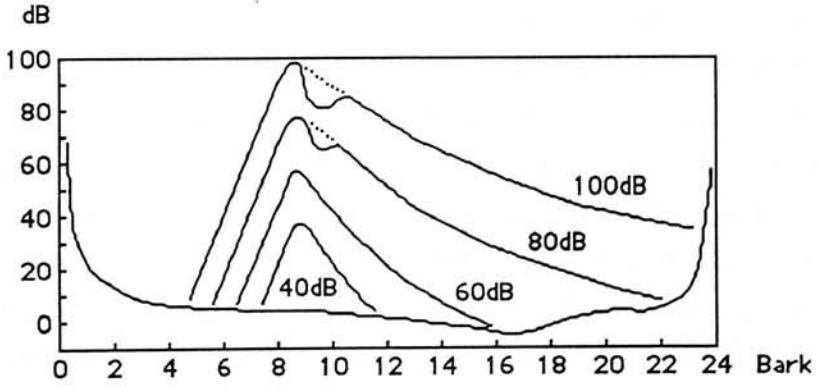

Figura 3.9: Curvas de mascaramento de um tom devido a uma banda crítica de ruído centrada em 1KHz. O parâmetro é a intensidade do ruído.

Na realidade, a curva de mascaramento exibe dependência da inclinação descendente com a amplitude de excitação (Fig. 3.9). Para níveis sucessivamente mais elevados desta, a inclinação varia desde 10dB/Bark até 5dB/Bark, aproximadamente. Para as excitações mais elevadas, verifica-se um pequeno efeito

não-linear do sistema auditivo que se traduz por uma depressão no início da inclinação descendente.

Zwicker refere também alguns resultados de mascaramento de tons devido a outros tons. Estes resultados encerram algumas incertezas devido a dificuldades na realização dos testes. De facto, quando a frequência do tom mascarante e a do tom mascarado são bastante próximas, verificam-se efeitos de batimento e a influência de tons combinados (combinação linear de dois tons como seja a sua soma e diferença). Estes aspectos iludem a percepção auditiva e comprometem a validade dos resultados em certas regiões espectrais. Por este motivo é que as bandas de ruído mascarante são consideradas preferíveis a tons puros mascarantes. Contudo, a informação útil das curvas de mascaramento de tons devido a tons confirmam [P5][P20] essencialmente as curvas obtidas anteriormente para o mascaramento de tons devido a bandas críticas de ruído, apesar do mascaramento máximo conseguido com tons puros ser menor do que o conseguido com bandas de ruído (este aspecto será retomado nos parágrafos 3.2.3.3.1 e 3.2.3.3.2). Em particular, a mesma depressão no início da inclinação descendente encontrada para estas curvas de mascaramento, e para níveis elevados de excitação, também é visível nas de mascaramento de tons por tons. O facto não é atribuído à audibilidade do tom mascarado, mas à de um tom diferença entre o mascarado e o mascarante.

A partir dos conceitos base anteriores, o cálculo do limiar de mascaramento segue os passos que se sintetizam a seguir. Supondo o espectro já convertido em unidades Bark, calcula-se inicialmente a intensidade de excitação, vista pela largura de uma banda crítica (1 Bark), para cada ponto do espectro. Este integral designa-se por intensidade crítica (*critical-band intensity*). Uma operação logarímtica converte seguidamente os valores da intensidade de excitação em valor crítico (*critical-band level*). O valor crítico máximo dentro dos limites de cada banda crítica é designado por ponto de excitação principal (*main excitation*). Se uma banda crítica tiver valores críticos idênticos, então o seu ponto de excitação principal é o seu ponto médio. Em cada ponto de excitação principal é colocada uma curva de mascaramento (com a inclinação superior corrigida para a intensidade da excitação principal) a uma distância, no eixo vertical, dada pelo índice de mascaramento. Este índice de mascaramento é um valor função da frequência que fornece a diferença entre a intensidade de ruído uniforme mascarante e o limiar de mascaramento de tons. Varia entre -2dB para baixas frequências e -6dB para altas frequências. Estas operações estão indicadas na Fig. 3.10 para o caso de um sinal de excitação formado por uma banda crítica de ruído ou por um conjunto de

11 tons. O diagrama de mascaramento final é dado pela reunião dos domínios inferiores de cada curva de mascaramento e o limiar absoluto de audição (ou mascaramento). Existe também algum mascaramento adicional considerando o arredondamento entre inclinações concorrentes de curvas de mascaramento adjacentes.

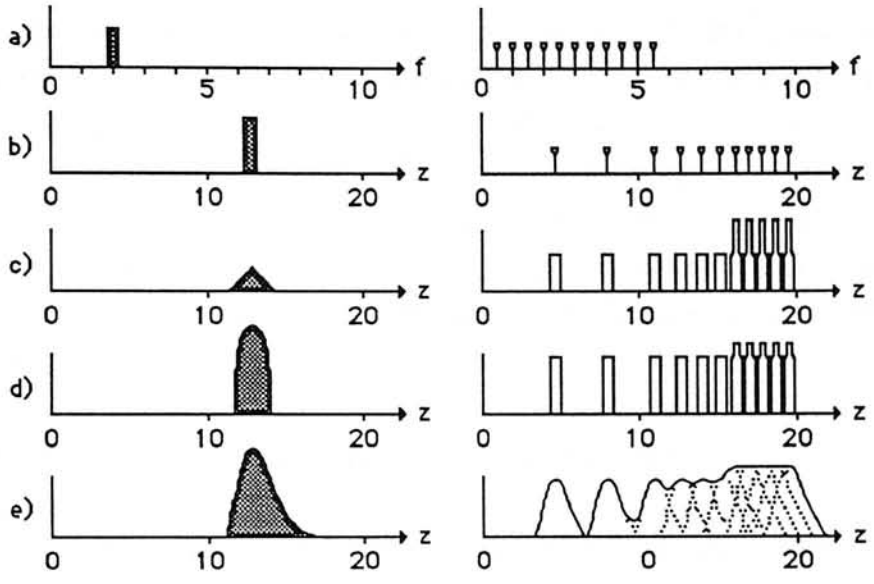

Figura 3.10: Determinação do limiar de mascaramento, segundo Zwicker [P5], para um sinal envolvendo uma banda crítica de ruído (lado esquerdo da figura) e para um sinal composto por 11 tons (lado direito da figura). a) representa o sinal usando o eixo das frequências graduado em KHz. b) idêntico a a) mas com o eixos das frequências convertido para unidades Bark. c) intensidade crítica do sinal. d) valor crítico do sinal. e) colocação de uma curva de mascaramento em cada ponto de excitação principal. Há um ligeiro arredondamento entre curvas adjacentes.

A metodologia de Zwicker inclui algumas premissas muito convenientes do ponto de vista de processamento e também assumidas por outros autores. Por exemplo, é reconhecido que as bandas críticas definem a selectividade espectral natural do sistema auditivo, a partir de circunstâncias em que o comportamento auditivo revela variações bruscas [P9] ou, equivalentemente, revela efeitos máximos. Porém, estes efeitos propagam-se, embora de forma atenuada, a bandas críticas vizinhas. A propagação destes efeitos é regida pela forma dos filtros naturais referidos a (mas não necessariamente centrados em) cada banda crítica, e vistos até agora como sendo as curvas de mascaramento.

Um outro aspecto importante diz respeito à aparente linearidade, para qualquer frequência, entre intensidade de excitação e efeito máximo de mascaramento. Isto é, um qualquer incremento na intensidade de excitação produz

o mesmo incremento no valor máximo do mascaramento por ela ocasionada. Este valor máximo corresponde, como se viu, ao pico das curvas de mascaramento e dista da intensidade de excitação o valor do índice de mascaramento. Por outras palavras, o índice de mascaramento só depende da frequência, apesar das curvas de mascaramento exibirem uma pequena dependência da inclinação superior com a intensidade da excitação (Fig. 3.9).

Estes aspectos são fundamentais e foram, de uma forma geral, corroborados por resultados psico-acústicos posteriores. Por outro lado, as dificuldades associadas à metodologia proposta por Zwicker devem-se sobretudo a duas razões. Por um lado, o índice de mascaramento é aplicado indistintamente para qualquer excitação, assumindo-se que só a sua intensidade é relevante. Como já referido, o índice de mascaramento foi calculado a partir de testes de mascaramento de tons devido a ruído. Esta abordagem é insuficiente tanto mais que Zwicker admite ([P5], pág. 63) que o mascaramento de tons devido a ruído é mais eficiente do que o mascaramento de tons devido a tons e, por uma extensão de conceitos, de ruído devido a tons. Por arrastamento, não é considerado o perfil tonal do espectro.

Por outro lado, o limiar de mascaramento avaliado assumiu sinais com duração superior a 200ms. Para sinais com duração inferior, Zwicker preconiza que o limiar de mascaramento deve aumentar na mesma proporção que a razão entre 200ms e a duração do sinal. Esta é também uma solução insatisfatória no contexto da codificação perceptual.

Apesar da metodologia original de Zwicker ter sido modificada [A28][A32] para superar as dificuldades que foram apontadas, a adoptada no capítulo 4 tem uma origem e naturezas distintas. Assim, passar-se-á a considerar uma diferente abordagem no cálculo do limiar de mascaramento, cujos créditos advêm dos resultados de compressão conseguidos no âmbito do trabalho desenvolvido até 1990, nos Laboratórios Bell da AT&T.

3.2.3.3 Modelos Psico-acústicos

3.2.3.3.1 Tons Mascarando Ruído

R. Hellman [P18] caracterizou quantitativamente a diferença entre a propriedade do ruído em mascarar tons e a propriedade dos tons em mascarar

ruído. Dos casos que considerou, interessam-nos fundamentalmente os que envolvem ruído confinado em bandas críticas, pelo que só a estes se passará a fazer referência.

Os testes realizados demarcaram três regiões do comportamento auditivo: uma em que o mascaramento era total, outra em que era nulo e uma outra em que o mascaramento era gradual entre estas duas últimas regiões. A região transitória designa-se por região de mascaramento parcial porque o sinal mascarante só é parcialmente efectivo. A região de mascaramento nulo revela-se quando o sinal mascarado tem intensidade igual ou superior ao mascarante, na mesma banda crítica. A região de mascaramento total é definida pelo ponto em que o sinal mascarado é-o completamente. O conjunto de diversos pontos relativos a esta última região é o que contém informação para o objectivo da codificação perceptual.

Infelizmente, o número de pontos considerado é bastante exíguo quer para o mascaramento de tons devido a ruído, quer para o mascaramento de ruído devido a ruído. No primeiro caso, só foi referido um ponto, concluindo-se que um tom puro, com a frequência de 1KHz, é completamente mascarado quando a sua intensidade é 5.5dB inferior à do ruído na banda crítica associada. Este valor difere em 2.5dB do publicado por Zwicker para a mesma circunstância, como foi referido no parágrafo 3.2.3.2. No segundo caso, foram considerados dois pontos correspondendo a tons mascarantes com intensidade de 70dB e com frequências de 1KHz e 1.4KHz, respectivamente. Os valores da intensidade mascarada das bandas críticas de ruído respectivas, distam 24dB e 25.5dB daquela intensidade. Schroeder, Atal e Hall [V4], considerando diversos tons com intensidade de 80dB, confirmaram estes dois últimos pontos e obtiveram outros que permitiram a obtenção da seguinte equação para a curva de mascaramento de bandas críticas de ruído devido a tons (TMN):

$$TMN_{dB}(b) = 15.5 + b \quad ; \quad 0.0 \leq b \leq 26.0 \tag{3.1}$$

Esta equação indica simplesmente que dada a intensidade de excitação de um sinal qualquer, composto só por componentes tonais, o ruído na banda crítica cuja frequência é **b** (em Bark), será completamente mascarado se tiver uma intensidade (15.5 + b) dB inferior à intensidade de excitação, na mesma banda crítica. Esta equação foi adoptada em vários trabalhos [A10][A11].

3.2.3.3.2 Ruído Mascarando Tons

Como foi referido no parágrafo anterior, os resultados de Hellman [P18] diferem dos de Zwicker [P5] quanto ao mascaramento de tons devido a bandas críticas de ruído. Concretamente, para a banda centrada em 1KHz, a diferença atinge os 2.5dB. Testes informais confirmaram a adequação do resultado de Hellman e sugeriram a aproximação que consiste em tomar como aceitável para todo o espectro o valor determinado para a nona banda crítica [A10][A11]. Deste modo, a equação que fornece a distância entre a intensidade de excitação devida a um sinal incoerente e o limiar de mascaramento de um tom é, em cada banda crítica:

$$NMT_{dB}(b) = 5.5 \tag{3.2}$$

Esta equação de mascaramento de tons devido a ruído (NMT) é usada como uma solução pragmática, para a circunstância concreta na codificação perceptual, de mascaramento do ruído de quantificação devido a um sinal incoerente.

3.2.3.3.3 Espraiamento entre Bandas Críticas

A estratégia de Zwicker para calcular o diagrama de mascaramento resume-se em quatro etapas essenciais:

1- convolução da energia espectral com uma janela rectangular de 1 Bark de largura,

2- determinação dos pontos de excitação principal,

3- colocação das curvas de mascaramento em função do índice de mascaramento,

4- obtenção do limiar de mascaramento por reunião e arredondamento das curvas de mascaramento.

Uma estratégia alternativa [A10], respeitado os atributos espectrais do sinal, envolve as seguintes fases principais:

1- cálculo da energia em cada banda crítica,

2- cálculo da energia mascarante através da convolução das energias críticas com uma função de espraiamento,

3- determinação do índice de mascaramento considerando o perfil tonal do espectro e realizando uma interpolação simples entre os valores das equações (3.1) e (3.2),

4- obtenção do limiar de mascaramento através da subtracção simples entre a energia mascarante e o índice de mascaramento.

Para a última estratégia, a fase 1 resume-se a um simples somatório para cada banda crítica. A fase 3 será abordada no parágrafo seguinte. A fase 4 tem implícitas as operações de ajuste para o limiar absoluto de audição (ou mascaramento) e a conversão do limiar de mascaramento para ruído de quantificação a injectar no espectro original do sinal. A fase 2 visa traduzir o efeito de mascaramento entre bandas críticas. Para tal, efectua um espraiamento da energia em cada banda crítica a bandas vizinhas, através de uma função de espraiamento, idêntica às curvas de mascaramento - que traduzem o mesmo efeito - consideradas por Zwicker. A operação de convolução com a função de espraiamento converte as energias críticas num diagrama de excitação que representa, na realidade, a distribuição de energia ao longo da membrana basilar.

A função de espraiamento adoptada no capítulo 4 foi proposta por Schroeder [P12]. Tem uma inclinação ascendente de 25dB/Bark e inclinação descendente de 10dB/Bark. Estas inclinações são fixas qualquer que seja a intensidade de excitação. Em particular, a inclinação descendente inclui no seu início, a depressão típica das curvas de mascaramento cuja intensidade de excitação é bastante elevada, como se ilustra na Fig. 3.9.

O espraiamento de energia, $S(z)$ (em dB), no ponto com frequência b_i (em Bark), devido à excitação existente na frequência b_j (em Bark), é dada pela função seguinte:

$$z = b_j - b_i \tag{3.3}$$

$$se\ \ 0.5 \le z \le 2.5\ \ então\ \ d = 8(z - 0.5)(z - 2.5)\ \ senão\ \ d = 0.0 \tag{3.4}$$

$$S(z) = 15.81 + 7.5\,(z + 0.474) - 17.5\sqrt{1 + (z + 0.474)^2} + d \tag{3.5}$$

A função de espraiamento (3.5) é válida quer para excitações de perfil tonal quer para excitações de perfil incoerente [V4].

3.2.3.3.4 Avaliação Tonal do Espectro

Usando codificação por transformada, um processo expedito de avaliar o carácter coerente (tonal) ou incoerente (ruidoso) das várias componentes do espectro, é quantificar a predizibilidade da sua evolução [A18]. Assim, considerando o raio da componente espectral **f**, avaliada no segmento transformado **t**, **r(t,f)**, assim como a fase da mesma componente **ϕ(t,f)**, a predição simples da componente espectral em **t**, considerando as ocorrências em **t-1** e **t-2**, será:

$$\hat{r}(t,f) = 2r(t-1,f) - r(t-2,f) \tag{3.6}$$

$$\hat{\phi}(t,f) = 2\phi(t-1,f) - \phi(t-2,f) \tag{3.7}$$

Uma medida de predizibilidade pode ser calculada a partir da distância euclidiana entre a componente predita e a correcta:

$$p(t,f) = \frac{dist\left[(\hat{r},\hat{\phi}),(r,\phi)\right]}{\hat{r} + r} \tag{3.8}$$

A equação (3.8) fornece valores entre **1.0** e **0.0**, indicando neste último caso que a predição foi eficiente, isto é, que a componente espectral tem uma evolução coerente.

Usando uma simples relação logarítmica, converte-se a medida de predizibilidade em medida de tonalidade:

$$T(t,f) = k_1 ln[p(t,f)] + k_2 \quad ; \quad k_1,k_2 \text{ contantes} \tag{3.9}$$

Para cada componente espectral **f**, a função **T(t,f)** caracterizará a tonalidade através de um valor compreendido entre **0.0** e **1.0**. No primeiro caso, a componente espectral é totalmente impredizível e, portanto, assimilada a ruído. No segundo caso, a componente espectral é completamente predizível (a partir das ocorrências em **t-1** e **t-2**) e, por isso, é identificada como sendo um tom.

Ponderando a predizibilidade em secções do espectro, a expressão (3.9) pode também ser usada para criar (ver capítulo 4) uma medida, **α(b)**, que caracteriza regiões espectrais (referidas à escala Bark) quanto à sua tonalidade média. O índice

de mascaramento final, **IM**, em cada região espectral, é calculado através de uma interpolação linear entre os valores fornecidos pelas equações (3.1) e (3.2):

$$IM(b) = \alpha(b)[TMN(b)] + [1 - \alpha(b)]NMT(b) \quad ; \quad 0.0 \leq b \leq 26.0 \qquad (3.10)$$

O valor IM considera deste modo a tonalidade espectral, para fornecer o limiar de mascaramento a partir do diagrama de excitação.

3.2.4 Mascaramento Binauricular

3.2.4.1 Estereofonia e Imagem Acústica

A experiência diária prova que a acuidade auditiva é significativamente aumentada pelo facto de possuirmos dois ouvidos. O efeito mais óbvio é a soma binauricular de intensidades, isto é, um mesmo som captado por dois ouvidos tem uma intensidade percebida dupla da sentida quando é captado por um único ouvido. A capacidade em detectar pequenas diferenças, quer em amplitude, quer em frequência (sensibilidade diferencial), é também aumentada binauricularmente. Por outro lado, a importante capacidade em localizar uma fonte sonora, resulta do processamento binauricular de sons que atingem os dois ouvidos com diferenças de intensidade, fase e conteúdo espectral. Este processamento, provavelmente executado a níveis mais elevados do sistema auditivo, combina toda a informação para produzir uma *imagem sonora* única. É por isso designado de *fusão binauricular*.

A geração de sinais acústicos a partir de dois únicos focos (*i.e.* estereofonia) é uma projecção do espaço acústico em duas dimensões, a não ser que sejam usadas técnicas de compensação espectral especiais [P5][P11][P13]. Do ponto de vista da codificação perceptual, esta restrição representa também uma conveniência, dado que é muito mais fácil caracterizar a percepção auditiva só no plano horizontal. Deste modo, referências futuras a imagem sonora situá-la-ão implicitamente neste plano e, como tal, a expressão localização toma-se sinónima de lateralização.

3.2.4.2 Efeito de Precedência e "Coktail Party"

O processo de fusão binauricular pode ser apreciado a partir da sensação auditiva resultante de um som acompanhado de bastantes reflexões, como as que resultam num recinto fechado. As reflexões atingem os ouvidos com um certo atraso relativamente ao sinal directo. Contudo, o ouvinte é capaz de localizar a fonte de som, baseado na direcção do sinal directo e não na das reflexões, se o atraso for curto. Esta propriedade do sistema auditivo em privilegiar a informação do sinal que chega primeiro, é conhecida por *efeito de precedência*. Se o atraso anterior for "grande" relativamente à duração do som (*e.g.* 30ms para voz e 5ms para impulsos), passar-se-á da sensação de fusão total - do sinal directo com as várias reflexões - para a de reverberação e, em casos extremos, para a sensação de ecos nítidos, ou seja, réplicas isoladas do sinal directo. Estas duas últimas situações "destroem" o efeito de precedência e mesmo a inteligibilidade do som, se este for contínuo (não necessáriamente estacionário).

Um outro exemplo do efeito de precedência é o que se verifica quando dois impulsos com a mesma intensidade são aplicados com atraso relativo, um a cada ouvido. Se o atraso for inferior a 5ms, os dois impulsos são fundidos num só cuja localização, no azimute, é determinada pelo impulso que ocorre em primeiro lugar. Se o segundo impulso for tornado suficientemente intenso, ele pode "apagar" o efeito de precedência devido à antecipação temporal do primeiro. Parece portanto possível realizar um balanço em amplitudes e atraso, de modo a alterar, ou mesmo anular, o efeito e precedência. Em resumo, o efeito de precedência serve o objectivo da localização espacial e de remoção de réplicas "irrelevantes" de um sinal acústico.

Um outro efeito interessante é o associado à descriminação de um sinal, na presença de outros, semelhantes ou de natureza diferente, como por exemplo, a voz de um orador no meio de uma multidão de vozes. Designa-se por efeito de *cocktail party* e está relacionado com o desmascaramento binauricular. De facto, a voz que se pretende ouvir é considerada sinal e todas as restantes vozes representam ruído. Como a localização do ruído é diversa e não coincide exactamente com a do sinal, então, devido ao processamento binauricular, o sinal não é tão efectivamente mascarado pelo ruído, e portanto, é mais notório. Este aspecto tem uma importância crucial para o capítulo 4 e revela que o mascaramento de um sinal devido a outro é máximo quando ambos coincidem no mesmo ponto, como acontece no mascaramento monauricular. Por outras palavras, o mascaramento de um sinal

devido a outro não depende só da sua relação em frequência, em intensidade e temporal, mas depende também da sua relação espacial [P2].

3.2.4.3 Modelos Psico-acústicos

3.2.4.3.1 Intensidade e Tempo Interauriculares

Considerando uma fonte sonora colocada fora do plano vertical médio aos dois ouvidos, facilmente se reconhece que as ondas sonoras chegam aos dois ouvidos com um atraso relativo porque a distância desde a fonte até cada um dos ouvidos é diferente, e também com uma diferença de amplitudes porque maior distância significa maior atenuação e, além disso, há a considerar efeitos de reflexão, difracção e sombreamento devido em particular, à cabeça e ao tronco.

A diferença entre os instantes de chegada aos dois ouvidos da mesma frente de onda acústica é designada por Diferença de Tempo Interauricular (ITD). A diferença de amplitude ou intensidade relativa aos mesmos instantes designa-se por Diferença de Intensidade Interauricular (IID). ITD e IID formam a base principal de informação para a localização de focos sonoros.

Considerando a velocidade de propagação no ar, e admitindo que a distância média entre os dois ouvidos é 23cm, o atraso máximo de propagação entre eles é cerca de 660μs, o que corresponde ao período aproximado da frequência de 1500Hz. Deste modo, para frequências inferiores a 1500Hz, a ITD funciona como um dado inequívoco na localização, enquanto que devido a efeitos de difracção [P20], a IID não representa um elemento significativo. Para frequências superiores a 1500Hz, a ITD conduz a ambiguidades e devido ao efeito de sombreamento da cabeça, os valores de IID assumem um papel preponderante na localização. A apreciação feita é só comparativa pois a ITD continua a ter significado a altas frequências e a IID também tem importância às baixas frequências. Em particular, a localização de sinais transitórios é conseguida com valores de ITD tão pequenos quanto 6μs [P11].

Por outro lado, existindo simultaneamente dados de ITD e IID, o resultado combinado é um balanço de efeitos. Isto é, uma ITD pode cancelar ou reforçar uma IID e vice-versa [P3]. Usando o exemplo referido anteriormente, se a um ouvido for aplicado um impulso e se ao outro ouvido for aplicado um outro impulso com um atraso relativamente ao primeiro, o efeito de precedência poderá tornar-se nulo,

isto é, a localização da imagem fundida dos impulsos poderá ser central, se o segundo impulso tiver uma intensidade superior à do primeiro. Na literatura referem-se razões de balanço como por exemplo, 25μs/dB [P20]. O balanço de efeitos é interpretado como um conflito entre vários aspectos fisiológicos, envolvendo principalmente o padrão de disparo dos nervos auditivos associado ao ITD e a latência média da resposta nervosa associada a IID e dependente da ITD [P2].

3.2.4.3.2 Diferença do Nível de Mascaramento

Verificou-se no parágrafo anterior que o efeito de precedência está particularmente relacionado com a IID e a ITD. O efeito *cocktail party* está sobretudo relacionado com uma coloração do espectro de cada sinal, de acordo com a sua direcção de incidência no ouvido. Como foi apontado no parágrafo 3.2.1, esta coloração do espectro deve-se à fisionomia da cabeça (sobretudo da aurícula) e do tronco.

O desmascaramento binauricular é uma consequência directa da diferente coloração espectral em cada ouvido, de um mesmo sinal inicial. Suponha-se, por exemplo, um estímulo acústico composto por uma banda de ruído e um tom com uma intensidade tal que é mascarado pela banda de ruído. Se este conjunto for apresentado a um só ouvido ou aos dois simultaneamente, o tom é mascarado pelo ruído. Porém, se se inverter a fase do ruído ou do tom num dos ouvidos, o tom deixa de ser mascarado e torna-se audível. Ademais, se se mantiver o ruído idêntico nos dois ouvidos e se se retirar o tom num deles, também cessa o efeito de mascaramento. Esta melhoria de detecção binauricular relativamente à monauricular, é descrita pela medida da Diferença do Nível de Mascaramento (MLD).

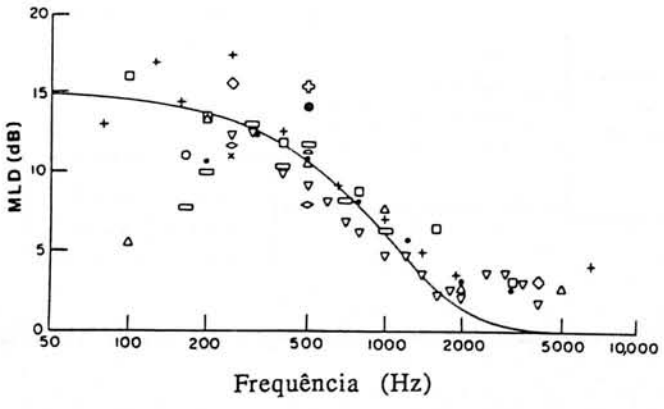

Figura 3.11: Amplitude da MLD. Gráfico adaptado de [P20].

Concretamente, a MLD (em dB) traduz a redução em amplitude que o sinal mascarado deve sofrer, em relação às condições de mascaramento monauricular, de modo a ser efectivamente mascarado em condições de audição binauricular. Um exemplo é ilustrado na Fig. 3.11 para o caso em que a fase do sinal mascarado é invertida binauricularmente. A curva média, deduzida a partir de vários testes, revela um efeito MLD particularmente intenso para baixas frequências, atenuando-se rapidamente para frequências próximas de 1500Hz. Esta frequência traduz a predominância do efeito da fase na melhoria da detecção e está intimamente ligada com a relação de sintonia dos padrões de disparo dos nervos auditivos, entre os dois ouvidos [P20]. A amplitude da MLD parece ser razoavelmente insensível à intensidade do estímulo mascarante [P5]. Se o ruído mascarante for não-correlacionado binauricularmente, contrariamente ao assumido acima, então a MLD decresce, a baixas frequências, de 15dB para 3dB.

Um resultado de fundamental importância no contexto do capítulo 4 é que a MLD e a percepção binauricular de uma forma geral, baseiam-se no processamento de sinais provenientes de bandas críticas homólogas dos dois ouvidos [P1]. Isto permite uma descrição do espectro idêntica, quer em questões monauriculares, quer em questões binauriculares. Fica assim corroborado o interesse e a conveniência de graduar o eixo das frequências em Bark.

Usando também esta última premissa, Durlach [P11][P14][P15] propôs um modelo que caracteriza a percepção binauricular, no que respeita ao efeito de MLD. Dado um estímulo acústico que incide sobre os dois ouvidos, as saídas de bandas críticas homólogas têm o processamento representado na Fig. 3.12.

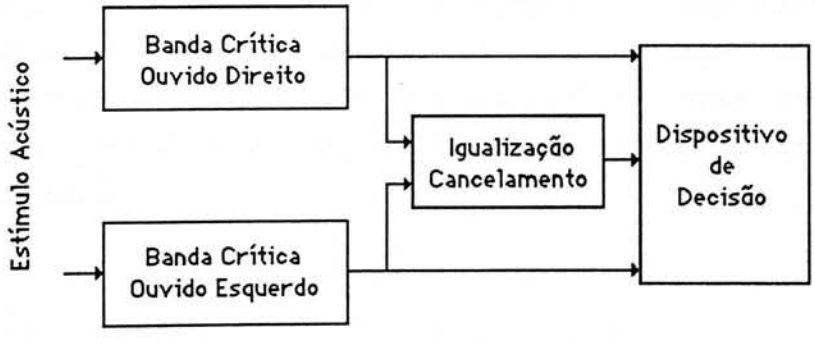

Figura 3.12: Diagrama de blocos simplificado do modelo de Equalização e Cancelamento de Durlach [P14].

Os sinais para processamento monauricular são directamente encaminhados para um dispositivo de decisão. São também conduzidos até um dispositivo de igualização e cancelamento. Aqui, ambos os sinais são transformados de modo que as componentes mascarantes comuns sejam identificadas (igualização). Estas são posteriormente canceladas, através de uma operação de subtracção entre os sinais transformados (cancelamento). A operação de cancelamento releva qualquer assimetria entre os dois sinais, pretendendo assim produzir um sinal binauricular. O dispositivo de decisão selecciona para processamento a entrada que exibir maior relação sinal-ruído.

Este modelo visa modelar resultados de testes psico-acústicos e, dada a sua complexidade, não tem grande interesse para o objectivo de codificação perceptual. Assim, considerar-se-á no capítulo 4 um modelo simples que pretende abordar aspectos fundamentais da percepção binauricular e prometedores do ponto de vista da compressão.

3.3 Conclusões

O capítulo que agora se conclui, identificou os aspectos dos sistema auditivo melhor conhecidos até ao momento e mais relevantes para o objectivo da codificação perceptual. Em particular, abordaram-se aspectos de mascaramento e identificou-se a sua expressão tanto no domínio das frequências e no dos tempos, como no espaço, cuja correlação com o conceito de *imagem acústica* foi também evidenciada.

Considerando a reprodução de sinais acústicos a partir de dois focos, como é o caso da estereofonia, a exploração dos mecanismos de mascaramento, nas suas várias vertentes, pode ser directamente convertida em ganho de codificação (compressão). No próximo capítulo, acentuar-se-á a tónica sobre soluções alicerçadas na identificação das direcções de mascaramento máximo e mascaramento mínimo.

4
CODIFICADOR PERCEPTUAL ESTEREOFÓNICO

4.1 Introdução

O presente capítulo descreve, em pormenor, a estrutura de um algoritmo para a codificação perceptual simultânea dos dois canais áudio de um par estereofónico.

Algumas das soluções apresentadas no capítulo 2 e enquadradas na codificação por transformada, são aqui retomadas e personalizadas no contexto estereofónico. Os aspectos de percepção auditiva, analisados no capítulo 3, são traduzidos em modelos psico-acústicos, de modo a proceder à extracção adequada de redundância e irrelevância, não só intracanal mas também intercanal. Neste último caso, o desmascaramento binauricular exige uma atenção particular já que é responsável pela superior acuidade auditiva, não existente no plano monauricular nem sendo sequer previsível a partir deste. De facto, o uso de codificadores perceptuais para codificar independentemente cada canal de um par estereofónico, sobretudo a muito baixas taxas de informação, conduz ao surgimento de artefactos graves na fase de reprodução. Estes estão relacionados com a localização, especialmente evidente a baixas frequências, como do ruído de quantificação que não acompanha a imagem dinâmica do sinal e, como tal, não é mascarado por este, apesar de monofónica e monauricularmente o ser.

A Fig. 4.1 representa, de forma simplificada, a estrutura do algoritmo codificador.

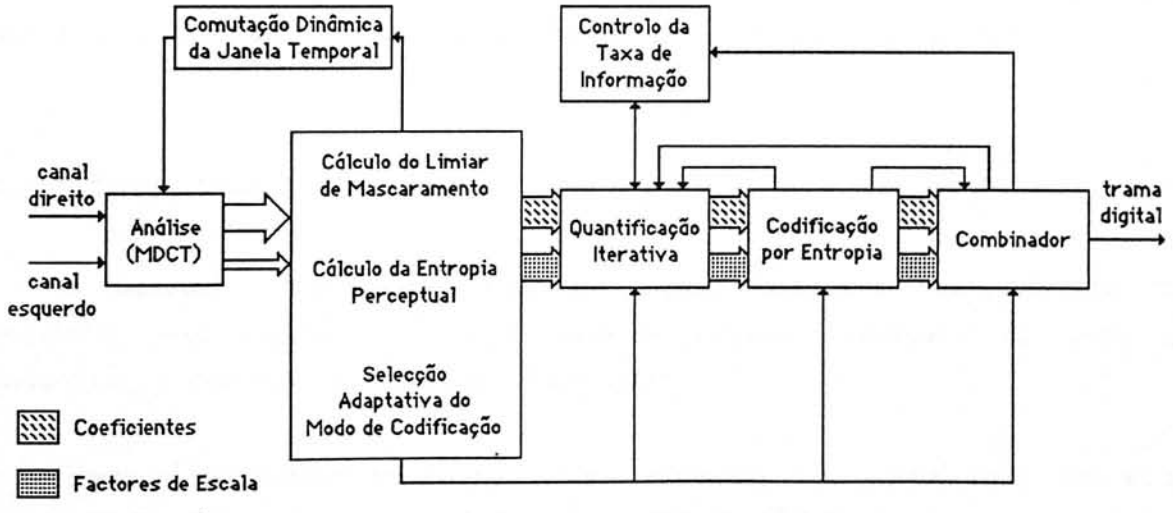

Figura 4.1: Estrutura simplificada do algoritmo de codificação.

Trata-se de um codificador por transformada que tem por objectivo comprimir transparentemente dois canais monofónicos, com a largura de banda de 20KHz e amostrados à frequência de 48KHz, numa taxa de informação total de 128Kbit/s. A transformada usada é a MDCT com diferentes resoluções para compatibilizar o ganho de codificação com as restrições do mascaramento temporal vistas no parágrafo 3.2.3.1. A quantificação e codificação dos coeficientes é realizada numa de duas possíveis bases de representação do sinal esterefónico, para o que é necessário calcular os limiares de mascaramento relativos a cada base. O processo de quantificação e codificação é iterativo, de modo a respeitar a taxa de informação constante, à saída da unidade codificadora.

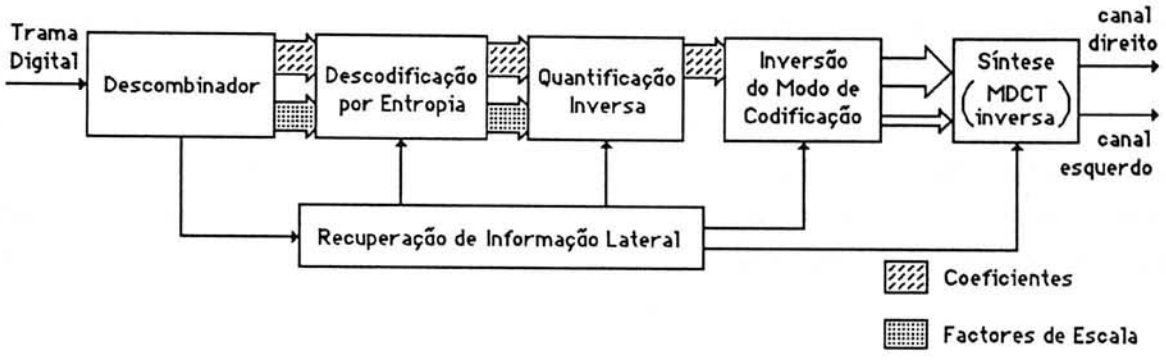

Figura 4.2: Estrutura simplificada do algoritmo de descodificação.

A unidade descodificadora está representada na Fig. 4.2. Dado que a sua função é unicamente reconstruir o sinal, a estrutura associada é bastante mais simples do que a da unidade codificadora, o que é, aliás, uma vantagem bem acolhida.

4.2 Terminologia

É necessário esclarecer, desde já, alguns conceitos omnipresentes na sequência deste capítulo e fornecer também algumas definições, de modo a simplificar a descrição do algoritmo codificador.

Como já foi referido no parágrafo 2.4.1, assume-se que o sinal áudio tem uma representação digital obtida por amostragem à taxa de 48 000 amostras por segundo. Cada amostra é quantificada linearmente em 16 *bits*. Amostras temporais consecutivas são concatenadas para formar os segmentos usados na transformação.

Serão considerados segmentos com comprimento de 1024 amostras ou de 256 amostras. O primeiro designar-se-á de *segmento longo* e a janela de amostragem associada chamar-se-á *janela longa*. De modo análogo, o segundo segmento será referido por *segmento curto* e a janela de amostragem associada por *janela curta*.

De acordo com a etapa de processamento, o espectro do sinal pode ser organizado em três formatos diferentes.

Admitindo já realizada a conversão do eixo das frequências para unidades Bark, o espectro pode ser organizado em *partições*. Estas têm uma resolução espectral aproximada de 1/3 de banda crítica. Às baixas frequências não é possível conseguir esta resolução e, por isso, adopta-se a resolução de uma linha espectral. Cada partição tem um único valor Bark associado que representa o seu ponto médio. As partições são fixas e definem a resolução espectral adequada para traduzir o efeito de mascaramento entre bandas críticas vizinhas.

O espectro pode também organizar-se em *bandas*. Cada banda agrupa um número de linhas espectrais que serão quantificadas na base de um mesmo factor de escala. Por sua vez, este é função do limiar de mascaramento. As bandas são fixas e definidas parametricamente.

Finalmente, o espectro é também organizado em *secções*. Cada secção envolve um número inteiro de bandas e representa a secção do espectro codificada com a mesma tabela de códigos de Huffman. Os limites de cada secção e o número total de secções variam de segmento para segmento.

4.3 Blocos de Análise e Síntese

O projecto das etapas de análise e síntese fundamenta-se na ponderação de vários objectivos já abordados em capítulos anteriores.

Considerando condições estacionárias, a codificação será tanto mais eficiente quanto maior for a resolução espectral da transformada porque mais rigorosa será a aplicação dos modelos psico-acústicos. A resolução espectral, assim como o ganho objectivo de codificação, aumenta com o comprimento da transformada.

Por outro lado, é importante conseguir uma boa selectividade espectral, de modo a obter coeficientes bastante descorrelacionados uns dos outros, o que permite uma quantificação mais efectiva de cada um deles. A resposta em frequência de cada linha espectral depende da transformada e da janela temporal de amostragem.

É igualmente importante dispor de uma descrição frequentemente renovada da dinâmica espectral do sinal. Este requisito, assim como a necessidade de solucionar problemas de fronteira de bloco, exige alguma sobreposição entre segmentos transformados adjacentes.

Finalmente, para minimizar a taxa total de informação, importa minimizar o número total de coeficientes espectrais codificados.

Os objectivos enunciados não se compatibilizam de forma trivial no que respeita a escolha do comprimento da transformada, a percentagem de sobreposição e a janela de amostragem temporal; o que justifica, aliás, as várias opções apresentadas no parágrafo 2.4.3.3. Um compromisso interessante é sugerido pela transformada MDCT, proposta por Princen e Bradley [T2][T3].

4.3.1 Transformada MDCT

4.3.1.1 Definição e Características

A MDCT é, na verdade, uma designação decorrente da sua similaridade com a familiar DCT. A designação original é Banco de Filtros Assegurando Cancelamento de Sobreposição Temporal (TDAC) e encerra a ideia de que a diferença entre bancos de filtros e transformada é meramente formal [T7]. O caso que nos interessa é o que envolve concatenação ímpar dos filtros [T3], cuja transformada directa tem a expressão (4.1).

$$X(k) = \sum_{n=0}^{N-1} h(n)\, x(n)\, cos\left[\frac{\pi}{2N}(2k+1)(2n+1+\frac{N}{2})\right] \quad ; \quad 0 \le k \le N-1 \qquad (4.1)$$

Nesta expressão, k é o índice espectral e $X(k)$ o coeficiente espectral correspondente, n é o índice temporal, N o comprimento da transformada, $x(n)$ é o

sinal no domínio dos tempos e $h(n)$ é a janela de análise (amostragem) temporal. A expressão (4.2) traduz a inversão de domínios.

$$y(n) = \frac{1}{N} f(n) \sum_{k=0}^{N-1} X(k) \cos\left[\frac{\pi}{2N}(2k+1)(2n+1+\frac{N}{2})\right] \quad ; \quad 0 \le k \le N-1 \qquad (4.2)$$

De modo análogo, f(n) representa a janela de síntese (amostragem) temporal e y(n) representa o segmento reconstruído.

A transformada MDCT fornece $\frac{N}{2}$ coeficientes únicos dado verificar-se a identidade (4.3).

$$X(k) = - X(N-1-k) \quad ; \quad 0 \le k \le \frac{N}{2}-1 \qquad (4.3)$$

Considerando também a propriedade de cancelamento temporal analisada no próximo parágrafo, demonstra-se [T2] que, mesmo considerando uma sobreposição de 50% entre segmentos adjacentes, a operação de transformada e, por conseguinte, qualquer sistema de codificação associado, manter-se-á criticamente amostrada(o), se as janelas de análise e síntese observarem as relações (4.4) e (4.5).

$$h(n) = f(n) \quad ; \quad 0 \le n \le N-1 \qquad (4.4)$$

$$\left[f(n + \frac{N}{2})\right]^2 + [f(n)]^2 = 2 \quad ; \quad 0 \le n \le \frac{N}{2}-1 \qquad (4.5)$$

Uma solução simples de (4.4) e (4.5) é uma janela em *seno* [A16][A30] que proporciona uma função de tranferência muito mais interessante do que as tradicionais janelas rectangulares.

$$h(n) = f(n) = \sqrt{2}\, sen\left[\frac{\pi}{N}(n+\frac{1}{2})\right] \quad ; \quad 0 \le n \le N-1 \qquad (4.6)$$

Na ausência de quantificação dos coeficientes, a MDCT com (4.6) garante reconstrução perfeita o que, como foi apontado no parágrafo 2.4.3.3, é uma propriedade importante para controlar a injecção e coloração do ruído de quantificação. A janela em *seno* é a que se considerará na sequência deste capítulo, pelo que implicitamente passará a estar associada à transformada MDCT.

4.3.1.2 Cancelamento de Sobreposição Temporal

O parágrafo anterior apresenta uma solução interessante de compromisso para os requisitos equacionados no parágrafo 4.3. De facto, para um comprimento da MDCT correspondente à resolução espectral desejada, beneficia-se da selectividade dada pela janela com a forma da função seno e beneficia-se de uma sobreposição de 50% entre segmentos adjacentes, o que garante uma monitoração eficiente da dinâmica do sinal, além de solucionar o problema de fronteira de bloco. Para além destas vantagens, o sistema conserva-se criticamente amostrado. Adicionalmente, pode-se contar com um algoritmo rápido de cálculo, como será demonstrado no parágrafo seguinte.

As propriedades enunciadas devem-se ao mecanismo de cancelamento de sobreposição temporal da MDCT que, dada a sua importância, será demonstrado a seguir, recorrendo a uma abordagem algo semelhante à usada em [T3].

Usando a seguinte igualdade:

$$\cos \alpha_n = \frac{e^{-j\alpha_n} + e^{j\alpha_n}}{2} \quad ; \quad \alpha_n = \frac{\pi}{2N}(2k+1)(2n+1+\frac{N}{2}) \tag{4.7}$$

as expressões (4.1) e (4.2) podem ser convertidas nas suas equivalentes (4.8) e (4.9), respectivamente.

$$X(k) = \sum_{p=0}^{N-1} h(p)\, x(p)\, \frac{e^{-j\alpha_p} + e^{j\alpha_p}}{2} \tag{4.8}$$

$$y(n) = \frac{1}{N} f(n) \sum_{k=0}^{N-1} X(k)\, \frac{e^{j\alpha_n} + e^{-j\alpha_n}}{2} \tag{4.9}$$

Por substituição de (4.8) em (4.9), resultará (4.10).

$$y(n) = \frac{1}{4N} f(n) \sum_{p=0}^{N-1} h(p)\, x(p) \left[\Sigma_1 + \Sigma_2 + \Sigma_3 + \Sigma_4 \right] \tag{4.10}$$

Com algumas deduções básicas, conclui-se que os integrais simbólicos da expressão anterior têm, para além da solução trivial, as soluções seguintes:

$$\Sigma_1 = \sum_{k=0}^{N-1} e^{-j(\alpha_p - \alpha_n)} = N\,(-1)^{l'}\,\delta(p-n-Nl') \quad ; \quad l' \text{ inteiro} \tag{4.11}$$

$$\Sigma_2 = \sum_{k=0}^{N-1} e^{-j(\alpha_p + \alpha_n)} = N\,(-1)^{l''}\,\delta(p-Nl''+n+1+\frac{N}{2}) \quad ; \quad l'' \text{ inteiro} \tag{4.12}$$

$$\Sigma_3 = \sum_{k=0}^{N-1} e^{j(\alpha_p + \alpha_n)} = N\,(-1)^{l'''}\,\delta(p-Nl'''+n+1+\frac{N}{2}) \quad ; \quad l''' \text{ inteiro} \tag{4.13}$$

$$\Sigma_4 = \sum_{k=0}^{N-1} e^{j(\alpha_p - \alpha_n)} = N\,(-1)^{l''''}\,\delta(p-n-Nl'''') \quad ; \quad l'''' \text{ inteiro} \tag{4.14}$$

Retomando (4.10) e associando (4.11) com (4.14) e (4.12) com (4.13), ter-se-á:

$$y(n) = \frac{1}{2}\,f(n)\left[(-1)^{l'}h(Nl'-n-1-\frac{N}{2})x(Nl'-n-1-\frac{N}{2}) + (-1)^{l''}h(Nl''+n)x(Nl''+n)\right] \tag{4.15}$$

l' e l'' são números inteiros e distintos dos anteriores. As soluções possíveis de (4.15) discretizam-se para as duas metades do segmento transformado, de acordo com as duas expressões (4.16) e (4.17).

$$y(n) = \frac{1}{2}\,f(n)\left[h(n)x(n) - h(\frac{N}{2}-1-n)x(\frac{N}{2}-1-n)\right] \quad ; \quad 0 \le n \le \frac{N}{2}-1 \tag{4.16}$$

$$y(n) = \frac{1}{2}\,f(n)\left[h(n)x(n) + h(\frac{3N}{2}-1-n)x(\frac{3N}{2}-1-n)\right] \quad ; \quad \frac{N}{2} \le n \le N-1 \tag{4.17}$$

Conclui-se imediatamente que a primeira metade do segmento reconstruído contém uma parcela de sobreposição temporal que é simétrica do sinal original (primeira metade), mas invertida nos tempos. De igual modo, a segunda metade do segmento reconstruído contém uma outra parcela de sobreposição temporal que corresponde à mesma metade do sinal original, unicamente invertida nos tempos.

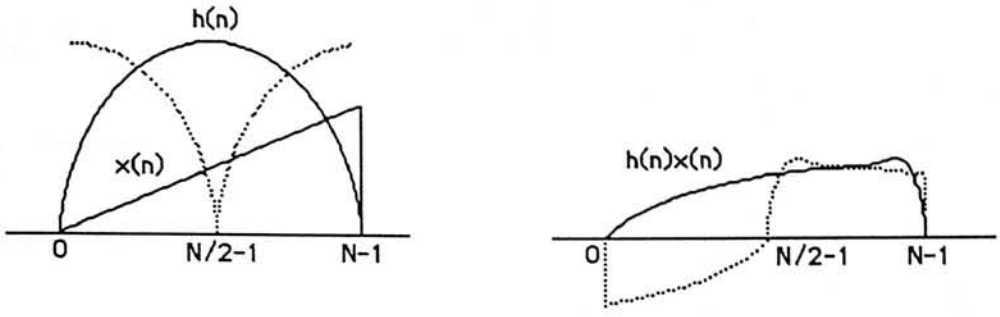

Figura 4.3: Componentes de sobreposição temporal resultantes da transformada MDCT de um segmento, em que o sinal é uma rampa (à esquerda). A janela de amostragem em *seno* é representada por $h(n)$. O produto $h(n)x(n)$ está representado, no lado direito da figura, pela curva a cheio. A linha ponteada representa as componentes de sobreposição temporal, evidenciadas em (4.16) e (4.17). A linha ponteada do lado esquerdo da figura ilustra a sobreposição entre segmentos consecutivos.

A Fig. 4.3 ilustra o exemplo em que o sinal $x(n)$ é uma rampa. A simetria dos termos invertidos revela-se particularmente conveniente. De facto, assumindo a sobreposição de 50% entre segmentos adjacentes, a primeira metade do segmento ilustrado será a segunda metade do segmento precedente que originará uma parcela, também invertida nos tempos, mas com sinal oposto ao do ilustrado. Como tal, ao sobreporem-se, cancelar-se-ão. O mesmo raciocínio aplica-se à segunda metade do segmento ilustrado.

Concretizando para a primeira metade do segmento transformado, o sinal $y(n)$ reconstruído, resultará da sobreposição da MDCT inversa do segmento m com a MDCT inversa do segmento $m-1$:

$$y(n) = y_m(n) + y_{m-1}(n+\frac{N}{2}) \quad ; \quad 0 \le n \le \frac{N}{2}-1 \tag{4.18}$$

Recorrendo a (4.16) e (4.17), ter-se-á:

$$y_m(n) = \frac{1}{2} f(n)\left[h(n)x_m(n) - h(\frac{N}{2}-1-n)x_m(\frac{N}{2}-1-n)\right] \tag{4.19}$$

$$y_{m-1}(n+\frac{N}{2}) = \frac{1}{2} f(n+\frac{N}{2})\left[h(n+\frac{N}{2})x_{m-1}(n+\frac{N}{2}) - h(N-1-n)x_{m-1}(N-1-n)\right] \tag{4.20}$$

Reconhecendo que:

$$x_{m-1}(n+\frac{N}{2}) = x_m(n) \quad ; \quad 0 \le n \le \frac{N}{2}-1 \tag{4.21}$$

$$x_{m-1}(N-1-n) = x_m(\frac{N}{2}-1-n) \quad ; \quad 0 \leq n \leq \frac{N}{2}-1 \tag{4.22}$$

resulta finalmente:

$$y(n) = \frac{1}{2}[A + B] \tag{4.23}$$

$$A = x_m(n)\left[f(n)h(n) + f(n+\frac{N}{2})h(n+\frac{N}{2})\right] \quad ; \quad 0 \leq n \leq \frac{N}{2}-1 \tag{4.23a}$$

$$B = x_m(\frac{N}{2}-1-n)\left[f(n+\frac{N}{2})h(N-1-n) - f(n)h(\frac{N}{2}-1-n)\right] \quad ; \quad 0 \leq n \leq \frac{N}{2}-1 \tag{4.23b}$$

A parcela (4.23b) é a componente de sobreposição temporal que se pretende eliminar. Para tal, o factor associado deverá anular-se, o que conduz à exigência de simetria da janela temporal de amostragem e à condição (4.4). A parcela (4.23a) representa o sinal original que se pretende recuperar. Para que se verifique uma reconstrução perfeita, o factor associado deverá ser constante e igual a 2, o que implica, através de (4.4), a condição (4.5).

Resta sublinhar que a propriedade de reconstrução perfeita da MDCT é válida na ausência de quantificação dos coeficientes. Quando existe quantificação, o ruído introduzido no domínio das frequências espraia-se em todo o segmento reconstruído. Dado que o erro de quantificação entre coeficientes do mesmo segmento ou de segmentos diferentes é não-correlacionado, não há lugar para um hipotético mecanismo de cancelamento do ruído de quantificação. Como tal, o erro de reconstrução para uma amostra temporal particular é a soma de ruídos ponderados pelas duas janelas que se sobrepõem.

4.3.1.3 Algoritmo Rápido de Cálculo

Por definição, a MDCT é uma transformada real. O cálculo dos coeficientes pode ser realizado recorrendo a um algoritmo rápido de cálculo da DCT, com as necessárias modificações [A16]. Relembrando o conceito e cálculo de uma medida de tonalidade para ponderação psico-acústica, expostos no parágrafo 3.2.3.3.4, surge evidente que, para além da informação de amplitude de cada componente espectral, é também necessário dispor de informação de fase. Por outras palavras, é necessário dispor de um transformada complexa. Uma solução simples consiste em

calcular paralelamente à MDCT, uma DFT, com o mesmo comprimento, de modo a fornecer informação para o algoritmo de caracterização tonal do espectro. Esta solução é ineficiente porque agrava a carga computacional e a complexidade de implementação. Contudo, foi adoptada na proposta ASPEC [A31] (parágrafo 2.5), o que poderá ter contribuído para a sua avaliação menos favorável aquando dos testes de Estocolmo, no âmbito do MPEG.

Realizando a transformação seguinte,

$$X(k) = \sum_{n=0}^{N-1} x'(n) \cos\left[\frac{\pi}{2N}(2k+1)(2n+1+\frac{N}{2})\right] = Re\left\{\sum_{n=0}^{N-1} x'(n) e^{-j\frac{\pi}{2N}(2k+1)(2n+1+\frac{N}{2})}\right\} \quad (4.24)$$

resulta evidente que, se a MDCT for tomada como a parte real da transformada complexa indicada, a vantagem da operação de transformação em conservar-se criticamente amostrada, apesar da sobreposição entre segmentos adjacentes, deve-se, na realidade, a uma operação de decimação no domínio das frequências e que consiste em anular as componentes imaginárias. A transformada complexa anterior não tem um algoritmo rápido de cálculo. Porém, é possível isolar uma exponencial complexa (4.25).

$$X(k) = Re\left\{e^{-j\frac{\pi}{2N}(2k+1)(1+\frac{N}{2})} \sum_{n=0}^{N-1} x'(n) e^{-j\frac{\pi}{N}(2k+1)n}\right\} = Re\left\{e^{-j\frac{\pi}{2N}(2k+1)(1+\frac{N}{2})} A(k)\right\} \quad (4.25)$$

O factor complexo tem módulo unitário para qualquer índice espectral, não contribuindo, portanto, para informação de amplitude. Por outro lado, para um **k** particular, a sua fase é constante, o que pode ser entendido como uma componente contínua (DC) cuja informação é irrelevante para caracterizar a evolução de fase do coeficiente **k**. Conclui-se, deste modo, que a informação tonal pode ser obtida a partir de **A(k)**. A transformada **A(k)** é a transformada de Fourier em Frequência Ímpar [T6] e pode ser calculada a partir de um algoritmo rápido do tipo FFT, usando uma estratégia de decimação no tempo [A13]. Finalmente, os coeficientes reais da MDCT são calculados a partir de **A(k)**:

$$X(k) = Re\{A(k)\}\cos\left[\frac{\pi}{2N}(2k+1)(1+\frac{N}{2})\right] + Im\{A(k)\}sen\left[\frac{\pi}{2N}(2k+1)(1+\frac{N}{2})\right] \; ; \; 0 \le k \le \frac{N}{2}-1 \quad (4.26)$$

A expressão (4.26) pode ser eficientemente calculada explorando a simetria e periodicidade das funções *seno* e *cosseno*.

Deste modo, conseguiu-se utilizar parte do esforço computacional associado à MDCT, para derivar informação tonal do espectro. Este é também um exemplo da simplificação dos diversos módulos de simulação do codificador perceptual, cuja necessidade se revelou imperiosa dado o grande número de variáveis e operações envolvidas.

4.4 Comutação Dinâmica de Janela de Amostragem Temporal

4.4.1 Detecção de Não-estacionaridades

A comutação dinâmica da janela de amostragem temporal é uma estratégia eficiente para melhorar a codificação perceptual em regiões do sinal caracterizadas por não-estacionaridades. Nestas regiões, entram em jogo aspectos de mascaramento temporal que impõem o aumento da resolução temporal do processo de codificação.

Tanto objectivamente (parágrafo 2.3) como perceptualmente (parágrafo 4.3), interessa aumentar o comprimento da transformada para melhorar o ganho de codificação. Janelas longas, com 1024 amostras, revelam-se adequadas, tanto para representar secções do sinal em que este se mantém, na maior parte dos casos, pseudo-estacionário, como para fornecer a resolução espectral (cerca de 47Hz) adequada à resolução das bandas críticas. O maior ganho de compressão assegurado por segmentos longos é só comprometido quando se torna necessário aumentar a resolução temporal do processo de codificação. Dos aspectos de mascaramento no domínio dos tempos, o efeito de pré-mascaramento é o mais restritivo e a partir de testes informais desaconselha resoluções temporais inferiores a 5ms. Janelas curtas de 256 amostras são, deste ponto de vista, razoavelmente adequadas.

Assumindo os dois comprimentos básicos referidos para a janela de amostragem e reflexamente, para a transformada utilizada, resta definir o critério de comutação. A relação sinal-ruído de reconstrução [A1] não é perceptualmente adequada para ordenar a comutação. A informação obtida a partir de um filtro diferenciador [A31], usado para detectar transições acentuadas do sinal, também não é perceptualmente eficiente para definir condições de comutação.

Para o codificador perceptual estereofónico, foi adoptado um critério de decisão que se baseia na medida de Entropia Perceptual (PE). A PE de um segmento particular fornece o limite inferior teórico da informação necessária para codificar transparentemente esse segmento. Dados os **L** coeficientes quantificados, $x_q(k)$, de um segmento codificado, a Entropia Perceptual é calculada através da expressão (4.27) [A10].

$$PE = \frac{\sum_{k=0}^{L-1} log_2 \left[2/x_q(k)/+1 \right]}{L} \qquad (4.27)$$

Intrinsecamente, a PE tem memória que está relacionada com uma protecção parcial de pré-mascaramento, estabelecida a partir do segmento precedente [A24]. Em consequência, a PE revela um súbito incremento, relativamente à PE do segmento anterior, quando se está em presença de uma não-estacionaridade do sinal. Esta propriedade importante é usada para activar um mecanismo de comutação de janela que procurará localizar o "início" da não-estacionaridade e decidir a melhor solução de comutação (parágrafo 4.4.3).

4.4.2 Tipos de Janelas

A utilização da MDCT impõe a observância de dois princípios: cancelamento da sobreposição temporal e conservação do sistema criticamente amostrado. Se as janelas longas e as janelas curtas forem definidas pela expressão (4.6), as janelas que realizem a transição entre estes dois tipos de janelas deverão também respeitar aqueles dois princípios. É possível efectuar a transição com base em janelas longas e curtas; porém, a sequência temporal é violada, o que favorece o aparecimento de pré-ecos. B. Edler [T1] sugeriu duas soluções que também respeitam a sequência temporal. Designam-se por janelas de *início* e *fim* de transição e têm o comprimento de um segmento longo.

$$h_{início}(n) = \sqrt{2}\ sen\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)\right] \quad ; \quad 0 \le n \le \frac{N}{2}-1 \qquad (4.28a)$$

$$h_{início}(n) = \sqrt{2} \quad ; \quad \frac{N}{2} \le n \le \frac{11N}{16}-1 \qquad (4.28b)$$

$$h_{início}(n) = \sqrt{2}\ sen\left[\frac{4\pi}{N}(n-\frac{9N}{16}+\frac{1}{2})\right] \quad ; \quad \frac{11N}{16} \le n \le \frac{13N}{16}-1 \tag{4.28c}$$

$$h_{início}(n) = 0 \quad ; \quad \frac{13N}{16} \le n \le N-1 \tag{4.28d}$$

O primeiro tipo de janelas, definido pela expressão (4.28), estabelece a transição entre janelas longas e janelas curtas. O segundo tipo, definido pela expressão (4.29), estabelece a transição inversa.

$$h_{fim}(n) = 0 \quad ; \quad 0 \le n \le \frac{3N}{16}-1 \tag{4.29a}$$

$$h_{fim}(n) = \sqrt{2}\ sen\left[\frac{4\pi}{N}(n-\frac{3N}{16}+\frac{1}{2})\right] \quad ; \quad \frac{3N}{16} \le n \le \frac{5N}{16}-1 \tag{4.29b}$$

$$h_{fim}(n) = \sqrt{2} \quad ; \quad \frac{5N}{16} \le n \le \frac{N}{2}-1 \tag{4.29c}$$

$$h_{fim}(n) = \sqrt{2}\ sen\left[\frac{\pi}{N}(n+\frac{1}{2})\right] \quad ; \quad \frac{N}{2} \le n \le N-1 \tag{4.29d}$$

A Fig. 4.4 representa, sem rigor de escala, os vários tipos de janelas. As janelas curtas são associadas em grupos de 4, de modo a representar o mesmo número de coeficientes que uma janela longa, se bem que representem um número diferente de amostras temporais. Em cada instante, ambos os canais do par estereofónico são codificados com o mesmo tipo de janela, por forma a permitir a extracção de redundância e irrelevância intercanal.

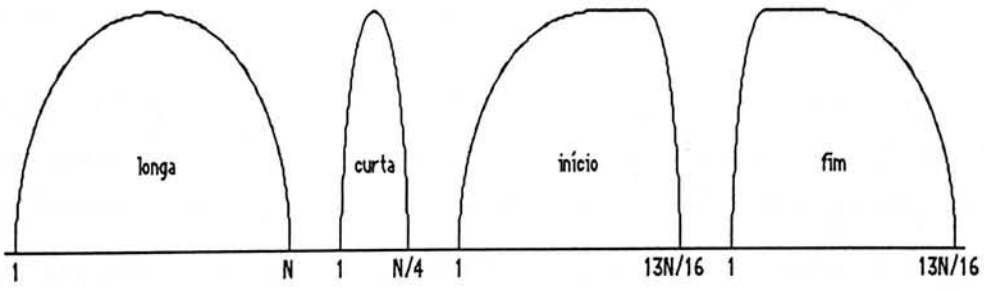

Figura 4.4: Aspecto das janelas de amostragem temporal. N equivale a 1024 amostras e o valor máximo atingido pelas janelas é $\sqrt{2}$. A janela *início* efectua a transição entre a janela *longa* e a *curta*. A janela *fim* realiza a transição inversa.

4.4.3 Alternativas de Comutação

Usando os valores da Entropia Perceptual, a estacionaridade do sinal é monitorada em duas fases. Primeiro, através dos valores da PE associados a segmentos longos. Se resultar denúncia de alguma não-estacionaridade, então esta é identificada por uma análise mais pormenorizada, baseada em segmentos associados a janelas curtas. Assim, a PE é calculada sempre para janelas longas, enquanto que, para janelas curtas, a PE só é calculada quando se torna necessário. Não obstante, a tonalidade é sempre calculada e actualizada para ambos os tipos de janelas, de modo a acompanhar a evolução do sinal.

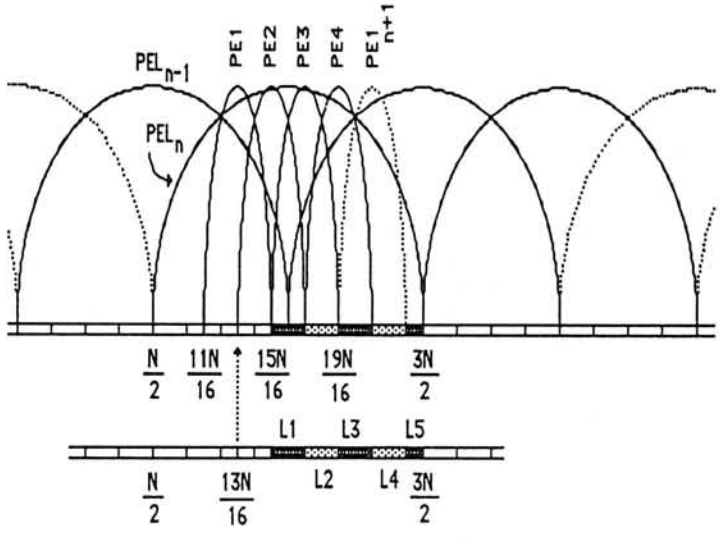

Figura 4.5: Quando o segmento desde $\frac{N}{2}$ até $\frac{3N}{2}$ está a ser codificado, é necessário dispôr das medidas de PE indicadas. A codificação do segmento *n-1* é só completamente definida quando se determinar a necessidade e tipo de comutação para o segmento *n*.

A Fig. 4.5 traduz toda a informação relevante que é necessário obter quando o segmento entre $\frac{N}{2}$ e $\frac{3N}{2}$ está a ser codificado. A primeira fase de monitoração referida anteriormente traduz-se no cálculo da diferença $PEL_n - PEL_{n-1}$. Se esta diferença for inferior a um limiar definido, o segmento é codificado com base na janela longa e passa-se à análise do segmento seguinte. Caso contrário, a existência de uma não-estacionaridade dentro desse segmento é declarada e são tomados outros passos, descritos a seguir, para decidir a alternativa correcta de comutação.

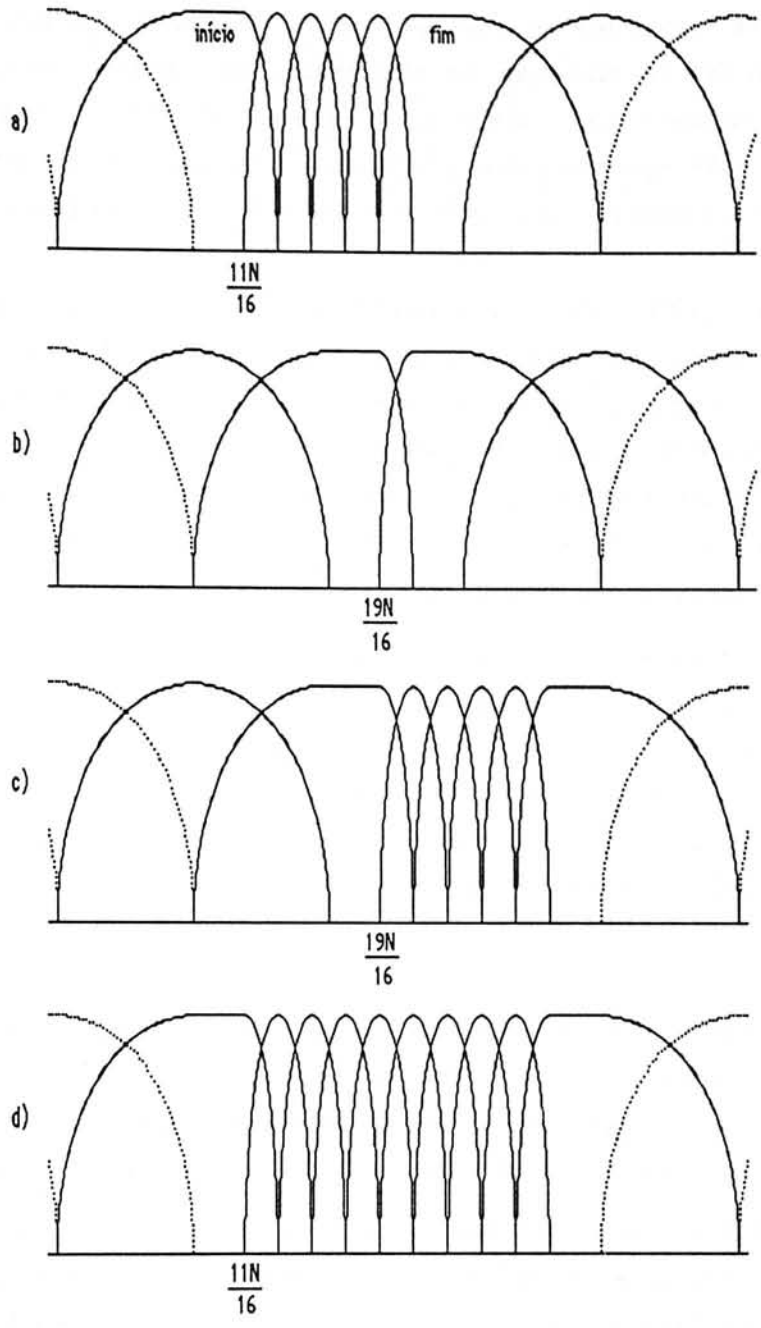

Figura 4.6: Sequências possíveis de codificação de acordo com a localização de não-estacionaridades do sinal.

Começa-se por calcular sequencialmente a PE de janelas curtas, assinaladas por PE2, PE3 e PE4. Correspondentemente, há 5 posições possíveis para o "começo" da não-estacionaridade, assinaladas por L1, L2, L3, L4 e L5. Como se tornará evidente em breve, se o "começo" tivesse ocorrido num ponto entre $\frac{N}{2}$ e $\frac{15N}{16}$, esta situação teria sido detectada ainda no segmento anterior. Resulta que o valor de PE1 não contém informação relevante para o segmento em análise. Contudo, o limiar de mascaramento associado a PE1 é necessário para o cálculo de PE2. Com base em

simulações, foi também observado e concluído que a informação de estacionaridade fornecida por janelas curtas, reside sobretudo na amplitude da PE individual e não na diferença entre a PE de janelas consecutivas, como acontece para janelas longas. Neste espírito, o objectivo é detectar a janela curta cuja PE exceda um limite predefinido, o que determina as 4 soluções de comutação ilustradas na Fig. 4.6.

Se a janela detectada for a correspondente a PE2 ou PE3, então o "começo" ocorre em L1 or L2. Estas posições situam-se confortavelmente no interior da sequência de janelas curtas e a codificação rege-se pela sequência final ilustrada na Fig. 4.6a. A necessidade de se colocar uma janela *início* antes das janelas curtas, obriga a atrasar a codificação do segmento anterior. Por esta razão, e considerando a sobreposição de 50% entre segmentos adjacentes, é necessário ter permanentemente em memória $\frac{3N}{2}$ (=1536) amostras temporais. Se a janela detectada for a associada a PE4, então o "começo" da não-estacionaridade ocorre em L3 e, na pior da hipóteses, pode situar-se nas proximidades da fronteira direita da última janela curta. Várias simulações provaram consistentemente que a colocação de uma janela *fim* de transição nestas circunstâncias, sob condições reais de codificação, degrada significativamente a reconstrução do sinal. Por esta razão, coloca-se uma outra sequência de 4 janelas curtas. A sequência final de codificação está ilustrada na Fig 4.6d.

Se não for detectada nenhuma janela curta, as restantes possibilidades são L4 ou L5. O problema passa a enquadrar-se no âmbito do futuro segmento a codificar e, por isso, torna-se necessário calcular a PE da sua primeira janela curta referida na Fig 4.5 por $PE1_{n+1}$. Se o seu valor for superior ao limite predefinido, então o "começo" da não-estacionaridade encontra-se em L4, caso contrário encontra-se em L5. No primeiro caso, a janela *início* pode ser seguida por uma janela *fim* porque o espraiamento do ruído de quantificação é equivalente ao de uma janela curta e o ganho de codificação resulta melhorado (Fig. 4.6b). No último caso, a sequência de codificação é a correspondente à Fig. 4.6c e equivale às mesmas condições referidas para a Fig. 4.6a. A prova deste facto pode ser obtida confirmando que $PE2_{n+1}$ é, na verdade, superior ao limite predefinido.

Dado que em cada instante se faz uso do mesmo tipo de janela para codificar ambos os canais do par estereofónico, resulta que a comutação de janelas nos dois canais ocorre quando pelo menos um a exige.

As várias opções de comutação apresentadas resultaram de numerosas simulações apoiadas no algoritmo monofónico exposto no próximo parágrafo e usando diversos trechos musicais, sobretudos ricos em ocorrências de "ataque". A Fig. 4.7 representa um ataque original de "Castanholas". Quando a codificação é realizada só com blocos longos, e à taxa total de 128Kbit/s para os dois canais, verifica-se que o espraiamento do ruído de quantificação é particularmente evidente até cerca de 6.5ms antes do ataque propriamente dito (Fig. 4.8). Este pré-eco é perfeitamente audível. Porém, quando para a mesma taxa, o algoritmo é livre de escolher a sequência de janelas, o sinal reconstruído é o reproduzido na Fig. 4.9. Neste caso, o ruído de quantificação traduz-se num pré-eco de cerca de 1ms que não é audível. É importante sublinhar que apesar do pré-eco para janelas curtas ser cerca de 2.5ms em média, este valor é reduzido pela margem de sobrecodificação (codificação para além do exigido pelo limiar de mascaramento) do segmento associado. Ainda para o caso da Fig 4.9, a maior fidelidade do sinal reconstruído corrobora também a necessidade de codificar sinais com grande largura de banda em segmentos mais curtos, nos quais aqueles se podem considerar pseudo-estacionários.

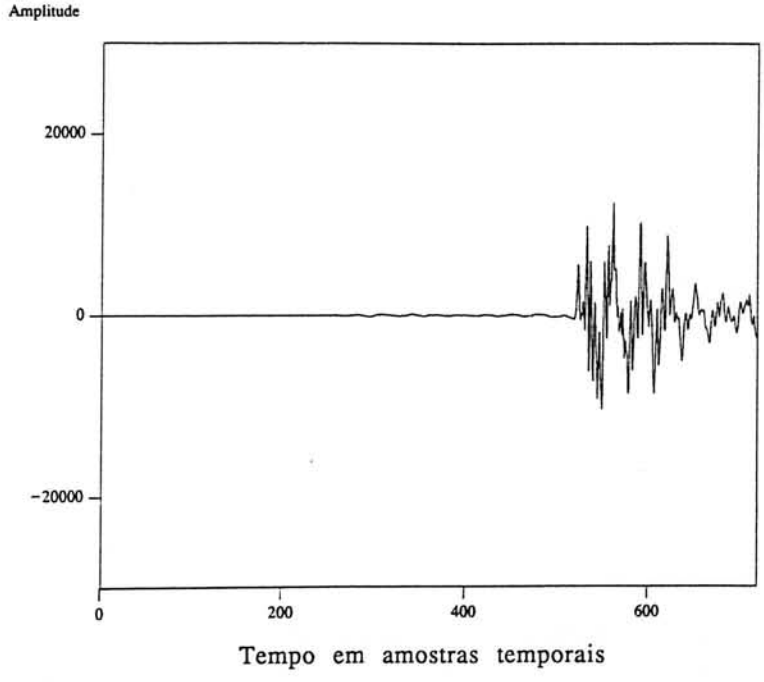

Figura 4.7: Reprodução de um ataque original do trecho musical "Castanholas".

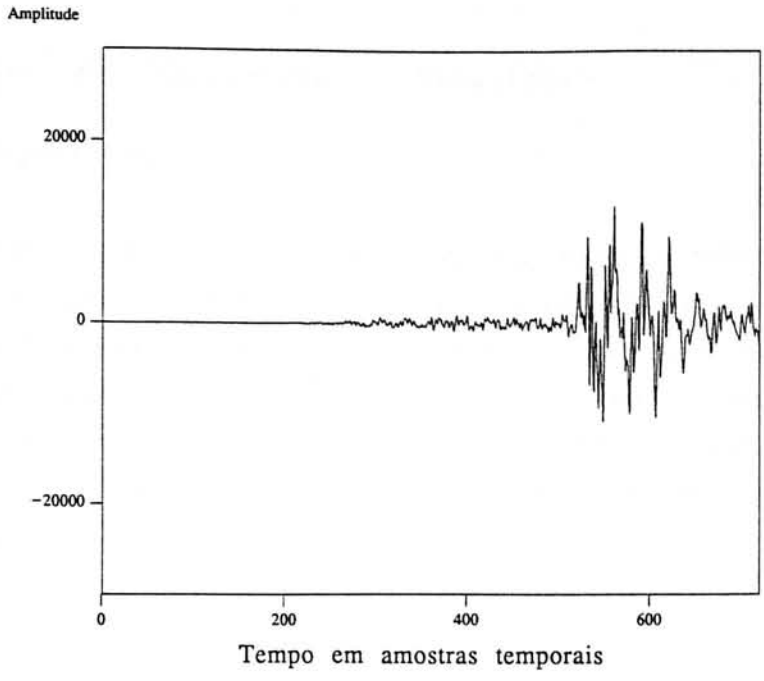

Figura 4.8: Sinal da Fig. 4.7 reconstruído quando a codificação é realizada só com janelas longas.

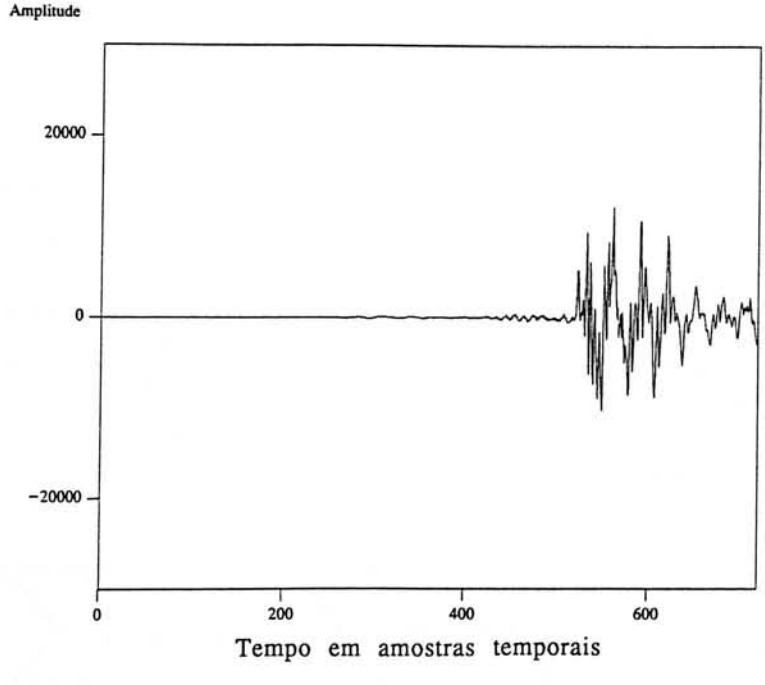

Figura 4.9: Sinal da Fig. 4.7 reconstruído quando a codificação é realizada com comutação de janelas.

4.5 Limiar de Mascaramento Monofónico

4.5.1 Objectivos

O limiar de mascaramento monofónico representa a máxima quantidade de ruído, convenientemente disposto ao longo do espectro, que é mascarado pelo sinal, quando este é apresentado simultaneamente a um ou aos dois ouvidos. Far-se-á de seguida, um resumo da metodologia desenvolvida para postular o limiar de mascaramento monofónico. Tomar-se-á como base, o modelo psico-acústico referido no parágrafo 3.2.3.3 para representar as propriedades analíticas monauriculares.

4.5.2 Distribuição de Energia e Predizibilidade

Representando, sem perda de generalidade, o número total de partições por *npart*, a frequência em Hertz por ω e em Bark por **b**, define-se a energia $\mathbf{e(p)}$ e a predizibilidade $\mathbf{P(p)}$, associadas a cada partição **p**, por (4.30) e (4.31), respectivamente.

$$e(p) = \frac{\sum_{i=0}^{npart-1} \left[S[B(i)-B(p)] \sum_{\omega=\omega_{inf}(i)}^{\omega_{sup}(i)} \left[r(\omega) \right]^2 \right]}{\sum_{j=0}^{npart-1} S[B(j)-B(p)]} \quad ; \quad 0 \le p \le npart-1 \tag{4.30}$$

$$P(p) = \frac{\sum_{i=0}^{npart-1} \left[S[B(i)-B(p)] \sum_{\omega=\omega_{inf}(i)}^{\omega_{sup}(i)} p(\omega) \left[r(\omega) \right]^2 \right]}{\sum_{i=0}^{npart-1} \left[S[B(i)-B(p)] \sum_{\omega=\omega_{inf}(i)}^{\omega_{sup}(i)} \left[r(\omega) \right]^2 \right]} \quad ; \quad 0 \le p \le npart-1 \tag{4.31}$$

As funções $\mathbf{r}(\omega)$ e $\mathbf{p}(\omega)$ foram introduzidas no parágrafo 3.2.3.3.4. A função espraiamento (3.5) recorre à função $\mathbf{B}(\mathbf{p})$ que fornece a frequência Bark associada à partição $\mathbf{p}$. A energia em cada partição é o somatório simples das influências das energias em todas as partições. Cada influência traduz a ponderação exercida pela função espraiamento cujo ganho é considerado para normalizar o somatório anterior. A medida de predizibilidade de cada coeficiente espectral é igualmente ponderada num somatório para fornecer a predizibilidade média de cada partição.

4.5.3 Tonalidade Espectral

Para cada partição, a medida de predizibilidade média, $\mathbf{P}(\mathbf{p})$, fornece um valor compreendido entre **1.0** e **0.0**. A conversão para a medida de tonalidade média é feita através da relação logarítmica (4.32), depois de limitar $\mathbf{P}(\mathbf{p})$ inferiormente a **0.05** e superiormente a **0.5** [A18].

$$\alpha(p) = -0.299-0.43\ln[P(p)] \quad ; \quad 0 \leq p \leq npart-1 \tag{4.32}$$

$\alpha(p)$ fornece para cada $\mathbf{p}$ um valor compreendido entre **0.0** e **1.0**, identificando, no primeiro caso, a partição $\mathbf{p}$ como sendo totalmente incoerente e no segundo caso, como sendo totalmente tonal.

4.5.4 Energia Máxima do Ruído de Quantificação

Retomando a terminologia de Zwicker, pode-se afirmar que (4.30) representa o diagrama de excitação originado pelo sinal monofónico. O limiar de mascaramento é determinado deduzindo o índice de mascaramento (3.10) ao diagrama de excitação. Por sua vez, o índice de mascaramento resulta das expressões de mascaramento de ruído devido a tons (3.1) e de mascaramento de tons devido a ruído (3.2), convenientemente ponderados pela tonalidade média (4.32) de cada partição.

Por definição de bandas críticas, as expressões (3.1) e (3.2) são também válidas para bandas sub-críticas e, em particular, para as partições, o que valida a afirmação anterior.

Na sequência de várias simulações, tornou-se evidente que o modelo psico-acústico apoiado nas expressões (3.1) e (3.2) tinha um desempenho inferior ao expectável para alguns trechos musicais. Concretamente, para sons eminentemente tonais, concluíu-se que a expressão (3.1) sobrecodificava as altas frequências e subcodificava as baixas frequências. Para sons eminentemente incoerentes, a expressão (3.2) também revelou favorecer a codificação das altas frequências em detrimento das baixas.

Dada a indisponibilidade de tempo para realizar testes psico-acústicos exaustivos, procurou-se "adaptar" as expressões referidas aos parâmetros do codificador perceptual, forçando a codificação no limiar de mascaramento e usando sons sintetizados e alguns trechos musicais mais críticos. Assume-se que o mesmo modelo psico-acústico é válido quer para segmentos longos, quer para segmentos curtos.

4.5.4.1 Modelo Psico-acústico Modificado

4.5.4.1.1 Tons Mascarando Ruído

O trecho musical particularmente útil para testar o modelo de tons mascarando ruído é composto por vários 'toques' de um "triângulo" cujo som é idêntico ao produzido por um diapasão. Adicionalmente, sintetizaram-se cinco trechos com diversas combinações de tons puros, desde 40Hz até 14KHz. Por um processo de ajuste iterativo, obteve-se a expressão (4.33) que é uma curva modificada de (3.1).

$$TMN'_{dB}(b) = 19.5 - \frac{18.0}{26.0} b \quad ; \quad 0 \leq b \leq 26.0 \tag{4.33}$$

Para as várias condições de teste ensaiadas, esta expressão conduziu a melhores resultados. Em particular, a PE média para o trecho "triângulo" desceu de 0.476bit/amostra (0.414 para o outro canal) para 0.465bit/amostra (0.404 para o outro canal), com um notório aumento de qualidade.

4.5.4.1.2 Ruído Mascarando Tons

Como já foi referido, a expressão (3.2) é assumida como uma solução pragmática para a circunstância concreta de interesse no contexto da codificação perceptual: mascaramento do ruído de quantificação, devido a um sinal incoerente. O melhor sinal incoerente disponível na biblioteca - com um total de 22 trechos musicais - é o correspondente a "castanholas" e compreende 43 'toques' acompanhados de eco e reverberação. Forçando ainda a codificação no limiar de mascaramento, concluíu-se que a expressão (4.34) era adequada para modelar o efeito pretendido de mascaramento.

$$NMT'_{dB}(b) = 6.56 - \frac{3.06}{26.0} b \quad ; \quad 0 \leq b \leq 26.0 \tag{4.34}$$

A PE média de castanholas melhorou de 0.375bit/amostra (0.326 para o outro canal) para 0.361bit/amostra (0.312 para o outro canal), o que se traduziu num aumento expressivo da qualidade de codificação. Tanto (4.34) como (4.33) representam uma pequena modificação das suas homólogas (3.2) e (3.1), através da inclinação das curvas iniciais, em torno de ponto a 8.5Bark e no sentido de beneficiar a codificação a baixas frequências.

4.5.4.2 Energia de Quantificação

O índice de mascaramento (3.10) tem um valor para cada partição dado pela interpolação simples entre os das expressões (4.33) e (4.34).

$$IM'(p) = \alpha[B(p)]TMN'[B(p)] + [1-\alpha[B(p)]]NMT'[B(p)] \quad ; \quad 0 \leq p \leq npart-1 \tag{4.35}$$

A energia de quantificação em cada partição obtém-se através da subtracção logarítmica entre a energia de excitação e o índice de mascaramento:

$$e_q(p) = e(p)\ 10^{-\frac{IM'(p)}{10}} \quad ; \quad 0 \leq p \leq npart-1 \tag{4.36}$$

A Fig. 4.10 representa o espectro de um segmento longo do trecho musical "Suzanne Vega". Sobreposto ao espectro está um conjunto de três curvas com uma disposição em escada e em que cada degrau corresponde, de facto, a uma partição. A

curva inferior representa a energia de quantificação obtida se, por hipótese, o sinal fosse totalmente tonal, isto é, se $\alpha[B(p)]= 1.0$ para qualquer p. A curva superior representa a energia de quantificação que se obteria se o sinal fosse totalmente incoerente. A curva intermédia traduz a ponderação de acordo com as reais características tonais do sinal. É possível concluir que para baixas frequências o sinal é essencialmente tonal. Para médias e altas frequências, o sinal tem um comportamento médio entre tom e ruído. Por outro lado, verifica-se também que o espraiamento do efeito de mascaramento produzido pela parte inferior do espectro, torna a parte superior do espectro irrelevante e passível, na maior parte, de uma quantificação nula.

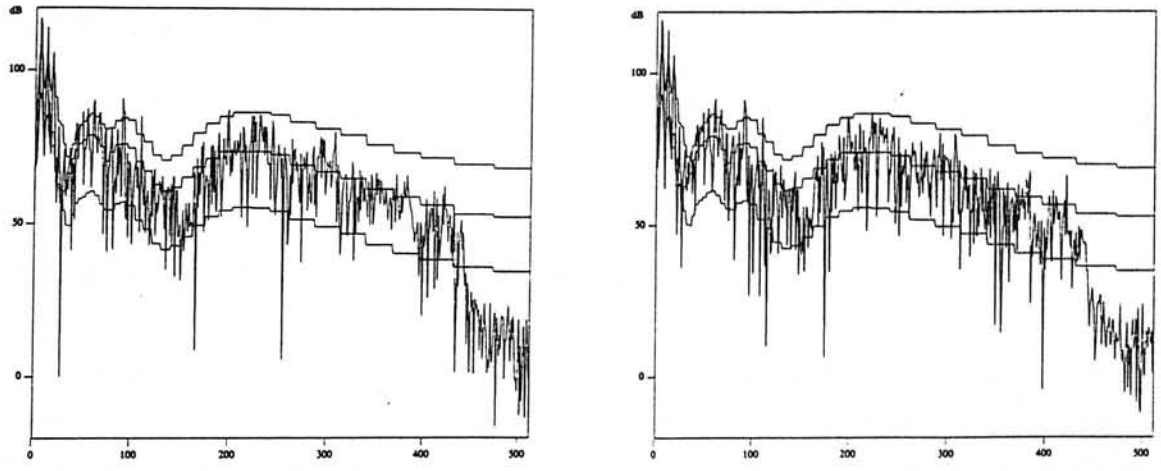

Figura 4.10: Espectro de um segmento longo de "Suzanne Vega" para o canal direito (gráfico na esquerda) e para o canal esquerdo (gráfico na direita). O eixo das abcissas representa a frequência em linhas espectrais. A curva (em escada) superior representa a energia máxima de quantificação que conduziria a uma codificação transparente se o sinal tivesse tonalidade média nula para todo o espectro. A curva inferior assume tonalidade média unitária para todo o espectro. A curva intermédia inclui a tonalidade real do sinal. É possível concluir que só algumas componentes espectrais, particularmente a baixas frequências, é que contêm informação relevante. As restantes ou são completamente mascaradas (sofrendo quantificação nula) ou admitem uma quantificação bastante efectiva (necessitando de muito poucos *bits* para representar os coeficientes).

A distribuição da energia de quantificação em cada linha espectral é definida por (4.37).

$$e_q(\omega) = \frac{e_q(p)}{\omega_{sup}(p)-\omega_{inf}(p)+1} \quad ; \quad 0 \le p \le npart\text{-}1 \quad ; \quad \omega_{inf}(p) \le \omega \le \omega_{sup}(p) \qquad (4.37)$$

O limiar de mascaramento final é obtido a partir dos valores fornecidos por (4.37) e considerando ainda o limiar absoluto de mascaramento e um controlo parcial de pré-eco.

4.5.4.3 Ajuste para o Limiar Absoluto de Audição

O limiar absoluto de audição ou mascaramento traduz a máxima quantidade de ruído que pode ser injectado, de forma imperceptível, em *qualquer* sinal. A curva da Fig. 3.1 é implementada em relação ao ponto de maior sensibilidade a 4KHz, que é de facto inaudível, se for concretizado como uma onda sinusoidal com amplitude de pico igual a meio dígito menos significativo (1/2 LSB).

Representando a energia do limiar absoluto de mascaramento por $absthr(\omega)$, a energia de quantificação ajustada virá:

$$e_{qa}(\omega) = max\left[e_q(\omega)\ ,\ absthr(\omega)\right] \quad ; \quad 0 \le \omega \le \pi \tag{4.38}$$

4.5.4.4 Controlo Parcial de Pré-eco

O controlo parcial de pré-eco, já referido no parágrafo 4.4.1 para justificar a memória intrínseca da medida de Entropia Perceptual, materializa um processo de protecção que conduz à sobrecodificação do segmento no qual se verifica o "começo" da não-estacionaridade do sinal. Postula-se [A24] que o pré-eco é inaudível se o ruído de quantificação for inferior ao limiar de mascaramento calculado, quando se "retira" a componente não-estacionária. Como não se pode calcular este limiar, usa-se o limiar do segmento precedente para o estimar.

$$e'_{qa(n)}(\omega) = min\left[e_{qa(n)}(\omega)\ ,\ Ke_{qa(n-1)}(\omega)\right] \quad ; \quad 0 \le \omega \le \pi \tag{4.39}$$

K representa uma constante e o índice $(n-1)$ refere o segmento anterior ao actual (n). Em rigor, **K** deveria ser **1.0** para se conseguir uma protecção efectiva, na ausência de outra qualquer. Como o objectivo é só uma protecção parcial, pois dispõe-se da estratégia de comutação de janelas de amostragem temporal, limita-se a diminuição da energia de quantificação impondo, por exemplo, **K=2.0**.

4.5.4.5 Energia de Quantificação em Bandas Espectrais

A quantificação dos coeficientes é efectuada com base em factores de escala. Cada factor de escala é único para a mesma banda ou grupo de coeficientes. Os limites de cada banda são definidos parametricamente e resultaram da caracterização estatística média de porções do espectro, passíveis de terem a mesma energia de quantificação. Designando o número total de bandas por $nband$, a energia de quantificação, em cada coeficiente da mesma banda, será dada por (4.40a) ou (4.40b)

$$e_{banda}(i) = \frac{\sum_{\omega=\omega_{inf}(i)}^{\omega_{sup}(i)} e'_{qa}(\omega)}{\omega_{sup}(i) - \omega_{inf}(i) + 1} \quad se\ banda(i)=1 \quad ; \quad 0 \le i \le nband\text{-}1 \tag{4.40a}$$

$$e_{banda}(i) = min\left[e'_{qa}\left[\omega_{inf}(i)\right], \ldots, e'_{qa}\left[\omega_{sup}(i)\right]\right] \quad se\ banda(i)=0 \tag{4.40b}$$

O parâmetro auxiliar $banda(i)$ identifica a banda i como sendo perceptualmente estreita ($banda(i)=1$) ou perceptualmente larga ($banda(i)=0$). A classificação deve-se a uma comparação simples da largura de cada banda em Hertz, com a correspondente largura em Bark.

4.5.5 Factores de Escala

Os factores de escala são as entidades que representam o limiar de mascaramento. Admite-se que, se a energia de quantificação (4.39) for repartida de igual modo por todos os coeficientes da cada banda (4.40), então verificar-se-á mascaramento total. Admitindo ainda que o ruído de quantificação tem amplitude de pico δ e que a distribuição probabilística de amplitudes é uniforme, então a sua energia será $\frac{\delta^2}{12}$ [A10]. A partir da energia de quantificação fornecida por (4.40), o passo de quantificação é imediato:

$$\delta(i) = \sqrt{12 e_{banda}(i)} \quad ; \quad 0 \le i \le nband\text{-}1 \tag{4.41}$$

A relação sinal-ruído mascarado (SMR) (factor de escala) transmitida ao descodificador, é obtida por quantificação logarítmica do passo de quantificação:

$$SMR(i) = nint\left[\frac{20}{1.5}\log_{10}\frac{\delta(i)}{min[absthr(\omega)]}\right] \quad ; \quad 0 \le i \le nband\text{-}1 \quad ; \quad 0 \le \omega \le \pi \qquad (4.42)$$

A função *nint* devolve o inteiro mais próximo do seu argumento.

4.6 Limiar de Mascaramento Estereofónico

4.6.1 Objectivos

Em geral, a experiência comum diz-nos que os sinais provenientes dos dois canais do par estereofónico soam de forma idêntica. Assumindo, como se fará até ao final deste capítulo, que o ouvinte se encontra simetricamente colocado relativamente aos focos sonoros, a afirmação anterior também significa que na maior parte do tempo de reprodução, a imagem acústica (localização subjectiva da fonte sonora, resultante da fusão binauricular) posiciona-se no ponto médio entre os focos sonoros. Assim, é evidente a existência de correlação intercanal que pode ser convertida em ganho de codificação.

O primeiro objectivo do limiar de mascaramento estereofónico é criar uma base de representação em que os dois canais se revelem mais descorrelacionados, permitindo a extracção de redundância intercanal.

O segundo objectivo relaciona-se com dificuldades de codificação na circunstância em que os sinais dos dois canais formam uma imagem acústica - estática ou dinâmica - coerente. Como já foi apontado na introdução do presente parágrafo, se os dois canais forem codificados independentemente, o ruído de quantificação surgirá descorrelacionado, não acompanhando a imagem do sinal, o que resultará, devido ao desmascaramento binaural, em artefactos graves aquando da reprodução do sinal estereofónico codificado. A solução consiste em correlacionar o ruído de quantificação nos dois canais, por forma a aplicar convenientemente a protecção de MLD.

O terceiro e último objectivo do limiar de mascaramento estereofónico visa extraír irrelevância intercanal devido a uma imagem acústica predominante. Isto é, se o sinal se tornar particularmente imponente numa direcção, haverá outras

direcções que deixarão de conter informação relevante pois serão mascaradas por aquela.

4.6.2 Vectores Soma e Diferença

Comparando a representação temporal dos sinais dos dois canais, verifica-se não haver correlação que mereça ser explorada [A2]. Porém, a representação espectral já exibe um número de peculiaridades que perpectiva vantagens na análise conjunta dos sinais [A3][A12]. De facto, uma possibilidade prática é converter a base original composta por canal *direito* (R) e canal *esquerdo* (L) numa outra base ortogonal, formada pelos vectores *soma* (S) e *diferença* (D). A conversão é efectuada através da combinação linear (4.43).

$$\begin{bmatrix} S \\ D \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} R \\ L \end{bmatrix} \qquad (4.43)$$

A operação de conversão é válida tanto no domínio dos tempos como no domínio das frequências, já que a operação de transformada é também uma operação linear. Contudo, é o domínio das frequências que será implicitamente assumido na sequência deste capítulo.

A similaridade sonora entre os canais direito e esquerdo denota, por si só, a grande correlação espectral entre estes. Os vectores R, L, S e D têm a dimensão da janela de amostragem temporal usada em cada instante (capítlo 4.2), que é suficiente, em qualquer caso, para traduzir a correlação espectral apontada numa maior concentração de energia no vector S ou no vector D. Esta descorrelação proporciona uma quantificação global mais efectiva e, mesmo que sejam só aplicados modelos monauriculares, resulta sempre um ganho de codificação superior ao obtido com codificação independente de canais [A12].

Em cada segmento, a base SD, em vez de fixa, poderia sofrer uma rotação de modo a maximizar a concentração de energia num dos vectores S ou D e a potenciar, desta forma, o ganho de codificação [V6]. Porém, a informação lateral necessária anula e por vezes supera a extracção de redundância conseguida com esta estratégia [A23].

Em média, a conversão fixa (4.43) revela ser eficaz na remoção de redundância e será, por isso, a adoptada no codificador perceptual estereofónico.

4.6.3 Protecção MLD

A conversão inversa de (4.43) é (4.44).

$$\begin{bmatrix} R \\ L \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} S \\ D \end{bmatrix} \qquad (4.44)$$

Uma análise simples prova que, quando os vectores S e D são quantificados, existe correlação entre o ruído de quantificação dos dois canais R e L reconstruídos. Assim, é possível planear a aplicação da protecção MLD. Dado que o desmascaramento binaural é mais efectivo quando a localização do ruído não acompanha a localização do sinal, e atendendo à disposição relativa entre os dois focos do par estereofónico e o ouvinte, a protecção de MLD é projectada para a circunstância mais adversa. Assim, supõe-se que a localização do sinal é perfeitamente *central* (coerência entre os dois canais) e que a localização do ruído de quantificação é totalmente *lateral* ou, equivalentemente, supõe-se que a localização do sinal é totalmente *lateral* (consonância de IID com ITD) e que a localização do ruído é totalmente *central.*

Com base na informação de Moore [P2] e outros autores [P20] e também nos nossos próprios testes e resultados, afigurou-se adequado prever uma protecção de MLD, particularmente necessária a baixas frequências, até cerca de 3KHz. Contrariamente à própria definição de MLD (protecção *relativa* ao limiar de mascaramento monauricular), e depois de testar alguns cenários com sinais coerentes e incoerentes, concluímos oportuno definir uma protecção de MLD genérica e *absoluta*, dependente só da energia de excitação e não dependente das propriedades tonais do sinal. A expressão (4.45) provou fornecer os melhores resultados de codificação.

$$MLD_{dB}(p) = 25.5\left[\cos\frac{\pi B(p)}{32.0}\right]^2 \quad ; \quad 0 \le B(p) \le 16.0 \quad ; \quad 0 \le p \le npart\text{-}1 \qquad (4.45)$$

Esta expressão é válida para as partições com frequência central até 16.0Bark, isto é, até 3KHz. Para frequências superiores não é necessária protecção de MLD. Ao modelo (4.45) corresponde a energia de quantificação associada à protecção MLD (4.46).

$$e_{qMLD}(p) = e(p)\ 10^{-\frac{MLD_{dB}(p)}{10}} \quad ; \quad 0 \le p \le npart\text{-}1 \tag{4.46}$$

No contexto da reprodução estereofónica, a energia (4.46) representa a quantidade máxima de ruído de quantificação em cada partição e em cada um dos vectores, S ou D, que evita o desmascaramento binauricular, para os cenários extremos de localização relativa entre sinal e ruído de quantificação.

4.6.4 Irrelevância Estereofónica

Uma imagem acústica resulta da conjugação de componentes espectrais dos dois canais, R e L, produzindo uma impressão espacial com várias direcções associadas a fontes sonoras virtuais. Esta impressão espacial é a projecção perceptual das IID e ITD físicas. Quando uma imagem acústica contém uma fonte sonora predominante, numa dada direcção, verifica-se que outras direcções da imagem perdem importância perceptual podendo ser completamente mascaradas pela dominante. Desta forma, uma imagem acústica resultante da reprodução estereofónica encerra alguma irrelevância intercanal que, se for correctamente identificada, poderá ser vantajosamente extraída.

A resolução temporal binauricular é avaliada em cerca de 6μs [P10] sendo, portanto, muito superior à resolução temporal monauricular, avaliada a partir dos nossos resultados em cerca de 1ms (parágrafo 4.4.3). Não é trivial actuar sobre a componente ITD de modo a intervir na imagem acústica, tanto mais que a percepção binauricular no domínio dos tempos (pré-mascaramento e pós--mascaramento binauricular) não teve ainda resultados de investigação tão conclusivos como a monauricular. Porém, é muito fácil actuar sobre a componente IID. Por coincidência, nos formatos vulgares de registo sonoro a IID é também a componente predominante na base das imagens acústicas. Neste contexto, tentar-se-á remover a irrelevância intercanal actuando exclusivamente sobre a IID.

A base de identificação das várias direcções é formada pelos vectores S e D. De facto, estes representam as componentes do sinal projectadas no eixo perpendicular ao eixo que une os focos sonoros e as componentes do sinal projectadas no eixo que une os focos sonoros, respectivamente. Retomando a terminologia anterior, o primeiro eixo associa-se à localização central e o segundo à localização lateral.

Para traduzir a extracção de irrelevância intercanal em grandezas tratáveis como os limiares de mascaramento, postulamos o seguinte princípio:

- Se se verificar um súbito aumento do sinal (e reflexamente do ruído de quantificação) no eixo lateral, é perceptualmente permitido um ligeiro incremento do ruído de quantificação no eixo central. O limite superior é o próprio ruído lateral.

Simétrica e equivalentemente:

- Se se verificar um súbito aumento do sinal (e reflexamente do ruído de quantificação) no eixo central, é perceptualmente permitido um ligeiro incremento do ruído de quantificação (correlacionado) no eixo lateral. O limite superior é o próprio ruído central.

Qualquer incremento deverá ser corrigido pelo limiar associado à protecção MLD do eixo oposto. Desta forma, a extracção de irrelevância é decidida a partir de um estado em que o desmascaramento binauricular está prevenido.

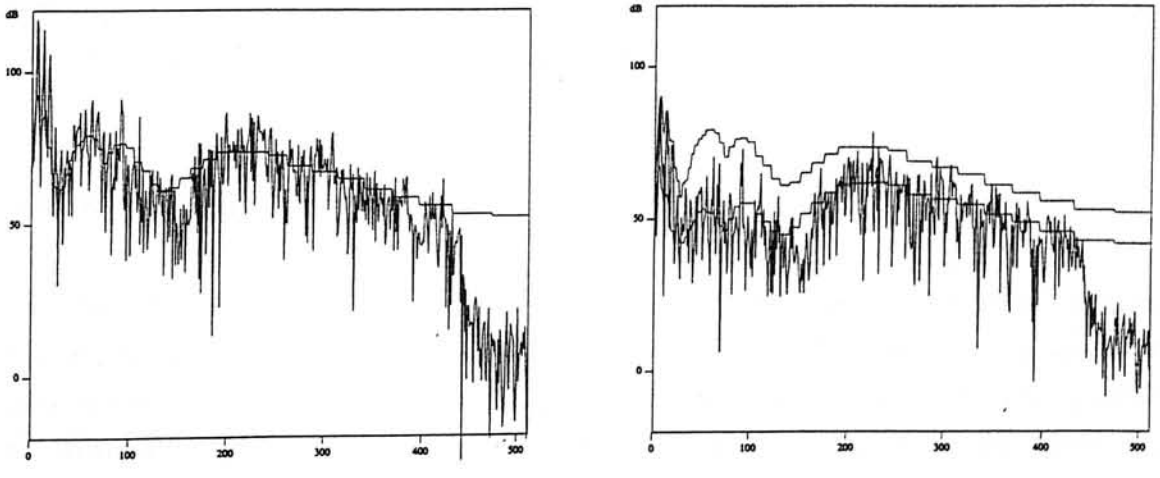

Figura 4.11: Energia de quantificação associada ao limiar de mascaramento estereofónico correspondente ao mesmo segmento da Fig. 4.10. O eixo das abcissas representa a frequência em linhas espectrais. O gráfico do lado esquerdo representa o espectro e a energia de quantificação do vector S e o do lado direito representa o espectro e a energia de quantificação do vector D. Dado que o sinal é composto fundamentalmente por componentes centrais, só há lugar para o mascaramento de componentes laterais devido a estas. Assim, a curva superior (em escada) da energia de quantificação associada ao vector D traduz a irrelevância intercanal, ao identificar apenas 4 componentes espectrais laterais com importância para a imagem perceptual do sinal.

A Fig. 4.10 representa a energia de quantificação monofónica avaliada num segmento longo do trecho "Suzanne Vega". Considerando o mesmo segmento, o espectro correspondente aos vectores S e D está reproduzido na Fig. 4.11.

A partir da Fig 4.10 é fácil concluir que se trata essencialmente de um sinal monofónico, ou seja, de um sinal com localização predominantemente central. Em consequência, a maior fracção da energia do sinal projecta-se no vector S. Concretizando o postulado anterior para extraír a irrelevância intercanal, pode-se apreciar a sua influência na energia de quantificação de cada vector, após esta haver sido calculada através do mesmo modelo implícito na Fig. 4.10. Para o vector S, verifica-se que não há qualquer liberdade oferecida por componentes mascarantes com localização lateral. Por este motivo, as curvas da energia de quantificação, obtidas antes e após a extracção de irrelevância, coincidem. Por outro lado, para o canal D, há duas curvas distintas. A inferior representa a energia de quantificação baseada no modelo intracanal. A curva superior traduz o mascaramento de componentes laterais devido a componentes centrais. A diferença entre as duas curvas representa a irrelevância intercanal. Concretamente, apenas 4 componentes espectrais no eixo lateral são consideradas relevantes para a imagem acústica (perceptual) do sinal. Assim, para o segmento considerado, se a codificação do sinal assentar na quantificação dos vectores S e D, explorando a irrelevância intercanal, resultará uma compressão muito mais efectiva do que a conseguida com quantificação dos vectores R e L.

4.6.5 Factores de Escala

Os factores de escala que representam o limiar de mascaramento estereofónico são calculados usando os mesmos passos descritos no parágrafo 4.5 para calcular o limiar monofónico. De facto, para cada vector, S ou D, são aplicáveis as expressões (4.30) até (4.42). De modo a incluir a extracção de irrelevância intercanal, antes de se distribuír a energia de quantificação no espectro original (4.37), é incluída a expressão (4.47) que também faz uso da expressão (4.46) cuja necessidade ficou exposta nos dois parágrafos anteriores.

$$e'_{q(S)}(p) = max\left[e_{q(S)}(p) \, , \, min\left[e_{qMLD(D)}(p) \, , \, e_{q(D)}(p) \right] \right] \; ; \; 0 \le p \le npart\text{-}1 \qquad (4.47a)$$

$$e'_{q(D)}(p) = max\left[e_{q(D)}(p) \, , \, min\left[e_{qMLD(S)}(p) \, , \, e_{q(S)}(p) \right] \right] \; ; \; 0 \le p \le npart\text{-}1 \qquad (4.47b)$$

As duas curvas de energia de quantificação referidas na Fig. 4.11, correspondem a $e_{q()}(p)$ e a $e'_{q()}(p)$.

4.7 Modos de Codificação Espectral

Após decidir o tipo de janela de amostragem temporal adequada para ambos os canais do par estereofónico (parágrafo 4.4), é necessário definir os vectores convenientes que deverão ser sujeitos a quantificação. As duas decisões não dependem entre si, o que permite simplificar a estrutura global do codificador e minimizar o atraso total de codificação. As hipóteses possíveis resumem-se à escolha dos vectores R e L (modo R/L) ou, alternativamente, dos vectores S e D (modo S/D).

O critério de decisão está ligado à necessidade e oportunidade do limiar de mascaramento estereofónico (parágrafo 4.6.1). Se o sinal estereofónico é tal que não gera uma imagem acústica definida, então a base S/D não representará, com grande probabilidade, o sinal mais descorrelacionado; não se justifica uma protecção para o desmascaramento binauricular (MLD) e não faz sentido a extracção de irrelevância intercanal. O modo de codificação mais conveniente será o modo R/L. Pelo contrário, se o sinal estereofónico é tal que gera uma imagem acústica definida, então estas três últimas razões virão negadas, o que torna o modo de codificação S/D oportuno e vantajoso.

A escolha do modo de codificação é feita com a precisão permitida pelos factores de escala. Isto é, o modo de codificação é determinado individualmente para cada banda de um mesmo segmento de sinal. Admite-se assim uma eventual consistência entre os dois canais do par estereofónico, para formar uma imagem acústica apenas a partir de fracções do espectro (determinadas pelas bandas). Uma solução simples, não unívoca - porque pode conduzir ao modo S/D mesmo quando não seja necessário - e funcional para determinar o modo de codificação para cada banda, é comparar logaritmicamente os factores de escala associados ao vector R e ao vector L. Se a diferença não ultrapassar um limite estabelecido (de acordo com os resultados das nossas simulações, igual a 2dB), adoptar-se-á o modo de codificação S/D. Na outra hipótese, resta adoptar o modo R/L (Fig. 4.12). A decisão do modo de codificação, além de ser adaptativa nas frequências, é também adaptativa nos

tempos pois a mesma banda, em segmentos subsequentes, pode ser codificada diferentemente.

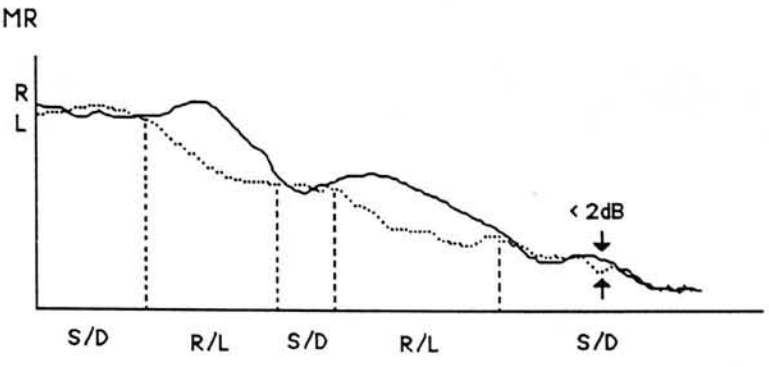

Figura 4.12: A escolha do modo de codificação é determinada em cada banda, a partir da diferença entre as relações sinal-ruído mascarado afectas aos canais direito e esquerdo. O gráfico representado é meramente ilustrativo, não traduzindo uma situação real nem representando o pormenor da divisão do espectro em bandas.

A Fig. 4.13 ilustra um sinal em que o canal direito é extraído do trecho "Suzanne Vega" e o canal esquerdo é extraído do trecho "Tracy Chapman". São trechos de vozes femininas distintos. Para cada caso, a curva (em degrau) intermédia representa representa a respectiva energia máxima de quantificação.

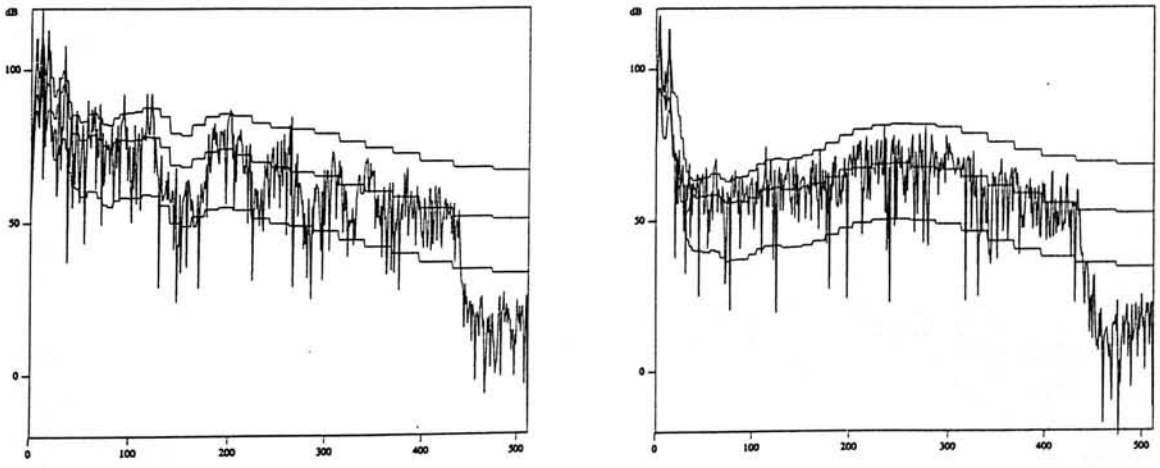

Figura 4.13: Espectro de um sinal composto, no canal direito, por um segmento longo de "Suzanne Vega" (gráfico na esquerda) e no canal esquerdo (gráfico na direita), por um segmento longo de "Tracy Chapman". As curvas em escada têm o mesmo significado referido para o sinal da Fig. 3.10. O eixo das abcissas representa a frequência em linhas espectrais.

A energia máxima de quantificação para os vectores S e D, considerando também irrelevância intercanal, está representada na Fig. 4.14.

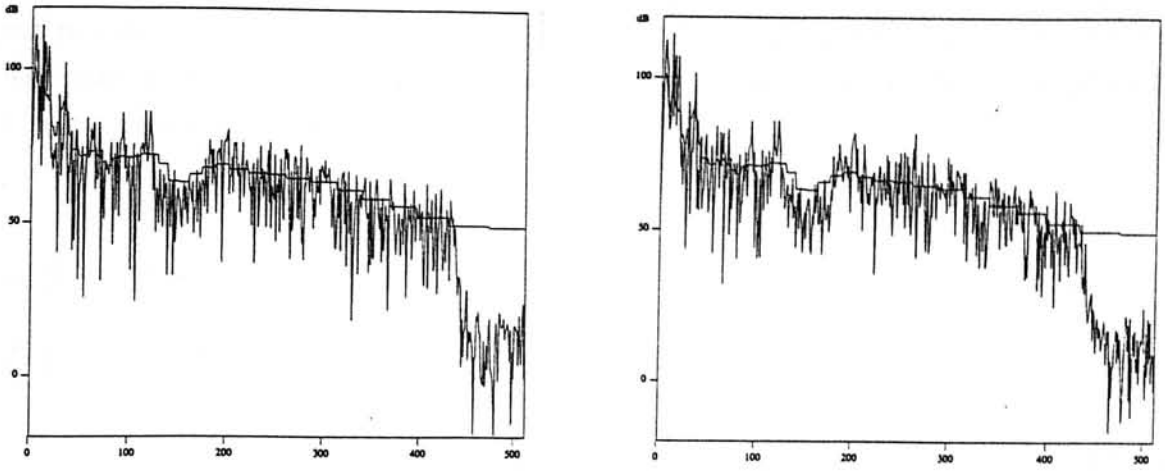

Figura 4.14: Vectores S (à esquerda) e D (à direita) corespondentes ao sinal da Fig 4.13. A curva em escada representada traduz em cada partição, a energia máxima de quantificação, incluindo já a irrelevância intercanal. O eixo das abcissas representa a frequência em linhas espectrais.

Pode-se concluir que este não é rigorosamente um sinal que origine uma imagem acústica. Contudo, o algoritmo identifica 15 bandas, de um total de 35, passíveis de serem codificadas no modo S/D. Concretizando para o segmento considerado e referindo apenas a alternância S e R (homóloga de L e D), podem-se identificar, na Fig. 4.15, as bandas que serão expressas pelos coeficientes e factores de escala associados a S e as restantes bandas que se verão representadas pelos coeficientes e factores de escala associados a R.

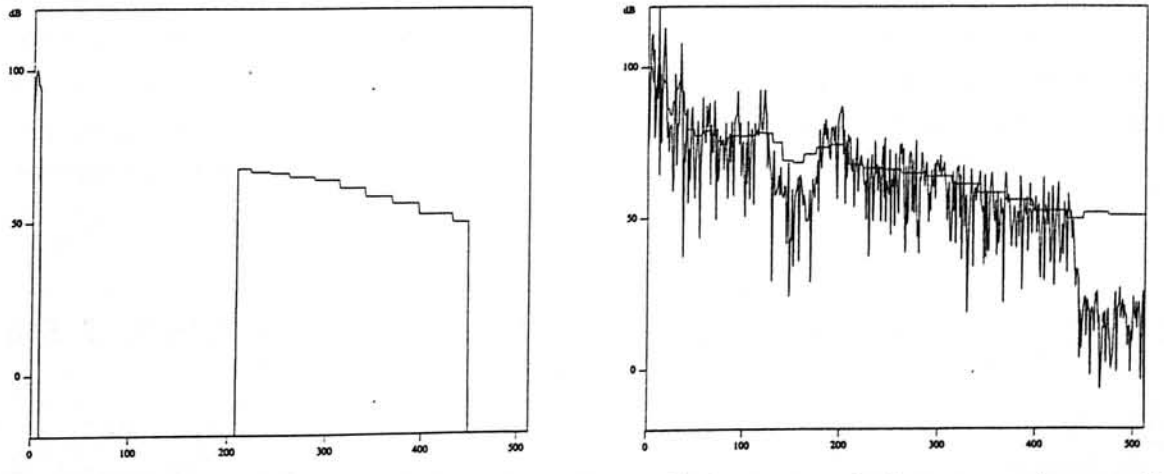

Figura 4.15: Espectro e energia de quantificação final para o sinal representado nas Fig. 4.13 e Fig. 4.14 e incluindo só a alternância de bandas em modo de codificação S e R. No gráfico da direita, as bandas codificadas com coeficientes e energia de quantificação correspondente ao vector S, estão identificadas no gráfico da esquerda, pela curva assinalada. As restantes bandas serão codificadas no modo R. O eixo das abcissas representa a frequência em linhas espectrais.

O sinal da Fig. 4.13 foi criado para evidenciar a alternância dos modos de codificação. No entanto, a partir de testes auditivos informais, concluíu-se também que, mesmo para um sinal bilingue como este, a comutação de modos de codificação é uma estratégia "benigna".

4.8 Códigos de HUFFMAN

4.8.1 Objectivos

Os códigos de Huffman ou códigos de comprimento variável [V6], representam um processo de codificação sem perdas, também conhecido por codificação por entropia, cujo ganho advém da atribuição de palavras de código com menor comprimento a entradas (ou ocorrências) com maior probabilidade. Para além disto, os códigos de Huffman são também autossincronizáveis, evitando a necessidade de separadores entre palavras de código.

O emprego de códigos de Huffman no algoritmo de codificação perceptual estereofónico visa melhorar o ganho global de compressão. Os coeficientes espectrais quantificados, os factores de escala e os indicadores binários do Modo de Codificação, são os elementos submetidos a codificação sem perdas. A construção adequada de tabelas de Huffman para cada caso, contribui decisivamente para a codificação a muito baixas taxas de informação. Para as situações referidas a seguir, os códigos de Huffman foram obtidos a partir de estísticas de ocorrências, iterativamente actualizadas com todos os 22 trechos musicais de uma biblioteca, até se observar uma relativa estabilização das probabilidades. O procedimento usado está exposto num documento interno da AT&T (Technical Memorandum [11224--880509-04TM]).

4.8.2 Codificação de Coeficientes Espectrais

Dado que os coeficientes espectrais quantificados representam a maior parte da informação a codificar, a estratégia de codificação sem perdas é estabelecida a partir destes. Em traços gerais, são definidas *secções* do espectro de acordo com a distribuição de amplitudes dos coeficientes. Cada secção de coeficientes é codificada com a mesma tabela de Huffman e envolve um número inteiro de bandas (e

reflexamente, o mesmo número inteiro de factores de escala). Em função de algum conhecimento prévio, foi definido um conjunto de 8 tabelas de Huffman só para coeficientes, com as características apontadas na Tabela 4.1.

Índice da Tabela de Huffman	MAV (n)	(2n+1)	Número de entradas	Tamanho da Tabela
0	0	1	-	-
1	1	3	4	81
2	2	5	4	625
3	3	7	4	2401
4	5	11	2	121
5	9	19	2	361
6	16	33	2	1089
7	SAÍDA(15)	33	2	1089

Tabela 4.1: Características das diferentes tabelas de Huffman para coeficientes espectrais.

A diferença entre as várias tabelas reside no Valor Máximo Absoluto (MAV) de cada entrada e no número de entradas que dá origem a um único símbolo de Huffman (ou código de Huffman). Ilustrando com o exemplo da tabela número 2, cada entrada (coeficiente) tem um valor compreendido entre -2 e +2. Cada código de Huffman comprime 4 entradas. Assim, a tabela conterá $(2MAV+1)^4 = 625$ códigos.

A tabela número 7 é uma tabela especial para secções que possuam coeficientes com MAV superior a 16. Os códigos de fronteira desta tabela representam saídas para um código adicional distinto (código de saída). O código de saída representa um valor positivo que se adiciona ao código de Huffman (apontador), com o mesmo sinal lógico deste. Dado que a tabela número 7 é bidimensional, poderá haver um ou dois códigos de saída para além de um apontador de Huffman. De modo a evitar uma tabela adicional para códigos de saída, que seria enorme dado o majorante teórico de um coeficiente quantificado, faz-se uso de uma regra simples e perfeitamente genérica [A15], que conduz à construção de códigos, também autossincronizáveis, mas não óptimos.

O código de saída constroi-se adicionando um cabeçalho à representação binária do número traduzido pelo código. Assim, cada código de saída terá um número de *bits* B dado por: $B = \min [4, \text{int}[\log_2 n + 1]]$; $0 \leq n < \infty$. A forma do cabeçalho está ilustrada pelos exemplos da Tabela 4.2.

Cabeçalho	B	Gama de Valores
0	5	0-15
10	6	16-31
110	7	32-63
1110	8	64-127
11110	9	128-255

Tabela 4.2: Alguns exemplos de códigos de saída.

A divisão do espectro em secções é realizada usando uma estratégia de custo mínimo. Inicialmente são definidas todas as secções possíveis - o limite máximo é uma secção por banda - em que cada secção tem a tabela de Huffman que melhor se adapta aos seus coeficientes. Como o início e o fim do espectro útil é conhecido, se K é o número inicial de secções, haverá K-1 separadores (e implicitamente K índices de tabelas de Huffman) que deverão ser transmitidos ao descodificador como informação lateral. O *preço* (em *bits*) para eliminar um separador é calculado. Este *preço* baseia-se numa estimativa do *preço* médio de cada tabela e traduz a canibalização de uma ou mais secções por outra que possui um MAV superior. O separador que exibe um *preço* menor é eliminado (os *preços* iniciais podem ser negativos). Os *preços* são recalculados antes de uma nova iteração. O algoritmo é repetido até que um número máximo de secções seja atingido e que o *preço* mais pequeno para eliminar um outro separador seja maior que um valor definido.

4.8.3 Factores de Escala e Indicadores de Modo de Codificação

Os factores de escala e os indicadores do Modo de Codificação têm tabelas de Huffman personalizadas.

Os valores fornecidos por (4.42) são números inteiros positivos e exibem uma grande correlação em bandas activas (*i.e.*, em que pelo menos um coeficiente é não-nulo) adjacentes. Por isso, antes de criar uma tabela de ocorrências, é aplicado DPCM entre factores de escala pertencentes a bandas activas consecutivas. Assim, a codificação do primeiro factor de escala é realizada em PCM. Os seguintes são codificados a partir de uma tabela bidimensional, contendo os valores em DPCM.

Um indicador do Modo de Codificação é uma variável binária que indica se uma dada banda do par esterofónico é codificada em modo R/L ou em modo S/D. Um segmento longo contém 35 bandas e um conjunto de 4 segmentos curtos contém

14*4=56 bandas. A tabela de Huffman para codificar o Modo de Codificação considera códigos, em que cada um condensa 7 indicadores. Assim, o Modo de Codificação para um segmento longo exprime-se em 5 códigos de Huffman e, para um conjunto de 4 segmentos curtos, exprime-se em 8 códigos.

4.9 Ajuste da Taxa de Informação

Tomado na sua estrutura básica, o codificador perceptual estereofónico é por natureza, um gerador de informação a taxa variável. Revela, portanto, uma vocação especial para ser incorporado num ambiente cujo modo de transferência seja assíncrono (ATM), como poderá acontecer nas futuras Redes de Digitais (de Banda Larga) com Integração de Serviços (RDIS), em fase de normalização no âmbito do CCITT. Neste caso, diz-se também que a codificação dos sinais áudio é realizada com qualidade constante porque se pode definir uma margem de segurança (sobrecodificação) permanente, relativamente ao limiar de mascaramento postulado em cada segmento de sinal codificado. Contudo, o ambiente de transferência encontrado nas actuais redes de comutação de circuitos é síncrono (STM), impondo, portanto, uma taxa de informação constante. Uma forma do codificador respeitar uma taxa constante de informação é variando a margem de segurança da codificação relativamente ao limiar de mascaramento. Assim, uns segmentos do sinal serão mais sobrecodificados do que outros. Se a taxa de informação for suficientemente pequena, alguns segmentos poderão inclusivamente ser subcodificados.

O codificador perceptual estereofónico dispõe de um bloco de controlo da taxa de informação (Fig. 4.1) para gerir um défice ou crédito de *bits* por segmentos consecutivos, de modo a manter a taxa média de informação constante.

Depois do limiar de mascaramento final haver sido calculado para cada segmento, o número total de *bits* para o codificar, com base neste limiar, é contabilizado. Este valor é comunicado ao bloco de gestão referido que, em função do défice ou crédito acumulado, determina um número máximo de *bits* que não deverá ser ultrapassado na codificação de toda a informação principal e lateral desse segmento. Seguidamente, é activado um processo iterativo de quantificação que sobrecodificará ou subcodificará o segmento, de modo que o número de *bits* realmente usado na sua codificação fique muito próximo mas aquém do número máximo de *bits* que lhe foi atribuído. O processo iterativo não é totalmente pacífico

pois dada a inclusão de códigos de Huffman, a função de aproximação é bastante hostil. Por esta razão, o algoritmo de quantificação iterativa identifica em primeiro lugar, uma solução satisfatória que será adoptada, caso não haja convergência.

O procedimento descrito é válido quer para segmentos longos, quer para grupos de 4 segmentos curtos. Neste caso, os coeficientes e os factores de escala são agrupados por concatenação ordenada de bandas homólogas.

4.10 Estrutura de uma Trama de Informação

A informação de um segmento codificado é combinada (Fig. 4.1) para formar uma trama com vários campos. Cada trama reune toda a informação necessária para reconstruir um segmento longo ou um conjunto de 4 segmentos curtos. Como já foi referido anteriormente, neste último caso, os indicadores do Modo de Codificação, os factores de escala e os coeficientes são agrupados por concatenação ordenada de bandas homólogas. A ordenação de vários campos da trama é realizada de acordo com a estrutura representada na Fig 4.13.

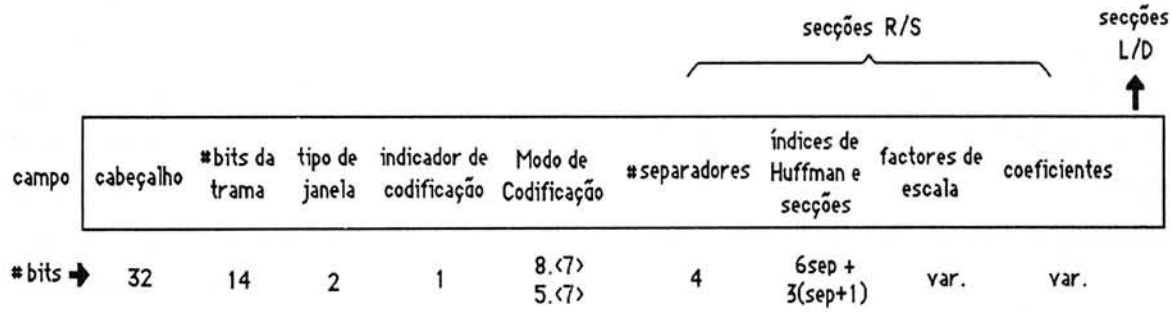

Figura 4.13: Formato da trama representativa de um segmento longo ou de um conjunto de 4 segmentos curtos. Está também indicado o número de *bits* associado a cada campo. <7> representa códigos de Huffman que condensam 7 *bits*. Os campos assinalados com comprimento variável compõem-se exclusivamente de códigos de Huffman.

Há campos que possuem um número fixo de *bits* e outros que possuem um número variável, devido sobretudo à utilização de códigos de Huffman. A Fig. 4.13 revela que os 5 primeiros campos da trama são comuns aos dois vectores do sinal esterofónico.

O cabeçalho tem funções de sincronização, além de também poder transportar informação diversa. O terceiro campo indica o tipo de janela (Fig. 4.4) de

amostragem temporal usada na codificação do segmento em questão. O indicador de codificação assinala se os dois vectores codificado são bilingues. Neste caso, o Modo de Codificação é R/L para todas as bandas, o que torna desnecessário o campo seguinte. Quando existe extracção de redundância e irrelevância intercanal, o campo de indicação do Modo de Codificação condensa o modo de codificação de todas as bandas. Seguidamente, surgem duas unidades de campos, idênticas para cada vector codificado e com a caracterização das secções de cada vector. Dado que o número máximo permitido de secções para cada vector é 16, usam-se 4 *bits* para a sua indicação. A localização de cada separador é definida por 6 *bits* e o índice da Tabela de Huffman usada para cada secção é expresso em 3 *bits*. Seguidamente, são colocados ordenadamente todos os códigos de Huffman respeitantes aos factores de escala e aos coeficientes espectrais.

O formato de trama descrito foi implementado no codificador perceptual estereofónico (ao nível de simulação) e revelou uma grande flexibilidade em se adaptar a diferentes condições como por exemplo, a codificação de um único canal.

4.11 Conclusões

Ficaram expostas, neste capítulo, as contribuições mais relevantes para a definição e simulação de um algoritmo de codificação perceptual estereofónico. Este algoritmo é capaz de extraír redundância e irrelevância intracanal e intercanal, de modo a codificar os sinais de um par estereofónico com uma relação taxa de informação/qualidade ainda não atingida por qualquer outro codificador, incluindo a norma em preparação no âmbito do MPEG.

No parágrafo 3 demonstrou-se a oportunidade da transformada MDCT e desenvolveu-se um algoritmo de cálculo rápido e eficiente, fornecendo simultaneamente informação psico-acúsica. Esta contribuição teve uma aplicação lateral que não cabe nesta tese porque está enquadrada no algoritmo de codificação MUSICAM, implementado pelos Laboratórios Bell.

O parágrafo 4 valorizou a estratégia de comutação de janelas de amostragem temporal para ultrapassar problemas de (des)mascaramento temporal, ao fazer uso da medida de Entropia Perceptual para detectar a necessidade da comutação e decidir, com rigor, a melhor solução perceptual de comutação.

O parágrafo 5 retomou uma estratégia de cálculo do limiar de mascaramento monofónico para melhorar o modelo psico-acústico subjacente.

Os parágrafos anteriores traduzem a base dos programas de simulação que foi necessário escrever para formar um sistema organizado e consistente, de modo a permitir e facilitar o desenvolvimento dos parágrafos seguintes.

O limiar de mascaramento estereofónico é um conceito inédito, apresentado nos parágrafos 6 e 7, e que resultou da caracterização do efeito de MLD e da postulação da irrelevância estereofónica. Os ganhos adicionais conseguidos, relativamente aos obtidos com o limiar de mascaramento monofónico, são bastante encorajadores como ficará ilustrado pelas estatísticas do capítulo 5.

O parágrafo 8 expôs a estratégia usada para aumentar o ganho global de codificação, através da codificação por Entropia.

O parágrafo seguinte caracterizou o procedimento desenvolvido para adaptar o codificador perceptual a ambientes de transferência síncrona (STM).

Finalmente, ilustrou-se no parágrafo 9 a estrutura física da trama implementada para codificar os vectores representativos do par estereofónico.

5
TESTES DE AUDIÇÃO

5.1 Objectivos

Não existe uma medida objectiva que traduza inequivocamente a qualidade perceptual de um sinal áudio submetido a um processo de codificação e descodificação. A melhor forma de avaliar a qualidade de um sinal áudio, relativamente à sua forma original, é caracterizar estatisticamente a sua qualidade subjectiva apreciada por ouvintes. Neste sentido, a realização de testes de audição é o único procedimento adequado para fornecer uma avaliação qualitativa do desempenho de um algoritmo codificador [A9].

Por convite, 18 ouvintes com algum treino auditivo ou simplesmente audiófilos, todos pertencentes aos Laboratórios Bell, acederam a participar nos testes de audição. Estes testes foram planeados para testar dois algoritmos codificadores funcionando a diferentes taxas. A escala de avaliação adoptada é a descrita na recomendação 562 do CCIR. Na Fig 6.1 pode-se apreciar a gama de valores de classificação e o respectivo significado.

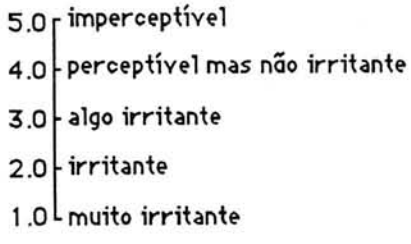

Figura 6.1: Escala de avaliação com 5 níveis de acordo com a Rec. 562 do CCIR. Mostra-se a correspondência com a avaliação subjectiva da existência de artefactos ou distorção entre as versões codificada e original do mesmo sinal.

5.2 Planeamento dos Testes

Os testes de audição decorreram numa câmara isolada em que dois ouvintes eram simultaneamente submetidos a um teste completo. Os ouvintes usaram auscultadores electrostáticos e podiam regular livremente o nível do sinal áudio.

O teste completo incluíu seis sessões em que cada uma envolvia um trecho musical particular. Na sua maior parte, os trechos foram extraídas de um Disco

Compacto emitido pela EBU (European Broadcasting Union) e destinado à realização de testes subjectivos [A4]. Os diferentes trechos, com duração entre 10 e 20 segundos cada um, são representativos de sons harmónicos (tonais), sons ricos em "ataques", sons contendo imagem esterofónica, ou simplesmente voz. O parágrafo seguinte indica a natureza de cada um. Cada sessão incluíu para além da versão original do sinal, três versões codificadas: em modo monofónico (codificação independente de cada canal) e à taxa total de 128Kbit/s, em modo estereofónico à taxa total de 96Kbit/s e neste mesmo modo à taxa total de 128Kbit/s.

No início de cada teste foram fornecidas indicações acerca da natureza, apresentação e modo de avaliação de cada sessão de audição. Foram também fornecidas fichas de avaliação, uma para cada sessão. Estes elementos encontram-se em apêndice (inclui-se só a ficha de avaliação respeitante ao último trecho musical).

Cada sessão incluíu 5 *conjuntos* de três sequências. Cada sequência era composta pela sucessão **A B C**. **A** era sempre o trecho musical original e era tomado como referência para a graduação máxima, isto é, para 5.0. **B** era a versão codificada ou a original. **C** era a negação de **B**. Esta disposição foi definida aleatoriamente por computador e só foi dada a conhecer, mesmo para os organizadores dos testes, eles próprios ouvintes, no final de todos os testes. Para cada *conjunto* foi pedido que os ouvintes atribuissem uma classificação para **B** e outra para **C**, tendo em consideração a referência **A**. O objectivo desta estratégia (que é idêntica à usada nos testes de Julho de 1990, em Estocolmo, no âmbito do MPEG) foi assim duplo: primeiro, levar o ouvinte a identificar qual de **B** ou **C** era o trecho original, o que deveria fazer indicando 5.0 numa destas duas alternativas. Em segundo lugar, pretendia-se levar o ouvinte a classificar a degradação do sinal codificado.

Cada *conjunto* compreendeu uma sequência repetida três vezes para permitir ao ouvinte alguma aprendizagem e alguma certeza na sua própria decisão. Sem o ouvinte saber, o primeiro *conjunto* de cada sessão dizia sempre respeito à versão codificada estereofonicamente a 96Kbit/s. Pretendeu-se, assim, criar um momento inicial de aprendizagem para a parte seguinte e mais importante do teste (em termos de estatísticas). Dos 4 restantes *conjuntos* de cada sessão, 2 representavam a codificação monofónica e 2 representavam a codificação estereofónica. Para cada sessão, esta ordenação foi também definida aleatoriamente.

Considerando os intervalos entre sessões, um teste completo envolveu cerca de uma hora.

5.3 Resultados e Conclusões

Considerando que a resposta fornecida por um ouvinte é uma variável aleatória com uma função densidade de probabilidade não conhecida, recorreu-se à desigualdade de Tchebycheff para obter uma estimativa do número mínimo de ouvintes necessários para a realização do teste. Esta estimativa pretendeu ser só um indicador inicial, não rigoroso, para a obtenção de um número suficiente de dados que permitisse, nas condições em que o teste global foi planeado, uma caracterização estatística com alguma credibilidade.

Assumiu-se que para cada condição de teste, a resposta x_i de um ouvinte caracteriza-se por uma média $E(x_i)=\mu$ e uma variância $\sigma_{xi}^2=1.0$. O valor da variância considera a escala de avaliação apresentada no parágrafo anterior. Se as N respostas dos ouvintes se assumirem como variáveis aleatórias do mesmo tipo, resultará para opinião média:

$$\overline{X}=\frac{1}{N}\sum_{i=0}^{N-1}E(x_i)=\mu \tag{5.1}$$

$$\sigma_{\overline{X}}^2=\frac{\sigma_{xi}^2}{N} \tag{5.2}$$

Para que um resultado conjunto possa ser admitido com uma confiança de 95% no intervalo $\varepsilon=\pm 0.25$ em torno de μ será, pela desigualdade de Tchebycheff:

$$P(\mu-\varepsilon<x<\mu+\varepsilon)\leq 1-\frac{\sigma_{xi}^2}{N\varepsilon^2} \tag{5.3}$$

Concretizando os valores adequados, resultará N=320 respostas por algoritmo de codificação. Dado que há 6 sessões de teste, cada uma envolvendo 2 respostas para o mesmo algoritmo de codificação, deduz-se que deveria haver pelo menos 27 ouvintes a participar nos testes.

O convite foi dirigido a 30 potenciais ouvintes, todos com alguma "educação" auditiva e ligados aos Laboratórios Bell. Porém, somente 18 participaram efectivamente nos testes. Os resultados que se apresentam a seguir, primeiro individualmente para cada trecho musical e finalmente, para cada algoritmo de codificação, baseiam-se no número de ouvintes indicado e nas expressões (5.4) e (5.5) para o cálculo da média e da variância.

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} x(i) \tag{5.4}$$

$$\sigma^2 = \frac{1}{N} \sum_{i=0}^{N-1} [x(i) - \mu]^2 \tag{5.5}$$

O valor x(i) representa um ponto de avaliação para condições particulares de teste representadas, no total, por N pontos.

Condição de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de Pontos	18	36	36
Identificações Erradas	2/18	1/36	4/36
Média	2.81	2.42	3.44
Variância	1.15	0.99	1.02

Tabela 5.1: Resultados para o trecho vocal "Suzanne Vega". Este trecho consiste simplesmente numa voz feminina.

Condição de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de Pontos	18	36	36
Identificações Erradas	5/18	9/36	15/36
Média	4.00	4.09	4.63
Variância	1.00	0.92	0.38

Tabela 5.2: Resultados para o trecho musical "Tracy Chapman". Este trecho consiste numa voz feminina com acompanhamento musical.

Condição de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de Pontos	18	36	36
Identificações Erradas	12/18	19/36	20/36
Média	4.75	4.62	4.75
Variância	0.44	0.51	0.44

Tabela 5.3: Resultados para o trecho tonal "Triângulo". Este trecho consiste num som idêntico ao produzido por um diapasão.

Condição de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de Pontos	18	36	36
Identificações Erradas	8/18	12/36	15/36
Média	4.56	4.61	4.72
Variância	0.56	0.51	0.36

Tabela 5.4: Resultados para o trecho harmónico "Trompete".

Condição de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de Pontos	18	36	36
Identificações Erradas	7/18	20/36	18/36
Média	4.38	4.63	4.61
Variância	0.80	0.64	0.66

Tabela 5.5: Resultados para o trecho tonal e difuso "Campaínhas". Este trecho consiste em toques diversos de diferentes campaínhas, porporcionando uma imagem acústica.

Condição de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de Pontos	18	36	36
Identificações Erradas	4/18	2/36	12/36
Média	4.02	2.60	4.61
Variância	0.89	1.11	0.66

Tabela 5.6: Resultados para o trecho ruidoso "Castanholas". Este trecho contém bastantes ocorrências de "ataque".

Dos diferentes trechos musicais, "Suzanne Vega" e "Castanholas" foram os casos em que os ouvintes melhor identificaram a versão codificada. De facto, o

primeiro é um sinal de voz extremamente dinâmico, com alterações súbitas de características tonais para ruidosas e vice-versa. O segundo trecho contém ataques bastante "puros" e por isso, difíceis de codificar. Mesmo para a codificação estereofónica a 128Kbit/s, a avaliação média de "Suzanne Vega" não é considerada de alta qualidade porque de acordo com a recomendação 562 de CCIR, fica aquém da pontuação mínima exigida para tal (4.0). Este resultado não é estranho ao facto da maior parte dos ouvintes revelar uma acuidade especial para sinais de voz isolados, dado pertencerem ao Departamento de Processamento de Voz dos Laboratórios Bell. A confirmar esta hipótese, a avaliação média para "Tracy Chapman" já é muito mais interessante (igual a 4.63) e a variância é surpreendentemente baixa (igual a 0.38), apesar de se tratar também de uma voz feminina, mas com acompanhamento musical que provavelmente terá inibido as "regras" perceptuais de avaliação evidenciadas na classificação do trecho anterior.

De todos os trechos musicais, "Castanholas" revelou a maior diferença de pontuação média entre a codificação monofónica e a estereofónica (a 128Kbit/s). Este é provavelmente o melhor exemplo do ganho conseguido com a extracção da irrelevância estereofónica e correspondente controlo do ruído de quantificação.

Considerando novamente a codificação estereofónica a 128Kbit/s, conclui-se que "Triângulo" e "Campaínhas" são os trechos musicais cujas avaliações médias, apesar de interessantes, são as menos credíveis porque estão próximas de terem resultado de escolha aleatória (dado o número de identificações incorrectas da versão codificada). Neste espírito, pode-se considerar que a sua codificação é transparente. Se um intervalo de confiança mais favorável for considerado, então "Trompete" e "Tracy Chapman" também se incluem nos trechos cuja codificação é transparente.

A tabela 5.7 exprime a classificação final por algoritmo de codificação, considerando todos os trechos musicais.

Algoritmo de Codificação	estéreo @ 96Kbit/s	duplo mono @ 128Kbit/s	estéreo @ 128Kbit/s
Número de pontos	108	216	216
Média	4.09	3.83	4.43
Variância	1.06	1.25	0.78

Tabela 5.7: Resultados de avaliação da qualidade de codificação de cada algoritmo, nas condições indicadas.

A média e a variância seguem uma evolução consistente e expectável, desde a codificação monofónica a 128Kbit/s até à estereofónica a 128Kbit/s. De facto, ambas as medidas melhoram. O processo de codificação monofónico à taxa total de 128Kbit/s não é considerado de alta qualidade pois a sua avaliação média é inferior a 4.0. O algoritmo de codificação estereofónico tem para a taxa mais favorável, uma avaliação média interessante que o designa de alta qualidade. Porém, a sua variância é bastante elevada, o que é uma consequência do exíguo número de pontos de amostragem e da influência dos trechos musicais mais críticos: "Suzanne Vega" e "Castanholas". Compreende-se assim que estes passaram a ser os trechos mais usados em testes posteriores ao de Julho de 1990 em Estocolmo, e relacionados com o apuramento da norma para áudio, em preparação no âmbito da ISO/MPEG.

Os diferentes trechos musicais, nas versões original e codificada, encontram-se registados numa Banda de Áudio Digital (DAT) disponível para demonstração.

6
CONCLUSÕES E FUTUROS DESENVOLVIMENTOS

A codificação de sinais áudio, com a perspectiva de reduzir a taxa de informação e sem comprometer a qualidade subjectiva, foi o tema central de investigação considerado nesta dissertação.

No capítulo 2 identificou-se a natureza do problema e concluíu-se que o conhecimento das características analíticas do sistema auditvo permite potenciar o ganho de compressão. Compararam-se algumas soluções e estratégias de codificação perceptual e referiu-se a norma em preparação no âmbito da ISO/MPEG.

O capítulo 3 concentrou-se sobre a informação fornecida pela psico-acústica e analisaram-se, em particular, as propriedades auditivas mais relevantes para o objectivo da compressão. Associou-se a percepção monauricular com a codificação monofónica e evidenciou-se a necessidade de incluir também aspectos da percepção binauricular no contexto da codificação estereofónica.

O projecto de um codificador perceptual estereofónico foi detalhado no capítulo 4. Destacaram-se as constribuições originais relativas à implementação eficiente da estrutura analítica do sinal, ao uso da Entropia Perceptual para alterar adaptativamente a resolução espectral/temporal do processo de codificação, ao melhoramento do modelo de mascaramento monauricular, à definição de um modelo de mascaramento binauricular e, finalmente, à adequação do algoritmo codificador a diferentes ambientes de transferência de informação. A utilização de códigos de Huffman contribuíu para melhorar a taxa de compressão.

O algoritmo codificador/descodificador desenvolvido, para além de ter originado um artigo [A13] a ser publicado na ICASSP92, motivou já um interesse prático concreto. De facto, o algoritmo está a ser objecto de registo de patente (exclusiva da AT&T) e é a base de uma proposta, formulada pelos Laboratórios Bell, para o fornecimento de equipamento de difusão de rádio digital para a Sociedade Norte-Americana de Difusão Radiofónica. Por outro lado, o mecanismo de detecção de não-estacionaridades através da Entropia Perceptual, descrito no parágrafo 4.4, foi adoptado recentemente no algoritmo ASPEC. Este algoritmo faz parte de uma estrutura híbrida de codificação ao nível 3 da norma em discussão no âmbito da ISO/MPEG.

O capítulo 5 descreveu os testes auditivos realizados para avaliar o desempenho do algoritmo codificador. A análise estatística dos resultados permitiu identificar um ganho significativo de codificação associado à extracção de redundância e irrelevância intercanal. A apreciação global do algoritmo codificador para a taxa total de 128Kbit/s, revela que a codificação é transparente para a maioria dos trechos musicais utilizados, excepto para alguns muito críticos em que a degradação é considerada "benigna".

A exiguidade de tempo não permitiu investigar as estratégias mais adequadas para protecção e correcção de erros introduzidos num canal real de transmissão. Esta será uma direcção de futuro trabalho para aumentar a robustez do algoritmo codificador desenvolvido, quer em ambientes de transferência síncrona (STM), quer em ambientes de transferência assíncrona (ATM).

Apesar de muito próximo, o algoritmo de codificação ainda não permite a desejável codificação transparente, em 64Kbit/s, de um trecho musical monofónico arbitrário. Este objectivo implica uma optimização dos parâmetros e modelos que condicionam directamente a eficiência de codificação. Assim, há por exemplo lugar para a investigação de novas estruturas de Bancos de Filtros/Transformadas e janelas de amostragem temporal que se adaptem melhor ao processo analítico do ouvido, em particular, em regiões não-estacionárias do sinal. De facto, concluíu-se que a comutação de janelas associadas a segmentos de diferentes comprimentos era uma solução perceptualmente eficaz. Esta conclusão é corroborada qualitativamente na psico-acústica pela verificação de um efeito de alargamento das bandas críticas [P20] quando o estímulo acústico tem características não--estacionárias. O efeito não foi desenvolvido no texto da tese porque ainda é considerado por alguns autores como uma hipótese. Torna-se portanto necessário investigar e caracterizar melhor a adaptação do modelo analítico do ouvido à dinâmica dos estímulos acústicos.

Continua também a haver lugar para o refinamento do modelo monauricular através da realização de testes psico-acústicos mais completos do que os publicados e considerando eventualmente algumas especificidades do algoritmo de codificação, como seja a natureza das componentes de sobreposição espectral e temporal não canceladas.

Há também necessidade de criar um leque vasto de trechos musicais que explore particularidades da percepção binauricular e que permita completar e aumentar a eficiência do modelo binauricular desenvolvido no capítulo 4.

Por outro lado, existe também a perspectiva de personalizar a análise perceptual do sinal áudio com base em modelos psico-acústicos, de modo a fornecer informação relevante para sistemas de reconhecimento de sons, ou mais especificamente, de sinais de voz.

7
REFERÊNCIAS

CODIFICAÇÃO DE VOZ

[V1] Barry G. Haskell e Raymond Steele; "*Audio and Video Bit-Rate Reduction*"; Proceedings of the IEEE, Vol. 69, N. 2, Feb. 1981.

[V2] James L. Flanagan *et al.*; "*Speech Coding*"; IEEE Trans. Communications, Vol. 27, N. 4, April 1979.

[V3] José M. Tribolet e Ronald E. Crochiere; "*Frequency Domain Coding of Speech*"; IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 27, N. 5, Aug. 1979.

[V4] M. R. Schroeder, B. S. Atal e J. L. Hall; "*Optimizing Digital Speech Coders by exploiting Masking Properties of the Human Ear*"; J. Acoustical Soc. America, Vol. 66, N. 6, Dec. 1979.

[V5] N. S. Jayant, J. D. Johnston e Y. Shoham; "*Coding of Wideband Speech*"; Eurospeech91.

[V6] N. S. Jayant e Peter Noll; "*Digital Coding of Waveforms*"; Prentice-Hall, Englewood Cliffs, 1984.

[V7] Rainer Zelinsky e Peter Noll; "*Adaptive Transform Coding of Speech Signals*"; IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 25, N. 4, Aug. 1977.

CODIFICAÇÃO DE ÁUDIO

[A1] A. Sugiyama *et al.*; "*Adaptive Transform Coding with an Adaptive Block Size*"; ICASSP 1990, pp. 1093-1096.

[A2] Dieter Bauer e Dieter Seitzer; "*Statistical Properties of High Quality Stereo Signals in the Time Domain*"; ICASSP 1989, pp. 2045-2048.

[A3] Dieter Bauer e Dieter Seitzer; "*Frequency Domain Statistics of High Quality Stereo Signals*"; AES 86th Convention, March 1989, preprint 2748(E-1).

[A4] EBU SQAM-*Subjective Quality Assessment Material: Recordings for Subjective Tests*. EBU Tech. 3253-E, April 1988.

[A5] E. F. Schroeder *et al.*; "*MSC: Stereo Audio Coding with CD-Quality and 256 Kbit/s*"; Trans. Consumer Electronics, Vol. 33, N. 4, Nov. 1987.

[A6] G. Davidson, L. Fielder e Mike Antill; "*High-Quality Audio Transform Coding at 128Kbit/s*"; ICASSP 1990, pp. 1117-1120.

[A7] G. Theile, G. Stoll e M. Link; "*Low Bit-Rate Coding of High-Quality Audio Signals*"; AES 82nd Convention, 1987, preprint 2796.

[A8] Hans Georg Musmann; "*The ISO Audio Coding Standard*"; GlobeCOM90, pp. 511-517, Dec. 1990.

[A9] H. Jakubowski e G. Spikofski; "*SQAM-The EBU Compact Disc for Subjective Assessments of Audio Systems*"; EBU Review, N. 227, Feb. 1988.

[A10] James D. Johnston; "*Estimation of Perceptual Entropy Using Noise Masking Criteria*"; ICASSP 1988, pp. 2524-2527.

[A11] James D. Johnston; "*Transform Coding of Audio Signals Using Perceptual Noise Criteria*"; IEEE J. Selected Areas in Communications, Vol. 6, N. 2, Feb. 1988.

[A12] James D. Johnston; "*Perceptual Transform Coding of Wideband Stereo Signals*"; ICASSP 1989, pp. 1993-1996.

[A13] James D. Johnston e Aníbal Ferreira; "*Sum-Difference Stereo Transform Coding*"; ICASSP 1992, a ser publicado

[A14] Joel Soumagne *et al.*; "*A Comparative Study of the Proposed High Quality Coding Schemes for Digital Music*"; ICASSP 1986, pp. 21-24.

[A15] J. Reeds; Comunicação Privada.

[A16] K. Brandenburg; "*High Quality Sound Coding at 2.5 bit/sample*"; AES 84th Convention, March 1988, preprint 2582(D-2).

[A17] K. Brandenburg e D. Seitzer; "*OCF: Coding High Quality Audio with Data Rates of 64Kbit/sec*"; AES 85th Convention; November 1988, preprint 2723(H-6).

[A18] K. Brandenburg e J. D. Johnston; "*Second Generation Perceptual Audio Coding: The hybrid Coder*"; AES 88th Convention, March 1990, preprint 2937(H-3).

[A19] K. Brandenburg *et al.*; "*ASPEC: Adaptive Spectral Entropy Coding of High Quality Music Signals*"; AES 90th Convention, February 1991, preprint 3011(A-4).

[A20] M. A. Krasner; "*Digital Encoding of Speech and Audio Signals based on the Perceptual Requirements of the Auditory System*"; MIT tech. Rep. 535, Lincoln Laboratories, 1979.

[A21] Raymond N. J. Veldhuis et al.; "*Subband Coding of Digital Audio Signals Without Loss of Quality*"; ICASSP 1989, pp. 2009-2012.

[A22] Reinhold Orglmeister; "*Data Reduction in High-Quality Audio Signals*"; AES 86th Convention, March 1989, preprint 2751(E-5).

[A23] R. G. van der Waal; Raymond N. J. Veldhuis; "*Subband Coding of Stereophonic Digital Audio Signals*"; ICASSP 1991; pp. 3601-3604.

[A24] S. Furui e M. M. Sondhi; "Advances in Speech Signal Processing"; Capítulo 4: "*Wideband Coding - Perceptual Considerations for Speech and Music*" por J. D. Johnston e K. Brandenburg; Bartlett Press Inc., 1991.

[A25] Sharad Singhal; "*High Quality Audio Coding Using Multipulse LPC*"; ICASSP 1990, pp. 1101-1104.

[A26] Wai-Yip Chan e Allen Gersho; "*High Fidelity Audio Transform Coding with Vector Quantization*"; ICASSP 1990, pp. 1109-1112.

[A27] Wai-Yip Chan e Allen Gersho; "*Constrained-Storage Vector Quantization in High Fidelity Audio Transform Coding*"; ICASSP 1991, pp. 3597-3600.

[A28] Y. F. Dehery, M. Lever e P. Urcun; "*A MUSICAM Source Codec for Digital Audio Broadcasting and Storage*"; ICASSP 1991, pp. 3605-3608.

[A29] Y. Mahieux, J. P. Petit e A. Charbonnier; "*Transform Coding of Audio Signals using Correlation between Successive Transform Blocks*"; ICASSP 1989, pp. 2021-2024.

[A30] Y. Mahieux e J. P. Petit; "*Transform Coding of Audio Signals at 64Kbit/s*"; GlobeCOM90, pp. 518-522, Dec. 1990.

[A31] ASPEC-ISO/MPEG Audio Coding Algorithm Report, 1990.

[A32] MUSICAM-ISO/MPEG Audio Coding Algorithm Report, 1990.

[A33] JIWP 10-CMTT/1/Test Group; "*Requirements on Low Bit Rate Digital Audio Coding Systems*"; CCIR, June 1991.

BANCOS DE FILTROS. TRANFORMADAS

[T1] B. Edler; "*Codierung von Audiosignalen mit Uberlappender Transformation und Adaptiven Fensterfunktionen*"; Frequenz, Sep. 1990.

[T2] John P. Princen e Alan Bernard Bradley; "*Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation*"; IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 34, N. 5, Oct. 1986.

[T3] J. P. Princen, A. W. Johnson e A. B. Bradley; "*Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation*"; ICASSP 1987, pp. 2161-2164

[T4] John W. Adams; "*A New Optimal Window*"; IEEE Trans. Signal Processing, Vol. 39, N. 8, Aug. 1989.

[T5] J. Spille e E. F. Schroder; "*Design of an Optimum Filterbank for High Quality Subband Audio Coding*"; AES 86th Convention, March 1989, preprint 2749(E-2).

[T6] Maurice Bellanger; "*Digital Processing of Signals*"; John Willey & Sons, 1989.

[T7] Miodrag Temerinac e Bernd Edler; "*LINC: a Commom Theory of Transform and Subband Coding*"; trabalho não publicado.

[T8] P. P. Vaidyanathan e Phuong-Quan Hoang; "*The Perfect-Reconstruction QMF Bank: New Architectures, Solutions and Optimization Strategies*"; ICASSP 1987, pp. 2169-2172.

[T9] Ronald E. Crochiere e Lawrence R. Rabiner; "*Multirate Digital Signal Processing*"; Prentice-Hall, Englewood Cliffs, 1983.

PSICO-ACÚSTICA

[P1] B. E. Mulligan, M. J. Mulligan e J. F. Stonecypher; "*Critical Band in Binaural Detection*"; J. Acoustical Soc. of America, Vol. 41, N. 1, 1967.

[P2] Brian C. J. Moore; "*An Introduction to the Psychology of Hearing*"; Academic Press, 1982.

[P3] Ervin R. Hafter e Samuel C. Carrier; "*Masking-Level Differences Obtained with a Pulsed Tonal Masker*"; J. Acoustical Soc. of America, Vol. 47, N. 4, 1970.

[P4] E. Zwicker; "*Subdivision of the Audible Frequency Range into Critical Bands*"; J. Acoustical Soc. of America, Vol. 33, N. 2, 1961.

[P5] E. Zwicker e H. Fastl; "*Psychoacoustics, Facts and Models*"; Springer-Verlag, 1990.

[P6] Frederic L. Wightman; "*Binaural Masking with Sine-Wave Maskers*"; J. Acoustical Soc. of America, Vol. 45, N. 1, 1969.

[P7] Frederic L. Wightman; "*Detection of Binaural Tones as a Function of Masker Bandwidth*"; J. Acoustical Soc. of America, Vol. 50, N. 2, 1971.

[P8] James M. Kates; "*An Adaptive Digital Cochlear Model*"; ICASSP 1991, pp. 3621-3624.

[P9] Jerry V. Tobias; "*Foundations of Modern Auditory Theory*"; Volume I, Academic Press, 1970.

[P10] Jerry V. Tobias; "*Foundations of Modern Auditory Theory*"; Volume II, Academic Press, 1972.

[P11] Manfred R. Schroeder; "*Models of Hearing*"; Proceedings of the IEEE, Vol. 63, N. 9, Sep. 1975.

[P12] M. R. Schroeder; In "*Recognition of Complex Acoustic Signals*", Life Sciences Research Report 5, edited by T. H. Bullock, pp. 323-328, 1977.

[P13] M. Konishi; "*Spatial Localization of Sound*"; Dahlem workshop on Recognition of Complex Acoustic Signals, pp. 127-143, 1976.

[P14] N. I. Durlach; "*Equalization and Cancellation Theory of Binaural Masking-Level Differences*"; J. Acoustical Soc. of America, Vol. 35, N. 8, Aug. 1963.

[P15] N. I. Durlach; "*Note on Binaural Masking-Level Differences at High Frequencies*"; J. Acoustical Soc. of America, Vol. 36, N. 3, Mar. 1964.

[P16] P. J. Metz, G. von Bismark e N. I. Durlach; "*Further Results on Binaural Unmasking and the EC Model II. Noise Bandwidth and Interaural Phase*"; J. Acoustical Soc. of America, Vol. 43, N. 5, 1968.

[P17] Rhona P. Hellman; "*Effect of Noise Bandwidth on the Loudness of a 1000-Hz Tone*"; J. Acoustical Soc. of America, Vol. 48, N. 2, 1970.

[P18] Rhona P. Hellman; "*Asymmetry of Masking between Noise and Tone*"; Perception and Psychophysics, Vol. 11, N. 3, 1972.

[P19] Robert W. Hawley e Jont B. Allen; "*Masking Models for Noise and Tone*"; Relatório Interno Bell Labs.

[P20] Stanley A. Gelfand; "*Hearing-An Introduction to Psychological and Physiological Acoustics*"; Marcel Dekker Inc., 1990.

8
APÊNCICE

INSTRUCTIONS

Thank you for participating in our tests of high quality wideband coders. Your task as a subject is to determine the difference between the original and coded signals presented to you, and to rate the quality of the coded signals. The quality of the coded signals is very good so you will have to listen for subtle differences in each case.
We will be testing several different coders whose coded signals will be presented in a random order.

Our test will consist of a complete presentation for each one of 6 different music samples. Between presentations there is a 3 minute break. Each presentation involves 5 sets and is organized as follows:

SET1-◆-◆-SET2-◆-◆-SET3-◆-◆-SET4-◆-◆-SET5

each SET is 1 ABC-◆-2 ABC-◆-3 ABC

A = original (reference for grade 5 , ≈ 10 seconds long)
B = original or coded signal (≈ 10 seconds long)
C = the opposite of **B**
\- = pause (≈ 1 second long)
◆ = short beep (≈ 0.5 seconds long)

Each set has a trio repeated three times. Within a particular set, the first segment, **A**, is always the original. We ask you to rate segments **B** and **C**. Of these, one will be the coded signal and the other will be the original.

At the end of each set, you should rate both signal **B** and signal **C** on the following scale:

1-very annoying degradation
2-annoying degradation
3-slightly annoying degradation
4-perceptable difference
5-indistinguishable from original

For each music sample use the corresponding rating sheet. Please rate the signal that you believe to be the original as a "**5.0**" and the other signal that you believe to be the coded according to the scale. One digit after the decimal point is allowed.

Please compare these instructions with the rating sheets. If you have questions please ask now.

Thank you for your collaboration.

J.D. Johnston
Aníbal Ferreira

music sample nº 6

(Castanets)

1-very annoying degradation
2-annoying degradation
3-slightly annoying degradation
4-perceptable difference
5-indistinguishable from original

set1
A 5.0
B ___.___
C ___.___

set2
A 5.0
B ___.___
C ___.___

set3
A 5.0
B ___.___
C ___.___

set4
A 5.0
B ___.___
C ___.___

set5
A 5.0
B ___.___
C ___.___