UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA
DEPARTAMENTO DE COMUNICAÇÕES

ALGORITMOS PARA REDUÇÃO DA TAXA DE BITS EM CODIFICADORES CELP

JOSÉ SINDI YAMAMOTO
Orientador: *Prof. Doutor FÁBIO VIOLARO*

Banca Examinadora: *FÁBIO VIOLARO (UNICAMP)*
ABRAHAM ALCAIM (PUC-CETUC)
AMAURI LOPES (UNICAMP)
JOÃO MARCOS TRAVASSOS-ROMANO (UNICAMP)
NORMONDS ALENS (USP)

Tese apresentada à Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas - UNICAMP, como parte dos requisitos exigidos para a obtenção do título de DOUTOR EM ENGENHARIA ELÉTRICA.

Novembro - 1993

Resumo

Neste trabalho, novos algoritmos de quantização vetorial dos coeficientes LPC e do sinal de excitação, aplicáveis à codificadores de voz do tipo CELP, foram desenvolvidos e avaliados através de testes subjetivos formais. Uma combinação destes algoritmos quando incorporada em um codificador CELP convencional, melhora significativamente a qualidade do sinal de voz sintetizado, obtendo-se uma boa qualidade mesmo a uma taxa tão baixa quanto 3,55 kbit/s. Adicionalmente, alguns dos algoritmos tem-se mostrado vantajosos na implementação de codificadores de voz de baixo atraso. Assim, codificadores de voz CELP à taxa de 6,8 kbit/s e atraso de 5 ms foram também implementados e avaliados formalmente em termos de qualidade subjetiva.

Agradecimentos

Ao Prof. Dr. Fábio Violaro, meu profundo agradecimento pela sua dedicação e atenção ao meu trabalho durante muitos anos, motivando discussões que abriram e encurtaram vários caminhos como também melhoraram a apresentação das informações contidas neste trabalho. Também, agradeço pela paciência demonstrada com várias das idéias que acabaram não vingando devido ao pouco tempo que pude dedicar ao trabalho em alguns períodos no decorrer destes anos.

Muito importante foi o privilégio que tive de conviver por vários anos com o Grupo de Processamento Digital de Voz do CPqD/TELEBRÁS. O engenheiro José Antônio Martins me proveu de importantes programas de quantização vetorial, e a engenheira Eliana De Martino me manteve informado, durante um longo período, sobre os últimos desenvolvimentos neste campo, além de ter me prestado uma inestimável ajuda na elaboração do plano de testes subjetivos. As engenheiras Flávia Martinho Ferreira e Roberta Abreu Mantegassi muito me influenciaram com constante incentivo e apoio. Meu grande agradecimento a eles e a todos os outros (que foram ou ainda são mantidos) membros deste grupo pela convivência e pronta ajuda que sempre me prestaram.

Também estou profundamente agradecido aos engenheiros Ralph R. Heirinch e Hélio Cesar Salles, gerentes, respectivamente, da Divisão de Enlaces Ópticos e Radioelétricos (DEOR) e Seção de Enlaces Radioelétricos (SER) do CPqD/TELEBRÁS, pela compreensão e atitudes que me permitiram dar continuidade ao trabalho e ainda terminá-lo no CPqD.

Finalmente, meu sincero agradecimento a todas as pessoas que gentilmente participaram do teste subjetivo como avaliadores, sem o que não seria possível concluir este trabalho adequadamente.

Índice

Capítulo 1

INTRODUÇÃO

Algoritmos de codificação de voz têm sido durante décadas uma importante área de pesquisa em desenvolvimento. Nos últimos 10 anos, o nível de atividades e interesse nesta área tem aumentado muito impulsionado pela demanda de uma grande gama de aplicações tais como transmissão de voz em faixa estreita (p. e. , Sistema Rádio Móvel Digital, Sistemas de Comunicação por Satélite, etc.), armazenamento--recuperação de voz e diversas outras formas de aplicação seja na rede telefônica pública como em redes privadas. Atualmente, diversos algoritmos de codificação de voz encontram-se ou estão em vias de serem adotados como padrão para estas aplicações. O padrão de codificação de voz mais antigo depois do PCM a 64 kbit/s, é o ADPCM a 32 kbit/s [1] adotado a nível do CCITT para aplicação geral na rede telefônica. Recentemente, como uma evolução neste campo de aplicação, foi também padronizado pelo CCITT o algoritmo de codificação de voz de baixo atraso a 16 kbit/s denominado LD-CELP [2]. Também a nível do CCITT, encontram-se em andamento atividades visando a padronização de um algoritmo de baixo atraso à taxa de 8 kbit/s. O RPE-LTP a 13 kbit/s [3] foi adotado como padrão europeu para aplicação na primeira geração de telefonia digital celular e, para a segunda geração, já se encontra em fase final de padronização um algoritmo de codificação de voz a meia-taxa. Ainda com relação à aplicação em telefonia celular, o VSELP a 8 e 6,7 kbit/s foi adotado nos E.U.A e Japão, respectivamente [4, 5]. No campo de aplicação em telefonia com privacidade de informação, foi adotado recentemente nos E.U.A um codificador CELP a 4,8 kbit/s (DOD 4.8 kbit/s-PFS1016) [6].

No desenvolvimento de um codificador de voz a baixas taxas, o principal ob-

jetivo é conseguir uma boa qualidade mantendo-se um nível de complexidade que permita a sua implementação em tempo-real e a um baixo custo. Embora os algoritmos de codificação de voz mais recentes sejam muito mais complexos que os algoritmos mais antigos, como o ADPCM, o rápido desenvolvimento da micro-eletrônica, principalmente na área de DSP's (Digital Signal Processors), tem permitido rapidamente a transformação de sofisticados algoritmos em produtos capazes de atender diversas aplicações a baixo custo.

Algumas aplicações específicas exigem outros requisitos. Por exemlo, para o caso de aplicações em rádio móvel digital, a robustez contra condições adversas de transmissão (taxa de erro de bits elevada tanto quanto 10^{-2}, ruído de fundo, etc.) é um dos requisitos mais importantes. Um baixo atraso de codificação torna-se essencial no caso de comunicações pessoais ("personal communications") para evitar uma série de problemas com eco. Em enlaces de longa distância como em comunicações via satélite, é também importante um baixo atraso de codificação para evitar um aumento na dificuldade de conversação já existente devido ao atraso de propagação muito elevado.

Atualmente, o rápido crescimento de demanda das aplicações, principalmente serviços de telefonia celular, tem motivado reduzir a taxa de bits dos codificadores de voz para 4,0 kbit/s ou menos ainda. Nos últimos anos, importantes avanços em algoritmos de codificação de voz a baixas taxas tem sido alcançados, obtendo-se uma produção de voz de boa qualidade a taxa de bits tão baixa quanto 4,8 kbit/s, sendo grande parte destes algoritmos baseados em dois sistemas originais conhecidos como *"Multi-Pulse Excited LPC" (MPE-LPC)* [7] e *"Code Excited Linear Prediction" (CELP)* [8]. Entretanto, apesar destes progressos, um pouco abaixo de 4,8 kbit/s já não se consegue atingir uma boa qualidade de voz [9] e a questão de atraso do algoritmo de codificação continua crítica em muitas aplicações. O desempenho do codificador CELP, por exemplo, degrada rapidamente para taxa de bits abaixo de 4,8 kbit/s.

O presente trabalho tem como objetivo melhorar a qualidade do sinal de voz sintetizado e diminuir o atraso de codificação dos codificadores de voz do tipo CELP a taxas de bits abaixo de 4,8 kbit/s. Para tanto, novos algoritmos de quantização vetorial dos diversos parâmetros destes codificadores são desenvolvidos. Estes algoritmos se classificam em:

- Algoritmos de baixo atraso para redução da taxa de bits dos coeficientes LPC;
- Algoritmos para redução conjunta da taxa de bits do sinal de excitação e dos coeficientes LPC;
- Algoritmos para redução da taxa de bits do sinal de excitação.

Inicialmente, no capítulo 2 são apresentados os métodos objetivos e subjetivos utilizados para a avaliação dos diversos algoritmos desenvolvidos. No capítulo 3, é apresentada uma visão geral dos modelos de codificação de voz MPE-LPC e CELP sobre os quais se aplicam e baseiam estes algoritmos. As contribuições efetivas deste traballho, na área de codificação de voz a baixas taxas e baixo atraso, iniciam-se no capítulo 4 com a introdução dos algoritmos de baixo atraso para redução da taxa de bits dos coeficientes LPC. Em seguida, são apresentados os algoritmos para redução conjunta da taxa de bits dos parâmetros do sinal de excitação e dos filtros de síntese de longo-prazo e curto-prazo, que se constituem na principal contribuição deste trabalho. No capítulo 5, são apresentadas as contribuições relativas à redução da taxa de bits do sinal de excitação. Finalmente, no capítulo 6. diversos codificadores de voz do tipo CELP a baixas taxas (< 4.0 kbit/s) e a taxas médias (< 7 kbit/s) com baixo atraso, de qualidade melhorada a partir da incorporação de uma combinação dos algoritmos apresentados nos capítulos 4 e 5, são implementados e avaliados formalmente em termos de MOS ("Mean Opinian Score").

Bibliografia

[1] Recomendação CCITT G.721,"32 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)", livro vermelho, Outubro de 1984.

[2] J.-H. Chen, R.V. Cox, Y.-C. Lin, N. Jayant e M.J. Melchner, "A Low-Delay CELP Coder for the CCITT 16 kbit/s Speech Coding Standard", IEEE Journal on Selected Areas in Communications, vol. 10, nº 5, junho de 1992, pág. 830-849.

[3] P. Vary, K. Hellwig, R. Hofmann, R.J. Sluyter, C. Galand, e M. Rosso,"Speech Codec for the European Mobile Radio System", IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1988, pág. 227-230.

[4] I.A. Gerson e M.A. Jasiuk,"Vector Sum Excited Linear Prediction (VSELP)", Advances in Speech Coding, Boston, MA: Kluwer, 1991, pág. 69-79.

[5] MPT/VSELP Functional Description - fonte : Motorola-1990.

[6] J.P. Campbell, Jr.T.E. Tremain e V.C. Welch,"The DOD 4.8 kbps Standard (Proposed Federal Standard 1016)", Advances in Speech Coding, Boston, MA: Kluwer, 1991, pág. 121-133.

[7] B. S. Atal, J. R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates", IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1982, pág. 614-617.

[8] M. R. Schroeder e B. S. Atal, "Code-Excited Linear Prediction (CELP) : high--quality speech at very low bit rates", IEEE Int. Conf. Acoust., Speech, Signal Process., Março de 1985, pág. 937-940.

[9] Atal B.S. e Caspers B.E., "Beyond Multipulse and CELP Towards High Quality Speech at 4 kbit/s", Advances in Speech Coding, Boston, MA: Kluwer, 1991, pág. 191-201.

Capítulo 2

MÉTODOS DE AVALIAÇÃO

2.1 INTRODUÇÃO

Um dos problemas mais importantes no desenvolvimento de codecs de voz a baixas taxas está em como avaliar adequadamente a *qualidade* do sinal de voz sintetizado. O conceito de qualidade de um sinal de voz é extremamente subjetivo e evasivo envolvendo não só inteligibilidade, mas também todos os aspectos dos processos de percepção humana da voz bem como gostos individuais. De fato, quando uma pessoa ouve uma voz sintetizada para dar a sua opinião em termos de qualidade (p. e. , uma nota variando de 1 a 5), além dos aspectos acústicos como nível de potência do sinal de voz sintetizado, ruído de quantização e qualquer outro tipo de distorção, os seus conhecimentos de lingüística, idioma, como também quaisquer outras informações relacionadas com o locutor são explorados. Métodos subjetivos de avaliação usando testes de opinião são os que fornecem medidas que melhor englobam todos estes fatores. Dentre os vários tipos de testes de opinião, um dos mais utilizados é o *teste de categorias*, também chamado de *ACR(Assessment Category Rating)*. Neste tipo de teste, várias pessoas, denominados avaliadores, ouvem um conjunto de frases de pequena duração sob diversas condições e fazem a seguinte atribuição de notas à qualidade do sinal de voz [1, 2, 3, 4] :

5 → qualidade *Excelente*
4 → qualidade *Boa*
3 → qualidade *Regular*
2 → qualidade *Pobre*
1 → qualidade *Péssima*

Após a coleta de notas de um número suficiente de avaliadores, calcula-se o *MOS(Mean Opinion Score)* no qual um valor médio das notas dos vários avaliadores é determinado. Entretanto, a realização dos métodos de avaliação subjetivos formais, isto é, segundo regras definidas que garantam uma boa confiabilidade dos resultados, exige um extenso trabalho que os torna custosos e demorados. Uma grande quantidade de arquivos de sinais de voz e sob diversas condições é utilizada. Para a geração destes arquivos de voz e a sua posterior audição, uma vez processados pelo codec em avaliação (e outros sistemas de referência sob diversas condições de operação), são necessários uma grande quantidade de locutores e avaliadores. Por outro lado, os métodos de avaliação objetivos são baratos e rápidos de serem realizados. Adicionalmente, o desempenho do codec é facilmente melhorado pela minimização direta de uma distorção definida por uma medida objetiva. Contudo, até agora, o problema de avaliação da qualidade de diferentes codecs através de uma medida objetiva universal não está resolvido.

Um compromisso entre os testes subjetivos formais e os testes objetivos, são os testes subjetivos informais. Neste caso, são utilizados apenas um pequeno número de arquivos de voz, locutores e avaliadores, de modo que o custo e o tempo de execução são muito mais reduzidos. Os testes subjetivos informais são muito úteis para se ter uma idéia inicial da potencialidade de desempenho do codec. Também são indispensáveis durante o desenvolvimento de qualquer algoritmo no processo de otimização dos diversos parâmetros.

No presente trabalho, métodos de avaliação objetivos acompanhados de testes subjetivos informais são empregados como meios comparativos de verificação de melhoria e determinação dos fatores de desempenho dos vários algoritmos investigados. Os testes subjetivos formais são utilizados na avaliação final dos codecs implementados.

2.2 MÉTODOS DE AVALIAÇÃO OBJETIVOS

Nesta seção é feita a descrição dos três métodos de avaliação objetivos utilizados no presente trabalho, a saber : Razão Sinal Ruído total ou de longo-prazo (RSR_{total}), Razão Sinal Ruído segmentar (RSR_{seg}) e Distância Cepstral (DC). Os dois primeiros métodos são definidos no domínio do tempo, enquanto que o terceiro método consiste de uma medida de distorção no domínio da frequência.

2.2.1 Razão Sinal Ruído Total

Sejam $x(n)$ e $y(n)$ o sinal de voz original e sintetizado, respectivamente. Então, a RSR_{total} é dada por :

$$RSR_{total} = 10 \log \frac{\sum_n x^2(n)}{\sum_n [x(n) - y(n)]^2} \tag{2.1}$$

onde a somatória é tomada sobre o total de amostras do sinal de voz de entrada.

A RSR_{total} é menos correlacionado com qualidade subjetiva que muitos outros métodos de avaliação objetivos tais como RSR_{seg} e DC. Entretanto, muitas vezes é utilizado durante o projeto e refinamentos de algoritmos de codificação de voz.

2.2.2 Razão Sinal Ruído Segmentar

Para o cálculo da RSR_{seg} utilizada neste trabalho, o sinal de voz, $x(n)$, é inicialmente dividido em L segmentos de 128 amostras. Em seguida, é calculada a potência média em dB do segmento l conforme a seguinte expressão :

$$P_{seg}(l) = 10 \log \frac{1}{128} \sum_{m=1}^{128} x^2[m + 128(l-1)], \qquad l = 1, 2, ..., L \tag{2.2}$$

A RSR_{seg} é, então, dada por :

$$RSR_{seg} = \frac{1}{L - M} \sum_{l=1}^{L} Q(l) \quad dB, \tag{2.3}$$

onde : M = número de segmentos tal que $P_{seg}(l) < P_{segmax} - 50dB$;
P_{segmax} = potência segmentar máxima em dB do arquivo de voz considerado.

$$Q(n) = \begin{cases} 10\log \dfrac{\sum_{m=1}^{128} x^2[m+128(l-1)]}{\sum_{m=1}^{128}\{x[m+128(l-1)] - y[m+128(l-1)]\}^2}, & \text{para } P_{seg}(l) \geq P_{segmax} - 50dB \\ 0, & \text{para } P_{seg}(l) < P_{segmax} - 50dB \end{cases} \quad (2.4)$$

De acordo com o cálculo de $Q(n)$, segmentos do arquivo de voz de potência muito baixa (p.e., períodos de silêncio) não são considerados no cálculo da RSR_{seg}. Este procedimento é justificado pelo fato de que segmentos de muito baixa potência pioram os valores de RSR_{seg}, enquanto que causam um impacto desprezível em termos de qualidade subjetiva.

2.2.3 Distância Cepstral

A Distância Cepstral é calculada através de análise cepstral como sendo [8] :

$$DC = \frac{10}{\ln 10}\{2\sum_{k=1}^{p}[c_x(k) - c_y(k)]^2\}^{\frac{1}{2}} \quad (2.5)$$

onde : p é a ordem do preditor utilizado na análise LPC;
$c_x(k)$ e $c_y(k)$ são os coeficientes cepstrais LPC do sinal de voz original e sintetizado, respectivamente.

Para o cálculo dos coeficientes cepstrais é feita a análise de predição linear de dois modelos, $A_x(z)$ e $A_y(z)$, correspondentes ao sinal original e sintetizado, respectivamente, ambos definidos como :

$$A(z) = 1 - \sum_{k=1}^{p} a(k)z^{-k} \quad (2.6)$$

onde : $a(k), k = 1, ..., p$ são os coeficientes LPC. Posteriormente, os coeficientes LPC $a(k)$ são transformados para coeficientes cepstrais $c(k)$ através da seguinte expressão [5]:

$$c(n) = \frac{1}{n}[\sum_{k=1}^{n-1}(n-k)c(n-k)a(k) + na(n)], \qquad n > 0 \quad (2.7)$$

Estudos comparativos de métodos de avaliação objetivos [6, 7, 8], mostram que a medida de Distância Cepstral é a que melhor se correlaciona com os resultados baseados em métodos de avaliação subjetivos.

2.3 MÉTODOS DE AVALIAÇÃO SUBJETIVOS

2.3.1 Testes Subjetivos Informais

Os testes subjetivos informais utilizados consistem na audição do sinal de voz sintetizado sobre o qual foram calculados os valores de RSR_{seg} e DC. Tendo em vista o reduzido número de avaliadores, nenhum valor de medida subjetiva como, por exemplo, o MOS é determinado. O objetivo principal destes testes é o de complementar as medidas de RSR_{seg} e DC no sentido de fornecer uma noção mais precisa da qualidade do sinal de voz sintetizado ou diagnosticar eventuais problemas.

2.3.2 Testes Subjetivos Formais

Os testes subjetivos formais [2, 3, 4] tornam-se imprescindíveis de serem realizados quando se deseja um valor numérico de medida que reflita precisamente a qualidade do sinal de voz sintetizado por um codec. Este valor de qualidade pode ser obtido sob diversas condições de operação do codec levando-se em conta efeitos tais como :

- Efeitos da taxa de erro de bits e nível de entrada;
- Efeito cascata e nível de entrada;
- Efeito de ruído ambiental e música;
- Dependência com o locutor.
- Etc..

O valor de medida mais utilizado é o MOS, obtido segundo regras estatísticas que garantem a sua validade, e um plano de testes ou experiências deve ser elaborado para cada uma destas condições de operação do codec. Os níveis de entrada são medidos utilizando-se um voltímetro de voz conforme recomendação CCITT P.56. Os arquivos de voz devem ser gravados em local apropriado com ruído ambiental de no máximo 30 dBA.

Nas experiências para avaliação subjetiva em termos de MOS, normalmente utiliza-se um sistema de referência denominado MNRU (Modulated Noise Reference Unit) como proposto na recomendação CCITT P.70. O MNRU consiste de sinal de voz com ruído branco tendo uma amplitude proporcional a do sinal de voz e razão sinal/ruído Q. Diz-se que um sinal de voz sintetizado tem *opinião equivalente* Q quando o seu valor de MOS obtido for igual ao do sistema MNRU com razão

sinal/ruído Q [9].

O plano de testes subjetivos formais, utilizado na avaliação final dos codecs implementados neste trabalho, é descrito no apêndice A. Por questões de custo e dificuldades operacionais, somente a experiência para avaliar a qualidade dos codecs na ausência de erro de bits é realizada.

2.4 ARQUIVOS DE VOZ

Os arquivos de voz são obtidos usando-se um sistema de aquisição como mostrado na figura 2.1, onde o filtro IRS (Intermediate Reference System) é o filtro da parte de transmissão do sistema IRS [10] e tem a função de simular a resposta em freqüência da cápsula transmissora do telefone e um meio de transmissão ideal até a entrada do codificador de voz. O microfone utilizado deve ter uma resposta plana e o sinal de voz filtrado pelo filtro IRS deve ser digitalizado através de um conversor A/D de 16 bits de resolução.

Os arquivos de voz, mesmo que sejam obtidos utilizando-se um mesmo sistema de aquisição e tenham sido normalizados para uma mesma potência ativa, apresentam entre si valores diferentes de fator de atividade e conteúdo espectral em função do locutor, idioma e tipo de conversação, de modo que os resultados de uma avaliação objetiva de um codec dependem do arquivo de voz utilizado. Para que os valores de RSR_{seg} e DC obtidos para diferentes algoritmos de codificação de voz ou durante as etapas do processo de otimização possam ser analisados, é necessário que estas medidas sejam feitas utilizando-se sempre um mesmo conjunto de arquivos de voz.

A função de transferência do filtro IRS, mostrada na figura 2.2, realça as altas freqüências do sinal de entrada e atenua as freqüências muito baixas, em torno da freqüência correspondente ao período de pitch. Assim, o filtro IRS torna o espectro do sinal de voz bastante desfavorável para redução da taxa de bits e representa a condição de pior caso que pode ser encontrado na prática.

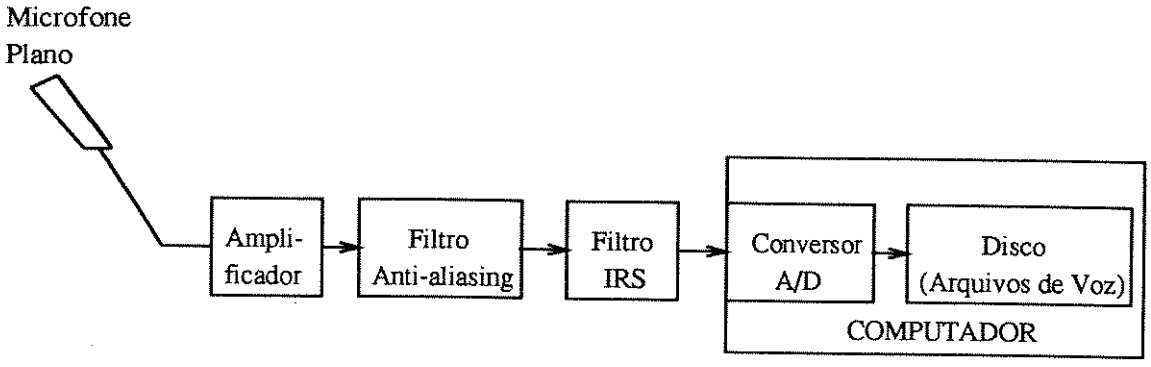

Figura 2.1: Sistema de aquisição de voz

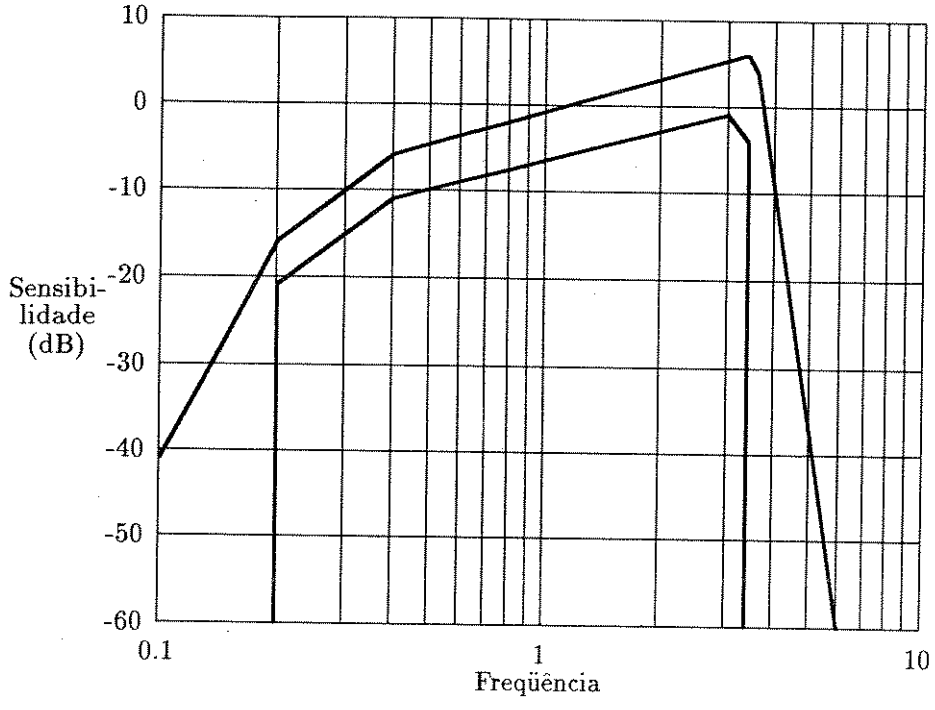

Figura 2.2: Função de transferência do filtro IRS (parte de transmissão do sistema IRS)

Bibliografia

[1] Recomendação CCITT : "Absolute Category Rating (ACR) Method for Subjective Testing of Digital Processes", Suplemento 14, Anexo A, Volume V, Livro Azul, 1988.

[2] Documento CCITT SG-XII : "Subjective Test Methodology for a Universal 16 kbit/s Coder", ver. 2.0, 28 de Fevereiro de 1989.

[3] Documento CCITT SG-XII :"Subjective Test Methodology for a 8 kbit/s Speech Coder", ver. 1.0, 1 de Abril de 1993.

[4] S. F. C. Neto,"Metodologias de Avaliação de Algoritmos de Codificação de Voz", Dissertação de Tese de Mestrado, UNICAMP/FEE/DECOM, Abril de 1993.

[5] J. D. Markel, A. H. Gray, Jr., "Linear Prediction of Speech", Springer-Verlag Berlin Heidelberg New York 1976, terceira edição, pág. 229-230.

[6] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voiceband codecs", IEEE Int. Conf. Acoust., Speech, Signal Process., 1982, Paris, pág.1000-1003.

[7] Contribuição CCITT, COM-XII-8-E, Abril/1985,"Proposal of Objective Quality Measure for Voiceband Codecs".

[8] Nobuhiko Kitawaki, Hiromi Nagabuchi and Kenzo Itoh, "Objective Quality Evaluation for low-bit-rate speech coding systems", IEEE Journal on Selected Areas in Communications, vol. 6, nº 2, fevereiro de 1988.

[9] CCITT : T. Watanabe, H. Nagabuchi e N. Kitawaki, "Law of Addition for Subjective Quality of Voiceband Codecs", Review of the Electrical Communications Laboratories, vol. 35, nº 4, 1987.

[10] Recomendação CCITT P.48 :"Specification for an Intermediate Reference System", vol. 5, Livro Azul, 1988.

Capítulo 3

MODELOS DE CODIFICAÇÃO DE VOZ MPE-LPC E CELP

3.1 INTRODUÇÃO

De uma maneira geral, os algoritmos de codificação de voz podem ser classificados em *codificadores de forma de onda, paramétricos* e *híbridos.* Os da primeira categoria tentam reproduzir amostra por amostra a forma de onda do sinal de voz original com a menor distorção possível. Os do segundo tipo procuram uma reprodução do espectro de freqüência a partir de um modelo que leva em conta os parâmetros mais críticos do mecanismo de reprodução da voz humana tais como freqüência fundamental das cordas vocais (período de "pitch"), classsificação dos sons em sonoros e não-sonoros, volume, etc. Já os codificadores da terceira classe, denominados codificadores híbridos, combinam algumas vantagens dos codificadores de forma de onda e paramétricos.

Os codificadores de forma de onda são capazes de reproduzir sinais de voz de boa qualidade à taxa de bits de até aproximadamente 16 kbit/s [1, 2, 3] e tem a vantagem de apresentarem baixo atraso e baixa complexidade de implementação. Além disto, normalmente permitem a transmissão de facsímile e de dados via modem a baixas taxas.

Um codificador paramétrico largamente conhecido é o *Vocoder LPC (Voice Coder Linear Predictive Coding)*, o qual consiste de um modelo de excitação e filtro

de síntese LPC como mostrado na figura 3.1 e atualizado a cada bloco de N amostras. Neste modelo de excitação, o sinal de voz é classificado basicamente em dois tipos de sons : sonoros e não-sonoros. Para os sons sonoros, a fonte de excitação é representada por um trem de pulsos periódicos segundo intervalos do período de pitch. Por outro lado, para os sons não-sonoros, um ruído aleatório representa a fonte de excitação. Este simples modelo de excitação é eficiente para reduzir a taxa de bits de um vocoder LPC até em torno de 500 bit/s [4, 5, 6, 7], mas às custas de uma sensível redução na qualidade do sinal de voz. Esta degradação deve-se a imprecisões na detecção do período de pitch, decisão sonoro/não-sonoro e falhas inerentes do próprio modelo. Por exemplo, a classificação do sinal de voz em apenas dois tipos, exclusivamente sonoros e exclusivamente não-sonoros, está longe de ser perfeita.

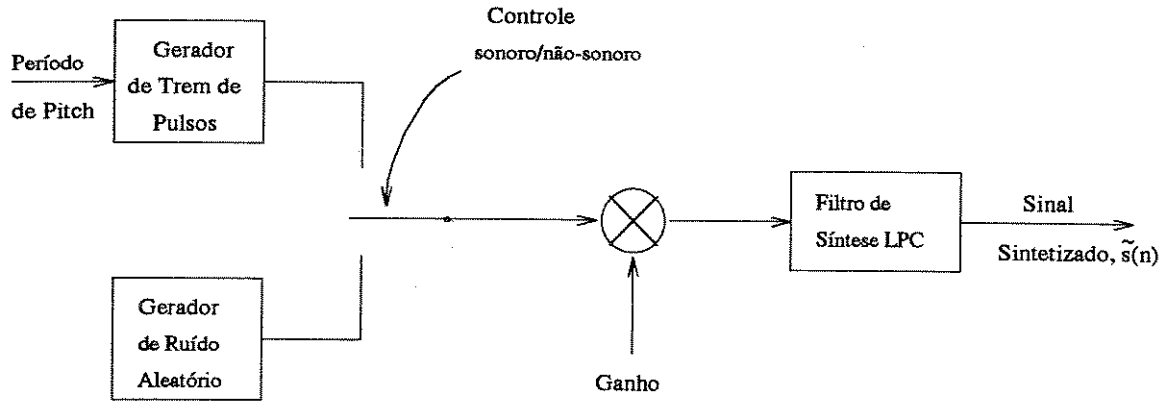

Figura 3.1: Vocoder LPC

O codificadores híbridos procuram contornar de alguma forma os problemas inerentes aos Vocoders LPC e assim melhorar a qualidade do sinal de voz sintetizado. Um dos codificadores híbridos mais conhecidos é o MPE-LPC ("Multipulse Excited Linear Predictor Coding") [8]. Ao contrário do Vocoder LPC, o MPE-LPC é um codificador LPC onde o sinal de excitação, denominado $multipulso(MP)$, consiste de um número fixo de pulsos por bloco, independentemente de o trecho do sinal de voz ser um som sonoro ou não-sonoro. Assim, o desempenho resultante devido ao modelo de excitação MP não depende do modelamento do sinal fonte em apenas dois tipos bem como não necessita de uma estimativa do período de pitch (a não ser que inclua um preditor de longo-prazo com o intuito de melhorar a qualidade do sinal de voz sintetizado como será visto mais adiante). O problema do modelo de excitação MP se resume na dificuldade em se determinar as amplitudes e as posições

dos pulsos que compõem o sinal de excitação.

Um outro modelo de excitação para um filtro de síntese LPC, é o de excitação por *dicionário de códigos (Codebook)*, e os codificadores de voz com este tipo de excitação são conhecidos como *CELP (Code Excited Linear Predictor)* [9]. Neste modelo, um conjunto de possíveis seqüências do sinal de excitação é previamente armazenado formando-se o dicionário de códigos. Para cada bloco de amostras do sinal de voz, é escolhida uma seqüência (aquela que melhor reproduz o sinal de voz original), dentre aquelas pertencentes ao dicionário, para servir como sinal de excitação para aquele bloco.

3.2 MODELO DE EXCITAÇÃO MULTIPULSO

Na figura 3.2 é mostrado o diagrama em blocos do modelo LPC de excitação MP. O diagrama é bastante parecido com o do Vocoder LPC, exceto pela ausência do gerador de pulsos, gerador de ruído e do comutador de sons sonoros/não-sonoros. A excitação para o bloco sintetizador LPC é produzida por um único gerador de excitação, que produz uma seqüência de pulsos localizados nos instantes m_1, m_2,...,m_i, ..., com amplitudes A_1, A_2, ...,A_i,..., respectivamente. Se o número de pulsos é aumentado arbitrariamente a valores grandes, de modo que haja um pulso a cada instante de amostragem, é possível obter-se uma réplica do sinal de voz original (ao custo de uma alta taxa de bits). Entretanto, é possível sintetizar todos os tipos de sons de voz, incluindo os sons sonoros e não-sonoros, com distorção quase imperceptível subjetivamente, empregando-se apenas poucos pulsos por bloco (por exemplo, 8 pulsos a cada bloco de 5ms).

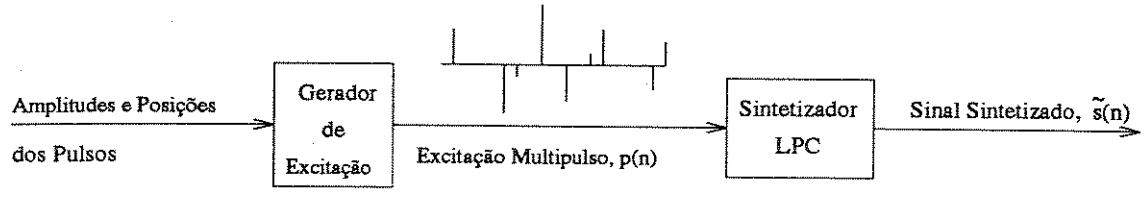

Figura 3.2: Sintetizador LPC com excitação Multipulso

3.2.1 Cálculo da Excitação Multipulso

Os processamentos envolvidos num codec utilizando o algoritmo MPE-LPC

terminam quando as posições m_i e amplitudes A_i dos pulsos de excitação são conhecidas em um dado intervalo de tempo. As amplitudes e posições dos pulsos são selecionadas por um procedimento de análise por síntese, de modo que o sinal de voz sintetizado satisfaça um critério de fidelidade relativo ao sinal original.

Muitas variações de métodos de cálculo das amplitudes e posições dos pulsos da excitação MP têm sido propostas na literatura [10, ...,18]. Neste trabalho, optou-se por uma implementação do método de Berouti et al. [10] onde a análise MP pode ser representada segundo a figura 3.3

A função de transferência de $H(z/\mu)$, denominado *filtro de síntese LPC com ponderação*, é dada por :

$$H(z/\mu) = \frac{1}{1 - \sum_{k=1}^{p} a_k \mu^k z^{-k}} \tag{3.1}$$

Na análise MP são utilizados dois filtros de síntese LPC com ponderação $H(z/\mu)$. O primeiro é excitado pelo sinal de resíduo, $r(n)$, recuperando-se o sinal de voz original mas com ponderação, $s_w(n)$. O segundo é excitado pelo sinal de excitação MP, $p(n)$, obtendo-se um sinal de voz sintetizado com ponderação, $\tilde{s}_w(n)$. A diferença entre o sinal de voz ponderado original, $s_w(n)$, e o sinal de voz ponderado sintetizado, $\tilde{s}_w(n)$, deve ser minimizada durante o processo de geração da excitação MP, $p(n)$, utilizando-se como medida o erro quadrático médio. Portanto, o fator de ponderação μ, utilizado nos dois filtros de síntese LPC, pondera o erro entre o sinal de voz original e sintetizado, de modo que a razão *sinal/ruído* seja menor na região das formantes e maior na região entre formantes, pois para a audição humana o ruído na região das formantes é mascarado pelo sinal de voz[1].

Em termos de implementação, o sinal de voz original é particionado em blocos de N amostras, sendo alocados em cada bloco N_p pulsos . Sejam $h_v(n)$ e $h_w(n)$ a resposta causal ao impulso do filtro inverso, $1/H(z)$, e do filtro de síntese LPC com ponderação, $H(z/\mu)$, respectivamente. Então, o sinal de resíduo, $r(n)$, é dado por:

$$r(n) = \sum_{k=-\infty}^{n} s(k) h_v(n-k), \qquad 0 \leq n < N \tag{3.2}$$

O sinal de voz original com ponderação, $s_w(n)$, é obtido passando-se o sinal de resíduo, $r(n)$, pelo filtro de síntese LPC com ponderação, $H(z/\mu)$:

[1] Pode-se também fazer $r(n) - p(n)$ e com a diferença obtida excitar um único filtro $H(z/\mu)$

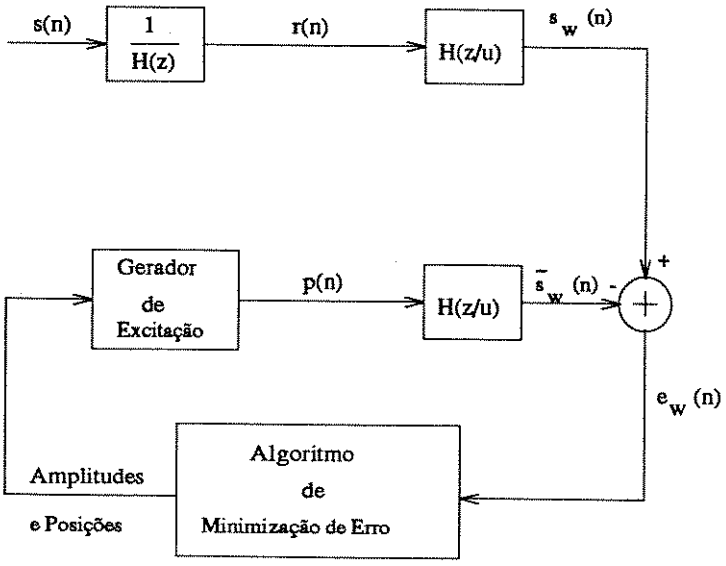

Figura 3.3: Diagrama em blocos do procedimento de análise Multipulso

$$s_w(n) = \sum_{k=-\infty}^{n} r(k)h_w(n-k), \qquad 0 \leq n < N \tag{3.3}$$

O sinal de voz sintetizado com ponderação, $\tilde{s}_w(n)$, é calculado passando-se o sinal de excitação, $p(n)$, através do segundo filtro de síntese com ponderação, $H(z/\mu)$:

$$\tilde{s}_w(n) = \sum_{k=-\infty}^{n} p(k)h_w(n-k), \qquad 0 \leq n < N \tag{3.4}$$

As amplitudes dos pulsos da excitação MP, $A_i = p(m_i)$, e as suas posições correspondentes, m_i, são obtidas através de um processo de minimização do erro quadrático médio ponderado dado por[2] :

$$\langle e_w^2 \rangle = \frac{1}{N} \sum_{n=0}^{N-1} [s_w(n) - \tilde{s}_w(n)]^2 \tag{3.5}$$

Nas equações (3.2) a (3.4), o limite inferior da somatória foi tomado como $-\infty$ ao invés de zero para levar em conta a memória dos filtros devido aos blocos anteriores. Para simplificar a obtenção do algoritmo de minimização de erro, considere-se que a cada bloco de N amostras, quando da aplicação do sinal $p(n)$ ao filtro $H(z/\mu)$, o mesmo tem suas condições iniciais zeradas (memória nula). Neste caso, deve-se descontar de $s_w(n)$ a memória do filtro de síntese $H(z/\mu)$ devido a excitação MP do

[2]Por simplicidade, daqui por diante é omitido da expressão do erro quadrático médio o termo $\frac{1}{N}$

bloco anterior. Com estas considerações resulta a figura 3.4, onde $d_o(n)$ é a resposta do filtro de síntese inferior à entrada nula após ter sido excitado no bloco anterior pela seqüência ótima $p(n)$, e $d(n)$ é o sinal de referência para a minimização de erro. Nestas condições resulta :

$$\tilde{s}_w(n) = \sum_{k=0}^{n} p(k)h_w(n-k), \qquad 0 \leq n < N \tag{3.6}$$

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} [d(n) - \tilde{s}_w(n)]^2 \tag{3.7}$$

A partir desta equação, dois métodos de determinação da excitação MP são possíveis : i) método com reotimização das amplitudes dos pulsos já alocados; ii) método simplificado, isto é, sem reotimização das amplitudes dos pulsos já alocados.

Determinação da Excitação MP com Reotimização de Amplitudes

Após serem colocados N_p pulsos nas posições m_i, $i = 1, ..., N_p$, substituindo-se a equação (3.4) em (3.7) obtém-se :

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} [d(n) - \sum_{i=1}^{N_p} A_i h_w(n - m_i)]^2, \tag{3.8}$$

onde A_i é a amplitude do i-ésimo pulso. Diferenciando-se esta expressão em relação às amplitudes A_i, obtém-se o seguinte conjunto de equações simultâneas:

$$\begin{bmatrix} \Phi(m_1,m_1) & \Phi(m_1,m_2) & \cdots & \Phi(m_1,m_{N_p}) \\ \Phi(m_2,m_1) & \Phi(m_2,m_2) & \cdots & \Phi(m_2,m_{N_p}) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(m_{N_p},m_1) & \Phi(m_{N_p},m_2) & \cdots & \Phi(m_{N_p},m_{N_p}) \end{bmatrix} \times \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ \vdots \\ \hat{A}_{N_p} \end{bmatrix} = \begin{bmatrix} \beta_{m_1} \\ \beta_{m_2} \\ \vdots \\ \beta_{m_{N_p}} \end{bmatrix} \tag{3.9}$$

onde:

$$\Phi(i,j) = \sum_{n=0}^{N-1} h_w(n-i)h_w(n-j), \qquad 0 \leq i,j < N \tag{3.10}$$

$$\beta_m = \sum_{n=0}^{N-1} d(n)h_w(n-m), \qquad 0 \leq m < N \tag{3.11}$$

e $\hat{A}_i$, $i = 1, ..., N_p$, são as incógnitas. Na forma matricial obtém-se:

$$\underline{\Phi}.\underline{\hat{A}} = \underline{\beta} \tag{3.12}$$

O método adotado para a resolução desta equação é o algoritmo de Cholesky baseado na decomposição triangular da matriz $\underline{\Phi}$ [19]. Este procedimento é aplicado de uma

maneira seqüencial à medida em que a posição m_i de cada novo pulso é determinada. Pode-se demonstrar que a melhor posição m_1 para um único pulso é o valor de m

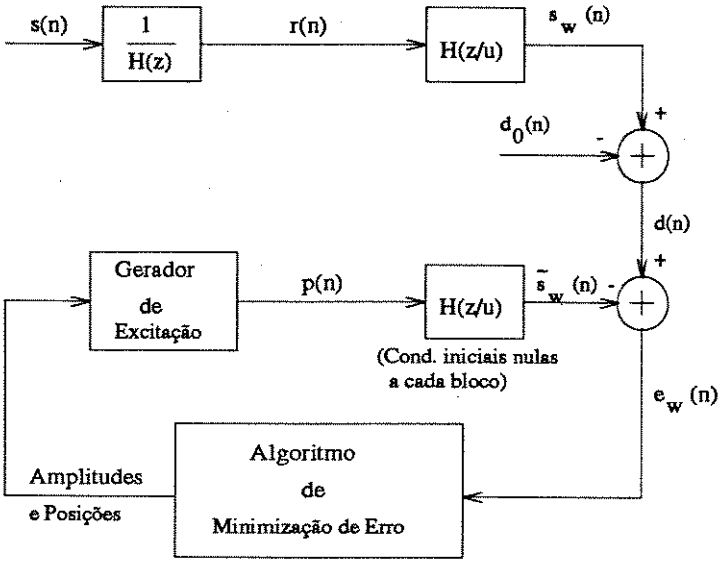

Figura 3.4: Diagrama em blocos do procedimento de análise MP com condições iniciais nulas do filtro de síntese LPC.

que maximiza a razão [10] :

$$\left| \frac{\beta_m}{\Phi(m,m)} \right| \tag{3.13}$$

Assim, a cada nova posição m_i do pulso determinada segundo este procedimento, resolvendo-se a equação (3.12) pelo método de Cholesky, determina-se o valor da amplitude $\hat{A}_{lp}$ (lp é o número de pulsos até então alocados), e os valores das amplitudes $\hat{A}_i$, $i = 1, ..., (lp-1)$ são re-otimizados de modo a minimizar o erro quadrático médio ponderado dado pela equação (3.7).

Neste procedimento de determinação e re-otimização da amplitude dos pulsos, o efeito dos pulsos que forem sendo alocados deve ser removido da seqüência $\{d(n)\}$ (que será usada para determinar a posição ótima do próximo pulso) :

$$d'(n) = d(n) - \sum_{i=1}^{lp} \hat{A}_i h_w(n - m_i) \tag{3.14}$$

onde lp é o número de pulsos já alocados. Substituindo-se o novo valor $d'(n)$ em (3.11), obtém-se uma fórmula para a atualização da função de correlação-cruzada β_m:

$$\beta'_m = \beta_m - \sum_{i=1}^{lp} \hat{A}_i \Phi(m_i, m) \tag{3.15}$$

Desta maneira, para a determinação da posição do próximo pulso, isto é, do $(lp+1)$--ésimo pulso, a razão a ser maximizada é:

$$\left| \frac{\beta'_m}{\Phi(m,m)} \right| \tag{3.16}$$

Determinação da Excitação MP sem Reotimização de Amplitudes

A geração da excitação MP sem reotimização de amplitudes é feita sequencialmente pulso por pulso, determinando-se primeiramente a posição e em seguida a amplitude. A determinação da posição do primeiro pulso, $\hat{m}$, é feita maximizando-se a equação (3.13) e a sua amplitude é dada por :

$$\hat{A}_{\hat{m}} = \frac{\beta_{\hat{m}}}{\Phi(\hat{m}, \hat{m})} \tag{3.17}$$

Analogamente ao caso da geração de excitação MP com reotimização dos pulsos, o efeito dos pulsos que forem sendo alocados deve ser removido da seqüência original $d(n)$ conforme a equação (3.14), substituindo-se $h_w(n)$ por $f(n)$. Assim, a posição dos demais pulsos é determinada maximizando-se a equação (3.16) e a amplitude é dada por :

$$\hat{A}_{\hat{m}} = \frac{\beta_{\hat{m}} - \sum_{i=1}^{l_p} \hat{A}_i \Phi(m_i - \hat{m})}{\Phi(\hat{m}, \hat{m})} \tag{3.18}$$

3.2.2 Modelo de Excitação Multipulso com Filtro de Síntese de Longo-Prazo

Para se obter um melhor desempenho, o modelo de excitação MP pode incluir um filtro de síntese de longo-prazo, $H_p(z)$, antes do filtro de síntese LPC com ponderação $H(z/\mu)$, conforme ilustrado na figura 3.5. Neste caso, para o cálculo da excitação MP, deve-se lembrar que $d_o(n)$ na figura 3.5 leva em conta a memória da combinação dos filtros de síntese LPC e longo-prazo, isto é, $H_p(z)H(z/\mu)$. Adicionalmente, na equação (3.6), deve-se empregar a resposta ao impulso desta combinação dos filtros, $f(n) = h_p(n) * h_w(n)$, ao invés de simplesmente $h_w(n)$.

O filtro de síntese de longo-prazo é, normalmente, de primeira ordem e com a seguinte função de transferência :

$$H_p(z) = \frac{1}{1 - \gamma z^{-M}} \tag{3.19}$$

onde : M é o período de pitch e
γ é o ganho de pitch (ganho do filtro de síntese de longo-prazo).

A determinação do período de pitch M e do ganho γ, pode ser feita em malha aberta ou em malha fechada. Um método em malha aberta bastante utilizado consiste em determinar como período de pitch M o valor de m que maximiza a seguinte função de autocorrelação normalizada [23] :

$$\frac{C(m)}{\sqrt{C_0(m)}}, \qquad M_{min} \leq m \leq M_{max} \tag{3.20}$$

onde :

$$C(m) = \sum_{n=0}^{N-1} r(n)r(n-m), \tag{3.21}$$

$$C_0(m) = \sum_{n=0}^{N-1} r^2(n-m), \tag{3.22}$$

e M_{min} e M_{max} especificam os valores mínimo e máximo para o período de pitch M e $r(n)$ poder ser um bloco de comprimento N do sinal de voz original($r(n) = s(n)$) ou resíduo LPC(vide eq. 3.2), sendo que a utilização do resíduo para $r(n)$ geralmente resulta numa melhor razão sinal/ruído. O ganho de pitch é, então, calculado a partir da seguinte expressão :

$$\gamma = \frac{\sum_{n=0}^{N-1} s(n)s(n-M)}{\sum_{n=0}^{N-1} s^2(n-M)} \tag{3.23}$$

A inclusão de um filtro de síntese de longo-prazo de primeira ordem conforme a expressão (3.17) com atualização dos seus parâmetros sendo efetuada a cada 5ms e baseada no sinal de resíduo LPC, resulta em um ganho de predição de aproximadamente 6.6 e 5.6 dB para vozes femininas e masculinas, respectivamente [28].

O algoritmo de determinação do período de pitch e ganho em malha fechada foi proposto por Singhal e Atal em [12]. Seja $d_1(n)$ um sinal de referência calculado subtraindo-se do sinal de voz original com ponderação, $s_w(n)$, a resposta da combinação dos filtros de síntese LPC e longo-prazo à entrada nula, $d'_o(n)$, após ter sido excitado no bloco anterior pela seqüência ótima $p(n)$, mas com o filtro de longo-prazo com as condições iniciais nulas, conforme mostrado na figura 3.6. Então, o período

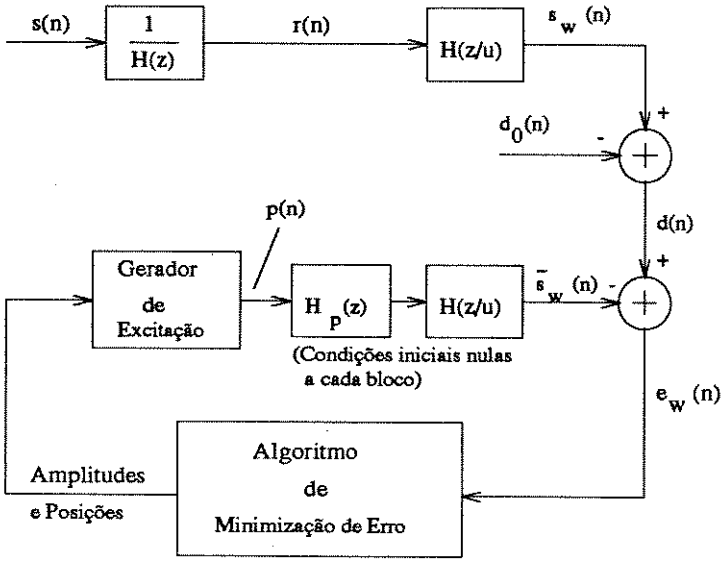

Figura 3.5: Diagrama em blocos do procedimento de análise MP com um filtro de síntese de longo-prazo em cascata com o filtro de síntese LPC.

e ganho de pitch são calculados a partir da minimização do erro quadrático médio entre o sinal de referência, $d_1(n)$, e a resposta à entrada nula, $\hat{s}_d(n)$, da combinação dos filtros de síntese LPC e longo-prazo mas, agora, com o filtro de síntese LPC com as condições iniciais nulas, conforme mostrado na figura 3.7 :

$$\langle e^2 \rangle = \sum_{n=0}^{N-1} [d_1(n) - \hat{s}_d(n)]^2 \tag{3.24}$$

Sendo $h_w(n)$ a resposta impulsiva do filtro de síntese LPC com ponderação, $H(z/\mu)$, tem-se que :

$$\hat{s}_d(n) = \gamma \sum_{k=0}^{N-1} r_p(k-M) h_w(n-k) \tag{3.25}$$

Substituindo-se (3.25) em (3.24) e derivando-se em relação a γ obtém-se :

$$\gamma = \frac{\sum_{k=0}^{N-1} r_p(k-M)\beta_k}{\sum_{k=0}^{N-1} r_p(k-M) \sum_{j=0}^{N-1} r_p(j-M)\Phi(k,j)} \tag{3.26}$$

onde $\Phi(k,j)$ é a função de autocorrelação do sinal $h_w(n)$ e β_k é a função de correlação-cruzada entre os sinais $d_1(n)$ e $h_w(n)$, conforme as equa,ões (3.10) e (3.11) (substituindo-se $d(n)$ por $d_1(n)$), respectivamente. Reescrevendo a equação (3.24)

levando em conta o ganho otimizado γ, resulta que o período de pitch M é igual ao valor de m entre M_{min} e M_{max} que maximiza :

$$\frac{[\sum_{k=0}^{N-1} r_p(k-m)\beta_k]^2}{\sum_{k=0}^{N-1} r_p(k-m) \sum_{j=0}^{N-1} r_p(j-m)\Phi(k,j)} \tag{3.27}$$

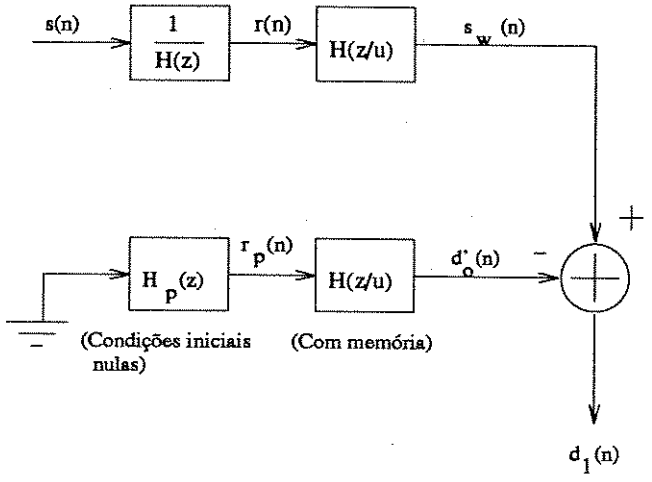

Figura 3.6: Determinação do sinal de referência para o cálculo do período de pitch em malha fechada

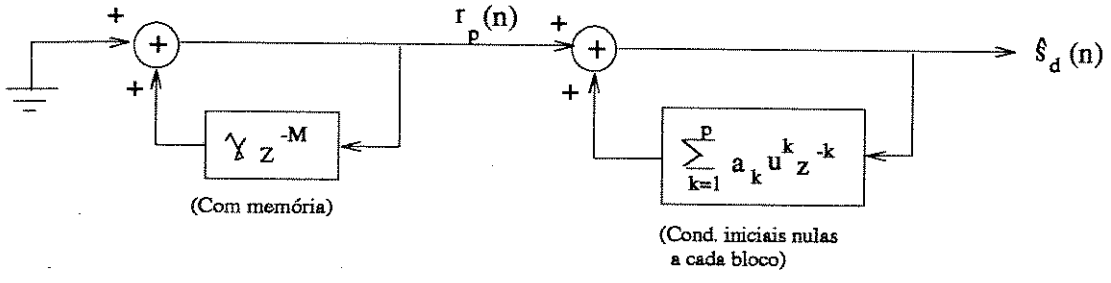

Figura 3.7: Modelo de análise para determinação do período de pitch em malha fechada

O procedimento de determinação de período de pitch em malha fechada é sub-ótimo, pois a minimização de erro é feita considerando-se uma entrada nula. Além disto, a resposta à entrada nula do filtro de síntese de longo-prazo, $r_p(n)$, está bem definida somente para $n \leq M$. Quando $n > M$, procura-se uma aproximação de

$r_p(n)$, por exemplo, repetindo-se o mesmo sinal do bloco anterior. Na seção 4.4.4, é proposto um novo método em malha fechada quase-ótimo, onde a minimização do erro é feita considerando-se o sinal de excitação realmente presente na entrada do filtro de síntese de longo-prazo.

Filtros de síntese de longo-prazo de ordem mais alta, fornecem ganhos de predição maiores, mas também é necessário um maior número de bits para a codificação dos coeficientes adicionais. Uma maneira de aumentar o ganho de predição sem aumentar a taxa de bits, consiste em utilizar filtros de síntese de longo-prazo de 1a ordem com atrasos fracionários [28, 29, 30].

3.2.3 Codificação da Posição dos Pulsos

Uma parte que merece bastante atenção num codificador baseado no modelo de excitação MP refere-se à representação dos N_p pulsos em um bloco de N amostras, pois existem no total $\begin{pmatrix} N \\ N_p \end{pmatrix}$ possíveis padrões.

Um método de codificação dos pulsos baseado em um esquema combinacional pode ser descrito da seguinte maneira [10, 20] :

1. Inicialização do algoritmo ($I = 0$);
2. Forma-se um vetor binário de comprimento N, cujos elementos são 0 ou 1, conforme a ausência ou presença de um pulso, respectivamente;
3. Percorre-se o vetor binário da posição N para a posição 1, à procura da posição n do próximo 1 no vetor e computa-se o índice I como sendo:

$$I = I + \begin{pmatrix} n-1 \\ m \end{pmatrix}, \tag{3.28}$$

   onde m é o número de 1's no vetor que ainda não foram encontrados acrescido de 1.
4. Se os N_p elementos foram encontrados, então termina-se o algoritmo, e o valor de I representa o código do padrão associado com os N_p pulsos. Em caso contrário, volta-se ao passo 3.

O problema inverso, que é determinar a posição dos pulsos a partir do código padrão I, é facilmente resolvido comparando-se o valor do código I com os mesmos

valores combinacionais $\binom{n-1}{m}$.

O procedimento para a determinação do código I do padrão associado com os N_p pulsos, bem como o procedimento inverso (i.e., a determinação da posição dos pulsos a partir do código I), utilizam valores numéricos representados pelo inteiro maior ou igual a $\log_2\binom{N}{N_p}$ bits.

Para exemplificar este método, sejam $N = 8$, $N_p = 3$ e o padrão de posições dos pulsos mostrado na Fig. 3.8, isto é, com pulsos nas posições 4, 6 e 7. O seguinte valor de índice I é obtido :

$$I = \binom{6}{3} + \binom{5}{2} + \binom{3}{1} = 33 \tag{3.29}$$

onde os valores combinacionais das posições 7, 6 e 4 valem, respectivamente, 20, 10 e 3. No decodificador determinam-se as posições dos pulsos a partir do valor do código I seguindo-se o seguinte raciocínio :

$\binom{7}{3} > 33 \Rightarrow$ não tem pulso na posição 8.

$\binom{6}{3} \leq 33 \Rightarrow$ tem pulso na posição 7.

$\binom{5}{2} \leq 13 \Rightarrow$ tem pulso na posição 6.

$\binom{4}{1} > 3 \Rightarrow$ não tem pulso na posição 5.

$\binom{3}{1} \leq 3 \Rightarrow$ tem pulso na posição 4.

Neste ponto, tem-se que $m = 0$ e, portanto, não tem pulso no restante das posições.

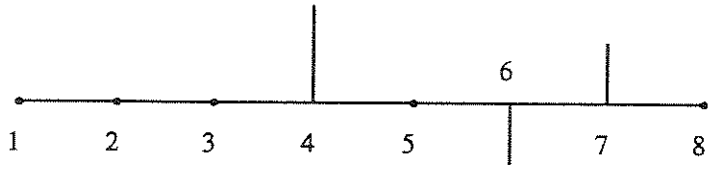

Figura 3.8: Exemplo de padrão de posições dos pulsos em uma excitação MP

3.3 MODELO DE EXCITAÇÃO POR DICIONÁRIO DE CÓDIGOS

No modelo de excitação por dicionário de códigos, utilizado em codificadores CELP [9], um número finito de seqüências ou vetores de sinal de excitação são contidos em um dicionário de códigos. Em outras palavras, no codificador CELP, o sinal de excitação é quantizado vetorialmente e, desta maneira, obtém-se uma grande redução na sua taxa de bits. O sinal de voz é sintetizado passando-se cada uma das seqüências de excitação do dicionário de códigos através de um filtro de síntese de longo-prazo, $H_p(z)$, em cascata com um filtro de síntese LPC com ponderação, $H(z/\mu)$, como mostrado na figura 3.9.

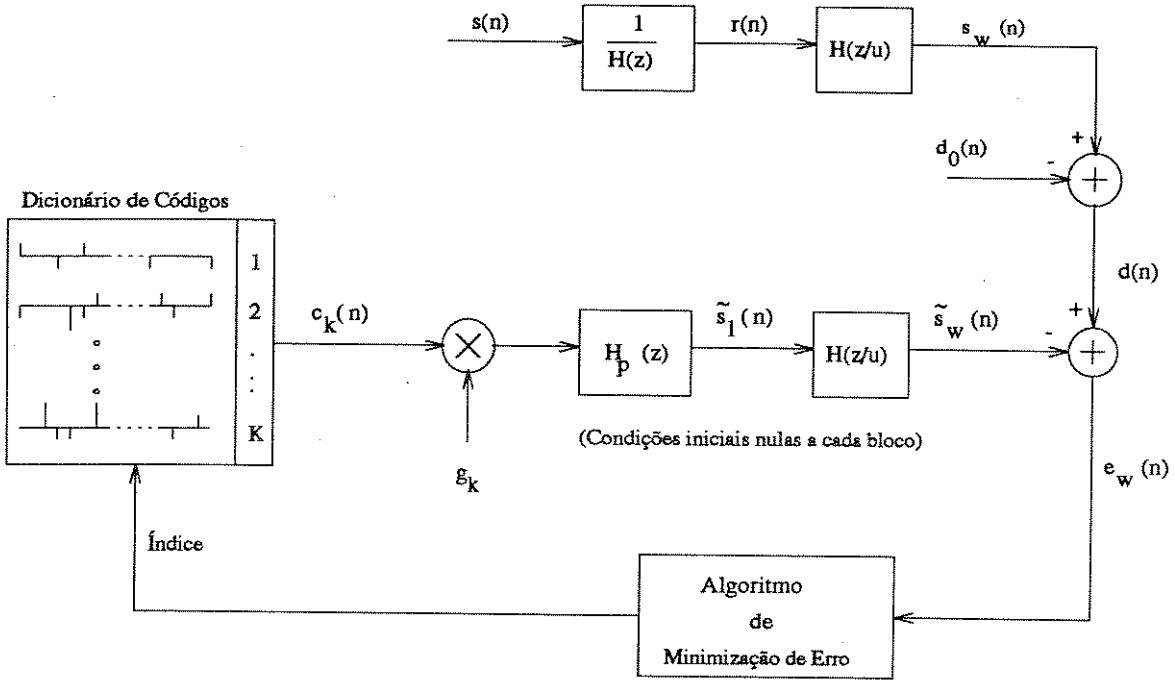

Figura 3.9: Diagrama em blocos do procedimento de análise CELP

Entretanto, para cada bloco do sinal de voz, uma seqüência de excitação ótima deve ser selecionada do dicionário de códigos através de um procedimento de análise por síntese utilizando-se um critério de erro quadrático médio ponderado. Assim, no modelo de excitação por dicionário de códigos tem-se basicamente dois problemas :

- Escolha da seqüência de excitação ótima;
- Projeto do dicionário de códigos das seqüências de excitação.

Atualmente existem várias estruturas de codificadores CELP, também chamado de VXC(Vector Excited Coding) [24] ou SELP(Stochastically Excited Linear Prediction) [25], constituindo uma classe de codificadores de mesmo nome. Dentro desta classe de codificadores CELP, além do codificador CELP [9] propriamente dito, mostrado na fig. 3.9, destacam-se o VSELP(Vector Sum Excited Linear Prediction) [26] e o LD-CELP(Low-Delay CELP) [27].

3.3.1 Escolha da Seqüência de Excitação Ótima

Seja um dicionário de códigos de *comprimento* K e de *dimensão* N (isto é, existem K diferentes seqüências de N amostras cada uma). No codificador CELP, a seqüência de excitação ótima é selecionada filtrando-se cada seqüência, $c_k(n), k = 1, ..., K, n = 0, ..., N-1$, multiplicada por um fator de ganho g_k, através do filtro de síntese de longo-prazo, $H_p(z)$, em cascata com o filtro de síntese LPC, $H(z/\mu)$.

O fator de ganho g_k pode ser determinado para cada seqüência $c_k(n)$ através de uma minimização do erro quadrático ponderado dado pela expressão (3.7), onde agora tem-se que :

$$\tilde{s}_w(n) = \sum_{i=0}^{N-1} g_k c_k(i) f(n-i), \tag{3.30}$$

onde $f(n)$ é a resposta impulsiva da combinação dos filtros de síntese de longo-prazo e LPC com ponderação :

$$f(n) = h_p(n) * h_w(n) \tag{3.31}$$

Assim, o sinal de referência $d(n)$ é calculado considerando-se a memória do filtro $F(z)$ resultante da combinação de $H_p(z)$ e $H(z/\mu)$:

$$F(z) = H_p(z) H(z/\mu) \tag{3.32}$$

A sequência de excitação ótima é, então, selecionada comparando-se o sinal de voz sintetizado pelo filtro $F(z)$ com o sinal de referência $d(n)$, utilizando-se como medida o erro quadrático médio ponderado. Substituindo-se (3.30) em (3.7), obtém-se :

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} [d(n) - \sum_{i=0}^{N-1} g_k c_k(i) f(n-i)]^2, \tag{3.33}$$

Diferenciando-se esta expressão em relação à g_k e igualando a zero resulta :

$$g_k = \frac{\sum_{n=0}^{N-1} d(n) \sum_{i=1}^{N} c_k(i) f(n-i)}{\sum_{n=0}^{N-1} [\sum_{i=0}^{N-1} c_k(i) f(n-i)]^2} \tag{3.34}$$

Usando-se este valor de g_k em (3.33), obtém-se a seguinte expressão para o erro quadrático médio ponderado :

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} d^2(n) - \frac{[\sum_{n=0}^{N-1} d(n) \sum_{i=0}^{N-1} c_k(i) f(n-i)]^2}{\sum_{n=0}^{N-1} [\sum_{i=0}^{N-1} c_k(i) f(n-i)]^2} \tag{3.35}$$

Assim, a seqüência de excitação ótima é aquela que minimiza o erro quadrático médio dado pela equação (3.35) ou que maximiza a expressão :

$$\frac{[\sum_{n=0}^{N-1} d(n) \sum_{i=0}^{N-1} c_k(i) f(n-i)]^2}{\sum_{n=0}^{N-1} [\sum_{i=0}^{N-1} c_k(i) f(n-i)]^2} \tag{3.36}$$

Uma outra alternativa de busca da seqüência de excitação ótima consiste em encontrar a seqüência $c_k(i)$, $0 = 1, ..., N-1$, que maximiza a seguinte expressão :

$$2\hat{g}_k \sum_{i=0}^{N-1} c_k(i) \sum_{n=0}^{N-1} d(n) f(n-i) - \hat{g}_k^2 \sum_{n=0}^{N-1} [\sum_{i=0}^{N-1} c_k(i) f(n-i)]^2 \tag{3.37}$$

onde $\hat{g}_k$ é o ganho g_k quantizado.

Num caso prático, a utilização da expressão (3.37) costuma fornecer um resultado ligeiramente superior, pois leva em consideração os efeitos da quantização do ganho do vetor de excitação.

3.3.2 Projeto do Dicionário de Códigos de Excitação.

Existem vários métodos propostos de projeto do dicionário de códigos de excitação. Um que tem apresentado bons resultados consiste em iniciar o processo

com um dicionário inicial gerado a partir de seqüências gaussianas e, em seguida, otimizá-lo utilizando uma técnica de agrupamento iterativo em conjunção com uma seqüência de treinamento de sinal de voz [21]. Como método de agrupamento pode ser utilizado o de Lloyd (algoritmo K-means) [22], onde o sinal de treinamento é agrupado em células não superpostas e calculados os centróides correspondentes que minimizam uma distorção média perceptualmente significativa. Estes centróides são, então, armazenados como vetores código. No caso do projeto do dicionário de códigos de excitação em codificadores CELP, o método de cálculo do centróide é definido de modo que o erro quadrático médio ponderado entre o sinal de voz original e sintetizado seja minimizado. Este método de cálculo do centróide pode ser assim descrito : inicia-se com uma célula de M elementos, cada elemento consistindo de uma seqüência de referência, $d_k(n)$, uma seqüência da resposta impulsiva truncada da combinação dos filtros de síntese LPC e longo-prazo, $f_k(n)$, e de um fator de ganho da seqüência de excitação, g_k, onde $1 \leq k \leq M$ e $0 \leq n < N$. O objetivo é encontrar a seqüência $c(n)$ que minimiza o erro quadrático médio entre $d_k(n)$ e $\sum_{i=0}^{N-1} g_k c(i) f_k(n-i)$. O erro quadrático médio ponderado E, para a célula, é definido como :

$$E = \frac{1}{M} \sum_{k=1}^{M} \sum_{n=0}^{N-1} [d_k(n) - \sum_{i=1}^{N} g_k c(i) f_k(n-i)]^2 \tag{3.38}$$

Tomando-se a derivada de E em relação a $c(j)$, $0 \leq j \leq N-1$, e igualando-se a zero, obtém-se a seguinte expressão para o cálculo do centróide :

$$\sum_{i=0}^{N-1} c(i) \sum_{k=1}^{M} g_k^2 \sum_{n=0}^{N-1} f_k(n-i) f_k(n-j) = \sum_{k=1}^{M} g_k \sum_{n=0}^{N-1} d_k(n) f_k(n-j), \qquad 0 \leq j \leq N-1 \tag{3.39}$$

Definindo-se :

$$\phi(j,i) = \sum_{k=1}^{M} g_k^2 \sum_{n=0}^{N-1} f_k(n-i) f_k(n-j) \tag{3.40}$$

$$\psi(j,0) = \sum_{k=1}^{M} g_k \sum_{n=0}^{N-1} d_k(n) f_k(n-j), \tag{3.41}$$

obtém-se a seguinte equação :

$$\sum_{i=0}^{N-1} c(i) \phi(j,i) = \psi(j,0) \tag{3.42}$$

Na forma matricial obtém-se :

$$\underline{\phi} . \underline{c}^t = \underline{\psi} \tag{3.43}$$

onde $\underline{c} = [c(0) \quad c(1)...c(N-1)]$. Esta equação pode ser resolvida usando uma decomposição de Cholesky da matriz $\underline{\phi}$.

Um problema que surge na determinação do centróide $c(n)$ utilizando-se este método, é que as seqüências de treinamento $d_k(n)$, derivadas a partir do sinal de voz original subtraindo-se o estado do filtro devido aos blocos anteriores, dependem do dicionário de códigos atual. Isto faz com que não se possa garantir que o erro E associado com os dicionários de códigos intermediários decresça monotonicamente. Na prática, entretanto, a redução do erro E é conseguida nas primeiras iterações, de modo que o processo pode ser interrompido tão logo um determinado critério de convergência seja satisfeito. Quando este procedimento é aplicado, o método de projeto do dicionário de códigos é denominado de *malha fechada.*

A convergência do erro E pode ser garantida se o sinal de treinamento $d_k(n)$ é calculado uma vez na primeira iteração utilizando-se o dicionário de códigos inicial e depois mantido constante nas iterações subsequentes. Neste caso, o projeto do dicionário de códigos é denominado de *malha aberta.*

Uma combinação de projeto em malha fechada e em malha aberta também é possível : uma série de iterações em malha aberta são executadas seguidas de uma iteração em malha fechada quando a seqüência de referência $d_k(n)$ é atualizada. Este processo de várias iterações em malha aberta seguido de uma iteração em malha fechada para a atualização de seqüência de referência, $d_k(n)$, é continuado até que um determinado critério de convergência seja satisfeito.

Bibliografia

[1] V. Iyengar e P. Kabal,"A Low Delay 16 kbit/s Speech Coder", IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1988, pág.243-246.

[2] J.D. Gibson, G.B. Haschke, "Backward Adaptive Tree Coding of Speech at 16 kbps", IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1988, pág. 251-254.

[3] F.M. Ferreira, "Codec ADPCM a 16 kbit/s com Quantização de Árvore", Dissertação de Tese de Mestrado, DECOM/FEE/UNICAMP, Outubro de 1990.

[4] D.Y. Wong, B.-H. Juang e A.H. Gray, "An 800 bit/s Vector Quantization LPC Vocoder", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-30, nº 5, Outubro de 1982, pág. 770-780.

[5] D.Y. Wong, B.-H. Juang e D.Y. Cheng, "Very Low Data Rate Speech Compression with LPC Vector and Matrix Quantization", IEEE Int. Conf. Acoust., Speech, Signal Process., 1983, Boston, pág. 65-68.

[6] S. Roucos, R.M. Schwartz, J. Makhoul, "A segment Vocoder at 150 B/s", IEEE Int. Conf. Acoust., Speech, Signal Process., 1983, Boston, pág. 61-64.

[7] J.A. Martins,"Vocoder LPC com Quantização Vetorial", Dissertação de Tese de Mestrado, DECOM/FEE/UNICAMP, Abril de 1991.

[8] B.S. Atal e J.R. Remde, "A new model of LPC excitation for producing natural--sounding speech at low bit rates", IEEE Int. Conf. Acoust., Speech, Signal Process., 1982, Paris, pág. 614-617.

[9] M.R. Schroeder e B.S. Atal, "Code-excited linear prediction (CELP) : High-quality speech at very low bit rates", IEEE Int. Conf. Acoust., Speech, Signal Process., Março de 1985, pág. 937-940.

[10] M. Berouti, H. Garten, P. Kabal e P. Mermelstein, "Efficient computation and encoding of the multipulse excitation for LPC", IEEE Int. Conf. Acoust., Speech, Signal Process., 1984, San Diego, CA, pág. 10.1.1-10.1.4.

[11] S. Ono, T. Araseki e K. Ozawa, "Improved pulse search algorithm for multi-pulse excited speech coder",IEEE Global Telecommun. Conf., 1984, pág. 287--291.

[12] S. Singhal e B.S. Atal, "Improving performance of multi-pulse coders at low bit rates",IEEE Int. Conf. Acoust., Speech, Signal Process., 1984, San Diego, CA, pág. 1.3.1-1.3.4.

[13] P. Kroon e E.F. Deprettere, "Experimental evaluation of different approaches to the multi-pulse coder", IEEE Int. Conf. Acoust., Speech, Signal Process., 1984, San Diego, CA, pág. 10.4.1-10.4.4.

[14] J.P. Lefevre e O. Passien, "Efficient algorithms for obtaining multipulse excitation for LPC coders", IEEE Int. Conf. Acoust., Speech, Signal Process., 1985, Tampa, FL, pág. 957-960.

[15] Y. Wabe, S. Tanaka, K. Ozawa e T. Araseki, "A multi-pulse LPC codec using digital signal processors",IEEE Int. Conf. Acoust., Speech, Signal Process., 1985, Tampa, FL, pág. 1429-1432.

[16] S. Singhal, "Reducing computation in optimal amplitude multipulse coders", IEEE Int. Conf. Acoust., Speech, Signal Process., Maio de 1986, Tokyo, Japão, pág. 2363-2366.

[17] K. Ozawa, S. Ono e T. Araseki, "A study on pulse search algorithm for multipulse excited speech coder realization", IEEE Journal on Selected Areas in Communications, vol. SAC-4, nº 1, Janeiro de 1986, pág. 133-141.

[18] S. Singhal e B.S. Atal, "Amplitude optimization and pitch prediction in multipulse coders", IEEE Transactions on Acoustics, Speech, and Signal Process., vol. 37, nº 3, Março de 1989, pág. 317-327.

[19] L.R. Rabiner e R.W. Schafer,"Digital Processing of Speech Signals", Prentice--Hall, Englewood Cliffs, N.Y., (1978), pág. 407-411.

[20] R. Montagna e M. Omologo,"Some Results on Multipulse Linear Predictive Coding", Proc. IEEE Global Telecommunications Conference, Houston, 1986, pág. 802-806.

[21] G. Davidson, M. Yong e A. Gersho,"Real Time Vector Excitation Coding of Speech at 4800 BPS", IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1987, Dallas, pág. 2189-2192.

[22] Y. Linde, A. Buzo, R. M. Gray,"An Algorithm for Vector Quantizer Design", IEEE Trans. on Comm., vol. COM-28, no1, Janeiro de 1990, pág. 84-95.

[23] M.R. Schoeder e B.S. Atal,"Adaptive Predictive Coding of Speech Signals", Bell Syst. Tech. J., vol. 49, Outubro de 1970, pág. 1973-1986.

[24] G. Davidson e A. Gersho,"Complexity Reduction Methods for Vector Excitation Coding", IEEE Int. Conf. on Acoust., Speech and Signal Process., Tokyo, Japão, pág. 3055-3058.

[25] W.B. Kleijn, D.J. Krasinski e R.H. Ketchun,"Improved Speech Quality and Efficient Vector Quantization in SELP", IEEE Int. Conf. on Acoust., Speech and Signal Process., Abril de 1988, pág. 155-158.

[26] I.A. Gerson e M.A. Jasiuk,"Vector Sum Excited Linear Prediction (VSELP)", Advances in Speech Coding, Boston, MA: Kluwer, 1991, pág. 69-79.

[27] J.-H. Chen, R.V. Cox, Y.-C. Lin, N. Jayant e M.J. Melchner, "A Low-Delay CELP Coder for the CCITT 16 kbit/s Speech Coding Standard", IEEE Journal on Selected Areas in Communications, vol. 10, no 5, junho de 1992, pág. 830-849.

[28] P. Kroon e B.S. Atal,"On Improving the Performance of Pitch Predictors in Speech Coding Systemas", Advances in Speech Coding, Boston, MA: Kluwer, 1991, pág. 321-327.

[29] P. Kroon e B.S. Atal,"Pitch Predictors with High Temporal Resolution", IEEE Int. Conf. Acoust., Speech, Signal Process., 1990, pág. 661-668.

[30] J. Marques, J. Tribolet, I. Trancoso e L. Almeida, "Pitch Prediction with Fractional Delays in CELP Coding", EUROSPEECH, Paris, Setembro de 1989, pág. 509-512.

Capítulo 4

QUANTIZAÇÃO VETORIAL DOS COEFICIENTES LPC

4.1 INTRODUÇÃO

Num codec de voz do tipo LPC, os benefícios provenientes de uma quantização eficiente dos coeficientes LPC são basicamente dois :

- A uma dada taxa de bits, a economia de bits feita para a transmissão dos coeficientes LPC pode ser usada para melhorar a quantização do sinal de excitação e assim melhorar o desempenho global do codec.
- Baixar a taxa de bits global do codec, mas mantendo-se o desempenho.

Assim, tem havido um considerável interesse neste campo e vários tipos de representação dos coeficientes LPC com propriedades mais adequadas para quantização foram propostos. Em [1] é mostrado que os coeficientes de reflexão constituem um conjunto de representação dos coeficientes LPC bastante vantajoso para quantização, pois, além de preservar a estabilidade do filtro, possuem uma propriedade natural de ordenação (isto é, se dois coeficientes forem trocados entre si, o filtro passa a não ter mais o mesmo comportamento), que pode ser usada no projeto de algoritmos de quantização mais eficientes. Além disto, demonstra-se que uma transformação não-linear dos coeficientes de reflexão, denominada Razão Log Área (Log Area Ratio-LAR), possui propriedades de quantização aproximadamente ótimas. Gray e Markel [2] propuseram a quantização do seno inverso dos coeficientes de reflexão. Um conjunto de parâmetros, conhecido como "Line Spectrum Pair" (LSP)

[3, 4], foi inicialmente proposto por Itakura como uma alternativa de representação espectral LPC. Atualmente, o LSP é tido como uma das representações dos coeficientes LPC mais eficientes para quantização e tem sido utilizado em diversos estudos [5, 6, 7]. Adicionalmente, vários algoritmos de quantização escalar e vetorial aproveitando as propriedades mais vantajosas destas transformações tem sido propostos [8, 9, 10, 11, 12].

Os algoritmos de quantização escalar utilizam em torno de 40 bits por bloco para quantizar 10 coeficientes LPC (normalmente transformados em LAR ou LSP) mantendo um nível de distorção aceitável. Esta taxa de bits pode ser eficientemente reduzida aplicando-se uma quantização vetorial ao conjunto dos coeficientes LPC [13, 14].

O dicionário de códigos dos coeficientes LPC é normalmente gerado a partir de um sinal de treinamento. Uma vez gerado o dicionário de códigos, a palavra código ótima pode ser selecionada fazendo-se, por exemplo, uma busca exaustiva de modo a minimizar uma determinada medida de distorção entre os coeficientes LPC (sob uma de suas formas de representação) originais e quantizados.

A escolha da medida de distorção é crucial para a quantização vetorial dos coeficientes LPC em termos de desempenho final e complexidade. Uma das medidas de distorção mais simples é o *Erro Quadrático Médio*. Entretanto, é mostrado em [15] que os melhores resultados são obtidos utilizando-se uma medida de distorção muito mais complexa, conhecida como *Itakura Saito Modificada*[16].

Um outro ponto importante na quantização vetorial dos coeficientes LPC é o algoritmo de busca do vetor ótimo dentro do dicionário de códigos. Normalmente esta busca é realizada em malha aberta empregando-se uma determinada medida de distorção. Este algoritmo não é muito eficiente quando aplicado a um codificador de voz LPC no sentido de minimizar a distorção entre o sinal de voz sintetizado e original. Para isto, estruturas em malha fechada devem ser empregadas. Neste sentido, em [19] é proposta uma otimização dos coeficientes LPC uma vez conhecido o sinal de excitação e, em [20], demonstra-se que um procedimento seqüencial de otimização conjunta dos coeficientes LPC quantizados escalarmente, parâmetros de excitação de curto-prazo (índice e ganho do vetor ótimo de excitação) e longo-prazo (período e ganho de pitch), melhora significativamente o desempenho de um codificador CELP.

No presente trabalho, é proposto um algoritmo de quantização vetorial conjunta em malha fechada dos coeficientes LPC e parâmetros de excitação de curto-prazo e longo-prazo, denominado *AMF (Análise em Malha Fechada)*. Adicionalmente, é proposto um algoritmo de quantização vetorial dos coeficientes LPC denominado *AEP (Análise do Erro de Predição)*. No algoritmo AEP, é feita uma análise do erro de predição entre o sinal de voz estimado e original. Portanto, trata-se de um algoritmo onde a busca do vetor ótimo é realizada em malha aberta e que não necessariamente minimiza a distorção entre o sinal de voz sintetizado e original. Contudo, como mostrado respectivamente nas seções 4.3.2 e 4.5.4, este algoritmo oferece uma complexidade menor e um desempenho superior em relação ao algoritmo tradicional de quantização vetorial utilizando a medida de distorção de Itakura Saito Modificada.

4.2 ALGORITMO TRADICIONAL DE QUANTIZAÇÃO VETORIAL E MEDIDA DE ITAKURA SAITO MODIFICADA

A medida de distorção de Itakura Saito Modificada aplica-se diretamente sobre os coeficientes LPC conforme o algoritmo tradicional de quantização vetorial mostrado na figura 4.1. Neste algoritmo, a busca do vetor ótimo é feita em mallha aberta, pois a medida de distorção resulta de uma comparação direta entre os diversos vetores de coeficientes LPC do dicionário de códigos e o vetor LPC original correspondente ao bloco do sinal de voz em análise. Sejam $\mathbf{a} = [a_1 \quad a_2 \ \ldots \ a_p]^t$ um vetor de coeficientes LPC original (sem quantização) e $\hat{\mathbf{a}}_l = [\hat{a}_{1,l} \quad \hat{a}_{2,l} \ \ldots \ \hat{a}_{p,l}]^t$ o l-ésimo vetor de coeficientes LPC quantizado de um dicionário de códigos de tamanho L. Então, a medida de Itakura Saito Modificada entre estes dois vetores LPC é definida como [16] :

$$d_{I,l}(\mathbf{a}, \hat{\mathbf{a}}_l) = (\mathbf{a} - \hat{\mathbf{a}}_l)^t R_{\mathbf{x}} (\mathbf{a} - \hat{\mathbf{a}}_l) \tag{4.1}$$

$$R_{\mathbf{x}} = \{r(i-k)/r(0),\ 0 \leq i,\ k \leq p-1\}, \tag{4.2}$$

onde : $r(i)$, $0 \leq i \leq p-1$ é a função de autocorrelação calculada sobre um bloco de N amostras do sinal de entrada $\mathbf{x}$, o mesmo bloco empregado no cálculo do vetor $\mathbf{a}$

Assim, uma vez calculada a matriz de autocorrelação $R_{\mathbf{x}}$ e o vetor LPC original $\mathbf{a}$, a

complexidade de um ciclo de busca do vetor ótimo, usando-se a medida de distorção de Itakura Saito Modificada, é de aproximadamente $p^2 + p$ multiplicações/adições.

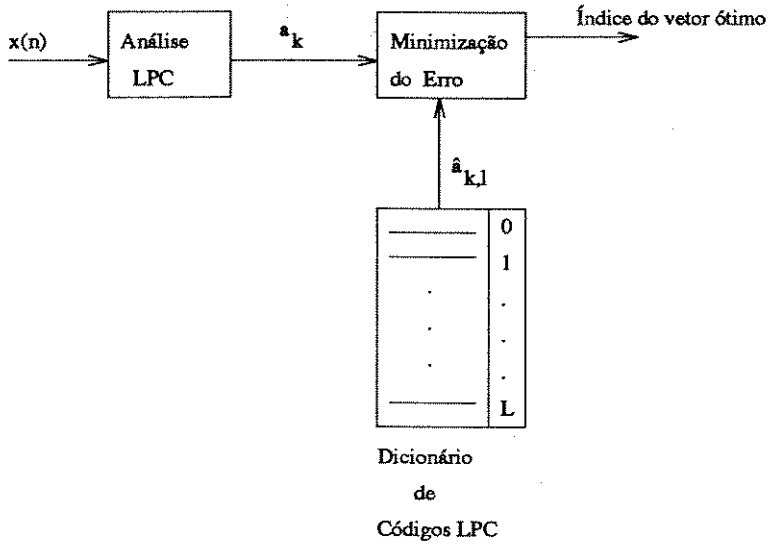

Figura 4.1: Agoritmo tradicional de quantização vetorial dos coeficientes LPC.

4.3 ALGORITMO POR ANÁLISE DO ERRO DE PREDIÇÃO

4.3.1 Princípio Básico

O esquema da quantização vetorial dos parâmetros LPC pelo algoritmo por análise do erro de predição (AEP) proposto neste trabalho é mostrado na figura 4.2. Neste esquema, cada um dos vetores $\mathbf{a}_l$, $l = 1, ..., L$, do dicionário de códigos contém um conjunto de p parâmetros LPC, isto é, representa um determinado preditor de ordem p. Para um bloco de sinal de voz de N amostras, o melhor preditor é selecionado de modo a minimizar uma medida de distorção entre o sinal de voz original e o sinal de voz estimado, $D = d(X_i - \tilde{X}_{i,l})$, onde :

$X_i = \{x_i(0), x_i(1), ..., x_i(N-1)\}$, é o bloco i do sinal de voz original;

$\tilde{X}_{i,l} = \{\tilde{x}_{i,l}(0), \tilde{x}_{i,l}(1), ..., \tilde{x}_{i,l}(N-1)\}$, é o bloco i do sinal de voz estimado utilizando-se o preditor l

Nestas condições, se C_p é o dicionário de códigos dos L preditores de ordem p,

e $\mathbf{a}_J$ $(J \leq L)$ é o vetor selecionado para o bloco i do sinal de voz de entrada, então tem-se que $\mathbf{a}_J$ é o vetor tal que :

$$d(X_i - \tilde{X}_{i,J}) \leq d(X_i - \tilde{X}_{i,l}), \qquad \forall\, l \neq J \quad e \quad l \leq L \tag{4.3}$$

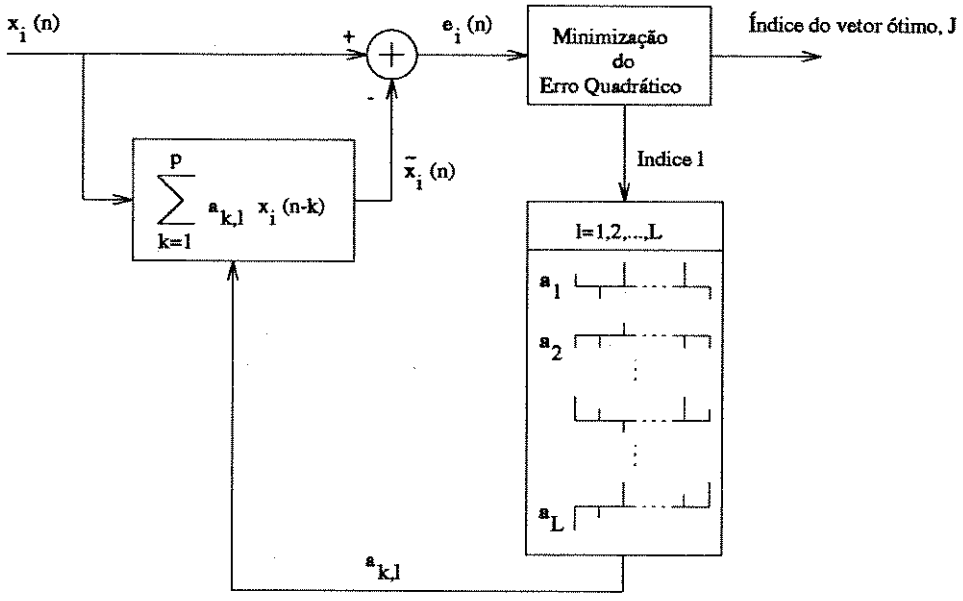

Figura 4.2: Quantização vetorial dos coeficientes LPC por Análise do Erro de Predição

4.3.2 Busca do Vetor Ótimo

A busca do vetor ótimo depende do tipo de distorção a ser utilizado entre o sinal de voz original e estimado. Em princípio, pode ser utilizado qualquer tipo de medida de distorção, seja no domínio do tempo ou no domínio da freqüência. Além disto, um mesmo tipo de distorção pode ser utilizado para diversos tipos de representação dos coeficientes LPC.

A medida de distorção a ser utilizada neste trabalho é o erro quadrático médio calculado como :

$$D_{i,l} = \sum_n [x_i(n) - \tilde{x}_{i,l}(n)]^2, \qquad l = 1, ..., L \tag{4.4}$$

onde : $x_i(n)$ são as amostras do bloco i do sinal de voz original;
$\tilde{x}_{i,l}(n)$ são as amostras do bloco i do sinal de voz estimado utilizando-se o preditor l.

A utilização do erro quadrático médio tem a vantagem prática de resultar uma equação linear que simplifica o algoritmo de busca do vetor ótimo. A partir da equação (4.4), pode-se fazer dois casos de análise, dependendo dos limites da somatória : covariância e autocorrelação.

Método da Covariância

Quando a somatória da equação (4.4) se estender de 0 até $N-1$, tem-se o tipo de análise de covariância, onde nenhuma restrição é feita sobre o sinal $x_i(n)$ fora do bloco de análise. Neste caso, tem-se que :

$$D_{i,l} = \sum_{n=0}^{N-1} [x_i(n) - \sum_{k=1}^{p} a_{k,l} x_i(n-k)]^2, \tag{4.5}$$

onde $a_{k,l}$ são os coeficientes LPC do preditor l. Desenvolvendo-se a equação (4.5) obtém-se :

$$D_{i,l} = \phi(0,0) - 2\sum_{k=1}^{p} a_{k,l}\phi(0,k) + \sum_{k=1}^{p} a_{k,l} \sum_{j=1}^{p} a_{j,l}\phi(j,k), \tag{4.6}$$

onde :

$$\phi(j,k) = \sum_{n=0}^{N-1} x_i(n-j)x_i(n-k) \tag{4.7}$$

Portanto, como $D_{i,l}$ é positivo, os coeficientes $a_{k,M}$ que formam o vetor ótimo, $\mathbf{a}_M$, são aqueles que maximizam a seguinte equação :

$$\Delta = 2\sum_{k=1}^{p} a_{k,l}\phi(0,k) - \sum_{k=1}^{p}\sum_{j=1}^{p} a_{k,l}a_{j,l}\phi(j,k) \tag{4.8}$$

Levando-se em conta que $\phi(j,k)$=$\phi(k,j)$, a equação (4.8) pode ser também escrita como :

$$\Delta = 2\sum_{j=1}^{p} a_{j,l}\phi(0,j) - \sum_{j=1}^{p} a_{j,l}^2\phi(j,j) - 2\sum_{k=1}^{p-1}\sum_{j=k+1}^{p} a_{k,l}a_{j,l}\phi(k,j) \tag{4.9}$$

A partir desta equação, tem-se que uma vez calculada a função $\phi(j,k)$, e lembrando-se que os valores de $a_{j,l}^2$ e $a_{k,l}a_{j,l}$ podem ser previamente calculados e devidamente armazenados, a complexidade deste algoritmo para um ciclo de busca é de cerca de $\frac{p^2+3p}{2}$ multiplicações/adições. Esta complexidade corresponde aproximadamente à metade do algoritmo tradicional de quantização vetorial usando-se a distorção de

Itakura-Saito Modificada. Além disto, como mostrado na seção 4.5.4, apresenta um desempenho superior em relação ao algoritmo tradicional, principalmente para baixo atraso de codificação.

Método de Autocorrelação

A análise de autocorrelação é obtida permitindo-se que o limite da somatória do erro se estenda de $-\infty$ a $+\infty$. Neste caso, o sinal $x(n)$ passa por um processo de janelamento que o torna igual a zero fora do bloco de análise. Sob estas condições demonstra-se que :

$$D_{i,l} = R(0,0) - 2\sum_{k=1}^{p} a_{k,l}R(k) + \sum_{k=1}^{p}\sum_{j=1}^{p} a_{k,l}a_{j,l}R(k-j), \tag{4.10}$$

onde:

$$R(k) = \sum_{n=k}^{N-1} x_i(n)x_i(n-k) \tag{4.11}$$

Assim, a complexidade do algoritmo AEP usando-se análise de autocorrelação é de aproximadamente $\frac{p^2+3p}{2}$ multiplicações/adições por ciclo de busca. Considerando-

-se que a complexidade do cálculo da função $R(k)$ é de apenas em torno de Np multiplicações e adições, enquanto que a função $\phi(k,j)$ exige cerca de $\frac{Np^2}{2}$ operações do mesmo tipo, usando-se análise de autocorrelação resulta numa diminuição de complexidade do algoritmo AEP. Entretanto, como mostrado na seção 4.5.4, resulta também numa degradação de desempenho, principalmente para baixo atraso de codificação, pois o efeito de erro maior no início e fim do bloco de análise torna-se mais significativo à medida que diminui o comprimento do bloco de análise.

4.4 ALGORITMO POR ANÁLISE DO SINAL DE VOZ SINTETIZADO

4.4.1 Princípios Básicos

A quantização vetorial dos coeficientes LPC pelo algoritmo proposto neste trabalho, denominado *quantização vetorial por análise do sinal de voz sintetizado ou em malha fechada* (AMF), aplica-se aos codificadores de voz do tipo CELP. Este algoritmo, ilustrado na figura 4.3, procura uma combinação ótima entre os vetores dos dicionários de códigos de excitação e dos coeficientes LPC de modo a minimizar

uma medida de distorção dada por $D = d(X_i - \hat{X}_{i,l})$, onde :

$X_i = \{x_i(0), x_i(1), ..., x_i(N-1)\}$, é o bloco i do sinal de voz original;

$\hat{X}_{i,l} = \{\hat{x}_{i,l}(0), \hat{x}_{i,l}(1), ..., \hat{x}_{i,l}(N-1)\}$, é o bloco i do sinal de voz sintetizado utilizando-se o preditor l

O princípio deste algoritmo é baseado no fato de que quando existe apenas um número limitado de seqüências de excitação e de coeficientes LPC, nem sempre uma melhor qualidade de sinal de voz é produzida pelo vetor ótimo de coeficientes LPC calculado isoladamente, e sim por aquele vetor LPC que melhor estiver "casado" com algum vetor do dicionário de códigos de excitação no sentido de minimizar a distorção entre o sinal de voz original e sintetizado.

No caso de se utilizar como medida de distorção o erro quadrático médio, tem-se para uma combinação da *k-ésima* sequência de excitação, $c_k(i)$, e *l-ésimo* filtro de síntese $f_l(n) = h_p^l(n) * h_w^l(n)$, a seguinte expressão :

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} d_l^2(n) - 2\hat{g}_{k,l} \sum_{n=0}^{N-1} d_l(n) \sum_{i=0}^{N-1} c_k(i) f_l(n-i) + \hat{g}_{k,l}^2 \sum_{n=0}^{N-1} [\sum_{i=0}^{N-1} c_k(i) f_l(n-i)]^2, \tag{4.12}$$

onde : $d_l(n)$ é o sinal de referência utilizando-se o preditor l (vide figura 4.3);
$\hat{g}_{k,l}$ é obtido quantizando-se o ganho $g_{k,l}$ dado por :

$$g_{k,l} = \frac{\sum_{n=0}^{N-1} d_l(n) \sum_{i=0}^{N-1} c_k(i) f_l(n-i)}{\sum_{n=0}^{N-1} [\sum_{i=0}^{N-1} c_k(i) f_l(n-i)]^2} \tag{4.13}$$

Assim, a combinação ótima é aquela que minimiza o erro quadrático médio dado pela expressão (4.12).

A equações (4.13) e (4.12) assemelham-se às equações (3.34) e (3.37), respectivamente, utilizadas no processo de análise por síntese num codificador CELP para a escolha do vetor ótimo de excitação. Entretanto, enquanto que num codificador CELP, para cada vetor de excitação existe apenas um filtro LPC, no algoritmo AMF existem L filtros diferentes. Consequentemente, para cada filtro l é necessário calcular um novo sinal de voz com ponderação, $s_w^l(n)$, e subtrair do mesmo a memória do filtro de síntese em análise, $d_0^l(n)$, devido à excitação ótima do bloco anterior,

obtendo-se assim o sinal de referência correspondente $d_l(n)$.

Num caso prático, para se obter a combinação ótima dos vetores de excitação e coeficientes LPC, é necessário fazer uma busca exaustiva de todas as combinações possíveis dos vetores pertencentes aos dois dicionários, podendo levar a uma complexidade muito elevada dependendo do tamanho dos dicionários.

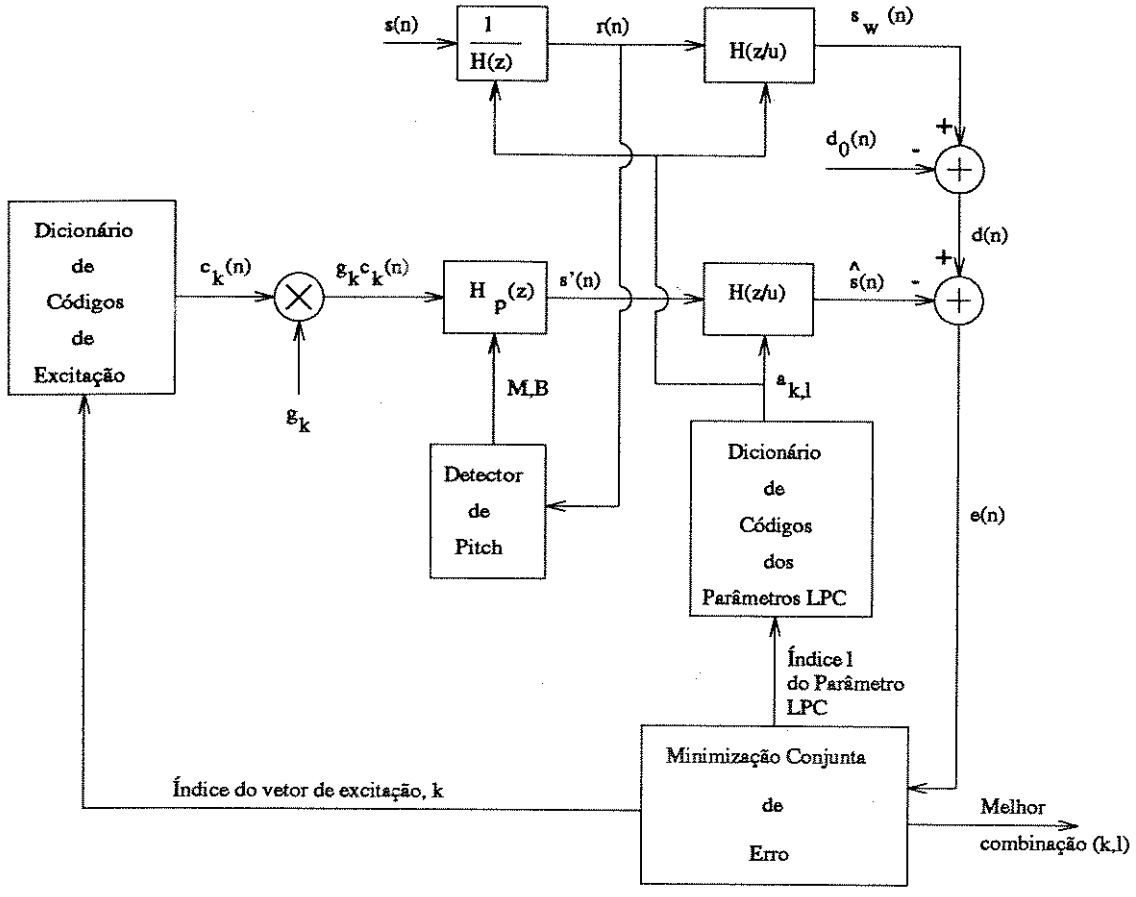

Figura 4.3: Quantização vetorial dos coeficientes LPC por Análise em Malha Fechada

4.4.2 Quantização Vetorial AMF com Busca Exaustiva

Na quantização AMF determina-se simultaneamente o filtro LPC e a seqüência de excitação conjuntamente ótimos. Substituindo-se a equação (4.13) em (4.12) e desconsiderando-se a quantização do ganho $g_{k,l}$, obtém-se que o erro quadrático médio da combinação de uma seqüência de excitação, $c_k(n)$, com um filtro de síntese

com resposta impulsiva $f_l(n)$, é dado por :

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} d_l^2(n) - \frac{[\sum_{n=0}^{N-1} d_l(n)\psi_{k,l}(n)]^2}{v_{k,l}}, \tag{4.14}$$

onde :

$$\psi_{k,l}(n) = \sum_{i=0}^{N-1} c_k(i) f_l(n-i) \tag{4.15}$$

$$v_{k,l} = \sum_{n=0}^{N-1} \psi_{k,l}^2(n) \tag{4.16}$$

Em codificadores CELP, o bloco de análise LPC é normalmente dividido em sub-blocos para a atualização do vetor de excitação. Neste caso, a equação (4.14) é calculada para cada sub-bloco de N amostras e o erro quadrático médio por bloco é igual a média aritmética dos erros por sub-bloco. Sejam K e L o tamanho dos dicionários de códigos de excitação e coeficientes LPC, respectivamente. Então, um algoritmo trivial, porém de elevado custo computacional, consiste em fazer uma busca exaustiva da combinação do vetor de excitação e coeficientes LPC que resulte no menor erro por bloco do seguinte modo :

1. Inicia-se com o primeiro filtro de síntese LPC, isto é, l=1;
2. Determina-se para cada sub-bloco do sinal de voz original, o vetor de excitação que minimiza o erro quadrático médio dado pela equação (4.14). Salva-se o resultado;
3. Determina-se o erro quadrático médio por bloco como sendo a média aritmética dos erros por sub-bloco calculados no passo 2. Salva-se o resultado;
4. Parte-se para o próximo filtro : $l = l + 1$
5. $l \geq L$? Sim, segue em frente. Não, volta ao passo 2;
6. A partir dos resultados salvos no passo 3, obtém-se o vetor LPC que resultou no menor erro por bloco, isto é, o filtro LPC ótimo;
7. A partir dos resultados salvos no passo 2, obtém-se as seqüências de excitação ótimas por sub-bloco para o filtro LPC ótimo determinado no passo 6.

Complexidade do Método AMF

Para pequenos dicionários de códigos, as funções $\psi_{k,l}(n)$ e $v_{k,l}$ podem ser previamente calculadas e devidamente armazenadas. Neste caso, a complexidade total de busca por ciclo do vetor ótimo LPC pelo algoritmo AMF é igual a soma das complexidades de cálculo dos sinais de voz original com ponderação, $s_w^l(n)$, e de referência, $d_l(n)$ (aprox. $2Np$ multiplicações/adições) somada à complexidade de solução da equação (4.14) para K vetores de excitação (aprox. $KN + N$ multiplicações/adições), resultando em aproximadamente $N(2p + K + 1)$ multiplicações/adições. Assim, uma busca exaustiva completa por sub-bloco de N amostras é feita em torno de $LN(2p + K + 1)$ multiplicações/adições.

Em termos práticos, este algoritmo é viável somente para pequenos dicionários de códigos, por exemplo, K=L=32. Então, se $p = 8$, $N = 40$ e o período de pitch mínimo é igual ou maior do que $N + 1$, tem-se uma complexidade de 62.720 multiplicações/adições para uma busca exaustiva completa por sub-bloco de $N = 40$ amostras. Para um bloco de análise LPC de 20 ms dividido em 4 sub-blocos para a determinação do vetor ótimo de excitação, esta complexidade corresponde a aproximadamente 25,1 MFLOPS ("Mega Floating-Point Operations per Second"), isto é, cerca de $12,5 \times 10^6$ mulitiplicações e $12,5 \times 10^6$ adições.

Para casos de dicionários de códigos maiores, um procedimento de busca por sub-dicionários é proposto neste trabalho. Este procedimento não-ótimo em termos de distorção do sinal de voz sintetizado, porém computacionalmente muito mais eficiente, é descrito na seção seguinte.

4.4.3 Quantização Vetorial AMF por Sub-Dicionário de Códigos

Vários algoritmos de quantização vetorial AMF sub-ótimos para a busca dos coeficientes LPC, porém computacionalmente mais eficientes, podem ser desenvolvidos. O algoritmo desenvolvido neste trabalho consiste em determinar inicialmente, em malha aberta e a cada bloco, um sub-conjunto de vetores do dicionário de códigos LPC original com as menores distorções usando-se o algoritmo tradicional de quantização vetorial ou o algoritmo AEP. Em outras palavras, a cada bloco constrói-se um sub-dicionário de códigos de coeficientes LPC x vezes menor do que o original. Analogamente, utilizando-se o vetor LPC do sub-dicionário que resultou na menor distorção, determina-se em malha fechada, a cada sub-bloco, um sub-dicionário de excitação y vezes menor do que o dicionário de códigos original usando-se o erro

quadrático médio ponderado como medida de distorção. Então, uma busca exaustiva da combinação ótima dos vetores é realizada sobre dois sub-dicionários de tamanho muito menor do que dos dicionários originais. Como mostrado nas figuras 4.12 e 4.13 da seção 4.5, para sub-dicionários de códigos x=y=4 (4 vezes menor que os dicionários de códigos originais LPC e de excitação), tem-se uma degradação de apenas 0.1-0.3 dB na *RSRseg* e menos do que 0.08 dB na *DC* em comparação com o procedimento de busca exaustiva sobre os dicionários de códigos originais.

4.4.4 Extensões da Quantização Vetorial AMF

A quantização vetorial pelo algoritmo AMF até agora descrita realiza uma otimização conjunta de somente dois conjuntos de parâmetros : LPC e de excitação de curto-prazo. Este algoritmo, doravante denominado AMF de referência (AMF-REF) pode ser estendido nos algoritmos AMF-1 e AMF-2, de modo a otimizar o período de pitch e a alocação de bits entre os coeficientes LPC e sinal de excitação.

Método AMF-1

No algoritmo AMF-1, a detecção do período de pitch é feita através de uma busca por bloco em duas etapas : inicialmente, através de um procedimento em malha aberta, determina-se uma lista de períodos de pitch candidatos a serem avaliadas posteriormente em malha fechada. A lista de períodos de pitch candidatos é criada selecionando-se os valores de m que resultam em maiores valores de correlação normalizada dada pela equação (3.20). Uma vez criada a lista de períodos de pitch candidatos, inicia-se a segunda etapa do procedimento, onde a determinação do período de pitch final é realizada em malha fechada por bloco, de modo a obter a melhor combinação de período de pitch, vetor de coeficientes LPC e de excitação. Este procedimento de detecção do período de pitch é aproximadamente ótimo, pois é feito em malha fechada com o sinal de excitação presente na entrada do preditor de longo-prazo ao invés de entrada *zero*. Como pode ser visto dos resultados de simulações mostrados nas figura 4.12 a 4.15 da seção 4.5, uma lista de quatro candidatos produz uma melhora de até aproximadamente 0.60 dB na *RSRseg* e de até 0.23 dB na *DC*.

Método AMF-2

No algoritmo AMF-2, além da otimização em malha fechada do período de

pitch, é feita uma alocação dinâmica de bits entre os coeficientes LPC e o sinal de excitação. A alocação dinâmica de bits baseia-se na existência de uma redundância significativa entre índice dos vetores LPC de blocos sucessivos. Esta redundância pode ser explorada para reduzir ainda mais a taxa de bits dos coeficientes LPC ou, então, melhorar a qualidade alocando-se um maior número de bits para o sinal de excitação. Na figura 4.4, é apresentada a razão entre número de mudanças de vetores LPC ocorridas e o número total de vetores LPC transmitidos num trecho de 6,14 s de voz em função do tamanho do dicionário de códigos. Adicionalmente, na figura 4.5, são mostrados os valores dos índices dos vetores LPC ao longo de um trecho de sinal de voz. Observa-se que a redundância é maior durante os trechos sonoros do sinal de voz. Uma alocação dinâmica de bits entre os coeficientes LPC e as seqüências de excitação em função da redundância dos vetores LPC, resulta numa melhoria adicional de até aproximadamente 0.5 dB e 0.18 dB na $RSRseg$ e DC, respectivamente, como pode ser visto nas figuras 4.16 e 4.17 da seção 4.5.

Complexidade do Método AMF por Sub-dicionário de Códigos

No algoritmo AMF por sub-dicionário de códigos, embora o tamanho dos sub--dicionários seja pequeno, os dicionários originais podem ser de tamanho tal que torne impraticável o armazenamento de todos os valores das funções $\psi_{k,l}(n)$ e $v_{k,l}$, calculados conforme as equações (4.15) e (4.16). Assim, é necessária uma redução de complexidade no cálculo do erro definido pelas equações (4.12) e (4.13). Esta redução pode ser conseguida usando-se a seguinte aproximação [21] :

$$\sum_{n=0}^{N-1}\{\sum_{i=0}^{N-1} c_k(i)f_l(n-i)\}^2 \simeq \varepsilon_l(0)\nu_k(0) + 2\sum_{i=1}^{N-1}\varepsilon_l(i)\nu_k(i) \tag{4.17}$$

onde :

$$\varepsilon_l(i) = \sum_{n=0}^{N-1} f_l(n)f_l(n-i) \tag{4.18}$$

$$\nu_k(i) = \sum_{n=0}^{N-1} c_k(n)c_k(n-i) \tag{4.19}$$

Portanto, substituindo-se estas equações em (4.12) e (4.13), o erro e ganho do vetor de excitação passam a serem calculados, respectivamente, como :

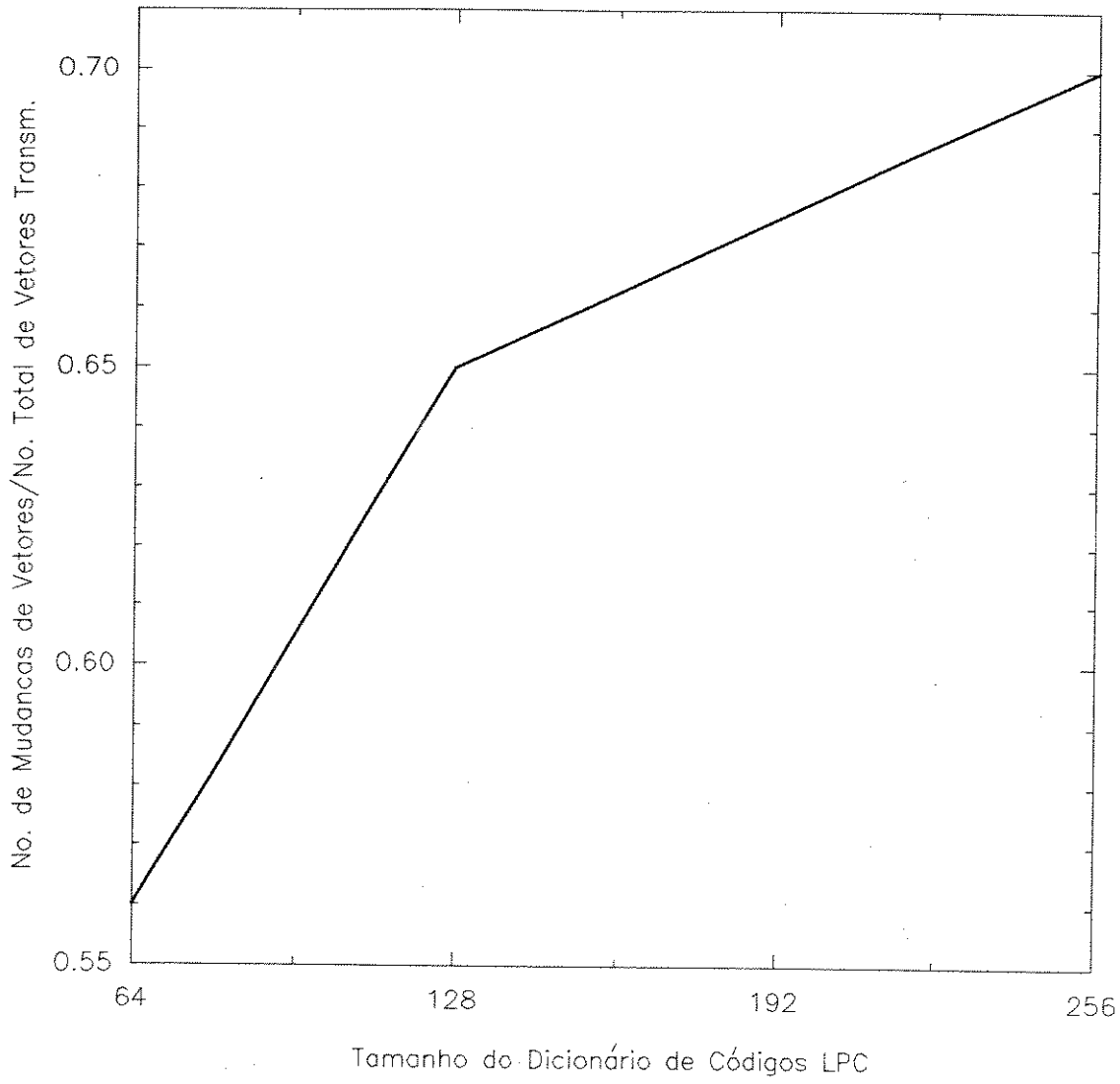

Figura 4.4: Curva da razão entre o número de mudanças de vetores LPC ocorridas e o número total de vetores LPC transmitidos.

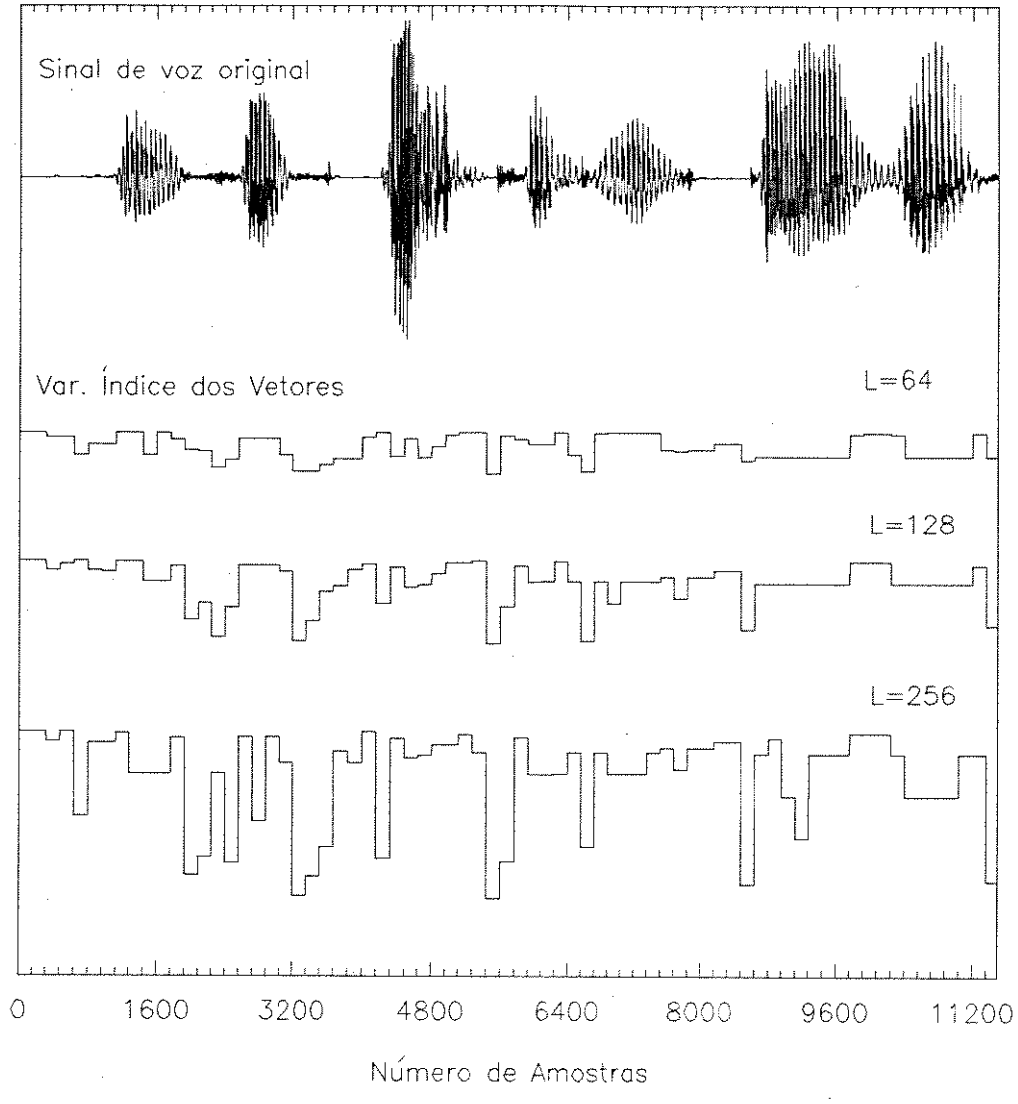

Figura 4.5: Histograma da variação dos índices dos vetores LPC ao longo de um trecho de sinal de voz.

$$\langle e_w^2 \rangle \simeq \sum_{n=0}^{N-1} d_l^2(n) - 2\hat{g}_{k,l} \sum_{n=0}^{N-1} d_l(n) \sum_{i=0}^{N-1} c_k(i) f_l(n-i) + \hat{g}_{k,l}^2 (\varepsilon_l(0)\nu_k(0) + 2 \sum_{i=1}^{N-1} \varepsilon_l(i)\nu_k(i)) \tag{4.20}$$

$$g_{k,l} \simeq \frac{\sum_{n=0}^{N-1} d_l(n) \sum_{i=0}^{N-1} c_k(i) f_l(n-i)}{\varepsilon_l(0)\nu_k(0) + 2\sum_{i=1}^{N-1} \varepsilon_l(i)\nu_k(i)} \tag{4.21}$$

Considerando-se que um vetor de excitação $c_k(n)$ consiste de somente w pulsos diferentes de zero, a complexidade da expressão : $\sum_{n=0}^{N-1} d_l(n) \sum_{i=0}^{N-1} c_k(i) f_l(n-i) = \sum_{i=0}^{N-1} c_k(i) \sum_{n=0}^{N-1} d_l(n) f(n-i)$, para todas as combinações possíveis dos vetores de excitação e LPC dos dois sub-dicionários de códigos, é de $\frac{N(N+1)L_s}{2} + wK_sL_s$ multiplicações/adições, onde K_s e L_s são, respectivamente, o tamanho dos sub-dicionários de excitação e LPC. Igualmente, como os vetores de excitação $c_k(n)$ consistem de poucos pulsos diferentes de zero, a maioria dos valores de $\nu_k(n)$ são também nulos. De fato, para $w = 4$ pulsos diferentes de zero no vetor de excitação $c_k(n)$, tem-se no pior caso somente 6 valores diferentes de zero para a função de autocorrelação $\nu_k(i)$. Uma maneira de calcular a expressão $\varepsilon_l(0)\nu_k(0) + 2\sum_{i=1}^{N-1} \varepsilon_l(i)\nu_k(i)$, supondo-se que o período de pitch é maior do que N, consiste em calcular previamente os valores de $\nu_k(i)$ e $\varepsilon_l(i)$ e armazená-los adequadamente. Assim, a complexidade de cálculo desta expressão para todas as combinações possíveis dos vetores dos dois sub-dicionários é igual a $(\lambda+1)K_sL_s$ multiplicações/adicões, onde λ é o número de valores de $\nu_k(i)$, $i = 1, \ldots, N-1$, diferentes de zero. Supondo-se que p é a ordem do filtro de síntese LPC em cascata com o filtro de síntese de longo-prazo de ordem 1, e que a resposta impulsiva desta combinação encontra-se previamente calculada e armazenada, são necessários $N(2pL_s + 1)$ multiplicações/adições para o cálculo do sinal de referência, $d_l(n)$. Finalmente, o cálculo da energia do sinal $d_l(n)$ requer NL_s multiplicações/adições. Somando-se todas estas operações, resulta que a complexidade total de busca do vetor ótimo por sub-dicionários de códigos do algoritmo AMF-REF é de $\frac{NL_s}{2}(N+1) + L_sK_s(w+\lambda+1) + N(2pL_s+1)$ multiplicações/adições. A título de ilustração, para um caso típico onde $w = 4$, $N = 40$, $K_s = L_s = 32$ e $p = 8$, obtém-se uma complexidade de aproximadamente 58.024 multiplicações/adições por sub-bloco. Para um bloco de análise LPC de 20 ms dividido em 4 sub-blocos para a determinação do vetor ótimo de excitação, esta complexidade corresponde a aproximadamente 23,21 MFLOPS.

No caso dos algoritmos AMF-1 e AMF-2 utilizando uma lista de até T períodos de pitch candidatos maior do que 40 amostras, tem-se uma complexidade de aproximadamente $\frac{NL_sT}{2}(N+1)+L_sK_s(wT+\lambda+1)+N(2pL_s+1)$ multiplicações/adições. Assim, para os mesmos valores de parâmetros do caso anterior e $T = 4$, obtém-se uma complexidade de 149.032 multiplicações/adições ou 59,61 MFLOPS. Em termos práticos, um codificador de voz com esta complexidade é perfeitamente factível de ser implementado em dois DSP's comerciais (p.ex., TMS320C30-40) ou mesmo em um único DSP de última geração(p. ex., TMS320C40). Através de simulações, tem-se verificado que o fato de se considerar somente períodos de pitch candidatos maior do que 40 amostras implica em uma degradação desprezível de desempenho.

4.5 SIMULAÇÕES

Nesta seção são apresentadas as simulações e respectivos resultados dos algoritmos AEP e AMF de quantização vetorial dos coeficientes LPC. As simulações consistem de comparações de desempenho de um modelo de codificação de voz CELP convencional quando os coeficientes LPC são quantizados por algoritmos tradicionais e pelos algoritmos AEP e AMF. Como medidas de desempenho são utilizadas a $RSRseg$ e DC acompanhadas de testes subjetivos informais.

4.5.1 Modelo CELP Convencional

O modelo de codificação de voz CELP convencional utilizado é mostrado na figura 3.9. O filtro de síntese consiste de um filtro de longo-prazo de 1a ordem em cascata com um filtro LPC de 8a ordem. O modelo usa um único dicionário de códigos de excitação, sendo que um vetor de excitação ótimo é determinado a cada sub-bloco de 40 amostras (5 ms) através de uma busca exaustiva. Nesta busca, a medida de distorção empregada é o erro quadrático médio ponderado.

O filtro de síntese de longo-prazo possui um retardo de 40 a 167 amostras, sendo, assim, empregados 7 bits para a representação do período de pitch. Com relação ao ganho do vetor de excitação e ganho de pitch, preferiu-se não quantizar estes parâmetros, pois uma melhor avaliação de desempenho dos algoritmos de quantização dos coeficientes LPC poderia ser obtida na ausência de outras degradações. Por outro lado, o filtro LPC teve os seus coeficientes quantizados através dos seguintes algoritmos :

1. Quantização escalar dos coeficientes transformados em LAR;
2. Quantização vetorial utilizado-se a distorção de Itakura Saito Modificada
3. Quantização vetorial pelos algoritmos AEP e AMF

Dicionários de excitação de diversos tamanhos ($K = 16, 32, ..., 1024$ vetores de dimensão igual a 40 amostras), gerados a partir de um ruído gaussiano com ceifagem central de 80%, foram utilizados nas simulações.

4.5.2 Quantização Escalar

A quantização escalar utilizada neste trabalho para fins de comparação de desempenho com os algoritmos AEP e AMF, é baseada (mas, não estritamente em conformidade) na quantização escalar dos coeficientes LPC do codec RPE-LTP a 13 kbit/s padronizado, através da especificação GSM 06.10 [17], para aplicação na primeira geração digital de telefonia celular européia. Os coeficientes de reflexão, $r(k)$, são obtidos segmentando-se, com janelas retangulares e sem superposição, o sinal de voz original em blocos de 160 amostras. Para efeitos de quantização e codificação, os coeficientes de reflexão são transformados em razão log-área, $LAR(k)$, definido como :

$$LAR(k) = \log_{10} \frac{1 + r(k)}{1 - r(k)}, \qquad k = 1, ..., p \tag{4.22}$$

É sabido que a razão log-área, $LAR(k)$, tem faixas dinâmicas e densidades de distribuição de amplitudes assimétricas e diferentes [1, 2]. Em função disto, os coeficientes $LAR(k)$ são limitados e quantizados diferentemente de acordo com a equação (4.23), onde $LAR_c(k)$ é a versão quantizada e codificada de $LAR(k)$ [17].

$$LAR_c(k) = Nint[A(k)LAR(k) + B(k)] \tag{4.23}$$

$$Nint[z] = int[z + 0.5sign(z)] \tag{4.24}$$

$$sign(z) = \begin{cases} 1, & z > 0 \\ -1, & z \leq 0 \end{cases} \tag{4.25}$$

A função $Nint$ define a operação de arredondamento ao valor do inteiro mais próximo. Os coeficientes $A(k)$, $B(k)$ e os diferentes valores máximos e mínimos de $LAR_c(k)$, bem como o número de bits empregados por coeficiente são dados na tabela 4.1.

Tabela 4.1 :Quantização das razões Log-Área

Índice k do LAR	$A(k)$	$B(k)$	Valor Mínimo de $LAR_c(k)$	Valor Máximo de de $LAR_c(k)$	Número de bits
1	20,000	0,000	-32	+31	6
2	20,000	0,000	-32	+31	6
3	20,000	4,000	-16	+15	5
4	20,000	-5,000	-16	+15	5
5	13,637	0,184	-8	+7	4
6	15,000	-3,500	-8	+7	4
7	8,334	-0,666	-4	+3	3
8	8,824	-2,235	-4	+3	3

Os coeficientes razão log-área quantizados e codificados, LAR_c, são decodificados de acordo com a seguinte expressão :

$$LAR'(k) = [LAR_c(k) - B(k)]/A(k) \tag{4.26}$$

Finalmente, os coeficientes de reflexão decodificados são determinados usando a transformada inversa da equação (4.22). A fim de evitar uma mudança brusca de filtro LPC de um bloco de sinal de voz para outro, os coeficientes de reflexão passam por um processo de interpolação.[1] De acordo com este processo de interpolação, se $r_0(k)$, $r_1(k)$ e $r(k)$, $k = 1, ..., p$, são, respectivamente, os coeficientes de reflexão do bloco atual, anterior e interpolado, tem-se que :

$$r(k) = 0.25r_0(k) + 0.75r_1(k), \qquad 0 \leq n \leq 12 \tag{4.27}$$

$$r(k) = 0.5r_0(k) + 0.5r_1(k), \qquad 12 < n \leq 26 \tag{4.28}$$

$$r(k) = 0.75r_0(k) + 0.25r_1(k), \qquad 26 < n \leq 39 \tag{4.29}$$

$$r(k) = r_0(k), \qquad n > 39 \tag{4.30}$$

onde n indica o número da amostra dentro do bloco de 160 amostras.

4.5.3 Quantização Vetorial

Utilizando-se a medida de distorção de Itakura Saito Modificada, foram projetados dicionário de códigos LPC de diversos tamanhos ($L = 16, 64, ..., 256$ vetores de dimensão igual a 8). Os dicionários de códigos consistem de células C_l, $l = 1, .., L$, cujo centróide é dado por [18] :

[1] Na especificação GSM 06.01, é feita uma interpolação dos coeficientes $LAR'(k)$ e posteriormente obtidos os coeficientes de reflexão decodificados. Neste trabalho, para fins de compatibilização com os procedimentos adotados na quantização vetorial dos coeficientes LPC, a interpolação é feita com os coeficientes de reflexão decodificados. De acordo com simulações realizadas, esta mudança não traz nenhuma alteração significativa de desempenho.

$$\hat{\mathbf{a}} = [\sum_{j:\mathbf{a}_j \epsilon C_l} R_{\mathbf{x}_j}]^{-1} \sum_{j:\mathbf{a}_j \epsilon C_l} R_{\mathbf{x}_j} \mathbf{a}_j \tag{4.31}$$

onde : $R_{\mathbf{x}_j}$ é a matriz de autocorrelação normalizada associada ao sinal de entrada $\mathbf{x}_j$ conforme definida pela equação (4.2);
$\mathbf{a}_j$ é o vetor de coeficientes LPC originais associado a $\mathbf{x}_j$.

Todos os dicionários de códigos foram projetados através do método LBG [18], utilizando-se no processso de "splitting" um limiar de perturbação dos vetores códigos igual a 1%, e finalizando-se o processo de agrupamento sempre que a diferença entre as distorções médias de duas iterações fosse menor do que 0,1%. Como medida de distorção no processo de agrupamento, empregou-se a medida de Itakura Saito Modificada, conforme equação (4.1), sendo, portanto, calculada diretamente sobre os coeficientes LPC originais e quantizados.

A seqüência de treinamento utilizada na geração dos dicionários de códigos é formada por vozes de 10 locutores, sendo 5 homens e 5 mulheres. Cada locutor contribuiu com a leitura de um texto durante 1 minuto, resultando em uma seqüência de treinamento com 10 minutos de voz. Para a digitalização do sinal de voz, utilizou-se uma freqüência de amostragem de 8 kHz, um conversor A/D de 16 bits e posteriormente uma normalização de amplitude máxima em 4095 níveis, correspondente a uma resolução de 13 bits, visando compatibilizar com o nível máximo de entrada de um codificador PCM lei-A=87,6.

Os dicionários de códigos assim gerados foram utilizados para a quantização vetorial dos coeficientes LPC empregando-se os seguintes algoritmos de busca :

1. Algoritmo tradicional, mostrado na figura 4.1, usando-se a medida de distorção de Itakura Saito Modificada;
2. Algoritmo AEP;
3. Algoritmo AMF.

A fim de evitar uma mudança brusca de filtro LPC de um bloco de sinal de voz para outro, os coeficientes de reflexão, uma vez selecionados do dicionário de códigos, seja nos esquemas AEP, AMF ou tradicional, passam por um processo de interpolação, do mesmo modo que no caso de quantização escalar descrito na seção anterior. Nos algoritmos AEP e AMF, os coeficientes são obtidos do dicionário de códigos na forma de coeficientes de reflexão. No caso do algoritmo tradicional, os

coeficientes LPC selecionados são primeiramente transformados em reflexão para posteriormente sofrerem o processo de interpolação.

4.5.4 Resultados Obtidos.

Comparação com AEP

Nas figuras 4.6 a 4.11, estão apresentados os resultados de desempenho objetivo em termos de RSR_{seg} e DC obtidas pelo modelo CELP com : i) quantização escalar dos coeficientes LPC transformados em log-área; ii) quantização vetorial tradicional e iii) quantização vetorial AEP. Os resultados apresentados nas figuras 4.8 e 4.9 foram obtidos com os coeficientes LPC quantizados vetorialmente a cada 160 amostras (20 ms), enquanto que nas figuras seguintes foram quantizados a cada 40 amostras (5 ms). Conforme os resultados obtidos, tem-se as seguintes principais conclusões :

1. O algoritmo AEP por covariância (AEP-COV) apresenta um desempenho objetivo ligeiramente superior ao AEP por autocorrelação (AEP-AUT) para blocos de análise LPC de 160 amostras, mas testes subjetivos informais indicam qualidade subjetiva aproximadamente equivalente. Para blocos de análise de 40 amostras, existe uma diferença acentuada de desempenho objetivo e subjetivo. Através de testes subjetivos informais, percebe-se que o algoritmo AEP-AUT introduz um nível de distorção espectral e principalmente de ruído de quantização mais elevado.

2. O algoritmo AEP-COV apresenta um desempenho objetivo em termos de RSR_{seg} equivalente ao algoritmo tradicional de quantização vetorial quando o bloco de análise LPC é de 160 amostras. Entretanto, devido à degradação de aproximadamente 0.5 dB menor em termos de DC (i.e., uma degradação menor na envoltória do espectro), a qualidade subjetiva do sinal de voz sintetizado é ligeiramente superior à do algoritmo tradicional. Para blocos de análise de 40 amostras (baixo atraso), o algoritmo AEP-COV apresenta valores de RSR_{seg} ligeiramente superiores e DC de aproximadamente 0.8 dB menor. Através de testes subjetivos informais, observa-se que esta diferença na DC resulta em uma qualidade ligeiramente superior para o sinal de voz sintetizado pelo algoritmo AEP.

3. Em relação à quantização escalar e no caso de blocos de análise de 160 amostras, os algoritmos AEP-COV e AEP-AUT apresentam uma degradação de aproxi-

madamente 0.2 e 0.6 dB em termos de RSR_{seg} e DC, respectivamente, a uma taxa de bits em torno de 4 vezes menor que a da quantização escalar.

4. Em relação à quantização escalar e no caso de blocos de análise de 40 amostras, que resulta numa taxa de bits praticamente equivalente à da quantização escalar, o algoritmo AEP-COV é superior 0.3 dB em termos de RSR_{seg}. Entretanto, em termos subjetivos, a qualidade do sinal de voz sintetizado pelo algoritmo AEP é prejudicado por uma degradação maior em termos de DC de cerca de 0.45 dB.

Um ponto importante que deve ser ressaltado, é que os dicionários de códigos utilizados não foram projetados utilizando-se o esquema AEP, mas sim o esquema tradicional de quantização vetorial. Caso seja utilizado um dicionário de códigos projetado considerando-se o esquema AEP, resultados ainda mais favoráveis ao algoritmo AEP devem ser obtidos.

Comparação com AMF

Nas figuras 4.12 e 4.13 são mostrados, respectivamente, os resultados de desempenho em termos de RSR_{seg} e DC obtidos pelo modelo CELP com quantização vetorial AMF. Os dicionários originais de códigos LPC e de excitação utilizados são do mesmo tamanho com 128 vetores. Estes resultados são dados em função do tamanho dos sub-dicionários de códigos LPC e de excitação. No caso de RSR_{seg}, observa-se que o emprego de sub-dicionários de códigos de excitação com mais do que 1/4 dos vetores do dicionário original, resulta numa melhora não muito significativa e tende a uma saturação. Em termos de DC, a diferença entre os resultados obtidos pelo dicionário original e sub-dicionário de códigos LPC com 1/4 dos vetores do primeiro, é de aproximadamente 0,27 dB, embora não presente uma tendência de saturação para sub-dicionários maiores como no caso da RSR_{seg}. Assim, na obtenção das curvas de RSR_{seg} e DC em função do tamanho dos dicionários originais de excitação e LPC, mostradas respectivamente nas figuras 4.14 e 4.15, foram utilizados sub-dicionários de tamanho igual a 1/4 dos respectivos dicionários originais.

O algoritmo AMF-REF(não inclui otimização conjunta do preditor de longo-prazo e nem alocação dinâmica de bits), apresenta valores de RSR_{seg} em torno de 0.8 dB superior ao algoritmo tradicional de quantização vetorial dos coeficientes LPC, enquanto que em termos de DC apresenta uma degradação maior de apenas

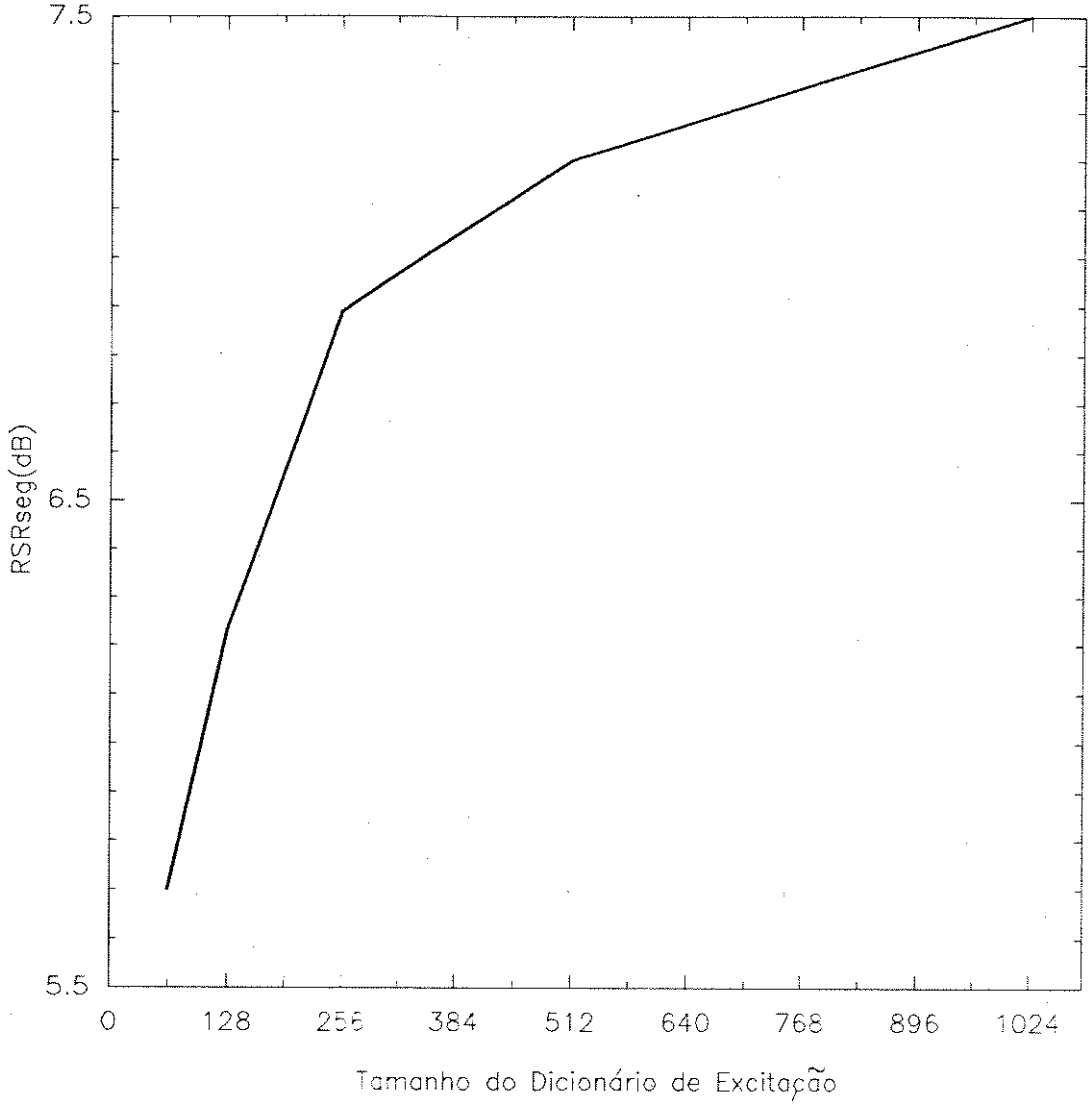

Figura 4.6: $RSR_{seg}(dB)$ com quantização escalar dos coeficientes LPC transformados em log-área em função do tamanho do dicionário de códigos de excitação.

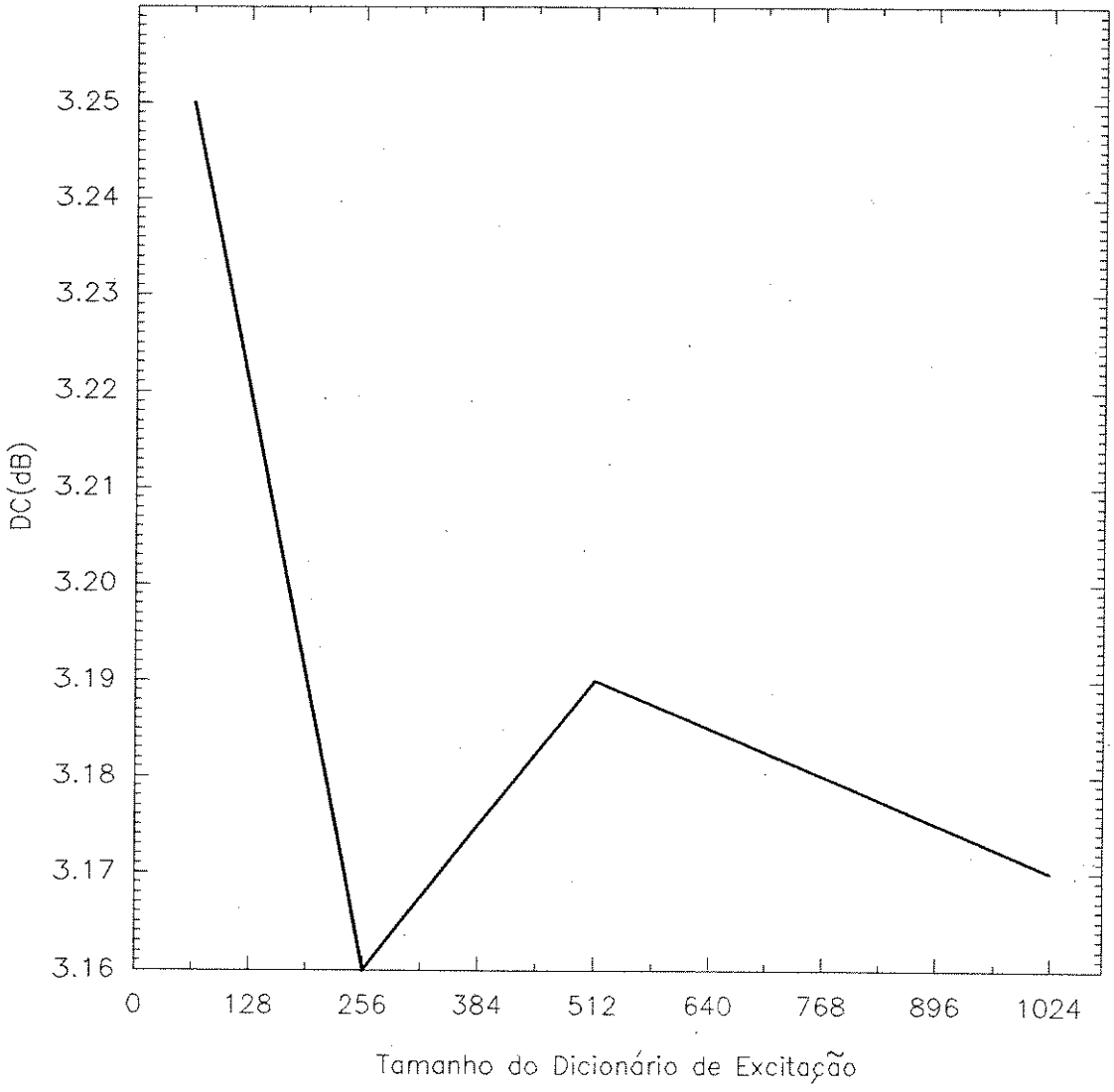

Figura 4.7: $DC(dB)$ com quantização escalar dos coeficientes LPC transformados em log-área em função do tamanho do dicionário de códigos de excitação.

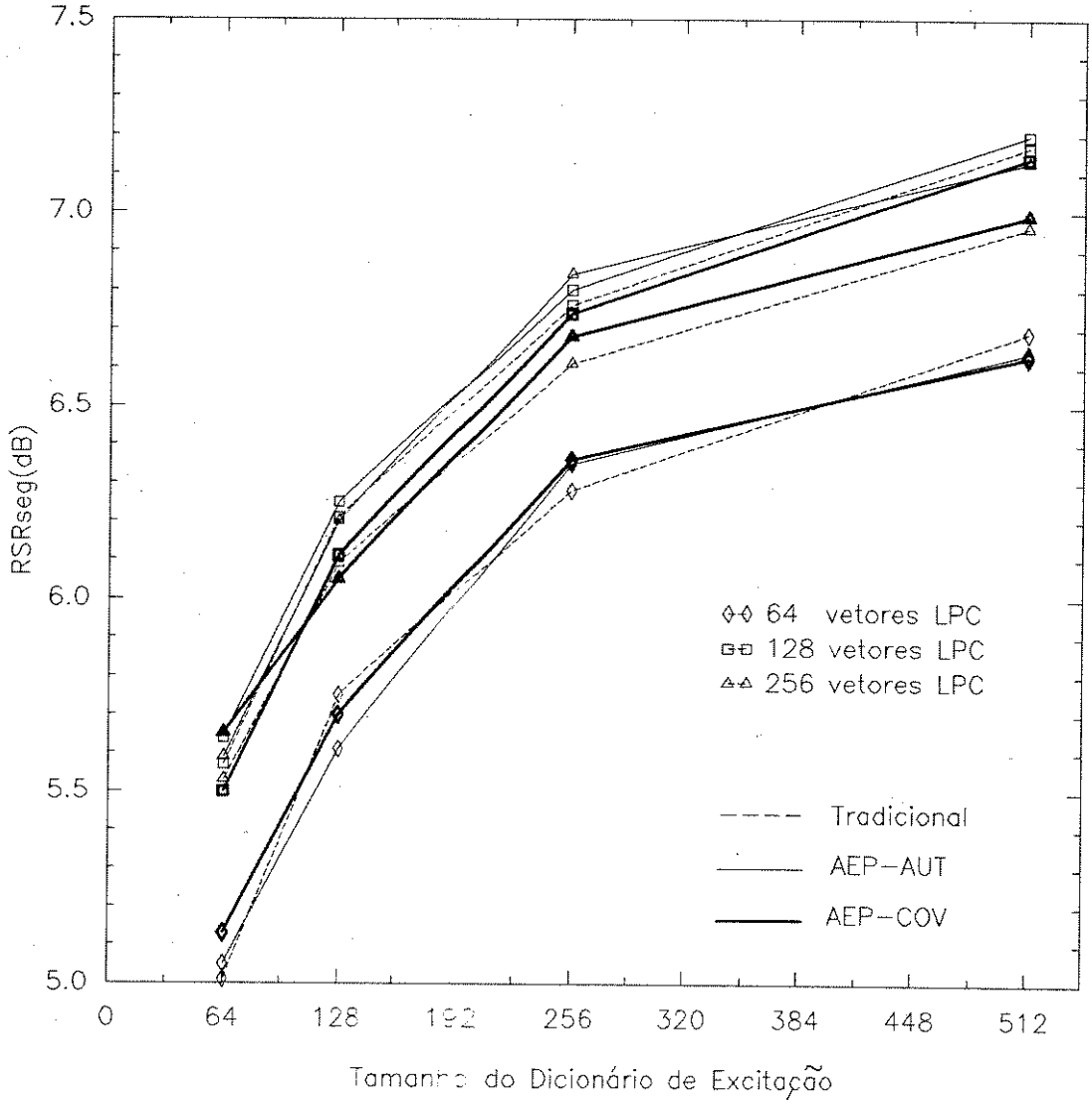

Figura 4.8: $RSR_{seg}(dB)$ para os algoritmos de quantização vetorial tradicional e AEP dos coeficientes LPC a cada 160 amostras em função do tamanho do dicionário de códigos LPC e de excitação.

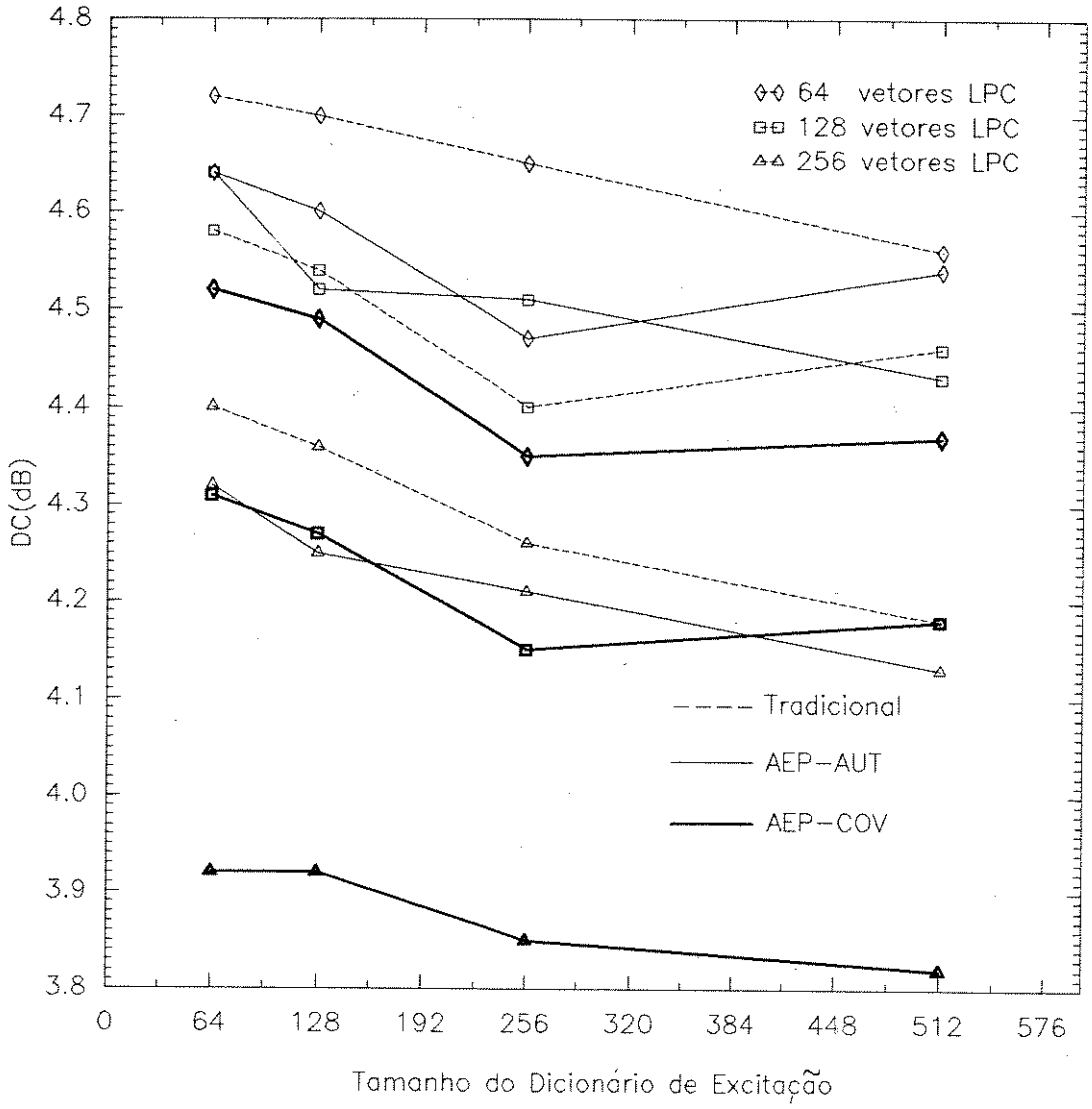

Figura 4.9: $DC(dB)$ para os algoritmos de quantização vetorial tradicional e AEP dos coeficientes LPC a cada 160 amostras em função do tamanho do dicionário de códigos LPC e de excitação.

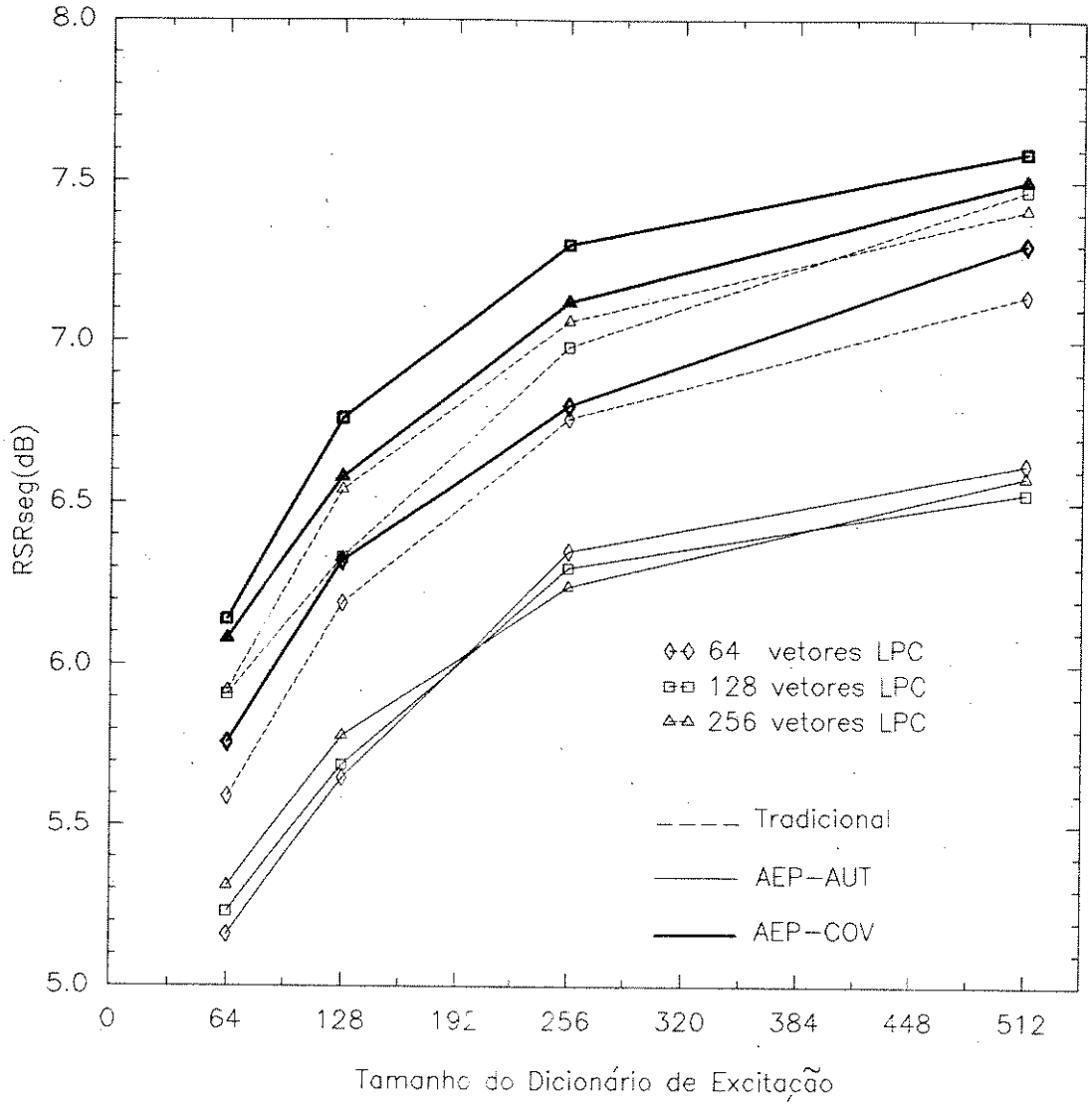

Figura 4.10: $RSR_{seg}(db)$ para os algoritmos de quantização vetorial tradicional e AEP dos coeficientes LPC a cada 40 amostras em função do tamanho dos dicionários de códigos LPC e de excitação.

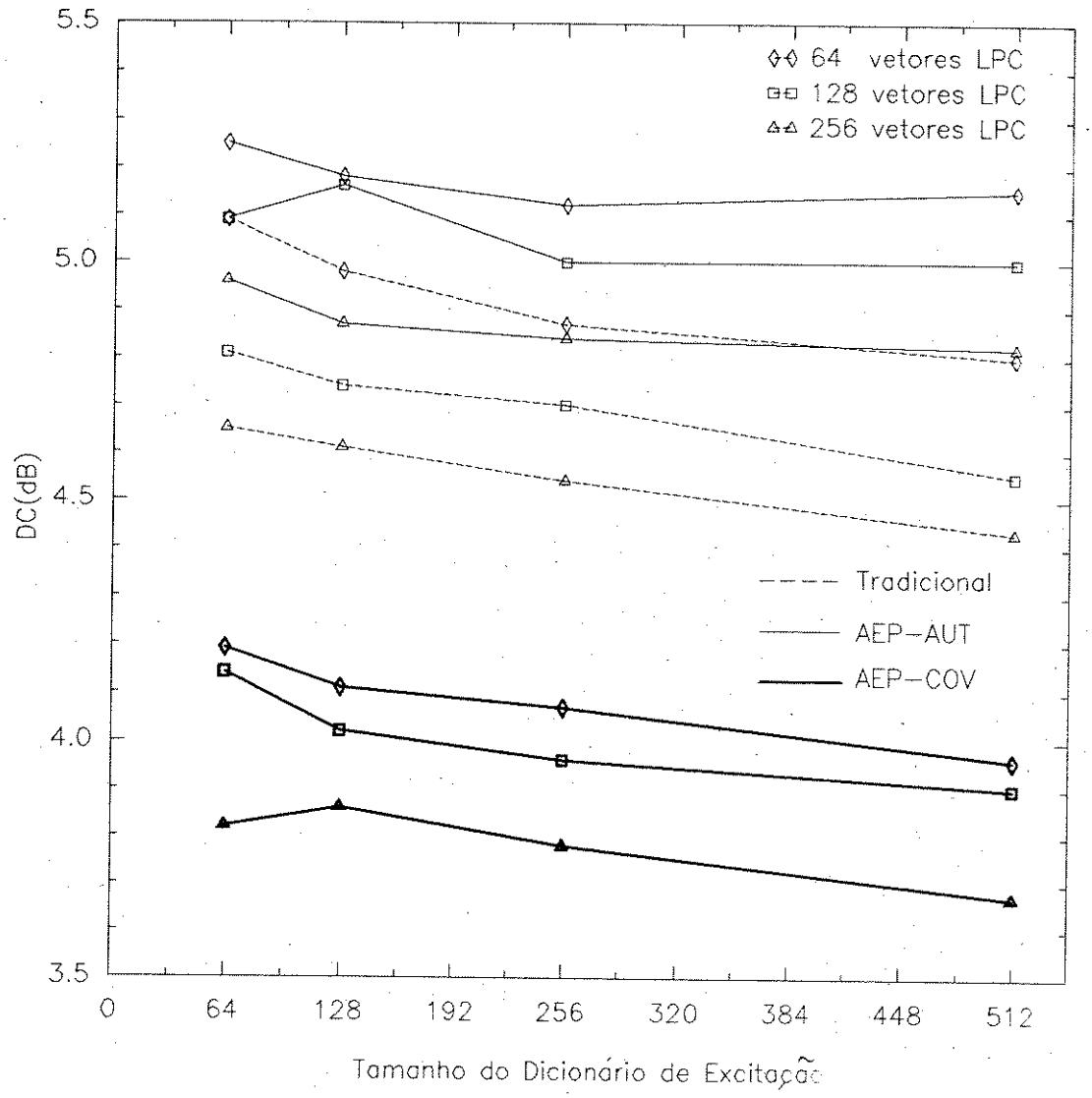

Figura 4.11: $DC(db)$ para os algoritmos de quantização vetorial tradicional e AEP dos coeficientes LPC a cada 40 amostras em função do tamanho dos dicionários de códigos LPC e de excitação.

0.2 dB. Este aumento na DC deve-se à utilização do erro quadrático médio como medida de distorção, o qual minimiza a distorção no domínio do tempo mas não necessariamente a distorção em termos de envoltória do espectro.

O algoritmo AMF-1(inclui otimização conjunta do preditor de longo-prazo) apresenta uma superioridade em relação ao algoritmo tradicional de aproximadamente 1.4 dB em termos de RSR_{seg}, enquanto que em termos de DC apresentam resultados próximos entre si.

Finalmente, o algoritmo AMF-2(inclui otimização conjunta do preditor de longo-prazo e alocação dinâmica dos bits entre os coeficientes LPC e o sinal de excitação) resulta em uma melhora global de aproximadamente 1.8 dB de $RSRseg$ e mantém praticamente a mesma DC, conforme mostrado nas figuras 4.16 e 4.17. Testes subjetivos informais mostram que o aumento na $RSRseg$ obtido pelo algoritmo AMF-REF, compensa a distorção espectral de aproximadamente 0.2 dB maior em relação ao algoritmo tradicional, obtendo-se um sinal de voz sintetizado de qualidade subjetiva notoriamente superior. No caso dos algoritmos AMF-1 e AMF-2, esta superioridade é maior ainda, pois há um aumento significativo em termos de RSR_{seg} sem aumentar a DC.

Uma sinopse de todos os resultados obtidos é mostrada nas figuras 4.18 e 4.19. Os resultados apresentados nestas figuras para os algoritmos de quantização vetorial (Tradicional, AEP e AMF), foram obtidos utilizando-se um dicionário de códigos LPC de 256 vetores.

4.6 CONCLUSÕES

Dois algoritmos de quantização vetorial dos coeficientes LPC, AEP e AMF, foram apresentados. Para um atraso de codificação de 20 ms, o algoritmo AEP apresenta uma pequena vantagem em termos de desempenho em relação ao algoritmo tradicional. Para atraso de 5 ms, esta superioridade de desempenho do algoritmo AEP torna-se mais significativa. Além disto, a complexidade do algoritmo AEP corresponde à aproximadamente metade da complexidade do algoritmo tradicional usando-se medida de distorção de Itakura Saito Modificada.

O algoritmo AMF é computacionalmente muito mais complexo que os algo-

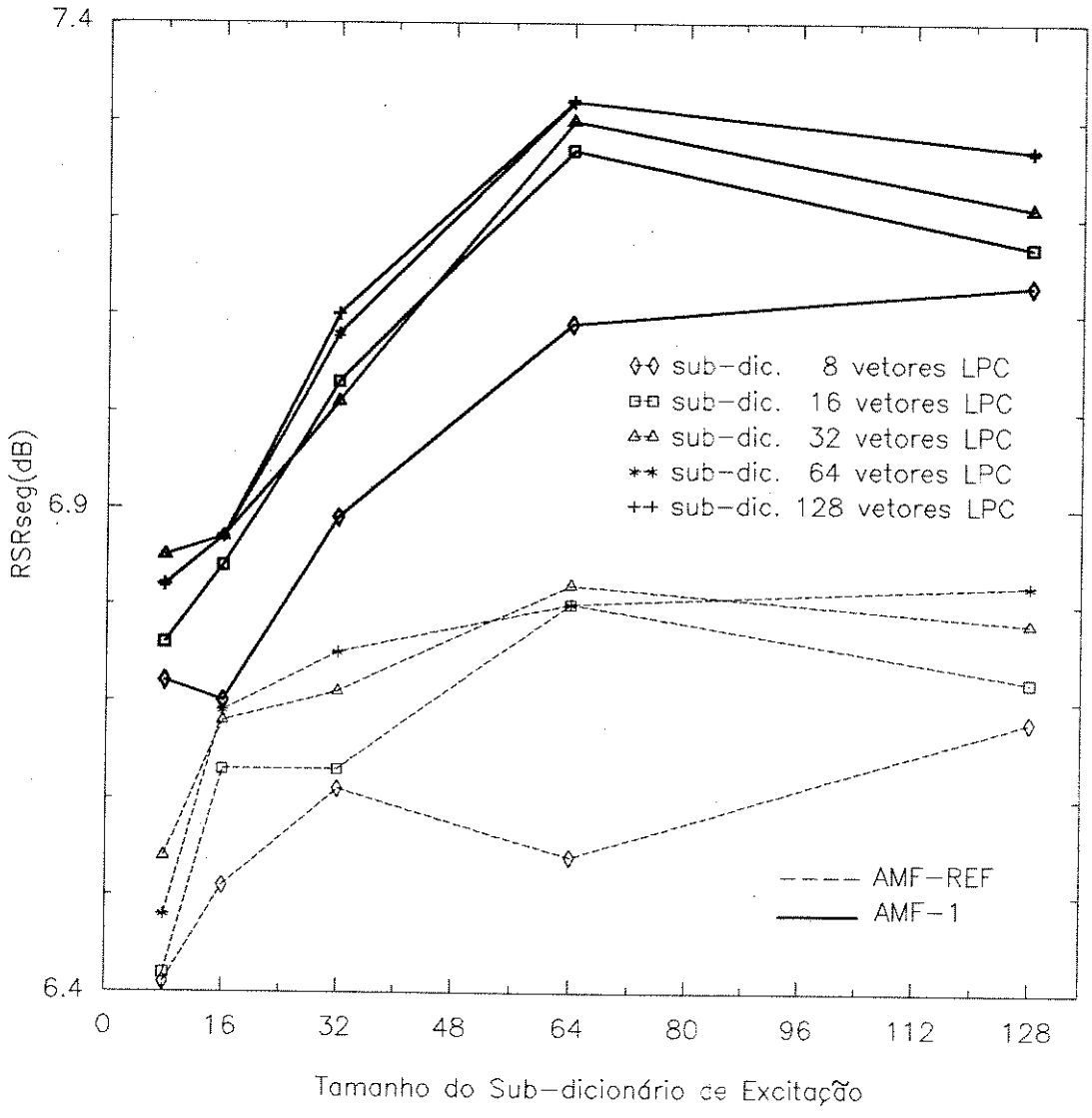

Figura 4.12: $RSR_{seg}(dB)$ para os algoritmos de quantização vetorial AMF dos coeficientes LPC em função do tamanho dos sub-dicionários de códigos para dicionários de códigos LPC e de excitação originais de 128 vetores.

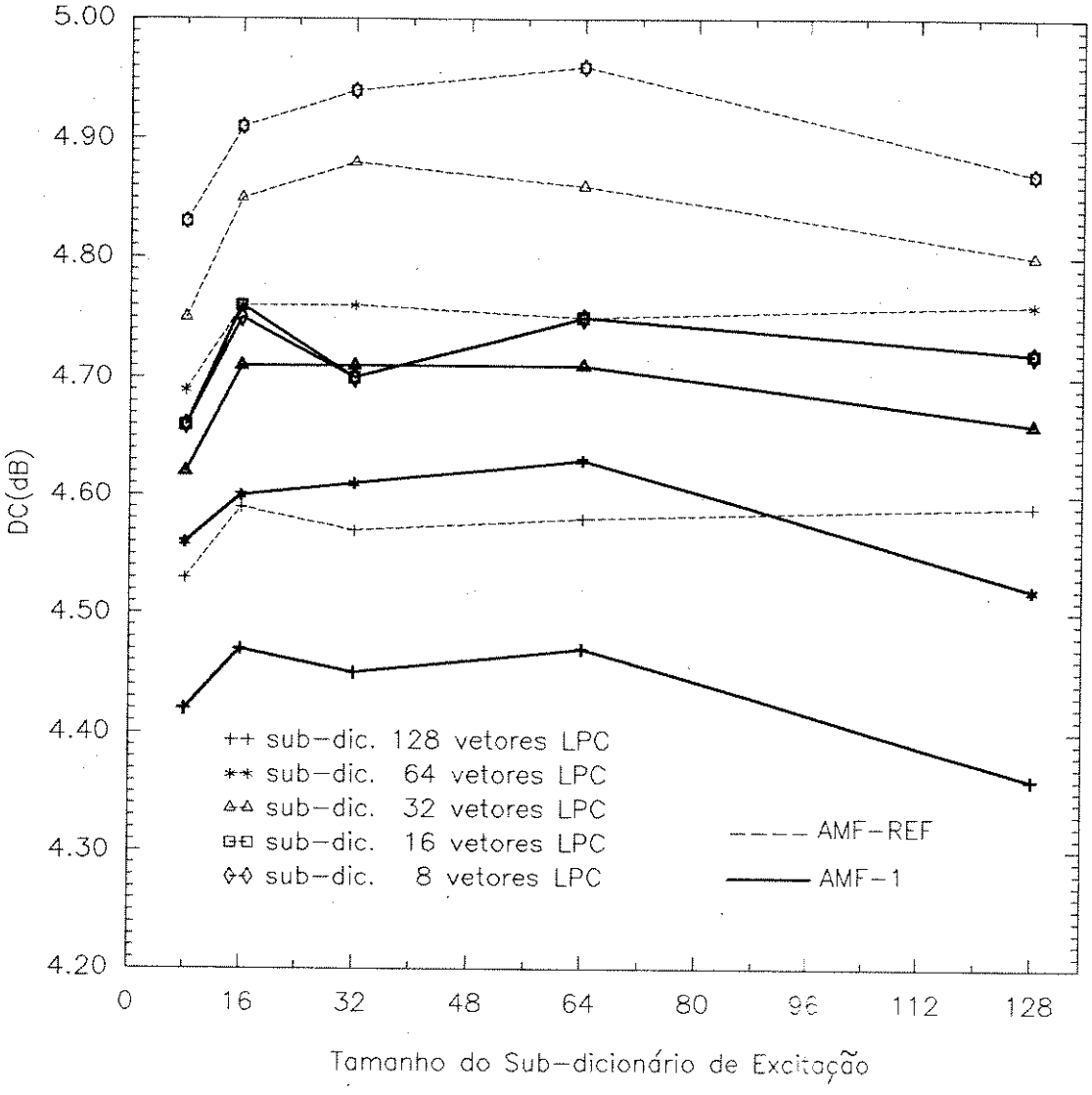

Figura 4.13: $DC(dB)$ para os algoritmos de quantização vetorial AMF dos coeficientes LPC em função do tamanho dos sub-dicionários de códigos para dicionários de códigos LPC e de excitação originais de 128 vetores.

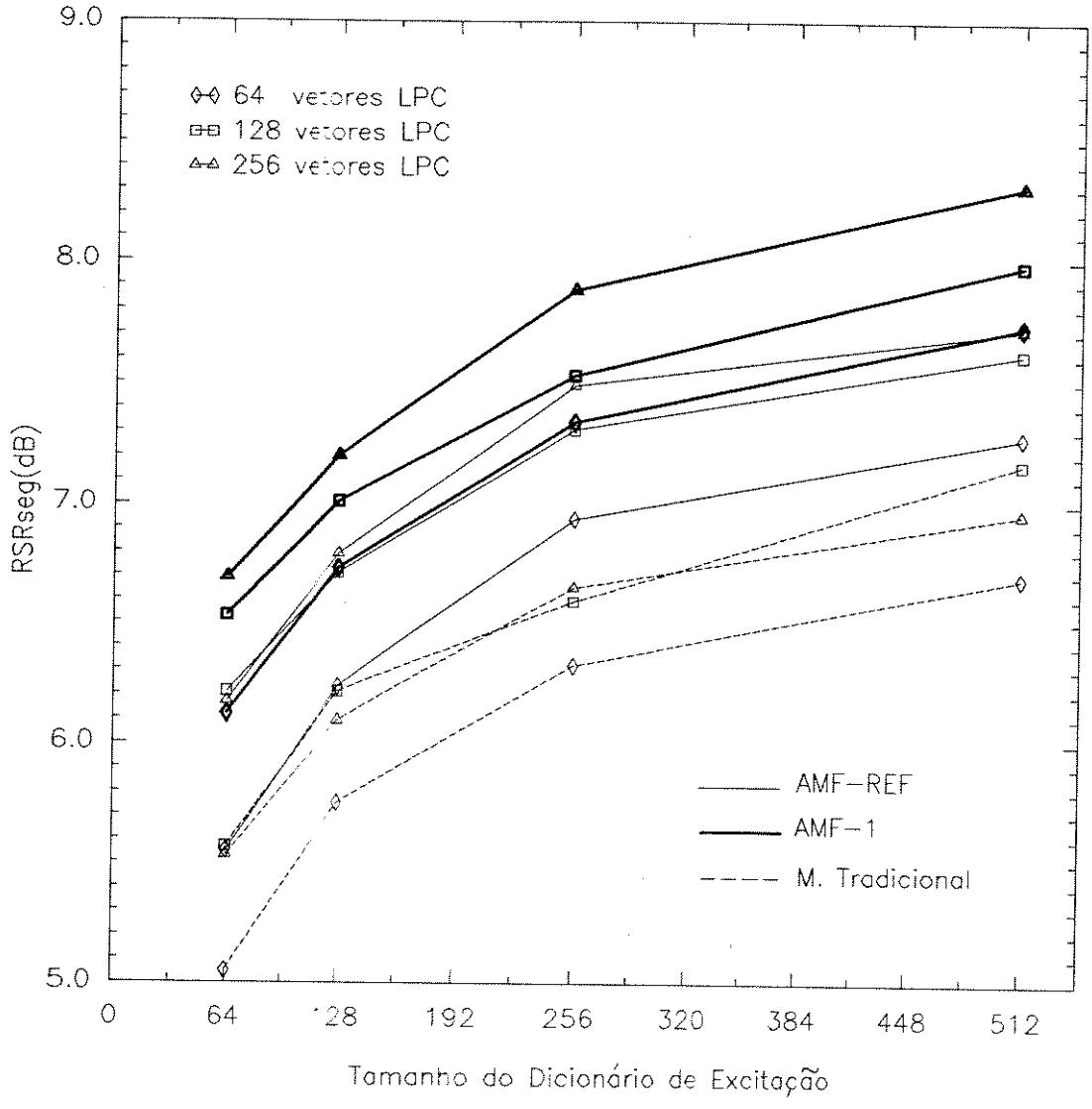

Figura 4.14: $RSR_{seg}(dB)$ para os algoritmos de quantização vetorial tradicional e AMF dos coeficientes LPC em função do tamanho dos dicionários de códigos LPC e de excitação utilizando-se sub-dicionários com 1/4 de vetores dos respectivos dicionários originais.

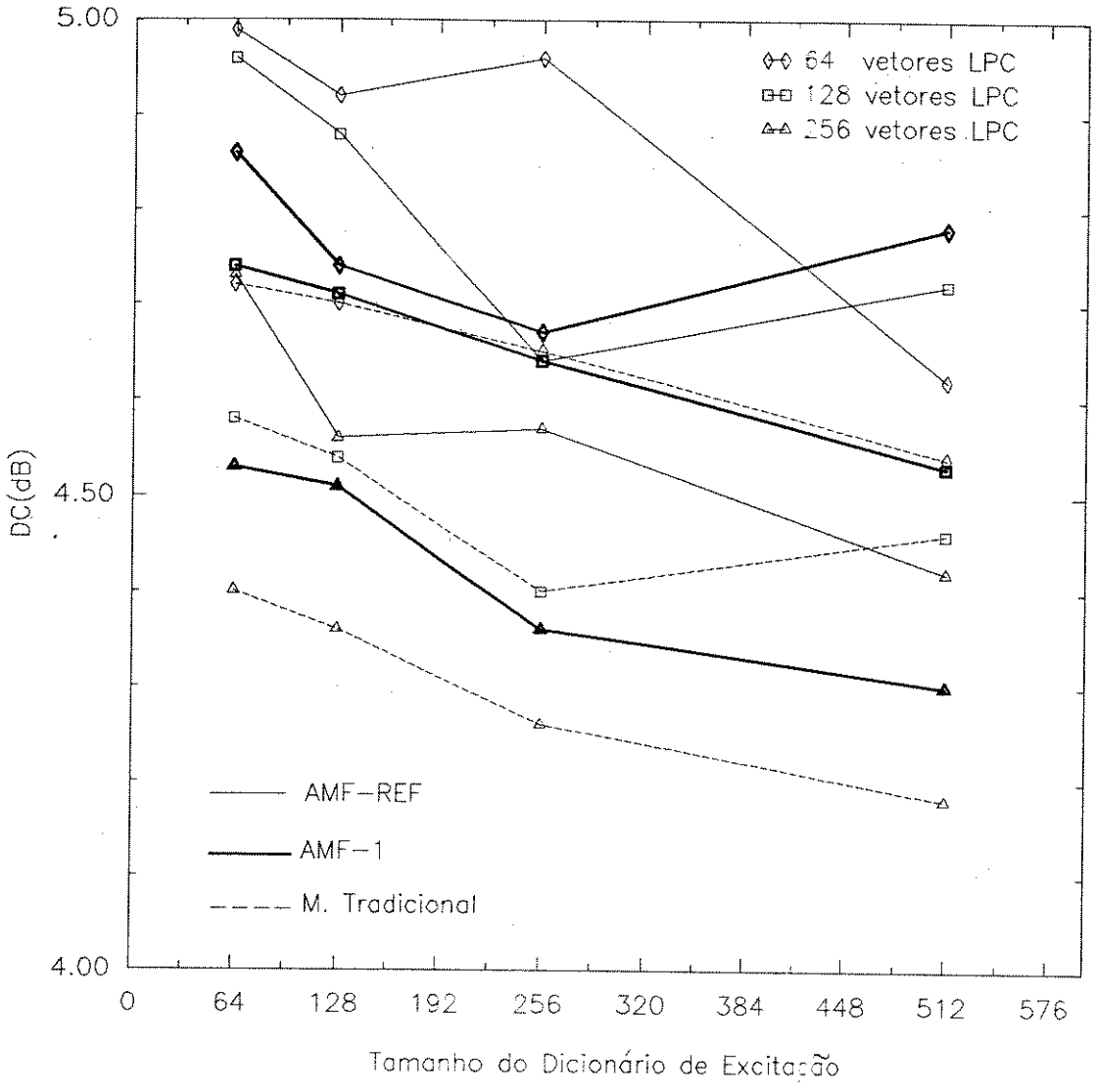

Figura 4.15: $DC(dB)$ para os algoritmos de quantização vetorial tradicional e AMF dos coeficientes LPC em função do tamanho dos dicionários de códigos LPC e de excitação utilizando-se sub-dicionários com 1/4 de vetores dos respectivos dicionários originais.

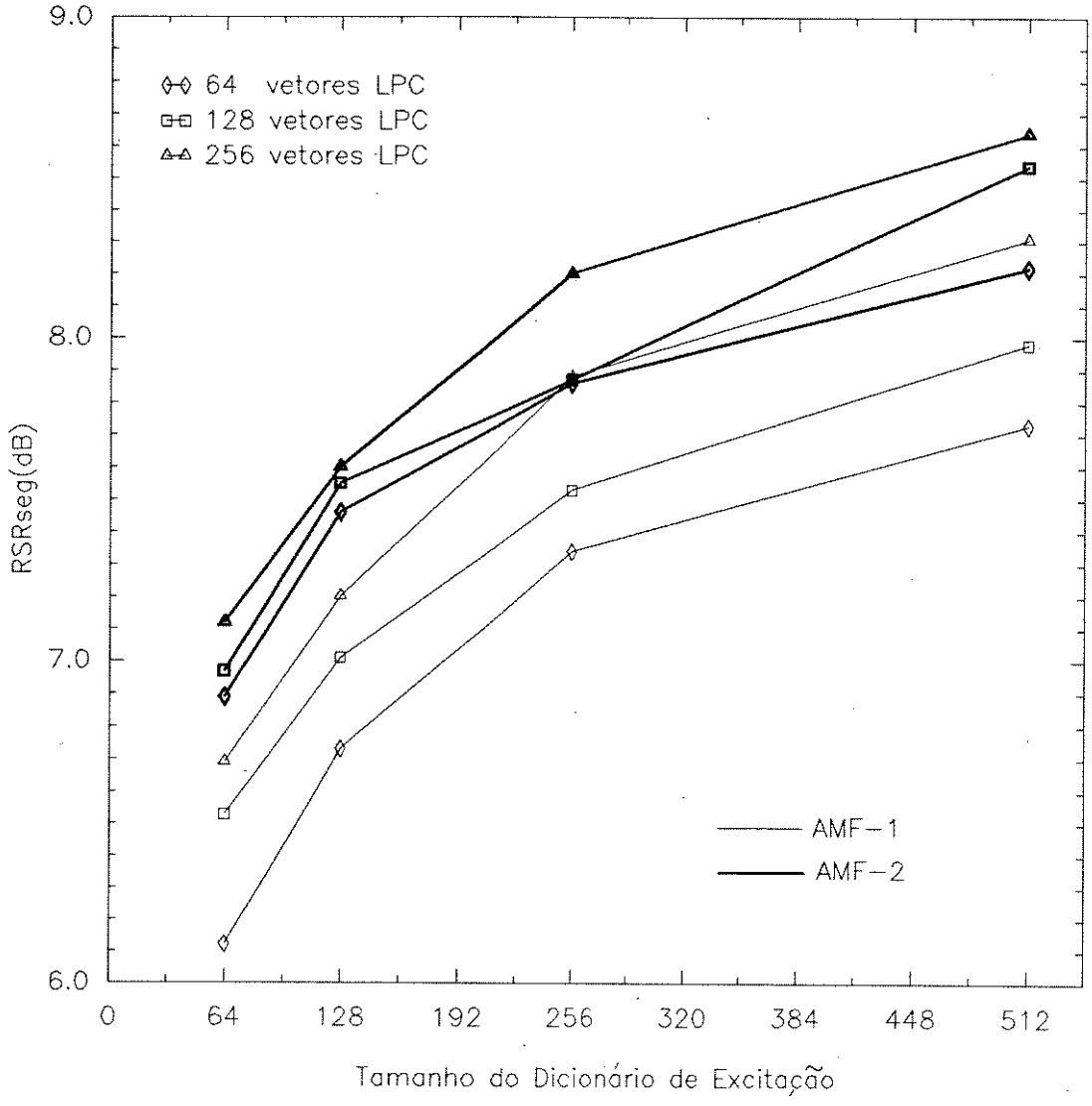

Figura 4.16: $RSR_{seg}(dB)$ para o algoritmo de quantização vetorial AMF dos coeficientes LPC com e sem alocação dinâmica de bits em função do tamanho dos dicionários de códigos LPC e de excitação utilizando-se sub-dicionários com 1/4 de vetores dos respectivos dicionários originais.

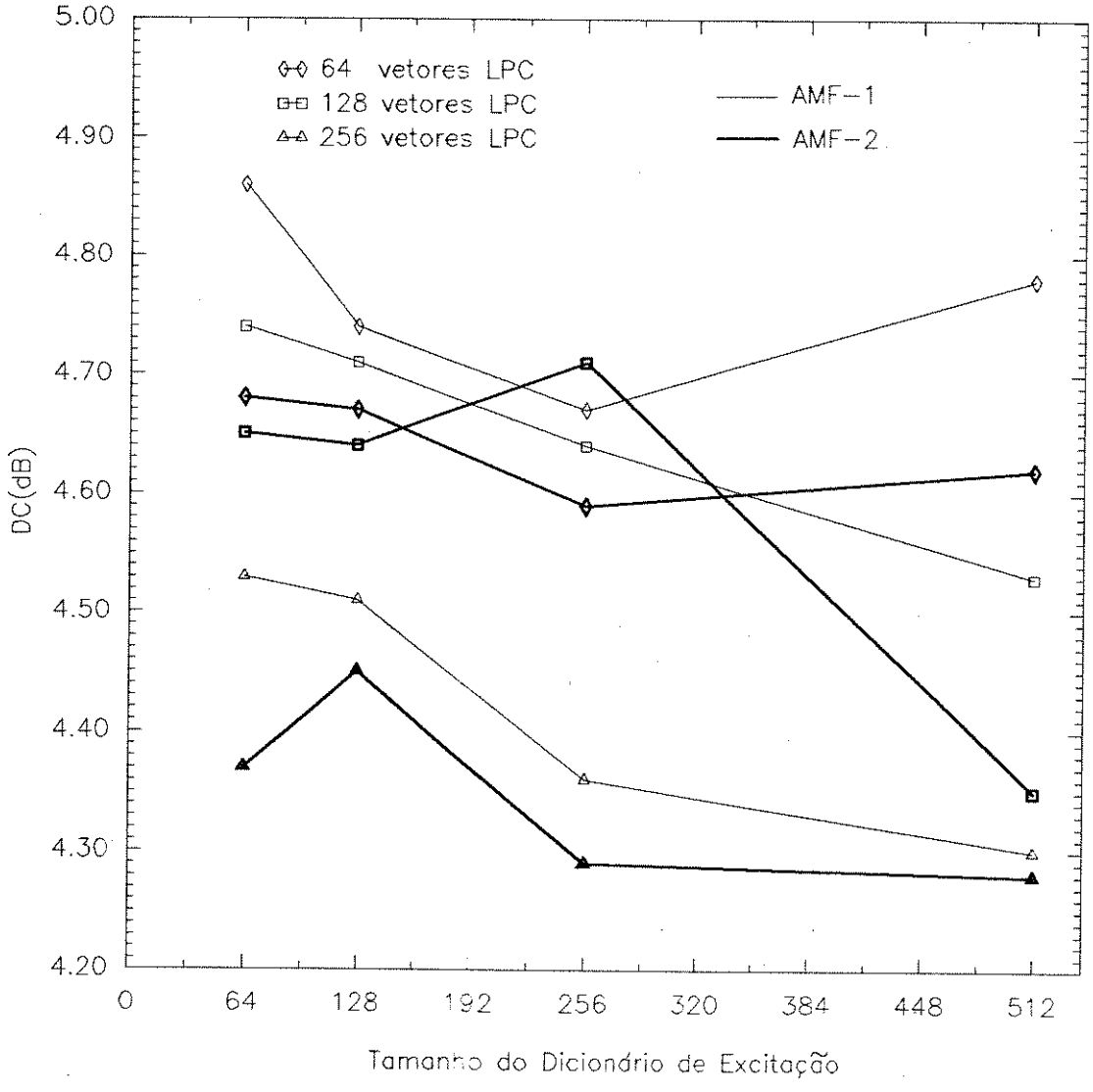

Figura 4.17: $DC(dB)$ para o algoritmo de quantização vetorial AMF dos coeficientes LPC com e sem alocação dinâmica de bits em função do tamanho dos dicionários de códigos LPC e de excitação utilizando-se sub-dicionários com 1/4 de vetores dos respectivos dicionários originais.

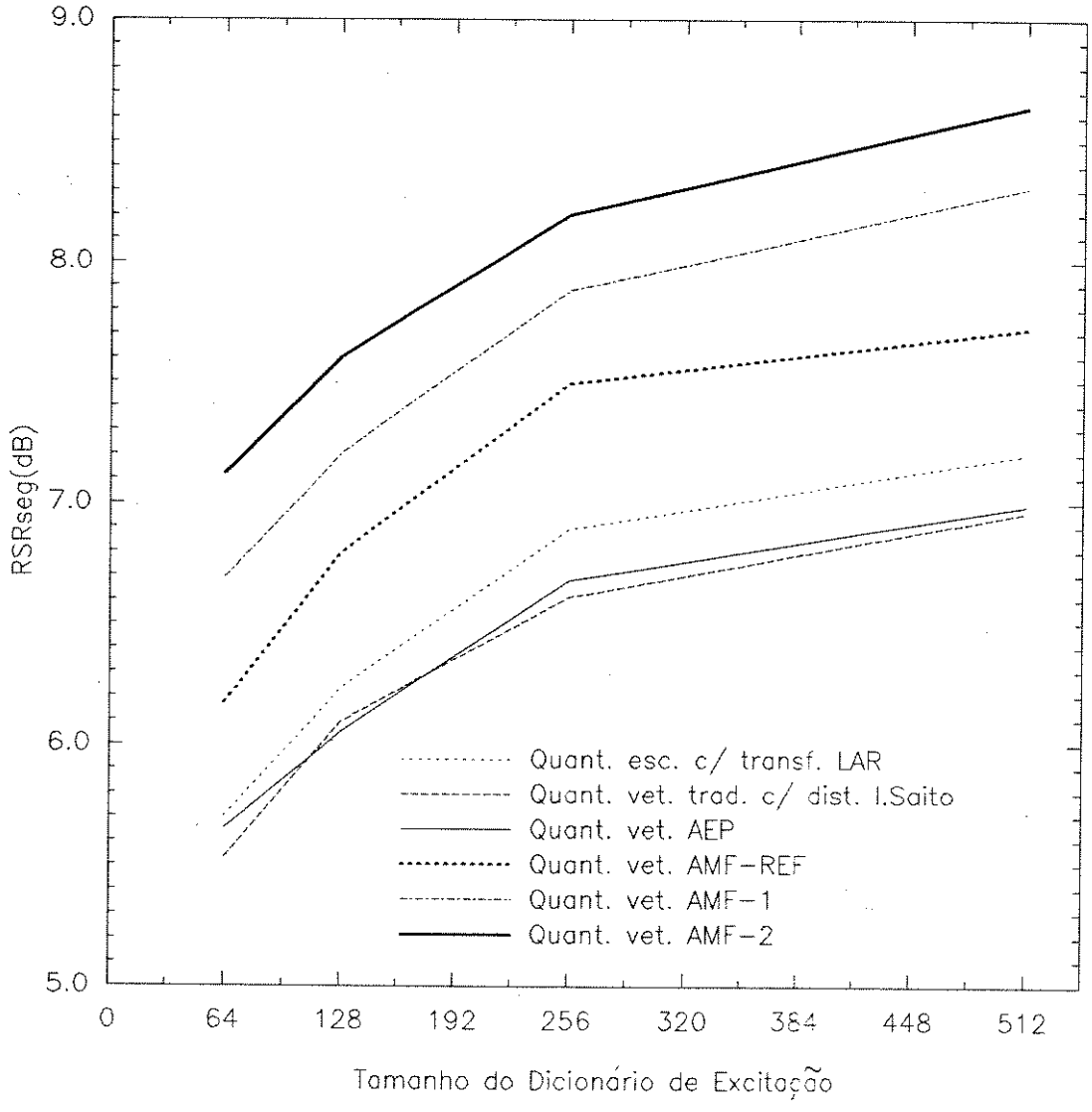

Figura 4.18: $RSR_{seg}(dB)$ para o algoritmo de quantização escalar dos coeficientes LAR e diversos algoritmos de quantização vetorial dos coeficientes LPC em função do tamanho do dicionário de códigos de excitação.

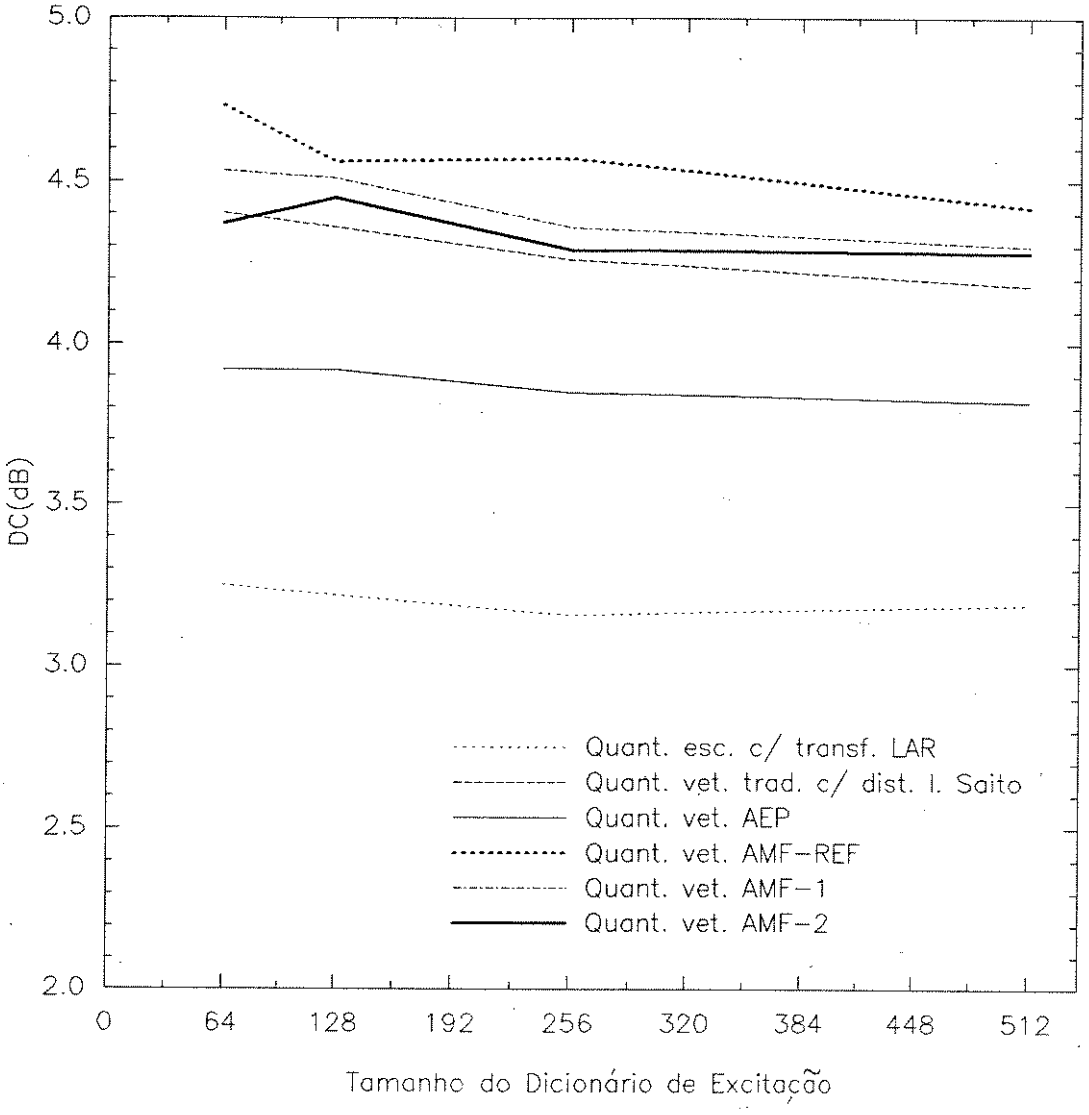

Figura 4.19: $DC(dB)$ para o algoritmo de quantização escalar dos coeficientes LAR e diversos algoritmos de quantização vetorial dos coeficientes LPC em função do tamanho do dicionário de códigos de excitação.

ritmos AEP e tradicional, mas apresenta um desempenho significativamente superior. Neste algoritmo, três casos de otimização conjunta são possíveis : i) coeficientes LPC e parâmetros de excitação a curto-prazo (AMF-REF); ii) coeficientes LPC e parâmetros de excitação de curto-prazo e longo-prazo(AMF-1); iii) caso ii) com alocação dinâmica de bits entre os coeficientes LPC e o sinal de excitação(AMF-2). Em qualquer dos casos, o aumento na RSR_{seg} é tal que compensa o pequeno aumento na DC resultando em uma qualidade do sinal de voz sintetizado significativamente superior que aquela do algoritmo tradicional usando a medida de distorção de Itakura Saito Modificada.

Nos algoritmos AEP e AMF, é importante ressaltar que a análise LPC não e feita no estilo convencional como no algoritmo de quantização vetorial tradicional que obriga o armazenamento de blocos de longa duração, aproximadamente 20 ms, de sinais de voz. A necessidade de armazenamento de blocos de sinais de voz durante a análise LPC ocasiona um atraso da mesma ordem de magnitude da duração do bloco ou, no caso de se considerar amostras passadas para reduzir o atraso, uma degradação de desempenho. Assim, no caso dos métodos AEP e AMF, em não havendo necessidade de armazenamento de blocos longos de sinais de voz, torna-os vantajosos para codificação de voz com baixo atraso. Em particular, com relação ao método AEP, foi mostrado que o seu desempenho melhora quando passa de um atraso de 20ms para 5ms.

Finalmente, ressalta-se que os dicionários de códigos LPC utilizados favorecem o algoritmo tradicional, pois o método LBG foi empregado comparando-se diretamente os coeficientes LPC e utilizando-se a medida de distorção de Itakura Saito modificada. Espera-se que resultados melhores que os apresentados neste trabalho sejam obtidos para os algoritmos AEP e AMF se for levado em conta, ao aplicar o método LBG, a estrutura de quantização vetorial daqueles algoritmos e a medida de distorção utilizada. A medida de distorção de Itakura Saito modificada não é necessariamente ótima em termos de minimização do ruído de quantização. Em outras palavras, na geração do dicionário de códigos LPC para o algoritmo AEP usando-se o método LBG, deveria ser utilizado o erro quadrático médio entre o sinal de voz original e sinal estimado durante o processo de agrupamento dos coeficientes LPC e determinação da centróide da célula. Analogamente, no algoritmo AMF deveria ser utilizado o erro quadrático médio ponderado entre o sinal de voz original e reconstruído para ambos os processos. Tanto no algoritmo AEP como no algoritmo AMF, equações específicas devem ser desenvolvidas para o cálculo

da centróide utilizando-se como medida de distorção o erro quadrático médio entre o sinal de voz original e estimado (algoritmo AEP) ou o erro quadrático médio ponderado entre o sinal de voz original e reconstruído (algoritmo AMF). Se, eventualmente, não for possível o desenvolvimento e emprego destas equações para o cálculo da centróide, poder-se-ia continuar empregando a equação 4.31, porém, às custas de uma degradação de desempenho dos algoritmos propostos.

Bibliografia

[1] R. Viswanathan, J. Makhoul, "Quantization Properties of Transmission Parameter in Linear Predictive Systems", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-23, nº 3, Junho de 1975, pág. 309-321.

[2] A.H. Gray,Jr., J. Markel, "Quantization and Bit Allocation in Speech Processing", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, nº 6, Dezembro de 1976, pág.459-473.

[3] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals", J. Acoust. Soc. Amm., vol. 57, 1975, pág. 535(A).

[4] N. Sugamura e F. Itakura,"Speech analysis and synthesis methods developed at ECL in NTT - From LPC to LSP,"Speech Commun., vol. 5, Junho de 1986, pág. 199-215.

[5] F.K. Soong e B.H. Juang, "Line spectrum pair (LSP) and speech data compression, IEEE Int. Conf. on Acoust., Speech, Signal Process., 1884, San Diego, CA, pág. 1.10.1-1.10.4.

[6] G.S. Kang e L.J. Fransen, "Application of line-spectrum pairs to low-bit-rate encoders", IEEE Int. Conf. on Acoust., Speech, Signal Process., 1985, Tampa, FL, pág. 244-247

[7] B.S. Atal, R.V. Cox e P. Kroon, "Spectral quantization and interpolation for CELP coders", IEEE Int. Conf. on Acoust., Speech, Signal Process., 1989, pág. 69-72.

[8] C.K. Un e S.C. Yang, "Piecewise Linear Quantization of LPC Reflection Coefficients", IEEE Int. Conf. on Acoust., Speech, Signal Process., Hartford, 1977, pág. 417-420.

[9] F.K. Soong e B.-H. Juang,"Optimal Quantization of LSP parameters", IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1988, pág. 394-397.

[10] N. Sugamura e N. Farvardin, "Quantizer Design in LSP Speech Analysis-Synthesis, IEEE Journal on Selected Areas in Communications, vol. 6, nº 2, Fevereiro de 1988, pág. 432-440

[11] S. Roucos, R. Schwartz e J. Makhoul, "Vector Quantization for Very-Low-Rate coding of Speech", IEEE Global Commun. Conf., Novembro de 1992, pág. 1074--1078.

[12] K.K. Paliwal e B.S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame", IEEE Int. Conf. on Acoust., Speech, Signal Process., Maio de 1991, pág. 661-664.

[13] B.-H. Juang, D.Y. Wong e A.H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding", IEEE Transactions on Acoust., Speech, and Signal Processing, vol. ASSP-30, nº 2, Abril de 1982, pág. 294-304.

[14] A. Buzo, A.H. Gray, Jr., R.M. Gray e J.D. Markel, "Speech Coding Based Upon Vector Quantization", IEEE Transactions on Acoust., Speech, and Signal Processing, vol. ASSP-28, nº 5, Outubro de 1980, pág. 562-574.

[15] J.A. Martins,"Vocoder LPC com Quantização Vetorial", Dissertação de Tese de Mestrado, DECOM/FEE/UNICAMP, Abril de 1991.

[16] J. Makhoul, S. Roucos e H. Gish, "Vector Quantization in Speech Coding", Proceedings of the IEEE, vol. 73, nº 11, Novembro de 1985, pág. 1551-1588.

[17] GSM Recommendation 06.10:GSM Full Rate Speech Transcoding 1988.

[18] Y. Linde, A. Buzo, R.M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, vol. COM-28, nº 1, Janeiro de 1980, pág. 84-95

[19] F. Tzeng,"Near Optimal Linear Predictive Speech Coding", IEEE Global Telecommun. Conf., 1990, pág. 962-966.

[20] W.P. LeBlanc, V. Cuperman, "Speech Coding at 4 and 8 kbit/s based on Iterative Sequential CELP Optimization", IEEE Global Telecommun. Conf., 1991, pág. 1874-1878.

[21] I.M. Trancoso, B.S. Atal, "Efficient Procedures for Finding the Optimum Innovation in Stochastics Coders", IEEE Int. Conf. on Acoust., Speech, and Signal Process., Maio de 1986, Tokyo, Japão, pág. 2375-2378.

Capítulo 5

REDUÇÃO DA TAXA DE BITS DO SINAL DE EXCITAÇÃO LPC

5.1 INTRODUÇÃO

A síntese de voz a baixas taxas usando técnicas LPC requer algoritmos eficientes de codificação tanto dos coeficientes LPC como do sinal de excitação. No capítulo 4, algoritmos de quantização vetorial conjunta dos coeficientes LPC e de excitação foram propostos. Entretanto, a taxa de bits para a codificação do sinal de excitação permanece relativamente elevada. Assim, por exemplo, no codificador CELP mostrado na figura 3.9, com quantização vetorial dos coeficientes LPC através de um dicionário de códigos de 512 vetores e operando a taxa de 3,5 kbit/s, aproximadamente 87% desta taxa corresponde ao sinal de excitação. Apesar disto, o número de bits utilizado não é suficiente para representar o sinal de excitação com precisão adequada de modo a sintetizar um sinal de voz de boa qualidade. Assim, a falta de uma representação eficiente do sinal de excitação continua sendo um grande obstáculo para uma sintetização de voz de boa qualidade a baixas taxas.

A alocação de bits para a representação do sinal de excitação é feita entre o período de pitch, ganho de pitch, índice e ganho do vetor de excitação. Com relação ao período e ganho de pitch, a alternativa que requer menor taxa de bits é a que faz a atualização destes parâmetros por bloco, embora a atualização por sub-bloco

resulte em melhor desempenho. O número de bits alocado para o índice do vetor de excitação depende do tamanho do dicionário e pode variar, dependendo da taxa de bits, de 5 a 12 bits por sub-bloco, enquanto que para a quantização escalar do ganho de excitação utilizam-se de 3 a 5 bits por sub-bloco. Assim, a taxa de bits necessária para a quantização do ganho de excitação representa uma porcentagem bastante alta da taxa de bits total do sinal de excitação, principalmente quando esta taxa é baixa.

Neste capítulo são propostos dois novos algoritmos para aumentar a eficiência de codificação do sinal de excitação num codificador CELP. O primeiro algoritmo é baseado na quantização vetorial de uma função de correlação-cruzada normalizada a partir da qual gera-se um sinal de excitação do tipo MP. A função de correlação--cruzada normalizada é enviada ao decodificador quantizada vetorialmente. Tendo--se esta informação, o sinal de excitação MP correspondente pode ser gerado no decodificador. O segundo algoritmo consiste numa alocação dinâmica de bits entre o ganho e índice do vetor ótimo de excitação, aproveitando-se da redundância do ganho de excitação entre sub-blocos sucessivos.

5.2 A FUNÇÃO DE CORRELAÇÃO-CRUZADA NORMALIZADA

A função de correlação-cruzada normalizada para a geração do sinal de excitação LPC foi proposta por Allen Gersho para a implementação de um codificador MPE- - LPC com todos os seus parâmetros quantizados vetorialmente [1]. Da equação (3.12), tem-se que no cálculo da excitação MP, as amplitudes A_i que minimizam o erro quadrático médio ponderado devem satisfazer as seguintes equações :

$$\sum_{i=1}^{N_p} A_i \Phi(m_j, m_i) = \beta_{m_j}, \qquad \begin{matrix} 1 \leq i, j \leq N_p \\ 0 \leq m_j, m_i < N \end{matrix} \tag{5.1}$$

onde :

$$\Phi(i,j) = \sum_{n=0}^{N-1} f(n-i) f(n-j), \qquad 0 \leq i, j < N \tag{5.2}$$

$$\beta_m = \sum_{n=0}^{N-1} d(n) f(n-m), \qquad 0 \leq m < N \tag{5.3}$$

A função de autocorrelação $\Phi(i,j)$ pode ser facilmente obtida do conjunto de coeficientes LPC do bloco de análise atual logo no início do processo de geração da

excitação. Calculando-se no codificador a função de correlação-cruzada normalizada, $R(m)$, definida como :

$$R(m) = \frac{\beta_m}{\Phi(0,0)}, \tag{5.4}$$

e enviando-a para o decodificador, o processo de geração da excitação MP pode ser feito no decodificador utilizando-se a função $R(m)$ e o conjunto de coeficientes LPC. O problema todo, portanto, reside em como enviar de modo eficiente a função $R(m)$ para o receptor.

A função $R(m)$ é uma forma de onda que poder ser manipulada de modo a conter o mesmo número de amostras que o próprio sinal de voz original ou o mesmo número de pulsos que a excitação MP. Entretanto, esta forma de onda possui uma faixa dinâmica menor devido à normalização e outras características que a tornam favorável para uma quantização vetorial com dimensão relativamente grande [1]. Por exemplo, o cálculo da função de autocorrelação de um sinal de voz, excitação MP ($p(m)$) e da própria função de autocorrelação normalizada $R(m)$, resulta em curvas como mostradas na figura 5.1. Observa-se que a função $R(m)$ apresenta valores de autocorrelação média maior do que para o sinal de excitação MP e sinal de voz original, principalmente para valores grandes do fator de ponderação μ. Isto mostra que a forma de onda da função $R(m)$ apresenta uma flutuação relativa menor que dos outros dois sinais.

Na figura 5.2 é mostrada a distribuição de probabilidade dos valores de autocorrelação de primeira ordem do sinal de voz original, excitação MP e da função $R(m)$ para diversos valores de μ. Para o caso da função $R(m)$, pode ser visto que quanto maior é o valor de μ, esta distribuição torna-se mais concentrada em torno de valores próximos de 1. Além disto, quando $\mu \geq 0,4$, a distribuição de probabilidade dos valores de autocorrelação para a função $R(m)$ é mais concentrada do que para o sinal de voz original e excitação MP. Todas estas características mostradas nas figuras 5.1 e 5.2 são favoráveis para a quantização vetorial da função de autocorrelação normalizada $R(m)$.

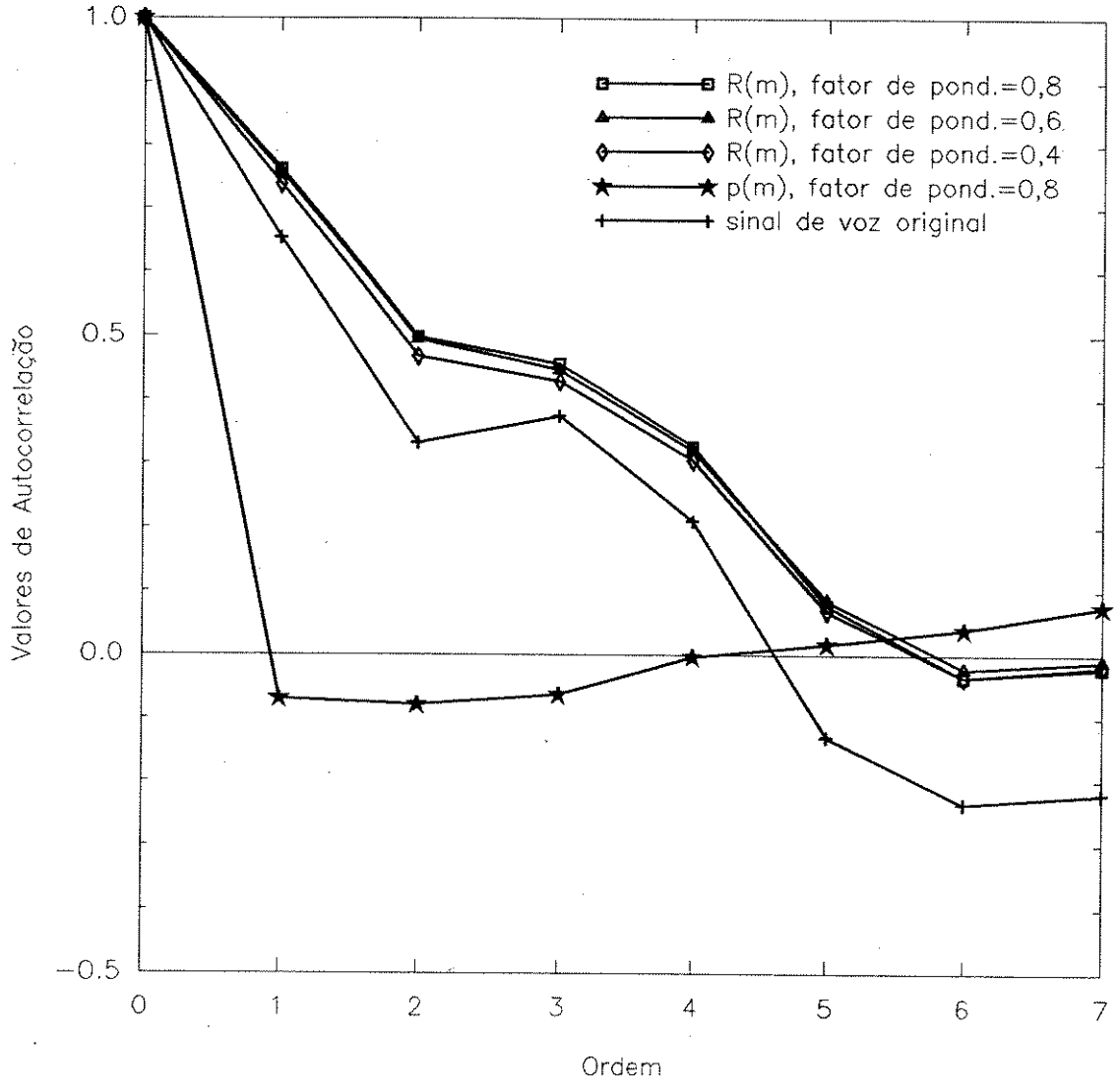

Figura 5.1: Valores de autocorrelação média em função da ordem.

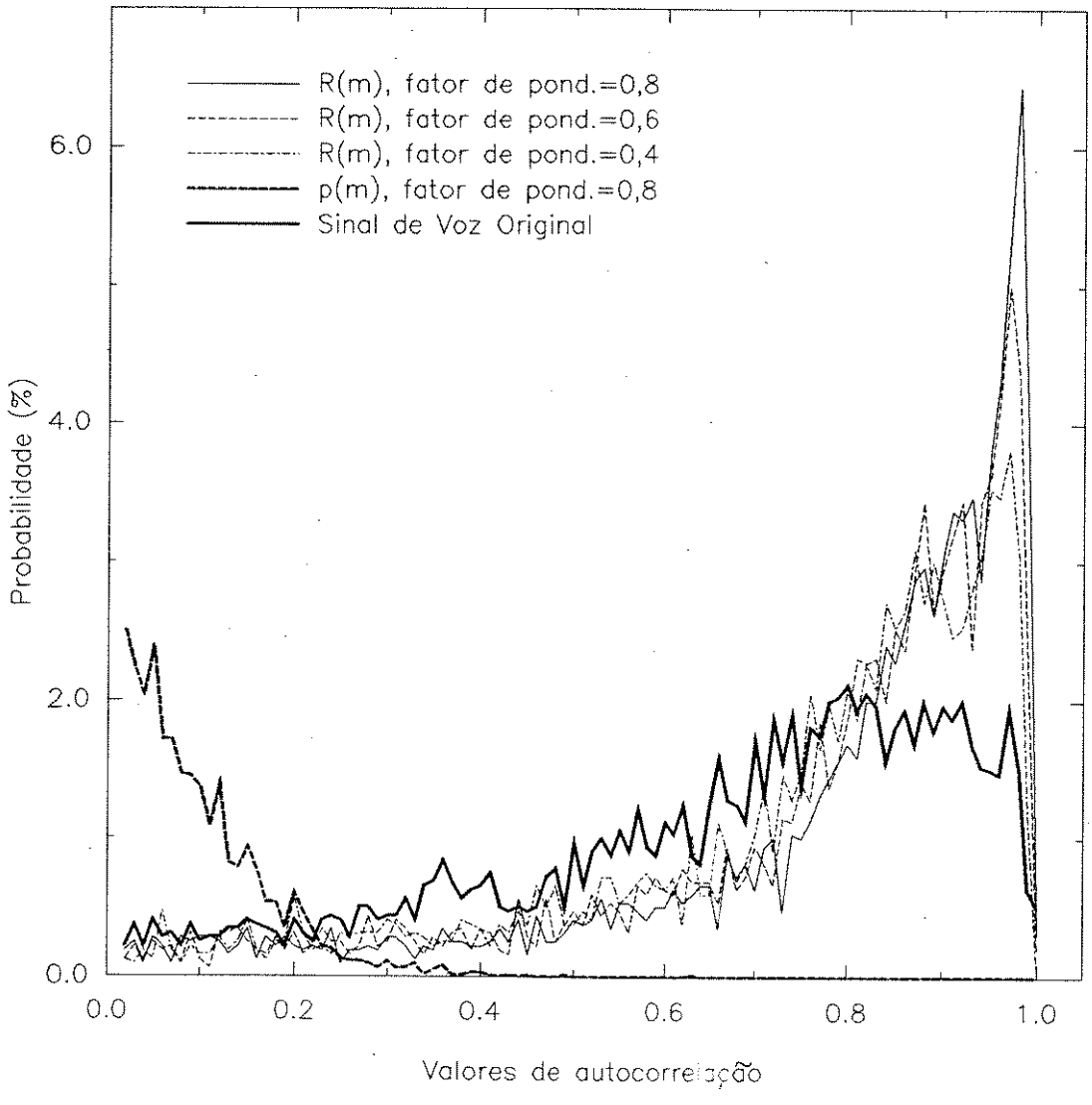

Figura 5.2: Distribuição de probabilidade dos valores de autocorrelação de primeira ordem.

5.3 PROJETO DO DICIONÁRIO DE CÓDIGOS DA FUNÇÃO DE CORRELAÇÃO-CRUZADA NORMALIZADA

O dicionário de códigos **R** da função de correlação-cruzada, $R(m)$, foi projetado utilizando-se o algoritmo LBG [2]. Uma seqüência de treinamento, consistindo de segmentos de 160 amostras da função $R(m)$, foi inicialmente gerada a partir do mesmo arquivo de voz empregado no projeto do dicionário de códigos dos coeficientes LPC. Posteriormente, cada segmento de 160 amostras de $R(m)$, calculado a cada bloco de análise LPC, foi dividido em 4 segmentos ou vetores menores de 40 amostras. Esta seqüência de treinamento foi usada pelo algoritmo LBG com um limiar de perturbação dos vetores códigos igual a 1%, e finalizando-se o processo de agrupamento quando a diferença entre distorções médias de duas iterações fosse menor do que 0,1%. Como medida de distorção no processo de agrupamento, foi empregado o erro quadrático médio, e a centróide $\hat{x}_i$ de cada célula C_i foi calculada como [2]:

$$\hat{x}_i = \frac{1}{K_i} \sum_{j:x_j \in C_i} x_j \tag{5.5}$$

onde : x_j são os vetores da seqüência de treinamento que estão na célula C_i;
K_i é o número de vetores na célula C_i.

5.4 CODIFICADOR MPE-QVR

5.4.1 Descrição Geral

Uma estrutura de codificador MPE-LPC com quantização vetorial da função $R(m)$, proposta por Gersho [1], consiste em calcular para cada sub-bloco de N amostras de sinal de voz, a função $R(m)$ e, posteriormente, aplicar a quantização vetorial através de uma comparação direta com vetores de correlação-cruzada normalizada de dimensão N de um dicionário de códigos **R**. Neste trabalho, é proposta uma nova estrutura de quantização vetorial onde a busca do vetor ótimo do dicionário **R** é realizada segundo um processo de análise por síntese. Um codificador de voz utilizando esta estrutura, denominado MPE-QVR (MPE com Quantização Vetorial da função $R(m)$) é mostrado na figura 5.3. Nesta estrutura de codificador, para cada vetor $\hat{R}(m)$ do dicionário de códigos **R**, calcula-se, inicialmente, a função de correlação-cruzada β_m conforme equação (5.4), onde o fator de normalização, $\phi(0,0)$, é obtido a partir dos coeficientes LPC quantizados do bloco de análise cor-

rente. Uma vez obtida a função de correlação-cruzada β_m, dois procedimentos de geração MP são possíveis :

- Os vetores do dicionário de códigos **R** contém amostras em todas as posições do sub-bloco de análise. Neste caso, o procedimento de geração da excitação MP inclui a determinação das posições dos pulsos e das amplitudes dos pulsos;
- Os vetores do dicionário de códigos **R** possuem as posições das amostras quantizadas, isto é, não possuem amostras em todas as posições. As posições das amostras dos vetores do dicionário **R** possuem a mesma distribuição de posições da excitação MP, de modo que durante a maximização da equação 3.13 somente estas posições são testadas para fazer a determinação seqüencial de amplitudes dos pulsos.

Por questões de complexidade e desempenho, o cálculo da excitação MP é feito sem a reotimização de amplitudes. Em termos de RSR_{seg}, constatou-se que o melhor desempenho com a reotimização de amplitudes, é obtido alocando-se 2 pulsos por sub-bloco de 40 amostras. Entretanto, a alocação de somente 2 pulsos a cada 40 amostras gera um ruído de fundo granular de baixa freqüência, auditivamente bastante perceptível, prejudicando a qualidade do sinal de voz sintetizado. Por outro lado, sem a reotimização de amplitudes, obtém-se uma RSR_{seg} equivalente ao caso com reotimização quando são alocados 5 pulsos por sub-bloco e o ruído de fundo torna-se menos perceptível.

O erro quadrático médio devido à excitação multipulso, $p_k(i)$, gerada a partir do k-ésimo vetor de correlação-cruzada é dado pela equação (4.20) substituindo-se $c_k(i)$ por $p_k(i)$. Além disto, para o caso do codificador MPE-QVR, tem-se que $l = 1$, resultando nas seguintes expressões :

$$\langle e_w^2 \rangle = \sum_{n=0}^{N-1} d^2(n) - 2g_k \sum_{n=0}^{N-1} d(n) \sum_{i=0}^{N-1} p_k(i) f(n-i) + g_k^2 (\varepsilon(0)\nu_k(0) + 2 \sum_{i=1}^{N-1} \varepsilon(i)\nu_k(i)) \tag{5.6}$$

$$g_k = \frac{\sum_{n=0}^{N-1} d(n) \sum_{i=0}^{N-1} p_k(i) f(n-i)}{\varepsilon(0)\nu_k(0) + 2 \sum_{i=1}^{N-1} \varepsilon(i)\nu_k(i)}, \tag{5.7}$$

onde : $\varepsilon(i) = \varepsilon_1(i)$

$\varepsilon_1(i)$ e $\nu_k(i)$ são dados pelas equações (4.18) e (4.19), respectivamente.

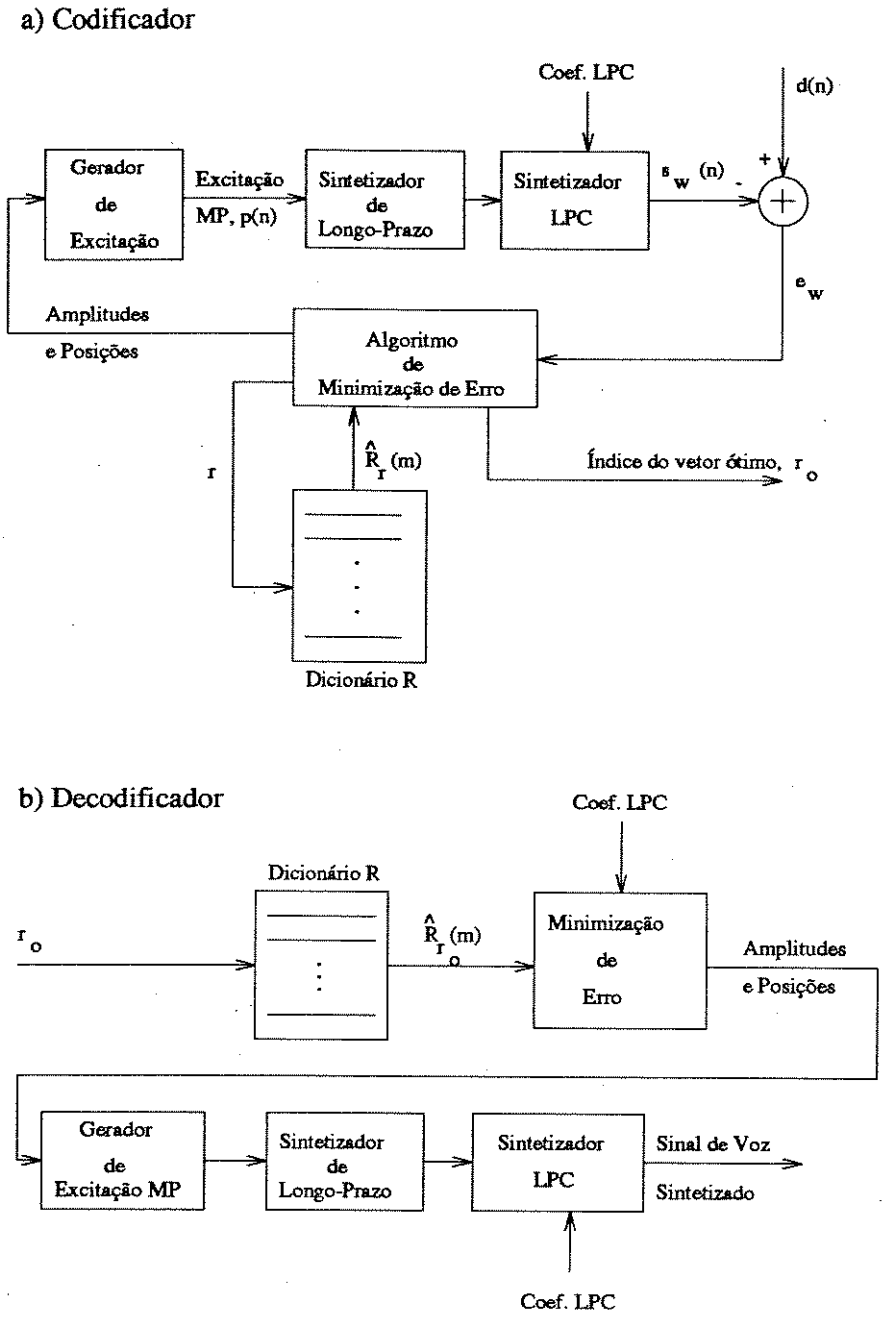

Figura 5.3: Estrutura do codificador e decodificador MPE-QVR.

5.4.2 Complexidade

A complexidade para a geração da excitação MP e escolha do vetor ótimo no codificador MPE-QVR, é igual a soma das complexidades de resolução das equações (3.18) e (5.6). Assim, para um sub-bloco de N amostras, dicionário de códigos de K vetores com w pulsos diferentes de zero e valores de $\nu_k(i)$ não nulos igual a λ, tem-se que esta complexidade é de aproximadamente $K(2w + \lambda + 1) + \frac{N^2+3N}{2}$ multiplicações/adições.

5.4.3 Desempenho

Na implementação do codificador MP-QVR realizada neste trabalho, os vetores do dicionário de códigos $\mathbf{R}$ contêm amostras em todas as posições e a excitação MP consiste de 5 pulsos por sub-bloco de 40 amostras. Nas figuras 5.4 e 5.5 são apresentados, respectivamente, os resultados de RSR_{seg} e RSR_{total} obtidos pelos codificadores CELP e MP-QVR em função da taxa de bits necessária para a transmissão dos parâmetros relativos à geração do sinal de excitação no decodificador. Observa-se que o codificador CELP apresenta valores de RSR_{seg} superiores em cerca de 1.8 dB em relação ao codificador MPE-QVR, de modo que o codificador MPE-QVR em si não apresenta vantagens em relação ao CELP. Entretanto, o codificador MPE-QVR pode ser combinado com o CELP, resultando no codificador $MPE - CELP$ que maximiza a RSR_{seg} mantendo a mesma taxa de bits.

5.5 CODIFICADOR MPE-CELP

5.5.1 Descrição Geral

O codificador MPE-QVR pode ser combinado com o codificador CELP numa única estrutura como mostrado na figura 5.6, resultando no codificador MPE-CELP, o qual utiliza um dicionário de excitação gaussiano ($\mathbf{G}$) e um outro de correlação-cruzada normalizada ($\mathbf{R}$). No codificador MPE-CELP, calcula-se o erro quadrático médio ponderado para cada um dos dois dicionário de código. Aquele dicionário que oferecer o menor erro é escolhido para a geração do sinal de excitação para o sub-bloco em análise. Assim, além do índice do vetor ótimo, deve ser transmitido o índice do dicionário escolhido (CE).

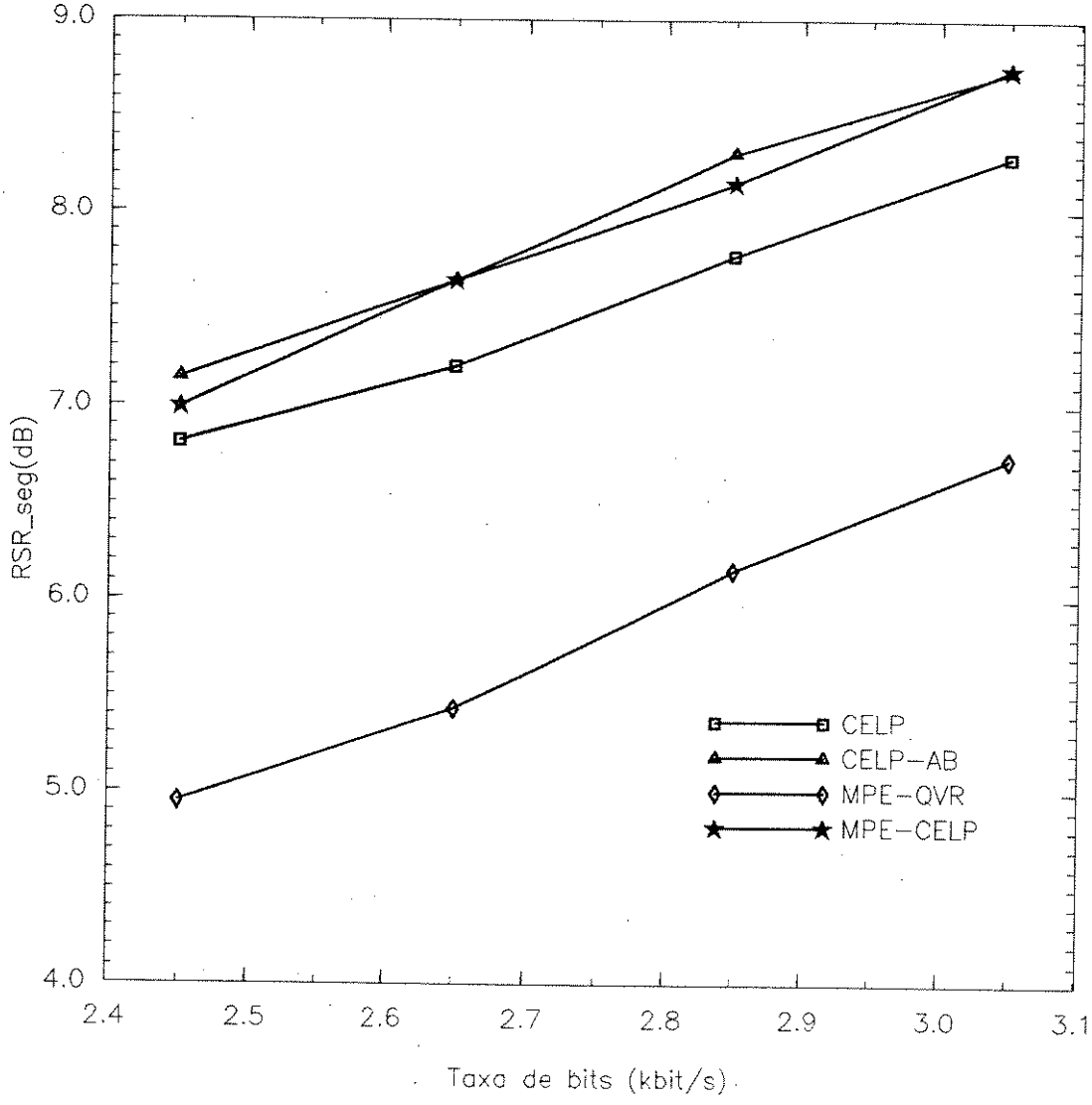

Figura 5.4: RSR_{seg} em função da taxa de bits do sinal de excitação para os codificadores CELP, MPE-VQR, CELP-AB e MPE-CELP.

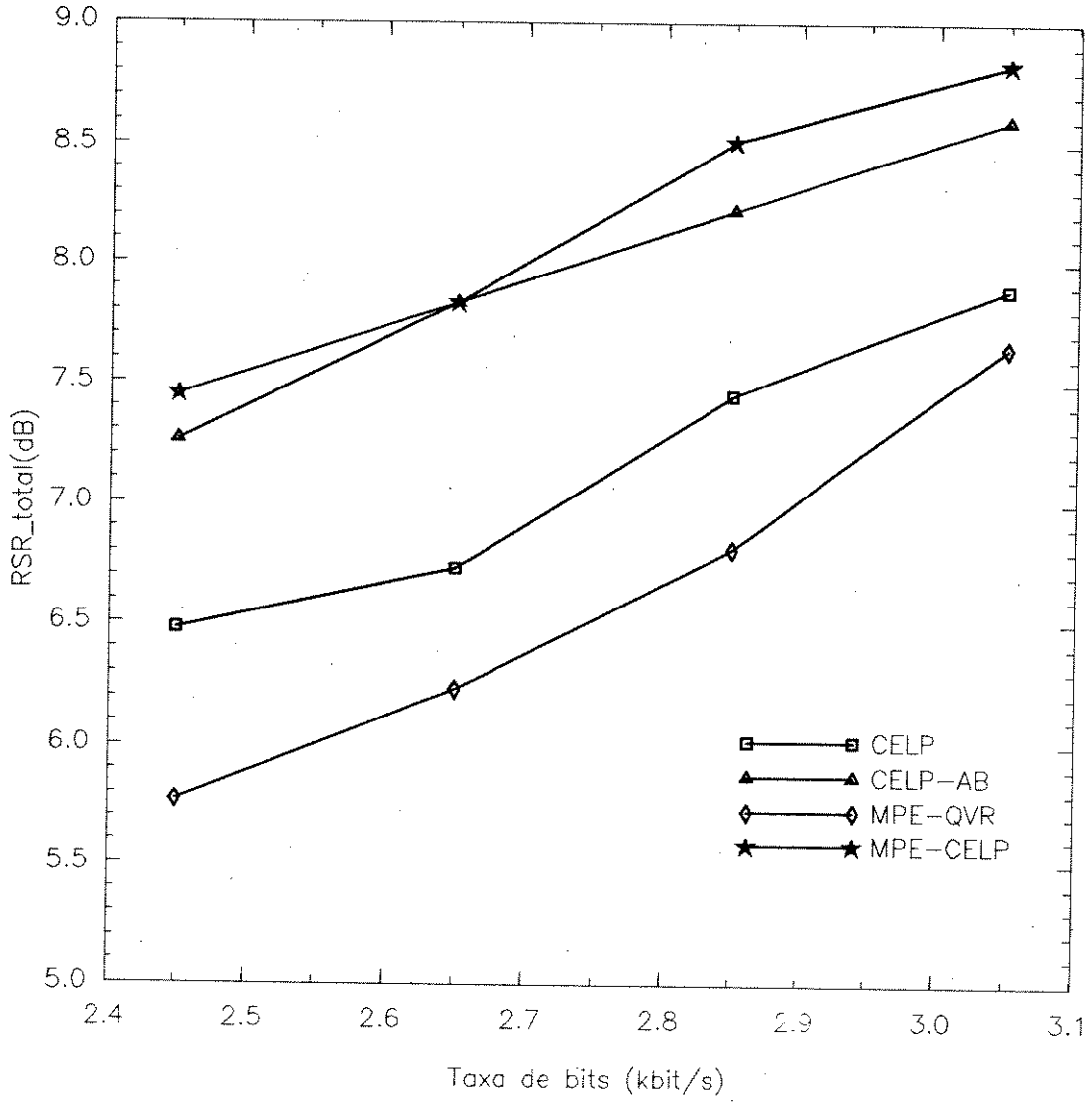

Figura 5.5: RSR_{total} em função da taxa de bits do sinal de excitação para os codificadores CELP, MPE-VQR, CELP-AB e MPE-CELP.

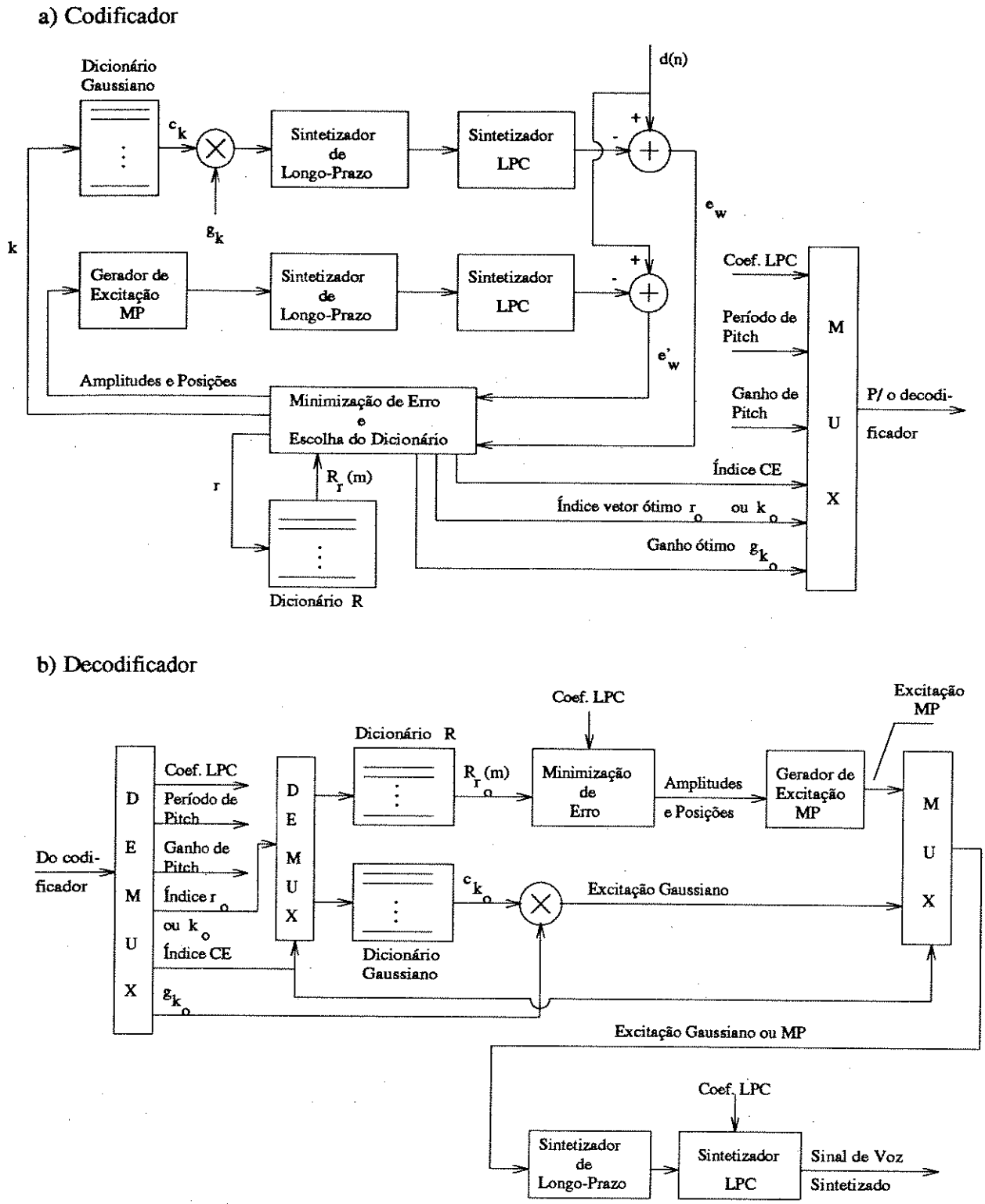

Figura 5.6: Estrutura do codificador e decodificador MPE-CELP.

5.5.2 Complexidade

A complexidade de busca do vetor ótimo e geração do sinal de excitação no codificador MPE-CELP é aproximadamente equivalente à soma das complexidades para o codificador CELP convencional e MPE-QVR. A complexidade para o CELP convencional é dado pela equação (4.20) para $l = 1$, isto é : $\frac{N(N+1)}{2} + Kw + (\lambda + 1)K + N$ multiplicações/adições. Esta complexidade somada à do codificador MPE-QVR, com exceção do termo correspondente ao cálculo da energia de $d(n)$, que é igual para os dois casos, resulta na seguinte complexidade para o codificador MPE-CELP : $K(3w + 2\lambda + 2) + N(N + 2)$ multiplicações/adições.

5.5.3 Desempenho

Conforme ilustrado na figura 5.7, o codificador MPE-CELP procura levar em conta as vantagens do codificador CELP para sons não-sonoros e do MPE-QVR para determinados trechos de sons sonoros. Observa-se que o codificador MPE-QVR normalmente apresenta uma RSR_{seg} inferior para trechos de voz não-sonoro, enquanto que para trechos sonoros ocorrem muitos segmentos onde a RSR_{seg} é superior em relação ao codificador CELP. Nestes trechos, o vetor ótimo é uma seqüência $R(m)$ a partir da qual é gerada uma excitação multipulso. Deste modo, o codificador MPE-CELP apresenta valores de RSR_{seg} e RSR_{total} de aproximadamente 0.6 e 1.0 dB, respectivamente, superiores aos do CELP convencional, conforme mostrado nas figuras 5.4. e 5.5.

5.6 ALOCAÇÃO DINÂMICA DE BITS ENTRE O GANHO E ÍNDICE DO VETOR DE EXCITAÇÃO

A quantização do ganho do sinal de excitação requer cerca de 4 bits por sub-bloco. Para sub-blocos de comprimento $N = 40$ amostras (5ms), este número de bits corresponde a 800 bit/s, que é uma taxa relativamente elevada quando trata-se de codificação de voz a baixas-taxas. Para diminuir esta taxa ainda mantendo-se o desempenho, é proposto neste trabalho um codificador CELP com alocação dinâmica de bits entre o ganho e índice do vetor de excitação, denominado CELP-AB. O codificador CELP-AB aproveita a redundância dos valores de ganho quantizados entre sub-blocos sucessivos para alocar um maior número de bits ao índice do vetor de excitação e, assim, melhorar o desempenho sem aumentar a taxa de bits ou diminuir a taxa de bits mantendo-se o desempenho. Através de simulações realizadas

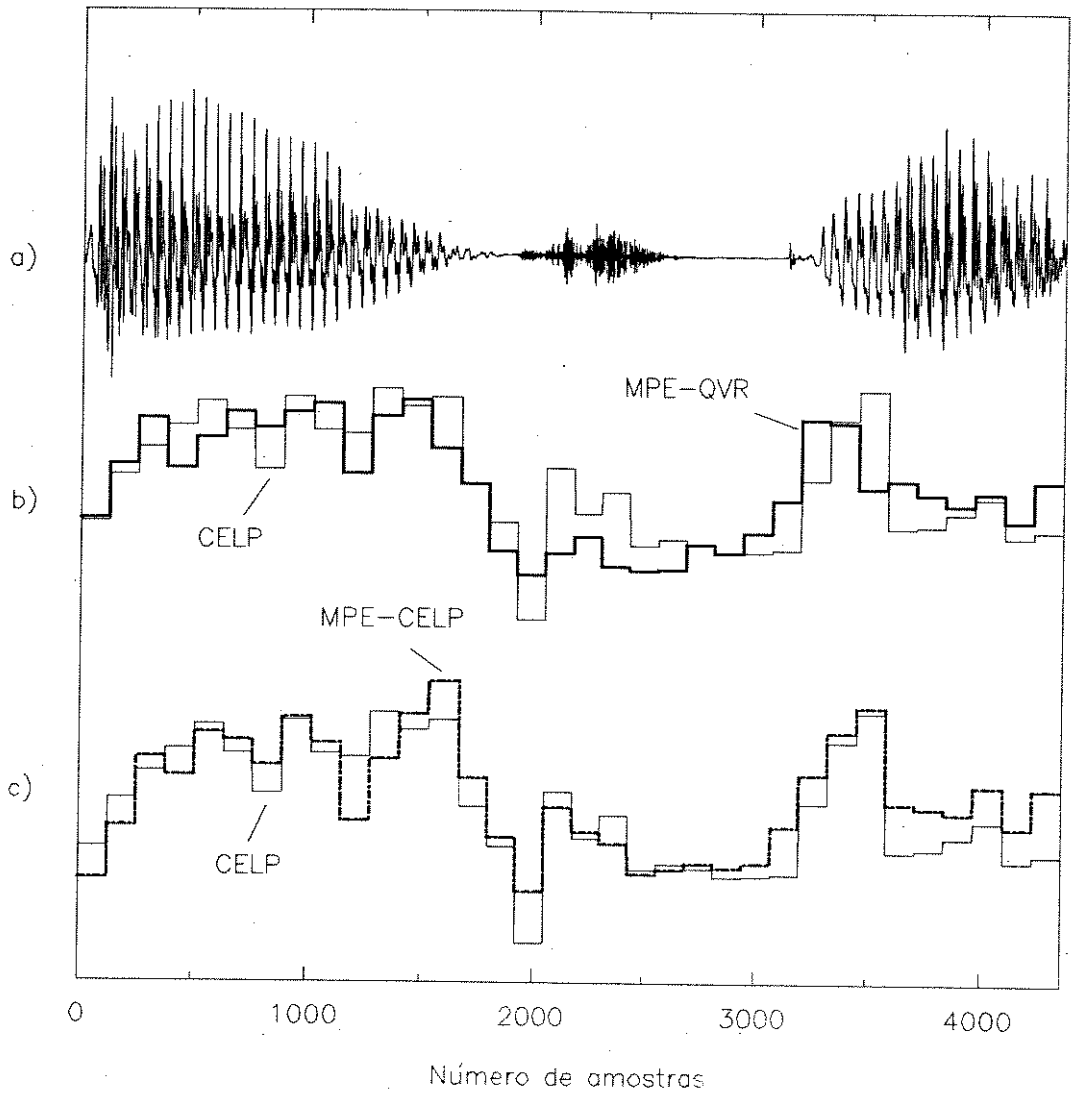

Figura 5.7: RSR_{seg} ao longo de um trecho de sinal de voz para os codificadores CELP, MPE-QVR e MPE-CELP

neste trabalho, constatou-se que esta redundância é de aproximadamente 25% para um quantizador de 4 bits (3 bits para a magnitude + 1 bit de sinal). No codificador CELP-AB implementado, são utilizados 3 bits (2 bits de magnitude + 1 bit de sinal) para a quantização propriamente dita do ganho de excitação e mais um bit para indicar a existência de redundância. Se o valor do ganho de excitação do sub-bloco corrente é igual ao anterior, então os 3 bits que seriam utilizados para a transmissão do ganho são alocados para a transmissão do índice do vetor ótimo de excitação. A alocação de 3 bits adicionais permite a utilização de um dicionário de excitação de tamanho 8 vezes maior do que quando os ganhos de sub-blocos subsequentes são diferentes. Esta alocação dinâmica de bits permite um aumento de até 0.6 e 1.0 dB de RSR_{seg} e RSR_{total}, respectivamente, conforme ilustrado nas figuras 5.4 e 5.5.

5.7 CONCLUSÕES

Dois algoritmos para diminuir a taxa de bits de transmissão dos parâmetros de excitação de curto-prazo num codificador CELP, mantendo-se ainda o seu desempenho, foram investigados :i) utilização da função $R(m)$ para geração de um sinal de excitação MP alternativo ao sinal de excitação gerado a partir do dicionário de códigos gaussiano; ii) alocação dinâmica de bits entre o ganho e índice do vetor de excitação. Destes dois algoritmos, o segundo mostra-se mais vantajoso sob o ponto de vista de complexidade. A utilização da função $R(m)$ somente, como no caso do codificador MPE-QVR, resulta num desempenho inferior ao do codificador CELP convencional.

Bibliografia

[1] H. Koyama e A. Gersho,"Fully Vector-Quantized Multipulse LPC at 4800 bps", IEEE Int. Conf. Acoust., Speech, Signal Process., Maio de 1986, Tokyo, Japão, pág. 445-448.

[2] Y. Linde, A. Buzo, R.M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, vol. COM-28, nº 1, Janeiro de 1980, pág. 84-95

Capítulo 6

EXEMPLOS DE CODECS A BAIXAS TAXAS E BAIXO ATRASO

6.1 INTRODUÇÃO

Neste capítulo são apresentados os seguintes codecs de voz implementados a partir da integração dos algoritmos de quantização dos coeficientes LPC e do sinal de excitação descritos, respectivamente, nos capítulos 4 e 5 :

- Codecs à taxa de bits entre 3,45 e 3,55 kbit/s :
  - CELP : codificador CELP com os coeficientes LPC quantizados escalarmente conforme descrito na seção 4.5.2;
  - CELP-AEP : codificador CELP com os coeficientes LPC quantizados vetorialmente pelo algoritmo AEP-COV (seção 4.3.2);
  - CELP-IS : codificador CELP com os coeficientes LPC quantizados vetorialmente pelo algoritmo tradicional usando medida de distorção de Itakura Saito Modificada;
  - CELP-AMF : codificador CELP-AB com quantização vetorial conjunta dos coeficientes LPC e sinal de excitação pelo algoritmo AMF-2 (seção 4.4.4);
  - MCELP-AMF : codificador MPE-CELP com quantização vetorial conjunta dos coeficientes LPC e sinal de excitação pelo algoritmo AMF-2.

- Codecs de baixo atraso à taxa de 6,8 kbit/s :
  - CELP-AEP-BA : codificador CELP com os coeficientes LPC quantizados vetorialmente pelo algoritmo AEP-COV a cada sub-bloco;
  - CELP-IS-BA : codificador CELP com os coeficientes LPC quantizados vetorialmente pelo algoritmo tradicional usando medida de distorção de Itakura Saito Modificada a cada sub-bloco.

Todos os codificadores CELP a 3,45-3,55 kbit/s implementados possuem a mesma estrutura mostrada na figura 3.9, diferindo entre si com relação aos algoritmos empregados na quantização vetorial dos coeficientes LPC e sinal de excitação. O codificador MCELP-AMF resulta da incorporação do algoritmo de quantização vetorial AMF-2 na estrutura do codificador MPE-CELP mostrada na figura 5.6. Os codificadores de baixo atraso a 6,8 kbit/s são obtidos realizando-se a quantização vetorial dos coeficientes LPC pelo algoritmo AEP por covariância a cada sub-bloco.

6.2 CARACTERÍSTICAS DOS CODECS À TAXA DE BITS ENTRE 3,45 E 3,55 KBIT/S.

Os codecs à taxa de bits entre 3,45 e 3,55 kbit/s possuem as seguintes características de parâmetros :

Codec CELP

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho (nº de vetores) 4
  - Dimensão 40
- Comprimento de bloco para análise LPC ..... 160 amostras
- Nº de sub-blocos para excitação LPC 4
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

Codec CELP-AEP

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho 512
  - Dimensão 40
- Dicionário LPC :
  - Tipo ..... Coef. de Reflexão
  - Tamanho 512
  - Dimensão 8
- Comprimento de bloco para análise LPC ..... 160 amostras
- Nº de sub-blocos para excitação LPC 4
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

Codec CELP-IS

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho 512
  - Dimensão 40
- Dicionário LPC :
  - Tipo ..... Coef. LPC
  - Tamanho 512
  - Dimensão 8
- Comprimento de bloco para análise LPC ..... 160 amostras
- Nº de sub-blocos para excitação LPC 4
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

Codec CELP-AMF

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho ..... 512 a 4096
  - Dimensão 40
- Sub-dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho 32
  - Dimensão 40
- Dicionário LPC :
  - Tipo ..... Coef. de Reflexão
  - Tamanho 512
  - Dimensão 8
- Sub-dicionário LPC :
  - Tipo ..... Coef. de Reflexão
  - Tamanho 32
  - Dimensão 8
- Tamanho da Lista de Per. de Pitch Candidatos 4
- Comprimento de bloco para análise LPC ..... 160 amostras
- Nº de sub-blocos para excitação LPC 4
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

Codec MCELP-AMF

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho ..... 512 a 2048
  - Dimensão 40
- Dic. de Correlação-Cruz. Normalizada :
  - Tamanho ..... 512 a 2048
  - Dimensão 40
- Sub-dicionário de Exc./Corr.-Cruz.:
  - Tamanho 32
  - Dimensão 40
- Dicionário LPC :
  - Tipo ..... Coef. de Reflexão
  - Tamanho 512
  - Dimensão 8
- Sub-dicionário LPC :
  - Tipo ..... Coef. de Reflexão
  - Tamanho Máximo 32
  - Dimensão 8
- Tamanho da Lista de Per. de Pitch Candidatos 4
- Comprimento de bloco para análise LPC ..... 160 amostras
- Nº de sub-blocos para excitação LPC 4
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

Em todos os codecs a 3,45-3,55 kbit/s, o bloco de análise LPC é de 160 amostras. Assim, estes codecs apresentam um atraso de codificação de cerca de 20 ms.

O sub-dicionário de excitação gaussiana/correlação-cruzada normalizada é constituída tanto por vetores de excitação gaussianos como por vetores de cor-

relação-cruzada num total de até 32 vetores. Deste total de vetores, a relação entre o número de vetores gaussianos e de correlação-cruzada normalizada depende do trecho do sinal de voz considerado, podendo existir mais vetores gaussianos do que de correlação-cruzada e vice-versa.

Nas tabelas 5.1 a 5.4 é apresentada a distribuição de bits entre os diversos parâmetros dos codecs a 3,45-3,55 kbit/s.

Tabela 5.1:Distribuição de bits por parâmetro do codec CELP

Parâmetros	*Nº de bits por sub-bloco*	Nº de bits por bloco
Coeficientes LPC		36
Período de Pitch		7
Ganho de Pitch		2
Ind. Vetor de Exc.	2	
Ganho Vetor de Exc.	4	
Total de bits	6	45

Considerando-se o tamanho do bloco de 160 amostras (20 ms) e sub-bloco de 40 amostras (5ms), tem-se que a taxa de bits total do codec é dada por :

$$\text{Taxa de bits total} = 6/5 + 45/20 = 3,45 \text{ kbit/s} \tag{6.1}$$

Tabela 5.2:Distribuição de bits por parâmetro dos codecs CELP-AEP e CELP-IS

Parâmetros	*Nº de bits por sub-bloco*	Nº de bits por bloco
Índice vetor LPC		9
Período de Pitch		7
Ganho de Pitch		2
Ind. Vetor de Exc.	9	
Ganho Vetor de Exc.	4	
Total de bits	13	18

$$\text{Taxa de bits total} = 13/5 + 18/20 = 3.5 \text{ kbit/s} \tag{6.2}$$

Tabela 5.3:Distribuição de bits por parâmetro do codec CELP-AMF

Parâmetros	*Nº de bits por sub-bloco*				Nº de bits por bloco	
	s/ rep. vetor LPC		c/ rep. vetor LPC			
	s/ rep. g. exc.	c/ rep. g. exc.	s/ rep. g. exc.	c/ rep. g. exc.	s/ rep. vetor LPC	c/ rep. vetor LPC
Índice vetor LPC					9	
Red. vetor LPC					1	1
Peridod de Pitch					7	7
Ganho de Pitch					2	2
Índice Vetor Exc.	9	12	11	12		
Ganho Vetor Exc.	3		3			
Red. Ganho Vetor Exc.	1	1	1	1		
Total de bits	13	13	15	13	19	10

$$\text{Taxa de bits total} = \begin{cases} 13/5 + 19/20 = 3,55 \text{ kbit/s}, & \text{ou} \\ 15/5 + 10/20 = 3,50 \text{ kbit/s}, & \text{ou} \\ 13/5 + 10/20 = 3,1 \text{ kbit/s} \end{cases} \tag{6.3}$$

Um melhor desempenho para o codificador CELP-AMF pode ser obtido mantendo-se ainda uma taxa de bits menor ou igual a 3,55 kbit/s. Para tanto, basta que seja feita uma alocação de até 14 bits para o índice do vetor de excitação quando ocorrerem simultaneamente uma repetição do vetor LPC e ganho de excitação entre blocos e sub-blocos sucessivos, respectivamente. Entretanto, por questões de complexidade, limitou-se em 12 o número máximo de bits a ser alocado para o índice do vetor de excitação.

Tabela 5.4:Distribuição de bits por parâmetro do codec MCELP-AMF

Parâmetros	*Nº de bits por sub-bloco*		Nº de bits por bloco	
	s/ rep. vetor LPC	c/ rep. vetor LPC	s/ rep. vetor LPC	c/ rep. vetor LPC
Índice Vetor LPC			9	
Red. vetor LPC			1	1
Peridod de Pitch			7	7
Ganho de Pitch			2	2
Índice Vetor Exc./Corr.-Cruz.	9	11		
Seleção Vetor Exc./Corr.-Cruz.	1	1		
Ganho Vetor Exc./Corr.-Cruz.	3	3		
Total de bits	13	15	19	10

$$\text{Taxa de bits total} = \begin{cases} 13/5 + 19/20 = 3,55 \text{ kbit/s}, & \text{ou} \\ 15/5 + 10/20 = 3,5 \text{ kbit/s}, \end{cases} \tag{6.4}$$

6.3 CARACTERÍSTICAS DOS CODECS DE BAIXO ATRASO A 6,8 KBIT/S

Os codecs de baixo-atraso a 6,8 kbit/s, CELP-AEP-BA e CELP-IS-BA, possuem as seguintes características de parâmetros :

CELP-AEP-BA

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho 4096
  - Dimensão 40
- Dicionário LPC :
  - Tipo ..... Coef. de Reflexão
  - Tamanho 512
  - Dimensão 8
- Comprimento de bloco para análise LPC ..... 40 amostras
- Nº de sub-blocos para excitação LPC 1
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

CELP-IS-BA

- Dicionário de Excitação :
  - Tipo ..... Gaussiano
  - Tamanho 4096
  - Dimensão 40
- Dicionário LPC :
  - Tipo ..... Coef. LPC
  - Tamanho 512
  - Dimensão 8
- Comprimento de bloco para análise LPC ..... 40 amostras
- Nº de sub-blocos para excitação LPC 1
- Ordem do filtro de síntese LPC 8
- Ordem do filtro de síntese de longo-prazo 1

Em todos os codecs a 6,8 kbit/s, a análise LPC é feita por sub-blocos de 40 amostras, resultando em codecs com atraso de aproximadamente 5 ms.

Na tabelas 5.5 é apresentada a distribuição de bits entre os diversos parâmetros dos codecs de baixo atraso CELP-AEP-BA e CELP-IS-BA à 6.8 kbit/s.

Tabela 5.5:Distribuição de bits por parâmetro dos codecs CELP-AEP-BA e CELP-IS-BA

Parâmetros	*Nº de bits por sub-bloco*
Índice vetor LPC	9
Período de Pitch	7
Ganho de Pitch	2
Ind. Vetor de Exc.	12
Ganho Vetor de Exc.	4
Total de bits	34

$$\text{Taxa de bits total} = 34/5 = 6,8\text{kbit/s} \tag{6.5}$$

6.4 PÓS-FILTRAGEM ADAPTATIVA

Para melhorar a qualidade perceptual do sinal de voz reconstruído, tanto os codecs a 3,45-3,55 kbit/s como os de baixo atraso a 6,8 kbit/s, incluem no decodificador um pós-filtro adaptativo [1-6]. O pós-filtro adaptativo ajusta a resposta em frequência de modo que os vales no espectro do sinal de voz reconstruído sejam atenuados enquanto que os picos são acentuados. O efeito auditivo de tal filtro é a de reduzir a perceptividade ao ruído de quantização, embora também introduza uma distorção causada pela redução na largura de faixa dos picos no espectro.

O pós-filtro utilizado neste trabalho é baseado em [5] e [6]. Um diagrama em blocos simplificado deste pós-filtro é mostrado na figura 6.1.

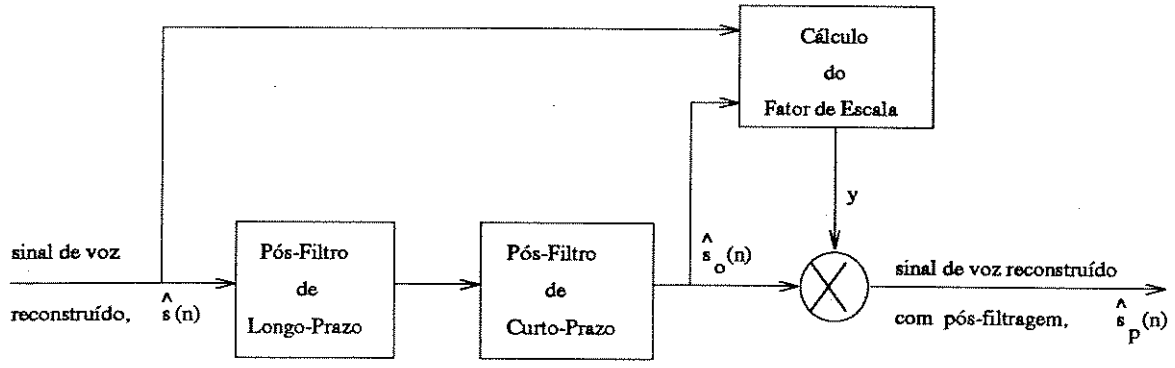

Figura 6.1: Diagrama em blocos do Pós-Filtro Adaptativo

O pós-filtro de longo-prazo é um filtro pente com picos espectrais localizados nas frequências múltiplas da frequência do pitch (inverso do período de pitch) do sinal de voz reconstruído. Seja M o período de pitch em número de amostras. Então,

a função de transferência do pós-filtro de longo-prazo é expressa como :

$$H_l(z) = g_l(1 + bz^{-1}), \tag{6.6}$$

onde os coeficientes g_l, b e o período de pitch M são atualizados conforme os intervalos de atualização do período de pitch no codificador. Assim, nos codecs a 3,5 kbit/s, aqueles parâmetros são atualizados por bloco de 160 amostras, enquanto que nos codecs de baixo atraso são atualizados a cada bloco de 40 amostras.

Os coeficientes g_l e b são determinados a partir do ganho de pitch. Se γ é o ganho de pitch, tem-se que :

$$b = \begin{cases} 0 & \text{se } \gamma < 0,6 \\ \lambda\gamma & \text{se } 0,6 \leq \gamma \leq 1 \\ \lambda & \text{se } \gamma > 1 \end{cases} \tag{6.7}$$

$$g_l = \frac{1}{1+b}, \tag{6.8}$$

onde λ é o fator de controle da intensidade da pós-filtragem.

O pós-filtro de curto-prazo tem a seguinte função de transferência :

$$H_s(z) = \frac{1 - \sum_{i=1}^{8} \overline{b}_i z^{-i}}{1 - \sum_{i=1}^{8} \overline{a}_i z^{-i}}, \tag{6.9}$$

onde : $\overline{b}_i = a_i \eta_1^i$, $i = 1, 2, ..., 8$
$\overline{a}_i = a_i \eta_2^i$, $i = 1, 2, ..., 8$
$\mu = \eta_3 k_1$

O filtro de pólos e zeros atenua as componentes de freqüências entre picos de formantes. Os coeficientes a_i's são do preditor de curto-prazo obtidos a partir dos coeficientes de reflexão quantizados vetorialmente, enquanto que k_1 é o primeiro coeficiente de reflexão. Desta maneira, nos codecs a 3,5 kbit/s os coeficientes $\overline{b}_i$ e $\overline{a}_i$ são atualizados a cada bloco de 160 amostras, enquanto que nos codecs de baixo-atraso, são atualizados a cada bloco de 40 amostras. Através de testes subjetivos informais, obteve-se a mesma combinação de valores de coeficientes utilizados na pós-filtragem do codec LD-CELP [6], tanto para os codecs a 3,45-3,55 kbit/s como para os de baixo-atraso a 6,8 kbit/s, a saber : $\lambda = 0.15$, $\eta_1 = 0.65$ e $\eta_2 = 0,75$

O sinal de voz reconstruído, quando passa pelo pós-filtro, normalmente apresenta uma magnitude média diferente do sinal reconstruído original. O fator de escala y tem a função de ajustar a magnitude média do sinal de voz reconstruído

com pós-fitragem de modo a mantê-la aproximadamente igual à do sinal de voz reconstruído original. Sejam $\hat{s}(n)$ e $\hat{s}_o(n)$ o sinal de voz reconstruído original e na saída do pós-filtro de curto-prazo, respectivamente. Então, o fator de escala y é calculado como :

$$y = \frac{\sum_{n=1}^{40} |\hat{s}(n)|}{\sum_{n=1}^{40} |\hat{s}_o(n)|} \tag{6.10}$$

O sinal reconstruído com pós-filtragem e magnitude corrigida passa a ser dado por :

$$\hat{s}_p(n) = y\hat{s}_o(n) \tag{6.11}$$

Esta operação de ajuste evita a ocorrência de grandes excursões ocasionais na saída do pós-filtro.

6.5 DESEMPENHO

Na figura 6.2 é apresentada a curva de MNRU obtida no experimento, e nas figuras 6.3, 6.4 e 6.5 são mostrados os resultados de desempenho dos codecs para voz feminina, masculina e feminina+masculina. Na tabela 6.1 são apresentados os valores numéricos de MOS e IC(Intervalos de Confiança) obtidos no teste para os diversos codecs.

Tabela 6.1 :Valores numéricos do teste subjetivo

CODECs	*MOS±IC*		
	Voz Fem.	Voz Masc.	Voz Fem.+Masc.
CELP	1,16±0,41	1,13± 0,20	1,15±0,11
CELP-AEP	2,63±0,28	2,25± 0,28	2,44±0,20
CELP-IS	2,74±0,40	2,03± 0,33	2,39±0,26
CELP-AMF	2,94±0,26	2,77± 0,25	2,86± 0,18
MCELP-AMF	2,80±0,28	2,50± 0,28	2,65±0,20
CELP-AEP-BA	3,04±0,29	2,33± 0,28	2,69±0,21
CELP-IS-BA	2,81±0,28	2,46± 0,33	2,64±0,22
RPE-LTP	4,16±0,30	4,24± 0,21	4,20±0,19

O codificador CELP-AMF apresenta o maior valor de MOS dentre todos os codificadores de voz a baixas taxas avaliados. Em relação ao CELP convencional, obtém-se uma melhora de qualidade de cerca de 1.71 em termos de MOS. Em seguida ao codificador CELP-AMF, tem-se o codificador MCELP-AMF, com um valor de MOS aproximadamente 1.50 superior em relação ao CELP convencional.

Os codificadores CELP-AEP e CELP-IS, bem como os codificadores de baixo atraso CELP-AEP-BA e CELP-IS-BA, apresentam praticamente a mesma qualidade.

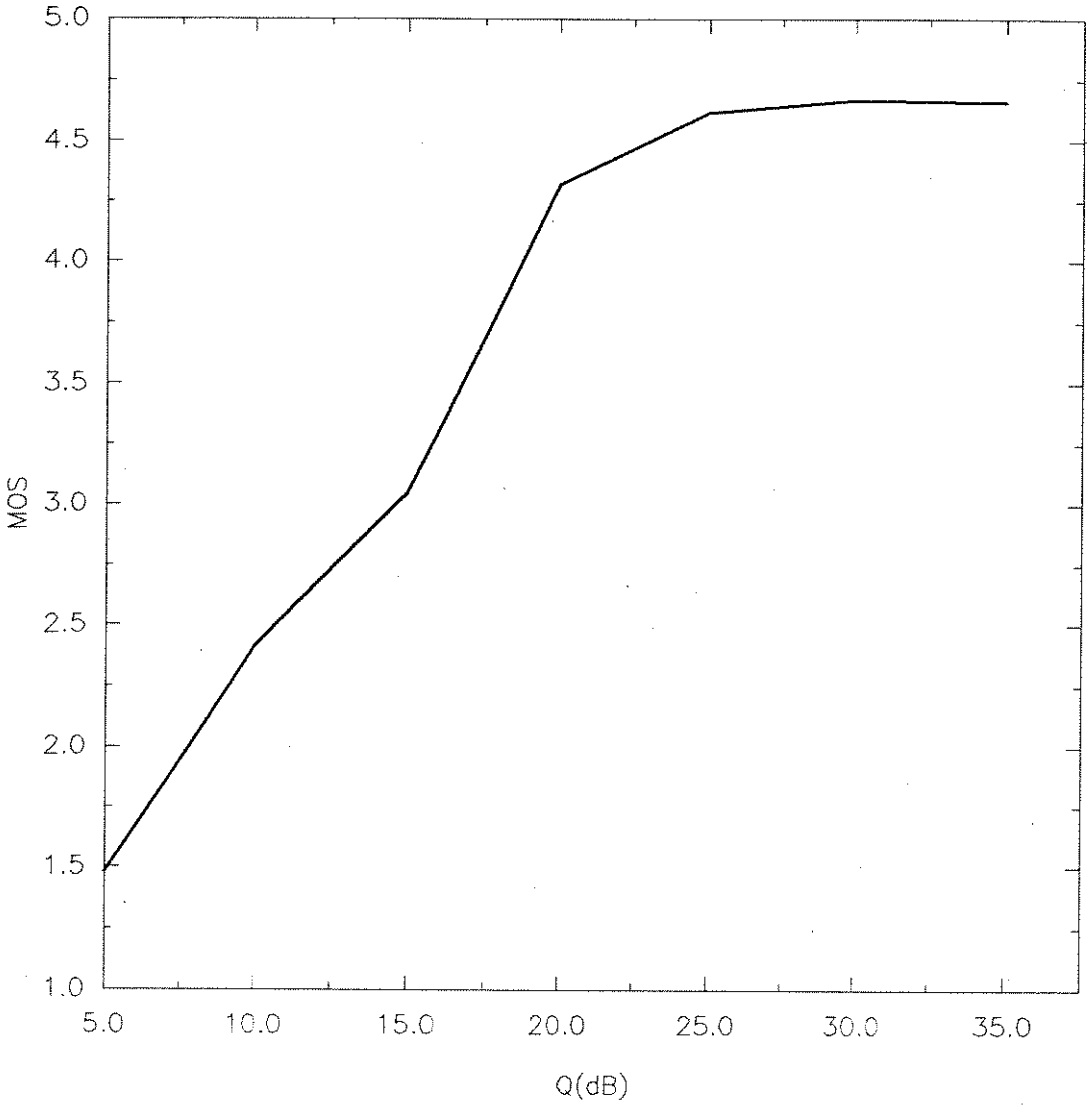

Figura 6.2: Curva de MNRU obtida no teste subjetivo

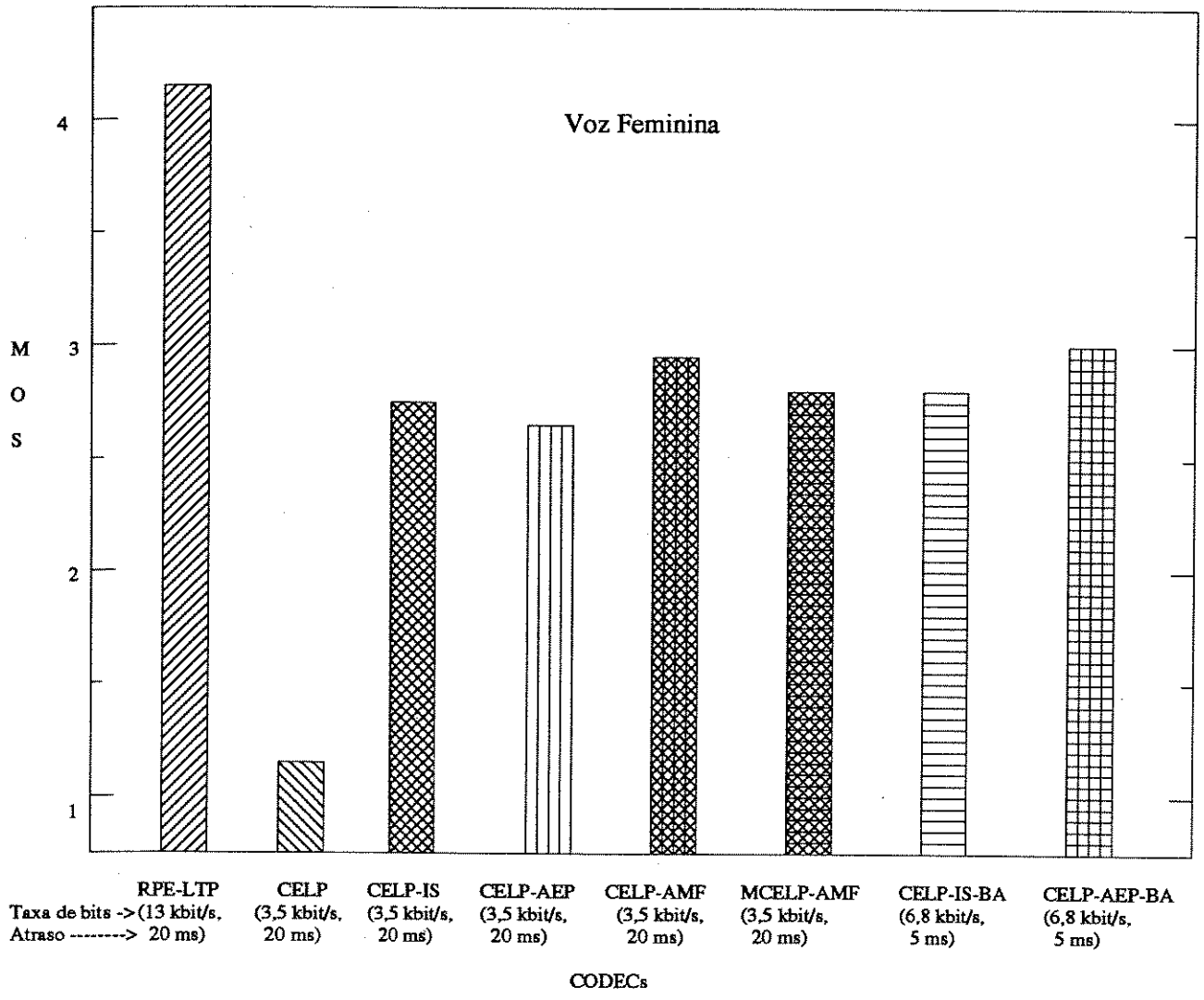

Figura 6.3: Resultados do teste subjetivo para voz feminina

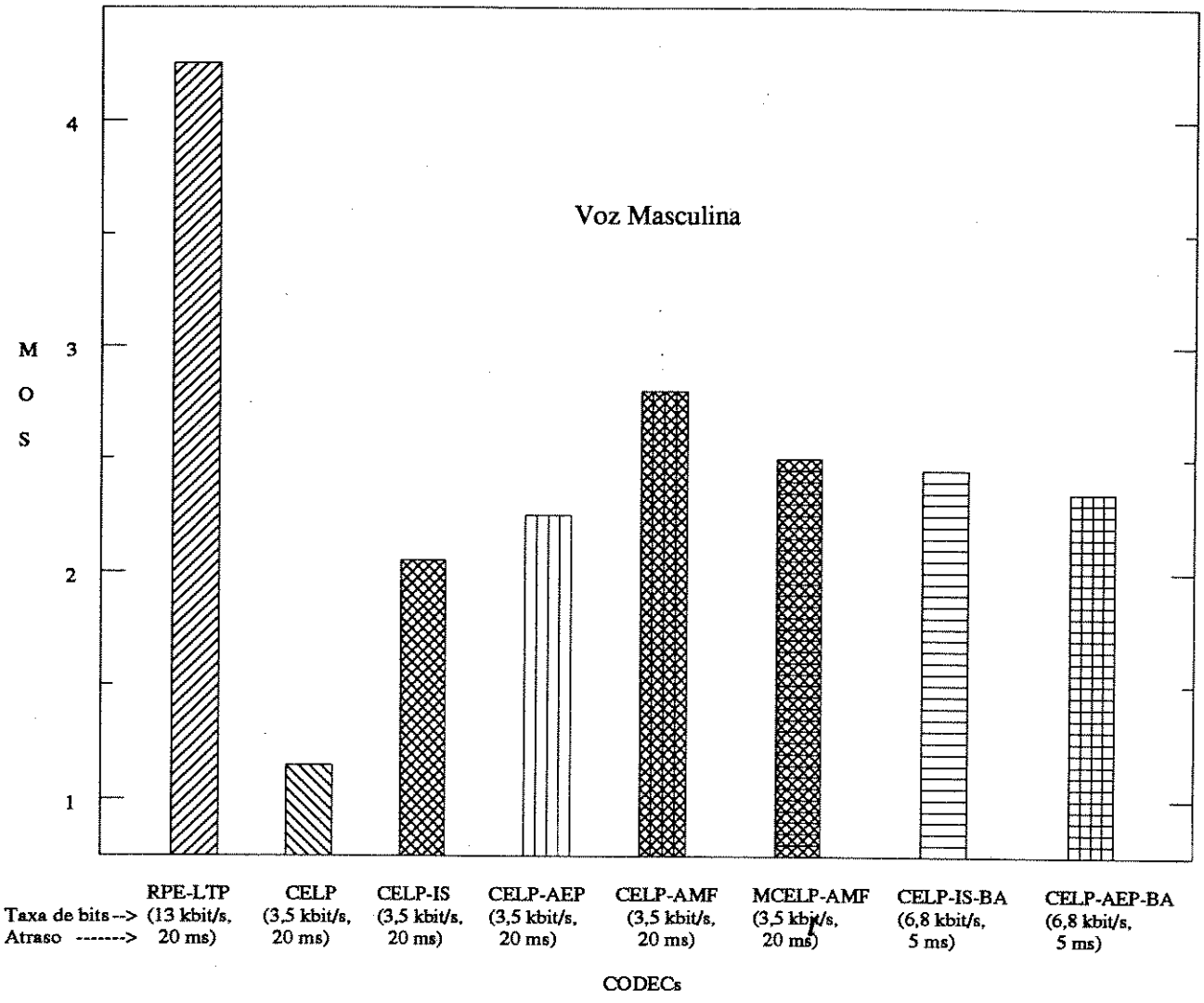

Figura 6.4: Resultados do teste subjetivo para voz masculina

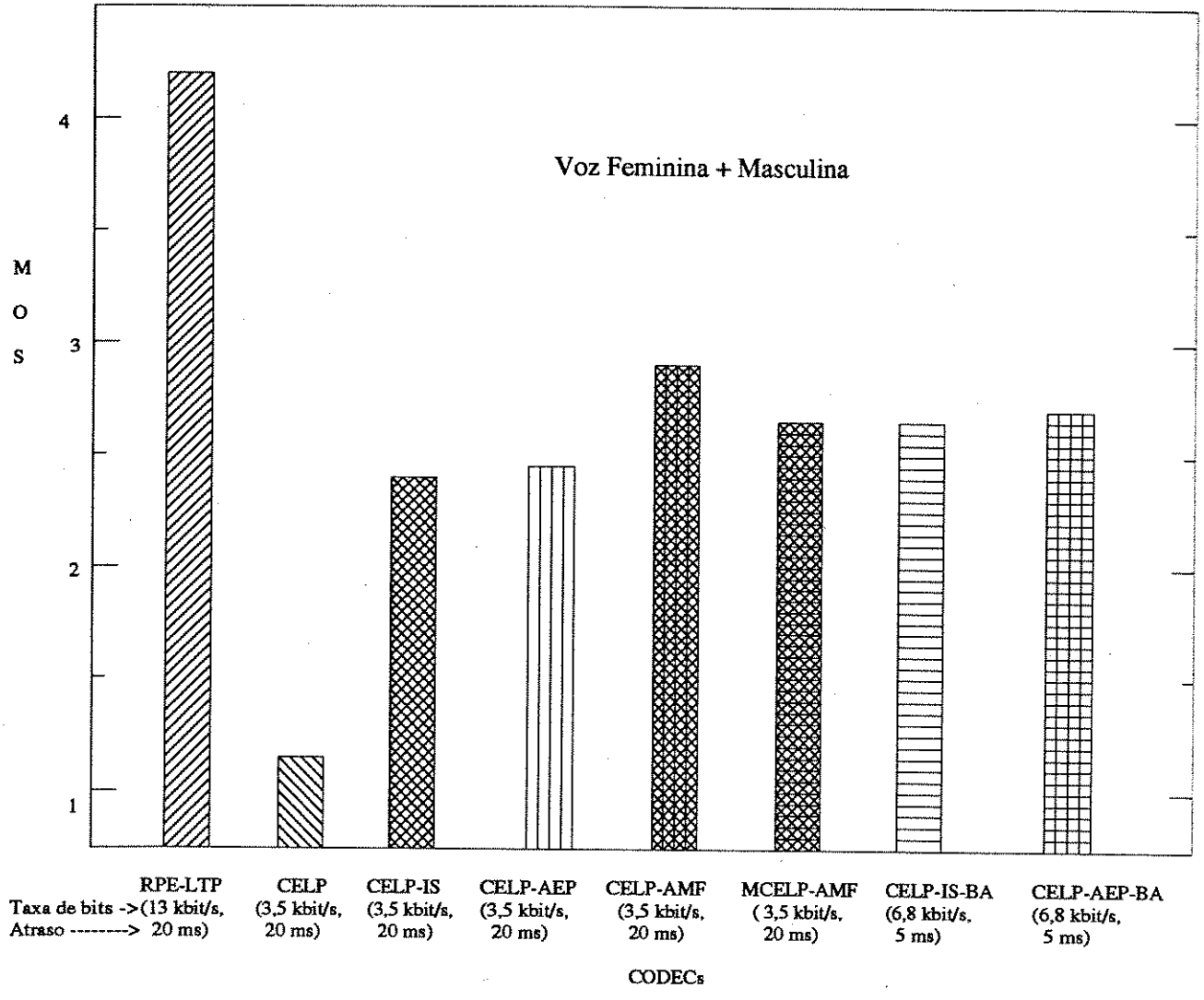

Figura 6.5: Resultados do teste subjetivo para voz feminina+masculina

Como já foi ressaltado no capítulo 4, o dicionário de códigos LPC foi gerado usando-se o algoritmo LBG sem levar em conta a estrutura e nem o tipo de distorção utilizada nos algoritmos AEP e AMF. Além disto, devido à questões operacionais, utilizou-se uma seqüência de treinamento sem filtragem IRS, enquanto que todos os arquivos de voz processados pelos codificadores e avaliados através de teste subjetivo formal passaram pela filtragem IRS. Assim, a qualidade do sinal de voz de todos os codificadores a baixas taxas e de baixo atraso, com exceção do CELP convencional, pode ser ainda melhorada reprojetando-se os respectivos dicionários de códigos, principalmente com relação aos codificadores que utilizam os algoritmos AEP e AMF. Finalmente, ressalta-se que dentre todos os codificadores implementados, o MCELP-AMF é o que pode ter o seu desempenho mais afetado devido a estes fatores, pois, além do dicionário de códigos LPC, emprega também o dicionário de códigos $\mathbf{R}$ que também foi projetado usando-se uma seqüência de treinamento sem filtragem IRS.

Bibliografia

[1] V. Ramamoortthy e N.S. Jayant,"Enhancement of ADPCM speech by adaptive postfiltering", Bell Syst. Tech. J., vol. 63, nº 8, Outubro de 1984, pág. 1465-1475.

[2] N.S. Jayant e V. Ramamoorthy, "Adaptive postfiltering of 16 kbit/s ADPCM speech, "Proc. IEEE Int. Comf. Acoust., Speech, Signal Process., Abril de 1986, pág. 829-832.

[3] Y. Yatsuzuka, S.Iizuka e T. Yamazaki, "A variable rate coding by APC with maximum likelihood quantization from 4.8 kbit/s to 16 kbit/s", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1986, pág. 3071-3074.

[4] P. Kroon e B.S. Atal, "Quantization procedures for 4.8 kbps CELP coders", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1987, pág. 1650-1654.

[5] J.-H. Chen e A. Gersho, "Real-time vector APC speech coding at 4800 bps with adadptive postfiltering", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Abril de 1987, pág.2185-2188.

[6] J.-H. Chen et al,"A Low-Delay CELP Coder for the CCITT 16 kbit/s Speech Coding Standadrt", IEEE Journal on Selected Areas in Communications, volume 10, nº 5, Junho de 1992, pág. 830-849.

Capítulo 7

CONSIDERAÇÕES FINAIS

7.1 CONCLUSÕES

Visando melhorar o desempenho de codificadores do tipo CELP para taxas de codificação abaixo de 4.0 kbit/s e diminuir o atraso de codificação para taxas médias, diversas inovações foram propostas para serem incorporadas a estes codificadores. Dentre estas inovações, a que resulta no melhor desempenho é o algoritmo AMF-2, que realiza uma quantização vetorial conjunta dos coeficientes LPC e parâmetros de excitação de curto-prazo e longo-prazo. Outras inovações propostas são os algoritmos de quantização vetorial dos coeficientes LPC em malha aberta, AEP, e os seguintes algoritmos de redução da taxa de bits do sinal de excitação : i) por alocação dinâmica de bits entre o índice e ganho do vetor de excitação; ii) utilização de dois dicionários de códigos, sendo um deles formado por vetores de uma função de correlação-cruzada normalizada a partir da qual gera-se uma excitação MP em malha fechada.

O algoritmo AEP apresenta uma qualidade subjetiva equivalente ou ligeiramente superior ao algoritmo tradicional usando a medida de distorção de Itakura Saito Modificada. Este bom desempenho, mesmo com blocos de análise de pequeno comprimento, aliado a sua baixa complexidade (aproximadamente 2 vezes menor que a do algoritmo tradicional com medida de Itakura Saito Modificada), torna o algoritmo AEP vantajoso na implementação de codificadores CELP de baixo atraso usando estrutura "forward", ao invés de "backward" que tem sido normalmente proposta na literatura. Neste ponto, vale ressaltar uma importante diferença, em

termos de desempenho, do algoritmo AEP para baixo atraso de codificação quando comparado com a estrutura backward : na estrutura backward, a redução do atraso de codificação é conseguida às custas de uma degradação significativa de desempenho do preditor e, para compensar esta degradação, requer-se uma alta taxa de bits para o sinal de excitação. Já quando se utiliza o algoritmo AEP, a redução no atraso de codificação é conseguida mantendo-se o desempenho de uma estrutura forward com atraso normal, não sendo necessário aumentar a taxa de bits do sinal de excitação. Assim, com o algoritmo AEP, pode-se obter um compromisso ótimo entre as taxas de bits do sinal de excitação e coeficientes LPC de modo a resultar numa taxa de bits global do codificador menor do que no caso backward. Esta mesma propriedade é esperada também do algoritmo AMF-REF e de suas estensões.

Dos algoritmos para redução da taxa de bits do sinal de excitação, o primeiro é mais vantajoso que o segundo em termos de complexidade, embora a complexidade de ambos seja compatível com a tecnologia atual de DSP's comerciais.

Estes algoritmos, quando integrados num codificador CELP, melhoram significativamente o seu desempenho, resultando em codificadores com qualidade em termos de MOS próximo de 3 à taxa de bits tão baixa quanto 3,5 kbits ou 6,8 kbit/s com atraso de 5 ms.

7.2 ATIVIDADES FUTURAS E COMPLEMENTARES

Diversas investigações devem ser ainda realizadas no sentido de otimizar os algoritmos propostos e obter codificadores a taxas abaixo de 4,0 kbit/s e a taxas médias mas com baixo atraso. Dentre estas investigações, destacam-se :

- Utilização de um sinal de treinamento mais adequado para a geração do dicionário de códigos LPC, contando adicionalmente com conversações normais e não simplesmente leitura de texto, com um número maior de locutores e passando pelo filtro IRS;
- Para cada um dos algoritmos de quantização vetorial propostos, utilizar um método de geração de dicionário de códigos LPC compatível que leve em conta a estrutura e medida de distorção utilizada por aqueles algoritmos;
- Na geração do dicionário de códigos da função $R(m)$, utilizar também um sinal de treinamento mais adequado;

- Utilização do algoritmo AMF-2 por sub-bloco na implemetação de codificadores CELP de baixo atraso ao invés do algoritmo AEP ou tradicional com medida de distorção de Itakura Saito Modificada;
- Utilização destes algoritmos em codificadores CELP com o dicionário de códigos de excitação projetado em malha fechada e com filtro de síntese de longo-prazo de 1a ordem com atraso fracionário;
- Incorporação destes algoritmos em outras estruturas de codificadores do tipo CELP como, por exemplo, o VSELP.

Apêndice A

PLANO DE TESTES SUBJETIVOS

A.1 INTRODUÇÃO

O presente plano de testes subjetivos teve como objetivo verificar a qualidade do sinal de voz sintetizado na ausência de erro de bits pelos codecs a taxa de bits entre 3,45 e 3,5 kbit/s e pelos codecs de baixo-atraso a 6,8 kbit/s, conforme descrito no capítulo 6. O método utilizado nos experimentos foi o ACR.

A.2 FATORES E CONDIÇÕES DE REFERÊNCIA

Fatores	**Número**	**Comentários**
Condições para o Codec :		
Codecs	8	CELP, CELP-IS, CELP-AEP, CELP-AMF, MCELP-AMF, CELP-IS-BA, CELP-AEP-BA, RPE-LTP
Taxa de erro de bits	1	0
Nível de Entrada	1	22 dB abaixo de 0dBm0 (G.711)
Condições de Referência :		
MNRU	8	Q=5, 10, 15, 20, 25, 30, 35, 40 dB
Condições Comuns :		
Nível de Audição	1	preferido
Número de locutores	4	2 masculinos e 2 femininos
Número de Avaliadores	16	8 homens e 8 mulheres
Modo de Audição		Via aparelho telefônico
Nível de Ruído Ambiental		< 30 dBA

Na tabela A.1 é apresentada a lista das condições por codec e MNRU.

Tabela A.1 :Lista das condições por codec e MNRU

Condições	Codecs/MNRU
01	CELP-AMF
02	CELP-IS
03	MCELP-AMF
04	CELP-IS-BA
05	MNRU: Q=35
06	CELP-AEP-BA
07	CELP
08	CELP-AEP
09	RPE-LTP
10	MNRU: Q=5
11	MNRU: Q=10
12	MNRU: Q=15
13	MNRU: Q=20
14	MNRU: Q=25
15	MNRU: Q=30
16	MNRU: Q=40

A.3 MATERIAL DE VOZ

Foram utilizados arquivos de voz com filtragem IRS conforme especificado pela recomendação P.48 do CCITT (vide seção 2.4). Todos os arquivos de voz foram gravados em um ambiente com menos de 500 ms de reverberação e ruído ambiental < 30 dBA.

Cada arquivo de voz consiste de duas sentenças curtas (pares de sentenças) escolhidas aleatoriamente de literaturas não-técnicas, jornais, de conversações cotidianas, etc. Cada sentença tem duração aproximada de 3 s e as duas sentenças de cada par são separadas entre si por uma pausa de aproximadamente 1 s. Os pares de sentenca tem duração, incluindo as pausas, de aproximadamente 8 s e separação entre si de aproximadamente 5 s, conforme ilustrado na figura A.1. A separação de 5 s entre os pares de sentenças é o tempo dado ao avaliador para dar a sua nota sobre a qualidade do par de sentenças que acabou de ouvir.

No total foram utilizados 64 arquivos ou pares de sentenças, isto é :

(8 condições de codecs + 8 MNRU)×(4 locutores)= 64 arquivos

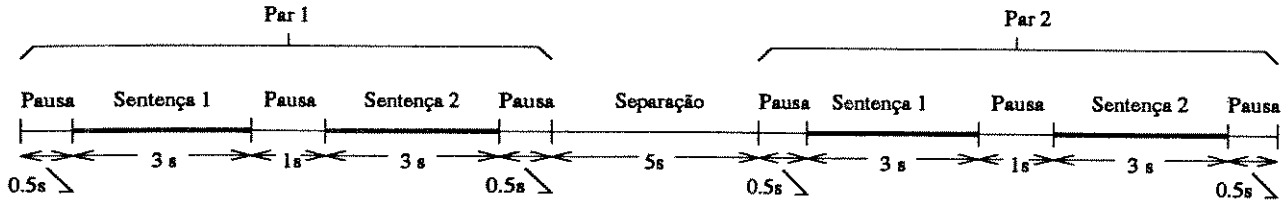

Figura A.1: Padrão de disposição das sentenças no tempo

A.4 RANDOMIZAÇÃO

O resultado de um teste subjetivo de opinião é influenciado pela ordem de apresentação das condições e pelo estado geral do avaliador. Assim, para um determinado número de avaliadores disponíveis, é muito importante que o teste seja limitado em *sessões* de duração máxima tal que não cause fatiga ao ouvinte. Já com relação a ordem de apresentação, um método que balanceia tanto a seleção das condições em segmentos bem como a ordem de apresentação destes segmentos é o princípio do "Latin-square"[1].

No presente plano de testes subjetivos, devido ao pequeno número de condições envolvidas, preferiu-se usar uma variação deste princípio, onde o teste foi dividido em 4 segmentos (A, B, C, D) de 16 condições, com cada segmento tendo a mesma ordem de apresentação aleatória das condições mas deslocada de pelo menos um par de sentenças em relação ao outro. Uma vez determinada esta ordem de apresentação, procedeu-se a uma escolha aleatória e balanceada dos locutores. Os segmentos foram, então, apresentados para audição em 4 ordens diferentes de acordo com o princípio do "Latin-square" :

ORDEM 1 :	P	CABD
ORDEM 2 :	P	DBAC
ORDEM 3 :	P	ADCB
ORDEM 4 :	P	BCDA

Para cada avaliador, inicia-se o experimento com o segmento de Prática (P) a fim de acostumá-lo com o objetivo envolvido. O segmento (P) consiste de 8 arquivos de voz com nível de qualidade variando aleatoriamente desde "Excelente" até "Ruim". Cada ordem de apresentação dos segmentos foi aplicado a um grupo de 4 avaliadores, totalizando 16 avaliadores para as 4 ordens. Adicionalmente, para

não causar fadiga ao avaliador, cada ordem foi aplicada em 4 sessões de 1 segmento, sendo que na primeira sessão foi apresentada também o segmento de Prática.

Na tabela A.2 tem-se a lista das condições e locutor dos 4 segmentos de testes e do segmento (P). Nesta tabela vale a seguinte notação :

IXYZ : X identifica o *Locutor*
YZ identifica a *Condição*

Tabela A.2 :Lista das Condições e Locutor por Segmento

P	A	B	C	D
I205	I201	I203	I213	I209
I302	I105	I301	I303	I413
I201	I315	I205	I101	I103
I110	I216	I115	I405	I401
I411	I110	I416	I215	I305
I108	I206	I310	I316	I415
I406	I302	I406	I410	I116
I314	I404	I202	I106	I210
	I212	I304	I402	I306
	I407	I412	I104	I102
	I114	I107	I312	I204
	I311	I214	I207	I112
	I108	I111	I314	I307
	I409	I408	I411	I414
	I313	I309	I208	I211
	I403	I113	I109	I308

A.5 DURAÇÃO DO TESTE

Cada avaliador teve que ouvir 64 (16 condições$\times$ 4 locutores) arquivos de voz ou pares de sentenças. Considerando-se um total de 8 s para um par de sentenças e 5 s para dar a nota, cada ouvinte gastou em média aproximadamente $(8+5)\times 64 = 13,87$ min para avaliar os 64 arquivos de voz. Como foram 16 avaliadores, a duração total do teste foi de aproximadamente 3,68 hs.

A.6 CÁLCULO DO MOS E IC

O valor de MOS foi obtida fazendo-se uma simples média aritmética das notas dos avaliadores :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{A.1}$$

onde x_i são as notas coletadas e N é o número total de notas para a condição que se deseja avaliar.

A precisão ou Intervalo de Confiança (IC) dos valores de MOS foi calculado para um nível de confiança de 95% conforme a seguinte equação :

$$IC = 1,96\sqrt{\frac{\sigma}{N}}, \tag{A.2}$$

onde σ é a estimativa do desvio padrão dado por :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{A.3}$$

Bibliografia

[1] J. Dénes e A.D. Keedwell,"Latin squares and their applications", Akadémiai Kiadó, Budapest; English Universities Press, London; Academic Press, New York, 1974.