MIGUEL ARJONA RAMÍREZ

BUSCA DE INOVAÇÕES E TRATAMENTO DE TRANSITÓRIOS EM CODIFICADORES DE VOZ CELP

Tese apresentada à Escola
Politécnica da Universidade de
São Paulo para obtenção do
título de Doutor em Engenharia.

São Paulo
1997

MIGUEL ARJONA RAMÍREZ

BUSCA DE INOVAÇÕES E TRATAMENTO DE TRANSITÓRIOS EM CODIFICADORES DE VOZ CELP

Tese apresentada à Escola
Politécnica da Universidade de
São Paulo para obtenção do
título de Doutor em Engenharia.

Área de Concentração:
Engenharia de Sistemas Eletrônicos

Orientadores:
Prof. Dr. Max Gerken
Prof. Dr. Normonds Alens

São Paulo
1997

Arjona Ramírez, Miguel
Busca de inovações e tratamento de transitórios
em codificadores de voz CELP.
São Paulo 1997.
110p.
Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Eletrônica.
1. Codificadores de Voz 2. Busca de Inovações 3. Tratamento de Transitórios I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Eletrônica II.t

Agradecimentos

Ao Prof. Dr. Normonds Alens, meu orientador inicial, presto uma homenagem póstuma pelo acompanhamento constante e atenção a aspectos objetivos e fundamentais durante a maior parte deste trabalho.
Ao Prof. Dr. Max Gerken, meu orientador, pela orientação na etapa final do trabalho, pelo zelo na manutenção da lógica e da coerência neste texto. Agradeço-lhe ainda a orientação ativa nas fases iniciais deste trabalho com relação a aspectos de filtros variantes no tempo.
Ao Prof. Dr. Ivandro Sanches pelo apoio dado às pesquisas em processamento de voz no âmbito do Laboratório de Processamento de Sinais e Sistemas (LPS) e pelas inestimáveis contribuições à integração de trabalhos de codificação e de reconhecimento, tendo como base o tratamento de sinais de voz ruidosos.
Ao Prof. Dr. Flávio Almeida de Magalhães Cipparrone pelo estímulo proporcionado especialmente nas últimas etapas deste trabalho.
Ao colega Mário Minami pelas discussões sobre processamento de voz e bancos de dados de sinais de voz e também pela ajuda na obtenção e aquisição destes.
Ao colega Ulisses Antonio Donato pelo desenvolvimento de uma implementação do codificador VSELP segundo a norma IS-54, pela colaboração na implementação de variantes e pela ajuda em vários aspectos de sistemas operacionais e de compiladores, incluindo ambientes com processadores digitais de sinais (DSP).
Ao colega Celso Setsuo Kurashima pelas discussões sobre codificadores de voz padronizados e em vias de padronização, bem como sobre procedimentos objetivos de medida da distorção causada pelos codificadores de voz e pela ajuda inestimável na obtenção de documentos e software a seu respeito.
Ao colega Rubem Dutra Ribeiro Fagundes pelas discussões sobre a interação entre reconhecimento de voz e processamento de linguagem natural e, especificamente, pelo auxílio prestado nas etapas iniciais deste trabalho com a aquisição de voz através de placas baseadas em DSPs.
À Fundação de Amparo à Pesquisa do Estado de São Paulo pela concessão e manutenção de Bolsa de Doutorado.
À minha mãe, Carmen, pelo estímulo.
À minha sogra, Sugiyo, pelo apoio e colaboração.
À minha esposa, Mariko, e à nossa filha, Karina, dedico este trabalho e agradeço toda a compreensão e inspiração.

PROGRAMA:
NÍVEL: MESTRADO: ( ) DOUTORADO: (X)

CANDIDATO: MIGUEL ARJONA RAMÍREZ

TÍTULO DO TRABALHO: "BUSCA DE INOVAÇÕES E TRATAMENTO DE TRANSITÓRIOS EM CONDIFICADORES DE VOZ CELP".
ORIENTADOR: PROF.DR. MAX GERKEN
DATA DA DEFESA: 18.12.1997

OBSERVAÇÕES:

A errata consolidando as observações dos membros de banca se encontra no anexo.

Max Gerken
Assinatura do Examinador
(caso necessário faça cópias deste impresso)

Errata da Tese "BUSCA DE INOVAÇÕES E TRATAMENTO DE TRANSITÓRIOS"

página 3, 2º item de contribuição, duas linhas finais
Onde se lê "compensado por um desempenho melhor para outros sinais.", leia-se "que é menos significativa quando expressa em unidades de distorção perceptiva.".
página 10 no início do parágrafo que contém a Equação (2.13)
Substitua "O sinal quantizado" por "O sinal residual quantizado".
página 11, parágrafo seguinte à Equação (2.13)
Substitua "igualdade dos erros de quantização do sinal residual e do sinal de voz" por "igualdade entre o erro de quantização do sinal residual e o erro de reconstrução do sinal de voz".
página 26, Equação (3.4)
Substitua o símbolo "$-a_p$" pelo símbolo "$-a_{p,l}$".
página 26, 2º item, em seguida à Equação (3.4)
Substitua "$b(k) = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$" por "$b = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$.".
página 43, Figura 3.11
Substitua o símbolo "u_0" pelo símbolo "$\tilde{u}_0$".
página 51, Equação (4.3)
Substitua o símbolo "$c(n)$" pelo símbolo "$e(n)$".
página 52, 2º parágrafo da seção 4.4
Complemente com a substituição de "excitação regular" por "excitação regular. Por exemplo, a posição do pulso de fase i_k é".
página 53, 3ª linha do 1º parágrafo da seção 4.5
Onde se lê "4 bits por 7,5 ms", leia-se "4 pulsos por 7,5 ms".
página 64, 1º parágrafo
Complete o parágrafo mediante a substituição de "como amplificação ou atenuação." por "como amplificação, atenuação ou compansão (seção 2.4).".
página 68, último parágrafo da seção 5.3
Apague a palavra "ponderada".
página 77, 1ª linha da Equação (6.26)
Substitua "se $j = 3$" por "se $j = 0$".
página 77, Equação (6.27)
Substitua "β_j" por "α_j".
página 78, Equação (6.30)
Apague o fator "G".
página 80, último parágrafo, 4ª linha
Onde se lê "0,2 dB", leia-se "0,25 dB".
página 80, último parágrafo, penúltima linha
Onde se lê "0,14 pontos inferior", leia-se "apenas 0,06 pontos superior".
páginas 81 e 82, Tabelas 6.1 a 6.6, linha "Exaustiva", coluna "Busca"
Substitua a palavra "Exaustiva" pela expressão "Exaustiva de posições".

página 82, Tabela 6.6, linha "Focalizada", coluna "TIMIT" sob "SNRSEG"
Substitua o valor "9,4216" pelo valor "9,4371".
página 82, Tabela 6.6, linha "Focalizada", coluna "TIMIT" sob "PSQM"
Substitua o valor "2,2800" pelo valor "2,0848".
página 82, Tabela 6.6, linha "Conjunta", coluna "TIMIT" sob "SNRSEG"
Substitua o valor "9,1816" pelo valor "9,1862".
página 104, 3º parágrafo, último período
Onde se lê "equivalência entre os dois métodos de busca pois, embora a busca conjunta incorra numa perda de relação sinal-ruído segmentada de 0,2 dB para alguns sinais, a melhora obtida sobre 1680 sinais da base de dados TIMIT acarreta uma redução de 0,14 pontos na distorção PSQM média.", leia-se "ligeira inferioridade da busca conjunta, que incorre numa perda média de relação sinal-ruído segmentada de 0,25 dB para um subconjunto de 1680 sinais da base de dados TIMIT. Visto pela distorção perceptiva, a busca conjunta causa um acréscimo de apenas 0,06 pontos na distorção média.".

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Os codificadores preditivos excitados por códigos (CELP) são descritos de forma variante no tempo com base em uma abordagem vetorial do método de análise-mediante-síntese, suscitando as questões da propagação dos estados e da busca do sinal de excitação.

Descrevem-se várias estruturas esparsas no tempo ou de dimensão reduzida para as inovações do dicionário fixo, em especial os multipulsos algébricos (ACELP) e os dicionários definidos por vetores-base (VSELP). Estes últimos admitem treinamento, deduzindo-se um algoritmo para tal.

Propõe-se uma busca conjunta de amplitude e posição dos multipulsos algébricos. Ela aproveita a natureza do dicionário para reduzir o número de vetores-código pesquisado e o número de coeficientes de autocorrelação calculados em comparação tanto com a busca exaustiva em posições quanto com a busca focalizada.

Analisam-se os transitórios em codificadores CELP a partir de dois pontos de vista baseados no decodificador e no analisador. A partir de um método de supressão de transitórios por ajuste do vetor de estados, propõe-se um método novo baseado numa estrutura não recorrente estendida (XFIR), revelando uma equivalência entre ajuste de estados e estrutura do filtro de síntese. Tanto a estrutura recursiva na forma direta I quanto a estrutura XFIR truncada são testadas dentro dos codificadores ACELP e VSELP treinado. Mais uma estrutura recorrente direta e duas em treliça são testadas dentro do codificador VSELP. Resulta um desempenho superior do codificador VSELP com a forma direta I recorrente e o codificador ACELP mostra-se menos sensível ao tipo de estrutura e à ordem da estrutura XFIR.

Abstract

Code-excited linear predictive (CELP) coders are introduced as time-variant systems, building upon a vector approach to the analysis-by-synthesis method. Attendant issues like state propagation and excitation search are considered.

Fixed codebook innovations are described which are sparse in the time domain or lie in a reduced dimension excitation space. Algebraic multipulse (ACELP) and basis vector (VSELP) codebooks are the main target. The latter may be trained and an algorithm is deduced to do so.

A joint amplitude and position search of an ACELP codebook is proposed. The features of this codebook are used to reduce both the number of searched codevectors and the number of computed autocorrelation coefficients in comparison to the full position search and the focused search.

Transients in CELP coders are analyzed from two standpoints, one at the decoder and the other at the analysis process. Starting from a state vector adjustment method, a novel transient suppression method is proposed which is based upon an extended finite impulse response (XFIR) structure, revealing an equivalence between state adjustment and synthesis filter structure. Both the direct form I recursive structure and the truncated XFIR structure are tested within the ACELP and the trained VSELP coders. One additional direct form recursive structure and two lattice recursive structures are tested inside the VSELP coder. The test results show that the VSELP coder performs better with the recursive direct form I structure and that the ACELP coder displays a reduced sensitivity to structure and impulse response length as well.

Capítulo 1

ORGANIZAÇÃO

Na seqüência apresentam-se a organização dos capítulos deste trabalho e a menção explícita das contribuições que contém.

No Capítulo 2 apresentam-se os processos básicos envolvidos na codificação da fala e os limites de compressão da informação que poderiam ser atingidos segundo vários critérios. Procura-se manter um contexto de processamento de sinais para a conceituação da quantização e da predição. Ademais, esta última é vinculada na origem à resposta impulsiva do filtro de síntese, a ser usada nos Capítulos 3 e 7. Finalmente, são apresentados alguns codificadores atuais usados em várias modalidades de comunicações telefônicas.

No Capítulo 3, amplia-se o contexto de processamento de sinais montado com a elaboração do método de análise-mediante-síntese vetorial para a introdução da codificação CELP. Ela é apresentada explicitamente de forma variante no tempo para a colocação da questão da propagação dos estados entre sub-blocos, a ser retomada no Capítulo 7. Em seguida, por etapas ilustradas, entram em cena a ponderação perceptiva, a matriz de resposta impulsiva do caminho de síntese e a remoção do efeito dos estados recebidos para, então, adentrar ao ambiente em que se processa a busca da excitação. Elabora-se geometricamente o processo de busca para a introdução das operações de projeção, filtragem regressiva ("backward filtering") e obtenção da energia transformada dos vetores-código. Fornece-se uma idéia da complexidade da busca e de sua evolução tecnológica. Descrevem-se os dicionários estocástico e adaptativo isolados e em conjunto, conduzindo à questão da ortogonalização. Finalizando, a partir da representação mais econômica do dicionário de códigos adaptativo sobre a tradicional representação "por extenso" do dicionário estocástico, abstraem-se propriedades de representações mais eficientes do ponto de vista do armazenamento, unindo os modelos apresentados em seguida no Capítulo 4 e preparando a descrição dos métodos de busca do Capítulo 6.

No Capítulo 4, representam-se vários dicionários fixos esparsos no tempo ou em dimensionalidade. No primeiro caso incluem-se os estocásticos ceifados no centro, os multipulso, os que utilizam pulsos regulares e os compostos por multipulsos algébricos e no segundo caso tem-se os compostos por vetores-base. Ademais, a maioria dos dicionários apresentados são algébricos no sentido de que suas componentes não nulas assumem exclusivamente os valores 1 e -1. Neste sentido generalizado, são algébricos

todos os dicionários que acabaram de ser mencionados, com as exceções do dicionário multipulso e do dicionário de pulsos regulares. Incluem-se também os dicionários com estrutura conjugada na categoria dos algébricos. Para comparação com o processo básico de busca dos codificadores CELP do Capítulo 3 e com o processo de busca conjunta de amplitude e posição do Capítulo 6, expõe-se um método de busca de inovações multipulso. Finalmente, mostra-se a possibilidade de treinar os dicionários definidos por vetores-base, tomando-se o caso do VSELP. O algoritmo de treinamento é derivado, servindo de referência para a execução de testes de avaliação dos métodos de tratamento de transitórios do Capítulo 7.

No Capítulo 5 descrevem-se os sinais individuais e os sinais agrupados em base de dados, os quais foram empregados para os testes dos codificadores padronizados e dos codificadores resultantes dos métodos propostos. Apresentam-se as características das medidas objetivas usadas nos testes: a relação sinal-ruído segmentada (SNRSEG) e a "perceptual speech quality measure" (PSQM). Ademais, descrevem-se os procedimentos de preparação dos sinais para as medidas objetivas, que são as desativações de pré-filtros e pós-filtros e a reamostragem com interpolação.

No Capítulo 6 apresentam-se três métodos de busca subótimos para o dicionário de multipulsos algébricos no contexto do codificador ACELP da recomendação G.723.1 da ITU-T. A busca focalizada é o método recomendado pela ITU-T, que restringe a busca a um subconjunto dinamicamente selecionado de vetores-código. A busca exaustiva de posições resulta ao levantar-se a restrição imposta na busca focalizada. A busca conjunta de amplitude e posição (AMPE-JPAS), que é o método proposto, aproveita a estrutura do dicionário ACELP para incrementar o grau de otimização da busca multipulsos do Capítulo 4 na direção da busca CELP básica do Capítulo 3. Ademais, o subconjunto pesquisado do dicionário de códigos possui um número reduzido de elementos escolhidos dinamicamente.

No Capítulo 7 tratam-se os transitórios gerados no sinal reconstruído e nos estados do filtro de síntese pelas variações dos parâmetros deste filtro e pela propagação do seu vetor de estados entre configurações adjacentes. Inicia-se exemplificando as condições para a existência de uma relação de inversão entre os filtros de análise e de síntese para duas estruturas selecionadas num contexto de grafos de fluxo de sinais. Esta ilustração também estabelece o sinal residual como referência de sinal de excitação do filtro de síntese. Toma-se então a excitação quantizada vetorialmente e analisam-se testes de desempenho de várias estruturas diretas e em treliça para o filtro de síntese em algumas implementações do codificador VSELP, sobressaindo-se a forma direta I. Em seguida, identificam-se 2 tipos de transitórios: transitórios nos estados do filtro de síntese no decodificador e transitórios no sinal residual resultante da filtragem inversa no analisador. Para suprimir os transitórios no decodificador, toma-se como base um método que ajusta os estados na passagem entre sub-blocos, cancelando as suas componentes que dependem diretamente dos coeficientes da configuração anterior do filtro. Na extensão proposta para o encadeamento de sub-blocos, esse mecanismo de supressão de transitórios apresenta, no espaço de estados, todas as propriedades de um filtro FIR transversal, a menos da limitação do número de variáveis de estado. Assim, propõe-se a interpretação do processo como um filtro FIR estendido (XFIR), conduzindo a implementações como filtro

FIR transversal mediante truncamento da sua resposta impulsiva. Adicionalmente, nota-se uma equivalência entre o método de transferência dos estados e a estrutura de implementação do filtro. Retomando os transitórios no analisador, mostra-se que o transitório atual depende apenas dos coeficientes atuais no caso do filtro de análise FIR transversal. Para todas as outras estruturas de filtro de análise, aparece uma dependência composta com os coeficientes anteriores do filtro. Assim, esta análise sugere mais complicações para as outras estruturas na presença de ruído de quantização do sinal de excitação.

Segue abaixo a relação das contribuições apresentadas neste trabalho.

- Na Seção 4.4 apresenta-se uma interpretação do dicionário ACELP nos moldes da representação vetor-gerador/janela explicitada na seção 3.11. Esta interpretação estabelece vínculos com as estruturas dos dicionários ceifados no centro e com os dicionários de pulsos regulares e auxilia na unificação do entendimento de métodos de busca, propiciando novas propostas.

- Na Seção 6.3 propõe-se um método de "busca conjunta de amplitude e posição em dicionários multipulsos algébricos" (AMPE-JPAS), que possui um critério de comparação melhorado em relação à busca multipulsos tradicional. Ao aplicar esta busca conjunta sobre o dicionário ACELP da recomendação G.723.1 da ITU-T, há uma redução no número de vetores-código pesquisados e no número de coeficientes de autocorrelação da resposta impulsiva calculados em comparação com a busca focalizada ("focused search") ao custo de uma pequena queda de desempenho em alguns casos, compensado por um desempenho melhor para outros sinais.

- Na Seção 7.6 apontam-se semelhanças entre filtros FIR e o mecanismo de supressão de transitórios nos estados do filtro de síntese. A interpretação de filtragem para a combinação "filtro + atualização de estados" é denominada "filtro de resposta impulsiva finita estendida (XFIR)". Na mesma seção propõe-se um mecamismo de filtragem sem ajuste de estados que aproxima o filtro XFIR por um filtro FIR mediante truncamento da resposta impulsiva. Este método é aplicado na Seção 7.8. Além disso, este método de supressão de transitórios aplica-se bem a codificadores com dicionário fixo ACELP, conforme resultados apresentados na Seção 7.10, podendo ser usado para redução da complexidade apesar de uma pequena perda de SNRSEG.

- Na Seção 7.5 comparam-se 2 métodos de supressão de transitórios em codificadores preditivos de voz, que ocorrem em subsistemas diferentes: um no decodificador e o outro no analisador. O primeiro envolve ajuste de estados e não restringe a estrutura do filtro de síntese ao passo que o segundo não ajusta os estados mas fixa a estrutura do filtro de análise, que tem que ser a estrutura FIR transversal para obter melhor desempenho. A base comum entre os 2 métodos pode ser aumentada e suas diferenças localizadas com o auxílio do método XFIR truncado proposto na Seção 7.6. Para isso basta observar que o método XFIR truncado não ajusta os estados mas fixa a estrutura do filtro de síntese na FIR transversal. Portanto, esta simplificação reduz o teste

dos dois métodos ao teste de desempenho de 2 estruturas do filtro de síntese: a FIR transversal e a IIR na forma direta I, respectivamente.

Capítulo 2

A CODIFICAÇÃO DO SINAL DE VOZ EM TELEFONIA

2.1 Introdução à codificação da fala

A linguagem é o meio mais importante de comunicação utilizado pela humanidade e sua manifestação primária é a fala. Dessa naturalidade de uso advém o interesse inicial em sua transmissão por meios artificiais.

Para auxiliar na sua compreensão, convém decompor o processo de comunicação pela fala em produção, transmissão e percepção. Esta decomposição recebe freqüentemente a denominação de cadeia da fala ("speech chain") [59].

Entretanto, outra conotação de "cadeia da fala", em vez da distribuição espacial, ressalta a distribuição temporal contínua dos sons que compõem a fala e a sua representação como uma cadeia discreta após o seu reconhecimento pelo receptor humano [65].

Considerando a cadeia da fala como uma sucessão de fones ou símbolos acústicos discretos, podemos calcular um limite inferior para a taxa de transmissão de um codificador de voz.

Em primeiro lugar, pode-se determinar a informação média de um único símbolo acústico ou fone, assumindo que sua realização seja independente dos fones vizinhos. Pela Teoria da Informação [37], essa informação é a entropia da fonte

$$H = -\sum_{i=1}^{N_A} p_i \log_2 p_i, \tag{2.1}$$

onde p_i para $i = 1, 2, \ldots, N_A$ é a distribuição de probabilidade de ocorrência do conjunto de N_A fones.

Tomando uma duração média t_f para os fones, a taxa de transmissão média será

$$I = \frac{H}{t_f}. \tag{2.2}$$

Para o português falado no Rio de Janeiro, temos as freqüências relativas de ocorrência dos 37 fones levantadas em [1] e reproduzidas na Tabela 2.1. Assumindo-as iguais à distribuição da probabilidade de ocorrência dos fonemas e tomando a

duração média dos fones $t_f = 100$ ms, tem-se uma taxa limitante inferior

$$I_{\text{inf}} = 48 \text{ bit/s}. \tag{2.3}$$

Fisicamente, o sinal de voz propaga-se através de uma onda acústica, que carrega um conjunto de informações.

Por um lado, o sinal de voz é produzido no aparelho fonador humano, que impõe restrições de largura de faixa ao sinal produzido [22], sendo, entretanto, as restrições mais severas impostas pelo aparelho auditivo. A faixa de audição máxima do ser humano denomina-se faixa de áudio e se estende de 20 Hz a 20 kHz. Ademais, o sinal de voz pode ter sua faixa de freqüências limitada a 10 kHz sem afetar sua percepção [39].

Por outro lado, a atenuação do canal acústico limita o alcance da fala, que só pôde ser ampliado com o recurso da telefonia. Por sua vez, o canal telefônico limita a faixa de freqüências do sinal de voz ao intervalo de 300 Hz a 3400 Hz, sendo que ambos os limites afetam a qualidade percebida da voz.

A capacidade C de transmissão de informação de um canal telefônico depende da largura de faixa das freqüências W de sinal de potência S a que ele consegue dar passagem e do seu sinal de ruído aditivo de potência N, com o qual o sinal terá que "competir" para se fazer notar. Para valores típicos de relação sinal-ruído $\text{SNR} = 10 \log\left(\frac{S}{N}\right) = 30$ dB e de faixa de passagem de $W = 3$ kHz, obtém-se, pelo Teorema 2 do clássico artigo de Shannon [66], uma taxa de informação de

$$\begin{aligned} C &= W \log_2\left(1 + \frac{S}{N}\right) \\ &= 30 \text{ kbit/s}. \end{aligned} \tag{2.4}$$

Considerando agora que o aparelho auditivo humano não consegue captar parte da informação acústica recebida, pode-se determinar uma taxa de informação mínima que imporia uma distorção não perceptível. Isto é, a essa taxa seria possível executar uma codificação transparente [39]. Essa taxa mínima foi chamada de "entropia perceptiva" por [38] e será denotada por H_p.

Os fenômenos que impedem o aparelho auditivo de captar parte da informação acústica são conhecidos como mascaramento, que podem ocorrer no domínio do tempo ou da freqüência. Este último é mais importante na codificação de voz (seções 3.4 e 3.5).

No mascaramento, um tom de dada freqüência, capaz de ser perfeitamente ouvido quando isolado, não é percebido quando emitido simultaneamente com um tom de freqüência próxima e de maior potência.

Para a faixa de freqüências do canal telefônico, Johnston [38] estimou

$$H_p = 10 \text{ kbit/s}. \tag{2.5}$$

2.2 A distribuição de amplitude do sinal de voz

A distribuição de amplitudes ou função densidade de probabilidade (fdp) do sinal de voz depende da faixa de freqüências considerada e das condições de gravação.

Tabela 2.1: Freqüências relativas dos fones do português falado no Rio de Janeiro.

Fone	Exemplo	Freqüência relativa (%)
a	*a*la	12,94
ϵ	*e*la	1,91
e	*e*le	4,82
i	v*i*da	8,57
$\supset$	l*o*ja	1,00
o	gl*o*bo	2,71
u	l*u*cro	5,49
$\tilde{a}$	maç*ã*	2,12
$\tilde{e}$	c*en*tro	2.30
$\tilde{\imath}$	s*im*	1,75
$\tilde{o}$	s*om*	0,75
$\tilde{u}$	*um*	1,27
j	do*i*s	3,13
w	ma*u*	3,19
$\tilde{j}$	t*em*	1,48
$\tilde{w}$	n*ão*	1,23
p	*p*ato	2,29
t	*t*ato	3,94
k	*c*ato	4,19
b	*b*ato	1,09
d	*d*ado	2,64
g	*g*ado	0,93
f	*f*ala	1,46
v	*v*ala	1,23
s	*s*ala	4,18
z	ca*s*a	1,81
r	ca*r*o	3,58
m	*m*orte	4,12
n	*n*orte	2,40
$\tilde{n}$	ni*nh*o	0,68
l	*l*ado	1,72
λ	a*lh*o	0,21
$\int$	*ch*ave	2,12
$\mathcal{Z}$	*j*ogo	1,32
R	ca*rr*o	2,06
$\hat{d}$	*d*ia	1,92
$\hat{t}$	*t*ia	1,44

Considerando a faixa de freqüências do canal telefônico e usando um microfone de alta qualidade, a fdp do sinal de voz é bem aproximada pela fdp *laplaciana ou exponencial bilateral* [37]:

$$S \sim L\left(0, \sigma_s^2\right) \Leftrightarrow p_S(s) = \frac{1}{\sqrt{2}\sigma_s} \exp\left(-\sqrt{2}\frac{|s|}{\sigma_s}\right). \tag{2.6}$$

Segundo [19], o estimador pelo método dos momentos (MME) de σ_s é

$$\hat{\sigma}_s = \sqrt{\frac{\sum_i S_i^2}{N}} \tag{2.7}$$

enquanto o estimador de máxima verossimilhança (MLE) é

$$\sigma_s^* = \sqrt{2}\frac{\sum_i |S_i|}{N}, \tag{2.8}$$

onde $\sigma_s^{*^2}$ pode ser usado como estimador da energia do sinal.

De fato, $\sigma_s^{*^2}$ foi usado para a compressão "quase-sem-perdas" do sinal residual para distribuição de bancos de dados de voz [24].

Um ajuste melhor à fdp experimental de longo prazo do sinal de voz pode ser obtido com a distribuição *gama* [37]:

$$S \sim \Gamma\left(0, \sigma_s^2\right) \Leftrightarrow p_S(s) = \frac{\sqrt[4]{3}}{\sqrt{8\pi\sigma_s|s|}} \exp\left(-\sqrt{3}\frac{|s|}{2\sigma_s}\right). \tag{2.9}$$

Mas, quando são considerados segmentos de centésimos de segundo do sinal de voz, suas fdps são mais bem descritas pela fdp *gaussiana ou normal*:

$$S \sim N\left(0, \sigma_s^2\right) \Leftrightarrow p_S(s) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{s^2}{2\sigma_s^2}\right). \tag{2.10}$$

2.3 Processos básicos da codificação

Para fins de processamento ou transmissão, o sinal de voz é capturado na forma de um sinal elétrico analógico, proporcional à pressão sonora.

O sinal elétrico é contínuo no tempo e na amplitude e, para ser convertido em uma representação digital, ambas as dimensões têm que ser discretizadas. A discretização no tempo denomina-se amostragem e a discretização em amplitude é chamada de quantização.

Geralmente, a operação de amostragem é executada antes da operação de quantização. A quantização, via de regra, consiste numa seqüência de comparações da amplitude amostrada com níveis de comparação projetados. Para evitar que variações temporais do sinal prejudiquem a precisão da quantização, o amostrador costuma ser um circuito de dois estados, amostragem e manutenção ("sample and hold").

A freqüência de amostragem tem seu valor fundamentalmente restrito pelo Teorema da amostragem de Nyquist, que afirma a possibilidade de reconstrução perfeita

de um sinal de faixa limitada a partir das amostras. Para tanto, a freqüência de amostragem não pode ser inferior ao dobro da maior freqüência de interesse do sinal.

De acordo com a seção 2.1, se o objetivo for a codificação transparente da fala, deve-se começar pela limitação do sinal de voz a uma faixa com limite superior em 10 kHz e amostrá-lo a 20 kHz no mínimo.

Em geral, os codificadores digitais na sua maioria amostram o sinal de voz na freqüência de 8 kHz, para o que pré-filtram o sinal em faixas cujo extremo inferior varia entre 50 e 200 Hz e cujo extremo superior pode atingir uns 3,6 kHz.

Nesses casos, dada a limitação natural ou forçada da faixa de freqüências do sinal de voz, a amostragem por si só é um processo sem perda de informação. De forma geral, as redundâncias do sinal podem ser aproveitadas pelos codificadores para a concentração da informação em um número reduzido de parâmetros como será visto em seguida.

Por outro lado, a quantização que segue a amostragem é um processo com perda de informação. Se as perdas forem imperceptíveis, elas representam apenas irrelevância removida do sinal [37].

2.4 A modulação por código de pulsos

A modulação por código de pulsos (PCM) é o esquema de codificação mais simples, que busca apenas a remoção da irrelevância.

Os codificadores PCM podem ser vistos como conversores analógico-digitais (A/D). Quando os níveis de representação são equidistantes, temos conversores A/D lineares.

Para o caso de codificadores telefônicos mais complexos, converte-se inicialmente o sinal de voz para uma representação digital através de conversores A/D lineares com quantizadores de 16 bits e freqüência de amostragem de 8 kHz, compondo uma taxa de 128 kbit/s. Em seguida, aplicam-se algoritmos para concentrar certas redundâncias do sinal de voz em parâmetros que são posteriormente quantizados. Em geral, esses algoritmos são denominados codificadores de voz por si sós.

Entretanto, para codificadores PCM "puros", pode-se aproveitar o efeito perceptivo de mascaramento do ruído de quantização por um sinal de amplitude maior.

Os quantizadores com distribuição exponencial dos níveis de representação são os que introduzem menor distorção no sinal reconstruído. Assim, estabeleceram-se em 1972 para a telefonia digital dois padrões de codificadores PCM à taxa de 64 kbit/s com distribuições de níveis globalmente logarítmicas embora lineares por segmento, que são a lei μ, adotada no Japão e nos Estados Unidos, e a lei A, adotada no resto do mundo, inclusive o Brasil [10].

As tabelas de quantização para as duas leis que compõem a recomendação G.711 encontram-se em [37]. O log PCM segundo a lei A apresenta maior resolução para pequenas amplitudes e uma faixa dinâmica equivalente ao PCM linear de 12 bits.

Pelo outro lado, o log PCM segundo a lei μ equivale em faixa dinâmica ao PCM linear de 13 bits mas apresenta maior ruído de quantização para pequenas amplitudes.

Entretanto, ambas as leis A e μ são consideradas equivalentes em qualidade global.

A complexidade adicional do log PCM sobre o PCM linear consiste na implementação das leis de compansão (compressão/expansão), que são lineares na origem e logarítmicas na saturação. A recomendação da ITU-T é o uso de "log-compansão digitalmente linearizável", que implica no uso de tabelas para a conversão entre os formatos PCM linear e log PCM. Em contraste, os primeiros sistemas PCM usavam as não-linearidades de diodos para aproximar as leis de compansão [37].

Tanto a passagem dos códigos pela tabela de quantização como a passagem dos sinais por diodos são de complexidades desprezíveis, podendo ser incorporadas no mesmo circuito integrado que realiza a conversão A/D.

2.5 A quantização do sinal residual

O processo de quantização de um sinal pode ser melhorado pela consideração dos erros de quantização cometidos nas amostras anteriores.

A diretriz básica do processo de quantização preditiva por análise-mediante-síntese é o aproveitamento máximo da informação disponível ao decodificador no instante da reconstrução, admitindo-se a validade de um modelo de predição. Os modelos de predição procuram remover a redundância do sinal de voz (seções 2.3 e 2.6). Assim, o codificador incorpora um decodificador completo numa malha de realimentação (Fig. 2.1).

Considerando esta figura, seja o valor predito da amostra atual dado por

$$\hat{s}(n) = P[s(n)|\tilde{s}(n-1), \tilde{s}(n-2), \ldots, \tilde{s}(n-p)], \tag{2.11}$$

onde $P[.|.]$ indica um preditor de ordem p. Esse preditor pode ser linear ou não-linear, fixo ou variável bem como sua determinação pode ser de forma progressiva ("forward") ou regressiva ("backward"), como será visto na seção 2.6.

O sinal diferencial ou sinal residual de predição é o erro de predição (Fig. 2.1), dado por

$$e(n) = s(n) - \hat{s}(n). \tag{2.12}$$

O sinal quantizado é dado por

$$\tilde{e}(n) = Q[e(n)], \tag{2.13}$$

onde o quantizador $Q[.]$ pode ser fixo ou adaptativo, com conseqüente erro de quantização do sinal residual

$$\varepsilon_e(n) = e(n) - \tilde{e}(n). \tag{2.14}$$

O sinal reconstruído (Fig. 2.1) é gerado como

$$\tilde{s}(n) = \tilde{e}(n) + \hat{s}(n). \tag{2.15}$$

Neste ponto, então, pode-se expressar o sinal predito de duas formas diferentes a partir das equações (2.12) e (2.15), respectivamente,

$$\hat{s}(n) = s(n) - e(n) \tag{2.16}$$

$$\hat{s}(n) = \tilde{s}(n) - \tilde{e}(n). \tag{2.17}$$

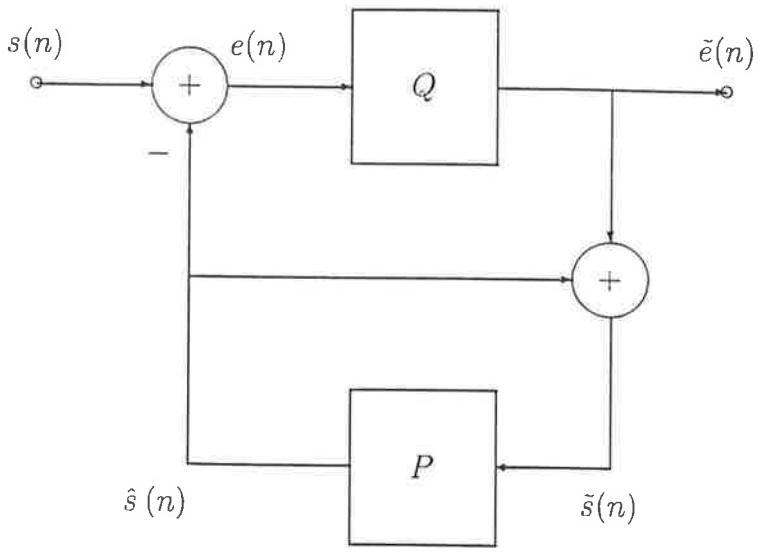

Figura 2.1: Codificador preditivo com quantizador escalar adaptativo ou fixo e com preditor adaptativo ou fixo, linear ou não-linear.

Igualando os segundos membros das Eqs. (2.16) e (2.17) e reordenando termos, obtém-se

$$s(n) - \tilde{s}(n) = e(n) - \tilde{e}(n). \tag{2.18}$$

Ainda, definindo naturalmente o erro de reconstrução do sinal de voz como

$$\varepsilon_s(n) = s(n) - \tilde{s}(n), \tag{2.19}$$

e usando a definição do erro de quantização do sinal residual na Eq. (2.14), pode-se reescrever a Eq. (2.18) como

$$\varepsilon_s(n) = \varepsilon_e(n). \tag{2.20}$$

Portanto, da igualdade dos erros de quantização do sinal residual e do sinal de voz (Eq. (2.20)) numa quantização preditiva por análise-mediante-síntese conclui-se que o efeito combinado do preditor e da realimentação é a propagação sem amplificação do erro de quantização do sinal de menor amplitude $e(n)$ para o sinal de maior amplitude $s(n)$. Ou seja, o quantizador de malha fechada apresenta um ganho de relação sinal-ruído sobre o quantizador de malha aberta.

Esse incremento em relação sinal-ruído é obtido pela ação do preditor e pode ser quantificado pelo seu *ganho de predição*, no caso, em malha fechada, que é definido como

$$G_P = \frac{\sigma_s}{\sigma_e}, \tag{2.21}$$

onde σ_s^2 e σ_e^2 são as energias do sinal de voz e do sinal residual, respectivamente, determinadas para um bloco de comprimento dado.

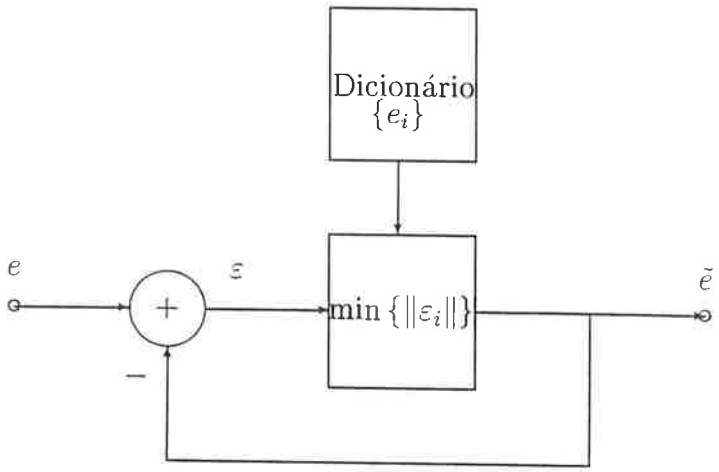

Figura 2.2: Quantizador escalar da excitação esquematizado em estilo vetorial.

Finalizando, convém detalhar o bloco do quantizador da Fig. 2.1 na Fig. 2.2. Como preparação para o tratamento da quantização vetorial da excitação usando a técnica da análise-mediante-síntese (seção 3.1), este esquema usa a designação "dicionário" para a "tabela de quantização".

2.6 Espectros de potência e predição linear

A redundância linear inerente ao sinal de voz pode ser representada tanto pelo seu espectro de potência quanto pela sua seqüência de autocorrelação.

Uma das primeiras medidas do espectro de potência de longo prazo da fala contínua foi obtida com um banco de filtros [20]. Nesse trabalho foram levantados espectros a intervalos de 125 ms, tomando-se sua média durante mais de um minuto de fala para dois conjuntos de 6 homens e 5 mulheres, respectivamente.

Os espectros de potência resultantes têm picos entre 250 e 500 Hz e acima desta freqüência decaem à razão de 8 a 10 db/oitava [52]. Estes dados globais podem ser estilizados por um filtro de segunda ordem com freqüência de corte de 500 Hz, que, a partir dela, decai a 12 dB/oitava ou mesmo por um filtro de primeira ordem, com decaimento de 6 db/oitava.

Efetivamente, filtros próximos dessa especificação foram usados como preditores em codificadores ADPCM ("Adaptive Differential Pulse Code Modulation") com preditor fixo e quantizador com degrau de quantização adaptativo (seção 2.5).

Em particular, os moduladores delta adaptativos CVSD ("continuously variable slope delta modulation") geralmente usam preditores fixos de 2ª ordem. Os moduladores delta adaptativos são codificadores ADPCM com quantizadores de dois níveis.

Porém, nos codificadores CVSD a adaptação do passo do quantizador é executada com período de 5 a 10 ms, que é da ordem do período fundamental ("pitch' ') [37].

Convencionou-se denominar este processo de adaptação de "compansão silábica".

A redundância do sinal de voz usada para essa adaptação do quantizador é a sua quase-periodicidade nos segmentos sonoros, que não comparece no espectro de longo prazo.

Quanto aos preditores fixos, o maior incremento de ganho de predição (seção 2.5) ocorre na passagem de nenhum preditor para um preditor de 1ª ordem. Na seqüência, obtêm-se ganhos incrementais cada vez menores até a ordem 4 ou 5 do preditor, a partir da qual eles cessam de ocorrer [52]. Apenas com o recurso a preditores adaptativos é possível atingir-se ganhos de predição típicos de 10 dB [39] até o máximo de 24 dB [37] com sinais de voz. Esses ganhos de predição só podem ser alcançados porque o espectro de curto prazo de blocos de duração da ordem de 20 ms do sinal de voz apresenta uma envoltória com um decaimento global de 12 db/oitava e com 4 ressonâncias ou formantes superpostas numa faixa de freqüências de 4 kHz.

A remoção dessa redundância pode ser efetuada por um preditor de ordem 10, em geral com adaptação a cada 20 ms, no caso da adaptação progressiva ("forward"). A adaptação progressiva consiste na determinação completa do preditor a partir de blocos do sinal de voz.

Esta adaptação é realizada inicialmente nos codificadores APC ("Adaptive Predictive Coding"), que, como propostos em [4], possuem quantizadores de 1 bit como os moduladores delta e duas etapas de predição: o preditor de curto prazo $P_2(z)$ (Fig. 2.3), correspondente ao preditor P da Fig. 2.1, e o preditor de longo prazo $P_1(z)$. O preditor $P_1(z)$ remove redundâncias de periodicidade, que se manisfestam na estrutura fina do espectro, e o preditor $P_2(z)$ remove redundâncias de formato da envoltória espectral.

O preditor de curto prazo $P_2(z)$ atua sobre o sinal residual $e(n)$ do preditor de longo prazo $P_1(z)$, porém, a ordem dos dois preditores pode ser invertida [26], resultando num codificador de propriedades bem diferentes [6].

Por outro lado, a adaptação regressiva ("backward") do preditor é efetuada a partir de amostras passadas do sinal reconstruído e não incorre em atraso algorítmico como a adaptação progressiva. Assim, a adaptação regressiva é apropriada para codificadores de voz de baixo atraso.

De fato, o codificador ADPCM G.726 [11] mantém um atraso nulo e atualiza incrementalmente os coeficientes do preditor a cada amostra por um método de gradiente denominado "least mean-absolute error (LMA)", que é uma versão do popular algoritmo "least mean square (LMS)" que opera com sinais algébricos [15].

Entretanto, o codificador LD-CELP G.728 [13] usa um filtro de síntese de ordem 50 em comparação com o correspondente do G.726 que tem apenas 2 pólos e 6 zeros. Dessa forma, torna-se muito complexo testar a estabilidade dos filtros resultantes da adaptação segundo os algoritmos LMA ou LMS. Assim, o G.728 implementa uma análise LPC regressiva, empregando um bloco de amostras passadas do sinal de voz reconstruído.

Claramente, para uma mesma taxa de quantização do sinal de excitação ou sinal residual, a adaptação regressiva consegue uma taxa total mais baixa pois não necessita transmitir os coeficientes do preditor como a adaptação progressiva.

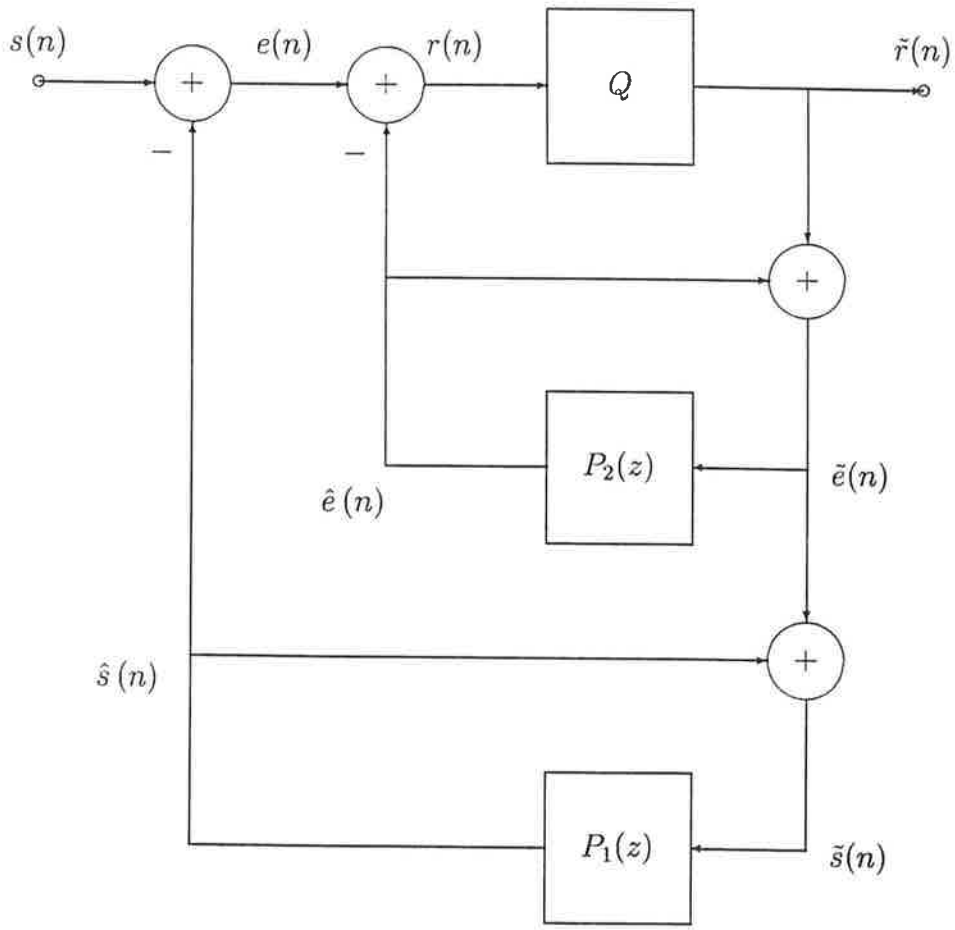

Figura 2.3: Codificador APC com preditor de longo prazo $P_1(z)$ e preditor de curto prazo $P_2(z)$.

Porém, o ganho de predição obtido com a adaptação regressiva é menor que o da adaptação progressiva e, quando a quantização da excitação atinge taxas inferiores a 1 bit/amostra [42], o sinal reconstruído distancia-se consideravelmente do original, tornando-se inadequado na tarefa de substituição.

Assim, como este trabalho aborda codificadores a taxas inferiores a 8 kHz, serão tratados apenas preditores com adaptação progressiva deste ponto em diante.

2.7 Predição linear e resposta impulsiva

Nesta seção será abordado um método alternativo de resolução das equações normais da predição linear pelo método da autocorrelação. Esta abordagem é mais robusta a erros numéricos [45], facilitando a determinação dos coeficientes do filtro de síntese na estrutura em treliça. Porém, a razão maior para o seu tratamento aqui é o vínculo que o algoritmo estabelece com a seqüência de resposta impulsiva do filtro de síntese, que é necessária no tratamento de transitórios causados por variações paramétricas (Capítulo 7).

Outro ponto que deve ser esclarecido é a conotação dos termos predição progressiva e regressiva, que serão usados no contexto de processos estocásticos estacionários. Isto é propiciado pelo uso do método da autocorrelação para a montagem das equações normais.

Um preditor linear de curto prazo de ordem p pode ser representado pela sua função de transferência

$$P(z) = -\sum_{j=1}^{p} a_j z^{-j}, \tag{2.22}$$

onde a ordem de predição será normalmente $p = 10$.

O sinal à saída do preditor é o sinal predito (Figs. 2.1 e 2.3)

$$\hat{s}(n) = -\sum_{j=1}^{p} a_j \tilde{s}(n-j). \tag{2.23}$$

O erro de predição é

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) \\ &= s(n) + \sum_{j=1}^{p} a_j \tilde{s}(n-j), \end{aligned} \tag{2.24}$$

que é o erro obtido em malha fechada por análise-mediante-síntese.

Entretanto, a análise preditiva (LPC) comumente efetuada opera em malha aberta. Operacionalmente, isto equivale a substituir as ocorrências $\tilde{s}(n-j)$ do sinal reconstruído na Eq. (2.24) pelas respectivas amostras $s(n-j)$ do sinal original de voz. Assim, obtém-se

$$e(n) = s(n) + \sum_{j=1}^{p} a_j s(n-j). \tag{2.25}$$

O erro de predição em malha aberta da Eq. (2.25) pode ser obtido à saída do filtro

$$A(z) = 1 + \sum_{j=1}^{p} a_j z^{-j} \tag{2.26}$$

alimentado pelo sinal de voz original. O filtro $A(z)$ é o filtro inverso da predição ou filtro de branqueamento.

A análise preditiva (LPC) normalmente determina o filtro de síntese $H_{\text{LPC}}(z)$ através dos coeficientes do denominador $A(z)$ da sua função de transferência. É o filtro de branqueamento associado à predição progressiva ("forward prediction"), que é dado pela Eq. (2.26), onde p é a ordem de predição.

Faz-se uma análise preditiva para cada bloco do sinal, cujas amostras são extraídas através de uma janela $w(n)$ como

$$s_l(n) = w(n).s(n + lL_B), \tag{2.27}$$

onde $l \geq 0$ é o número do bloco de análise e L_B é o espaçamento entre blocos consecutivos. Usa-se geralmente a janela de Hamming, cujo comprimento é $L_w = L_B$ para o caso em que não há sobreposição entre blocos consecutivos, sendo maior quando há sobreposição.

O desenvolvimento a ser feito nesta seção aplica-se apenas ao método da autocorrelação, que se baseia na seqüência de autocorrelação do sinal janelado $s_l(n)$, que é dada por

$$r(\lambda) = \sum_{n=-\infty}^{\infty} s_l(n) s_l(n + \lambda), \tag{2.28}$$

onde, em princípio, o índice λ varia de $-\infty$ a $+\infty$ com $r(\lambda)$ assumindo valores nulos para $|\lambda| > L_w$, sendo L_w o comprimento da janela, que pode ser maior que L_B. Além disso, para uma dada ordem de predição p, o algoritmo de estimação do preditor linear necessita apenas os coeficientes $r(\lambda)$ com índices na faixa $|\lambda| \leq p$.

No formalismo de espaço vetorial de polinômios que será usado, define-se o produto interno de dois polinômios como

$$\langle A(z), B(z) \rangle = \frac{1}{2\pi j} \oint_C A(z) R(z) B(z^{-1}) z^{-1} dz, \tag{2.29}$$

onde C representa qualquer curva fechada situada na região de convergência da função $A(z)R(z)B(z^{-1})$ e cujo interior contenha a origem do plano complexo z e $R(z)$ é a transformada z da seqüência de autocorrelação (Eq. (2.28)) do sinal de voz janelado.

O critério do mínimos quadrados conduz às equações normais, que podem ser escritas como

$$\langle A(z), z^{-j} \rangle = 0, \quad j = 1, 2, \ldots, p. \tag{2.30}$$

Conforme mostrado em [47] e abordado de forma condensada em apêndice de [2], as equações normais podem ser resolvidas mediante um método iterativo na ordem de predição p de tal forma que, ao final do processo, terão sido resolvidos todos os subsistemas normais de ordens m menores ou iguais a p.

Uma vez resolvido o sistema normal de ordem $m-1$, introduz-se um "coeficiente de combinação linear" k_m para expressar a solução do sistema de ordem m como uma combinação linear entre os preditores progressivo $A_{m-1}(z)$ e regressivo ("backward predictor") $B_{m-1}(z)$ de ordem $m-1$ como

$$A_m(z) = A_{m-1}(z) + k_m B_{m-1}(z). \tag{2.31}$$

onde, no caso do método da autocorrelação de predição linear em que temos uma matriz de autocorrelação com estrutura Toeplitz simétrica, o filtro de branqueamento da predição regressiva é expresso em função do filtro de branqueamento progressivo como

$$B_{m-1}(z) = z^{-m} A_{m-1}(1/z). \tag{2.32}$$

Tendo começado com o protótipo de solução (2.31), já foram resolvidas $m-1$ das m equações normais. Basta resolver-se a equação normal com a m-ésima potência de z^{-1}

$$\langle A_{m-1}(z) + k_m B_{m-1}(z), z^{-m} \rangle = 0.$$

que acaba fixando o valor do coeficiente k_m em

$$k_m = -\langle A_{m-1}(z), z^{-m} \rangle / \sigma^2_{m-1}, \tag{2.33}$$

onde σ^2_{m-1} é a variância do resíduo de predição de ordem $m-1$.

A variância do resíduo de predição de ordem m é o produto interno

$$\sigma^2_m = \langle A_m(z), 1 \rangle = \langle A_{m-1}(z), 1 \rangle + k_m \langle B_{m-1}, 1 \rangle. \tag{2.34}$$

Dada a forma como o sinal foi segmentado e como foi obtida a seqüência de autocorrelação (Eq. (2.28)), a matriz de autocorrelação gerada tem estrutura Toeplitz simétrica (Eq. (2.32)). Assim, pode-se usar o algoritmo de Levinson-Durbin ou o algoritmo de Le Roux-Gueguen [45], dependendo da escolha dos parâmetros. O algoritmo de Le Roux-Gueguen também é conhecido como algoritmo de Schur [64], que o derivou anteriormente em outro contexto.

No caso do algoritmo de Levinson-Durbin, os parâmetros são os coeficientes de predição $a^m_j, j = 1, 2, \ldots, m$, e k_m, $m = 1, 2, \ldots, p$, onde

$$A_m(z) = 1 + \sum_{j=1}^{m} a^m_j z^{-j}.$$

A cada iteração, o algoritmo de Durbin

i) usa a Eq. (2.33) para calcular k_m;

ii) usa a Eq. (2.31) para calcular $a^m_j, i = 1, 2, \ldots, m$;

iii) usa a Eq. (2.34) para calcular o resíduo.

Por sua vez, no caso do algoritmo de Le Roux-Gueguen, os parâmetros são k_m, $m = 1, 2, \ldots, p$ e

$$e_i^m = \langle A_m(z), z^{-i} \rangle, \tag{2.35}$$

para $i = I, I+1, \ldots, 0, J, J+1, \ldots, K$, onde

$$I = \min\{-(p-m-1), 0\} \tag{2.36}$$

e

$$\begin{cases} m+1 \leq p \rightarrow \begin{cases} J = m+1 \\ K = p \end{cases} \\ m+1 > p \rightarrow \begin{cases} J = 0 \\ K = 0 \end{cases} \end{cases} \tag{2.37}$$

Considerando que, aplicando $A_m(z)$ ao sinal de voz $s(n)$, obtém-se o sinal residual de predição $u_m(n)$ e que, ao aplicar-se z^{-i}, tem-se o sinal atrasado $s(n-i)$, os parâmetros e_i^m podem ser interpretados como seqüência de correlação cruzada entre o sinal residual da predição de ordem m e o sinal de voz, ou seja,

$$e_i^m = \mathrm{E}\left[u_m(n) s_l(n-i)\right]. \tag{2.38}$$

Introduzindo a Eq. (2.31) em (2.35) e expandindo, obtém-se

$$e_i^m = \underbrace{\langle A_{m-1}(z), z^{-i} \rangle}_{e_i^{m-1}} + k_i \langle B_{m-1}(z), z^{-i} \rangle \tag{2.39}$$

Levando em conta a relação (2.32) de simetria especular de coeficientes entre os preditores regressivos e progressivos, pode-se reescrever o segundo produto interno em (2.39) da forma seguinte

$$\langle B_{m-1}(z), z^{-i} \rangle = \langle z^{-m} A_{m-1}(z^{-1}), z^{-i} \rangle. \tag{2.40}$$

Considerando a estrutura Toeplitz da matriz de autocorrelação do sinal em análise e, em seguida, a sua simetria, seguem as igualdades

$$\begin{aligned} \langle z^{-(m-j)}, z^{-i} \rangle &= \langle z^{j}, z^{-(-m+i)} \rangle \\ &\quad \langle z^{-j}, z^{-(m-i)} \rangle \end{aligned} \tag{2.41}$$

Aplicando em (2.40) o conjunto de relações obtidas de (2.41) com $j = 0, 1, \ldots, m-1$, obtém-se a relação

$$\langle B_{m-1}(z), z^{-i} \rangle = \underbrace{\langle A_{m-1}(z), z^{-(m-i)} \rangle}_{e_{m-i}^{m-1}}. \tag{2.42}$$

Portanto, resulta possível expressar a Eq. (2.39) de forma recorrente para o cálculo das correlações cruzadas da ordem atual em função das antecedentes como

$$e_i^m = e_i^{m-1} + k_m e_{m-i}^{m-1} \tag{2.43}$$

para $i = I, I+1, \ldots, 0, J, J+1, \ldots, K$.

A inicialização das variáveis do algoritmo de Le Roux-Gueguen é feita com os coeficientes de autocorrelação já identificados na Equação (2.28) ou na Equação (2.35) tomando $m = 0$. Explicitamente, tem-se

$$\begin{aligned} e_i^0 &= \langle 1, z^{-i} \rangle \\ &= r(i) \\ &= r(|i|), \end{aligned}$$

para $i = -(p-1), -(p-2), \ldots, -1, 0, 1, \ldots, p$.

Por exemplo, para a predição linear de ordem $p = 4$, o algoritmo de Le Roux-Gueguen por extenso fica

$$\begin{array}{llll} k_1 = -\dfrac{e_1^0}{e_0^0} & k_2 = -\dfrac{e_2^1}{e_0^1} & k_3 = -\dfrac{e_3^2}{e_0^2} & k_4 = -\dfrac{e_4^3}{e_0^3} \\ e_0^1 = e_0^0 + k_1 e_1^0 & e_0^2 = e_0^1 + k_2 e_2^1 & e_0^3 = e_0^2 + k_3 e_3^2 & e_0^4 = e_0^3 + k_4 e_4^3 \\ & & & \\ e_2^1 = e_2^0 + k_1 e_{-1}^0 & e_3^2 = e_3^1 + k_2 e_{-1}^1 & e_4^3 = e_4^2 + k_3 e_{-1}^2 & \\ e_{-1}^1 = e_{-1}^0 + k_1 e_2^0 & e_{-1}^2 = e_{-1}^1 + k_2 e_3^1 & & \\ & & & \\ e_3^1 = e_3^0 + k_1 e_{-2}^0 & e_4^2 = e_4^1 + k_2 e_{-2}^1 & & \\ e_{-2}^1 = e_{-2}^0 + k_1 e_3^0 & & & \\ & & & \\ e_4^1 = e_4^0 + k_1 e_{-3}^0 & & & \end{array}$$

Em particular, os coeficientes de correlação cruzada e_i^m com deslocamentos $i > 0$ representam a redundância linear restante após a predição de ordem m enquanto aqueles com deslocamento $i \leq 0$ aproximam a resposta impulsiva do filtro $1/A_m(z)$. Eles coincidem com essa resposta impulsiva se o sinal tiver sido gerado por um modelo auto-regressivo AR(m) perfeito.

Para poder calcular a resposta impulsiva e_i^p com $i = 0, -1, \ldots, -(p-1)$ na última iteração do algoritmo de Le Roux-Gueguen, é necessário iniciar com o dobro dos coeficientes de autocorrelação do sinal e continuar efetuando os cálculos a cada iteração sobre um intervalo dobrado de índices. Isto é, o conjunto de parâmetros envolvidos é e_i^m, com $i = -(2p-m-1), -(p-m-2), \ldots, 0, m+1, m+2, \ldots, 2p$ e $m = 0, 1, \ldots, p$.

Observa-se, ainda, que a correlação cruzada e_0^m com deslocamento nulo coincide com a variância do sinal residual após a predição de ordem m, conforme desenvolvimento abaixo.

Pela definição (2.35), para $i = 0$, tem-se

$$e_0^m = \langle A_m(z), 1 \rangle. \tag{2.44}$$

Porém, como $A_m(z) = 1 + \sum_{j=1}^m a_j^m z^{-j}$ é a solução do sistema de equações normais de ordem m, redimensionando as Equações (2.30), obtém-se

$$\langle A_j(z), z^{-j} \rangle = 0$$

para $j = 1, 2, \ldots, m$, que, por combinação linear, fornecem

$$\sum_{j=1}^{m} a_j^m \langle A_j(z), z^{-j} \rangle = 0,$$

e, de forma equivalente,

$$0 = \sum_{j=1}^{m} \langle A_j(z), a_j^m z^{-j} \rangle. \tag{2.45}$$

Finalmente, adicionando membro a membro a Eq. (2.45) à Eq. (2.44) e associando os dois produtos internos, obtém-se

$$e_0^m = \langle A_m(z), A_m(z) \rangle, \tag{2.46}$$

que é precisamente a variância do sinal residual $u_m(n) = 1 + \sum_{j=1}^{m} a_j^m s_l(n-j)$ remanescente após a predição de ordem m.

2.8 Codificadores telefônicos

O desempenho da grande maioria dos codificadores de voz para telefonia digital encontra-se representado na Figura 2.4, extraída de [16], que está baseada em vários ensaios subjetivos e não em um único ensaio. Está apresentada aqui apenas para comparações aproximadas de qualidade relativa. Este é o desempenho para voz sem ruído de fundo. GSM, JDC e IS-XXX são padrões celulares regionais da Europa, Japão e América do Norte, respectivamente. FS são normas federais dos EUA e G.XXX são recomendações da ITU-T (International Telecommunication Union - Telecommunication Standardization Sector). Por exemplo, G.711 representa o codificador log PCM a 64 kbit/s (seção 2.4) e ambos G.726 e G.727 representam o codificador ADPCM a 32 kbit/s (seção 2.6). A diferença entre as recomendações G.726 e G.727 situa-se no tipo de equipamento de multiplex a que se destinam.

No biênio de 1995-1996 surgiram novos padrões internacionais (ITU-T G.729, G.729A e G.723.1) e novos padrões regionais (codificadores melhorados à taxa completa para os sistemas móveis europeu, GSM-EFR, e da América do Norte, IS-641).

O codificador G.723.1 [32], que possui duas taxas de operação, 5,3 e 6,3 kbit/s, comparecendo com dois pontos na Fig. 2.4, aplica-se a comunicações multimídia, tendo sido desenvolvido originalmente para videofones a baixa taxa de transmissão. O codificador à taxa mais baixa é um CELP algébrico (ACELP) e o codificador à taxa alta é um multipulsos com quantização de máxima verossimilhança (MP-MLQ), sendo possível alternar entre as duas taxas a cada bloco de 30 ms. O atraso algorítmico deste codec é de 37,5 ms.

O codificador G.729 [33] é um ACELP com estrutura conjugada (CS-ACELP) que opera à taxa de 8 kbit/s com atraso algorítmico de 15 ms e com qualidade de voz suficientemente alta para uso na rede telefônica fixa (convencional). Embora tenha sido originalmente projetado para aplicações móveis, aplica-se também a comunicações multimídia.

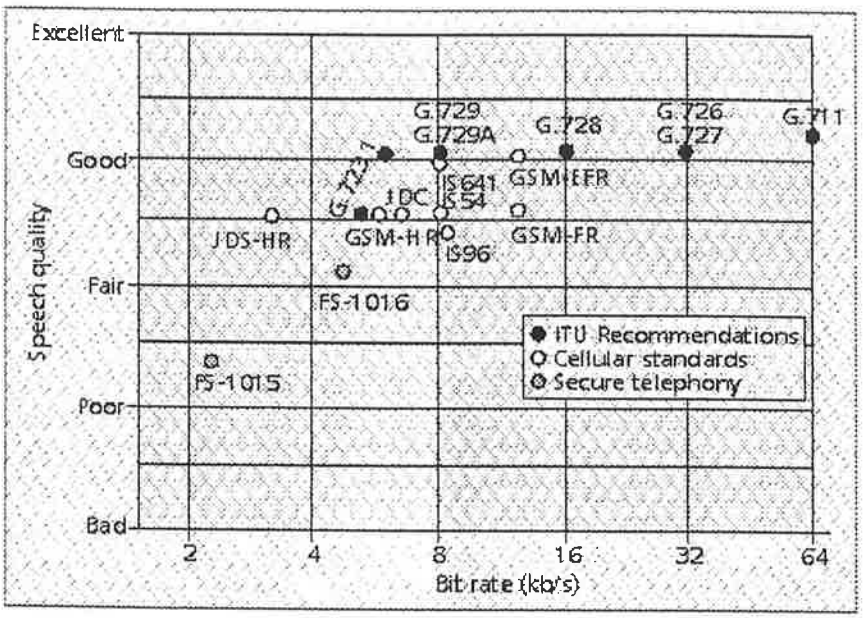

Figura 2.4: Comparação de qualidade entre diferentes codificadores padronizados.

O codificador do Anexo A da recomendação G.729 é uma versão computacionalmente mais eficiente do codificador CS-ACELP [34]. Em processadores digitais de sinais de ponto fixo, a complexidade do G.729A é de 10,5 MIPS em comparação com os 20 MIPS do G.729. Aproximadamente 5 MIPS de redução de complexidade são atribuídos a um novo método de busca das inovações no dicionário algébrico [61].

O codificador IS-54 foi projetado para a telefonia celular digital TDMA ("time division multiple access") da América do Norte, apresentando qualidade de voz insuficiente para a rede telefônica fixa. Ele é um codificador CELP com excitação por soma vetorial (VSELP) que opera à taxa de 8 kbit/s.

Para atender a demanda por melhor qualidade de voz, na norma IS-136 de telefonia celular TDMA foi incluído o codificador melhorado à taxa completa de 8 kbit/s, definido pela norma IS-641. Ele é um codificador ACELP nos moldes do G.729, porém com uma busca fixa de inovações mais eficiente [30].

Capítulo 3

CODIFICADORES CELP

Os codificadores de voz por análise-mediante- síntese baseados na predição linear (LPAS) constituem uma vasta classe. A primeira apresentação formal de um codificador típico desta classe foi feita por Schroeder e Atal [62] *apud* Kroon [41], que denominaram seu processo de codificação de *"Code-excited linear prediction"* (CELP).

Os codificadores preditivos necessitam de um sinal de excitação para poder gerar um sinal de voz reconstruído à saída do filtro de síntese.

Em particular, os codificadores CELP selecionam a intervalos regulares vetores de um conjunto definido para compor o sinal de excitação. A denominação CELP origina-se na imagem de um dicionário ("codebook") para conter esses vetores ou códigos.

Inicialmente, os códigos eram amostras de um processo de ruído branco de média nula e variância unitária. Em seguida, percebeu-se que eles poderiam ter formas diversas. Neste aspecto, uma boa definição globalizante desta classe de codificadores aparece em [58].

Porém, devido à imprecisão carregada pelo termo CELP ao longo de sua existência, Kroon e Kleijn [42] preferem o uso da denominação LPAS.

Entretanto, aqui será usado o termo CELP por brevidade de referência e por popularidade também.

3.1 Análise-mediante-síntese

O princípio da análise-mediante-síntese já foi visto na quantização escalar diferencial (seção 2.5). No caso dos codificadores CELP, toma-se apenas uma decisão por sub-bloco sobre a determinação de todo o conjunto de amostras de excitação para o filtro de síntese.

Efetivamente, dada a natureza vetorial da excitação CELP, é mais natural considerar o filtro de síntese como um bloco que engloba o preditor e abrir o bloco do quantizador (Fig. 2.2) para transformar o diagrama do codificador da Fig. 2.1 na Fig. 3.1, conforme sugestão em [39]. Além disso, a malha de realimentação do quantizador, que se fecha no comparador na Fig. 2.2, percorre o filtro de síntese na Fig. 3.1.

Ademais, essa configuração ressalta o sintetizador que está encerrado no analisador. Rigorosamente, ao subsistema delimitado na Fig. 3.1 por linhas tracejadas apenas falta o mecanismo de obtenção do vetor de excitação $\tilde{e}$ para se tornar no decodificador. Esse mecanismo é, essencialmente, o acesso a um vetor pelo índice numa tabela de vetores.

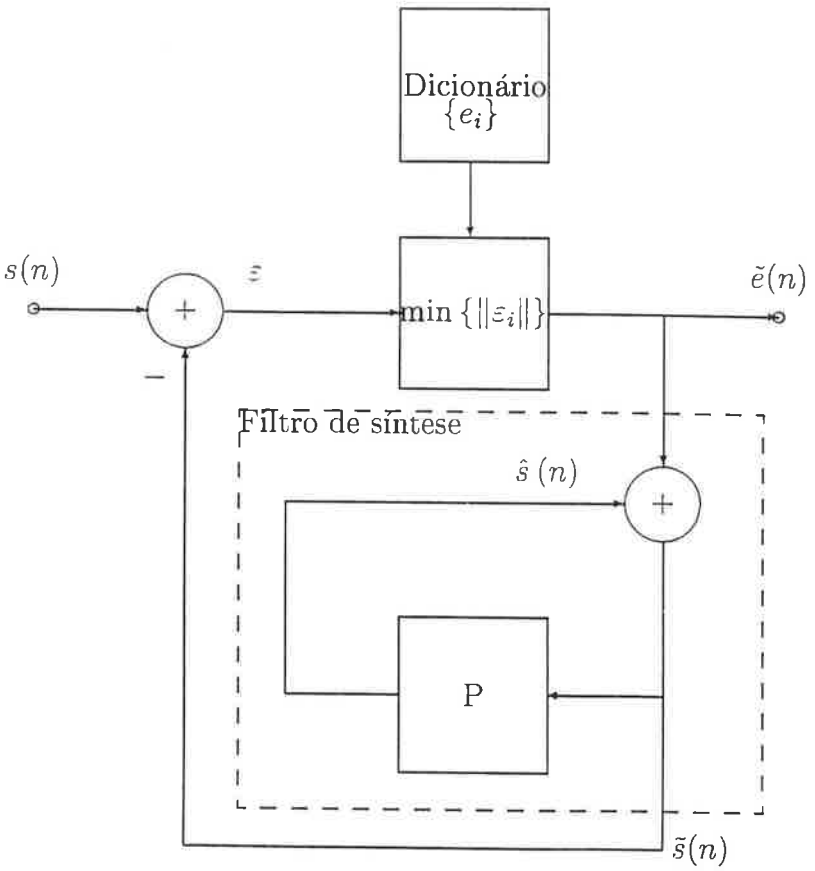

Figura 3.1: Codificador preditivo com quantizador vetorial.

Nesta configuração, torna-se evidente que o erro que interessa ao quantizador vetorial da excitação é o erro de reconstrução e não o erro de predição somente, como indicado na Fig. 2.1 para o quantizador escalar.

3.2 A segmentação do sinal de voz

Consideram-se três segmentações do sinal de voz na sua análise e síntese [42]

- janela de análise preditiva,
- bloco de filtragem ("frame"),
- sub-bloco de excitação ("subframe").

O bloco de filtragem representa o conjunto de amostras reconstruídas sob uma mesma configuração de coeficientes do filtro de síntese. A configuração de coeficientes do filtro é calculada com as amostras obtidas através da janela preditiva. O bloco de filtragem subdivide-se em sub-blocos, em geral em número de quatro, para cada qual sendo buscado um vetor de excitação.

No caso geral em que há interpolação das representações consecutivas de coeficientes do filtro, a configuração do filtro de síntese associada a um bloco é o alvo a ser atingido no fim do bloco. O comprimento do bloco de filtragem é de 160 e 240 amostras no VSELP [21] e no ACELP/MP-MLQ [32], respectivamente.

A janela de análise preditiva é o conjunto de amostras que são extraídas do sinal de voz com periodicidade igual ao comprimento do bloco de filtragem, englobando coeficientes de ponderação $w(n)$ (seção 2.7). No caso de coeficientes iguais tem-se a janela retangular. Geralmente, a janela de análise estende-se além da fronteira posterior do bloco a que está associada. Este "avanço no futuro" é benéfico ao processo de interpolação de representações de conjuntos consecutivos de coeficientes do filtro de síntese.

Como exemplo, a janela de análise do codificador ACELP/MP-MLQ da recomendação G.723.1 tem a forma de Hamming e tem um comprimento de 180 amostras centradas no meio das últimas 60 de cada bloco de filtragem.

Por último, os sub-blocos de excitação são subdivisões do bloco de síntese que possuem comprimentos iguais à dimensão do dicionário de códigos. Em correspondência, o sinal de voz original também tem que sofrer o mesmo processo de segmentação para poder servir de referência durante o processo de busca do melhor vetor-código a cada sub-bloco.

3.3 Composição dos codificadores CELP

Nesta seção elabora-se uma representação variante no tempo do codificador CELP inicialmente apresentada pelo autor em [2].

Os codificadores CELP compõem-se de dois subsistemas:

- um filtro $H_{\mathrm{LPC}}(z)$, que modela a envoltória do espectro de curto prazo do sinal de voz;
- um gerador da excitação $e(n)$, que modela a estrutura fina do espectro de curto prazo do sinal de voz.

Os parâmetros do filtro $H_{\mathrm{LPC}}(z)$ são determinados a partir da *análise preditiva* ou *"linear predictive coding"* (LPC) (seção 2.7) de segmentos de centésimos de segundo de duração, representados por blocos de N amostras do sinal de voz (seção 3.2).

A seqüência de excitação é gerada em sub-blocos de comprimento L, menor que os blocos de filtragem.

Os segmentos da seqüência de excitação são especificados numa representação forma-ganho [41], onde a forma é dada por vetores-código

$$c_i(n), \quad n = 0, \ldots, L-1$$

e o ganho g é um fator de escala (Fig. 3.2). As formas de onda normalizadas e os ganhos são quantizados independentemente num processo subótimo, porém, de complexidade mais tratável.

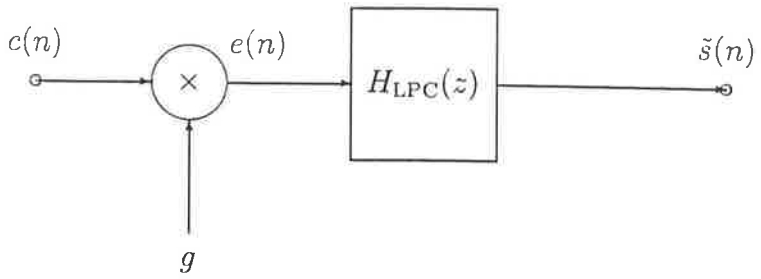

Figura 3.2: Síntese CELP de um segmento de voz.

Os vetores-código são escolhidos de um conjunto

$$\{c_i(n),\ \ i = 1, \ldots, M\},$$

denominado dicionário de códigos ("codebook") através de *análise mediante síntese.*

Para a caracterização completa do processo de síntese CELP (Fig. 3.2), convém considerar

- a seqüência de excitação $c(n)$, $n = 0, 1, \ldots$, formada pela concatenação dos vetores-código $c_{i(m)}$ selecionados a cada sub-bloco m de excitação;
- a seqüência de ganhos da excitação $g(n)$, $n = 0, 1, \ldots$. definida a cada instante n pelo ganho g_m determinado para o sub-bloco m atual;
- o filtro linear e variante no tempo

$$H_{\mathrm{LPC},l}(z) = \frac{1}{1 + \sum_{k=1}^{p} a_{kl} z^{-k}}, \tag{3.1}$$

  onde a_{kl}, $k = 1, \ldots, p$, são os coeficientes de predição linear para o bloco l de análise LPC;
- a seqüência de voz sintética $\tilde{s}(n)$, $n = 0, 1, \ldots$, formada pela concatenação dos vetores $\tilde{s}_m$ reconstruídos a cada sub-bloco m de excitação.

Há dois tipos de transições na síntese CELP:

- transições de bloco com a mudança dos coeficientes do filtro;
- transições de sub-bloco com a mudança do ganho e do vetor de excitação.

Para assegurar uma certa suavidade ao processo de filtragem, na passagem pelos dois tipos de transições mantém-se o estado do filtro $H_{\mathrm{LPC},l}(z)$.

Seja $x_m(n)$ o vetor variável que contém os p estados do filtro $x_m^{(k)}(n)$, $k = 1, \ldots, p$ no instante n do sub-bloco m. Serão de interesse três fases do sub-bloco m, cujos estados serão representados por

$$\begin{cases} x_m(0) & \text{as condições iniciais para o sub-bloco } m, \\ x_m(n), \quad n = 1, \ldots, L-1 & \text{os estados ao longo do sub-bloco } m, \\ x_m(L) & \text{os estados ao fim do sub-bloco } m, \end{cases}$$

e impõe-se a cada transição de sub-bloco a condição

$$x_m(0) = x_{m-1}(L). \tag{3.2}$$

Em seguida, apresenta-se uma das possíveis representações de estados do filtro LPC, que descreve completamente o processo de filtragem.

Denotando a seqüência de vetores de estado do filtro $H_{\mathrm{LPC},l}(z)$ por $x(n)$, para $n \geq 0$, pode-se, a cada instante n, obter o estado seguinte $x(n+1)$ a partir do estado atual $x(n)$ e da entrada atual $g(n)c(n)$ através de

$$x(n+1) = A_l x(n) + b g(n) c(n), \tag{3.3}$$

onde

- a matriz de transição de estados do filtro $H_{\mathrm{LPC},l}(z)$, implementado na forma direta I, durante o bloco l é

$$A_l = \begin{bmatrix} -a_{1l} & -a_{2l} & \cdots & -a_{p-1,l} & -a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}; \tag{3.4}$$

- o vetor de coeficientes da entrada é b, dado por

$$b(k) = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$$

Completa-se a representação de estados do filtro LPC com a equação amostral da seqüência de saída

$$\tilde{s}(n) = x^{(1)}(n), \tag{3.5}$$

onde $x^{(1)}(n)$ é o valor da 1ª das variáveis de estado $x^{(k)}(n)$, $k = 1, \ldots, p$, do filtro LPC no instante n.

3.4 Efeitos perceptivos na audição

Em testes de percepção, foram caracterizados os seguintes efeitos de mascaramento auditivo [5]:

- nas faixas espectrais das formantes, o ruído de quantização é parcial ou totalmente encoberto pelo sinal de voz;
- o ouvido tolera mais ruído durante os sons de transição quando a largura de faixa das formantes é maior.

Mais quantitativamente, para sons vocálicos, o ruído de quantização é audível em todas as faixas de freqüência em que a sua envoltória espectral estiver menos do que 3 dB abaixo da envoltória espectral do sinal de voz [63].

Conforme foi visto para a análise LPC (seção 2.7), o critério de minimização do erro quadrático total tende a produzir um sinal de erro com espectro uniforme. De uma forma geral, o erro de quantização de outros preditores que adotam um critério de minimização quadrático também possui espectro plano [5].

Para codificadores CELP, pode-se manter a análise LPC com o seu erro de quantização de espectro plano e procurar compensar o perfil espectral do erro de reconstrução aplicando às regiões das formantes pesos menores no cálculo do erro total associado a cada vetor-código do dicionário.

3.5 Ponderação espectral perceptiva

Deve-se efetuar uma remodelagem do espectro do ruído de quantização ("quantization noise shaping") para aproveitar os efeitos de mascaramento auditivo no domínio da freqüência (seção 2.1).

Pretende-se alterar o espectro da seqüência de erro pela inserção de um filtro antes do cálculo do erro quadrático (Fig. 3.3). Seja $W(z)$ a função de transferência desse filtro de ponderação espectral perceptiva ("perceptual weighting filter").

Para impedir a passagem do erro de reconstrução na faixa das formantes, $W(e^{j\omega})$ deve ter atenuação máxima nessas regiões espectrais (seção 3.4). Portanto, o filtro LPC inverso é um bom candidato a filtro de ponderação espectral.

Entretanto, seria desejável que $W(z)$ tivesse algum parâmetro para acentuar a passagem das formantes caso o filtro LPC inverso as atenue em excesso.

Considerando os dois últimos parágrafos e o efeito de atenuação de ressonâncias pela expansão de faixa, costuma-se dar ao filtro de ponderação espectral ajustável a forma

$$W(z) = \frac{H_{\mathrm{LPC}}(z/\gamma_2)}{H_{\mathrm{LPC}}(z/\gamma_1)}. \tag{3.6}$$

A expansão de faixa associada a um dado fator de expansão γ é [2]

$$\Delta B = -(f_a/\pi)\ln\gamma, \tag{3.7}$$

onde γ é o coeficiente de expansão de faixa e f_a é a freqüência de amostragem do sinal de voz.

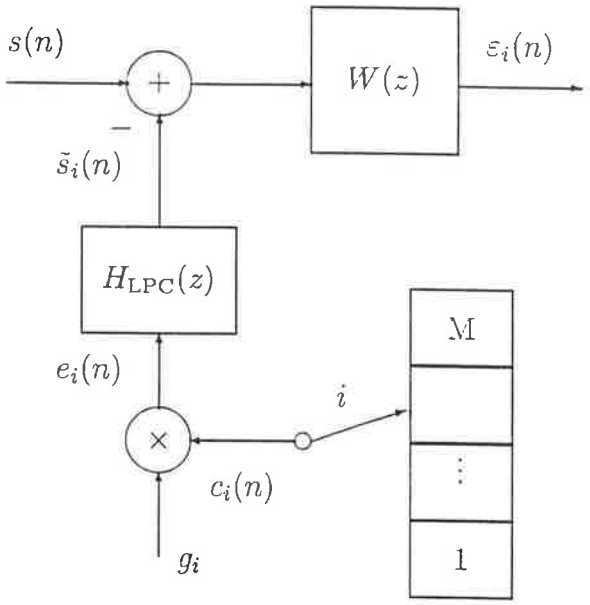

Figura 3.3: Busca com erro ponderado perceptivamente dentre M vetores-código.

Tome-se $\gamma_2 = 0,8$ e $\gamma_1 = 1$, por exemplo. O valor de γ_2 acarreta uma expansão $\Delta B = 568$ Hz, conforme o módulo da resposta em freqüência do filtro expandido à Fig. 3.4. O filtro de ponderação espectral está representado na Fig. 3.5, tornando clara a tolerância de erros maiores nas formantes.

Em outros casos, para manter mais rigorosamente uma curva de ponderação espectral, pode-se fazer $\gamma_1 \neq 1$. Por exemplo, para o codificador G.723.1 [32], $\gamma_2 = 0,5$ e $\gamma_1 = 0,9$.

Retornando à aplicação do filtro de ponderação, o esquema de busca da Fig. 3.3 demanda a execução de duas filtragens por vetor-código. Pode-se prescindir de uma delas no caso em que $\gamma_1 = 1$ com o deslocamento do filtro $W(z)$ para a entrada do somador de comparação, como indicado na Fig. 3.6, onde $y(n)$ é o segmento de voz ponderada e $H(z) = H_{\mathrm{LPC}}(z).W(z)$ é o filtro de síntese ponderado.

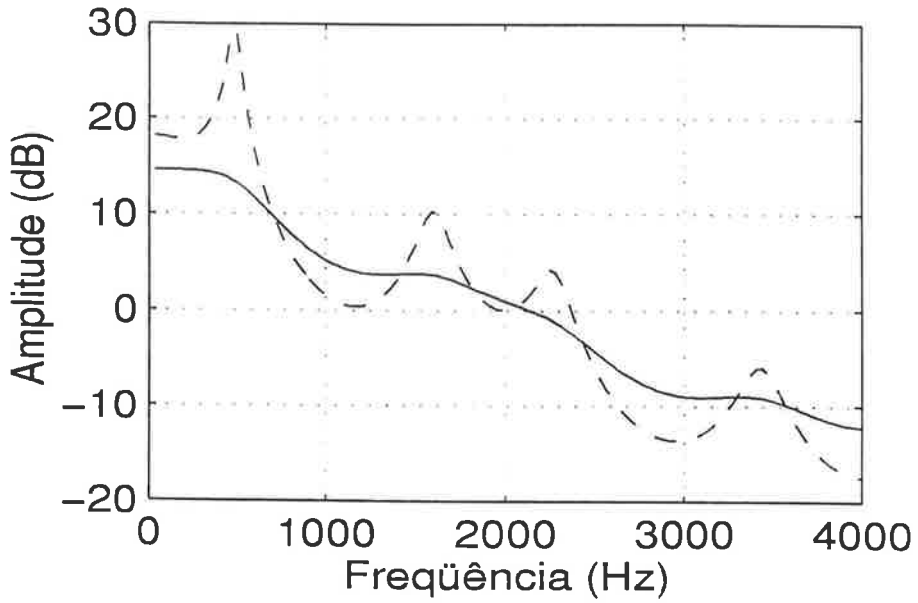

Figura 3.4: Resposta em freqüência de um filtro de síntese LPC (curva tracejada) e seu filtro expandido (curva contínua) com coeficiente de expansão $\gamma = 0,8$.

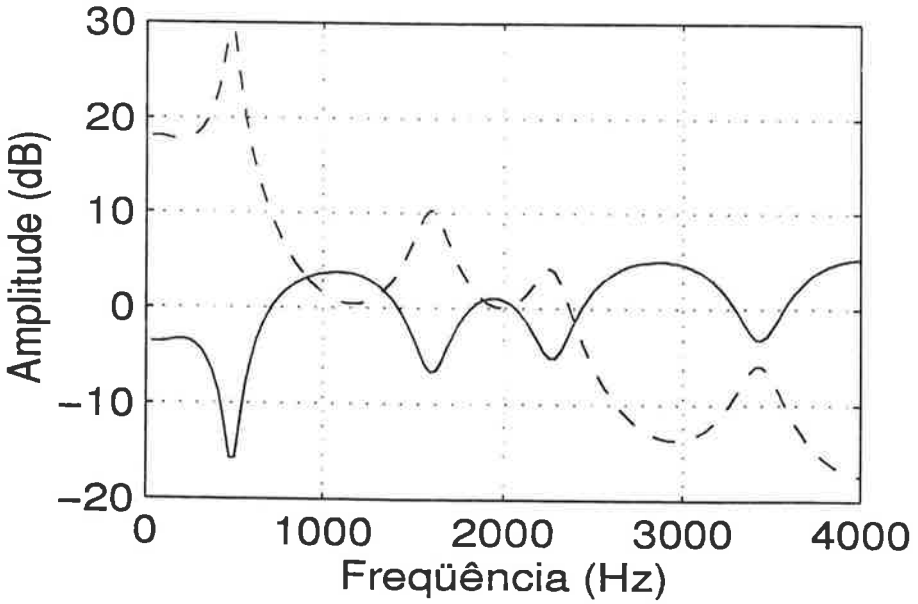

Figura 3.5: Resposta em freqüência de um filtro de síntese LPC (curva tracejada) e seu filtro de ponderação (curva contínua) com coeficientes de expansão $\gamma_1 = 1$ e $\gamma_2 = 0,8$.

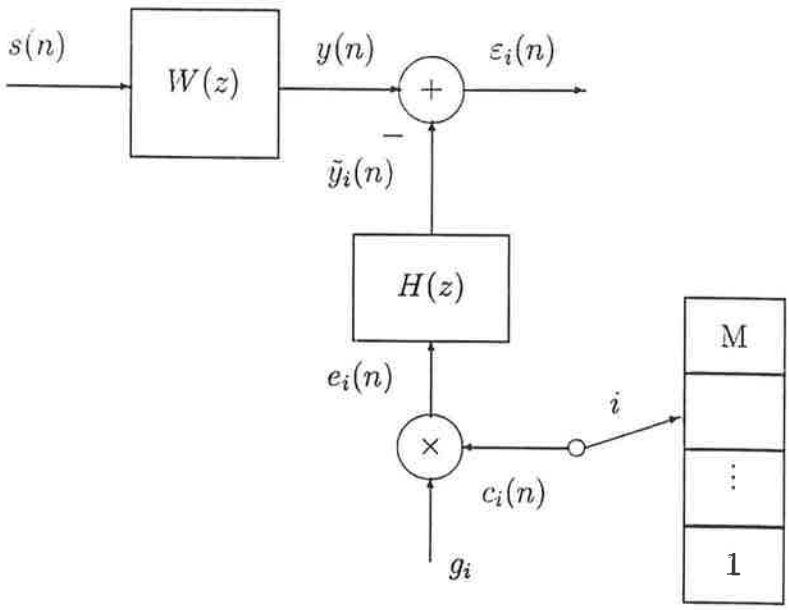

Figura 3.6: Busca com comparação no domínio ponderado dentre M vetores-código.

3.6 Matriz de resposta impulsiva

Dado o filtro de síntese ponderado

$$H(z) = H_{\mathrm{LPC}}(z).W(z), \tag{3.8}$$

seja $h(n)$, para $n \geq 0$, sua seqüência de resposta impulsiva com condições iniciais nulas.

Tome-se o vetor de resposta impulsiva do filtro $H(z)$ no sub-bloco de duração L como

$$h = [h(0)\ h(1)\ \ldots\ h(L-1)]^T. \tag{3.9}$$

que é a sua seqüência de resposta impulsiva truncada às primeiras L amostras.

A matriz de resposta impulsiva do filtro $H(z)$ é uma matriz Toeplitz triangular inferior que contém o seu vetor de resposta impulsiva na 1ª coluna:

$$H = \begin{bmatrix} h(0) & 0 & \cdots & 0 & 0 \\ h(1) & h(0) & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ h(L-2) & \ddots & h(1) & h(0) & 0 \\ h(L-1) & h(L-2) & \cdots & h(1) & h(0) \end{bmatrix}. \tag{3.10}$$

Adicionalmente, quando $H(z)$ é um filtro de síntese ponderado na forma

$$H(z) = \frac{1}{A(z/\gamma_2)}, \tag{3.11}$$

com

$$A(z/\gamma_2) = 1 + \sum_{k=1}^{p} a_k \gamma_2^k z^{-k},$$

tem-se

$$h(0) = 1.$$

A matriz de resposta impulsiva permite escrever a resposta total do filtro expandido $H(z)$ num sub-bloco como

$$\tilde{y}_i = He_i + y_0, \tag{3.12}$$

onde

- H é a matriz de resposta impulsiva do filtro $H(z)$ com $x_m(0) = 0$;
- y_0 é a resposta à excitação nula de $H(z)$ truncada a um sub-bloco com $x_m(0) = x_{m-1}(L)$;

sendo

$$\begin{cases} x_m(0) & \text{as condições iniciais para o sub-bloco } m, \\ x_m(n), \quad n = 1, \ldots, L-1 & \text{os estados ao longo do sub-bloco } m, \\ x_m(L) & \text{os estados ao fim do sub-bloco } m. \end{cases}$$

Note-se que o uso da matriz de resposta impulsiva H (Eq. (3.10)) não acarreta nenhuma imprecisão numérica porque cada vetor-código e_i na Eq. (3.12) tem dimensão que coincide exatamente com o número de colunas da matriz H assim como todas as amostras de cada vetor-alvo reconstruído $\tilde{y}_i$ podem ser calculadas porque a sua dimensão coincide com o número de linhas da matriz H.

Durante o processo de busca (Fig. 3.6), o erro de reconstrução é dado por

$$\varepsilon_i = y - \tilde{y}_i. \tag{3.13}$$

Das equações (3.12) e (3.13), tem-se

$$\varepsilon_i = y - y_0 - He_i. \tag{3.14}$$

Definindo o vetor-alvo da busca como

$$u = y - y_0 \tag{3.15}$$

e a sua reconstrução como

$$\tilde{u}_i = He_i, \tag{3.16}$$

pode-se escrever trivialmente o erro de reconstrução como

$$\varepsilon_i = u - \tilde{u}_i. \tag{3.17}$$

A busca do vetor-alvo u, obtido após a subtração da descarga y_0 do sub-bloco de voz ponderada y, está ilustrada na Fig. 3.7, onde $H_C(z)$ denota o filtro de síntese ponderado $H(z)$ com estado inicial $x_m(0) = x_{m-1}(L)$ e $H_D(z)$ indica o filtro $H(z)$ com estado inicial nulo.

Nos casos em que o fator de expansão $\gamma_1 \neq 1$ para o filtro de ponderação $W(z)$, a função de transferência do filtro de síntese ponderado não pode ser simplificada para a forma da Eq. (3.11) como tratado na seção 3.5.

Adicionalmente, para o codificador G.723.1, já mencionado a este respeito na seção 3.5, em cascata com o filtro de síntese ponderado da Eq. (3.8), comparece ainda um filtro $P(z)$ de modelagem harmônica do ruído ("harmonic noise shaping") [32], resultando num filtro de síntese ponderado combinado para a análise em malha fechada

$$H_S(z) = H_{\mathrm{LPC}}(z).W(z).P(z). \tag{3.18}$$

Porém, como a busca se restringe ao sub-bloco em questão, sem perda de precisão, pode-se calcular a resposta impulsiva truncada no comprimento de um sub-bloco do filtro $H_S(z)$ na Eq. (3.18) e, daí, retoma-se a Eq. (3.12) do desenvolvimento acima sem mais alterações.

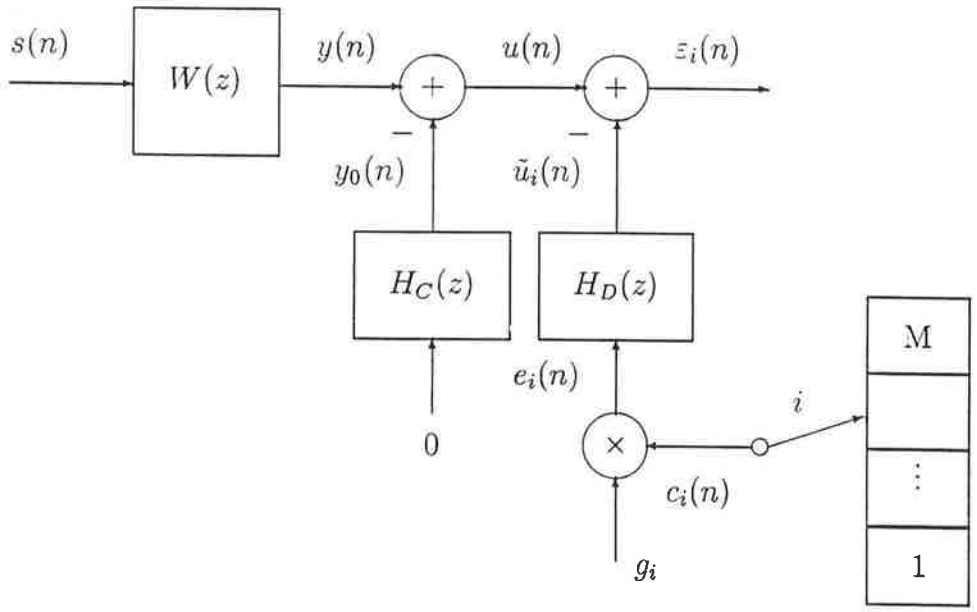

Figura 3.7: Modelo de busca da excitação para um codificador CELP com dicionário fixo único contendo M vetores-código.

3.7 Dicionário estocástico e complexidade de busca

O dicionário de códigos proposto originalmente por Atal e Schroeder é constituído por $M = 1024$ vetores-código aleatórios de dimensão $L = 40$ com amostras independentes e identicamente distribuídas com distribuição gaussiana de média $\mu = 0$ e de variância $\sigma^2 = 1$. Esse dicionário contém as formas de onda de energia unitária admissíveis. Um dicionário escalar de ganhos completa a quantização da forma de onda de energia arbitrária necessária para poder representar o sub-bloco de sinal de voz.

Com esses dicionários não estruturados a busca do melhor vetor-código tinha que ser exaustiva sobre o conjunto de M formas de onda normalizadas. Dessa forma, nos primórdios da codificação CELP, eram necessárias mais de 320 megaoperações de multiplicação e acumulação por segundo de fala [68].

Atualmente, graças aos avanços algorítmicos nos processos de busca e à engenhosidade na estruturação dos dicionários de códigos, os codificadores que necessitam 30 MIPS (megainstruções por segundo) já são considerados de alta complexidade [16].

3.8 Critério fundamental de busca da excitação num codificador CELP

A partir desta seção, consideram-se os sinais dentro de um sub-bloco, sendo todos de mesmo comprimento L e podendo, portanto, ser representados por vetores de dimensão L.

Em princípio, cada vetor-código c_i no dicionário é filtrado, gerando o vetor-código filtrado q_i, que, a menos de um fator de ganho G_i, determina o correspondente vetor reconstruído $\tilde{u}_i$ e conseqüente vetor de erro

$$\varepsilon_i = u - \tilde{u}_i. \tag{3.19}$$

Busca-se o vetor-código c_ξ pelo critério

$$\xi = \underset{i}{\text{argmin}} \{\|\varepsilon_i\|^2\} \tag{3.20}$$

de minimização do erro quadrático total no sub-bloco.

Tomando-se a norma quadrada da Eq. (3.19), obtém-se

$$\varepsilon_i^T \varepsilon_i = u^T u - 2\tilde{u}_i^T u + \tilde{u}_i^T \tilde{u}_i. \tag{3.21}$$

Além do vetor-alvo u, na Figura 3.8, o outro vetor dado é o vetor-código filtrado, que determina a direção do vetor-alvo reconstruído, ainda desconhecido, isto é,

$$\tilde{u}_i = G_i q_i, \tag{3.22}$$

onde G_i é o fator de ganho.

Substituindo as ocorrências do vetor-alvo reconstruído na Eq. (3.21) pela sua expressão na Eq. (3.22), tem-se

$$\varepsilon_i^T \varepsilon_i = u^T u - 2G_i q_i^T u + G_i^2 q_i^T q_i, \tag{3.23}$$

que expressa o erro quadrático total $\varepsilon_i^T \varepsilon_i$ para o índice i em função da única variável independente G_i.

Assim, pode-se obter o valor G_i do ganho que minimiza $\varepsilon_i^T \varepsilon_i$ igualando a zero a derivada deste em relação àquele, que. a partir da Eq. (3.23), fornece

$$\frac{d\left(\varepsilon_i^T \varepsilon_i\right)}{dG_i} = -2q_i^T u + 2G_i q_i^T q_i = 0, \tag{3.24}$$

de cuja 2ª igualdade pode-se obter o valor do ganho

$$G_i = \frac{q_i^T u}{q_i^T q_i}. \tag{3.25}$$

Apenas confirmando que o valor determinado de G_i em (3.25) realmente minimiza o erro quadrático total para o índice i, a 2ª derivada é $\frac{d^2\left(\varepsilon_i^T \varepsilon_i\right)}{dG_i^2} = 2q_i^T q_i$, que é um valor positivo.

Substituindo na Eq. (3.23) a expressão do ganho achado da Eq. (3.25), segue

$$\begin{aligned} \varepsilon_i^T \varepsilon_i &= u^T u - 2\frac{q_i^T u}{q_i^T q_i} q_i^T u + \left(\frac{q_i^T u}{q_i^T q_i}\right)^2 q_i^T q_i \\ &= u^T u - \frac{\left(q_i^T u\right)^2}{q_i^T q_i}. \end{aligned} \tag{3.26}$$

Como a norma quadrada do vetor-alvo reconstruído é

$$\begin{aligned} \tilde{u}_i^T \tilde{u}_i &= G_i^2 q_i^T q_i \\ &= \left(\frac{q_i^T u}{q_i^T q_i}\right)^2 q_i^T q_i \\ &= \frac{\left(q_i^T u\right)^2}{q_i^T q_i}, \end{aligned} \tag{3.27}$$

tem-se que a norma quadrada do vetor erro ε_i na Eq. (3.26) é

$$\varepsilon_i^T \varepsilon_i = u^T u - \tilde{u}_i^T \tilde{u}_i. \tag{3.28}$$

Portanto, da Eq. (3.28), que é a expressão do Teorema de Pitágoras, depreende-se que $\varepsilon_i \perp \tilde{u}_i$, mostrando que a ortogonalidade ilustrada na Fig. 3.8 está correta.

Mais importante ainda operacionalmente é notar que, sendo u constante, a Eq. (3.28) permite reescrever o critério de busca (3.20) da seguinte forma equivalente

$$\xi = \underset{i}{\text{argmax}} \left\{\|\tilde{u}_i\|^2\right\} \tag{3.29}$$

ou ainda, a partir da Eq. (3.26),

$$\xi = \underset{i}{\text{argmax}} \left\{ \frac{\left(q_i^T u\right)^2}{q_i^T q_i} \right\}. \tag{3.30}$$

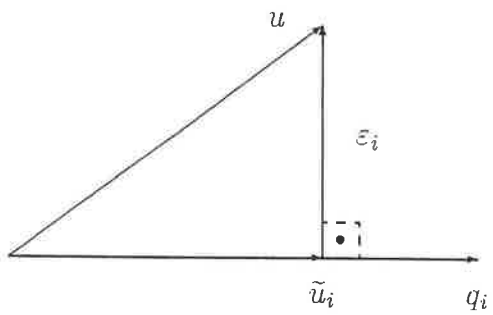

Figura 3.8: Busca da excitação por análise-mediante-síntese vista como processo projetivo.

Para referência, denota-se a correlação cruzada com deslocamento relativo ("lag") nulo entre o vetor-alvo e um dado vetor-código filtrado por

$$C_i = q_i^T u \tag{3.31}$$

e a energia do vetor-código filtrado por

$$\sigma_{q_i}^2 = q_i^T q_i, \tag{3.32}$$

de tal forma que a norma quadrada da projeção $\tau_i = \|\tilde{u}_i\|^2$ na Eq. (3.27) pode ser expressa como

$$\tau_i = \frac{C_i^2}{\sigma_{q_i}^2}. \tag{3.33}$$

Considerando que o vetor-código c_i filtrado é

$$q_i = Hc_i, \tag{3.34}$$

onde H é a matriz de resposta impulsiva, pode-se escrever a correlação em (3.31) como

$$\begin{aligned} C_i &= u^T q_i \\ &= u^T H c_i. \end{aligned} \tag{3.35}$$

A equação (3.35) permite multiplicar a matriz de resposta impulsiva à esquerda em primeiro lugar, resultando o vetor

$$t = \left(u^T H\right)^T, \tag{3.36}$$

que pode ser interpretado como uma filtragem regressiva no tempo do vetor-alvo. Este expediente permite calcular cada correlação C_i como

$$C_i = t^T c_i. \tag{3.37}$$

executando, assim, apenas uma única filtragem para a busca sobre todos os vetores-código sem ter que filtrá-los um a um.

Embora a expressão (3.37) tenha grande potencial para simplificar a busca, a simplificação não pode se realizar enquanto não englobar o cálculo da energia em (3.32), que, com (3.34), pode se expressar por

$$\sigma_{q_i}^2 = c_i^T \left(H^T H\right) c_i. \tag{3.38}$$

Com a introdução da matriz de autocorrelação da resposta impulsiva

$$\Phi = H^T H. \tag{3.39}$$

a energia pode ser representada de forma mais compacta como

$$\sigma_{q_i}^2 = c_i^T \Phi c_i. \tag{3.40}$$

expressão útil na busca de inovações esparsas no tempo (Capítulo 6).

3.9 Dicionário adaptativo

O dicionário de códigos tem sido tratado implicitamente como uma entidade fixa. Entretanto, para reconstruir segmentos quase-periódicos de voz, recorre-se a uma estrutura adaptativa para poder redefinir os vetores-código à medida que o sinal varia.

Este dicionário de códigos adaptativo têm sua constituição básica definida pelo vetor gerador b de excitações atrasadas, que, no sub-bloco m, encontra-se no estado

$$b_m = \left[\begin{array}{cccc} e(mL-K+1) & e(mL-K+2) & \cdots & e(mL) \end{array} \right], \tag{3.41}$$

onde $e(n)$, $0 \leq n \leq mL$ é a seqüência de excitações passadas e K é o comprimento do vetor gerador b_m, representado na Fig. 3.9, onde ainda é usado um dicionário fixo composto por M vetores-código.

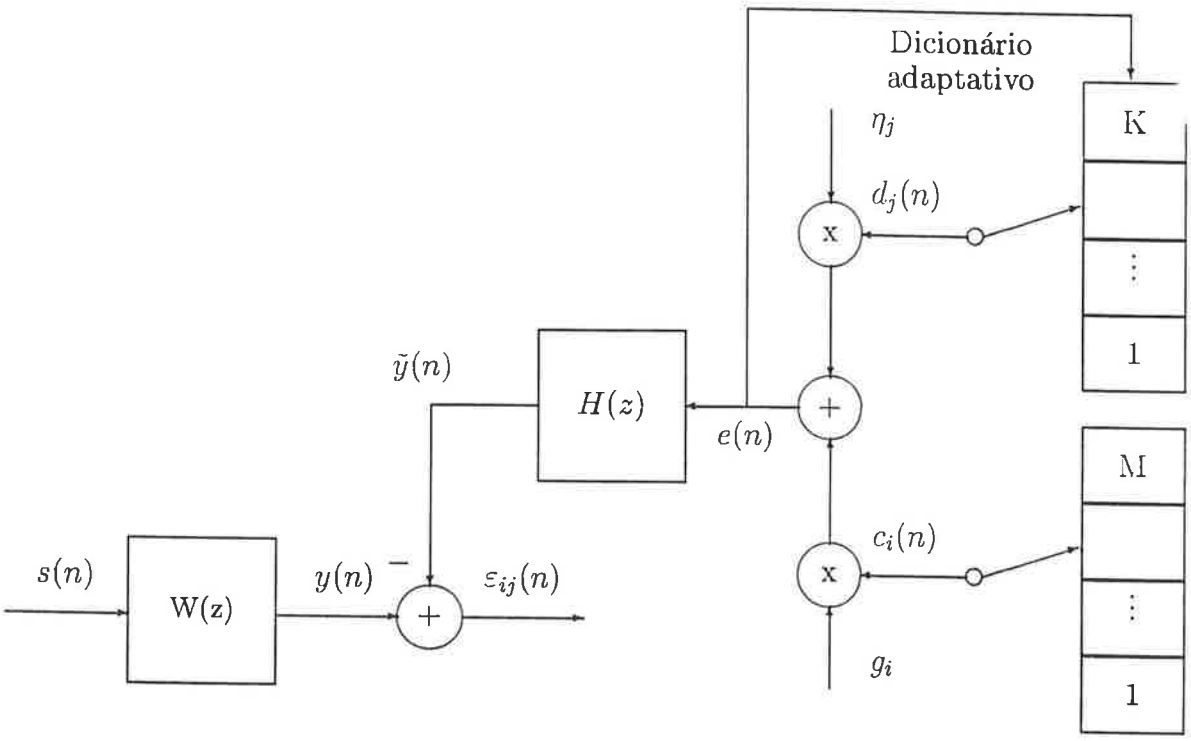

Figura 3.9: Atualização de estados do codificador CELP.

O vetor b é atualizado ao fim de cada sub-bloco de busca da excitação mediante os dois procedimentos seguintes

i) esquecimento das L amostras mais antigas;

ii) anexação do vetor e de excitação mais recente.

O vetor gerador atualizado é

$$b_{m+1}(k) = \begin{cases} b_m(k+L) & 1 \leq k \leq K-L \\ e(k-K+L) & K-L \leq k \leq K-1 \end{cases} \tag{3.42}$$

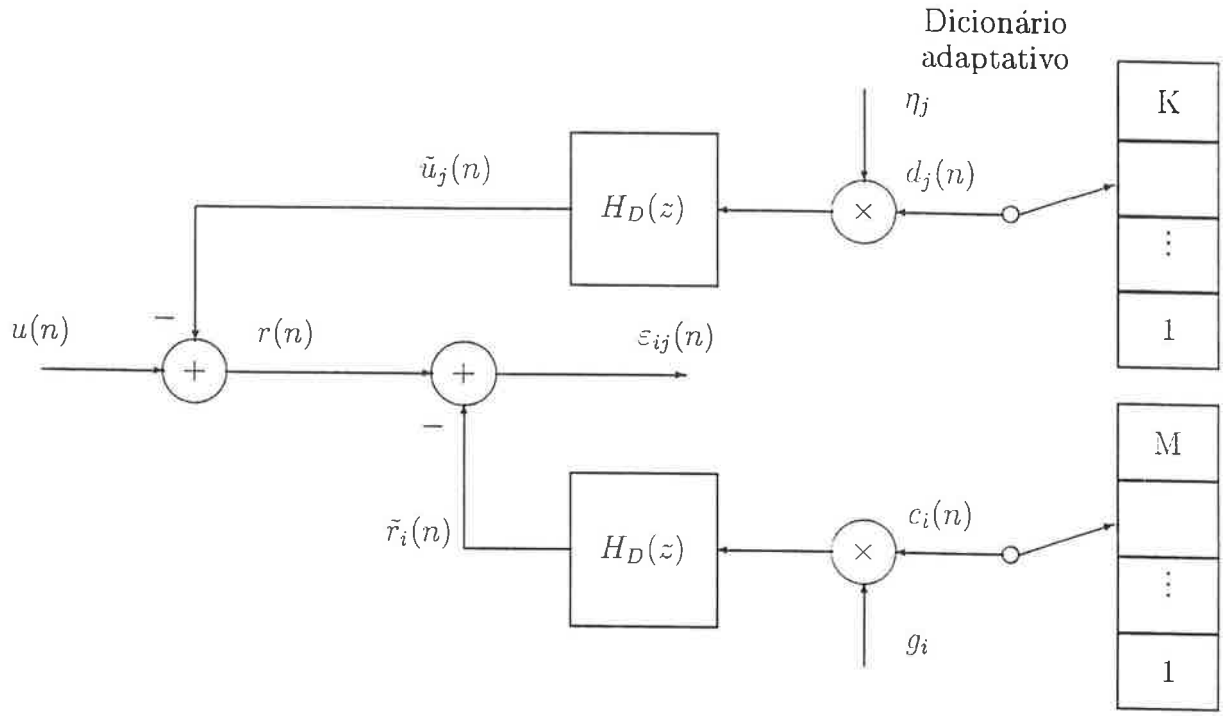

Figura 3.10: Busca no codificador CELP com dicionário adaptativo.

Os vetores-código são segmentos de comprimento L retirados do vetor gerador b com sobreposição de $L-1$ amostras entre vetores adjacentes. Isto é,

$$d_j(n) = b(K-j+n+1) \quad \begin{matrix} n = 0, 1, \ldots, L-1 \\ j = L, L+1, \ldots, K. \end{matrix} \tag{3.43}$$

De forma típica, podem-se usar sub-blocos de comprimento $L = 40$ e freqüência de amostragem $f_a = 8$ kHz. Isto significa que o índice mínimo é $j = 40$, correspondente a uma freqüência fundamental de 200 Hz.

Há falantes com freqüência fundamental acima de 200 Hz. Assim, seguindo [9], redefine-se o índice mínimo de atraso como $J = 20$.

Adotando um dicionário de códigos de 128 níveis, deve-se projetar o comprimento do vetor b das excitações atrasadas em $K = 147$. Assim, consegue-se cobrir uma faixa de períodos fundamentais de 2,500 ms a 18,375 ms com passo de 0,125 ms ou uma faixa de freqüências fundamentais equivalente de 54,42 Hz a 400 Hz.

Os vetores-código definidos pela Equação (3.43) serão chamados de *completos*. Porém, ao se estender a definição (3.43) para índices j tais que $J \leq j \leq L-1$, constata-se que faltam amostras antigas de b para completar as amostras dos vetores-código d_j correspondentes. Por este motivo, denominam-se estes vetores-código de *incompletos*.

Geralmente, complementa-se a definição do dicionário de códigos com a hipótese de que os vetores-código incompletos sejam periódicos de períodos iguais aos seus

respectivos índices. Resultam, pois, definidos como

$$d_j(n) = \begin{cases} b(K-j+n+1) & n = 0, 1, \ldots, j-1 \\ b(K-2j-n+1) & n = j, j+1 \ldots, L-1 \\ & j = J, J+1, \ldots . L-1 \end{cases} \quad (3.44)$$

A busca da excitação no dicionário adaptativo é efetuada antes da busca no dicionário fixo, cuja contribuição é assumida nula durante a busca adaptativa. Ou seja, assume-se durante a busca adaptativa que o vetor-código fixo filtrado $\tilde{r}_i(n)$ na Figura 3.10 é o vetor nulo. Assim, por simplicidade, passa-se a representar o vetor-erro ponderado por ε_j em vez de ε_{ij} como está indicado na Figura 3.10.

Consideradas essas observações, busca-se a excitação pelo mesmo processo descrito na seção 3.8 para o dicionário de códigos fixo. Isto é, seleciona-se o vetor-código $d = d_P$ tal que

$$P = \underset{j=J,\ldots,K}{\operatorname{argmin}} \varepsilon_j^T \varepsilon_j, \quad (3.45)$$

sendo $\varepsilon_j^T \varepsilon_j$ o erro quadrático ponderado de cada reconstrução (Eq. (3.20)).

Ou, ainda, pelo critério equivalente (3.29),

$$P = \underset{j=J,\ldots,K}{\operatorname{argmax}} \|\tilde{u}_j\|^2, \quad (3.46)$$

onde $\tilde{u}_j = \eta_j v_j$ é o vetor-alvo residual reconstruído pelo vetor-código atrasado d_j, sendo v_j seu vetor filtrado associado.

Alternativamente, o atraso P pode ser determinado usando-se apenas o sinal de voz ou o sinal residual da predição. Nesses casos, o algoritmo é denominado detetor de período fundamental ("pitch") [29]. Na literatura dos codificadores CELP, esses algoritmos são chamados de detetores de "pitch" em malha aberta [42] em oposição aos detetores de "pitch" em malha fechada, que são os algoritmos de análise-mediante-síntese que usam dicionários adaptativos conforme abordado nesta seção.

Os codificadores CELP mais recentes, tais como o ACELP [32], usam uma combinação de deteção de "pitch" em malha aberta com busca adaptativa. Neles o valor estimado em malha aberta para o "pitch" define um subconjunto reduzido de índices do dicionário adaptativo, que são buscados em malha fechada. Desta forma, combina-se a precisão maior da busca em malha fechada com a complexidade reduzida da busca em malha aberta.

A implementação na síntese do preditor de "pitch" pode ser através de um dicionário adaptativo ou através de um filtro de síntese de longo prazo $H_L(z)$, associado a um preditor de longo prazo $P_1(z)$ (seção 2.6), onde

$$H_L(z) = \frac{1}{1 - P_1(z)}. \quad (3.47)$$

O preditor de longo prazo

$$P_1(z) = \sum_{k=-L_p}^{L_p} \eta_k z^{-(L+k)} \quad (3.48)$$

é um filtro de ordem $L + L_p$ com $2L_p + 1$ coeficientes não nulos ou, resumidamente, um preditor de ordem $2L_p + 1$. Geralmente, usam-se preditores de longo prazo de ordens 1, 3 ou 5. O codificador G.723.1 usa um preditor de longo prazo de ordem 5 [32]. Além disso, no caso excepcional do codificador CELP de baixo atraso G.728, usa-se um único preditor de ordem 50 que cobre simultaneamente o curto prazo e o longo prazo (seção 2.6).

O dicionário adaptativo definido nesta seção pode ser implementado equivalentemente por um preditor de longo prazo de primeira ordem no caso dos vetores-código completos.

Finalizando, o dicionário de códigos atrasados é tão importante que um codificador, o "vocoder" auto-excitado (SEV) [58], funciona somente com este tipo de estrutura.

3.10 Busca em dois dicionários de códigos

Os codificadores com dois dicionários de códigos requerem uma análise muito mais complexa do que aquela dos codificadores de dicionário único se o objetivo for uma busca ótima.

Na busca ótima em dois dicionários de códigos, para cada par de vetores-código

$$(c_i, d_j), \quad 1 \leq i \leq M, \quad J \leq j \leq K,$$

onde M e $K - J + 1$ são, respectivamente, o número de níveis dos dicionários de códigos 1 e 2. Assume-se que o dicionário 2 seja adaptativo de excitações atrasadas, como é usual.

Como um processo desta complexidade é praticamente irrealizável, recorre-se a buscas seqüenciais. Às vezes, implementam-se algumas correções, tentando aproximar-se da solução ótima.

Examinam-se três procedimentos de busca em 2 dicionários de códigos:

- Buscas seqüenciais simples;
- Buscas seqüenciais com determinação conjunta dos ganhos;
- Buscas seqüenciais com ortogonalidade.

A determinação conjunta dos ganhos é um refinamento das buscas seqüenciais que pode ser implementado com pouca complexidade adicional pois basta ser executada uma única vez por busca.

Por outro lado, as buscas seqüenciais com ortogonalidade requerem uma ortogonalização a cada vetor-código fixo comparado no caso de dicionários fixos não estruturados. Assim, isto implica em mais de mil ortogonalizações por busca (seção 3.7).

Entretanto, quando os dicionários fixos são estruturados como no caso do VSELP em que os vetores-código são definidos por somas de vetores-base (seção 4.8), o número de ortogonalizações é reduzido. Por exemplo, no VSELP basta ortogonalizar os 14 vetores-base filtrados em vez dos 256 vetores-código filtrados resultantes da composição de seus dois dicionários fixos.

A abordagem seguida abaixo está baseada na dissertação de mestrado do autor [2] que foi adaptada para o contexto desta tese, tratando-se de uma abordagem própria de resultados que não aparecem unificados na literatura.

3.10.1 Buscas seqüenciais simples

Uma busca seqüencial simples em dois dicionários de códigos compõe-se apenas de duas buscas, sendo uma em cada dicionário.

A primeira busca usa o vetor-alvo u (Fig. 3.10 e Fig. 3.11), definido pela Eq. (3.15).

A segunda busca usa como vetor-alvo o erro residual da 1ª busca (Fig. 3.11), que é dado por

$$r = u - \tilde{u}. \tag{3.49}$$

O vetor de erro ponderado ao final da 2ª busca (Fig. 3.11) é

$$\varepsilon = r - \tilde{r}. \tag{3.50}$$

As buscas da excitação seguem o processo descrito na seção 3.8. Desta forma, em primeiro lugar, busca-se o vetor-código atrasado $d = d_P$ (seção 3.9) de modo que o vetor de voz ponderado gerado pelo vetor atrasado d pode ser obtido como

$$v = Hd. \tag{3.51}$$

Assim, o vetor reconstruído $\tilde{u}$ é a projeção euclidiana do vetor-alvo u sobre o vetor ponderado v, dada por

$$\tilde{u} = \frac{u^T v}{v^T v} v. \tag{3.52}$$

Por sua vez, o ganho η_P é definido pela coordenada da projeção $\tilde{u}$ na direção de projeção v, ou seja,

$$\tilde{u} = \eta_P v \tag{3.53}$$

pode ser calculado como

$$\eta_P = \frac{u^T v}{v^T v}, \tag{3.54}$$

resultado que pode ser obtido comparando-se as Equações (3.53) e (3.52).

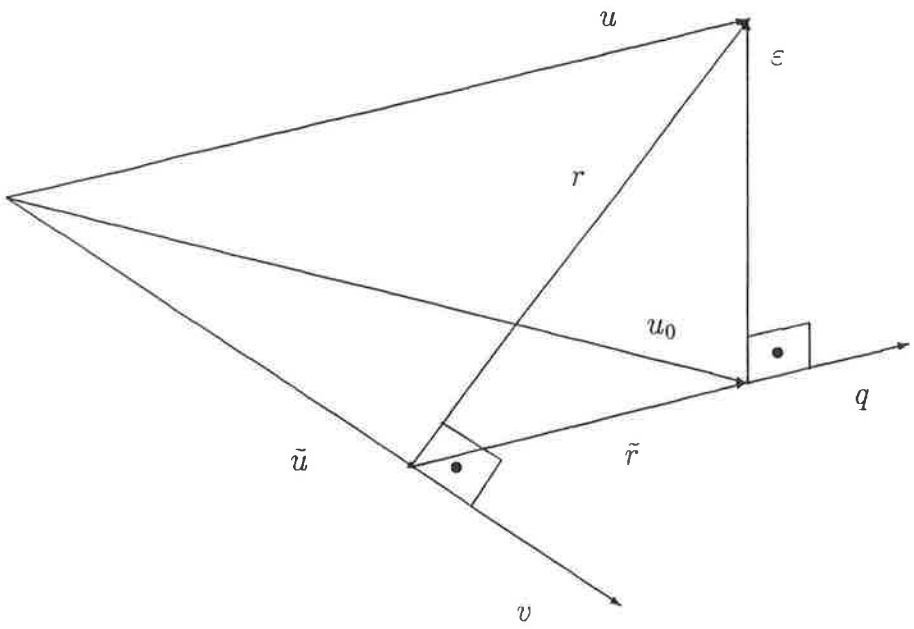

Figura 3.11: Representação geométrica dos vetores na busca seqüencial com vetor-alvo u e vetor-alvo residual r.

Para a busca no dicionário de códigos fixo, executam-se os mesmos procedimentos:

- busca do vetor-código $c = c_i$ (Fig. 3.10);

- filtragem de c, resultando no vetor de voz ponderado

$$q = Hc; \tag{3.55}$$

- reconstrução do vetor-alvo residual r baseada em q, dada pelo vetor

$$\tilde{r} = \frac{r^T q}{\|q\|^2} q; \tag{3.56}$$

- determinação do ganho g tal que

$$\tilde{r} = gq, \tag{3.57}$$

que resulta

$$g = \frac{r^T q}{\|q\|^2}. \tag{3.58}$$

Após as duas buscas, pode-se reconstruir o vetor-alvo u pela filtragem da excitação composta

$$e_0 = \eta_P d + gc, \tag{3.59}$$

resultando na reconstrução composta

$$\tilde{u}_0 = \eta_P v + gq. \tag{3.60}$$

3.10.2 Buscas seqüenciais com determinação conjunta dos ganhos

No método de buscas seqüenciais simples, devido às projeções realizadas (Eq. (3.52) e Eq. (3.56)), temos

$$r \perp v \iff r^T v = 0 \tag{3.61}$$

$$\varepsilon \perp q \iff \varepsilon^T q = 0. \tag{3.62}$$

Porém, com as condições (3.61) e (3.62) apenas, não é possível garantir que o erro final ε seja ortogonal ao plano $[v\ q]$ (Fig. 3.11) porque não sabemos se a ortogonalidade "$\varepsilon \perp v$" se verifica ou não.

Para garantir a ortogonalidade de ε a $[v\ q]$, dada a Eq. (3.62), basta impor a condição

$$\varepsilon \perp v \iff \varepsilon^T v = 0. \tag{3.63}$$

Revendo o procedimento descrito na seção 3.10.1, verifica-se que é possível melhorar a reconstrução (3.60) se, em vez de tomar ganhos η_P e g pelos seus valores dados em (3.54) e (3.58), respectivamente, forem usadas suas expressões gerais (3.53) e (3.57).

Da Fig. 3.11, extrai-se a condição de decomposição do vetor-alvo

$$u = \tilde{u} + \tilde{r} + \varepsilon, \tag{3.64}$$

onde, introduzindo os ganhos através das Eqs.(3.53) e (3.57), tem-se

$$u = \eta_P v + gq + \varepsilon. \tag{3.65}$$

Tomando os produtos internos membro a membro de (3.65) por v e por q separadamente, obtém-se

$$u^T v = \eta_P(v^T v) + g(q^T v) + (\varepsilon^T v) \tag{3.66}$$
$$u^T q = \eta_P(v^T q) + g(q^T q) + (\varepsilon^T q), \tag{3.67}$$

respectivamente.

Impondo as condições (3.62) e (3.63) às equações (3.66) e (3.67), obtém-se o sistema de equações lineares

$$\begin{bmatrix} v^T v & v^T q \\ v^T q & q^T q \end{bmatrix} \begin{bmatrix} \eta_P \\ g \end{bmatrix} = \begin{bmatrix} u^T v \\ u^T q \end{bmatrix}, \tag{3.68}$$

cuja resolução fornece os novos valores dos ganhos.

3.10.3 Buscas seqüenciais com ortogonalidade

Na busca com ganhos determinados conjuntamente da seção anterior, a ortogonalidade do vetor-erro ao plano $[\ v \quad q\ ]$ dos vetores-código filtrados é imposta *a posteriori*.

Podem-se aproximar um pouco mais as buscas seqüenciais da busca conjunta ótima se a 2ª busca já for realizada com a imposição da ortogonalidade.

A busca adaptativa com vetor-alvo u, resultando no vetor-código filtrado v e no ganho η_P, é executada inicialmente como na seção 3.10.2.

A busca no dicionário de códigos fixo envolve as seguintes etapas (Fig. 3.12):

- cada vetor-código c_i é filtrado, produzindo o vetor ponderado

$$q_i = Hc_i; \tag{3.69}$$

- o plano $[\ v \quad q_i\ ]$ é representado por uma base ortogonal através da determinação da componente w_i de q_i ortogonal a v, que é

$$w_i = q_i - \frac{v^T q_i}{v^T v} v; \tag{3.70}$$

- projetando-se o vetor-alvo u sobre o vetor ortogonal w_i, obtém-se o vetor

$$t_i = \frac{w_i^T u}{w_i^T w_i} w_i, \tag{3.71}$$

cuja norma quadrada é

$$\|t_i\|^2 = \frac{(w_i^T u)^2}{w_i^T w_i}; \tag{3.72}$$

- seleciona-se o vetor-código c_ξ de índice ξ tal que

$$\xi = \underset{i=1,\ldots,M}{\text{argmax}} \|t_i\|^2; \tag{3.73}$$

- determinação conjunta dos ganhos η_P e g pelo sistema (3.68), com $q = Hc_\xi$, resultando a excitação total

$$e = \eta_P d + g c_\xi.$$

O critério de busca (3.73) aqui utilizado é equivalente àquele da seção 3.8, conforme se mostra em seguida.

Tomando-se o ganho η_P da busca adaptativa inicial e a projeção t_i do vetor-alvo u sobre cada vetor ortogonal w_i dada pela Eq. (3.71), pode-se expressar a correspondente projeção de u em cada plano $[\ v \quad q_i\ ]$ como

$$\tilde{u}_i = \eta_P v + t_i. \tag{3.74}$$

A cada projeção, o vetor-alvo u pode ser recomposto (Fig. 3.12) como

$$u = \tilde{u}_i + \varepsilon_i, \tag{3.75}$$

onde ε_i é o vetor-erro correspondente.

Considerando que, por construção,

$$\varepsilon_i \perp \tilde{u}_i,$$

tomando o produto interno de cada membro da Eq. (3.75) por si próprio, obtém-se

$$\|u\|^2 = \|\tilde{u}_i\|^2 + \sigma_{\varepsilon_i}^2, \tag{3.76}$$

onde

$$\sigma_{\varepsilon_i}^2 = \|\varepsilon_i\|^2$$

é o erro quadrático ponderado total associado a cada vetor-código c_i.

Como durante o processo de busca $\|u\|^2$ é constante, decorre da condição (3.76) que a ocorrência do mínimo de σ_{ε_i} coincide com a ocorrência do máximo de $\|\tilde{u}_i\|^2$, ou seja,

$$\underset{i=1,\ldots,M}{\text{argmin}}\ \sigma_{\varepsilon_j} = \underset{i=1,\ldots,M}{\text{argmax}} \|t_i\|^2. \tag{3.77}$$

Portanto, o critério de busca pelo menor erro quadrático ponderado total equivale ao critério de busca pela maior projeção (Eq. (3.73)).

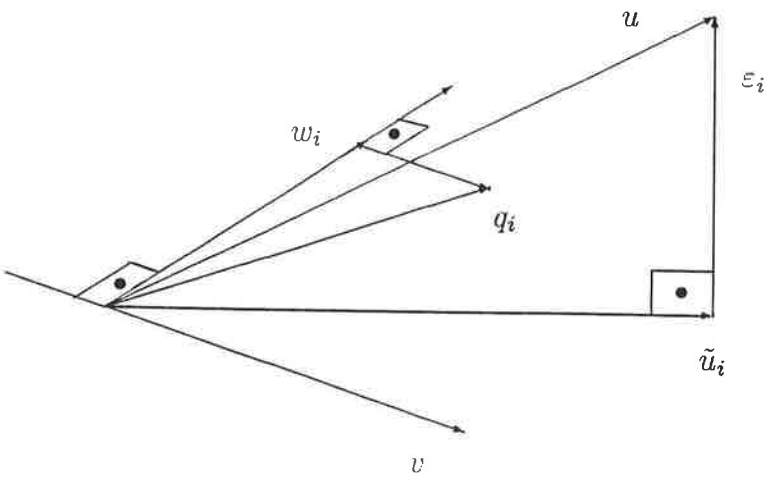

Figura 3.12: Representação geométrica dos vetores na busca seqüencial com ortogonalidade.

3.11 Representações de dicionários de códigos

Para reduzir a complexidade da busca, conforme mencionado na seção 3.3, costuma-se usar uma decomposição forma-ganho do dicionário de códigos de excitação.

Desta forma, o dicionário de códigos de excitação passa a ser um dicionário-produto, cujo dicionário de códigos é o produto cartesiano do dicionário de formas de onda normalizadas com o dicionário (ou tabela de quantização no caso escalar) de fatores de ganho [46].

Ampliando esta discussão, no projeto do dicionário de códigos de formas normalizadas há quatro características dependentes a considerar [42]:

i) tamanho ou número de vetores-código;

ii) dimensão ou comprimento dos sub-blocos;

iii) complexidade da busca;

iv) população de códigos.

Para melhorar o desempenho da codificação sem aumentar excessivamente o número de códigos, costuma-se acrescentar um dicionário adaptativo para capturar as características periódicas do sinal de voz. Na verdade, isto constitui outro dicionário-produto cuja busca se estende sobre a soma dos números de códigos de seus componentes em vez da multiplicação (seção 3.9). Porém, considerando a execução em seqüência dos processos de busca, o conjunto pode ser visto como um dicionário multiestágio.

Como otimização, os dicionários multiestágios admitem ortogonalização, que pode ser executada a cada junção entre estágios ou na sua composição final como já foi visto no caso da combinação do vetor-código fixo com o vetor-código adaptativo na seção 3.10.

As características da população influenciam muito na complexidade da busca dos vetores-código. A característica primordial é a sua forma de existência, isto é, se eles são perenes ou efêmeros.

Os dicionários de códigos em que cada vetor-código é perene podem ser imaginados como uma tabela de formas de onda. Esta tem sido a imagem adotada até aqui para o dicionário de códigos fixo gaussiano. Entretanto, ele poderia ser definido por uma semente que, processada por um gerador de números pseudo-aleatórios, se tornasse num vetor-código da dimensão projetada.

De fato, já foram usadas "sementes" para o dicionário adaptativo, que foram denominadas de "vetores geradores". Embora haja um processo de "geração" nos dois casos, o vetor gerador do dicionário adaptativo é variável e a semente de valor determinado do dicionário aleatório sempre gera o mesmo vetor-código de dimensão L. Por sua vez, a dimensão do vetor gerador do dicionário adaptativo, geralmente, é $K > 2L$, onde L é a dimensão do dicionário, cujo tamanho é próximo de K. Portanto, durante o processo de busca adaptativa ocorre uma proliferação acumulada, porém, não necessariamente simultânea, de K valores para algo próximo de KL valores ao

passo que, durante a busca no dicionário aleatório, a proliferação é de M valores para ML valores.

O mecanismo de compactação envolvido na representação do dicionário adaptativo pelo seu vetor gerador chama-se "sobreposição" entre vetores-código adjacentes, sendo que, no caso do dicionário adaptativo, os vetores-código são extraídos por uma janela móvel horizontal de comprimento L, que desliza sobre o vetor gerador.

Na figura 3.13 pode-se visualizar como a propriedade populacional de sobreposição é herdada pelo dicionário adaptativo e por alguns dicionários fixos.

No Capítulo 4 apresentam-se os dicionários ceifados no centro (seção 4.1) e os dicionários de multipulsos algébricos (seção 4.4), que ensejam a extensão do conceito de vetor gerador.

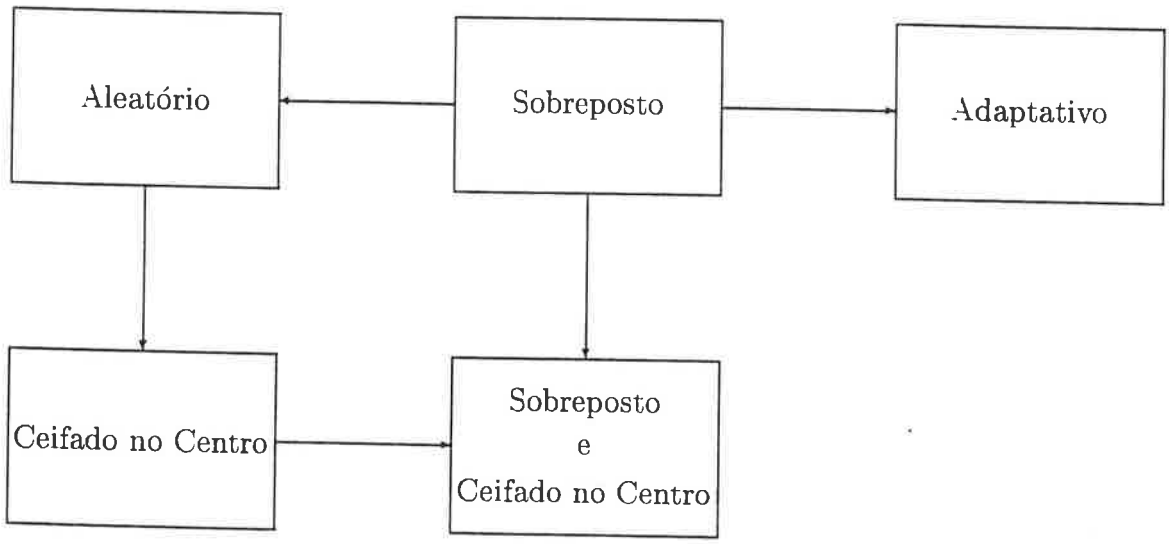

Figura 3.13: Herança da propriedade de sobreposição entre dicionários de códigos.

Capítulo 4

MODELOS DE INOVAÇÕES

As inovações surgem no contexto da representação de processos estocásticos estacionários no sentido amplo $r(n)$ por um sistema de fase mínima $H_{\mathrm{LPC}}(z)$ com um processo de ruído branco $e(n)$ à sua entrada (Fig. 4.1).

O processo estocástico $e(n)$ é denominado de inovações do processo $r(n)$, o filtro H_{LPC} é o filtro de inovações ("innovations filter") e seu filtro inverso $A(z) = \frac{1}{H_{\mathrm{LPC}}(z)}$ é o filtro de branqueamento ("whitening filter") [50].

Neste capítulo o termo "inovações" será usado com certa impropriedade porque o sinal usado é o vetor-alvo residual $r(n)$ após a busca no dicionário de códigos adaptativo e o filtro de inovações é o filtro de síntese, que é um sistema de fase mínima para o sinal de voz original completo $s(n)$.

Entretanto, a distinção maior que se pretende fazer é entre a componente periódica do sinal de voz e a componente fixa, que é nova em relação ao sinal de voz passado. Ademais, a componente fixa $e(n)$ do sinal de excitação aproxima-se bem de um sinal amostral de um processo de ruído branco.

No Capítulo 3 foi apresentado o dicionário de inovações gaussiano. Neste capítulo, serão apresentados os dicionários de inovações com ceifagem central, os dicionários multipulso, os dicionários que utilizam pulsos regulares, os dicionários compostos por multipulsos algébricos e os que são compostos por vetores-base. Ademais, este último admite treinamento, que também será abordado.

Há ainda o método das autocorrelações no domínio do tempo, que simplifica a busca ao custo da manutenção de um segundo dicionário auxiliar de autocorrelações

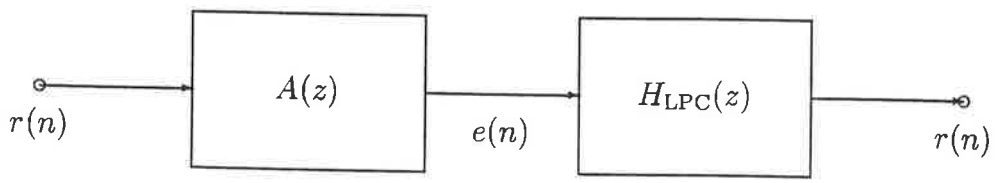

Figura 4.1: Análise e síntese do vetor-alvo residual por inovações.

[69]. Também é possível definir os vetores-código fixos em domínios transformados, como o domínio da freqüência (Fourier) e o domínio dos vetores singulares (SV) [69].

Finalmente, o dicionário de inovações, que tem sido tomado como sinônimo de fixo, também pode se tornar variável no domínio SV [2] [53] [54].

4.1 Dicionários de códigos ceifados centralmente

Os dicionários de códigos com ceifagem central são gerados a partir de um dicionário estocástico gaussiano de média nula e variância unitária.

Gera-se cada vetor-código ceifado f_i pela aplicação ao seu vetor-código estocástico associado c_i da função de ceifagem central ("center-clipping")

$$f_i(n) = \begin{cases} c_i(n) & \text{se } |c_i(n)| > A_{cc} \\ 0 & \text{caso contrário} \\ & n = 0, 1, \ldots, L-1 \\ & i = 1, 2, \ldots, M \end{cases} \quad (4.1)$$

onde

- A_{cc} é o nível de ceifagem central;
- L é o comprimento do sub-bloco de excitação e
- M é o número de vetores em cada dicionário de códigos.

Para a geração de um dicionário de códigos ternário, quantizam-se os vetores-código ceifados centralmente com a função signum, ou seja,

$$t_i(n) = \begin{cases} \text{sign}(f_i(n)) & n = 0, 1, \ldots, L-1 \\ & i = 1, 2, \ldots, M. \end{cases} \quad (4.2)$$

O dicionário fixo do codificador CELP do Departamento de Defesa dos E.U.A. (DoD) [9] é um dicionário de códigos ternários obtidos de sinais gaussianos por ceifagem central dentro de $A_{cc} = 1,2$ desvios-padrão, resultando numa esparsidade média de 77%. Ademais, seus 512 vetores-códigos de dimensão $L = 60$ podem ser obtidos por sobreposição de $L-2$ amostras a partir de um vetor gerador de dimensão $K = 1084$.

Os vetores-código ceifados assim como os ternários têm espectros mais brancos do que os gaussianos originais.

4.2 Estruturas multipulso

Os sinais de excitação multipulso são sinais esparsos, compostos por uma pequena quantidade de pulsos. Podem ser representados de forma genérica por

$$c(n) = G \sum_{k=0}^{M-1} \alpha_k \delta(n - m_k), n = 0, 1, \ldots, L-1, \quad (4.3)$$

onde L é o comprimento do sub-bloco de busca da excitação, M é o número de pulsos, α_k e m_k são, respectivamente, as amplitudes e as posições dos pulsos e G é o fator de ganho do sinal de excitação.

Este modelo de sinais de excitação foi proposto inicialmente por Atal e Remde para melhorar a naturalidade da voz produzida por sintetizadores de voz ou vocoders LPC [7].

O número de pulsos utilizado é pequeno em comparação com a duração do sub-bloco, tornando o sinal de excitação altamente esparso. Por exemplo, [7] sugere uma densidade de 8 pulsos por 10 ms no máximo. Isto representa uma esparsidade de 90%, que é mais elevada que aquela imposta ao dicionários de códigos ternários da seção 4.1 acima.

4.3 Pulsos regulares

A excitação multipulso é chamada de "excitação de pulsos regulares" se os pulsos componentes estão sobre uma grade com espaçamento S constante e suas posições são

$$m_k = kS + i, \tag{4.4}$$

onde o vetor de excitação de pulsos regulares é identificado pelo indice i, que pode ser interpretado como a fase da grade e pode assumir os valores $i = 0, 1, \ldots, S-1$ no caso do dicionário completo.

4.4 Multipulsos algébricos

Embora a codificação CELP algébrica (ACELP) possa ser aplicada para taxas de transmissão por volta de 10 kbit/s de sinais de voz de faixa larga [43], enfoca-se abaixo sua aplicação à transmissão de sinais de voz de faixa telefônica a taxas próximas de 5 kbit/s.

Cada inovação do ACELP toma um único pulso de cada fase de excitação regular

$$m_k = j_k S + i_k + i_g, \tag{4.5}$$

onde o espaçamento é $S = 8$, j_k indica a colocação do pulso escolhido da fase i_k e as fases são $i_0 = 0, i_1 = 2, i_2 = 4$ e $i_3 = 6$. O deslocamento ("shift") $i_g = 0$ ou $i_g = 1$ indica a seleção da grade par ou da grade ímpar, respectivamente. As ordens de posicionamento $j_k = 0, 1, \ldots, 7$ dos pulsos de cada fase é que determinam suas posições na grade (Tabela 4.1).

Os vetores de pulsos regulares podem ser vistos como dicionário gerador no seu conjunto, sendo que há dois dicionários geradores, identificados com a grade par e com a grande ímpar, respectivamente, pois de cada grade emerge um conjunto de vetores-código (seção 3.11). Numa dada grade i_g, o vetor de localizações $j = (j_0, j_1, j_2, j_3)$ define um caminho no grafo da grade. Os nós do grafo associado a uma grade são as posições dos pulsos das fases com ramos entre todos os pares de nós de fases adjacentes (Tabela 4.1).

Tabela 4.1: Grade de posições pares do ACELP.

Fase	Posições
$i_0 = 0$	0, 8, 16, 24, 32, 40, 48, 56
$i_1 = 2$	2, 10, 18, 26, 34, 42, 50, 58
$i_2 = 4$	4, 12, 20, 28, 36, 44, 52, (60)
$i_3 = 6$	6, 14, 22, 30, 38, 46, 54, (62)

Cada caminho do grafo define o conjunto das posições dos pulsos do vetor-código c_i na grade selecionada. Isto é, o índice i do vetor-código é composto por duas componentes $i = (i_p, i_s)$, sendo

$$i_p = (m_0, m_1, m_2, m_3) \tag{4.6}$$

a componente que define as posições e a componente i_s codifica os sinais dos 4 pulsos (Equação 4.5).

A busca da excitação é facilitada pela natureza ternária dos sinais, cujos pulsos não nulos têm amplitudes $\alpha_k \in \{-1, 1\}$. Portanto, geometricamente, as excitações admissíveis para o codificador G.723.1 situam-se nos vértices de hipercubos centrados na origem, totalmente contidos em subespaços de dimensão 4.

Entretanto, a busca da excitação no G.723.1 se dá por um processo de análise-mediante-síntese no estilo CELP, onde os pulsos que compõem o sinal de excitação são usados conjuntamente. Ao contrário, no estilo multipulso de busca, define-se a posição de um pulso por vez.

A composição do sinal de excitação é muito robusta nesta estruturação multipulsos algébrica como adotada no G.723.1, pois o tratamento dos segmentos periódicos é dado no estilo de dicionário de códigos adaptativo com ênfase harmônica. Ademais, quando o retardo da excitação periódica é menor do que o comprimento de um sub-bloco, o próprio vetor fixo de excitação sofre um processo de periodização com a soma de uma imagem atrasada.

Note-se ainda que, ao contrário dos codificadores harmônicos, este tratamento especial da periodicidade é feito sem se perder a capacidade de representação dos segmentos de voz surdos.

4.5 ACELP com várias densidades de pulsos

Para pesquisar o efeito da densidade de pulsos sobre o desempenho do codificador ACELP, foram testadas uma densidade menor e outra maior que a densidade padronizada de 4 bits por 7,5 ms. Essas densidades foram de 3 e de 8 pulsos por sub-bloco de 7,5 ms, respectivamente.

O dicionário ACELP de 3 pulsos gera a mesma taxa de bits para as posições que o dicionário padrão, 12 bit/grade/sub-bloco, que se distribui entre 3 fases com 16 posições cada. Na verdade, a fase i_2 é constituída pelo mesmo vetor gerador da fase

Tabela 4.2: Grade de posições pares do ACELP de 3 pulsos.

Fase	Posições
$i_0 = 0$	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, (60)
$i_1 = 2$	2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, (62)
$i_2 = 0$	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, (60)

Tabela 4.3: Grade de posições pares do ACELP de 8 pulsos.

Fase	Posições
$i_0 = 0$	0, 16, 32, 48
$i_1 = 2$	2, 18, 34, 50
$i_2 = 4$	4, 20, 36, 52
$i_3 = 6$	6, 22, 38, 54
$i_4 = 8$	8, 24, 40, 56
$i_5 = 10$	10, 26, 42, 58
$i_6 = 12$	12, 28, 44, (60)
$i_7 = 14$	14, 30, 46, (62)

i_0 (Tabela 4.2). Assim, pode-se afirmar, equivalentemente, que há 2 fases e tomam-se 2 posições distintas da primeira e 1 posição da segunda.

Em contraste, o dicionário ACELP de 8 pulsos projetado, para manter a mesma cobertura de posições por grade, necessita a taxa de 16 bit/grade/sub-bloco para codificar as posições apenas (Tabela 4.3). Adicionalmente, compondo esta taxa com aquela necessária para transmitir os 8 sinais algébricos e os parâmetros restantes da excitação e da predição linear, atinge-se a taxa de 6400 bit/s para o codificador como um todo (Tabela 4.4).

Como a taxa de 6400 bit/s é apenas ligeiramente superior do que a taxa do codificador multipulsos MP-MLQ ("Multipulse Maximum Likelihood Quantization"), decidiu-se incluí-lo também nos testes.

Tabela 4.4: Distribuição de bits do índice do ACELP para três densidades de pulsos

Densidade (pulsos por sub-bloco)	Três	Quatro	Oito
Localizações (bits em j)	12	12	16
Grade (bits em i_g)	1	1	1
Sinais (bits em i_s)	3	4	8
Total (bits em i)	16	17	25
Taxa do codificador (bit/bloco)	156	160	192
Taxa do codificador (bit/s)	5200	5333	6400

Tabela 4.5: Desempenho de três codificadores ACELP com 3, 4 e 8 pulsos por sub-bloco e do codificador MP-MLQ.

Sinal	Codificador	SNRSEG (dB)	PSQM
v	ACELP-3	12.5567	2,6046
	ACELP-4	13,0996	2,5359
	ACELP-8	15,0415	2,4707
	MP-MLQ	15,1241	2.3381
sa1	ACELP-3	8,9733	2,6155
	ACELP-4	9,4280	2,5742
	ACELP-8	10,4688	2.3702
	MP-MLQ	10,7078	2.3164

O codificador MP-MLQ tem excitação fixa composta por multipulsos algébricos em número de 6 para os sub-blocos pares e 5 para os ímpares [32].

Para os testes usaram-se os sinais "v" e "sa1" (seção 5.2), sendo que o sinal "sa1" foi retirado do conjunto da falante FAKS0 da região dialetal 1 da partição de teste da base de dados TIMIT.

O codificador ACELP com densidade de 4 pulsos por sub-bloco é o codificador padronizado, porém, empregando a busca exaustiva de posições (seção 6.1), que é também a que é usada pelos codificadores ACELP de 3 e 8 pulsos.

Pelos resultados dos testes (Tabela 4.5), o ACELP-3 apresenta uma perda considerável de 0,5 dB em relação ao ACELP-4. Considerando, adicionalmente, que a complexidade da busca exaustiva do ACELP-3 é a mesma que a do ACELP-4, não é compensador reduzir a densidade de pulsos.

Com relação ao ACELP-8, o aumento de desempenho é considerável em relação ao ACELP-4, sendo de pelo menos 1 dB. Entretanto, como o seu desempenho é, mesmo assim, ligeiramente inferior ao do MP-MLQ, é preferível abandoná-lo em favor deste, que apresenta complexidade menor.

4.6 A busca da excitação multipulso

A busca da excitação multipulso (Eq. (4.3)) tem como meta global a minimização da norma do vetor de erro (seção 3.8). Neste caso, o vetor de excitação é

$$e(n) = \sum_{k=0}^{M-1} A_k \delta(n - m_k), \tag{4.7}$$

onde se indicam explicitamente as amplitudes completas $A_k = G\alpha_k$ dos pulsos individuais.

O correspondente vetor-alvo reconstruído é dado por

$$\tilde{u}(n) = \sum_{k=0}^{M-1} A_k h(n - m_k), \tag{4.8}$$

onde $h(n)$, para $n = 0, 1, \ldots, L-1$. é a seqüência de resposta impulsiva do filtro de síntese ponderado truncada no comprimento L do sub-bloco (seção 3.6).

Sendo u o vetor-alvo, o vetor-erro de reconstrução ponderado pode ser expresso por

$$\varepsilon(n) = u(n) - \sum_{k=0}^{M-1} A_k h(n - m_k). \tag{4.9}$$

Anulando-se o gradiente do erro quadrático ponderado $\varepsilon^T\varepsilon$ em relação às amplitudes, obtém-se [71]

$$\begin{bmatrix} \Phi(m_0, m_0) & \Phi(m_0, m_1) & \cdots & \Phi(m_0, m_{M-1}) \\ \Phi(m_1, m_0) & \Phi(m_1, m_1) & \cdots & \Phi(m_1, m_{M-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(m_{M-1}, m_0) & \Phi(m_{M-1}, m_1) & \cdots & \Phi(m_{M-1}, m_{M-1}) \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_{M-1} \end{bmatrix} = \begin{bmatrix} \xi_{m_0} \\ \xi_{m_1} \\ \vdots \\ \xi_{m_{M-1}} \end{bmatrix}, \tag{4.10}$$

onde a primeira matriz à esquerda dos coeficientes do sistema linear (4.10) é a submatriz $\Phi(m, m)$ da matriz de autocorrelação da resposta impulsiva Φ, introduzida pela Eq. (3.39) na seção 3.8, sendo a M-pla $m = (m_0, m_1, \ldots, m_{M-1})$ e

$$\xi_{m_k} = u^T H(:, m_k), \tag{4.11}$$

sendo $H(:, m_k)$ a coluna m_k da matriz de resposta impulsiva H.

Convém salientar que a equação matricial (4.10) presta-se ao recálculo das amplitudes dos pulsos uma vez que suas posições tenham sido previamente determinadas. A minimização do erro $\varepsilon_i^T\varepsilon_i$ em relação às posições dos pulsos produz um sistema não linear de equações sem solução fechada [67]. Assim, a solução geralmente adotada é a determinação da posição de um único pulso por vez.

Na verdade, a Eq. (4.10) pode ser usada num processo iterativo, sendo tomadas apenas as posições determinadas até a iteração em consideração.

A posição do primeiro pulso é determinada por

$$m_1 = \underset{i=0,1,\ldots,L-1}{\operatorname{argmin}} \varepsilon_i^T\varepsilon_i \tag{4.12}$$

O vetor-alvo é atualizado entre iterações

$$u' = u - \sum_{k=0}^{l-1} A_k H(:, m_k), \tag{4.13}$$

onde l é o número de pulsos já determinados.

Substituindo-se o novo vetor-alvo em (4.11), obtém-se as correlações cruzadas atualizadas

$$\xi_i' = \xi_i - \sum_{k=0}^{l-1} A_k \Phi(m_k, i), \tag{4.14}$$

usadas na determinação da posição do pulso seguinte por

$$m_l = \underset{i=0,1,\ldots,L-1}{\text{argmin}} \; \varepsilon_i'^T \varepsilon_i', \tag{4.15}$$

4.7 Dicionários fixos conjugados

O uso de mais de um dicionário fixo possibilita uma busca multiestágio, que é menos complexa do que uma busca sobre um único dicionário do tamanho do produto cartesiano dos dicionários componentes, reduzindo o número de vetores-código buscados (seção 3.11). Além disso, o codificador torna-se menos vulnerável a erros de troca de bits na transmissão, porque um dos dois vetores-código pode ser recebido sem erro com grande probabilidade.

Tomando-se como modelo o CELP com estrutura conjugada da NTT [27], o vetor-código conjugado resultante de dois subdicionários é

$$c_i = \theta_1 c_{i_1 j} + \theta_2 c_{i_2 k}, \tag{4.16}$$

onde θ_1 e θ_2 são os sinais dos vetores e $c_{i_1 j}$ e $c_{i_2 k}$ são os vetores de excitação dos dois subdicionários.

Esta composição conjugada também foi aplicada ao quantizador de ganhos CS-CELP e esse quantizador de ganhos foi retido na versão final do codificador CS-ACELP da recomendação G.729 da ITU-T [33].

4.8 Dicionários estruturados por vetores-base

O codificador preditivo de voz excitado por soma vetorial (VSELP) contém um dicionário de códigos atrasados adaptativo e dois dicionários de códigos fixos [25].

O vetor de excitação total num sub-bloco (Fig.4.2) é dado por

$$x = \eta d + g_1 c_1 + g_2 c_2, \tag{4.17}$$

onde

- d é o vetor-código atrasado e η é o seu ganho;
- c_1 e c_2 são vetores-código selecionados de cada um dos dois dicionários de códigos fixos, com ganhos g_1 e g_2, respectivamente.

Ademais, os dicionários de códigos fixos 1 e 2 são definidos pelos conjuntos de vetores-base

$$\{b_{1k}, \quad k = 1, \ldots, M_B\}$$

e

$$\{b_{2k}, \quad k = 1, \ldots, M_B\},$$

respectivamente.

Os vetores-código c_1 e c_2 são produzidos como

$$c_1 = \sum_{k=1}^{M_B} \theta_{1k} b_{1k} \tag{4.18}$$

$$c_2 = \sum_{k=1}^{M_B} \theta_{2k} b_{2k}, \tag{4.19}$$

respectivamente.

As coordenadas θ_{lk} assumem apenas valores binários

$$\theta_{lk} \in \{-1, 1\}, \quad l = 1, 2.$$

Portanto, os vetores-código fixos c_1 e c_2 são conjugados (seção 4.7), sendo que os vetores dos subdicionários são constantes e representados pelos vetores-base. Os parâmetros que distinguem os vetores-código num mesmo dicionário fixo são, pois, os "ganhos" θ_{lk}.

Assim, cada coordenada pode ser representada por um bit. Define-se

$$\nu_{lk} = 1 \quad \leftrightarrow \quad \theta_{lk} = 1; \tag{4.20}$$

$$\nu_{lk} = 0 \quad \leftrightarrow \quad \theta_{lk} = -1. \tag{4.21}$$

Os bits ν_{1k} e ν_{2k} são os dígitos binários dos índices i_1 e i_2, respectivamente, ou seja, estes índices são representados binariamente pelas seqüências de bits

$$i_1 = \sum_{k=1}^{M_B} \nu_{1k} 2^{k-1} \tag{4.22}$$

$$i_2 = \sum_{k=1}^{M_B} \nu_{2k} 2^{k-1}. \tag{4.23}$$

No VSELP original [25], a dimensão de cada dicionário de códigos fixo é

$$M_B = 7.$$

4.9 Treinamento dos dicionários fixos do VSELP

O treinamento pode ser visto como um algoritmo de quantização vetorial generalizado, que consiste na iteração das duas fases seguintes

i) Os parâmetros são determinados a partir do sinal de voz $s(n)$, dados os vetores-base $b_{1k}(n)$ e $b_{2k}(n)$, $k = 1, 2, \ldots, 7$

$$[h_m(n), g_m, \eta_m, u_m(n), d_m(n)] = \mathcal{C}\left(s(n) | b_{1k}(n), b_{2k}(n)\right) \tag{4.24}$$

para os sub-blocos $m = 1, 2, \ldots, M$. Os parâmetros calculados são as respostas impulsivas $h_m(n)$ dos filtros de síntese ponderados, o vetor de ganhos g_m com os ganhos relativos a cada vetor-base, o ganho η_m do vetor-código adaptativo, os vetores-alvo $u_m(n)$ e os vetores-código $d_m(n)$ selecionados do dicionário adaptativo.

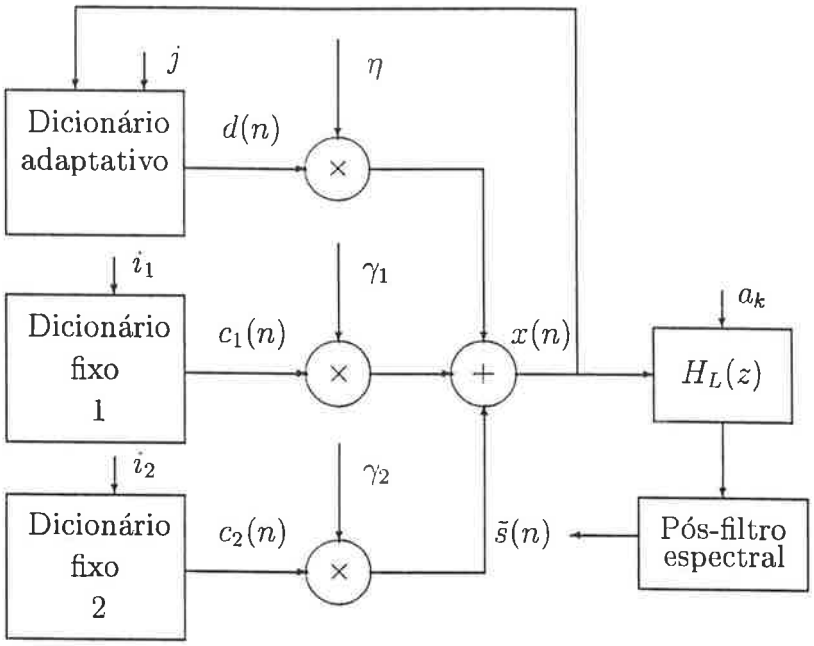

Figura 4.2: Decodificador de voz VSELP.

ii) Os vetores-base são calculados a partir da acumulação apropriada dos parâmetros fornecidos acima ao longo de toda a base de sinais escolhida para treinamento:

$$[b_{1k}(n), b_{2k}(n)] = \mathcal{S}\left(h_m(n), g_m, \eta_m, u_m(n), d_m(n)\right). \tag{4.25}$$

Note-se que o processo *i)* equivale a uma codificação a menos do fornecimento de alguns parâmetros adicionais, que são os vetores-alvo.

A fase *ii)* é o cálculo dos vetores-base o treinamento propriamente dito.

Fase de codificação

Começando com a codificação VSELP (seção 4.8), tem-se vetores- base

$$b_{lk}(n), \quad n = 0, 1, \ldots, L-1 \begin{cases} k = 1, 2, \ldots, M_B & \rightarrow \text{nº de vetores-base por dicionário} \\ l = 1, 2 & \rightarrow \text{nº de dicionários fixos} \end{cases} \tag{4.26}$$

de dimensão $L = 40$, inicialmente preenchidos com vetores amostrais gaussianos de média nula e variância unitária, por exemplo.

Os conjuntos de vetores-código dos dicionários para o sub-bloco m são definidos por

$$c_{lm} = \sum_{k=1}^{M_B} \theta_{lkm} b_{lk}, \quad l = 1, 2 \tag{4.27}$$

onde as ordenadas binárias θ_{lkm} assumem os valores -1 ou 1.

Com o codificador VSELP determinam-se os parâmetros do sinal $s(n)$, composto de M sub-blocos de $L = 40$ amostras.

A cada bloco q de $4L$ amostras, o codificador VSELP calcula um filtro LPC com largura de faixa expandida, cuja resposta impulsiva truncada

$$h_m(n), \quad n = 0, 1, \ldots, L-1$$

é usada durante as buscas da excitação nos sub-blocos $m = 4q-3, 4q-2, \ldots, 4q$.

A cada sub-bloco m, o codificador VSELP identifica um vetor-alvo $u_m(n)$, de dimensão L, que é aproximado após as buscas nos dicionários adaptativo e fixos pelo vetor reconstruído

$$\tilde{u}_m = g_{1m}H_m c_{1m} + g_{2m}H_m c_{2m} + \eta_m H_m d_m, \tag{4.28}$$

onde

$$H_m = \begin{bmatrix} h_m(0) & 0 & \cdots & 0 & 0 \\ h_m(1) & h_m(0) & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ h_m(L-2) & \ddots & h_m(1) & h_m(0) & 0 \\ h_m(L-1) & h_m(L-2) & \cdots & h_m(1) & h_m(0) \end{bmatrix} \tag{4.29}$$

é a matriz de resposta impulsiva truncada no comprimento do sub-bloco do filtro expandido no sub-bloco m.

O vetor reconstruído no sub-bloco m (Eq.(4.28)) pode ser expresso em função dos vetores-base mediante

$$\tilde{u}_m = \sum_{k=1}^{M_B} \gamma_{1km} H_m b_{1k} + \sum_{k=1}^{M_B} \gamma_{2km} H_m b_{2k} + \eta_m H_m d_m, \tag{4.30}$$

onde

$$\gamma_{lkm} = g_{lm}\theta_{lkm},$$

para $k = 1, 2, \ldots, M_B$, $l = 1, 2$, são os ganhos associados a cada vetor-base na composição da excitação.

Para simplificar a notação, na Eq.(4.30) cada vetor-base será expresso por um único índice como b_{k+lM_B} bem como seu respectivo ganho será denotado por $\gamma_{k+lM_B,m}$, resultando

$$\tilde{u}_m = \sum_{k=1}^{2M_B} \gamma_{km} H_m b_k + \eta_m w_m, \tag{4.31}$$

onde aparece o vetor-código adaptativo filtrado, que se determina por

$$w_m = H_m d_m. \tag{4.32}$$

Fase de cálculo dos vetores-base

Agora pode-se definir o problema do treinamento como sendo a determinação dos $2M_B$ vetores-base $b_k, k = 1, 2, \ldots, 2M_B$, cada um composto por L amostras $(b_k(n), n = 0, 1, \ldots, L-1)$ de forma a minimizar o erro quadrático total de reconstrução

$$\|\varepsilon\|^2 = \sum_{m=1}^{M} \|u_m - \tilde{u}_m\|^2, \tag{4.33}$$

dados γ_{im}, H_m, η_m, para $i = 1, 2, \ldots, 2M_B$ e $m = 1, 2, \ldots, M$.

Simplificando, dada a constância da contribuição $\eta_m w_m$ do dicionário adaptativo em relação aos vetores-base, podem-se tomar os vetores-alvo como

$$u'_m = u_m - \eta_m w_m \tag{4.34}$$

e os vetores-alvo reconstruídos da Equação (4.31) têm que ser, conseqüentemente, redefinidos como

$$\tilde{u}_m = \sum_{k=1}^{2M_B} \gamma_{km} H_m b_k. \tag{4.35}$$

Na seqüência, desenvolve-se o erro quadrático de reconstrução restrito ao sub-bloco m

$$\|\varepsilon_m\|^2 = \|u'_m - \tilde{u}_m\|^2. \tag{4.36}$$

Inicia-se por

$$\begin{aligned} \|\varepsilon_m\|^2 &= (u'_m - \tilde{u}_m)^T (u'_m - \tilde{u}_m) \\ &= \|u'_m\|^2 - 2\tilde{u}_m^T u'_m + \|\tilde{u}_m\|^2. \end{aligned} \tag{4.37}$$

Como as variáveis são as componentes de b_k, pode-se abstrair da constante $\|u'_m\|^2$, reescrevendo a Eq.(4.37) como

$$\|\varepsilon_m\|^2 = \|\tilde{u}_m\|^2 - 2\tilde{u}_m^T u'_m + \text{ cte.} \tag{4.38}$$

Em seguida, desenvolvem-se ambas as parcelas variáveis do erro quadrático de reconstrução no sub-bloco.

Primeiramente, a parcela linear nas variáveis pode ser expressa como

$$\tilde{u}_m^T u'_m = u_m'^T \tilde{u}_m = \sum_{k=1}^{2M_B} \gamma_{km} u_m'^T H_m b_k \tag{4.39}$$

Em segundo lugar, a parcela quadrática nas variáveis pode ser desenvolvida como

$$\|\tilde{u}_m\|^2 = \sum_{k=1}^{2M_B} \sum_{j=1}^{2M_B} \gamma_{km}\gamma_{jm} b_k^T H_m^T H_m b_j. \tag{4.40}$$

Tomando-se o gradiente da parcela linear (Eq.(4.39)) em relação a cada vetor base, obtém-se

$$\frac{\partial \tilde{u}_m^T u'_m}{\partial b_k} = \gamma_{km} u_m'^T H_m \tag{4.41}$$

para $k = 1, 2, \ldots, 2M_B$.

Tomando-se o gradiente da parcela quadrática (Eq.(4.40)) em relação a cada vetor base, obtém-se

$$\frac{\partial \|\tilde{u}_m\|^2}{\partial b_k} = 2 \sum_{j=1}^{2M_B} \gamma_{km}\gamma_{jm} H_m^T H_m b_j \tag{4.42}$$

para $k = 1, 2, \ldots, 2M_B$.

Substituindo as Eqs.(4.41) e (4.42) na Eq.(4.37) e adaptando a definição (3.39) de matriz de autocorrelação da resposta impulsiva para

$$\Phi_m = H_m^T H_m, \tag{4.43}$$

obtém-se o sistema de equações lineares

$$A_m b = c_m, \tag{4.44}$$

onde

$$A_m = \begin{bmatrix} \gamma_{1m}^2\Phi_m & \gamma_{1m}\gamma_{2m}\Phi_m & \cdots & \gamma_{1m}\gamma_{2M_B-1,m}\Phi_m & \gamma_{1m}\gamma_{2M_B,m}\Phi_m \\ \gamma_{1m}\gamma_{2m}\Phi_m & \gamma_{2m}^2\Phi_m & \ddots & \gamma_{2m}\gamma_{2M_B-1,m}\Phi_m & \gamma_{2m}\gamma_{2M_B,m}\Phi_m \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \gamma_{1m}\gamma_{2M_B-1}\Phi_m & \ddots & \ddots & \gamma_{2M_B-1,m}\gamma_{2M_B-1,m}\Phi_m & \gamma_{2M_B-1,m}\gamma_{2M_B,m}\Phi_m \\ \gamma_{1m}\gamma_{2M_B}\Phi_m & \ddots & \ddots & \gamma_{2M_B,m}\gamma_{2M_B-1,m}\Phi_m & \gamma_{2M_B,m}\gamma_{2M_B,m}\Phi_m \end{bmatrix}, \tag{4.45}$$

$$c_m = \begin{bmatrix} \gamma_{1m}u_m'^T H_m \\ \gamma_{2m}u_m'^T H_m \\ \cdots \\ \cdots \\ \gamma_{2M_B-1,m}u_m'^T H_m \\ \gamma_{2M_B,m}u_m'^T H_m \end{bmatrix} \tag{4.46}$$

e o vetor de variáveis é

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ b_{2M_B-1} \\ b_{2M_B} \end{bmatrix}. \tag{4.47}$$

Ainda pode-se definir a matriz A_m de forma mais compacta através de operações matriciais com a introdução do vetor de ganhos dos vetores-base

$$\gamma_m \triangleq \begin{bmatrix} \gamma_{1m} & \gamma_{2m} & \cdots & \gamma_{2M_B-1,m} & \gamma_{2M_B,m} \end{bmatrix}^T \tag{4.48}$$

como

$$A_m = (\gamma_m \gamma_m^T) \otimes \Phi_m, \tag{4.49}$$

onde "⊗" representa o produto tensorial de Kronecker.

Finalmente, obtém-se o sistema de equações lineares relativo ao sinal completo

$$Ab = c, \tag{4.50}$$

onde

$$A = \sum_{m=1}^{M} A_m \tag{4.51}$$

$$c = \sum_{m=1}^{M} c_m. \tag{4.52}$$

O algoritmo desenvolvido acima foi usado para o treinamento do codificador VSELP em testes de métodos de tratamento de transitórios relatados na seção 7.9. Registram-se em seguida dados disponíveis na literatura sobre incremento de desempenho com o treinamento.

Primeiramente, para o VSELP da Motorola, fizeram um treinamento em 16 iterações a partir de vetores-base gaussianos [25] e obtiveram uma melhora de desempenho em relação sinal-ruído segmentada ponderada de 0,64 dB. Ou seja, os desempenhos inicial e final e o incremento foram

$$\text{WSNRSEG}(0) = 13,41 \text{ dB}$$

$$\text{WSNRSEG}(16) = 14,05 \text{ dB}$$

$$\Delta\text{WSNRSEG}_{\text{treino}} = 0,64 \text{ dB}.$$

Em segundo lugar, há o caso do codificador CELP com estrutura conjugada (CS-CELP) da NTT [27], que, fazendo um treinamento com 8 iterações, obteveram os incrementos em SNR segmentada e não segmentada dados abaixo

$$\Delta\text{SNRSEG}_{\text{treino}} = 0,5 \text{ dB}$$

$$\Delta\text{SNR}_{\text{treino}} = 1,5 \text{ dB}.$$

Capítulo 5

METODOLOGIA

5.1 Condicionamento dos sinais

Tanto o sinal de voz original antes de ser passado ao algoritmo de codificação quanto o sinal de voz reconstruído após a saída do decodificador requerem um tratamento especial ou condicionamento. Em geral, esses condicionamentos envolvem uma filtragem dos sinais além de algum tratamento de amplitude como amplificação ou atenuação.

Costuma-se incorporar um filtro passa-altas para atenuar as primeiras harmônicas da freqüência do sinal de distribuição de energia elétrica presente no sinal de voz original. Este filtro tem freqüência de corte entre 100 e 200 Hz e comparece nos codificadores VSELP da norma IS-54 [21] e ACELP/MP-MLQ da recomendação G.723.1 [32], usados posteriormente.

Esse condicionamento deve ser levado em conta no uso do codificador de voz pois, se considerado no treinamento dos quantizadores de parâmetros nele embutidos, pode causar uma queda significativa de desempenho se não for utilizado. Os quantizadores de parâmetros representativos de preditores LPC, como os pares de raias espectrais (LSPs) incluem-se nesta categoria.

Por outro lado, para poder aplicar as medidas objetivas de qualidade e de distorção, convém desativar este pré-filtro. A principal razão para isso é o atraso gerado por qualquer filtro que apareça no caminho do sinal.

A neutralização do pré-filtro pode ocorrer segundo os dois procedimentos descritos em seguida. No codificador ACELP, o pré-filtro simplesmente não é solicitado, injetando-se o sinal de voz original diretamente no codificador. Por outro lado, na implementação utilizada do codificador VSELP, optou-se por manter o pré-filtro ativo e extrair o sinal de voz pré-filtrado para ser usado como sinal original nos cálculos de desempenho e de distorção. Coerentemente, o sinal a ser usado como sinal reconstruído nesses cálculos deve ser o sinal de entrada do pós-filtro, que pode estar ativo ou não.

O pós-filtro é aplicado no decodificador sobre o sinal de voz reconstruído. A sua função é a redução do ruído de codificação nas regiões do espectro menos relevantes para o sinal de voz. Em geral, ele é composto por uma cascata de uma seção de

curto prazo $H_{\mathrm{cp}}(z)$ e uma seção de longo prazo $H_{\mathrm{lp}}(z)$ como

$$H_{\mathrm{pf}}(z) = G_{\mathrm{pf}} H_{\mathrm{cp}}(z) H_{\mathrm{lp}}(z). \tag{5.1}$$

onde G_{pf} é um fator geral de ganho.

O pós-filtro de curto prazo enfatiza as regiões das formantes, podendo utilizar a mesma forma adotada para o filtro de ponderação perceptiva na Equação 3.6 da seção 3.5. Porém, como não há interesse em alterar a declividade espectral ("spectral tilt"), torna-se necessário um filtro FIR de primeira ordem em cascata [14], resultando

$$H_{\mathrm{cp}}(z) = \frac{H_{\mathrm{LPC}}(z/\alpha_2)}{H_{\mathrm{LPC}}(z/\alpha_1)} \left(1 - \mu z^{-1}\right), \tag{5.2}$$

sendo, para o codificador ACELP [32], $\alpha_1 = 0,65$, $\alpha_2 = 0.75$ e o fator de declividade é

$$\mu = 0,25 \frac{r(1)}{r(0)}, \tag{5.3}$$

onde $r(k)$ é a seqüência de autocorrelação (seção 2.7) do sinal reconstruído não pós-filtrado.

O pós-filtro de longo prazo é comumente um filtro FIR, embora também possa ser um filtro recorrente, mas sempre com um polinômio de grau maior do que zero no numerador da sua função de transferência. Para o codificador ACELP [32], o pós-filtro de longo prazo é

$$H_{\mathrm{lp}} = g_0 + g_f z^{M_f} + g_b z^{-M_b}, \tag{5.4}$$

onde os atrasos M_f e M_b são os valores absolutos dos índices positivo e negativo, respectivamente, dos máximos positivos da função de correlação cruzada entre vetores extraídos do sinal de excitação composto. Esses atrasos são selecionados na faixa de $P-3$ a $P+3$ amostras, sendo P o índice escolhido do dicionário de códigos adapatativo.

5.2 Sinais e bases de dados de voz

Vários sinais de voz foram usados para testes e para treinamento de codificadores de voz ao longo deste trabalho.

Para executar os primeiros testes dos algoritmos utilizados, em geral recorreu-se ao sinal "v". Ele consiste num trecho de voz sonora masculina de 320 ms de duração, amostrado à taxa de 8 kHz com quantização PCM lei μ de 8 bits linearizada em 13 bits (seção 2.4) e relação sinal-ruído média de 39 dB, que corresponde às duas primeiras sílabas da frase "*um galo*".

Para efetuar testes um pouco mais representativos, usaram-se três conjuntos de sinais: o sinal "flavia", o sinal "3m3f" e um subconjunto de 13 sinais da base de dados TIMIT.

O sinal "flavia" é assim denominado porque nos foi transmitido pela engenheira Flávia Martinho Ferreira do Grupo de Processamento Digital de Voz do CPqD/TELEBRÁS em 1995 durante testes de comparação da nossa implementação do codificador VSELP segundo a norma IS-54 com a implementação que ela tinha desenvolvido para o processador digital de sinais TMS320C30.

O sinal contém duas frases, sendo uma pronunciada por um falante masculino e a outra por um falante feminino, com duração conjunta de 15,84 s, conforme texto abaixo

flavia-m: `Certas surpresas são difíceis de realizar, muitas coisas podem sair erradas. Algumas vezes os erros são realmente engraçados.`

flavia-f: `A profissão de aeromoça parece ser excitante, mas é também muito cansativa. Lidar com pessoas faz parte do trabalho.`

O conjunto de sinais "3m3f" contém uma frase pronunciada por cada um de três falantes masculinos e três falantes femininos com duração total de 21 s. Este é um dos conjuntos de sinais que acompanha a implementação de referência do codificador CELP FS1016 [9].

O conjunto de 13 sinais da base de dados TIMIT abaixo descrita foi extraído da região dialetal da cidade de NY na partição de teste e foram previamente reamostradas na freqüência de 8 kHz após uma filtragem passa-baixas para uma faixa de 3,5 kHz efetuada com um filtro elíptico de 11ª ordem. A duração total deste conjunto de sinais é de 52 s.

Relaciona-se abaixo o texto das 13 frases utilizadas:

sa1: `She had your dark suit in greasy wash water all year.`

sa2: `Don't ask me to carry an oily rag like that.`

si576: `But even mother's loving attitude will not always prevent misbehavior.`

si602: `But the ships are very slow now, and we don't get so many sailors any more.`

si941: `The farmer's life must be arranged to meet the demands of crops and livestock.`

si1283: `Living in a shelter the radioactivity of fallout decays rapidly at first.`

si1354: `Within the larger social system are the structural and functional subsystems.`

si1564: `Husbandry was bounded by snake-rail fences, and there were grazing cattle.`

si2194: He had fallen into a soft job, and now the job was gone and he was stranded.

sx113: A muscular abdomen is good for your back.

sx134: December and January are nice months to spend in Miami.

sx284: Jeff thought you argued in favor of a centrifuge purchase.

sx396: The fish began to leap frantically on the surface of the small lake.

Essas frases foram pronunciadas por um conjunto de 3 falantes femininos e 3 falantes masculinos, tendo uma duração conjunta de 52 s.

Os três falantes femininos e suas respectivas frases são

FDRW0: sa1, si1283 e sx113;

FLNH0: si941 e sx134;

FMGD0: si1564 e si2194.

Os três falantes masculinos e suas respectivas frases são

MCMJ0: si602 e sx284;

MJDH0: sa2 e si1354;

MPAM1: si576 e sx396.

O treinamento e vários testes de codificação foram realizados com o conjunto completo de sinais das partições de teste das bases de dados TIMIT [23] e NTIMIT [36]. Essas partições possuem sinais correspondentes num total de 1680 cada, com uma duração de 1h 26min 27s. A base de dados TIMIT contém ao todo 6300 frases obtidas através de leitura em ambiente silencioso ao passo que a base de dados NTIMIT possui as mesmas frases transmitidas cada uma por uma ligação telefônica com características próprias.

5.3 Relação sinal-ruído segmentada

Dados um sinal de voz $s(n)$, $n = 0, 1, \ldots, N-1$ e um sinal reconstruído $\tilde{s}(n)$, $n = 0, 1, \ldots, P-1$, fragmentam-se os dois sinais em

$$K = \left\lfloor \frac{\min\{N, P\}}{M} \right\rfloor$$

segmentos de comprimento M, obtendo o conjunto de segmentos de referência

$$s_m(n), \quad n = 0, 1, \ldots, M-1, \quad m = 1, 2, \ldots, K$$

e o conjunto de segmentos reconstruídos

$$\tilde{s}_m(n), \quad n = 0, 1, \ldots, M-1, \quad m = 1, 2, \ldots, K.$$

Neste trabalho usaram-se segmentos de comprimento $M = 128$, que têm duração de 16 ms, para efetuar as medidas, conforme sugerido em [37].

Seguindo o procedimento adotado por José Sindi Yamamoto [71] para o cálculo da relação sinal-ruído segmentada, selecionam-se apenas os segmentos com energia maior que um nível de 50 dB abaixo do máximo

$$E_m = 10 \log \left(\max_m \sum_{n=0}^{M-1} s_m^2(n) \right).$$

Desta forma, selecionam-se, em ordem q crescente, todos os segmentos que satisfaçam a relação

$$10 \log \left(\sum_{n=0}^{M-1} s_{m(q)}^2(n) \right) > E_m - 50.$$

Portanto, o número de segmentos selecionado é

$$Q = \text{argmax}\{m(q)\}$$

e, após esta pré-seleção de energia, resultam os conjuntos de segmentos

$$s_{m(q)}(n), \quad n = 0, 1, \ldots, M-1, \quad q = 1, 2, \ldots, Q$$

e

$$\tilde{s}_{m(q)}(n), \quad n = 0, 1, \ldots, M-1, \quad q = 1, 2, \ldots, Q.$$

A partir dos dois conjuntos de segmentos pré-selecionados, obtém-se o conjunto de segmentos do sinal de erro como

$$e_m(n) = s_{m(q)}(n) - \tilde{s}_{m(q)}(n), \quad n = 0, 1, \ldots, M-1, \quad q = 1, 2, \ldots, Q.$$

Finalmente, a relação sinal-ruído segmentada ponderada é calculada como

$$\text{SNRSEG}(s, \tilde{s}) = \frac{10}{Q} \sum_{q=1}^{Q} \log \frac{\sum_{n=0}^{M-1} s_{m(q)}^2(n)}{\sum_{n=0}^{M-1} e_{m(q)}^2(n)}.$$

5.4 A medida objetiva de distorção PSQM

A PSQM ("perceptual speech quality measure") [35] mede a distorção de um sinal de voz codificado em relação a um sinal de voz original. A PSQM simula experimentos de julgamento por ouvintes humanos da qualidade da fala codificada [8].

No cálculo da PSQM, os sinais original e reconstruído são mapeados em representações psicofísicas ou internas do aparelho auditivo humano. O equivalente psicofísico da freqüência são as faixas críticas ("critical bands") ou escala Bark de freqüências e o equivalente da intensidade sonora são os sones comprimidos.

A transformação do domínio físico para o psicofísico é executada em três operações:

- mapeamento tempo-freqüencial:
- conversão não linear da escala de freqüências:
- conversão não linear da escala de intensidades (compressão).

O mapeamento do domínio do tempo para o domínio tempo-freqüência é implementado através de uma transformada de Fourier de curto prazo ("short-term Fourier transform") com uma janela de Hanning

$$w(n) = 0,5\left(1 - \cos\frac{2\pi n}{N_F}\right) \tag{5.5}$$

para $n = 0, 1, \ldots, N_F - 1$, onde $N_F = 512$ amostras para uma freqüência de amostragem de 16 kHz[1] com sobreposição de 50% entre blocos adjacentes. Isto resulta numa representação tempo-freqüencial com resolução constante em ambos os domínios do tempo e da freqüência.

O blocos de índice i do sinal original $x_i(n)$ e do sinal reconstruído $y_i(n)$ são janelados, resultando os blocos

$$\begin{aligned} x_{w_i}(n) &= w(n).x_i(n) \\ y_{w_i}(n) &= w(n).y_i(n). \end{aligned} \tag{5.6}$$

Os espectros de potência amostrados de $x_{w_i}(n)$ e $y_{w_i}(n)$ são calculados após as transformadas rápidas de Fourier (FFT)

$$x_{w_i}(n) \overset{\text{FFT}}{\Longrightarrow} X_i(k)$$

$$y_{w_i}(n) \overset{\text{FFT}}{\Longrightarrow} Y_i(k)$$

como

$$\begin{aligned} P_{x_i}(k) &= X_i^*(k)X_i(k) \\ P_{y_i}(k) &= Y_i^*(k)Y_i(k). \end{aligned} \tag{5.7}$$

Em seguida, converte-se a escala de freqüência linear em Hz na escala não linear de faixas críticas, resultando nos espectros de potência por faixa dentro de cada bloco. O espectro de potência do sinal de voz reconstruído é escalado em cada bloco, quando ambos os sinais original e reconstruído são filtrados na faixa telefônica e têm ruído Hoth adicionado para simular o ambiente de audição. Finalmente, os sinais são filtrados com a função de transferência do ouvido externo para o ouvido interno.

São usadas $N_b = 56$ faixas críticas de largura $\Delta z = 0,312$ Bark, onde a primeira faixa começa em 15,6 Hz e a última faixa termina em 4193 Hz.

[1] A recomendação P.861 também admite um comprimento de bloco $N_F = 256$ para uma freqüência de amostragem de 8 kHz, porém o software que a acompanha apenas opera na freqüência de amostragem de 16 kHz.

As densidades espectrais de potência no domínio Bark na faixa j do bloco i são calculadas através de

$$\begin{aligned} P'_{x_i}(j) &= S_p \frac{\Delta f_j}{\Delta z} \frac{1}{l_l(j) - l_f(j) + 1} \sum_{k=l_f(j)}^{l_l(j)} P_{x_i}(k) \\ P'_{y_i}(j) &= S_p \frac{\Delta f_j}{\Delta z} \frac{1}{l_l(j) - l_f(j) + 1} \sum_{k=l_f(j)}^{l_l(j)} P_{y_i}(k), \end{aligned} \tag{5.8}$$

onde $l_f(j)$ e $l_l(j)$ são a primeira e a última amostras da FFT na faixa em Hertz correspondente à faixa Bark j, Δf_j é a largura da faixa j em Hertz, Δz é sua largura em Barks e S_p é o fator de calibração de potência nas faixas críticas.

Após filtradas com a resposta de freqüência $F(j)$ "intermediate reference system" (IRS) do aparelho telefônico médio e submetidas ao ruído ambiente Hoth [35], as densidades espectrais de potência no domínio Bark do sinal original e do sinal reconstruído na faixa j do bloco i são

$$\begin{aligned} \mathrm{PH}_{x_i}(j) &= H(j).F(j).P'_{x_i}(j) \\ \mathrm{PH}_{y_i}(j) &= H(j).F(j).P'_{y_i}(j) \end{aligned} \tag{5.9}$$

As funções amostradas de densidade de intensidade ("loudness") comprimida são obtidas através da função de Zwicker [35] como

$$\begin{aligned} L_{x_i}(j) &= S_l . \left(\frac{P_0(j)}{0,5}\right)^{\gamma} . \left[\left(0,5 + 0,5.\frac{\mathrm{PH}_{x_i}(j)}{P_0(j)}\right)^{\gamma} - 1\right] \\ L_{y_i}(j) &= S_l . \left(\frac{P_0(j)}{0,5}\right)^{\gamma} . \left[\left(0,5 + 0,5.\frac{\mathrm{PH}_{y_i}(j)}{P_0(j)}\right)^{\gamma} - 1\right], \end{aligned} \tag{5.10}$$

onde $P_0(j)$ é o limiar de audibilidade na faixa j e S_l é o fator de calibração da intensidade nas faixas críticas. Se forem encontrados valores negativos de $L_{x_i}(j)$ ou $L_{y_i}(j)$, eles são zerados.

O valor ótimo $\gamma = 0,001$ do expoente γ foi determinado com as bases de dados de vários experimentos de avaliação de qualidade de fala codificada.

Para utilização na modelagem cognitiva, são calculadas as intensidades comprimidas momentâneas (em Sones comprimidos) pelas somatórias das densidades nas Eqs. (5.10) como

$$\begin{aligned} L_{x_i} &= \sum_{j=1}^{N_b} L_{x_i}(j).\Delta z \\ L_{y_i} &= \sum_{j=1}^{N_b} L_{y_i}(j).\Delta z. \end{aligned} \tag{5.11}$$

A primeira aproximação da distorção PSQM é obtida com a densidade amostrada de perturbação ruidosa $N_i(j)$, que é obtida a partir do valor absoluto da diferença entre as intensidades comprimidas L_{x_i} e L'_{y_i}, sendo esta última a versão escalada em intensidade ("loudness scaling") de L_{y_i} relativamente a L_{x_i}.

Finalmente, a medida PSQM é obtida pela aplicação a $N_i(j)$ de dois efeitos cognitivos

- processamento de assimetria:
- processamento dos intervalos de silêncio.

Quando uma nova componente tempo-freqüencial é introduzida num sinal de voz, a qualidade subjetiva degrada-se mais que quando uma componente de igual intensidade é bloqueada pelo codificador de voz. Esta assimetria é mais acentuada nos intervalos de silêncio.

Os intervalos de silêncio são determinados através de um limiar de potência calculado a partir de um fator de calibração global e as distorções calculadas sobre os blocos ativos e sobre os blocos silenciosos são ponderadas para o calculo final da PSQM.

5.5 Aplicação da PSQM

Antes de aplicar a PSQM, os sinais de voz original e reconstruído têm que ser reamostrados à taxa de 16 kHz. Como filtro antiimagem, usou-se o filtro FIR simétrico com 118 coeficientes de 24 bits definido em [31]. A representação em número fixo de bits permite implementações bit-exatas do filtro em hardware.

Capítulo 6

BUSCA CONJUNTA DE AMPLITUDE E POSIÇÃO

Neste capítulo, propõe-se um novo processo de busca conjunta de atributos de inovações compostas por multipulsos algébricos (AMPE-JAPS). Como ele será aplicado à excitação algébrica do ACELP da recomendação G.723.1 [32], apresenta-se inicialmente o processo de busca focalizada usado na sua implementação de referência.

Ademais, como referência de complexidade e de desempenho, apresenta-se um processo de busca exaustiva das inovações, inserido nos moldes da implementação de referência.

6.1 A busca exaustiva de posições no dicionário ACELP

A busca básica da excitação CELP envolve (seção 3.8):

- o cálculo do produto interno C_i do vetor-alvo filtrado regressivamente com o vetor-código c_i;
- o cálculo da energia $\sigma_{q_i}^2$ do vetor-código filtrado q_i causado pelo vetor-código c_i.

Dada a natureza dos vetores-código ACELP, o produto interno C_i compõe-se de quatro amostras do vetor-alvo residual r filtrado regressivamente

$$C_i = \sum_{k=0}^{3} \alpha_k t(m_k), \tag{6.1}$$

onde

$$t = \left(r^T H\right)^T$$

é o vetor-alvo residual filtrado regressivamente.

Embora o objetivo da busca seja a maximização da razão $\tau_i = \frac{C_i^2}{\sigma_{q_i}^2}$, faz-se, em primeiro lugar, uma simplificação: a maximização de C_i.

Desta forma, facilmente determina-se que

$$\alpha_k = \text{sign}\left(t(m_k)\right), \tag{6.2}$$

para $k = 0, 1, 2, 3$, maximiza C_i, cujo valor máximo é

$$C = \sum_{k=0}^{3} |t(m_k)|. \tag{6.3}$$

De fato, as equações (6.2) são aplicadas no codificador G.729 [33], cujo dicionário ACELP possui uma única grade cobrindo tanto as posições pares como as ímpares.

Entretanto, o codificador ACELP G.723.1 possui grades separadas para as posições pares e ímpares, adotando-se, para simplificar a busca, os mesmos sinais para ambas, que são definidos como

$$\alpha(m_k) = \begin{cases} \text{sign}\left(t(m_k)\right) & \text{se } |t(m_k)| > |t(m_k+1)| \\ \text{sign}\left(t(m_k+1)\right) & \text{caso contrário} \end{cases} \tag{6.4}$$

para $m_k = i_k, i_k + 8, \ldots, i_k + 7 \times 8$ e $i_k = 0, 2, 4, 6$ (seção 4.4).

Na Eq. (6.4) usa-se a notação $\alpha(m_k)$ para denotar um sinal de sinais algébricos sobre todos os pontos da fase i_k da grade par. Para referir-se ao sinal da fase i_k de um caminho específico i sobre a grade, será mantida a notação simplificada α_k.

Não obstante a grande simplificação ocasionada pela álgebra de sinais, a busca como um todo tem sua complexidade determinada pelo cálculo das energias dos vetores-código filtrados, que não pode se beneficiar da filtragem regressiva logo de início.

Porém, dada a alta esparsidade dos vetores-código, é compensador o cálculo da matriz Φ de autocorrelação da resposta impulsiva (seção 3.8) porque a energia pode, então, ser determinada por adições ou subtrações de poucos elementos de Φ.

Especificamente, a energia do vetor-código filtrado $q_i = Hc_i$ é

$$\sigma_{q_i}^2 = \sum_{k=0}^{3} (\alpha_k)^2 \phi(m_k, m_k) + \sum_{k=0}^{3} \sum_{k \neq l = 0}^{3} \alpha_k \alpha_l \phi(m_k, m_l). \tag{6.5}$$

Notando que Φ é simétrica (Eq. (3.39)) e que as amplitudes são ± 1, a energia pode ser calculada por

$$\sigma_{q_i}^2 = \sum_{k=0}^{3} \phi(m_k, m_k) + 2 \sum_{k=0}^{2} \sum_{l=k+1}^{3} \alpha_k \alpha_l \phi(m_k, m_l). \tag{6.6}$$

A matriz de autocorrelação Φ possui $L_2^2 = 4096$ elementos, onde $L_2 = 64$ é o comprimento do sub-bloco estendido até a próxima potência de 2.

Como o algoritmo de referência usa, por simplicidade, apenas a grade de posições pares, a submatriz de autocorrelação de interesse fica reduzida a $\left(\frac{L_2}{2}\right)^2 = 1024$ elementos.

Entretanto, considerando-se que é selecionada apenas uma posição de cada fase da grade (Tabela 6.1) e lembrando-se de que a submatriz é simétrica, decorre que

se devem calcular todos os 32 elementos da diagonal e todas as $\binom{4}{2}$ combinações duas a duas das fases com todos os 8^2 pares de posições. Ou seja, o número de elementos de Φ que têm que ser calculados reduz-se a

$$\begin{aligned} N_\Phi &= 32 + \binom{4}{2} \times 8^2 \\ &= 416. \end{aligned} \tag{6.7}$$

A correlação C_i (Eq. (6.1)) pode claramente ter seu valor acumulado parcialmente por pulso. O mesmo acontece com a energia $\sigma^2_{q_i}$, o que pode ser visto claramente reescrevendo a Eq. (6.6) como

$$\begin{aligned} \sigma^2_{q_i} &= \phi(m_0, m_0) \\ &+ \phi(m_1, m_1) + 2\alpha_0\alpha_1\phi(m_0, m_1) \\ &+ \phi(m_2, m_2) + 2\alpha_0\alpha_2\phi(m_0, m_2) + 2\alpha_1\alpha_2\phi(m_1, m_2) \\ &+ \phi(m_3, m_3) + 2\alpha_0\alpha_3\phi(m_0, m_3) + 2\alpha_1\alpha_3\phi(m_1, m_3) + 2\alpha_2\alpha_3\phi(m_2, m_3), \end{aligned} \tag{6.8}$$

onde as parcelas a serem acumuladas a cada pulso estão na mesma linha. Nota-se que a complexidade combinatória cresce com a fase dos pulsos, característica que permite que a busca seja implementada por laços encaixados ("nested loops") associados às fases dos pulsos.

Essa estrutura de fluxo da implementação de referência do ACELP G.723.1 facilita a busca paralela nas grades par e ímpar. Isso é feito até a acumulação do terceiro pulso, quando a maior correlação acumulada determina a seleção da grade par ou da grade ímpar.

A busca exaustiva das posições varre todas as possíveis combinações com um único pulso de cada fase:

$$\begin{aligned} N_{\mathrm{EX}} &= 8^4 \\ &= 4096. \end{aligned} \tag{6.9}$$

6.2 A busca focalizada

Com a busca ACELP implementada dentro de quatro laços encaixados (seção 6.1), é possível reduzir sua complexidade controlando a entrada e a permanência no quarto laço.

A entrada no quarto laço, que acumula o efeito de cada pulso de fase i_3, apenas é permitida se a correlação acumulada até o terceiro pulso no caminho de busca $C_{i,3} = \sum_{k=0}^{2} \alpha_k t(m_k)$ satisfizer

$$C_{i,3} > C_L, \tag{6.10}$$

onde o limiar de correlação C_L é o máximo entre os valores determinados para a grade par, C_{Lp}, e para a grade ímpar, C_{Li},

$$C_L = \max\{C_{Lp}, C_{Li}\}.$$

O limiar de correlação, C_{Lp}, para a grade par é

$$C_{Lp} = \frac{1}{2}\left(\frac{1}{4096}\sum_{i=0}^{4095} C_{i,3} + \max_{i=0,1,\ldots,4095}\{C_{i,3}\}\right). \tag{6.11}$$

De forma análoga, define-se o limiar C_{Li} para a grade ímpar.

Os cálculos do valor médio e do valor máximo das correlações, indicados na Eq. (6.11), podem ser efetuados apenas sobre os pontos da grade porque cada posição comparece na determinação de um número equivalente de vetores-código. Assim, a Eq. (6.11) pode ser escrita como

$$C_{Lp} = \frac{1}{2}\left(\frac{1}{8}\sum_{j=0}^{7}\sum_{i=0}^{2}\alpha(8j+i)t(8j+i) + \sum_{i=0}^{2}\max_{j=0,1,\ldots,7}\{\alpha(8j+i)t(8j+i)\}\right). \tag{6.12}$$

A permanência no quarto laço é controlada a cada sub-bloco, restringindo-se o número de caminhos que serão trilhados integralmente. Este número é variável dependendo da história das últimas buscas.

Na seção 6.4 apresentam-se os números de comparações efetivamente efetuadas para alguns sinais de voz de teste.

6.3 A busca conjunta de amplitude e posição

A busca conjunta de amplitude e posição em dicionários compostos por multipulsos algébricos (AMPE-JPAS) é descrita nesta seção para o dicionário fixo definido na recomendação G.723.1 para a taxa de 5,3 kbit/s.

Neste dicionário, cada inovação multipulso algébrica contém $M = 4$ pulsos e são emitidas na taxa de uma a cada 7,5 ms. Ilustrando geometricamente a descrição apresentada na seção 4.4, as excitações admissíveis para o codficador ACELP situam-se nos vértices de hipercubos com centro na origem, totalmente contidos em subespaços de dimensão 4.

Tendo como motivação a composição dos vetores-código filtrados a partir de respostas impulsivas deslocadas dada por

$$q(n) = G\sum_{k=0}^{M-1}\alpha_k h(n-m_k), n = 0,1,\ldots,L-1, \tag{6.13}$$

o processo de busca conjunta seleciona, uma por vez, cada uma das M respostas impulsivas deslocadas que vai definir por acumulação uma direção parcial sobre a qual o vetor-alvo residual proporciona a maior projeção.

O algoritmo de busca seleciona uma posição de pulso e sua respectiva amplitude por iteração, envolvendo 4 iterações ao todo. O processo inicia-se com a determinação das projeções do vetor-alvo residual r sobre vetores de resposta impulsiva atrasados na primeira iteração. Nas iterações seguintes, projeta-se o vetor r sobre as direções parciais de projeção, que são diagonais filtradas de hipercubos centrados na origem.

Esse procedimento mantém as posições dos pulsos obtidos, que são usadas conjuntamente, sujeitando as respectivas amplitudes a uma redefinição para a determinação da direção parcial de projeção seguinte.

Na primeira iteração, seleciona-se a posição $i = m_0$ que maximize a norma quadrada da projeção

$$\tau_i = \frac{\left(r^T H(:,i)\right)^2}{\Phi(i,i)}, \tag{6.14}$$

onde H é a matriz de resposta impulsiva do filtro de síntese ponderado e Φ é sua matriz de autocorrelação (seção 3.8).

A cada iteração $j = 1, 2, 3$, seleciona-se a posição m_j que satisfaça

i) $\tau_i^j = \frac{\left(r^T P^j(:,i)\right)^2}{\mathcal{P}^j(i,i)}$;

ii) $\upsilon_i^j = \frac{\left(r^T S^j(:,i)\right)^2}{\mathcal{S}^j(i,i)}$;

iii) $I = \underset{i}{\text{argmax}} \left\{\tau_i^j\right\}$;

iv) $J = \underset{i}{\text{argmax}} \left\{\upsilon_i^j\right\}$;

v) $m_j = \begin{cases} I & \text{se } \tau_I^j = \max\{\tau_I^j, \upsilon_J^j\} \\ J & \text{caso contrário} \end{cases}$

onde

$$p^j(n,i) = h(n,i) + q_j(n) \tag{6.15}$$

$$\mathcal{P}^j = (P^j)^T P^j \tag{6.16}$$

$$s^j(n,i) = h(n,i) - q_j(n) \tag{6.17}$$

$$\mathcal{S}^j = (S^j)^T S^j \tag{6.18}$$

Explicitam-se em seguida as matrizes de origem dos elementos usados na Equação (6.15) e na Equação (6.17). A matriz de resposta impulsiva do filtro de síntese ponderado (seção 3.6) é dada em função dos seus elementos como

$$H = [h(n,i)],$$

onde n é o índice temporal e i é o índice de atraso. Analogamente, a matriz de direções parciais primárias de projeção para a iteração j é dada por

$$P^j = \left[p^j(n,i)\right]$$

e a matriz de direções parciais secundárias de projeção para a iteração j é dada por

$$S^j = \left[s^j(n,i)\right].$$

Além disso, o conjunto de índices de atraso i pesquisado depende da ordem j da iteração. Para a primeira iteração ($j = 0$), todos os atrasos no intervalo do sub-bloco

são pesquisados, isto é, $i = 0, 1, \ldots . L-1$. Para as iterações seguintes ($j = 1, 2, 3$), são eliminadas da busca todas as localizações associadas às fases dos pulsos definidos nas iterações anteriores. Isto é, elimina-se da busca uma linha da grade de posições (Tabela 6.1) por iteração, sendo que a paridade da grade é definida pela paridade do pulso buscado na primeira iteração.

Dentre as direções parciais de projeção admissíveis na iteração j, definidas pelas Equações (6.15) e (6.17), seja a direção parcial de projeção selecionada expressa por

$$f_j = H(:, m_j) + \sigma_{j-1} q_{j-1}, \tag{6.19}$$

onde

$$\sigma_{j-1} = \begin{cases} 1 & \text{se diagonal principal} \\ -1 & \text{se diagonal secundária} \end{cases} \tag{6.20}$$

é o sinal da diagonal escolhida no plano formado pelo j-ésimo pulso com o subvetor-código disponível c_{j-1} e

$$q_{j-1} \triangleq H c_{j-1} \tag{6.21}$$

é o subvetor-código filtrado.

Projetando-se o vetor-alvo residual r sobre a direção parcial de projeção selecionada f_j, obtém-se o vetor-alvo parcialmente reconstruído

$$\tilde{r}_j = A_j f_j, \tag{6.22}$$

onde A_j é o ganho com sinal, do qual se toma apenas o sinal

$$\beta_j = \text{sign}(A_j), \tag{6.23}$$

enquanto o valor absoluto do ganho

$$G_j = |A_j| \tag{6.24}$$

será convertido no parâmetro ganho propriamente dito apenas após a busca do último pulso.

Assim, o subvetor-código filtrado após a busca do j-ésimo pulso é dado por

$$q_j = \beta_j f_j. \tag{6.25}$$

Generalizando, o subvetor-código filtrado pode ser expresso em função das respostas impulsivas atrasadas selecionadas e dos sinais após cada pulso buscado como

$$q_j = \begin{cases} \beta_j H(:, m_j) & \text{se } j = 3 \\ \beta_j H(:, m_j) + \beta_j \sum_{l=0}^{j-1} \prod_{m=l}^{j-1} \sigma_m \beta_m H(:, m_l) & \text{caso contrário.} \end{cases} \tag{6.26}$$

Finalmente, o vetor-código selecionado é $c = c_3$

$$c = \sum_{j=0}^{3} \beta_j \delta(n - m_j), n = 0, 1, \ldots, L-1, \tag{6.27}$$

cujo vetor-código filtrado é $q = q_j$, dado por

$$q = \beta_3 H(:, m_3) + \beta_3 \sum_{l=0}^{2} \prod_{m=l}^{2} \sigma_m \beta_m H(:, m_l) \tag{6.28}$$

com o correspondente vetor-alvo reconstruído

$$\tilde{r} = Gq. \tag{6.29}$$

Substituindo o índice de fase k dos pulsos pelo índice de busca j na Eq. (6.13), obtém-se

$$q(n) = G \sum_{j=0}^{M-1} \alpha_j h(n - m_j), n = 0, 1, \ldots, L-1. \tag{6.30}$$

Assim, observa-se na comparação da Eq. (6.30) com a Eq. (6.26) que a composição dos sinais de cada pulso é dada por

$$\alpha_j = \begin{cases} \beta_j & \text{se } j = 3 \\ \beta_j \prod_{m=j}^{2} \sigma_m \beta_m & \text{caso contrário.} \end{cases} \tag{6.31}$$

Como apreciação comparativa, note-se que este processo de busca usa o mesmo critério da busca básica do CELP (seção 3.8), porém, aplicado a cada direção parcial de projeção admissível, $P^j(:, i)$ e $S^j(:, i)$. A diferença da busca conjunta com a busca CELP está na natureza das direções parciais de projeção, que não são propriamente vetores-código filtrados. Porém, a quarta e última direção de projeção constitui, a menos de um sinal algébrico, um vetor-código filtrado. Portanto, apesar da semelhança da busca conjunta com a busca multipulso (seção 4.6), esta montagem do vetor-código filtrado sendo inexistente na busca multipulso, classifica a busca conjunta mais próxima da busca CELP num dicionário de códigos multiestágio em que cada estágio contribui com um dos pulsos (seção 3.11).

Para maior abrangência dos métodos de redução da complexidade da busca, devem ser citados os métodos que usam dispositivos de pré-seleção, seja através de comparações euclidianas diretas com o próprio vetor-alvo residual [28] ou obtendo vetores-código filtrados com o uso da resposta impulsiva altamente truncada [49]. Além desses, outros métodos mais recentes de buscas eficientes de dicionários compostos por multipulsos algébricos são mencionados no Capítulo 8.

6.4 Complexidade dos processos de busca

Em contraste com as buscas focalizada e exaustiva nas posições, a busca conjunta de atributos de cada pulso não precisa usar elementos fora da diagonal principal da matriz de autocorrelação da resposta impulsiva (seção 6.3) como a busca exaustiva em posições e a busca focalizada. Ademais, a busca conjunta pode aproveitar a informação de comprimento do sub-bloco menor que $L_2 = 64$.

Os elementos diagonais de autocorrelação são tomados da matriz Φ na iteração $j = 0$ e das matrizes P^j e S^j nas iterações seguintes (seção 6.3). Portanto, o número de autocorrelações calculadas é igual ao número de buscas N_{JPAS}, isto é

$$N_{\text{AUTO}} = N_{\text{JPAS}}, \tag{6.32}$$

Tabela 6.1: Grade de posições pares do ACELP.

Fase	Posições
$i_0 = 0$	0, 8, 16, 24, 32, 40, 48, 56
$i_1 = 2$	2, 10, 18, 26, 34, 42, 50, 58
$i_2 = 4$	4, 12, 20, 28, 36, 44, 52, (60)
$i_3 = 6$	6, 14, 22, 30, 38, 46, 54, (62)

que será calculado em seguida.

Nota-se, em primeiro lugar, que deve ser buscada a posição de cada pulso para o sinal positivo e o negativo da combinação com o vetor filtrado acumulado. Assim, pode-se contar o número de buscas assumindo a amplitude do pulso dada e, depois, dobrar o resultado obtido para as iterações $j = 1, 2, 3$. Tem-se uma exceção no caso da iteração $j = 0$ em que o vetor filtrado acumulado é nulo.

A escolha da grade par ou da grade ímpar é feita com a determinação do pulso mais significativo, que envolve, portanto, $L = 60$ comparações.

Observando a grade de posições pares (Tabela 6.1), constata-se que a busca do 2º pulso mais significativo envolve, por ordem de fase de varredura da grade, (0,8,7,7), (8,0,7,7), (8,8,0,7) ou (8,8,7,0) buscas. Assim, tem-se, em média, 22,5 comparações para determinar o 2º pulso.

Prosseguindo, na busca do 3º pulso mais significativo há, por ordem de fase de varredura da grade de posições, (0,0,7,7), (0,8,0,7), (8,0,0,7) (8,0,7,0) ou (8,8,0,0) comparações. Isto resulta numa média de 15 comparações.

Por último, para a busca do último pulso, são necessárias 7,5 comparações em média.

Acumulando essas contagens parciais, totaliza-se o número médio de buscas

$$\begin{aligned} N_{\text{JPAS}} &= 60 + 2(22,5 + 15 + 7,5) \\ &= 150, \end{aligned} \tag{6.33}$$

que pode ser expresso em função do comprimento L do sub-bloco de excitação como

$$\begin{aligned} N_{\text{JPAS}} &= L + 2\left(\frac{3}{4}\frac{L}{2} + \frac{1}{2}\frac{L}{2} + \frac{1}{4}\frac{L}{2}\right) \\ &= \frac{5}{2}L. \end{aligned} \tag{6.34}$$

Além do valor médio, pode-se obter a distribuição completa do número de vetores-código pesquisados por vez. Para tanto, observa-se que os números de localizações na seqüência de fases pesquisada podem ser agrupados nas seis classes seguintes com correspondentes números de vetores-código pesquisados em cada uma.

i) (8,8,7,7)

$$N_{\text{JPAS}} = 60 + 2(22 + 14 + 7) = 146$$

ii) (8,7,8,7)

$$N_{\text{JPAS}} = 60 + 2(22 + 15 + 7) = 148$$

iii) (8,7,7,8)

$$N_{\text{JPAS}} = 60 + 2(22 + 15 + 8) = 150$$

iv) (7,8,8,7)

$$N_{\text{JPAS}} = 60 + 2(23 + 15 + 7) = 150$$

v) (7,8,7,8)

$$N_{\text{JPAS}} = 60 + 2(23 + 15 + 8) = 152$$

vi) (7,7,8,8)

$$N_{\text{JPAS}} = 60 + 2(23 + 16 + 8) = 154$$

Nota-se, ainda, que cada dupla (8,8) de números de localizações pode resultar de uma das duas duplas ordenadas de fases (i_0, i_1) ou (i_1, i_0) e cada dupla (7,7) pode advir de uma das duas duplas ordenadas de fases (i_2, i_3) ou (i_3, i_2) Assim, cada quarteto ordenado acima enumerado pode ocorrer de 4 maneiras distintas, resultando na distribuição de probabilidade representada na Fig. 6.1. Desta distribuição, infere-se o número médio de 150 vetores-código pesquisados, de acordo com o valor obtido indenpendentemente pela Equação (6.33).

Efetuaram-se testes de codificação com buscas exaustiva, focalizada e conjunta para os três sinais "v", "flavia" e "sa1" e para o conjunto de sinais da partição de teste da base de dados TIMIT (seção 5.2), cujas estatísticas de números de buscas e de coeficientes de autocorrelação calculados por sub-bloco encontram-se nas Tabelas 6.2, 6.3, 6.4 e 6.5, respectivamente. O sinal "sa1" foi retirado do subconjunto colhido da falante FAKS0 da região dialetal 1 da partição de teste da base de dados TIMIT.

Os sinais "v", "flavia" e "sa1" contêm 40, 2112 e 528 sub-blocos de excitação, respectivamente, enquanto os 1680 sinais que compõem a partição de teste da base de dados TIMIT possuem em conjunto 688.244 sub-blocos.

Os testes com o curto sinal "v" apresentam resultados contraditórios. A busca exaustiva perde para a busca focalizada em SNRSEG, embora ganhe em PSQM (Tabela 6.6). A busca conjunta também ganha em SNRSEG da busca exaustiva e ganha em PSQM da busca focalizada.

A redução de complexidade proporcionada pela busca conjunta é notável tanto quando expressa pelo número de comparações por busca quanto pelo número de coeficientes de autocorrelação que têm que ser calculados. Quanto ao desempenho, afora o sinal "v", nota-se que a busca conjunta perde apenas 0,2 dB em relação às buscas exaustiva e focalizada. Sobretudo, para o maior conjunto de sinais, a distorção média medida pela PSQM é 0,14 pontos inferior à distorção média da busca focalizada (Tabela 6.6).

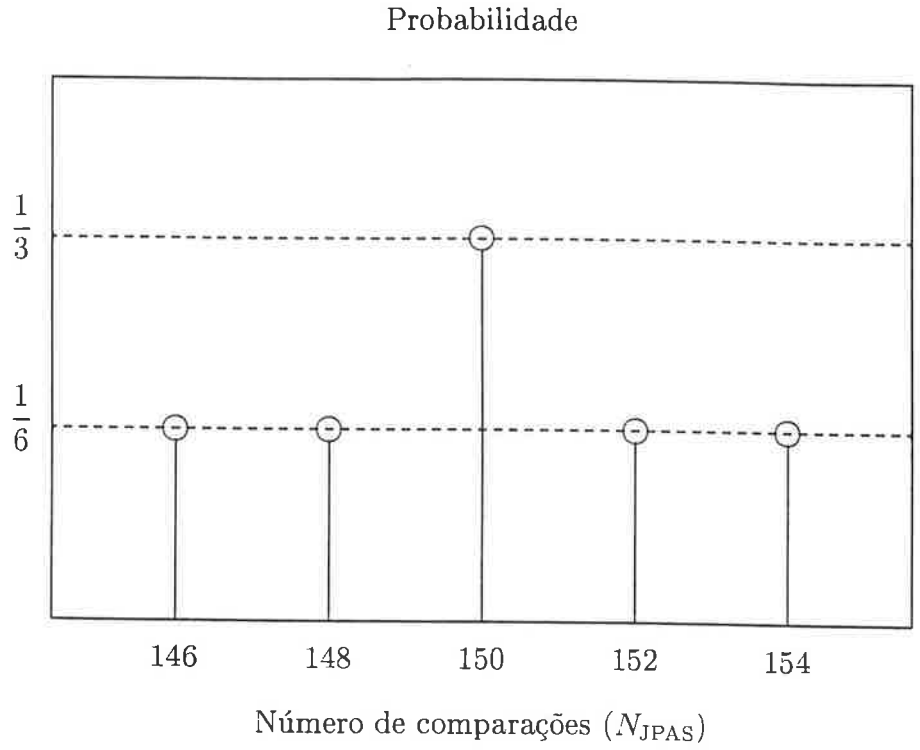

Figura 6.1: Distribuição de probabilidade do número de comparações por busca conjunta AMPE-JPAS.

Tabela 6.2: Estatísticas de comparações e cálculos de autocorrelações por sub-bloco para 3 algoritmos de busca sobre o sinal "v".

Busca	Comparações			Autocorrelações		
	Mínimo	Médio	Máximo	Mínimo	Médio	Máximo
Focalizada	208	389,33	656	416	416,00	416
Exaustiva	4096	4096,00	4096	416	416,00	416
Conjunta	146	150,05	154	146	150,05	154

Tabela 6.3: Estatísticas de comparações e cálculos de autocorrelações por sub-bloco para 3 algoritmos de busca sobre o sinal "flavia".

Busca	Comparações			Autocorrelações		
	Mínimo	Médio	Máximo	Mínimo	Médio	Máximo
Focalizada	144	448,57	1072	416	416,00	416
Exaustiva	4096	4096,00	4096	416	416,00	416
Conjunta	146	149,66	154	146	149,66	154

Tabela 6.4: Estatísticas de comparações e cálculos de autocorrelações por sub-bloco para 3 algoritmos de busca sobre o sinal "sa1" e medidas.

Busca	Comparações			Autocorrelações		
	Mínimo	Médio	Máximo	Mínimo	Médio	Máximo
Focalizada	152	466,69	1032	416	416,00	416
Exaustiva	4096	4096,00	4096	416	416,00	416
Conjunta	146	149.70	154	146	149,70	154

Tabela 6.5: Estatísticas de comparações e cálculos de autocorrelações por sub-bloco para 3 algoritmos de busca sobre a partição de teste da base de dados TIMIT.

Busca	Comparações			Autocorrelações		
	Mínimo	Médio	Máximo	Mínimo	Médio	Máximo
Focalizada	72	458,48	2040	416	416,00	416
Exaustiva	4096	4096,00	4096	416	416,00	416
Conjunta	146	149.62	154	146	149,62	154

Tabela 6.6: Desempenho das buscas ACELP para três sinais e para a partição de teste da base TIMIT.

Busca	SNRSEG (dB)				PSQM			
	v	flavia	sa1	TIMIT	v	flavia	sa1	TIMIT
Focalizada	13,8168	9,1492	9,2661	9,4216	2,6698	2,2958	2,5580	2,2800
Exaustiva	13,0996	9,1734	9,4280	9,4467	2,5359	2,2855	2,5742	2,0695
Conjunta	13,2048	8,9516	9,1834	9,1816	2,6516	2,3586	2,6484	2,1434

Capítulo 7

DINÂMICA DA FILTRAGEM E TRANSITÓRIOS

7.1 Motivação

O estímulo latente para abordar com mais detalhes o processo de filtragem em codificadores preditivos veio ao autor quando da primeira redação sobre os aspectos variantes no tempo da filtragem relacionados com o mecanismo de transferência dos estados em [2]. Esse assunto está coberto na seção 3.3.

Cumulativamente, o estímulo que realmente disparou esta linha de pesquisa apareceu ao analisar a conformidade numérica de implementações distintas do codificador VSELP da norma IS-54 [21]. Esperava-se que as diferenças nos resultados numéricos fossem causadas por diferenças computacionais entre implementações diferentes. Porém, elas deviam-se a decisões de alto nível dos projetistas quanto à estrutura usada para o filtro de síntese. Os resultados desse trabalho sobre estruturas de filtros foram publicados em [55] e [56] e serão abordados na seção 7.4.

Interessantemente, um dos resultados desta pesquisa é a unificação da questão da transferência do vetor de estados do filtro de síntese com a questão da sua estrutura quando se tratam os transitórios gerados pela variação paramétrica do filtro.

7.2 O sinal ideal de excitação

Tendo-se determinado um filtro de predição e não estando restritos pela taxa de transmissão, o melhor sinal de excitação é o sinal residual.

O sinal residual obtém-se da passagem do sinal de voz original pelo filtro de análise. O filtro de análise é o inverso do filtro de síntese e apresenta um efeito de uniformização espectral ou branqueamento (Capítulo 4).

Nesta seção introdutória pretende-se mostrar que, na condição de transmissão perfeita do sinal residual, a reconstrução do sinal de voz também é perfeita, qualquer que seja a estrutura adotada para os filtros de análise e de síntese.

Para fixar a argumentação, serão abordados os casos de duas estruturas de filtros, a forma direta I e a estrutura em treliça.

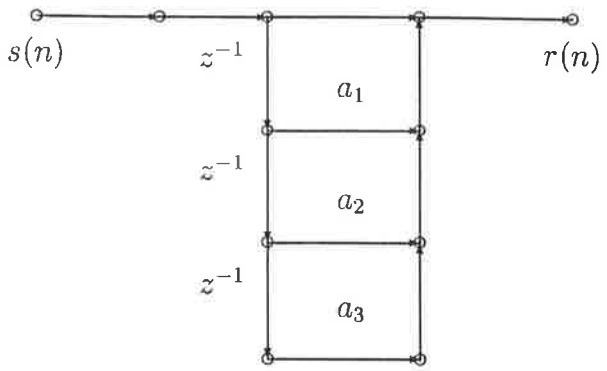

Figura 7.1: Filtro de análise transversal.

Seja feita, sem perda de generalidade, uma predição linear de ordem $p = 3$, obtendo-se o filtro de síntese

$$H(z) = \frac{1}{A(z)}, \tag{7.1}$$

onde o filtro inverso, usado na análise, é dado por

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}. \tag{7.2}$$

Na implementação transversal do filtro de análise (Fig. 7.1), o sinal residual de saída é dado em função do sinal de voz de entrada por

$$r(n) = s(n) + a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3), \tag{7.3}$$

onde se prescinde da representação individual dos estados, que são precisamente as três amostras atrasadas do sinal de voz que acabaram de ser utilizadas.

A implementação do filtro de síntese deve ser feita na forma direta I. Entretanto, embora trivial, será feita a sua derivação algébrica com considerações do fluxo do sinal. Esta abordagem conduzirá a um procedimento que poderá ser seguido também quando a estrutura do filtro não tiver uma inversão conhecida.

Para a obtenção do sinal de voz reconstruído $\tilde{s}(n)$, devem-se usar as suas próprias amostras passadas, que são, por hipótese, reconstruções perfeitas das amostras passadas do sinal original $s(n)$. Algebricamente, substituem-se as ocorrências do sinal $s(n)$ na Eq. (7.3) pelas correspondentes amostras do sinal $\tilde{s}(n)$ e extrai-se a amostra atual deste como (Fig. 7.3)

$$\tilde{s}(n) = r(n) - a_1 \tilde{s}(n-1) - a_2 \tilde{s}(n-2) - a_3 \tilde{s}(n-3). \tag{7.4}$$

Tem-se, portanto, a reconstrução perfeita da amostra $s(n)$ por $\tilde{s}(n)$, dadas as reconstruções perfeitas de $s(n-1)$, $s(n-2)$ e $s(n-3)$.

Nota-se, ademais, que a reconstrução perfeita resiste a uma mudança paramétrica. Sejam a'_1, a'_2 e a'_3 os novos coeficientes de predição linear a partir do instante

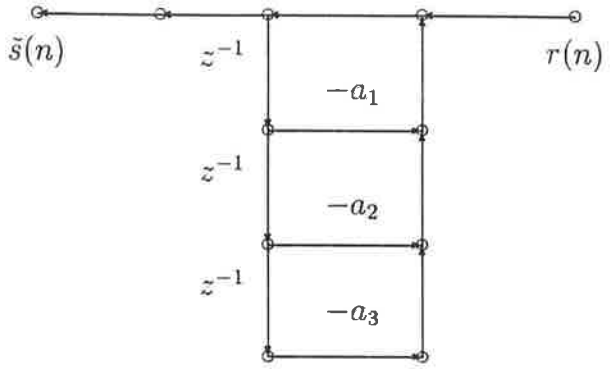

Figura 7.2: Filtro de síntese na forma direta I obtido por inversão do fluxo de sinal do filtro de análise transversal.

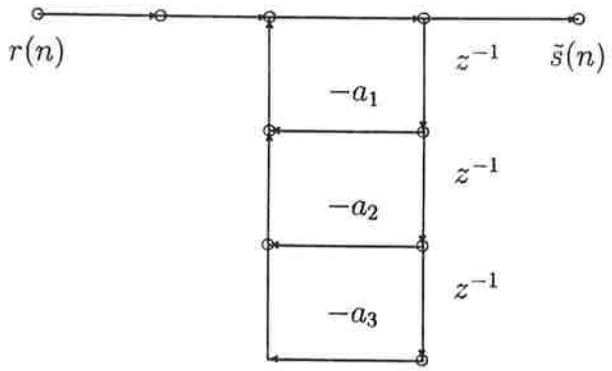

Figura 7.3: Filtro de síntese na forma direta I.

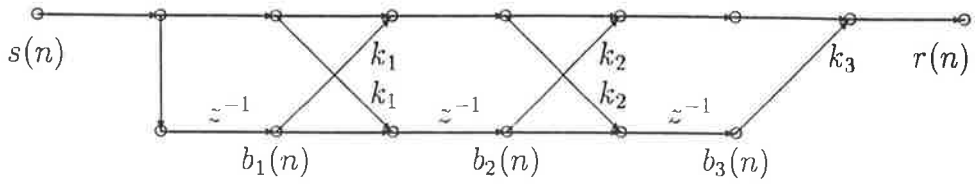

Figura 7.4: Filtro de análise em treliça.

$n+1$. Assim, as saídas dos filtros de análise e síntese são, respectivamente,

$$r(n+1) = s(n+1) + a_1' s(n) + a_2' s(n-1) + a_3' s(n-2), \quad (7.5)$$
$$\tilde{s}(n+1) = r(n+1) - a_1' \tilde{s}(n) - a_2' \tilde{s}(n-1) - a_3' \tilde{s}(n-2). \quad (7.6)$$

Assim, os filtros de análise e de síntese, mais do que terem funções de transferência inversas, têm que implementar fluxos de sinal invertidos.

As regras gerais para efetuar tal inversão encontram-se em [48] para grafos de fluxo de sinal em geral. Para aplicação a filtros digitais, aplica-se o seguinte procedimento simplificado [51]:

i) O sentido de fluxo do sinal é invertido no caminho direto de ligação entre a entrada e a saída (isto é, os nós de entrada e de saída são invertidos).

ii) Os ganhos em todos os ramos do caminho direto são substituídos pelos seus inversos.

iii) Os ganhos dos ramos situados fora do caminho direto que entram nos nós somadores localizados no caminho direto devem sofrer uma troca de sinal algébrico.

Estas regras estão diretamente associadas com o procedimento algébrico seguido para a forma direta I, podendo-se verificar que a aplicacão das regras *i)* e *ii)* à Fig. 7.1 produz a Fig. 7.2, que resulta na Fig. 7.3 através da regra *iii)*. Como aplicação não trivial do processo de inversão, toma-se em seguida a estrutura em treliça.

Na implementação em treliça do filtro de análise (Fig. 7.4), o sinal residual de saída é dado em função do sinal de voz de entrada por

$$r(n) = s(n) + k_1 b_1(n) + k_2 b_2(n) + k_3 b_3(n), \quad (7.7)$$

onde os estados do filtro de análise são calculados por

$$b_1(n+1) = s(n) \quad (7.8)$$
$$b_2(n+1) = k_1 s(n) + b_1(n) \quad (7.9)$$
$$b_3(n+1) = k_2 k_1 b_1(n) + k_2 s(n) + b_2(n). \quad (7.10)$$

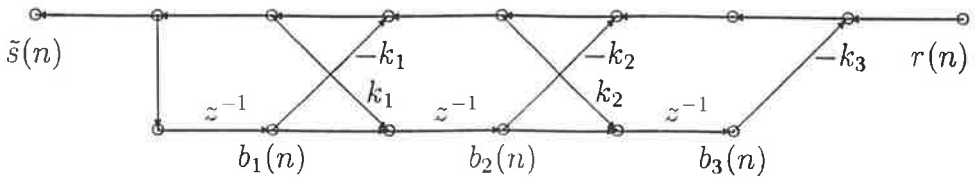

Figura 7.5: Filtro de síntese em treliça obtido pela inversão do fluxo de sinal do filtro de análise.

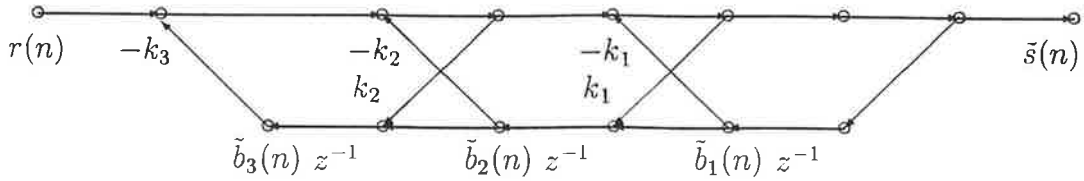

Figura 7.6: Filtro de síntese em treliça.

Tomando-se como referência o papel dual representado pela entrada do filtro de análise e pela saída do filtro de síntese (Fig. 7.6), calcula-se esta última a partir de

$$\tilde{s}(n) = r(n) - k_1\tilde{b}_1(n) - k_2\tilde{b}_2(n) - k_3\tilde{b}_3(n), \tag{7.11}$$

onde os estados do filtro de síntese também têm que ser reconstruções perfeitas dos estados correspondentes do filtro de análise, sendo calculados por

$$\tilde{b}_1(n+1) = \tilde{s}(n) \tag{7.12}$$
$$\tilde{b}_2(n+1) = k_1\tilde{s}(n) + \tilde{b}_1(n) \tag{7.13}$$
$$\tilde{b}_3(n+1) = k_2k_1\tilde{b}_1(n) + k_2\tilde{s}(n) + \tilde{b}_2(n). \tag{7.14}$$

Por outro lado, seguindo as regras de inversão de fluxo de sinal acima, passa-se, pelas regras *i)* e *ii)*, da Fig. 7.4 à Fig. 7.5 , que se transforma através da regra *iii)* na Fig. 7.6. Finalmente, a Figura 7.6 pode ser descrita pelas Equações (7.11) a (7.14).

7.3 Modelos para o sinal de excitação

Na seção anterior abordou-se a relação entre os filtros de análise e de síntese de um sinal de voz acoplados pelo sinal de excitação ideal, que é o sinal residual da predição. Porém, os codificadores de voz não podem representar em toda sua precisão o sinal residual, recorrendo a métodos de quantização que podem envolver um modelo para os sinais de excitação usados em seu lugar.

Assumindo o sinal excitação como uma seqüência de amostras irrestritas e independentes, a sua quantização vetorial acarreta ruído de quantização.

Entretanto, ao ver o processo de quantização da excitação do ponto de vista de um modelo ou molde em que se encaixa a forma de onda normalizada do sinal com o ganho não quantizado, o ruído de "quantização" é um ruído induzido pelo modelo. Pensando na codificação CELP, o índice do dicionário de códigos é um parâmetro essencialmente representável por um número inteiro enquanto o ganho é um número real.

Estendendo-se este ponto de vista operacional do codificador, pode-se decompor o erro de reconstrução (ou ruído de quantização em sentido generalizado) em três componentes

i) Ruído induzido pelo modelo;

ii) Ruído de processamento do algoritmo;

iii) Ruído de quantização dos parâmetros.

Quanto ao item *ii)*, a especificação geral mais importante é o tipo de aritmética a ser usada. Na implementação final do codificador, esta decisão vai influenciar a escolha do processador digital de sinais (DSP) [44].

7.4 Estruturas para o filtro de síntese

Consideram-se, em seguida, aspectos estruturais envolvidos na implementação do filtro de síntese nos codificadores CELP (Capítulo 3). Algumas conseqüências da representação dos sinais de excitação nos codificadores, relacionadas na seção 7.3, serão apreciadas nos resultados dos testes sobre implementações de codificadores CELP. Porém, de início, o tratamento dos fluxos dos sinais nos filtros segue o estilo da seção 7.2.

Nos codificadores CELP, no início de cada busca de excitação, decompõe-se linearmente o filtro de síntese ponderado $H(z)$ em dois subsistemas (Fig. 7.7):

- o filtro de síntese ponderado $H_C(z)$ carregado com os estados recebidos do sub-bloco anterior, que é estimulado pela excitação nula e
- o filtro de síntese ponderado $H_D(z)$, que sempre parte da condição nula e é excitado por cada um dos vetores-códigos.

Ao fim de cada busca por todos os dicionários de códigos do codificador, quando pode ser montado o sinal composto de excitação, este sinal composto é empregado para excitar o filtro de síntese ponderado $H(z)$, que parte do estado final atingido no sub-bloco anterior.

Assim, os processos de filtragem $W(z)$, $H_C(z)$ são os únicos relevantes para o tratamento de transitórios. Por conseguinte, a estrutura de $H_D(z)$ é irrelevante, como também pode ser compreendido pela sua possibilidade de representação pela matriz de resposta impulsiva (seção 3.6).

Como $H_C(z)$ é um componente de $H(z)$, ambos são implementados na mesma estrutura. Coerentemente, o filtro de ponderação $W(z)$ também é implementado na mesma estrutura.

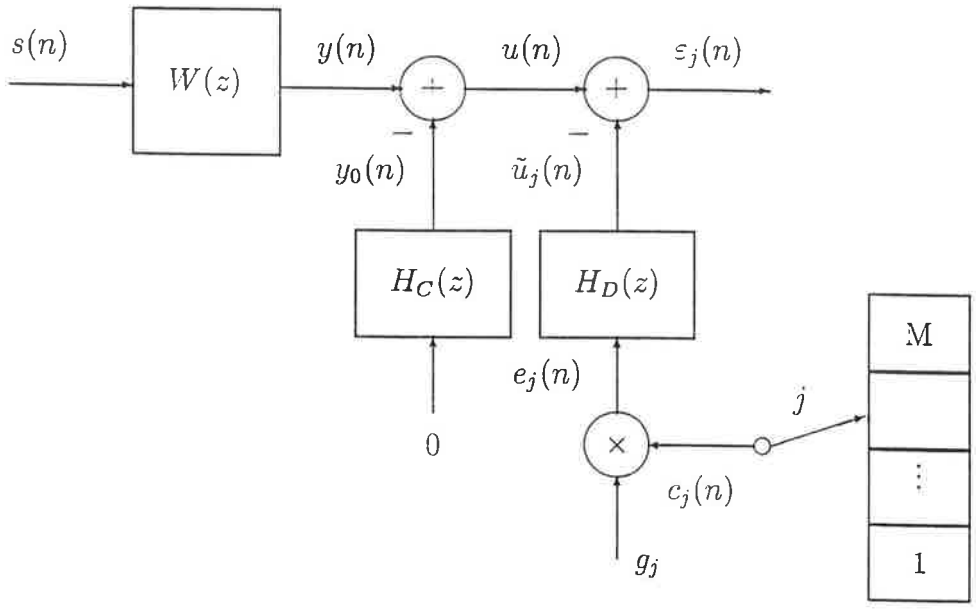

Figura 7.7: Modelo de busca da excitação para um codificador CELP com dicionário fixo único.

Foi testado o comportamento de mais duas estruturas, uma direta e uma em treliça, além da forma direta I e da treliça padrão com dois multiplicadores. Essas duas estruturas adicionais são a forma direta II transposta e a treliça normalizada [55] e [56]. A forma direta II transposta do filtro de síntese de ordem $p = 3$ é mostrada na Fig. 7.8.

Na implementação em treliça com dois multiplicadores do filtro de ponderação $W(z)$, tem que ser usada a topologia da Fig. 7.9 porque sua função de transferência apresenta polinômios de grau p tanto no numerador quanto no denominador.

A treliça normalizada [47] foi testada porque as estruturas ortogonais variantes no tempo são consideradas estáveis contanto que suas seções invariantes no tempo sejam estáveis em primeiro lugar [18]. A topologia da treliça normalizada é composta de seções ortonormais e está mostrada para ordem $p = 3$ na Fig. 7.10, onde

$$\begin{aligned} s_i &= k_i \\ &= \text{sen}\theta_i \\ c_i &= \sqrt{1 - k_i^2} \\ &= \cos\theta_i \quad i = 1, 2, \ldots, p, \end{aligned} \tag{7.15}$$

e k_i, $i = 1, 2, \ldots, p$ são os coeficientes de reflexão da estrutura em treliça da Fig. 7.9.

Implementaram-se quatro codificadores VSELP com filtros variantes no tempo de ordem $p = 10$ nas estruturas direta I, direta II transposta, treliça com dois multiplicadores e treliça normalizada. Cada um desses codificadores foi programa em linguagem C e no ambiente MATLAB em microcomputador PC. Cada um dos

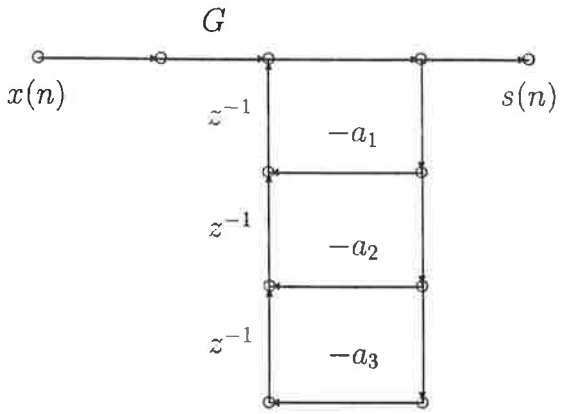

Figura 7.8: Filtro de síntese na forma direta II transposta.

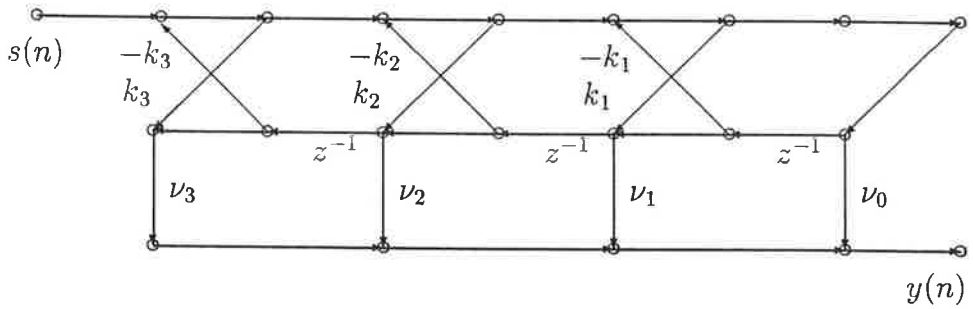

Figura 7.9: Filtro de ponderação em treliça com dois multiplicadores.

quatro pares de implementações sofreu um processo iterativo de correções de código e testes de desempenho até a obtenção da coincidência numérica.

Numa etapa final, o código fonte em linguagem C foi compilado para o processador digital de sinais (DSP) TMS320C30, tendo sua correção funcional testada.

Usaram-se dois conjuntos de sinais de voz para os testes: o conjunto 3m3f e o conjunto de 13 sinais da base de dados TIMIT (seção 5.2).

Os desempenhos dos pares de implementações no PC e no DSP em linguagem C para cada uma das quatro estruturas e para os dois conjuntos de sinais de teste encontram-se na Tabela 7.1.

Conclui-se desses resultados que as formas diretas apresentam desempenho melhor do que as estruturas em treliça. Chen [12] já tinha relatado resultados preliminares nesta direção.

7.5 Transitórios em codificadores CELP

Num codificador CELP há dois processos de decodificação, sendo que um ocorre no próprio decodificador e outro tem lugar no codificador propriamente dito. Este último é devido ao procedimento de análise-mediante-síntese utilizado durante a

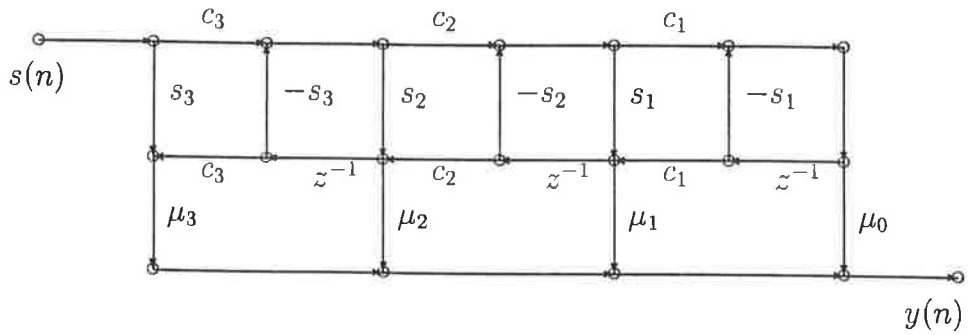

Figura 7.10: Filtro de ponderação em treliça normalizada.

Tabela 7.1: Comparação de qualidade entre 4 codificadores com 4 estruturas de filtros para 2 implementações e 2 conjuntos de sinais.

Estrutura (IIR, ordem 10)	SNRSEG (dB) PC	SNRSEG (dB) DSP	Sinal de voz
Forma direta I	10,86	10,82	3m3f
Forma direta II transposta	10,69	10,68	3m3f
Treliça com 2 multiplicadores	10,52	10,53	3m3f
Treliça normalizada	10,32	10,25	3m3f
Forma direta I	10,65	10,61	TIMIT
Forma direta II transposta	10,56	10,37	TIMIT
Treliça com 2 multiplicadores	10,46	10,37	TIMIT
Treliça normalizada	10,17	10,04	TIMIT

busca da excitação nos dicionários de códigos.

No sinal de voz reconstruído pelo decodificador, podem-se identificar componentes transitórias. Neste contexto, existe um método proposto em [72] para eliminação de transitórios considerando uma única transição dos parâmetros do filtro.

Na seção 7.6 estende-se esse método para aplicação a uma seqüência de alterações de valores dos parâmetros nas interfaces entre sub-blocos. Ademais, através da análise das relações encontradas para o método estendido, emerge uma implementação que usa um filtro FIR transversal, com resposta impulsiva suficientemente longa, como filtro de síntese.

Alternativamente, os transitórios podem ser analisados num hipotético processo de análise residente no codificador, que produz o sinal residual a partir do sinal de voz original. Uma análise formal deste processo [40] conclui que o filtro inverso implementado como um filtro FIR transversal é a única estrutura, dentre as realizáveis formas diretas e em treliça, que não gera componente transitória de origem paramétrica. Isto ocorre porque sua resposta à excitação nula depende apenas dos parâmetros do filtro no sub-bloco atual e do sinal de voz passado enquanto outras estruturas dependem adicionalmente dos parâmetros passados do filtro (seção 7.7).

Portanto, na abordagem a partir do processo de análise do sinal de voz, conclui-se pela superioridade de desempenho do filtro de síntese implementado como IIR na forma direta I.

Chega-se, assim, a uma contradição quanto à melhor estrutura para a implementação do filtro de síntese. Tenta-se resolver esse conflito com testes de codificadores CELP selecionados, aos quais são aplicadas ambas as classes de técnicas de supressão de transitórios.

Na seqüência, aborda-se inicialmente a supressão de transitórios do ponto de vista do decodificador, onde é detalhado o método proposto. Prossegue-se com a consideração dos transitórios no contexto do processo de análise do sinal de voz, que subsidia a interpretação dos testes apresentados logo depois.

A análise dos resultados desses testes de transitórios em estruturas de filtragem (seção 7.9) levanta a questão do treinamento dos dicionários fixos, onde se aplica o processo deduzido na seção 4.9. Ademais, apresentam-se os resultados de treinamentos de dicionários fixos com estrutura VSELP sobre um conjunto de sinais selecionado da base de dados TIMIT (seção 5.2).

Finalmente, aborda-se a questão da precisão usada na implementação do codificador em decorrência do uso de aritmética de ponto fixo e da variação da taxa de bits transmitidos. Nesse contexto, usando o codificador CELP da recomendação G.723.1, apresentam-se os resultados dos testes sobre uma extensa base de dados, que inclui também sinais corrompidos com ruído de canais telefônicos.

7.6 Supressão de transitórios no decodificador

Considere-se o filtro de síntese descrito no espaço de estados mediante as matrizes $\{A_m, B_m, C_m, D_m\}$, assumidas constantes durante o sub-bloco m

$$\begin{aligned} x(n+1) &= A_m x(n) + B_m e(n) \\ \tilde{s}(n) &= C_m x(n) + D_m e(n), \end{aligned} \tag{7.16}$$

onde $e(n)$ é o sinal de excitação, $x(n)$ é o vetor de estados e $\tilde{s}(n)$ é o sinal de voz reconstruído.

Generalizando os resultados obtidos em [72] para uma seqüência de comutações dos parâmetros dos filtros nos instantes de passagem entre sub-blocos consecutivos $1, 2, \ldots, m$, e iniciando o processo de decodificação a partir do estado nulo, obtém-se a seguinte seqüência de vetores de estado após a passagem do sub-bloco m para o sub-bloco $m+1$:

$$\begin{aligned} x(n) = A_{m+1}^{n-mL} \sum_{j=0}^{mL-1} \Big[& A_m^{mL-1-j} B_m - \\ & A_{m+1}^{mL-1-j} B_{m+1} \Big] e(j) + \sum_{j=0}^{n-1} A_{m+1}^{n-1-j} B_{m+1} e(j), \end{aligned} \tag{7.17}$$

onde L é o comprimento do sub-bloco. O transitório nas variáveis de estado na passagem do sub-bloco m para o sub-bloco $m+1$ é dado pela somatória das parcelas indicadas entre colchetes na Eq.(7.17).

O método proposto em [72] procura eliminar os transitórios entre duas configurações dos parâmetros do filtro pela aplicação do incremento

$$\Delta x(mL) = -\sum_{j=(m-1)L}^{mL-1} \left[A_m^{mL-1-j} B_m - A_{m+1}^{mL-1-j} B_{m+1}\right] e(j) \tag{7.18}$$

ao vetor de estados no instante de transição $n = mL$. Ou seja, em palavras, nesta interpretação consideram-se transitórias aquelas parcelas das variáveis de estado que excedem os valores que seriam atingidos pelo filtro atual quando excitado por toda a seqüência passada de excitações. Conseqüentemente, a eliminação dessas parcelas transitórias permite que o estado inicial do sub-bloco atual seja justamente aquele que seria atingido se o filtro atual tivesse estado presente desde o início da reconstrução do sinal, conforme explicitado logo em seguida.

Eliminando-se totalmente o transitório, a evolução do vetor de estados no sub-bloco $m+1$ torna-se

$$x(n) = \sum_{j=0}^{n-1} A_{m+1}^{n-1-j} B_{m+1} e(j). \tag{7.19}$$

A supressão do transitório nos estados aconteceria naturalmente se a estrutura do filtro fosse tal que

$$A_m^{mL-1-j} B_m = A_{m+1}^{mL-1-j} B_{m+1} \tag{7.20}$$

para $j = 0, 1, \ldots, mL-1$. Isto implica na dupla igualdade

$$A_m = A_{m+1} \text{ e } B_m = B_{m+1}.$$

Neste ponto passa-se à extensão do método original. Ele é atingido com a aplicação da Eq.(7.17) às consecutivas alterações de valor dos parâmetros. Conclui-se assim que o transitório nas variáveis de estado é nulo em qualquer caso se, e somente se,

$$A_k \equiv A_{k+1} \text{ e } B_k \equiv B_{k+1}, k = 1, 2, \ldots, m,$$

independentemente do valor dos coeficientes do filtro. Estas condições podem ser satisfeitas por um filtro FIR transversal desde que se desprezem as amostras mais antigas do sinal de excitação.

De fato, as variáveis de estado do filtro FIR transversal são amostras passadas do próprio sinal de excitação, que não é influenciado de forma alguma pela mais recente variação dos parâmetros.

Conseqüentemente, para simplificar as referências ao método estendido, passa-se a usar a denominação de filtro XFIR, que guarda de certa forma a semelhança e a distinção mantida com o filtro FIR propriamente dito.

Assim, a nova implementação proposta neste trabalho requer simplesmente a aplicação de um filtro FIR para a eliminação dos transitórios. A resposta impulsiva deste filtro FIR reproduz a princípio a do filtro XFIR e depois torna-se uma versão truncada quando esta cresce além do tamanho projetado para o filtro FIR. Deve-se apenas definir a ordem deste filtro. Para tanto, são fornecidos subsídios na seção 7.8.

A principal vantagem desta abordagem reside na memória de estados finita do filtro FIR ao passo que o método estendido requer memória crescente para armazenar o sinal de excitação passado na sua integridade.

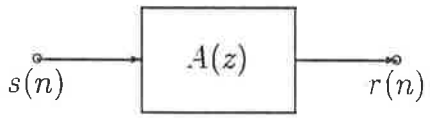

Figura 7.11: Filtragem inversa do sinal de voz.

7.7 Efeitos variantes no tempo no analisador

O processo de análise de um codificador CELP envolve uma estimação dos parâmetros de predição linear, sendo seguido pelos processos de busca da excitação. Os coeficientes de predição linear a_k determinam o filtro inverso

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \tag{7.21}$$

onde $p = 10$ é a ordem da predição linear, e o filtro de síntese ponderado é dado por

$$H(z) = \frac{1}{A(z/\gamma)}, \tag{7.22}$$

onde $0 < \gamma < 1$ é o coeficiente de expansão de faixa.

Na figura 7.7, usou-se a notação $H_C(z)$ para designar o filtro $H(z)$ carregado, ou seja, com estado inicial não necessariamente nulo e designou-se por $H_D(z)$ o filtro $H(z)$ com estado inicial nulo ou descarregado.

A busca da excitação tem como referência um vetor-alvo $u(n)$, derivado do sinal de voz $s(n)$, sem considerar qualquer sinal de excitação como referência no processo (Fig. 7.7).

Entretanto, abstraindo-se apenas a dinâmica de variação dos parâmetros do filtro de síntese, pode-se derivar um sinal de excitação de referência. Esta referência é o sinal residual $r(n)$, que se obtém pela passagem do sinal de voz $s(n)$ através do filtro inverso do filtro de síntese (Fig. 7.11). Observa-se que este processo de análise foi abordado seguido pelo processo de síntese na seção 7.2, onde o filtro $A(z)$ e seu inverso foram submetidos a variações paramétricas e estruturais.

Tome-se a estrutura FIR transversal do filtro inverso como base de comparação e seja sua representação no espaço de estados dada por

$$\begin{aligned} x(n+1) &= A_m x(n) + B_m s(n) \\ r(n) &= C_m x(n) + D_m s(n). \end{aligned} \tag{7.23}$$

Considere-se agora uma implementação alternativa do filtro de análise como, por exemplo, a estrutura em treliça

$$\begin{aligned} \overline{x}(n+1) &= \overline{A}_m \overline{x}(n) + \overline{B}_m s(n) \\ r(n) &= \overline{C}_m \overline{x}(n) + \overline{D}_m s(n). \end{aligned} \tag{7.24}$$

Seja T_m a transformação similar do estado da primeira realização no estado da segunda, isto é,

$$\overline{x}(n) = T_m x(n). \tag{7.25}$$

A segunda representação de estados pode ser ordenadamente expressa em função da primeira representação e da matriz de transformação similar como

$$\{\overline{A}_m, \overline{B}_m, \overline{C}_m, \overline{D}_m\} = \{T_m A_m T_m^{-1}, T_m B_m, C_m T_m^{-1}, D_m\}. \tag{7.26}$$

Iterando a equação (7.24) de $n = -p$ a $n = 0$, onde p é a ordem do filtro, obtém-se

$$x(0) = A_{m-1}^p x(-p) + \sum_{j=1}^{p} A_{m-1}^{j-1} B_{m-1} s(-j). \tag{7.27}$$

Introduzindo-se a matriz de controlabilidade $\mathcal{C}_m$

$$\mathcal{C}_m = \left[\begin{array}{cccc} B_m & A_m B_m & A_m^2 B_m & \cdots & A_m^{p-1} B_m \end{array} \right] \tag{7.28}$$

em (7.27) e notando-se que, para o filtro FIR transversal de ordem p, vale $A_m^p = 0$, o estado inicial do bloco atual de análise (estado final do bloco anterior) relaciona-se com os valores anteriores do sinal através de [40]

$$x(0) = \mathcal{C}_{m-1} s_{m-1}, \tag{7.29}$$

onde

$$s_{m-1} = \left[\begin{array}{cccc} s(-1) & s(-2) & \cdots & s(-p) \end{array} \right]^T.$$

Manipulando-se a definição (7.28) e as relações de transformação (7.26), resulta que a matriz de controlabilidade da segunda realização é dada em função daquela da primeira por

$$\overline{\mathcal{C}}_{m-1} = T_{m-1} \mathcal{C}_{m-1}. \tag{7.30}$$

No filtro FIR transversal, o estado inicial é o próprio vetor de entradas passadas. Isto pode ser visto nesta formulação pela construção da matriz de controlabilidade (Eq.(7.28)) do filtro FIR transversal, que resulta

$$\mathcal{C}_{m-1} = I_p.$$

Em conseqüência, a matriz de controlabilidade da segunda implementação (Eq. (7.30)) torna-se

$$\overline{\mathcal{C}}_{m-1} = T_{m-1}. \tag{7.31}$$

O transitório decorrente da variação dos coeficientes do filtro na passagem do sub-bloco $m-1$ para o sub-bloco m pode ser obtido fazendo-se $s_m = 0$. Resulta assim a componente transitória do sinal residual para o próprio sub-bloco de sinal de voz s_m. Essa componente transitória é dada por [40]

$$r_m = \mathcal{O}_m x(0), \tag{7.32}$$

onde $\mathcal{O}_m$ é a matriz de observabilidade do primeiro filtro, que se expressa por extenso como

$$\mathcal{O}_m = \left[\begin{array}{cccc} C_m^T & A_m^T C_m^T & A_m^{T\,2} C_m^T & \cdots & A_m^{T\,p-1} C_m^T \end{array} \right]^T . \tag{7.33}$$

Manipulando-se a definição (7.33) e as relações de transformação (7.26), resulta que a matriz de observabilidade da segunda realização é dada em função daquela da primeira por

$$\bar{\mathcal{O}}_m = \mathcal{O}_m T_m^{-1}. \tag{7.34}$$

Assim, o sinal residual no bloco atual decorrente apenas do estado inicial, para a realização em filtro FIR transversal, é

$$r_m = \mathcal{O}_m s_{m-1} \tag{7.35}$$

e, para a segunda realização, tem-se

$$\bar{r}_m = \mathcal{O}_m T_m^{-1} T_{m-1} s_{m-1}, \tag{7.36}$$

onde s_{m-1} é o sub-bloco anterior do sinal de voz.

Tendo-se em vista todas as formas diretas e em treliça realizáveis, tem-se que a estrutura FIR transversal é a única em que ambas as transformações similares são idênticas e independentes dos coeficientes do filtro e dadas por $T_{m-1} = T_m = I$. Assim, esta é a única dessas estruturas de filtro inverso que apresenta efeitos transitórios independentes dos valores anteriores dos coeficientes do filtro.

Conseqüentemente, em decorrência da inversão estrutural, a estrutura IIR na forma direta I é a única que apresenta efeitos transitórios independentes de configurações paramétricas anteriores sobre o filtro de síntese.

7.8 Testes de supressão de transitórios

Um codificador CELP com estrutura VSELP contendo dois dicionários de códigos fixos e um adaptativo foi usado na parte inicial dos testes. Ademais, os vetores-base dos dicionários fixos foram aqueles definidos em [21].

Aplicaram-se os procedimentos de supressão dos transitórios nas variáveis de estado ao filtro de ponderação perceptiva e ao filtro de síntese ponderado, ambos no codificador, e também ao filtro de síntese do decodificador.

O conjunto de sinais usado nesses testes iniciais consiste em 13 frases da base de dados TIMIT (seção 5.2).

Os testes com os filtros variantes no tempo implementados na forma direta I recorrente revelaram que a supressão de transitórios nas variáveis de estado não garante uma reconstrução mais precisa na saída (Tabela 7.2).

Adicionalmente, o filtro FIR transversal foi testado para verificar a obtenção de reconstrução equivalente ao filtro XFIR. Uma comparação de filtros FIRs de várias ordens com a referência XFIR revela que uma resposta impulsiva de um bloco de comprimento (160 amostras) já fornece uma SNRSEG muito próxima ao passo que uma duração de dois blocos satisfaz um critério mais conservador.

Tabela 7.2: Desempenho de reconstrução do sinal de codificadores VSELP com filtros de síntese IIR na forma direta I, FIR estendido e FIR transversal em função da ordem dos filtros para o subconjunto de sinais TIMIT.

Estrutura	Ordem do filtro	SNRSEG (dB)
IIR	10	10,2003
XFIR	∞	9,5900
FIR	40	8,9708
FIR	80	9,5639
FIR	160	9,5961
FIR	320	9,5901
FIR	480	9,5901

Os testes com estruturas IIR foram restritos à forma direta I devido a resultados anteriores que indicam sua superioridade em codificadores VSELP da norma IS-54 quando testados com as estruturas em treliça, treliça normalizada e com a forma direta II transposta (seção 7.4).

7.9 Treinamento do quantizador da excitação

Os testes de reconstrução de sinais na seção 7.8 usaram os vetores-base definidos na norma IS-54 para os dicionários fixos do codificador VSELP [21], que foram treinados sobre uma base de dados não disponível com distribuição inicial gaussiana das amostras dos vetores-base [25].

Como o filtro de síntese do codificador VSELP padrão é implementado como IIR na forma direta I, existe a possibilidade de que um vício estatístico em favor desta estrutura tenha se acumulado durante o processo de treinamento dos dicionários fixos.

Conseqüentemente, treinaram-se os vetores-base com o subconjunto de sinais descrito na seção 7.8, sendo que os vetores-base foram inicialmente escolhidos de duas distribuições, que são a gaussiana e a laplaciana de médias nulas e de variâncias unitárias (seção 2.2).

O treinamento com distribuição inicial gaussiana foi realizado num computador PC Pentium com Windows NT e o de distribuição laplaciana foi executado numa estação de trabalho SPARC 10, necessitando por volta de 2 dias e 15 horas e 2 dias e 13 horas de tempo ininterrupto, respectivamente, por iteração.

Treinaram-se os dicionários fixos de acordo com o algoritmo apresentado na seção 4.9. Os resultados desses testes estão nas Tabelas 7.3 e 7.4 e revelam que [57]:

- o filtro IIR na forma direta I é superior ao FIR transversal;
- os vetores-base distribuídos laplacianamente no início conduzem a melhores dicionários do que os distribuídos gaussianamente no início.

Tabela 7.3: Resultados em SNRSEG dos treinamentos dos dicionários fixos do VSELP com a partição de teste da base de dados TIMIT para 8 iterações.

Estrutura do filtro	Distribuição inicial	Ordem do filtro	SNRSEG (dB)	
			inicial	final
FIR	Gaussiana	160	7,2547	8,0056
FIR	Laplaciana	160	7,3249	8,0078
FIR	Gaussiana	40	6,5379	7,1242

Tabela 7.4: Distorção resultante em PSQM após os treinamentos dos dicionários fixos do VSELP com a partição de teste da base de dados TIMIT para 8 iterações.

Estrutura do filtro	Distribuição inicial	Ordem do filtro	PSQM	
			inicial	final
FIR	Gaussiana	160	2,4270	2,2928
FIR	Laplaciana	160	2,3416	2,2940
FIR	Gaussiana	40	2,6672	2,5511

Os incrementos de desempenho em SNRSEG obtidos ao longo dos treinamentos encontram-se nas Figuras 7.12, 7.14 e 7.16 e os incrementos de distorção em PSQM estão nas Figuras 7.13, 7.15, e 7.17 para as seguintes condições, respectivamente:

- distribuição inicial gaussiana e filtro FIR160;
- distribuição inicial laplaciana e filtro FIR160;
- distribuição inicial gaussiana e filtro FIR40.

Comparando-se os incrementos de desempenho com aqueles registrados na literatura (seção 4.9), nota-se que os treinamentos efetuados foram eficazes, pois

FIR160(Gauss): $\mathbf{\Delta SNRSEG_{treino}} = 0{,}7509$ dB

FIR160(Gauss): $\mathbf{\Delta PSQM_{treino}} = -0{,}1342$

FIR160(Laplace): $\mathbf{\Delta SNRSEG_{treino}} = 0{,}6829$ dB

FIR160(Laplace): $\mathbf{\Delta PSQM_{treino}} = -0{,}0476$

FIR40(Gauss): $\mathbf{\Delta SNRSEG_{treino}} = 0{,}5863$ dB

FIR40(Gauss): $\mathbf{\Delta PSQM_{treino}} = -0{,}1161$.

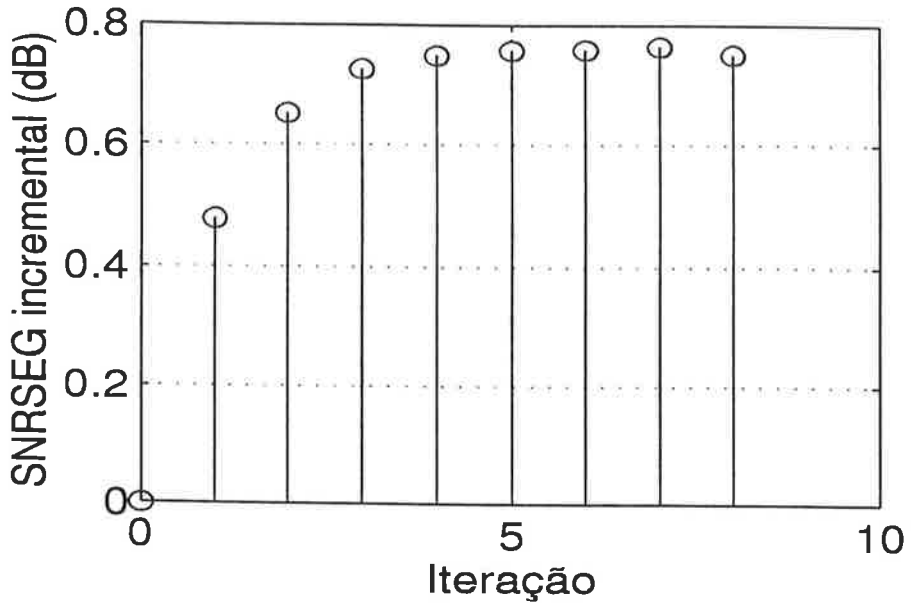

Figura 7.12: Desempenho incremental para filtro FIR160 ao longo do treinamento dos vetores-base do codificador VSELP IS-54 com filtro FIR em 8 iterações sobre a partição de teste do TIMIT a partir de população inicial distribuída gaussianamente.

Tabela 7.5: Desempenho do ACELP sobre a partição de teste da base de dados TIMIT para filtros IIR direta I e FIR com comprimentos 60 e 240.

Medida	IIR-I	FIR-I 240	FIR-I 60
SNRSEG (dB)	9,4371	9,1472	9,1728
PSQM	2,0848	2,1619	2,1575

7.10 Influência do comprimento da resposta impulsiva

O comprimento da resposta impulsiva nas implementações FIR transversais do filtro de síntese pode ser consideravelmente reduzido sem afetar o desempenho.

Os resultados com o codificador ACELP da recomendação G.723.1 (Tabela 7.5) revelam uma equivalência de desempenho para comprimentos da resposta impulsiva truncados na duração do bloco ou do sub-bloco, inclusive com uma ligeira superioridade deste último caso.

Entretanto, deve-se registrar que estes resultados são dependentes da estrutura dos dicionários de códigos fixos, pois com o codificador VSELP verifica-se uma queda considerável de desempenho quando a resposta é truncada no comprimento do sub-bloco como se depreende dos dados constantes da seção 7.9.

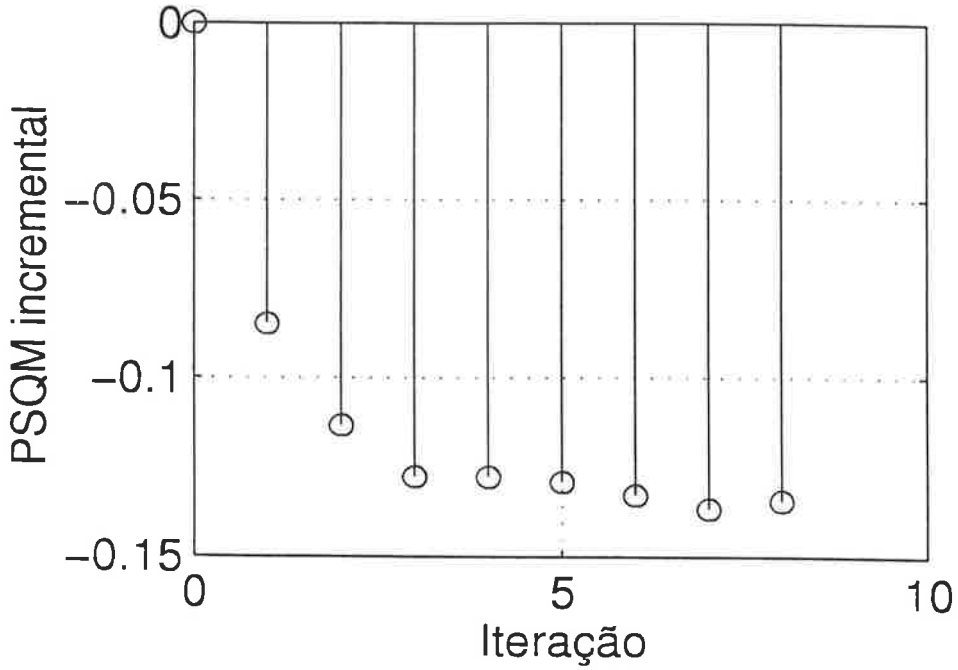

Figura 7.13: Distorção incremental para filtro FIR160 ao longo do treinamento dos vetores-base do codificador VSELP IS-54 com filtro FIR em 8 iterações sobre a partição de teste do TIMIT a partir de população inicial distribuída gaussianamente.

7.11 Influência de ruídos na linha telefônica

Adaptou-se uma implementação do codificador MP-MLQ/ACELP definido pela norma G.723.1 [32] para poder assumir estruturas alternativas IIR direta I e FIR transversal no filtro de síntese.

Esta configuração permite testar a influência da precisão numérica bem como de ruídos presentes na ligação telefônica sobre o trabalho de reconstrução do sinal. Ademais, este codificador permite o teste de duas taxas de transmissão, que são 5,3 e 6,3 kbit/s.

Selecionaram-se os sinais limpos da base de dados TIMIT [23] e as suas correspondentes versões corrompidas com ruído de linha telefônica foram tomadas da base de dados NTIMIT [36].

Estes testes foram realizados com o conjunto completo de sinais das partições de teste das bases de dados TIMIT e NTIMIT (seção 5.2).

A medida de distorção objetiva PSQM, descrita na seção 5.4, (Tabela 7.7), que incorpora características perceptivas do aparelho auditivo humano, foi usada em conjunto com a medida de qualidade de reconstrução SNRSEG (Tabela 7.6).

Os resultados dos testes indicam que os efeitos transitórios comparativos entre estruturas do filtro de síntese numa dada taxa são mais significativos para os sinais limpos que para os sinais ruidosos.

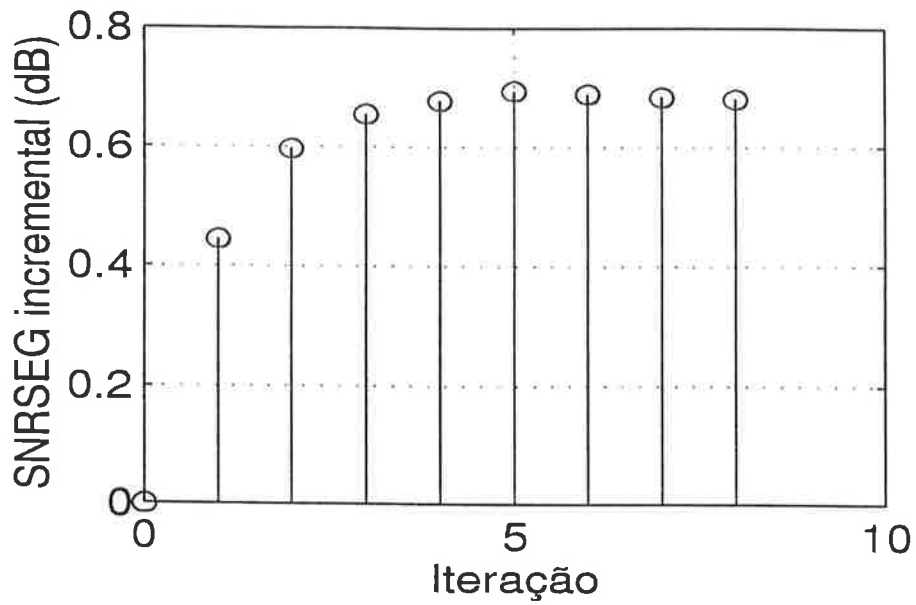

Figura 7.14: Desempenho incremental para filtro FIR160 ao longo do treinamento dos vetores-base do codificador VSELP IS-54 com filtro FIR em 8 iterações sobre a partição de teste do TIMIT a partir de população inicial distribuída laplacianamente.

7.12 Considerações finais

Em termos absolutos, a adoção de um modelo para o sinal de excitação, independente dos detalhes específicos do quantizador usado, revelou-se determinante para o desempenho ocasionado pelas diversas estruturas de filtros de síntese.

Entretanto, em termos relativos, vantagens comparativas de implementação podem compensar as perdas inerentes, que, dependendo da estrutura do dicionário de inovações, podem ser consideravelmente reduzidas. Um caso de perdas reduzidas de desempenho ocorreu com o ACELP com resposta impulsiva truncada (seção 7.10).

Tabela 7.6: Desempenho do codificador G.723.1 para duas estruturas de filtro de síntese medida em SNRSEG (dB) sobre toda a extensão da partição de teste da base de dados.

		SNRSEG (dB)	
Base de dados	Taxa (kbit/s)	IIR-I	FIR-I (ordem 240)
TIMIT	5,3	9,4216	9,1297
TIMIT	6,3	10,7622	10,4157
NTIMIT	5,3	7,1988	7,0793
NTIMIT	6,3	8,4504	8,3212

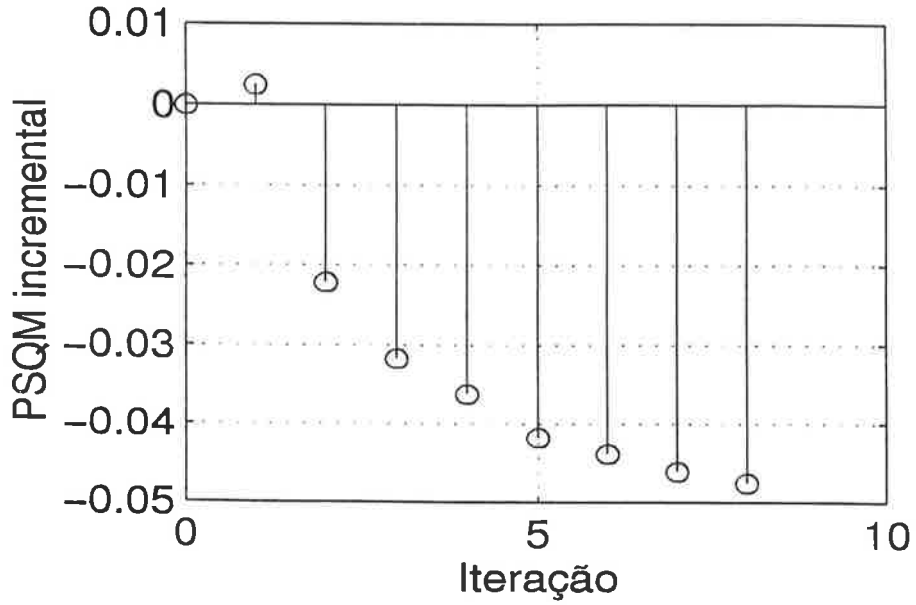

Figura 7.15: Distorção incremental para filtro FIR160 ao longo do treinamento dos vetores-base do codificador VSELP IS-54 com filtro FIR em 8 iterações sobre a partição de teste do TIMIT a partir de população inicial distribuída laplacianamente.

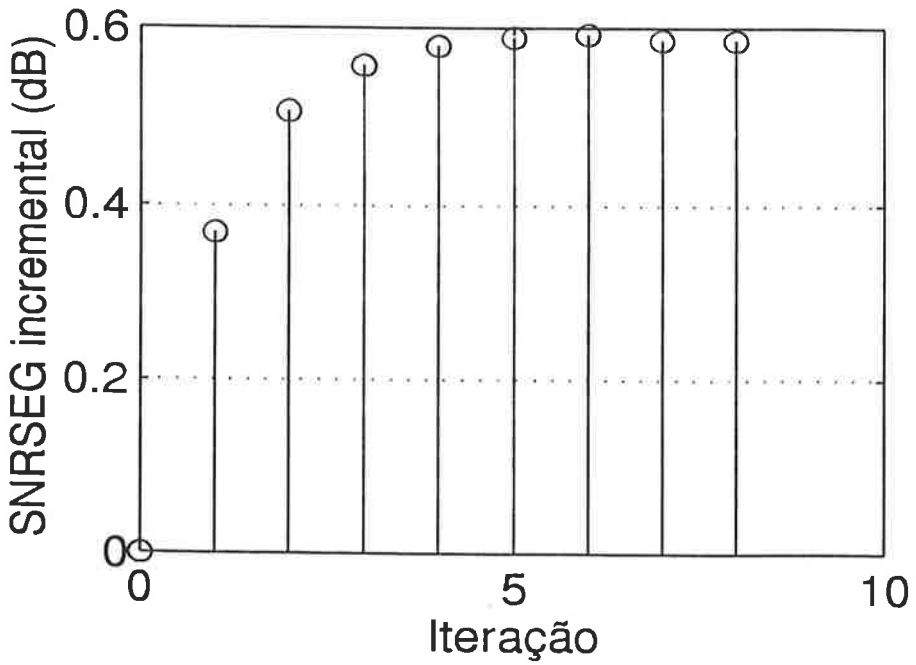

Figura 7.16: Desempenho incremental para filtro FIR40 ao longo do treinamento dos vetores-base do codificador VSELP IS-54 com filtro FIR em 8 iterações sobre a partição de teste do TIMIT a partir de população inicial distribuída gaussianamente.

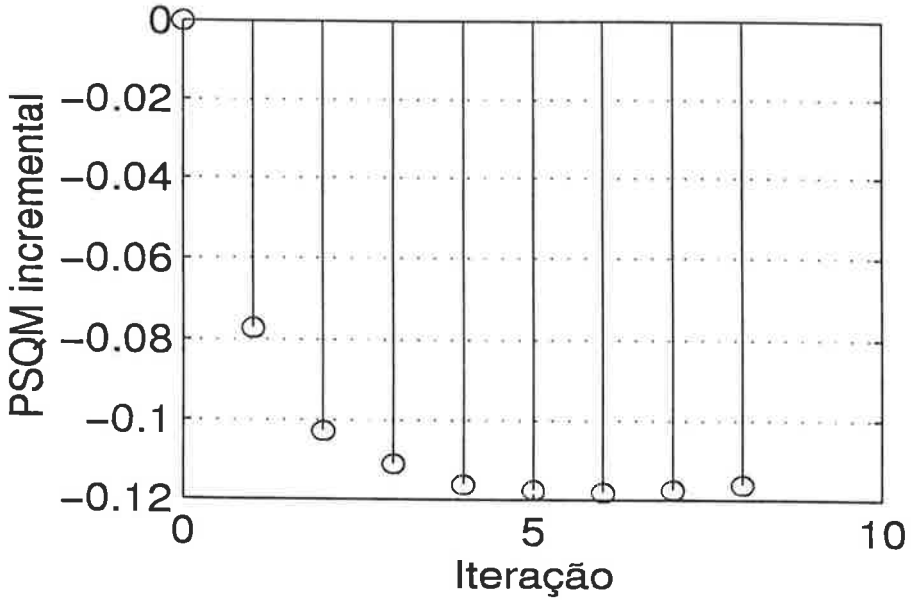

Figura 7.17: Distorção incremental para filtro FIR40 ao longo do treinamento dos vetores-base do codificador VSELP IS-54 com filtro FIR em 8 iterações sobre a partição de teste do TIMIT a partir de população inicial distribuída gaussianamente.

Tabela 7.7: Distorção introduzida pelo codificador G.723.1 para duas estruturas de filtro de síntese medida em PSQM sobre toda a extensão da partição de teste da base de dados.

		PSQM	
Base de dados	Taxa (kbit/s)	IIR-I	FIR-I (ordem 240)
TIMIT	5,3	2,2800	2,3821
TIMIT	6,3	2,0292	2,1851
NTIMIT	5,3	2,4213	2,4876
NTIMIT	6,3	2,0727	2,1409

Capítulo 8

CONCLUSÃO

Apresentaram-se os codificadores preditivos usados em comunicações com ênfase nos codificadores CELP. Para esta classe de codificadores, descreveram-se de forma abrangente as características comuns dos dicionários adaptativos e dos fixos estruturados.

Na busca conjunta AMPE-JPAS proposta no Capítulo 6, o número de elementos pesquisados corresponde a 4% dos elementos pesquisados na busca exaustiva de posições e menos de 40% do número médio de elementos pesquisados na busca focalizada.

O número de coeficientes de autocorrelação calculados pelo AMPE-JPAS é 36% do número calculado pela busca focalizada. Exemplificando, em testes com sinais de voz, o número de vetores-código pesquisados por busca pelo AMPE-JPAS é próximo de 40% daquele pesquisado pela busca focalizada e apenas 3,7% do número testado pela busca exaustiva de posições. Em relação ao desempenho, os resultados sinalizam uma equivalência entre os dois métodos de busca pois, embora a busca conjunta incorra numa perda relativa de relação sinal-ruído segmentada de 0.2 dB para alguns sinais, a melhora obtida sobre 1680 sinais da base de dados TIMIT acarreta uma redução de 0,14 pontos na distorção PSQM média.

Como continuação deste trabalho de pesquisa de buscas eficientes de inovações, ainda no contexto dos codificadores ACELP, é interessante confrontar os três métodos de busca analisados com as buscas eficientes do codificador melhorado para a taxa completa da telefonia celular digital TDMA, norma IS-641, e o codificador para transmissão multimídia simultânea de voz e de dados (DSVD) da recomendação G.729A da ITU-T (seção 2.8). Alerta-se nesta questão que estes codificadores ACELP operam sobre uma grade de posições diferente, sendo necessária sua substituição pelas grades do G.723.1 antes dessa comparação. De outra forma, devem-se inserir essas técnicas de busca no contexto do codificador G.723.1 para possibilitar comparações diretas.

Como base de comparação preliminar, podem ser considerados os dados publicados em [30] sobre o codificador IS-641, relatando que o método de busca de inovações usado percorre apenas 9% do espaço admíssivel de excitações enquanto, em relação ao codificador G.729A, a busca é restrita a 320 inovações, representando 3,9% dos vetores-código [61]. Este último codificador tornou-se mais eficiente ao custo de uma

queda de aproximadamente 0,2 dB da relação sinal-ruído.

Além das buscas eficientes de inovações para o filtro de síntese, abordaram-se os transitórios causados pela variância no tempo deste filtro bem com seu tratamento através da seleção da sua estrutura e das características populacionais dos dicionários fixos.

No Capítulo 7, em várias situações de teste, confirma-se na média a previsão do estudo de transitórios no analisador: o filtro de síntese na forma direta I com simples transferência dos estados entre sub-blocos tem o melhor desempenho. O treinamento dos vetores-base do dicionário VSELP não consegue elevar o desempenho da estrutura FIR transversal sobre o do IIR na forma direta I. Ademais, o desempenho do filtro de síntese FIR transversal cai muito quando o comprimento da resposta impulsiva passa da dimensão do bloco para a dimensão do sub-bloco. Entretanto, com o codificador ACELP, o desempenho mantém-se com este nível de truncamento da resposta impulsiva. Ainda, a perda de desempenho situa-se na média de 0,3 dB ao usar-se o FIR transversal no ACELP contra a perda de 0.9 dB para o VSELP para situações equivalentes. Portanto, no caso do ACELP, outras considerações de implementação podem determinar a preferência estrutural. Finalmente, analisam-se as diferenças de desempenho para o ACELP na presença de ruído telefônico. Nesse caso, a distinção estrutural reduz-se a 0,1 dB.

Além das técnicas de supressão de transitórios abordadas no Capítulo 7, que se aplicam nos instantes de transição entre sub-blocos, existem propostas que antecipam a ocorrência dos transistórios no próprio processo de busca da excitação [17].

Ademais, as medidas de desempenho realizadas foram comparações com um sinal de voz original. Como extensão do trabalho, é interessante avaliar o sinal isoladamente. Essa situação de avaliação do sinal isolado ocorre na síntese de voz. É notável que, em testes subjetivos de sintetizadores de voz, se tenham verificado correlações negativas entre distorções temporais e o índice médio de opinião (MOS) [60].

Bibliografia

[1] ALCAIM, A.; SOLEWICS, J.; DE MORAES, J.A. Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, v.7, n. 1, p.23-41, Dez. 1992.

[2] ARJONA RAMÍREZ, M. *Codificador preditivo de voz por análise- mediante-síntese.* São Paulo, 1992. Dissertação de Mestrado - Escola Politécnica da Universidade de São Paulo.

[3] ATAL,B.S.; SCHROEDER, M.R. Predictive coding of speech signals. In: CONF. COMMUN. AND PROCESS., 1967. *Proceedings.* , p.360-1, s.l., IEEE, 1967.

[4] ATAL, B.S.; SCHROEDER, M.R. Adaptive predictive coding of speech signals. *Bell Syst. Tech. J.*, v. 49, n. 8, p.1973-86, Oct. 1970.

[5] ATAL, B.S.; SCHROEDER, M.R. Predictive coding of speech signals and subjective error criteria. *IEEE Trans. Acoust., Speech, Signal Processing*, v.27, n.3, p.247-54, June 1979.

[6] ATAL, B.S. Predictive coding of speech at low bit rates. *IEEE Trans. Commun.*, v.30, n.4, p.600-14, Apr. 1982.

[7] ATAL, B.S.; REMDE, J.R. A new model of LPC excitation for producing natural-sounding speech at low bit rates. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Paris, 1982. *Proceedings.* v.1, p.614-7, s.l., IEEE, 1982.

[8] BEERENDS, J.G.; STEMERDINK, J.A. A perceptual speech-quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, v.42, n.3, p.115-23, March 1994.

[9] CAMPBELL, J.P.; TREMAIN, T.E.; WELCH, V.C. The proposed federal standard 1016 4800 bps voice coder: CELP. *Speech Technology*, v.5, n.2, p.58-64, Apr. May 1990.

[10] CCITT *Recommendation G.711, Pulse code modulation (PCM) of voice frequencies, Blue Book,* v.III.4, ITU-T, Geneva, 1989, p.175-84.

[11] CCITT *40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM). Recommendation G.726*, ITU-T, Geneva, 1990.

[12] CHEN, J.-H.; LIN, Y.-C.; COX, R.V. A fixed-point 16 kb/s LD-CELP algorithm. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Toronto, 1991. *Proceedings.* v.1, p.21-4, s.l., IEEE, 1991.

[13] CHEN, J.-H. A low-delay CELP coder for the CCITT 16 kb/s speech coding standard. *IEEE J. Select. Areas Commun.*, p.830-49, June 1992.

[14] CHEN, J.-H.; GERSHO, A. Adaptive postfiltering for quality enhancement of coded speech. *IEEE Trans. Speech Audio Processing*, v.3, n.1, p.59-71, Jan. 1995.

[15] CHEN, J.-H. Low-Delay Coding of Speech. In: KLEIJN, W.B.; PALIWAL, K.K. (Ed.) *Speech Coding and Synthesis*, p.209-56, Amsterdam, Elsevier Science, 1995.

[16] COX, R.; KROON, P. Low bit-rate speech coders for multimedia applications. *IEEE Spectrum*, v.33, n.12, p.34-41, Dec. 1996.

[17] CUCCHI, S.; FRATTI, M.; RONCHI, M. On improving performance of analysis by synthesis speech coders. *IEEE Trans. Speech Audio Processing*, v.4, n.3, p.243-7, May 1996.

[18] CZARNACH, R.; RABENSTEIN, R.; SCHÜSSLER, H.W. On the stability of certain time-varying digital filters. *Proc. European Signal Processing Conference*, Erlangen, p. 77-80, 1983.

[19] DUDEWICZ, E.J.; MISHRA, S.N. *Modern Mathematical Statistics.* New York, John Wiley & Sons, 1988.

[20] DUNN, H.K.; WHITE, S.D. Statistical measurements on conversational speech. *J. Acoust. Soc. Am.*, v.11, p.278-88, Jan. 1940.

[21] ELECTRONIC INDUSTRIES ASSOCIATION *Cellular system*, Report IS-54, Dec. 1989.

[22] FLANAGAN, J.L. *Speech analysis, synthesis and perception.* 2.ed. Berlin, Springer, 1972.

[23] FISHER, W.M.; DODDINGTON, G.R.; GOUDIE-MARSHALL, K.M. "The DARPA speech recognition research database: specifications and status", *Proc. DARPA Speech Recognition Workshop*, Palo Alto, Feb. 1986. *Proceedings.* v.1, p.93-9, s.l., DARPA, 1986.

[24] GAROFOLO, J.; ROBINSON, T.; FISCUS, J. The development of file formats for very large speech corpora: Sphere and shorten. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Adelaide, 1994. *Proceedings.* v.1, p.113-6.

[25] GERSON, I.A.; JASIUK, M.A. Vector sum excited linear prediction (VSELP) speech coding at 8 kbps. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Albuquerque, 1990. *Proceedings.* v.1, p.461-4, s.l., IEEE, 1990.

[26] GIBSON, J.D. Adaptive prediction for speech encoding. *IEEE ASSP Magazine*, v.1, n.3, p.12-26, July 1984.

[27] HAYASHI, S.;KATAOKA, A.; MORIYA, T. 8 kbit/s Short and Medium Delay Speech Codecs Based on CELP Coding. *European Trans. on Telecommun.*, v.5, n.5, p.49/583-56/90, Sept.-Oct. 1994.

[28] HERNÁNDEZ, L.A.; CASAJÚS-QUIRÓS; F.J., FIGUEIRAS-VIDAL, A.R.; GARCÍA-GÓMEZ, R. On the behavior of reduced complexity code-excited linear prediction (CELP). In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Tokyo, 1986. *Proceedings.* v.1, p.469-72, s.l., IEEE. 1986.

[29] HESS, W. *Pitch determination of speech signals.* Berlin, Springer, 1983.

[30] HONKANEN, T.; VAINIO, J.; JÄRVINEN, K.; HAAVISTO, P. Enhanced full rate codec for IS-136 digital cellular system. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Munich, 1997. *Proceedings.* v.2, p.731-4, s.l., IEEE, 1997.

[31] ITU-T *CCITT Software Tool Library Manual.* ITU-T, Geneva, 01 June 1992.

[32] ITU-T *Recommendation G.723.1 Dual rate speech coder for multimedia applications transmitting at 5.3 and 6.3 kbit/s.* ITU-T, Geneva, March 19, 1996.

[33] ITU *Recommendation G.729 Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP).* ITU-T, Geneva, March 19, 1996.

[34] ITU *Recommendation G.729 - Annex A: Reduced complexity 8 kbit/s using CS-ACELP speech codec.* ITU-T, Geneva, Nov. 8, 1996.

[35] ITU *Recommendation P.861 Objective quality measurement of telephone-band (300-3400 Hz) speech codecs.* ITU-T, Geneva, August 30, 1996.

[36] JANKOWSKI, C.; KALYANSWAMY, A.; BASSON, S.; SPITZ, J. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Albuquerque, 1990. *Proceedings.*, v.1, p.109-12, s.l., IEEE, 1990.

[37] JAYANT, N.S.; NOLL, P. *Digital coding of waveforms.* Englewood Cliffs. Prentice-Hall, 1984.

[38] JOHNSTON, J.D. Estimation of perceptual entropy using noise masking criteria. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, New York, 1988. *Proceedings.*, p.2524-7, s.l., IEEE, 1988.

[39] KLEIJN, W.B.; PALIWAL, K.K. An Introduction to Speech Coding. In: KLEIJN, W.B.; PALIWAL, K.K. (Ed.) *Speech Coding and Synthesis*, p.1-47, Amsterdam, Elsevier Science, 1995.

[40] KROON, P. *Time-domain coding of (near) toll quality speech at rates below 16 kb/s.* Delft, May 1985. Ph. D. dissertation - Delft University of Technology.

[41] KROON, P.; DEPRETTERE, E.F. A class of analysis-by-synthesis predictive coders for high quality speech synthesis at rates between 4.8 and 16 kbits/s. *IEEE J. Select. Areas Commun.*, v.6, n.2, p.353-63, Feb. 1988.

[42] KROON, P.; KLEIJN, W.B. Linear-Prediction based Analysis-by-Synthesis Coding. In: KLEIJN, W.B.; PALIWAL, K.K. (Ed.) *Speech Coding and Synthesis*, p.79-119, Amsterdam, Elsevier Science, 1995.

[43] LAFLAMME, C.; SALAMI, R.; ADOUL, J-P. 9.6 kbit/s ACELP coding of wideband speech. In: ATAL, B.S. et al. (Ed.) *Speech and audio coding for wireless and network applications*, p.147-52, Dordrecht, Kluwer, 1993.

[44] LAPSLEY, P.; BLALOCK, G. How to estimate DSP processor performance. *IEEE Spectrum*, v.33, n.7, p.74-8, July 1996.

[45] LE ROUX, J.; GUEGUEN, C. A fixed point computation of partial correlation coefficients. *IEEE Trans. Acoust., Speech. Signal Processing*, v.25, n.8, p.257-9. June 1977.

[46] MAKHOUL, J.; ROUCOS, S.; GISH, H. Vector quantization in speech coding. *Proc. IEEE*, v.73, n.11, p.155-88, Nov. 1985.

[47] MARKEL, J.D.; GRAY, A.H. *Linear prediction of speech* . Berlin, Springer, 1976.

[48] MASON, S.J. Feedback theory - Some properties of Signal Flow Graphs. *Proc. I.R.E.*, v.41, p.1144-56, Sept. 1953.

[49] OHMURO, H.; IKEDO, J.; MORIYA, T.; KATAOKA, A.; HAYASHI, S.; MANO, K. Dual-pulse CS-CELP: A toll-quality low-complexity speech coder at 7.8 kbit/s. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Atlanta, 1996. *Proceedings*. v.1, p.558-61, s.l., IEEE, 1996.

[50] PAPOULIS, A. *Probability, random variables, and stochastic processes*. 3.ed. New York, McGraw-Hill, 1991.

[51] PORAT, B. *Digital processing of random signals*. Englewood Cliffs, Prentice-Hall, 1994.

[52] RABINER, L.R.; SCHAFER. R.W. *Digital processing of speech signals*. Englewood Cliffs, Prentice-Hall, 1978.

[53] RAMÍREZ, M.A.; ALENS, N. Codificador preditivo de voz excitado por vetores singulares. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, Natal, 1993. *Anais*, v.1, p.86-91, Natal, UFRN, 1993.

[54] RAMÍREZ, M.A.; ALENS, N. Singular-vector-excited linear predictive speech coder. In: IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE, Houston, 1993. *Proceedings*, v.2, p.1294-8, Piscataway, IEEE, 1993.

[55] RAMÍREZ, M.A.; DONATO, U.A.; ALENS, N. Filter structures for a DSP VSELP speech coder In: INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING APPLICATIONS & TECHNOLOGY, Boston, 1995. Proceedings, v.II, p.1854-8, Newton, DSP Associates, 1995.

[56] RAMÍREZ, M.A.; DONATO, U.A.; ALENS, N. Filter structures for a DSP VSELP speech coder In: XIth CHILEAN CONGRESS OF ELECTRICAL ENGINEERING, Punta Arenas, 1995. Proceedings, v.II, p.G-082-7, Universidad de Magallanes, 1995.

[57] RAMÍREZ, M.A.; ALENS, N.; GERKEN, M. Transitórios em Codificadores de Voz CELP. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, Recife, Set. 1997. *Anais*. p.200-4, Recife, UFPE, 1997.

[58] ROSE, R.C.; BARNWELL, T.P. Design and performance of an analysis-by-synthesis class of predictive speech coders. *IEEE Trans. Acoust., Speech. Signal Processing*, v.38, n.9, p.1489-1503, Sept. 1990.

[59] RLE-MIT *Speech Communication Group*. http://rleweb.mit.edu/G-spe.htm, Oct. 1997.

[60] SAGISAKA, P.; IWAHASHI, W.B. Objective Optimization in Algorithms for Text-to-Speech Synthesis. In: KLEIJN, W.B.; PALIWAL, K.K. (Ed.) *Speech Coding and Synthesis*, p.685-706, Amsterdam, Elsevier Science, 1995.

[61] SALAMI, R.; LAFLAMME, C.; BESSETTE, B.; ADOUL, J.-P. Description of ITU-T recommendation G.729 Annex A: reduced complexity 8 kbit/s CS-ACELP codec. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Munich, 1997. *Proceedings*. v.2, p.775-8, s.l., IEEE, 1997.

[62] SCHROEDER, M.R.;ATAL, B.S. Code-excited linear prediction (CELP): High quality speech at very low bit rates. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Tampa, 1985. *Proceedings*. v.2, p.437-40, s.l., IEEE, 1985.

[63] SCHROEDER, M.R. Predictive coding of speech: historical review and directions for future research. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Tokyo, 1986. *Proceedings*. v.4, p.3157-64, s.l., IEEE, 1986.

[64] SCHUR, J. Über Potenzreihen, die im Innern des Einheitskreisen beschränkt sind. *Z. für die Reine und Angewandte Mathematik*, v.147, p.205-32, 1917.

[65] SCLIAR-CABRAL, L. Problemas de processamento lexical com exemplos do português. In: WORKSHOP COGNITIVE SCIENCES AND CONCEPTION OF INFORMATION SCIENCE, Florianópolis, http://www.ctai.rct-sc.br/labiutil/leipzar0.html, Feb. 1996.

[66] SHANNON, C.E. Communication in the presence of noise. *Proc. I.R.E.*, v.37, p.10-21, Jan. 1949.

[67] SINGHAL, S. Reducing computation in optimal amplitude multipulse coders. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Tokyo, 1986. *Proceedings*. v.4, p.2363-6.

[68] TRANCOSO, I.M.; ATAL, B.S. Efficient procedures for finding the optimum innovation in stochastic coders. In: IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESSING, Tokyo, 1986. *Proceedings*. v.4, p.2375-8.

[69] TRANCOSO, I.M.; ATAL, B.S. Efficient procedures for selecting the optimum innovation in stochastic coders. *IEEE Trans. Acoust., Speech, Signal Processing*, v.38, n.3, p.385-96, Mar. 1990.

[70] VORAN, S.; SHOLL, C. Perception-based objective estimators of speech quality. In: IEEE SPEECH CODING WORKSHOP, Annapolis, 1995. *Proceedings*. p.13-4.

[71] YAMAMOTO, J.S. *Algoritmos para redução da taxa de bits em codificadores CELP*. Campinas, 1993. Tese de Doutorado - Faculdade de Engenharia Elétrica - Universidade de Campinas.

[72] ZETTERBERG, L.H.; ZHANG, Q. Elimination of transients in adaptive filters with application to speech coding. *Signal Processing*, v.15, p.419-28, 1988.