Capítulo 7

Métodos de Codificação de Fala

Homer Dudley em 1939, com a introdução do *vocoder* de canal e o sintetizador electrónico *Voder*, iniciou a moderna investigação sobre codificação de sinais de fala. Neste capítulo, resumem-se alguns dos métodos de codificação mais importantes entretanto propostos, com incidência na codificação de sinais de banda telefónica. Levando em consideração os atributos descritos no capítulo 5, os receptores empregam um dos modelos de síntese apresentados no capítulo 4, que tiram partido das características dos sinais de fala descritos no capítulo 2. O emissor inclui o respectivo método de estimação de parâmetros descrito no capítulo 3. Muitos dos modelos de síntese referidos foram na realidade desenvolvidos no contexto da codificação. Os parâmetros

transmitidos deverão ser quantificados eficientemente, utilizando nomeadamente as técnicas apresentadas no capítulo 6.

Seguindo uma perspectiva histórica, nas secções 7.1 e 7.2 os codificadores serão divididos respectivamente em codificadores de forma de onda (*waveform-approximating coders*), de baixa complexidade e atraso, caracterizados por seguirem o sinal amostra-a-amostra com grande qualidade, e em codificadores de fonte ou *vocoders*, que descrevem o sinal de fala de um modo paramétrico, conseguindo uma diminuição do débito binário à custa da diminuição da qualidade dos sinais sintetizados e do aumento do atraso e da complexidade. O excelente artigo de Flanagan *et al.* [Flanagan (79)] inclui uma descrição pormenorizada destas duas classes de codificação, já conhecidas na década de 70.

Seguidamente, nas secções 7.3 e 7.4, são descritos dois métodos de codificação que surgiram na década de 80, em que a divisão anterior é esbatida: a codificação análise-por-síntese baseada em predição linear (LPAS - *Linear Prediction Analysis-by-Synthesis*) e a codificação sinusoidal. Estes codificadores aproveitam as vantagens dos dois tipos de codificadores anteriores, pelo que são conhecidos por codificadores híbridos. Por um lado são codificadores de forma de onda, pois tentam reproduzir o sinal amostra-a-amostra, e por outro lado utilizam um modelo paramétrico com o objectivo de diminuir o débito binário, tal como nos *vocoders*. Uma descrição pormenorizada destes dois métodos de codificação pode ser encontrada respectivamente nos capítulos 3 [McAulay (95)] e 4 [Kroon (95)] do livro de [Kleijn (95)]. Trancoso *et al.* apresentam também uma comparação entre estes dois métodos [Trancoso (90)]. Como mostra a figura 7.1, estes codificadores vieram preencher a zona de débito binário não ocupada pelas duas classes anteriores, conhecida no fim da década de 70 por *lacuna na codificação*

[Atal (79)]. Nesta figura, a qualidade é medida através da escala MOS, sendo a qualidade *má* raramente atribuída e a *excelente* normalmente só atribuída a codificadores de sinais de banda larga (50-7000 Hz).

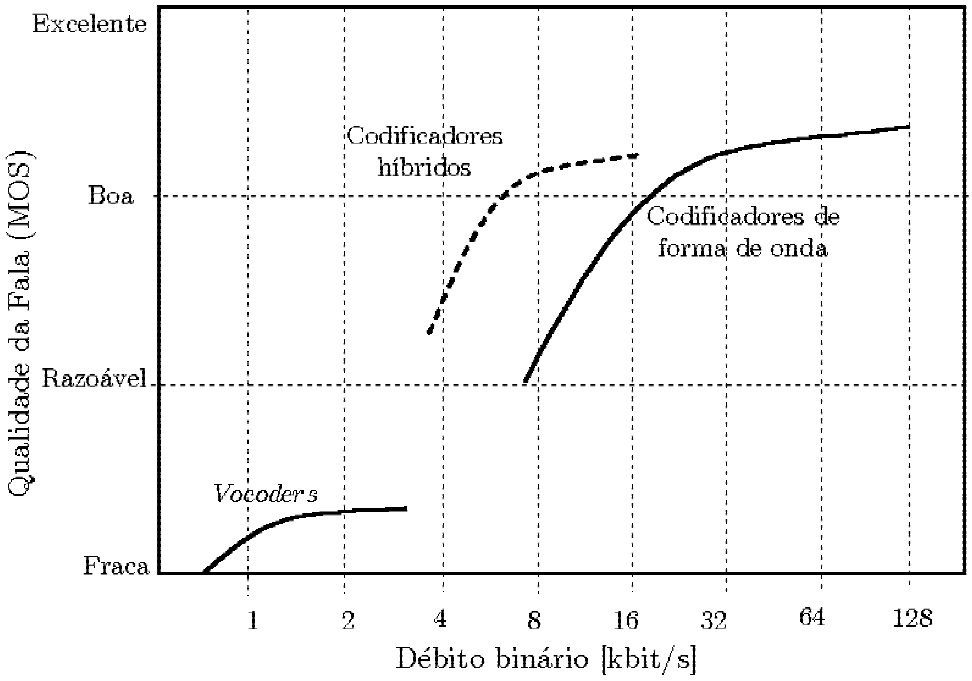

Figura 7.1
Qualidade da fala do sinal sintetizado, função do débito binário e do tipo de codificação, para sinais de banda telefónica.

Nas secções 7.5 e 7.6, resumiremos dois métodos de codificação propostos já na década de 90: o codificador por predição linear com excitação mista (MELP - *Mixed Excitation Linear Prediction*) e o codificador por interpolação da forma de onda (WI - *Waveform Interpolation*). Estes dois métodos de codificação permitiram uma qualidade na gama de débito binário entre os 2400 a 4800 bit/s que nem mesmo os codificadores híbridos alcançavam. Uma descrição pormenorizada da codificação WI pode ser encontrada no capítulo 5 [Kleijn (95-b)] do livro de [Kleijn (95)].

Na secção 7.7 serão apresentadas diversas estratégias para codificar sinais de fala a muito baixo débito binário, em que se espera

um grande desenvolvimento na década de 2000 a 2010. Como complemento a esta secção, pode-se consultar uma pesquisa bibliográfica que cobre desde o ano de 1956 até 1991, correspondente a codificadores de fala de débito binário até aos 2400 bit/s [Jaskie (92)].

A divisão cronológica seguida é baseada na data da introdução do modelo e não na data em que o método de codificação obteve maturidade para emergir enquanto norma. Estes dois factos ocorrem normalmente com cerca de uma década de diferença, pelo que a codificação de forma de onda e os *vocoders* resistiram até ao fim da década de 80, enquanto que os codificadores híbridos foram dominantes durante a década de 90, com grande ênfase para uma das variantes do método LPAS, o codificador CS-ACELP (*Conjugate Structure Algebraic Code-Excited Linear Prediction*), que por esse facto é tratado numa subsecção. Estes métodos, apresentados na tabela 7.1, são na sua maioria baseados em predição linear (LPC), diferindo essencialmente nas estratégias de modelação do resíduo de predição.

Muitos dos codificadores de fala utilizam sistemas de detecção de actividade da voz (VAD - *voice activity detection*) (*e.g.*, G.729 [Benyassine (97)], sistema GSM [Vahatalo (99)]). Quando da ausência de voz o codificador transmite a informação do ruído de fundo com menor débito binário, ou simplesmente o receptor injecta *ruído de conforto*. A detecção é baseada na comparação do nível de energia do sinal com uma estimação do ruído de fundo, tendo em consideração características do sinal de fala tal como a periodicidade. Estes sistemas podem ainda ser utilizados quando da introdução de canceladores de eco e de sistemas de redução de ruído, em que é necessária uma boa estimativa do ruído de fundo.

Método de Codificação	Débito binário [kbit/s]	Estimação espectral
Forma de onda		
PCM *companding* [Smith (57)]	64	
(A)DPCM [Cutler (52)]	16-40	
(A)DM [Schouten (52)]	16	
APC [Atal (70)]	10	LPC
ATC [Zelinski (77)]	16- 40	transformada
SBC [Crochiere (77)]	16 - 64	
Vocoder		
Canal[Dudley (39)] [Bell Labs (56)]	0,4 – 1,2	canal
Formantes [SRDE Labs (61)]	0,4 – 1,2	formantes
LPC[Atal (71)]	2,4	LPC
LPAS		
Multi-pulso[Atal (82)]	16	LPC
RPE [Deprettere (85)]	13	LPC
CELP [Atal (84)]	4,8 - 12,2	LPC
Sinusoidal		
Sinusoidal [Almeida (82)]	1,5 – 9,6	sinusoidal
MELP		
MELP [McCree (91)]	2,4	LPC
WI		
WI [Kleijn (91)]	2,4 – 4,8	LPC
Muito baixo débito binário		
Quantificação vectorial [Wong (81)]	0,8 – 1,5	LPC,formantes,canal
Interpolação [McAulay (81)]	0,5 – 1	LPC,formantes
Super tramas [Fette (91)]	0,6 – 0,8	LPC
Quantificação matricial [Tsao (85)]	0,2 – 0,6	LPC,formantes
Quantificação segmental [Roucos (82)]	0,2 – 0,6	LPC,formantes
Codificação fonética [Schwartz (80)]	0,1 – 0,6	LPC,formantes

Tabela 7.1
Métodos de codificação de sinais de fala.

As conclusões do capítulo são apresentadas na secção 7.8, discutindo-se a evolução dos codificadores de sinais de fala desde a década de 70 a 2000 e perspectivando-se os futuros desenvolvimentos. Nesta secção são ainda apresentadas, na figura 7.18 e tabela 7.10, algumas das normas de codificação e respectivos atributos.

7.1 Codificação de forma de onda

Os codificadores de forma de onda podem dividir-se em codificadores no domínio do tempo e no domínio da frequência. A sua característica é a aproximação, à medida que o número de bits de codificação vai aumentando, do sinal de saída ao de entrada, pelo que podem ser aferidos objectivamente através da relação sinal-ruído.

7.1.1 Codificadores no domínio do tempo

Os codificadores de forma de onda no domínio do tempo têm um funcionamento amostra-a-amostra. Os codificadores sem qualquer predição correspondem aos codificadores utilizando modulação por código de impulsos (PCM), normalmente empregando quantificação uniforme ou pseudo-logarítmica (*companding*). A lei-μ, pseudo-logarítmica, foi proposta em 1957 por Smith [Smith (57)] [Rabiner (78)], embora só em 1972, juntamente com a lei-A, tenham sido constituídas como a recomendação G.711 do CCITT para utilização em centrais telefónicas. Esta foi a primeira norma em codificação de fala e utiliza 8 bits de codificação por amostra, a que corresponde um débito binário de 64 kbit/s.

A vantagem da utilização de *companding*, em relação à quantificação uniforme, é manter a relação sinal-ruído constante para uma gama alargada da potência do sinal de entrada, dependendo esta apenas do número de intervalos de quantificação e logo do débito binário ($\approx$38 dB para 8 bits de codificação por amostra, conforme a equação 6.16). O *companding* tem também uma melhor adequação à distribuição do sinal de fala, já que este tem maiores ocorrências para valores baixos. A relação sinal-ruído da quantificação uniforme só supera a da quantificação empregando *companding* para potências

normalizadas do sinal de entrada maiores do que -15 dB, impossíveis de atingir para um sinal de fala. Um outro factor importante conseguido pelo *companding* é o aumento da qualidade perceptual em relação ao PCM uniforme. É sabido que o aparelho auditivo é menos sensível ao ruído em zonas de maior potência, pelo que se quantifica as amostras com valores menores com maior acuidade em detrimento das amostras com valores mais elevados. Pode-se então concluir do melhor desempenho objectivo (aumento médio da SNR) e subjectivo (melhoria da qualidade perceptual) do PCM com *companding* em relação ao PCM uniforme.

Os esquemas diferenciais datam de 1952 [Rabiner (78)] (DM - *Delta Modulation* [Schouten (52)]) (DPCM - *Differential Pulse Code Modulation* [Cutler (52)]), sendo posteriormente propostas diversas variantes com quantificação adaptativa. No entanto, só em 1983 o CCITT aprovou a recomendação G.721, utilizando modulação por código de impulsos diferencial adaptativo (ADPCM - *Adaptive Differential Pulse Code Modulation*) [Jayant (84)]. A informação transmitida corresponde à diferença entre o valor da amostra e uma sua predição (resíduo de predição), codificada com 4 bits por amostra através de um quantificador adaptativo, totalizando um débito binário de 32 kbit/s. A adaptação tem como base amostras quantificadas (AQB - *Adaptive Quantization with Backward Estimation*), não sendo necessário enviar informação adicional, pois esta pode ser determinada no receptor.

Na recomendação G.721, cujo esquema de blocos é apresentado na figura 7.2, a predição é efectuada por um preditor adaptativo com 2 pólos e 6 zeros. Os coeficientes de predição são estimados com uma variante do algoritmo LMS, minimizando o valor absoluto do erro e não o erro quadrático. Como a informação para adaptação assenta nas

amostras passadas do sinal de saída, também não é necessária a transmissão da informação dos coeficientes de predição. A estabilidade do filtro com apenas dois pólos é facilmente verificada, melhorando-se a predição com a introdução dos 6 zeros. A aplicação principal deste codificador é nas redes telefónicas, duplicando o número de conversações conseguidas pela recomendação G.711.

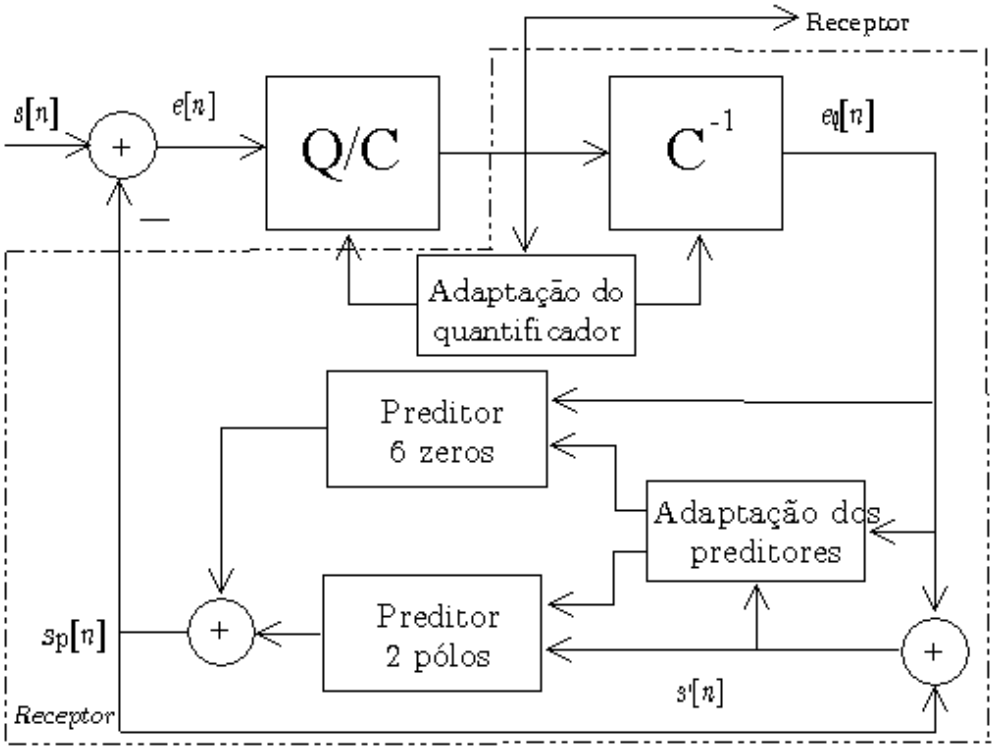

Figura 7.2
Codificador ADPCM, Norma G.721, G.723, G.726.

A recomendação G.723, de 1988, utiliza o mesmo método de codificação da recomendação G.721, mas produz dois débitos binários diferentes, 24 kbit/s e 40 kbit/s, correspondendo a 3 e 5 bits de codificação por amostra. A recomendação G.726[1] de 1990 é uma unificação das normas G.721 e G.723, que tecnicamente desapareceram, incluindo adicionalmente a codificação a 16 kbit/s, correspondente a 2 bits de codificação por amostra.

[1] Demonstração na página da disciplina (/demos/cod_forma/cod_forma.html)

A recomendação G.727 também de 1990 é uma extensão da recomendação G.726, incluindo os mesmos débitos, mas a codificação é embebida possibilitando extrair bits entre o emissor e o receptor, atributo necessário quando da sobrecarga de sistemas de transmissão de pacotes com multiplexagem de diversas fontes. O codificador ADPCM básico utiliza 2, 3 ou 4 bits por amostra, sendo o erro de quantificação transmitido em PCM com 3, 2 ou 1 bit por amostra. Os bits do codificador básico não podem ser extraídos mantendo uma qualidade mínima, enquanto que em situações de sobrecarga na rede os bits que representam o erro de quantificação podem ser extraídos, principiando naturalmente pelos bits de menor peso.

Em 1970 Atal e Schroeder propuseram o codificador por predição adaptativa (APC - *Adaptive Predictive Coding*) [Atal (70)], cujo esquema de blocos foi apresentado na figura 3.1 e que utilizava como predição dois filtros de predição linear: um explorava a correlação entre amostras consecutivas, modelada através de um filtro LPC e o outro explorava a quase-periodicidade existente nas zonas vozeadas, modelada através de um preditor de longa duração. A grande inovação de princípio do APC em relação ao ADPCM é a de que se utiliza um preditor LPC por trama (5 ms) estimado a partir do sinal de entrada, e não o algoritmo LMS. Embora aumentando o atraso e tendo que transmitir os coeficientes de predição, este modela melhor os sinais de fala que a estrutura de 6 zeros e 2 pólos utilizada na norma G.726. Note-se que é complexo verificar a estabilidade do filtro resultante utilizando o algoritmo LMS quando o preditor tem mais de 2 pólos, sendo mais fácil verificar a estabilidade para o preditor LPC.

O resíduo de dupla predição tem uma menor gama dinâmica que o sinal original e uma característica espectral branqueada. Este resíduo, na simulação em computador descrita por Atal, era codificado em

modulação delta adaptativa (ADM - *Adaptive Delta Modulation*) [Jayant (84)], ou seja, com um codificador diferencial adaptativo de apenas 1 bit por amostra. Os sinais eram amostrados a 6,67 kHz e apesar de a simulação não quantificar os coeficientes de predição, pressupunham os autores ser possível atingir um total de cerca de 10 kbit/s com uma qualidade bastante razoável, o que deixaria 3,33 kbit/s para quantificar os coeficientes de predição. Sabe-se hoje que este valor é mais do que suficiente. Uma outra contribuição do APC foi a inclusão de um esquema de mascaramento auditivo do ruído de quantificação, enaltecendo-o nas zonas dos formantes a que corresponde uma energia mais alta mas tentando minimizá-lo nas zonas entre formantes, perceptualmente mais sensíveis.

Atal comparou perceptualmente a qualidade do codificador APC com a de um codificador PCM lei-μ com 6 bits por amostra, tendo como entradas os mesmos sinais e a mesma frequência de amostragem. Mesmo com o codificador PCM a funcionar com um débito binário de 40 kbit/s, a qualidade foi considerada muito parecida com a do APC, embora a comparação não seja inteiramente justa pois este último não era totalmente quantificado.

7.1.2 Codificadores no domínio da frequência

No domínio da frequência destacam-se dois tipos de codificadores de forma de onda: o codificador por transformada adaptativo (ATC - *Adaptive Transform Coding*) e o codificador sub-banda (SBC - *Sub-band Coding*).

O codificador ATC, proposto por Zelinski e Noll [Zelinski (77)], divide o espectro do sinal num número suficientemente grande (na ordem das 256) de bandas de frequência, de forma a ser garantida uma

resolução capaz de discriminar a estrutura de riscas típica das zonas vozeadas. A síntese é produzida por transformada em bloco (com *overlap add*). Para quantificar eficientemente estas bandas os bits disponíveis são distribuídos adaptativamente por trama, quantificando com maior acuidade (3-5 bits por coeficiente) as baixas frequências e as bandas de maior energia correspondentes aos formantes e com menor acuidade as altas frequências e as zonas entre formantes. As zonas entre formantes, contudo, embora não contribuam para a inteligibilidade, são perceptualmente bastante sensíveis ao ruído de quantificação, pois este não é mascarado por uma energia de sinal elevada.

Para uma codificação eficiente, várias transformadas são possíveis. A transformada de Karhunen-Loève (KLT) é óptima, mas é computacionalmente complexa e requer funções de base dependentes do sinal de entrada. A DFT é também uma alternativa, computacionalmente simples e com funções de base sinusoidais fixas, mas não atinge a qualidade obtida pela KLT. A transformada mais utilizada é a DCT (*Discrete Cosine Transform*), também computacionalmente simples e com funções de base fixas, mas que se aproxima da qualidade obtida com a KLT quando o sinal tem uma auto-correlação elevada [O'Shaughnessy (99)]. A DCT é muito utilizada na codificação de imagens (JPEG) e vídeo (H.261, MPEG-1, MPEG-2).

Na codificação SBC o espectro do sinal é dividido num pequeno número de bandas (2 a 8) [Crochiere (77)], sendo o sinal de cada banda decimado e codificado utilizando técnicas no domínio do tempo. Como exemplo, a recomendação G.722 do ITU-T para codificação de sinais de áudio ou sinais de fala de banda larga (50-7000 Hz), que data de 1988, como mostrado na figura 7.3 divide o espectro em duas sub-bandas, codificando cada uma com codificadores ADPCM similares ao da recomendação G.727. O sinal da banda alta é codificado com 16 kbit/s

e o da banda baixa com 48 kbit/s, num total de 64 kbit/s. Na codificação de banda baixa estão embebidos codificadores de 40 e 32 kbit/s. No receptor são gerados os sinais das duas bandas e somados para produzir o sinal de banda larga.

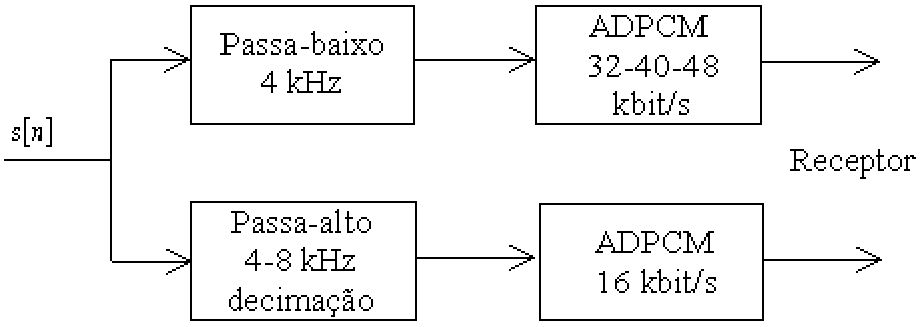

Figura 7.3
Codificador SBC, recomendação G.722 - emissor.

Uma vez que a qualidade da codificação da banda de baixas frequências é praticamente igual à da recomendação G.711, a qualidade final é melhor que a deste para o mesmo débito binário, pois o sinal torna-se mais inteligível com a inserção da banda de altas frequências. Repare-se que em testes subjectivos comparando fala de banda telefónica com fala de banda larga, esta última obtém uma qualidade na escala MOS em alguns casos 1 ponto acima.

Uma descrição pormenorizada da codificação de banda larga, incluindo uma melhor descrição sobre a recomendação G.722, pode ser encontrada no capítulo 8 [Adoul (95)] do livro de [Kleijn (95)].

7.2 *Vocoders*

Os codificadores de fonte ou *vocoders* utilizam um modelo simplificado da produção da fala e não reproduzem fielmente o sinal de entrada, mas apenas o representam de modo a manter as suas

características mais importantes, tais como a envolvente espectral, a estrutura fina do espectro e a energia global.

O primeiro codificador de sinais de fala a ser proposto, por Dudley em 1939 [Dudley (39)] [Jaskie (92)], foi um *vocoder* de canal. Este tipo de codificadores representa os sinais de fala no domínio da frequência, através da energia de um conjunto de filtros passa-banda contíguos, que dividem a banda total num número fixo de canais. É transmitida a energia correspondente a cada canal, a decisão de vozeamento e o valor da frequência fundamental, sendo estes parâmetros utilizados no receptor para controlar um banco de filtros ressonantes, que sintetizam o sinal. O primeiro *vocoder* a utilizar a tecnologia de processamento digital de sinais foi também um codificador deste tipo, proposto pelos Laboratórios Bell em 1956, com um débito binário de 800 bit/s [Jaskie (92)]. Além da informação sobre a decisão de vozeamento e da frequência fundamental, esta arquitectura não tira partido de outras particularidades dos sinais de fala, pelo que a qualidade dos sinais sintetizados é inferior à obtida por outros métodos de codificação entretanto propostos.

Em 1961 foi proposto pelos Laboratórios SRDE em Inglaterra um *vocoder* de formantes, com um débito binário de 864 bit/s [Jaskie (92)]. Estes são também codificadores no domínio da frequência, mas, ao contrário dos codificadores de canal, tiram melhor partido das características do sinal de fala. O espectro dos sinais de fala tem picos nas zonas dos formantes, cuja caracterização é importante em termos da inteligibilidade. Os valores das frequências dos formantes e as respectivas larguras de banda são codificados juntamente com a decisão de vozeamento e da frequência fundamental. Estes parâmetros são utilizados no receptor para sintetizar o sinal, controlando um banco de filtros sintonizáveis com as frequências dos formantes.

Tanto o tracto vocal como o período de vibração das cordas vocais variam lentamente durante a produção da fala, podendo considerar-se como tendo parâmetros constantes num curto intervalo de tempo. Para simular o tracto vocal pode ser utilizado um filtro LPC [Atal (71)], cujos parâmetros são reestimados a intervalos que variam tipicamente entre os 20 e os 30 ms. Este filtro é excitado (modelo fonte-filtro), como ilustra a figura 7.4, com ruído branco para produzir sons não vozeados, ou com um trem de pulsos com período igual ao da vibração das cordas vocais, para produzir sons vozeados.

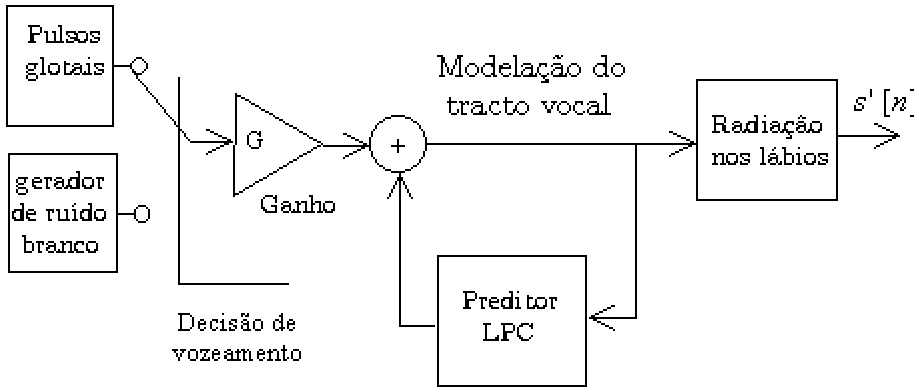

Figura 7.4
Esquema de blocos (sintetizador) de um *vocoder* LPC.

Tirando partido da coarticulação e da pequena variação do tracto vocal ao longo do tempo, os parâmetros de controlo são interpolados sincronamente com os períodos glotais, minimizando as variações entre tramas que podem ser audíveis. Devido ao pequeno número de parâmetros de controlo deste modelo conseguem-se implementar codificadores com débitos binários bastante baixos. São no entanto pouco robustos porque o modelo da excitação é demasiado simples e devido a possíveis erros na decisão de vozeamento e na estimação da frequência fundamental, resultando numa qualidade perceptual, na melhor das hipóteses, *razoável*.

Como exemplo deste tipo de codificação encontra-se a norma FS-1015 [Tremain (82)] [FED-STD-1015 (84)] de 1984, do

Departamento de Defesa dos Estados Unidos da América, coincidente com a norma NATO STANAG 4198. Esta utiliza uma trama de 22,5 ms e produz um débito de 2400 bit/s, capaz de ser transmitido sobre linhas telefónicas com a tecnologia de *modems* então existente. A distribuição de bits dos diversos parâmetros é mostrada na tabela 7.2.

Parâmetros	Zonas Vozeadas	Zonas não Vozeadas	bits por segundo
Coeficientes LPC	41	20	1823/889
RMS (*Root Mean Square*)	5	5	222
Frequência fundamental & V/UV	7	7	310
Protecção contra erros	0	20	0/889
Sincronismo	1	1	45
Não usado	0	1	0/45
Total	54	54	2400

Tabela 7.2
Distribuição dos bits na norma FS-1015.

7.3 Codificação análise-por-síntese baseada em predição linear

Se por um lado os *vocoders* não conseguem melhor que uma qualidade *razoável* mesmo acima dos 2,4 kbit/s, por outro os codificadores de forma de onda têm uma degradação acentuada abaixo dos 16 kbit/s. Esta lacuna na codificação, entre os 2,4 e os 16 kbit/s, permaneceu até ao início da década de 80. Em 1982 Atal e Remde [Atal (82)] propõem o codificador *multi-pulso* que altera completamente este cenário. A inovação introduzida em relação ao APC consiste na modelação da excitação da predição de curta duração através apenas dos impulsos *mais importantes* (vectorial) e não amostra-a-amostra, sendo estes impulsos escolhidos por meio de um procedimento de análise-por-síntese (cadeia fechada) baseada em predição linear (LPAS - *Linear Prediction Analysis-by-Synthesis*), cujo esquema de blocos é apresentado na figura 7.5. Este procedimento garante uma elevada qualidade quer em termos da SNR quer da qualidade perceptual.

A introdução de um filtro de predição de longa duração [Singhal (84)], tal como já proposto no APC, acrescentou ao codificador *multi-pulso* um aumento significativo da qualidade. Por outro lado, comparando com o *vocoder* LPC que utiliza o mesmo tipo de preditor, o aumento da informação sobre o resíduo de predição torna o modelo LPAS mais robusto e produzindo melhor qualidade.

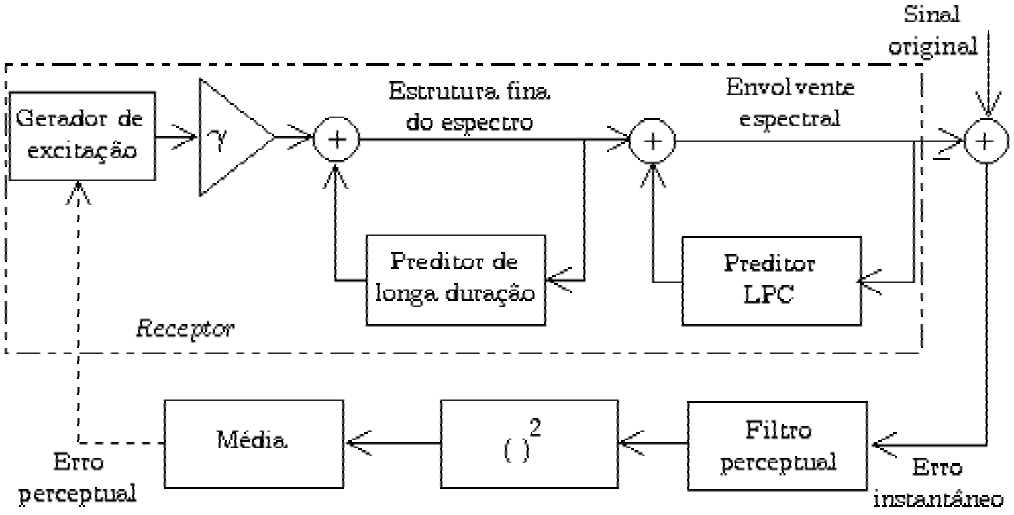

Figura 7.5
Esquema de blocos de um codificador LPAS.

Uma outra inovação proposta por Atal e Rende consiste na introdução de um filtro perceptual antes do cálculo da energia do erro de predição. A potência do erro de predição tende a distribuir-se igualmente em frequência. O filtro perceptual é um filtro de de-ênfase aos picos da envolvente espectral (formantes), que melhora a relação sinal-ruído perceptual à custa da distorção do próprio sinal. A ideia por detrás deste filtro é a de que o ouvido humano suporta melhor o ruído nas zonas de frequência em que existe uma energia razoável do sinal, mas é bastante crítico em relação a zonas de frequência com baixa energia. Representando o filtro perceptual $W(z)$ pela equação:

$$W(z) = \frac{A(z)}{A(\gamma z)}, \tag{7.1}$$

em que $A(z)$ representa o filtro inverso de síntese LPC e γ um parâmetro entre 0 e 1 que controla o grau de mascaramento auditivo. Se γ =1, então $W(z)$=1 e não existe mascaramento auditivo. Se γ=0, então $W(z)$ corresponde ao filtro inverso LPC. Neste caso é retirada importância ao ruído na banda dos formantes e é minimizado o ruído que está fora dessas bandas. São tipicamente utilizados valores de γ entre 0,75 e 0,95. A figura 7.6 ilustra um exemplo da envolvente espectral de uma trama, estimada a partir do filtro LPC e a respectiva resposta do filtro perceptual com γ = 0,8.

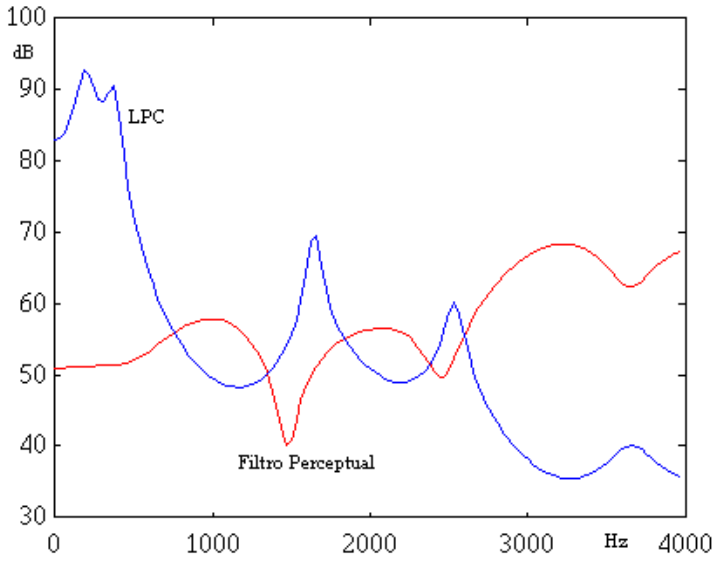

Figura 7.6
Filtragem Perceptual
Resposta em frequência de um filtro LPC de ordem 16
e resposta do filtro perceptual correspondente, com um valor de γ=0,8.

Existem fundamentalmente três tipos de codificadores LPAS, que se distinguem pela estratégia de geração da excitação: Os codificadores *multi-pulso*; por excitação regular de impulsos (RPE - *Regular Pulse Excitation*); e por predição linear com excitação por código (CELP - *Code-Excited Linear Prediction*).

7.3.1 Codificador *multi-pulso*

O codificador *multi-pulso* [Atal (82)] modela o resíduo de dupla predição à custa dos impulsos *mais importantes*. A escolha em bloco destes impulsos é bastante complexa, pelo que alternativamente estes podem ser escolhidos sequencialmente do modo seguinte: inicialmente, sem qualquer impulso, é subtraído ao sinal original a contribuição da memória dos preditores devida à excitação na trama anterior; seguidamente é colocado um impulso em todas as posições da trama e calculado o erro perceptual, sendo escolhido aquele que o minimiza. A amplitude do impulso escolhido é calculada e quantificada, sendo a sua contribuição subtraída ao sinal de entrada. A escolha de novos impulsos continua até se atingir um erro razoavelmente baixo ou até serem escolhidos um número fixo de impulsos.

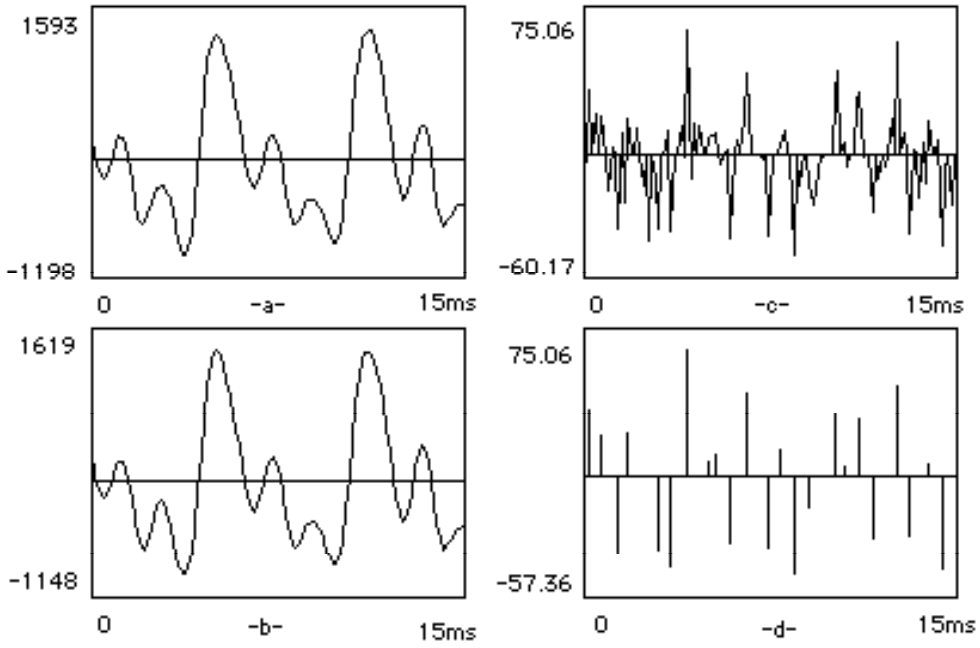

Figura 7.7
Formas de onda típicas do codificador *multi-pulso*.
a) Sinal original. b) Sinal sintetizado. c) Resíduo de dupla predição.
d) Modelação por impulsos (8 impulsos por cada 5 ms).

Dez impulsos por cada 10 ms são suficientes para produzir fala com *boa* qualidade, variando o débito binário total entre os 13 e os 16

kbit/s. As formas de onda típicas deste codificador são mostradas na figura 7.7. Impondo restrições nas posições dos impulsos (MP-MLQ - *Multipulse excitation with maximum likelihood quantizer*) foi aprovada em 1995 a recomendação G.723.1 [Cox (96)] do ITU-T, produzindo apenas 6,3 kbit/s. As aplicações desta recomendação são em videotelefone e transmissão de voz sobre IP (*Internet protocol*). Repare-se que este tipo de comunicações, quando se estabelecia (à data da sua aprovação) sobre a rede telefónica através de *modems*, normalmente não excedia débitos binários superiores a 14,4 kbit/s.

7.3.2 Codificador RPE

Uma variante à codificação multi-pulso é a codificação com excitação regular de impulsos [Deprettere (85)] [Kroon (86)], em que os impulsos, como mostrado na figura 7.8, estão sujeitos a grelhas pré-determinadas com as posições igualmente espaçadas.

↑...↑...↑...↑...↑...↑...↑...↑...↑...
.↑...↑...↑...↑...↑...↑...↑...↑...↑..
..↑...↑...↑...↑...↑...↑...↑...↑...↑.
...↑...↑...↑...↑...↑...↑...↑...↑...↑

Figura 7.8
Grelhas possíveis na codificação RPE (L=40, n=4)
Os ↑ representam a presença de impulsos e os ... a ausência de impulsos.

Encontra-se um exemplo deste tipo de codificação (RPE-LT) na norma da rede GSM de telefones celulares móveis a 13 kbit/s (GSM 06.10) [Vary (88)], adoptada por Portugal. Nesta, a dimensão da trama correspondente ao filtro LPC é de 20 ms, sendo as grelhas escolhidas em 4 subtramas de 5 ms (L=40), havendo 1 impulso de 4 em 4 amostras (n=4 grelhas possíveis, codificadas com 2 bits). As amplitudes dos respectivos impulsos são estimadas resolvendo um

conjunto de equações lineares. A tabela 7.3 mostra a distribuição dos bits para esta norma, sendo as formas de onda ilustradas na figura 7.9.

Parâmetros	bits por trama	bits por segundo
Atrasos do Preditor de Longa Duração	4x7=28	1400
Ganho do Preditor de Longa Duração	4x2=8	400
Codificação da grelha (RPE)	4x2=8	400
Amplitudes dos impulsos (RPE)	4x45=180	9000
Coeficientes LPC (LAR)	36	1800
Total	260	13000

Tabela 7.3
Distribuição dos bits no codificador RPE-LT, norma *full-rate* GSM.

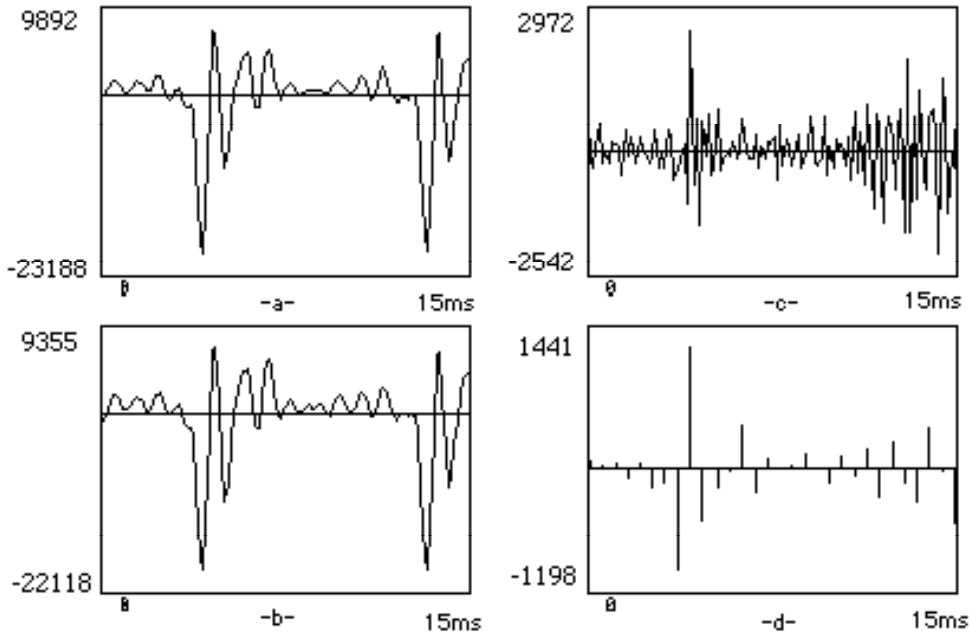

Figura 7.9
Formas de onda típicas do codificador RPE-LT.
a) Sinal original. b) Sinal sintetizado. c) Resíduo de dupla predição.
d) Modelação do resíduo pelo codificador RPE-LT.

7.3.3 Codificador CELP

A codificação por predição linear com excitação por código [Atal (84)] [Schroeder (85)], quantifica vectorialmente cada trama da excitação de dupla predição, conseguindo débitos binários bastante

reduzidos. Embora o codificador CELP pudesse provir de uma evolução do codificador *multi-pulso*, na sua origem estão estudos sobre distorção e considerações de mascaramento auditivo no âmbito do codificador APC. O princípio seguido é o de que o resíduo de dupla predição se aproxima de ruído branco, pelo que inicialmente foi designada por codificação estocástica. Realizações deste tipo de ruído são armazenadas num dicionário designado de estocástico, sendo escolhido aquele (palavra de código estocástica) que produz o menor erro. O diagrama de blocos deste codificador é apresentado na figura 7.10.

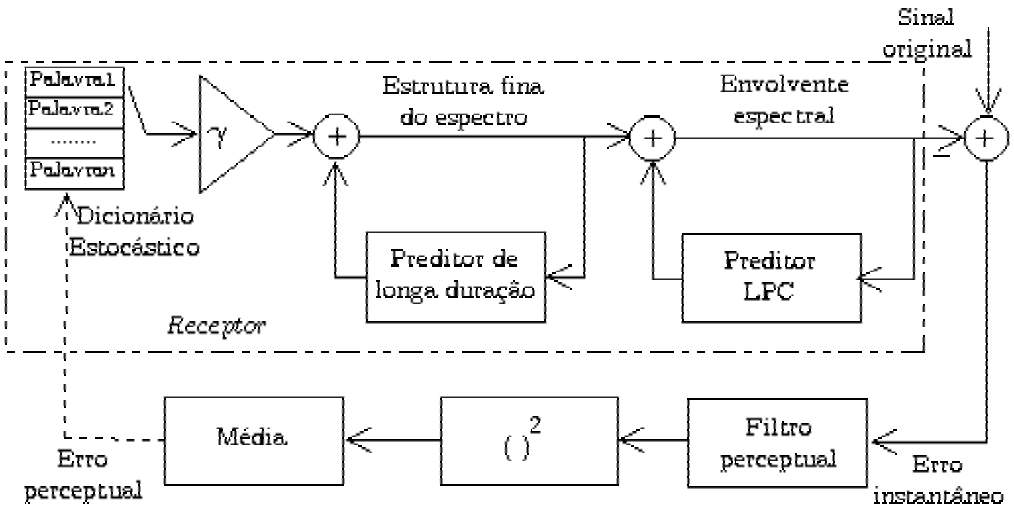

Figura 7.10
Diagrama de blocos de um codificador CELP (original).

O codificador proposto por Atal e Schroeder utilizava um dicionário com 1024 palavras de código de duração 5 ms (subtrama da trama do preditor linear com duração de 20 ms), sendo também quantificado um factor de escala para toda a palavra de código, correspondendo a um total de cerca de 3000 bit/s.

Originalmente o codificador CELP apenas utilizava análise-por-síntese para determinar a palavra de código estocástica, sendo os coeficientes do preditor de longa duração calculados em cadeia aberta. Em 1988, Kleijn *et al.* [Kleijn (88)] propuseram um método de análise-por-síntese válido também para a estimação dos parâmetros do

preditor de longa duração, aumentando a qualidade mas também a complexidade. Neste, é utilizado um segundo dicionário em que cada palavra de código é uma versão deslocada da excitação anterior do filtro LPC, existindo tantas palavras de código como períodos possíveis da frequência fundamental. Este dicionário é refeito por cada trama, pelo que se designa de dicionário adaptativo. A palavra de código adaptativa óptima e respectivo ganho são então calculados e a sua contribuição retirada ao sinal de entrada antes de se proceder à procura da palavra de código estocástica. Uma vez que não é necessária uma estimação explicita da frequência fundamental, outra vantagem deste método é a de não impor o atraso devido à janela de análise. O atraso de algoritmo é então devido apenas à janela de análise do filtro LPC. O diagrama de blocos deste codificador é apresentado na figura 7.11.

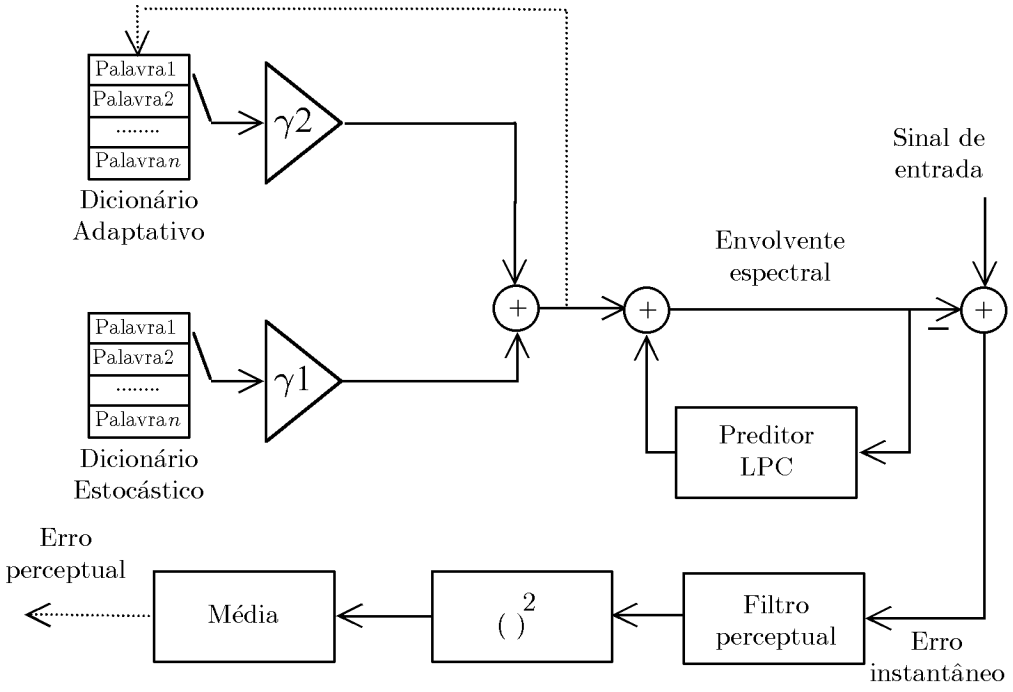

Figura 7.11
Diagrama de blocos de um codificador CELP com dicionário adaptativo.

Apesar do filtro perceptual, a qualidade subjectiva do sinal sintetizado pode ainda ser melhorada diminuindo o ruído entre

formantes resultante da quantificação, à custa do aumento do ruído na zona dos formantes que será por estas mascarado, através da introdução no receptor de um filtro de pós-processamento dado por:

$$H(z) = \frac{A(z/\beta)}{A(z/\alpha)}\left((1-\mu)-\mu z^{-1}\right), \qquad (7.2)$$

em que $0 \leq \beta < \alpha \leq 1$ e $0 < \mu < 1$, sendo μ o parâmetro responsável pela atenuação do declive espectral. Valores típicos [Ribeiro (91-a)] destes parâmetros são β=0,5, α=0,8 e μ=0,5. Note-se contudo que devido à degradação objectiva da qualidade, esta diminui quando o filtro é utilizado por codificadores em cascata.

A norma FS-1016 a 4600 bit/s do DoD (mais 200 bits/s para codificação de canal) [Campbell (90)] de 1991 é um exemplo de um codificador CELP, com tramas de 30 ms para cálculo dos coeficientes LPC e 4 subtramas de 7,5 ms para cálculo da informação dos dois livros de código. A distribuição dos bits por trama e por segundo deste codificador é apresentada na tabela 7.4.

Parâmetros	bit por trama	bit/s
Coeficientes LPC	34	1133,3
Dicionário adaptativo	4*7	933,3
Ganho adaptativo	4*5	666,3
Dicionário estocástico	4*9	1200
Ganho estocástico	4*5	666,3
Total	**138**	**4600**

Tabela 7.4
Distribuição dos bits por trama de 30 ms, norma FS-1016 a 4600 bit/s.

A relação entre a qualidade e o débito binário deste tipo de codificação é muito boa, pelo que a maioria das normas com débito binário superior a 3400 bit/s recentemente aprovadas pelos diversos organismos internacionais, são variantes do codificador CELP que diminuem a complexidade na procura nos livros de código. As formas de onda típicas deste codificador são apresentadas na figura 7.12.

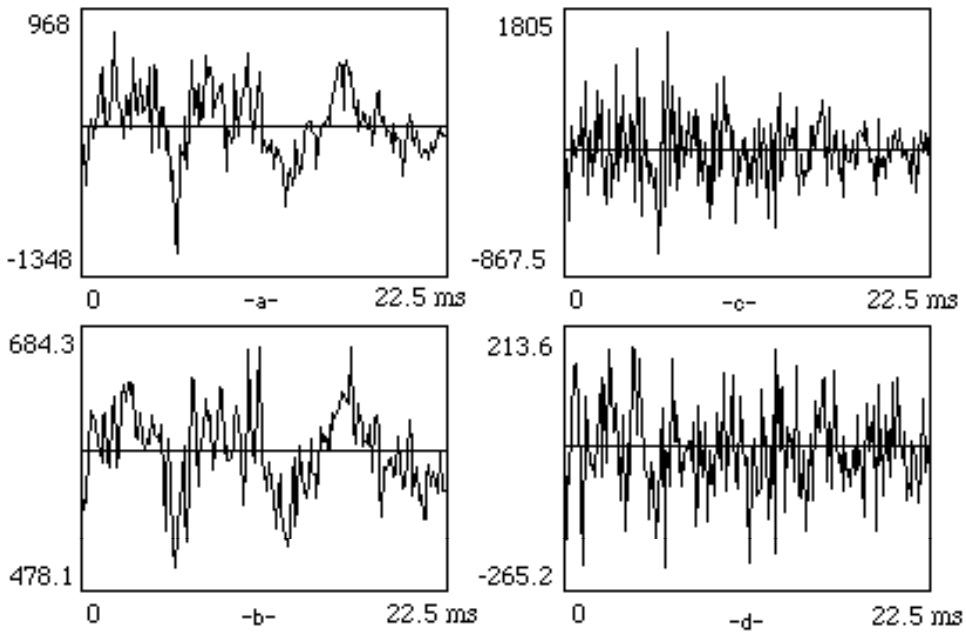

Figura 7.12
Formas de onda típicas do codificador CELP.
a) Sinal Original. b) Sinal sintetizado. c) Resíduo de dupla predição.
d) Modelação do resíduo pelo codificador CELP.

Uma das variantes do codificador CELP é o codificador VSELP (*Vector Sum Excited Linear Prediction*) [VSELP (89)], proposto pela Motorola e seleccionado como a norma IS-54 para uso nos Estados Unidos da América em sistemas telefónicos, com um débito de 7,95 kbit/s. A diferença principal reside na utilização não de um livro de código estocástico mas de dois livros de código de pequena dimensão (128 palavras de código), com as palavras de código constituídas pela combinação linear de 7 vectores de base, obtida pela decomposição do índice (± 1) da palavra de código correspondente. Os vectores são ortogonalizados em relação à excitação previamente seleccionada, melhorando a qualidade do sinal sintetizado. Outra vantagem desta estrutura é a robustez na presença de erros no canal, já que um erro na transmissão produz apenas uma pequena diferença na palavra de código gerada. Uma outra implementação do codificador VSELP é a norma GSM-HR (*Half-Rate*) de 1994 (GSM 06.06), com um débito binário de

5,6 kbit/s, que duplica o tráfego do GSM-FR mantendo a qualidade. Este codificador utiliza um livro de código de 9 bits nas zonas vozeadas e dois livros de código de 7 bits nas zonas não vozeadas.

7.3.4 Codificador LD-CELP

Uma das desvantagens dos codificadores que funcionam por tramas, em relação aos que codificam o sinal amostra-a-amostra, é o aumento do atraso devido à necessidade da leitura de uma trama completa antes da estimação dos parâmetros correspondentes. Com a utilização de livros de código adaptativos, os parâmetros que passaram a exigir uma dimensão maior da janela de análise são os coeficientes LPC, com janelas de duração típica de 20 ms ou mais. Como descrito no subcapítulo 5.3, o atraso total é de no mínimo 3 tramas de análise, pelo que o codificador terá no mínimo um atraso total de 60 ms. Embora o atraso não seja um problema em aplicações de armazenamento, é um atributo importante em aplicações de transmissão bidireccional em tempo real. No entanto, existe um compromisso que envolve o atraso, a qualidade e o débito binário. Para determinado codificador, a redução do atraso é conseguida diminuindo a janela de análise, mas esta diminuição envolve o aumento do débito binário, uma vez que aumenta a informação a emitir por unidade de tempo. Para manter o débito binário seria necessário diminuir o número de bits atribuídos a cada trama, o que por sua vez diminuiria a qualidade do sinal sintetizado. Por este motivo a redução do atraso em codificadores de baixo débito mantendo boa qualidade tem sido um desafio para os investigadores.

Em 1988 o ITU-T propôs normalizar um codificador com um débito binário de 16 kbit/s, com qualidade equivalente ao da norma G.726 a 32 kbit/s, ao mesmo tempo que deveria ter um atraso máximo

de apenas 2 ms. Em 1992 foi aprovada a recomendação G.728, um codificador CELP que gozava destas características, denominado por LD-CELP (*Low delay* CELP) [Chen (95)], cujo esquema de blocos é apresentado na figura 7.13. As diferenças principais em relação a um codificador CELP tradicional são a não utilização do preditor de longa duração e a adaptação do preditor LPC ser efectuada com base em amostras passadas do sinal de saída (BA-LPC) (*Backward Adaptation* LPC) e não com base no sinal original. Esta adaptação é representada no diagrama de blocos da figura 7.13 através das setas a tracejado. O único parâmetro a ser transmitido é o índice da trama que minimiza o erro perceptual. Para salvaguardar o atraso máximo de 2 ms a dimensão das palavras de código é de apenas 0,625 ms, a que correspondem 5 amostras, codificadas com 10 bits ou seja 2 bits por amostra, perfazendo um total de 16 kbit/s.

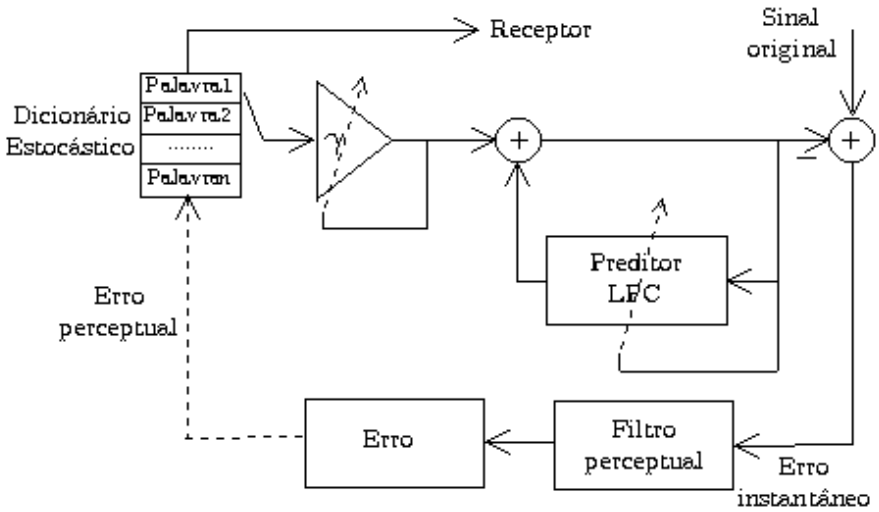

Figura 7.13
Diagrama de blocos do codificador LD-CELP.

A não utilização do preditor de longa duração deve-se à grande sensibilidade na presença de erros de canal quando da adaptação com base no sinal de saída. Segundo reportado por Chen, um erro de apenas 1 bit pode causar uma divergência permanente entre o emissor e o receptor. O período fundamental das vozes femininas corresponde

tipicamente a menos de 50 amostras, pelo que a ordem do preditor LPC foi aumentada para 50, o que cobre o último período fundamental e substitui o preditor de longa duração.

Todos os codificadores de pequeno atraso anteriores ao LD-CELP adaptavam o preditor com base num algoritmo de gradiente, como utilizado no ADPCM G.726. O problema deste tipo de adaptação é que não garante a estabilidade do filtro resultante sem um aumento significativo da complexidade, pelo que este era limitado a um máximo de dois pólos. A vantagem da utilização de BA-LPC é que a estabilidade é facilmente preservada (ver secção 3.1.8). Uma vez que um filtro de maior ordem pode ser utilizado, a utilização de BA-LPC alcança um melhor desempenho que os algoritmos de gradiente. Uma descrição pormenorizada deste codificador, incluindo uma versão a 8 kbit/s, pode ser encontrada no capítulo 6 [Chen (95)] de [Kleijn (95)].

7.3.5 Codificador CS-ACELP

Uma variante do codificador CELP é o modelo CS-ACELP (*Conjugate Structure Algebraic* CELP), proposto por Laflame *et al.* [Laflame (91)] e que utiliza um livro de código algébrico e não estocástico. As palavras de código são esparsas, com os impulsos não nulos tomando os valores de +1 ou -1, o que torna o procedimento de procura muito rápido. As posições dos impulsos são determinadas pelo índice da palavra de código, permitindo a eliminação total do seu armazenamento. Esta técnica aumenta também a robustez na presença de erros de canal, já que um erro de um bit não faz gerar uma palavra de código muito diferente da original. Devido à diminuição da complexidade e da memória de armazenamento esta variante tornou-se extremamente popular, resultando em várias normas seguidamente apresentadas, todas posteriores ao ano de 1994.

7.3.5.1 Recomendações G.729, G.729A e G.723.1

O ITU-T iniciou em 1990 os trabalhos que conduziriam em 1995 à recomendação G.729, um codificador com um débito binário de 8 kbit/s e uma qualidade equivalente à da recomendação G.726 a 32 kbit/s, ao mesmo tempo que deveria ter pequeno atraso e baixa complexidade. O primeiro objectivo deste codificador era a sua utilização em telefones públicos via rádio.

O codificador escolhido [Salami (98)][Cox (96)(97)] opera tramas de 10 ms, transmitindo a informação dos livros de código adaptativo e algébrico em subtramas de 5 ms. Em cada trama é estimado um filtro preditor de ordem 10, que necessita de 5 ms de informação posterior, resultando num atraso de algoritmo de 15 ms. Para transmitir a informação do filtro LPC os seus coeficientes são transformados em coeficientes LSF, que são por sua vez quantificados vectorialmente com 18 bits por trama.

Para cada subtrama é inicialmente calculado em cadeia aberta o período fundamental. Em torno deste valor (± 3) é posteriormente determinada em cadeia fechada a palavra de código adaptativa óptima. Este procedimento é um bom compromisso entre a qualidade e a complexidade. É utilizado um atraso fraccionário com uma resolução de 1/3 do período de uma amostra no intervalo [19□:84□] e com resolução inteira no intervalo [85:143]. Esta escolha de resolução permite uma codificação com 8 bits, sendo um bom compromisso entre a qualidade e o débito binário. A palavra de código adaptativa da segunda subtrama é codificada diferencialmente com 5 bits.

As 40 amostras correspondentes à palavra de código algébrica estão divididas em 4 subconjuntos, 3 dos quais com 8 amostras e um com 16. É permitida uma amostra não nula em cada um destes

subconjuntos, que poderá ter um sinal positivo ou negativo. As posições das amostras nestes subconjuntos são apresentadas na tabela 7.5.

impulso	sinal	Posições	bits
1	±1	0; 5; 10; 15; 20; 25; 30; 35	3+1
2	±1	1; 6; 11; 16; 21; 26; 31; 36	3+1
3	±1	2; 7; 12; 17; 22; 27; 32; 37	3+1
4	±1	3; 8; 13; 18; 23; 28; 33; 38 4; 9; 14; 19; 24; 29; 34; 39	4+1

Tabela 7.5
Posições dos impulsos dos subconjuntos da palavra de código algébrica na recomendação G.729.

O número total de bits por cada subtrama é de 17, correspondendo a um total de 34 bits por trama. A procura no livro de código é produzida em quatro ciclos, cada um abrangendo mais um impulso. O número total de combinações de impulsos corresponde a 2^{13}=8192 (17 menos 4 bits de sinal). Para reduzir a complexidade, a procura é focada nas combinações que com maior probabilidade possam coincidir com a óptima. Dado que cada vez que se entra no último ciclo são testadas 16 possibilidades, estes só são realizados se a contribuição dos três primeiros impulsos for já razoável. O número máximo de entradas neste último ciclo é limitado a 90, a que correspondem 90x16=1440 posições testadas, 18% do valor total, sendo a qualidade praticamente equivalente. Repare-se ainda que a palavra de código só fica completamente determinada com um total de 17 bits, a que corresponderiam 2^{17}=128k palavras de código, o que aumentaria enormemente a complexidade e memória necessária para o armazenamento, caso se tratasse de um livro de código estocástico.

Os ganhos das palavras de código adaptativa e algébrica são quantificados vectorialmente com 7 bits, correspondendo a um total de 14 bits por trama. São produzidas 100 tramas por segundo, cada uma com 80 bits, de modo a perfazer um débito binário de 8 kbit/s. A distribuição de bits por trama é apresentado na tabela 7.6.

Parâmetros	bits por trama	bits por segundo
Atrasos do pred. long. duração	8; 5	1300
Paridade 1 atraso pred. long. dur.	1	100
Palavra de código algébrica	17; 17	3400
Ganhos adaptativo e algébrico	7; 7	1400
Coeficientes LPC (LSF)	18	1800
Total	70	8000

Tabela 7.6
Distribuição dos bits na recomendação G.729.

Como Anexo A desta recomendação, é proposta uma variante com metade da complexidade total, em que a procura no livro de código algébrico é limitada apenas a 320 possibilidades (20 entradas no último ciclo), 4% do valor total. Esta variante é denominada de recomendação G.729A [Salami (97)], tendo compatibilidade bit-a-bit entre a informação transmitida em relação à recomendação G.729.

Para tornar mais flexível a recomendação G.723.1 (ver secção 7.3.1) dedicada à transmissão de videotelefone e voz sobre IP, quando a comunicação se estabelece a um débito binário baixo, foi aprovada uma variante agora a 5,3 kbit/s, utilizando o modelo CS-ACELP. A principal diferença entre esta recomendação e a G.729 é o aumento da trama para 30 ms, sendo esta dividida em 4 subtramas de 7,5 ms. Uma comparação entre estes três codificadores (G.723.1, G.729 e G.729A) pode ser encontrada em [Cox (96)(97)].

7.3.5.2 Normas GSM-EFR e GSM MR-ACELP

A norma GSM-EFR (EFR - *Enhanced Full-Rate*) a 12,2 kbit/s, normalizada nos Estados Unidos em 1995 e na Europa em 1996 (GSM 06.60), é outro exemplo bem sucedido da utilização do codificador ACELP. A trama corresponde a 20 ms e é dividida em 4 subtramas de 5 ms. Também com base no modelo ACELP é aprovada em 1998 uma norma para codificação AMR [Ekudden (99)], sendo o débito binário

continuamente adaptado às condições do canal de rádio. Para canais de baixo ruído é utilizado um débito binário de 12,2 kbit/s correspondentes ao GSM-EFR. Para canais ruidosos aumenta-se o número de bits de codificação do canal, à custa de uma diminuição do débito binário de codificação da fonte. Esta norma, denominada de MR-ACELP (*Multi-Rate* ACELP), tem 8 modos de funcionamento a que correspondem 8 débitos binários possíveis entre 4,75 kbit/s e 12,2 kbit/s, distribuídos como mostra a tabela 7.7.

Parâmetros	bit/s	bit/s	bit/s	bit/s	bit/s	bit/s	bit/s	bit/s
Coeficientes LPC	1150	1150	1300	1300	1300	1350	1300	1900
Dicionário adaptativo	1000	1000	1200	1200	1300	1400	1300	1500
Dicionário algébrico	1800	1800	2200	2800	3400	3400	6200	7000
Ganhos	800	1200	1200	1400	1400	1800	1400	1800
Total	**4750**	**5150**	**5900**	**6700**	**7400**	**7950**	**10200**	**12200**

Tabela 7.7
Distribuição dos bits por trama de 20 ms, norma GSM MR-ACELP.

Esta norma é um exemplo de como a redução do débito binário devido a uma codificação mais eficiente pode ser útil na minimização do impacto de erros de canal, caso os bits poupados sejam utilizados para correcção de erros. Assim, pode-se obter pior qualidade na presença de erros de canal com um codificador de alto débito, do que a obtida com um codificador de menor débito mas qualidade razoável que utilize códigos de correcção, mesmo para débitos binários totais mais baixos[2].

7.4 Codificação sinusoidal

Os codificadores ATC [Zelinsky (77)] que processem por exemplo uma trama de 32 ms (256 amostras), modelarão o espectro através de 128 coeficientes complexos, simplesmente quantificando com mais níveis os coeficientes correspondentes às zonas de maior amplitude. Nas zonas vozeadas, estas zonas coincidem ou estão muito próximas das

[2] Demonstração na página da disciplina (demos/canal/canal.html)

harmónicas do sinal. Tirando partido da estrutura do sinal de fala, as amplitudes e fases destas harmónicas são realmente os únicos parâmetros a serem transmitidos, dando origem aos codificadores sinusoidais, que são também codificadores híbridos. O número de harmónicas, considerando a banda telefónica (4 kHz), é de 10 a 80, dependendo da frequência fundamental (50 Hz-400 Hz), dando ideia da baixa eficiência do codificador ATC na modelação das zonas vozeadas.

Os primeiros codificadores sinusoidais foram propostos por Hedelin em 1981 [Hedelin (81)] e por Almeida e Tribolet em 1982 [Almeida (82)], utilizando o modelo de síntese descrito na secção 4.3. Os primeiros codificadores sintetizavam o sinal no domínio da frequência, utilizando o mesmo tipo de transformação por blocos (com *overlap-add*) do ATC. Este modelo assume que o sinal numa trama é estacionário, restringindo as amplitudes a serem constantes. De forma a manter a continuidade do sinal nas fronteiras das tramas passou-se a sintetizar o sinal no domínio do tempo, o que facilita a variação das frequências e das amplitudes (equação 4.5):

$$s'(t) = \sum_{k=1}^{K(t)} A_k(t) \sin\left(\varphi_k(t)\right), \tag{7.3}$$

em que $A_k(t)$ e $\varphi_k(t)$ representam, respectivamente, as amplitudes e fases da k-ésima sinusóide, cuja frequência é dada pela derivada da fase $\varphi_k(t)$, sendo $K(t)$ o número de sinusóides na banda considerada.

Restringindo as frequências a serem harmónicas da frequência fundamental (figura 7.14), o codificador sinusoidal degenera num codificador harmónico. O codificador harmónico descrito por Almeida e Tribolet [Almeida (84)] reproduzia as zonas puramente não vozeadas com o modelo ATC e as zonas puramente vozeadas com o modelo harmónico, controlando dinamicamente a distribuição de bits entre os

dois modelos nas zonas de vozeamento misto, correspondendo a componente não vozeada ao resíduo do modelo harmónico. No entanto a versão do codificador implementada, com um débito binário de 4800 bit/s, empregava uma decisão binária de vozeamento para comutar entre os dois modelos. Em 1989, Marques e Almeida propõem manter a restrição harmónica nas zonas não vozeadas [Marques (89) (90)], utilizando um número suficiente de sinusóides de modo a cobrir razoavelmente todo o espectro. Em 1984, McAulay e Quatieri [McAulay (84)] propõem um modelo sinusoidal válido para as zonas não vozeadas, não impondo a restrição harmónica mas mantendo apenas a continuidade da fase.

Os parâmetros transmitidos pelos codificadores sinusoidais podem variar para diferentes débitos binários. Para débitos binários elevados, transmitem-se as frequências, amplitudes e fases no inicio das tramas, conseguindo-se uma *boa* qualidade. Para débitos de 8 kbit/s, podem-se transmitir as frequências e respectivas amplitudes, sendo a fase definida como o integral da frequência e mantendo apenas a sua continuidade [McAulay (84)]. Para um débito binário de 4800 bit/s, Almeida e Tribolet [Almeida (83)] impõem que as frequências sejam harmónicas da frequência fundamental e codificam as amplitudes com base no modelo de predição linear, por amostragem da envolvente espectral, mas transmitem a informação da fase. Para débitos binários perto dos 2400 bit/s é ainda perdida a relação de fases, mantendo-se apenas a sua continuidade. Nesta situação, os parâmetros a transmitir são idênticos aos do *vocoder* LPC, pelo que os dois modelos são inter-operáveis[3] [McAulay (90)]. Os codificadores sinusoidais cobrem então todo o leque de débito binário que vai desde os modelos paramétricos de codificação de fonte à codificação de forma de onda, ou

[3] Demonstração na página da disciplina (/demos/vocharm/vocharm.html)

seja, tal como os modelos LPAS cobrem a lacuna de codificação existente até ao final da década de 70.

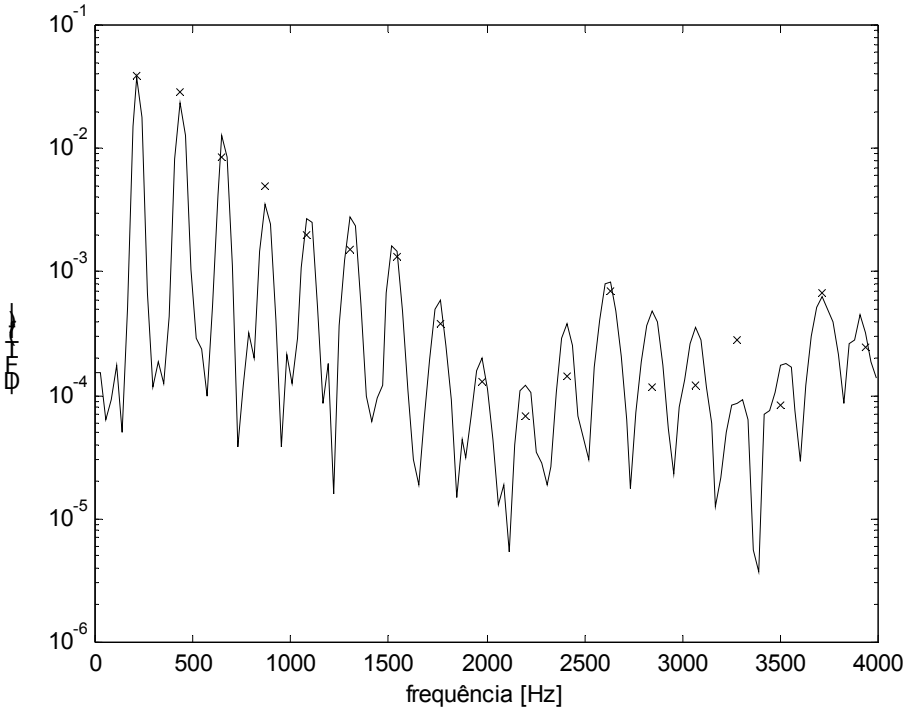

Figura 7.14
Módulo da transformada de Fourier de uma trama de 32 ms (256 amostras) de uma zona vozeada (janela de Hamming). Os X representam o módulo da transformada de Fourier do sinal, estimada através de uma janela de dimensão de 4,625 ms igual ao periodo fundamental (37 amostras).

Em 1988, é proposto por Griffin e Lim uma variante do codificador harmónico [Griffin (88)], designado por *vocoder com excitação multibanda* (*multiband excitation vocoder*), produzindo um débito binário de 8 kbit/s. A principal diferença em relação à codificação harmónica *clássica* é a transmissão de uma decisão de vozeamento por cada harmónica, sendo transmitidas as fases das harmónicas consideradas vozeadas. A síntese é produzida no domínio do tempo para as harmónicas vozeadas, permitindo uma melhor interpolação entre tramas, e por transformada para as bandas não vozeadas. As amplitudes das harmónicas não são representadas através

dos coeficientes LPC, mas através de um codificador diferencial adaptativo ao longo das harmónicas.

Em 1990, uma versão do codificador harmónico a funcionar a um débito binário de 4,15 kbit/s foi adoptada como a norma Inmarsat-M pela Corporação de Satélites Marítimos Internacional (Inmarsat), para comunicações via satélite de navios para terra. A norma MPEG-4 (*Moving Picture Expert Group*) emprega um modelo misto de codificação, utilizando nas zonas vozeadas um codificador sinusoidal e nas zonas não vozeadas um codificador CELP. Este codificador, denominado de HVXC (*Harmonic Vector eXcitation Coding*), funciona a débitos binários escaláveis de 2000 e 4000 bit/s e ainda com um débito binário variável entre os 1200 e os 1700 bit/s. É também possível variar a frequência fundamental e a velocidade de audição, o que facilita a procura de trechos do sinal de fala.

7.5 Codificação MELP

Tradicionalmente os *vocoders* utilizam dois tipos de excitação: pulsos com período igual ao da frequência fundamental nas zonas vozeadas ou ruído branco nas zonas não vozeadas. É transmitida muito pouca informação sobre o sinal de excitação do filtro de predição linear (resíduo de predição), conseguindo-se débitos binários bastante baixos mas diminuindo a qualidade final do sinal sintetizado. Uma das causas desta diminuição na qualidade é a utilização de um critério binário na decisão de vozeamento, pois além da possibilidade de erros na decisão, nas zonas vozeadas persiste sempre uma variação nas formas de onda entre períodos fundamentais devido à turbulência do ar proveniente dos pulmões, que ocorre especialmente durante a fase aberta da glote. A utilização de pulsos invariantes na excitação dos *vocoders* faz com que o sinal sintetizado adquira uma qualidade tonal, perdendo naturalidade.

Para tentar solucionar este problema, Kwon e Goldberg [Kwon (84)] propõem a supressão da decisão de vozeamento e a utilização de excitação mista, com uma relação de amplitudes entre pulsos periódicos e ruído estimada com base no resíduo óptimo, sendo esta transmitida para o receptor. Kang e Fransen [Kang (85-a)] apresentam um conjunto de experiências com vista à melhoria da excitação sem transmissão adicional de informação, no qual se encontra a utilização de excitação mista. O *vocoder* com excitação multibanda [Griffin (88)] transmite a informação de uma decisão de vozeamento por harmónica. No contexto da codificação harmónica, Marques e Abrantes [Marques (94)] propõem a utilização de codificadores harmónicos híbridos, em que a excitação é composta por misturas de sinusóides e de sinais aleatórios de banda estreita. McAulay e Quatieri [McAulay (95)] propõem a transmissão de uma probabilidade de vozeamento, sendo as harmónicas abaixo de uma frequência proporcional à probabilidade de vozeamento sintetizadas como vozeadas e acima dessa frequência como não vozeadas.

Em 1991 foi proposto por McCree e Barnwell um *vocoder* LPC empregando excitação mista [McCree (91)], em que a componente vozeada é filtrada passa-baixo e a componente não vozeada é filtrada passa-alto, sendo a frequência de corte determinada automaticamente no emissor. Esta solução não evita o aparecimento de distorções tonais de curta duração nas baixas frequências. A introdução de variações aleatórias ($\pm$25%) no período fundamental, controladas por um terceiro estado na decisão de vozeamento, destrói a distorção tonal e modela melhor as variações erráticas dos períodos fundamentais, muitas vezes encontradas durante as transições de vozeamento. Uma das vantagens deste terceiro estado é uma maior facilidade na decisão de vozeamento, já que as zonas de menor periodicidade em que alguma dúvida sobre a

decisão poderia surgir, são modeladas com esta variação errática. Em 1992, McCree e Barnwell adicionam a este *vocoder* a capacidade de decisão de vozeamento por bandas [McCree (92)]. De forma a manter o débito binário baixo, a decisão de vozeamento é efectuada em apenas 5 bandas, sendo esta informação codificada com 4 bits, para além da decisão de vozeamento global.

O sinal sintetizado através do filtro de predição linear tem normalmente formantes com maior largura de banda que o sinal original, resultando numa atenuação mais rápida entre períodos fundamentais [McCree (95)]. O resíduo apresenta, por esse facto, formantes muito próximas dos formantes do sinal de fala original, sendo inteligível. Kang e Fransen [Kang (85-a)] exploram este facto para modelar a excitação vozeada através de um filtro idêntico ao de síntese por predição linear, mas em que cada coeficiente é multiplicado por um factor de escala inferior à unidade, reduzindo o ganho nos formantes. McCree e Barnwell [McCree (93)] utilizam um filtro de pós-processamento de envolvente, equivalente aos já utilizados nos codificadores CELP (equação 7.2), para reduzir o ruído entre os formantes. O objectivo dos autores não é a de reduzir este ruído, mas sim o de realçar os formantes mais fracos no sinal sintetizado em relação aos do sinal de fala original. Este codificador, cujo esquema de blocos do receptor é apresentado na figura 7.15, foi denominado de codificador por predição linear com excitação mista (MELP - *Mixed Excitation Linear Prediction*).

Para modelar melhor o resíduo de predição nas zonas vozeadas, pode-se representá-lo através da sua expansão em série de Fourier, calculado em frequências múltiplas da frequência fundamental. A síntese da excitação é produzida através de uma DFT inversa, de dimensão igual à do período fundamental e utilizando valores

interpolados entre tramas consecutivas. Para não aumentar em demasia o débito binário, McCree *et al.* [McCree (96)] propõem que se quantifiquem vectorialmente as amplitudes das 10 primeiras harmónicas com 8 bits, sendo as amplitudes das restantes mantidas constantes e todas as fases colocadas a zero.

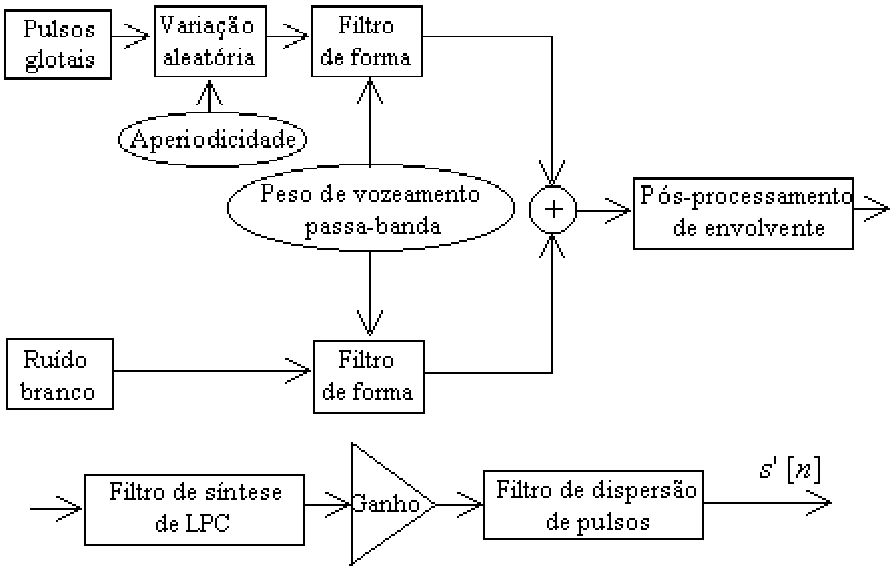

Figura 7.15
Síntese LPC com excitação mista (MELP).

Em 1997 o Departamento de Defesa dos Estados Unidos da América normalizou o codificador MELP incluindo todas as características anteriormente descritas e produzindo um débito binário de 2400 bit/s [Supplee (97)], para substituir as normas FS-1015 e FS-1016. Este codificador utiliza quantificação vectorial dos coeficientes de predição linear para eliminar a informação redundante e com a redução no débito binário conseguida em relação à norma FS-1015, introduz as melhorias atrás descritas na definição da excitação, obtendo a qualidade equivalente ou mesmo superior à da norma FS-1016. A distribuição de bits para cada um dos parâmetros deste codificador é apresentada na tabela 7.8.

Parâmetros	Zonas Vozeadas	Zonas não Vozeadas
Coeficientes LPC	25	25
Ganho	8	8
V/UV & Frequência Fundamental	7	7
Vozeamento passa-banda	4	0
Aperiodicidade	1	0
Resíduo de predição	8	0
Protecção contra erros	0	13
Sincronismo	1	1
Total	**54**	**54**

Tabela 7.8
Distribuição dos bits no codificador MELP, nova norma do DoD a 2400 bit/s, por trama de 22,5 ms.

7.6 Codificação WI

As zonas vozeadas dos sinais de fala, quase periódicas, podem ser interpretadas como uma sucessão de ciclos correspondentes a períodos fundamentais, que mesmo para ciclos bastante separados no tempo variam ligeiramente entre si. Dada esta grande correlação é possível extrair esparsamente ciclos de forma de onda e obter os valores intermédios através de interpolação. Esta é a principal motivação da codificação por interpolação de forma de onda (WI - *Waveform Interpolation*), proposta por Kleijn [Kleijn (91)] em 1991.

Se para um instante de tempo for conhecido o respectivo ciclo correspondente ao períodos fundamental, denominado de *forma de onda característica* e se for conhecida a fase correspondente neste ciclo, o sinal é reconstruído sem distorção. O sinal de fala é então descrito por uma função a duas dimensões, $u(t,\phi)$, em que as formas de onda características são representadas ao longo do eixo da fase ϕ, normalizadas em duração e periódicas em 2π. A evolução temporal das formas de onda características é representada ao longo do eixo do tempo t, pelo que a função $u(t,\phi)$ é denominada de *envolvente da forma de onda característica*. Um exemplo desta função é apresentada na figura 7.16.

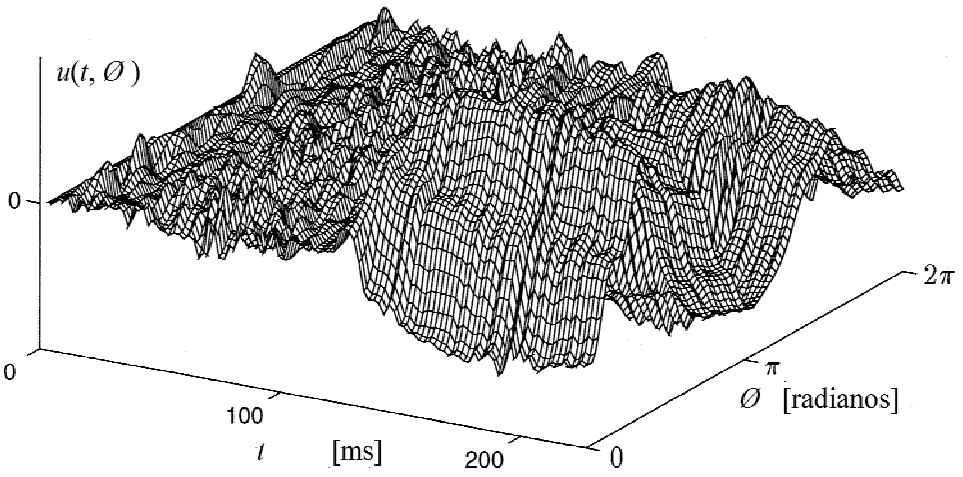

Figura 7.16
Envolvente da forma de onda característica
correspondente à palavra "*cheat*" (adaptado de [Kleijn (95)]).

A forma de onda característica é estimada através de uma janela adaptativa com a duração do período fundamental. É normalmente utilizado o resíduo de predição linear e não directamente o sinal de fala. A razão principal da utilização do resíduo de predição é a extensão periódica da forma de onda característica ao longo do eixo da fase. Esta extensão pode levar a descontinuidades perceptualmente audíveis na zona das fronteiras. Para as minimizar as fronteiras devem conter o mínimo de energia possível e, uma vez que o resíduo de predição apresenta pulsos bem definidos e uma região de baixa energia entre estes, a estimação da forma de onda característica é facilitada neste domínio. Para além da continuidade no eixo das fases, deverá também ser imposta a máxima continuidade ao longo do eixo dos tempos. Esta condição pode ser observada variando a fase da forma de onda característica, de modo a que se maximize a correlação com a forma de onda característica estimada no instante anterior.

Uma das maneiras de representar de um modo eficiente a forma de onda característica é através da série de Fourier. Repare-se que,

coincidindo a dimensão da janela com o período fundamental, as frequências da série correspondem às $K(t)$ harmónicas do sinal. A envolvente da forma de onda característica será então descrita por:

$$u(t,\phi) = \sum_{k=1}^{K(t)} \alpha_k(t)\sin(k\phi) + \beta_k(t)\cos(k\phi). \qquad (7.4)$$

O emissor deverá calcular, normalmente a intervalos regulares, a informação da forma de onda característica, a frequência fundamental, a energia e os coeficientes de predição linear. No receptor, a envolvente da forma de onda característica é reconstruída amostra-a-amostra utilizando interpolação. Para uma reconstrução perfeita, a forma de onda característica deverá ser amostrada pelo menos uma vez por período fundamental, o que implica que a largura de banda da evolução é de, no máximo, $f_p(t)/2$, denotando $f_p(t)$ a frequência fundamental em função do tempo.

O resíduo de predição é obtido convertendo $u(t,\phi)$ num sinal a uma dimensão, sendo a fase $\phi(t)$ especificada em cada amostra do sinal, como o integral da frequência,

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} 2\pi f_p(t)dt. \qquad (7.5)$$

Após reconstruir o resíduo de predição o sinal é sintetizado através do filtro de predição linear. Alternativamente, os coeficientes de Fourier da forma de onda característica correspondentes ao sinal de saída podem ser calculados directamente a partir de $\alpha(t)$, $\beta(t)$ e dos coeficientes de predição linear [Keijn (95)].

A transformada de Fourier da função $u(t,\phi)$, aplicada ao longo do tempo e para uma dada fase ϕ, mostra a evolução da forma de onda característica. Para zonas vozeadas em que esta evolução é lenta

(*SEW - Slowly Evolving Waveform*) a maior parte da energia está abaixo de 20 Hz [Keijn (95)]. Nas zonas não vozeadas a evolução da forma de onda característica é rápida (*REW - Rapidly Evolving Waveform*), correspondendo às frequências mais elevadas de evolução. Isto quer dizer que é possível separar por filtragem ao longo do tempo a componente vozeada (*SEW*) e a componente não vozeada (REW).

Antes de 1994 os codificadores WI utilizavam codificação CELP para representar as zonas não vozeadas. A separação da forma de onda característica nas componentes *SEW* e *REW* permitiu que a codificação WI fosse estendida às zonas não vozeadas [Kleijn (95-a)]. Assim, consegue-se manter uma boa qualidade nas zonas vozeadas mesmo amostrando a *SEW* a um ritmo baixo (*e.g.*, 40 Hz), sendo os valores intermédios obtidos por interpolação. A *REW*, contudo, deverá ser amostrada com maior ritmo, mas permite uma menor precisão na sua representação, uma vez que nas zonas não vozeadas a envolvente espectral transporta a maior parte da informação perceptual. O débito binário total situa-se numa gama entre os 2400 e os 4800 bit/s.

Um codificador WI foi submetido à nova norma do DoD a 2400 bit/s [Kleijn (95)]. Embora o codificador seleccionado tenha sido um codificador MELP [Supplee (97)], a qualidade do codificador WI proposto foi julgada uma das melhores, tendo sido mesmo o codificador com melhor desempenho em termos da reconhecibilidade do orador. De notar que a este débito binário o espectro da componente *SEW* é representado apenas pela sua amplitude, através de um quantificador vectorial de 7 bits, sendo esta informação transmitida em intervalos regulares de 25 ms. O espectro da *REW* é também representado apenas pela sua amplitude, a intervalos regulares de 6,25 ms. A componente *REW* é quantificada vectorialmente em tramas alternadas e codificada com 3 bits, sendo os restantes valores transmitidos com um código de 1

bit, de modo a repetir a informação imediatamente anterior ou posterior. A distribuição de bits por trama é apresentada na tabela 7.9.

Parâmetros	bit/trama
Coeficientes LPC	30
RMS	2x4
V/UV & Frequência Fundamental	7
SEW	7
REW	2x4
Total	**60**

Tabela 7.9
Distribuição dos bits no codificador WI, por trama de 25 ms, proposto (não seleccionado) para a nova norma do DoD a 2400 bit/s.

7.7 Codificação a muito baixo débito binário

Nos *vocoders* LPC, tipicamente com um débito binário de 2400 bit/s, a quantificação da envolvente espectral é a que consome a maior parte do débito binário. Para codificação a débitos binários mais baixos é utilizado como síntese *vocoders* (*e.g.* LPC, formantes), mas em que são colocados grandes esforços na quantificação mais eficiente da informação sobre a envolvente espectral. Na descrição destes métodos começaremos pela quantificação vectorial, que tira partido da correlação intra-trama e descreveremos seguidamente as metodologias que tiram partido da correlação inter-tramas, provocando um aumento do atraso: a interpolação entre tramas; a quantificação de super-tramas; a quantificação matricial; e a quantificação segmental, na qual se inclui a codificação fonética. O aumento do atraso conjuntamente com a *fraca* qualidade que estes codificadores auferem, tornam-nos de difícil aceitação quando utilizados em comunicação bidireccional, tendo no entanto algumas aplicações a nível do armazenamento. As fronteiras de definição das diversas metodologias não estão claramente definidas, resultando esta divisão apenas de uma tentativa de sistematização.

7.7.1 Quantificação vectorial

Na norma FS-1015, os 41 bits que são utilizados por trama para representar a envolvente espectral modelam 2^{41} configurações diferentes do tracto vocal. Wong *et al.* [Wong (81)(82)] propõem um quantificador vectorial com apenas 1024 vectores de código, com livros de código separados para representar as tramas vozeadas e as não vozeadas e utilizando procura em árvore binária para diminuir a carga computacional. Alterando unicamente a quantificação da norma FS-1015 obtêm um codificador independente do orador com um débito binário de 800 bit/s.

Kang e Fransen [Kang (85)] empregam também a metodologia de análise e síntese da norma FS-1015 e um quantificador vectorial de 12 bits. A medida de distância utilizada na procura do melhor vector de código tem em consideração a posição relativa dos coeficientes LSF, adaptando os coeficientes w_k da equação 6.21, já que esta determina a posição dos formantes, perceptualmente mais importantes que as zonas entre formantes. A energia é transmitida com 5 bits por trama e a frequência fundamental com 5 bits, uma vez de 3 em 3 tramas. O débito binário total é de 800 bit/s.

7.7.2 Interpolação entre tramas

Os formantes do tracto vocal e, consequentemente, a envolvente espectral, têm uma variação lenta de trama para trama. Esta correlação inter-tramas pode ser explorada para transmitir a informação apenas num subconjunto das tramas, sendo a informação das outras tramas obtida por interpolação. Seguindo este princípio, McAulay [McAulay (81)] propõe um *vocoder* de canal em que o sinal é dividido em 50 tramas por segundo, sendo a informação da envolvente

espectral transmitida em tramas alternadas. A informação nas tramas intermédias é obtida através de um algoritmo denominado de preenchimento da trama (*frame-fill*), utilizando para tal uma de 4 hipóteses diferentes, codificadas com 2 bits: (1) repetição da trama anterior; (2) repetição da trama seguinte; (3) interpolação linear com pesos idênticos das 2 tramas adjacentes; (4) interpolação linear atribuindo 33,3% do peso à trama anterior. O débito binário obtido é de 800 bit/s. Cada canal tem uma largura de banda fixa de 120 Hz, sendo os ganhos obtidos por amostragem da envolvente espectral. Nas zonas vozeadas a envolvente espectral é modelada através de 3 formantes, enquanto que nas zonas não vozeadas é utilizado um preditor LPC de ordem 6.

A segmentação do sinal de fala em trajectórias lineares de formantes é utilizada por Zolfaghari e Robinson para codificar fala a um débito binário de 500 bit/s [Zolfaghari (97-a)(97-b)], sendo a síntese final efectuada através de um modelo sinusoidal. A segmentação é obtida utilizando programação dinâmica. Holmes [Holmes (98)] utiliza também trajectórias lineares de formantes, mas obtendo a segmentação através de técnicas de reconhecimento [Rabiner (89) (93)]. Na síntese é utilizado um sintetizador paralelo de formantes, resultando um débito variável entre os 600 e os 1000 bit/s.

7.7.3 Quantificação de super-tramas

A empresa Motorola desenvolveu um codificador que utiliza a análise e síntese da norma FS-1015, operando a 600 bit/s [Fette (91)]. Este agrupa 4 tramas naquilo que designam de *super-trama*, estando disponíveis 35 bits para codificar o espectro das 4 tramas. É escolhido 1 de 8 quantificadores possíveis, que produzem desde a quantificação

escalar da trama central para sons estacionários, até à quantificação vectorial e diferencial nas zonas de variação rápida do espectro.

A norma NATO STANAG 4479 baseia-se em princípios muito semelhantes, para reduzir o débito binário da norma FS-1015 para 800 bit/s [Mouy (92)(95)]. Este codificador transmite a informação conjunta de 3 tramas, baseado nos seguintes princípios: numa zona estacionária, o valor médio do espectro deve ser bem quantificado e numa zona de transição deve ser dado ênfase às variações e não tanto à definição exacta do espectro. Para atingir estes objectivos, o codificador testa nas 3 tramas de entrada 8 esquemas diferentes de quantificação, escolhendo aquele que minimiza determinado critério de distância espectral. São utilizados 3 bits para codificar qual o esquema escolhido e 32 bits para codificar o espectro das 3 tramas. Por exemplo, para um som estável a meio de um segmento fonético, os 32 bits são utilizados para codificar o valor médio do espectro. Se as 2 primeiras tramas apresentarem um espectro semelhante e houver uma variação acentuada na terceira trama, são utilizados 24 bits para codificar a informação média das 2 primeiras tramas e 8 bits para codificar diferencialmente a terceira trama, etc.

7.7.4 Quantificação matricial

A quantificação matricial é uma extensão da quantificação vectorial a um conjunto fixo de vectores de coeficientes, cada um deles representando a envolvente espectral numa trama. Uma vez que existe uma grande dependência entre os espectros de tramas sucessivas, apenas um subconjunto de todas as permutações é possível, pelo que o débito binário conseguido pela quantificação matricial será mais baixo que o da quantificação vectorial.

Tsao e Gray [Tsao (85)] utilizam um livro de código matricial, treinado com apenas um orador, que contém informação sobre a envolvente espectral e sobre a energia. São geradas 512 matrizes, representando 3 tramas. Sendo a dimensão de cada trama de 20 ms, são necessários 150 bit/s para emitir a informação da energia e da envolvente espectral. O livro de código é treinado através do algoritmo LBG, que agrupa conjuntos de matrizes, utilizando como critério de distância a soma das distâncias espectrais por trama. Repare-se que o número de matrizes de código pode ser reduzido porque o número de tramas que cada matriz representa é pequeno (3 tramas, 60 ms) e porque o livro de código é treinado com um único orador.

7.7.5 Quantificação segmental

Ao contrário da quantificação matricial, que representa o espectro de um número fixo de tramas, a quantificação segmental utiliza segmentos de dimensão variável. Estes são mais flexíveis na modelação das variações da dimensão dos segmentos fonéticos [Shiraki (88)], diminuindo o número de elementos do livro de código. Esta diminuição não leva, obrigatoriamente, a uma diminuição do débito binário, uma vez que é necessário transmitir a duração de cada segmento. Deste tipo de modelação resulta também um débito binário variável, mas a principal desvantagem é a dificuldade na segmentação do sinal de entrada. Os segmentos podem ser definidos automaticamente, *e.g.*, quantificação vectorial, ou ter uma relação directa com unidades fonéticas, dando origem aos *vocoders* fonéticos.

7.7.5.1 Segmentos definidos automaticamente

Roucos e Wilgus [Roucos (82)] propuseram um algoritmo baseado nas variações espectrais em tramas sucessivas, para determinar

zonas estáveis e zonas de transição. A segmentação produzida correspondia grosso modo a *difones*, definidos como o segmento entre o centro de um segmento fonético ao centro do segmento fonético seguinte, sendo reportada uma média de 11 segmentos por segundo. Para produzir o livro de código, Roucos segmentou uma base de dados de sinais de fala provenientes de um único orador e escolheu aleatoriamente os segmentos. Foi utilizado um livro de código de 13 bits cuja procura era feita em duas etapas, de modo a diminuir a quantidade de cálculo. Na primeira etapa era procurado o segmento que minimizava a distorção num livro de código de 8 bits, em que cada segmento era representado pelo centróide correspondente a um conjunto de outros 32 segmentos. Na segunda etapa era procurado o segmento óptimo dentro do conjunto escolhido na primeira etapa. O número de distâncias calculadas foi assim reduzido de 8192 para apenas 256+32=288. Roucos propôs ainda um algoritmo de segmentação e quantificação conjunta, baseado em programação dinâmica. O débito binário total era de 230 bit/s.

De modo a que o codificador pudesse ser utilizado por um orador diferente do orador com o qual o livro de código foi treinado, Roucos e Wilgus propuseram um algoritmo de normalização [Roucos (84)] que transformava os segmento do livro de código, através da minimização do erro quadrático entre os segmentos de treino e os segmentos transformados. A desvantagem principal deste codificador, em relação a um codificador independente do orador, é a necessidade de adaptar o livro de código a um novo orador antes da sua efectiva utilização.

7.7.5.2 Codificadores fonéticos

Os codificadores fonéticos [Schwartz (80)] são um caso particular dos *vocoders* segmentais, em que os segmentos são segmentos fonéticos,

difones ou mesmo sílabas. Para treinar os livros de código é necessário que os segmentos estejam etiquetados com razoável precisão, o que ainda só é possível com recurso à segmentação manual. A segmentação do sinal de entrada é obtida com técnicas de reconhecimento de sinais de fala.

Para além do índice do segmento fonético reconhecido é transmitida a informação de carácter prosódico da sua duração, energia e frequência fundamental. As palavras de código contêm, normalmente, informação espectral baseada em coeficientes LPC, que é utilizada conjuntamente com a informação prosódica para sintetizar o sinal de saída. Naturalmente, da codificação fonética resulta um debito binário variável, função do número de unidades fonéticas por segundo.

Com esta cascata de reconhecimento e síntese os codificadores fonéticos atingem débitos binários extremamente baixos. Em relação aos codificadores em que os segmentos são definidos automaticamente, pode-se ainda tirar partido da estrutura da língua, nomeadamente da correlação entre segmentos fonéticos consecutivos, para diminuir o débito binário necessário à codificação da sequência fonética. No entanto, devido à quantidade e tipo de informação transmitida, o sinal sintetizado, embora inteligível, pode não manter as características do orador de entrada, sendo extremamente dependente dos oradores com os quais foram produzidos os livros de código. Um esquema de blocos possível da codificação fonética é apresentado na figura 7.17.

No trabalho reportado por Schwartz *et al.* [Schwartz (80)], os sinais de fala eram codificados com cerca de 100 bit/s, sendo 60 a 75 bit/s utilizados para representar a sequência fonética e os restantes para representar o valor da duração e de uma frequência fundamental por segmento fonético. As palavras de código correspondiam a difones e

continham informação sobre a envolvente espectral e sobre a trajectória do ganho, sendo o sinal reconstruído por síntese LPC. O treino dos modelos de reconhecimento e dos livros de código, contudo, foram produzidos com apenas um orador.

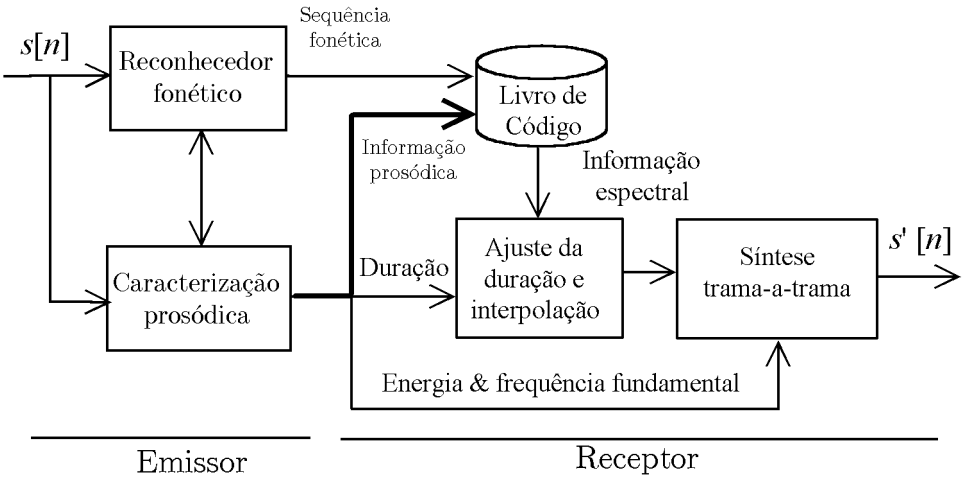

Figura 7.17
Esquema de blocos de um codificador fonético.

Picone e Doddington [Picone (89)] propuseram um codificador fonético, com um débito de 170 bit/s para transmitir a informação da sequência fonética e da respectiva duração. O codificador utilizava modelos HMM [Rabiner (93)] para segmentar o sinal, produzindo uma média de 12,3 segmentos por segundo no *corpus* de teste utilizado. De modo a tornar o codificador menos dependente do orador, Picone utilizou 2, 4 ou 8 modelos por cada segmento fonético (máximo de 64 segmentos diferentes), codificados respectivamente com 7, 8 e 9 bits por segmento. Metade dos modelos foi treinada com vozes masculinas e a outra metade com vozes femininas. A taxa de reconhecimento obtida é de cerca de 35%. A qualidade, utilizando apenas dois modelos por segmento fonético, ou seja, diferenciando apenas o género do orador, é significativamente ruidosa, sendo melhorada à medida que se aumentou o número de modelos por segmento fonético.

Ribeiro e Trancoso [Ribeiro (98)(99)] descrevem um codificador fonético para o Português Europeu, produzindo um débito binário médio de 443 bits/s. A sequência fonética é codificada com um código de entropia e modelos de bigramas de segmentos fonéticos (*i.e.*, modelos que tiram partido da probabilidade conjunta de dois segmentos fonéticos), resultando num débito médio de 59 bit/s. A duração de cada segmento é também codificada com um código de entropia de 56 bit/s. A energia e a frequência fundamental são actualizadas em tramas de 22,5 ms, sendo codificados com 328 bit/s.

O livro de código contém informação duplicada para oradores do género feminino e masculino, sendo o género identificado através do valor médio da frequência fundamental, parâmetro já transmitido. Utilizando uma frequência de decisão de 157 Hz obteve-se um erro de identificação do género de 5%.

Em vez de transmitir a energia e a frequência fundamental por trama, estas podem ser transmitidas uma vez por segmento, resultando um codificador com um débito binário médio total de 270 bit/s, mas a qualidade do sinal sintetizado torna-se demasiado tonal.

É também apresentada uma versão com adaptação ao orador, produzindo um débito binário médio de 559 bits/s[4]. Para adaptar o sinal sintetizado ao orador de entrada são transmitidos, para os segmentos correspondentes às vogais e glides, perceptualmente mais importantes que as consoantes, os valores médios dos coeficientes LSF, denominados pelos autores de *informação específica do orador*. As palavras de código são modificadas de modo a que os valores médios dos coeficientes correspondam aos valores transmitidos. É ainda proposto um método de adaptação incremental, que tira partido da

[4] Demonstração na página da disciplina (/cmr/fonetica/demo_cph.html)

correlação intra-orador, sendo as palavras de código modificadas de modo a incorporar a informação específica do orador. Com este procedimento não é necessário o treino dos livros de código previamente à utilização por um novo orador.

7.8 Conclusões e perspectivas futuras

Este capítulo apresentou as técnicas de codificação de sinais de fala de banda telefónica, segundo a ordem cronológica do seu aparecimento, exemplificadas com algumas das normas aprovadas pelos principais organismos internacionais. Como se pode verificar pela figura 7.18, em cerca de cada 10 anos foram normalizados codificadores com uma redução do débito binário para metade, sendo praticamente mantida a qualidade[5].

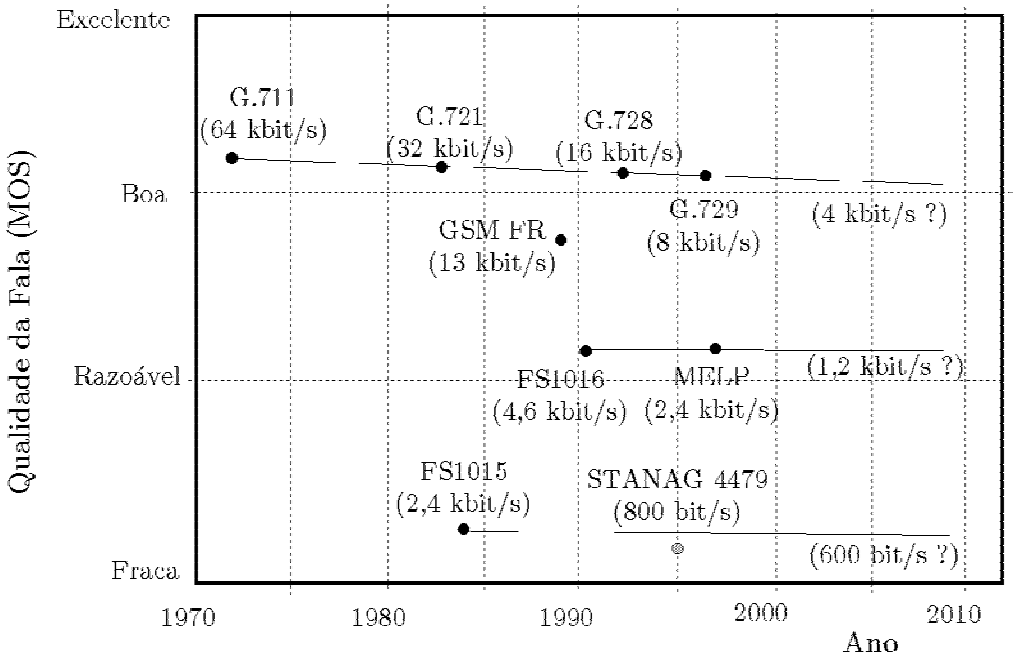

Figura 7.18
Evolução da qualidade nos codificadores de fala função da data da sua normalização (débito binário inserido).

[5] Demonstração na página da disciplina (/demos/codificacao/decada_f.html)

A única norma existente (G.711) em 1972 utilizava codificação PCM *companding* para codificar fala a 64 kbit/s. Com a introdução de um esquema diferencial com predição foi possível baixar este débito binário, mas a codificação do resíduo de predição amostra-a-amostra mantinha o débito binário acima dos 16 kbit/s. Ao tirar partido da correlação intra-tramas, quer na estimação dos coeficientes de predição quer na quantificação do resíduo de predição, quebrou-se esta barreira e, particularmente com o codificador ACELP, atingiu-se *boa* qualidade na escala MOS a débitos cada vez mais baixos.

No outro extremo do débito binário, a primeira norma (FS-1015) data de 1984, codificando o sinal de fala a 2400 bit/s. Este débito é atingido utilizando um modelo de síntese totalmente paramétrico, embora a qualidade seja apenas pouco acima da *fraca*.

Com o desenvolvimento de métodos de quantificação mais potentes, nomeadamente da quantificação vectorial, foi possível baixar o débito binário mantendo a mesma qualidade, ou, em alternativa, transmitir mais informação sobre o resíduo de predição, melhorando a qualidade. São exemplos da primeira abordagem os codificadores a muito baixo débito binário apresentados na subsecção 7.7.1. e da segunda abordagem o codificador MELP da nova norma DoD, utilizando os bits economizados na modelação mista da excitação.

A técnica comum à maioria dos métodos de codificação (APC, *Vocoder* LPC, LPAS, MELP, WI) é a utilização do filtro (LPC) de predição linear para modelar a envolvente espectral. Estes métodos diferem essencialmente nas estratégias de modelação do resíduo de predição. O codificador APC modela este resíduo amostra-amostra, atingindo consequentemente *boa* qualidade mas também o débito binário mais elevado. O *vocoder* LPC, no outro extremo do débito

binário, utiliza uma modelação binária (vozeado/não vozeado). O método LPAS, por seu lado, modela vectorialmente o resíduo de predição, sendo este estimado através de síntese, conseguindo igualmente *boa* qualidade mas com uma redução apreciável no débito binário em relação ao APC. Para além da correlação entre amostras consecutivas de que o preditor LPC tira partido, o APC e o LPAS tiram também partido da correlação entre períodos glotais nas zonas vozeadas, através da introdução do preditor (LT) de longa duração. O codificador MELP, ao utilizar uma modelação mista para o resíduo, quer através de um terceiro estado de vozeamento quer através de uma decisão binária por bandas de frequência, consegue melhor qualidade que o *vocoder* LPC com um débito de 2400 bit/s, tradicionalmente ocupado por este último. O codificador WI atinge uma gama de débito binário entre o LPAS e o *vocoder* LPC, explorando a correlação entre períodos glotais nas zonas vozeadas, não através do preditor LT, mas através de interpolação. Repare-se ainda que mesmo a codificação sinusoidal, desde que modele as amplitudes através do preditor LPC, pode ser interpretada como modelando o resíduo de predição através de um conjunto de frequências e respectivas fases, sendo mantidas constantes as amplitudes.

O capítulo termina com a descrição das técnicas utilizadas para codificação a muito baixo débito binário. Estas utilizam, devido ao baixo débito, a modelação binária do resíduo de predição, típica dos *vocoders* LPC, diferindo essencialmente no modo como tiram partido da correlação inter-tramas para codificar a envolvente espectral. Devido ao aumento do atraso e à diminuição da qualidade deram até ao ano 2000 origem apenas à criação de uma norma (NATO STANAG 4479). Os débitos binários mais baixos são alcançados com recurso aos codificadores fonéticos, em que é transmitida a representação linguística

sobre os sinais de fala. Na tabela 7.10 são listados alguns dos atributos dos codificadores normalizados.

Codificador	Débito binário [kbit/s]	Frame/ *look ahead* [ms]	Memória RAM [byte]	Complexidade [MIPS]	Ano
ITU-T					
G.711 - PCM	64	0,125/0	1	<<1	1972
G.726 - ADPCM	16,24,32,40	0,125/0	50	1,25	1990
G.722 - SBC banda larga	64	1,5/0	2000	10	1988
G.728 - LD-CELP	16	0,625/0	4000	30	1992
G.729 - CS-ACELP	8	10/5	6000	25	1995
G.723.1 - CS-ACELP	5,3	30/7,5	4000	20	1995
G.723.1 - MP-MLQ	6,3	30/7,5	4000	15	1995
G.729A - CS-ACELP	8	10/5	4000	15	1996
GSM					
GSM-FR RPE-LT	13	20/0	2000	4,5	1988
GSM-HR VSELP	5,6	20/5	8000	30	1994
GSM-EFR ACELP	12,2	20/0	6000	25	1995
DoD					
FS-1015 - *Vocoder* LPC	2,4	22,5/90	4000	8,7	1984
FS-1016 - CELP	4,6	30/7,5	4000	17	1991
STANAG 4479	0,8	67,5/90	4000	10	1994
MELP	2,4	22,5/23	24000	21	1997

Tabela 7.10
Alguns atributos de codificadores do ITU-T, GSM e DoD
(Adaptado de [Cox (95)(96)],[Kohler (97)]).

Na figura 7.19 são apresentadas, em grandes blocos, as aproximações à codificação de sinais de fala e as relações entre estas (é omitida a codificação de forma de onda). De modo a baixar o débito binário para cerca de 100-200 bit/s, deverão no futuro ser utilizadas, conjuntamente com a caracterização do orador, técnicas de reconhecimento de fala contínua (*i.e.*, não só o reconhecimento da sequência fonética mas também de palavras, frases, marcas prosódicas, etc.), que convertam o sinal acústico na sua representação ortográfica. O receptor será implementado através de sintetizadores de

texto-para-fala e a voz modificada de modo a aproximar-se da do orador de entrada.

Se a tendência de diminuição do débito binário mantendo a qualidade se mantiver ao ritmo das últimas décadas, como mostrado na figura 7.18, deverá ser atingida durante a década de 2000-2010 uma qualidade *boa* com um débito binário de 4 kbit/s. O ITU-T tem este desafio lançado desde 1995. Também os codificadores com um débito binário de 1200 bit/s poderão alcançar uma qualidade acima do *razoável*. Prevê-se ainda um grande esforço para melhorar a qualidade dos codificadores a débito binário abaixo dos 800 bit/s, nomeadamente utilizando quantificação segmental.

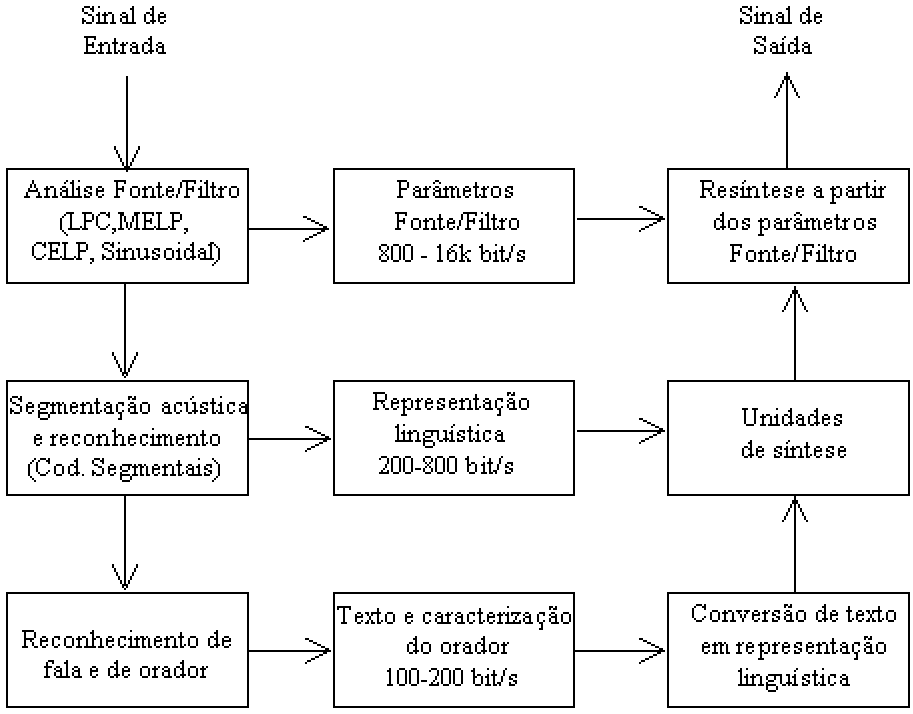

Figura 7.19
Aproximações à codificação de sinais de fala

Os codificadores de fala apresentados foram na sua maioria desenvolvidos para funcionar na rede telefónica pública. Mais recentemente tem sobressaído a utilização de codificadores de fala

funcionando através do protocolo IP (VoIP)[6], nomeadamente [Davidson (00)] utilizando as recomendações G.711 a 64 kbit/s, G.729 a 8 kbit/s e G.723.1 a 6,3 e 5,3 kbit/s. Nas ligações IP o tráfego é segmentado em pacotes e o canal partilhado por vários utilizadores, podendo cada pacote tomar caminhos diferentes até ao destinatário. Nas zonas de silêncio, por exemplo, o canal pode ser utilizado por outros utilizadores. Se por um lado a melhor utilização da infra-estrutura da rede IP torna os preços mais baratos, é também esta melhor utilização que a torna bastante sensível, pois um pacote pode não chegar ao destinatário a tempo de ser descodificado em tempo real, ou simplesmente ser eliminado devido ao aumento do tráfego em determinado troço, ou ainda por quebra de uma ligação que pode durar um tempo longo. Para aplicações que não envolvam a transmissão em tempo real estes pacotes são retransmitidos, mas para aplicações em tempo real e bidireccionais como o VoIP (*Voice over IP*), serão considerados perdidos os pacotes que não cheguem ao destinatário com um atraso máximo pré-definido, podendo a conversação tornar-se inteligível. Este tipo de situação leva a repensar as estratégias de codificação e descodificação, devendo emergir uma nova geração de codificadores. Por um lado a informação entre pacotes deverá estar correlacionada de modo a que na falta de um pacote (ou rajada de pacotes) seja possível sintetizar o sinal com a informação dos pacotes adjacentes. Por outro lado não deverá haver uma dependência entre pacotes que a perda de um (ou rajada) iniba a síntese em pacotes adjacentes.

Uma outra área do desenvolvimento na codificação de fala será baixar o débito binário dos codificadores de sinais de banda larga. Este desafio foi já iniciado no seio do ITU-T, pretendendo-se alcançar com

[6] Demonstração na página da disciplina (/demos/voip/voip.html)

um débito de 24 kbit/s a qualidade obtida pela recomendação G.722 a 56 kbit/s, e com 16 kbit/s a qualidade do G.722 a 48 kbit/s. Repare-se que quando da utilização de filtros LPC, a ordem de predição deverá aumentar acima dos 10 coeficientes, típico na codificação de sinais de banda telefónica, de modo a modelar os formantes que se situem acima dos 3400 Hz.