CODIFICAÇÃO CELP E ANÁLISE ESPECTRAL DE VOZ

Ranniery da Silva Maia

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. Sergio Lima Netto, Ph.D.

Prof. Fernando Gil Vianna Resende Junior, Ph.D.

Prof. Eduardo Antônio Barros da Silva, Ph.D.

Prof. Abraham Alcaim, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2000

MAIA, RANNIERY DA SILVA

Codificação CELP e análise espectral de voz [Rio de Janeiro] 2000

XVI,115 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2000)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1.Processamento da Fala 2.Codificação 3.Análise espectral

I.COPPE/UFRJ II.Título (série)

Aos meus pais, Oswaldo e Telma, e aos meus irmãos, Rosiery e Amaury.

Agradecimentos

Meus sinceros agradecimentos:

- aos Professores Sergio Lima Netto e Fernando Gil Vianna Resende Junior, pela orientação durante todo o período de pesquisa;
- ao Professor Abraham Alcaim, da PUC-RJ, pelo fornecimento de artigos que foram de alta importância para o desenvolvimento deste trabalho;
- aos colegas Amaro Azevedo de Lima, Cássio Barboza Ribeiro e Rodrigo Caiado de Lamare, pelas importantes contribuições;
- aos funcionários, alunos e professores do Laboratório de Processamento de Sinais da COPPE/UFRJ, pelas diversas ajudas que me foram dadas;
- ao CNPq, pela bolsa de estudos fornecida.

Ranniery da Silva Maia

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CODIFICAÇÃO CELP E ANÁLISE ESPECTRAL DE VOZ

Ranniery da Silva Maia

Março/2000

Orientadores: Sergio Lima Netto
Fernando Gil Vianna Resende Junior

Programa: Engenharia Elétrica

Neste trabalho é apresentado um sistema CELP para a codificação da fala a uma taxa de 4,67 kbps. São abordadas diferentes formas para a quantização escalar dos parâmetros do filtro de síntese e características relacionadas aos dicionários adaptativo e fixo. Também é apresentada uma análise acerca da estacionaridade de sinais de fala em sub-bandas, com possíveis aplicações para codificação.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

CELP CODING AND SPECTRAL ANALYSIS OF SPEECH

Ranniery da Silva Maia

March/2000

Advisors: Sergio Lima Netto
Fernando Gil Vianna Resende Junior

Department: Electrical Engineering

In this work, a CELP system for speech coding at a bit rate of 4.67 kbps is developed. We present different quantization schemes for the synthesis filter parameters and a discussion on the characteristics of the adaptive and fixed codebooks. Also, a subband stationarity analysis of speech signals is presented with possible applications in speech coding.

Sumário

1 Introdução **1**

1.1 Proposta do trabalho 2

1.2 Organização da dissertação 2

2 Codificação da fala **4**

2.1 Introdução 4

2.2 A fala humana 4

2.2.1 Características 4

2.2.2 Digitalização 7

2.3 Técnicas de Codificação 8

2.3.1 Codificadores de forma de onda 8

2.3.2 Codificadores fonte 9

2.3.2.1 O *vocoder* LPC 9

2.3.3 Codificadores híbridos 10

2.3.3.1 Codificação por excitação de multi-pulsos (MPE) . . 12

2.3.3.2 Codificação por excitação de pulsos regulares (RPE) 13

2.3.4 Comparação entre os grupos de codificadores 13

2.4 A técnica CELP 13

2.4.1 A idéia inicial 14

2.4.2 CELP com dicionário adaptativo 18

2.4.2.1 O filtro de síntese $H(z)$ 18

2.4.2.2 O filtro de ponderação do erro $W(z)$ 20

2.4.2.3 O dicionário fixo 21

2.4.2.4 O dicionário adaptativo 21
2.5 Padrões de codificação CELP 24
2.5.1 DoD-CELP a 4,8 kbps 24
2.5.2 LD-CELP a 16 kbps 25
2.5.3 CS-ACELP a 8 kbps 26
2.5.4 Qualidade subjetiva dos padrões 28
2.6 Conclusão 29

3 Quantização dos parâmetros do filtro de síntese 30
3.1 Introdução 30
3.2 O problema de quantização 30
3.3 Parametrização de $H(z)$ 32
3.3.1 Coeficientes de reflexão 32
3.3.2 Logaritmos da razão das áreas 33
3.3.3 Freqüências do espectro de linha 34
3.4 Quantização escalar dos parâmetros de $H(z)$ 37
3.4.1 Quantização dos LAR 38
3.4.2 Quantização das LSF 40
3.4.3 Quantização diferencial das LSF 43
3.5 Avaliação dos quantizadores 46
3.5.1 Medida de distorção 46
3.5.2 Formas de avaliação 47
3.5.2.1 Avaliação sem o efeito da interpolação 47
3.5.2.2 Avaliação com o efeito da interpolação 51
3.6 Conclusão 55

4 Implementação de um sistema CELP 56
4.1 Introdução 56
4.2 Considerações iniciais 57
4.3 A análise LPC 58
4.3.1 Características gerais 58

4.3.2 Suavização espectral 58
4.3.3 Quantização e interpolação 61
4.4 Obtenção da melhor excitação 62
4.5 O dicionário adaptativo 65
4.5.1 Procedimento de busca rápida 67
4.6 O dicionário fixo 68
4.6.1 Procedimento de busca rápida 70
4.6.2 Excitação restrita 71
4.7 Quantização dos ganhos e alocação de bits 72
4.8 Resumo das características 74
4.9 Avaliação do sistema implementado 76
4.9.1 Medidas objetivas de qualidade usadas 77
4.9.1.1 Razão sinal-ruído segmentada perceptual 77
4.9.1.2 Distância cepestral 78
4.9.1.3 Distância de Itakura 78
4.9.2 Resultados 78
4.10 Conclusão 81

5 Análise de sinais de fala em sub-bandas 83
5.1 Introdução 83
5.2 Método utilizado para a análise 84
5.3 Medidas de distância 85
5.3.1 Distância espectral logarítmica 85
5.3.2 Distância cepestral 86
5.3.3 Distância de Itakura 87
5.3.4 Distancia de Itakura-Saito 89
5.3.5 Distância entre CPL no formato LAR 90
5.4 Experimentos 91
5.4.1 Base de dados usada 91
5.4.2 Aplicação das medidas 91
5.4.3 Bancos de filtros usados 92

5.4.4 Resultados 94
5.5 Conclusão 98

6 Conclusão **100**
6.1 Contribuições do trabalho 100
6.2 Propostas para trabalhos futuros 102

Referências Bibliográficas **103**

A Bases de dados utilizadas **108**

B Partições e dicionários dos quantizadores **110**

Lista de Figuras

2.1 Diagrama de blocos ilustrando simplificadamente como a fala é produzida. 5

2.2 Exemplos de tipos diferentes de sons para segmentos de 25 ms de um sinal de fala: (a) sonoro; (b) surdo. 6

2.3 Espectro de potência de um modelo do trato vocal humano com a indicação dos formantes. 7

2.4 Modelo elétrico digital simplificado de produção da fala humana. . . . 9

2.5 Diagrama de blocos do *vocoder* LPC: (a) codificador; (b) decodificador. 11

2.6 Diagrama de blocos dos codificadores híbridos que utilizam o procedimento de análise-por-síntese. 12

2.7 Comportamento taxa×qualidade dos principais grupos de codificadores da fala (figura reproduzida de [1]). 14

2.8 Processo de remoção de correlações de um segmento de sinal de fala. . 15

2.9 Exemplos de remoção de correlações: (a) segmento sonoro de sinal de fala; (b) segmento mostrado em (a) sem correlações de curto termo; (c) segmento mostrado em (a) sem correlações de curto e longo termo. 16

2.10 Primeiro sistema de codificação CELP: (a) codificador; (b) decodificador. 17

2.11 Sistema de codificação CELP com dicionário adaptativo: (a) codificador; (b) decodificador. 19

2.12 Espectros de amplitude dos filtros $H(z)$ e $W(z)$ para um segmento sonoro de um sinal de fala. 20

2.13 Comportamento taxa×qualidade de alguns algoritmos padrões de codificação da fala. 28

3.1 Modelo de tubos com iguais comprimentos para o trato vocal humano. 33
3.2 Gráfico com as raízes de $A(z)$, $P(z)$ e $Q(z)$ indicadas respectivamente por estrelas, círculos e triângulos, para um segmento sonoro de 25 ms. 36
3.3 Histogramas dos valores assumidos por cada LAR para a base de dados considerada: de l_1 (a) a l_{10} (j). 39
3.4 Histogramas dos valores assumidos por cada LSF para a base de dados considerada: de w_1 (a) a w_{10} (j). 42
3.5 Histogramas dos valores assumidos pelas diferenças entre LSF consecutivas para a base de dados considerada: de Δw_1 (a) a Δw_{10} (j). . . 45
3.6 Procedimento de determinação da DE_q para a avaliação da quantização dos parâmetros de $H(z)$ sem levar em conta a interpolação entre blocos. 48
3.7 Detalhes da forma como são determinados os coeficientes LPC para cada bloco, para a avaliação dos quantizadores sem levar em conta a interpolação entre blocos. 48
3.8 Procedimento de determinação da DE_c usada para a avaliação da quantização dos parâmetros de $H(z)$, levando-se em conta os efeitos da interpolação entre blocos. 52
3.9 Detalhes da forma como são determinados os coeficientes LPC para cada sub-bloco, para a avaliação dos quantizadores levando-se em conta a interpolação entre blocos. 52

4.1 Janela binomial usada para suavização espectral: (a) resposta ao impulso; (b) resposta em amplitude. 60
4.2 Diagrama de blocos detalhado do codificador CELP implementado. . 63
4.3 Respostas em amplitude dos filtros interpoladores usados na busca da melhor excitação no dicionário adaptativo: (a) para a faixa de resolução em oitavas; (b) para a faixa de resolução em quartas. 67

4.4 Forma de determinação das seqüências candidatas no dicionário fixo com superposição utilizado. 70
4.5 Distribuições do ganho G_a, considerando apenas valores na faixa $[0; 2]$. 73
4.6 Distribuições do ganho $\tilde{G}_f$, considerando apenas valores na faixa $[-0,05; 0,05]$. 73

5.1 Procedimento utilizado para analisar a estacionaridade de sinais de fala em sub-bandas. 84
5.2 Respostas em magnitude dos filtros do banco uniforme. 93
5.3 Respostas em magnitude dos filtros do banco não-uniforme. 94

Lista de Tabelas

3.1 Características de cada LAR l_i para a base de dados considerada. . . 40

3.2 Alocação de bits para a quantização uniforme dos LAR, com indicação dos limites superiores e inferiores inferidos pelos histogramas. 41

3.3 Características de cada LSF w_i para a base de dados considerada. . . 41

3.4 Alocações de bits para quantizações não-uniformes das LSF. 43

3.5 Características das diferenças entre LSF, Δw_i, para a base de dados considerada. 44

3.6 Alocações de bits para quantizações diferenciais não-uniformes das LSF. 46

3.7 Desempenho dos quantizadores para blocos de 20 ms com sub-blocos de 5 ms, sem levar em conta os efeitos da interpolação entre blocos. . 50

3.8 Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 6 ms, sem levar em conta os efeitos da interpolação entre blocos. . 50

3.9 Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 7,5 ms, sem levar em conta os efeitos da interpolação entre blocos. 51

3.10 Desempenho dos quantizadores para blocos de 20 ms com sub-blocos de 5 ms, levando-se em conta os efeitos da interpolação entre blocos. . 53

3.11 Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 6 ms, levando-se em conta os efeitos da interpolação entre blocos. . 54

3.12 Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 7,5 ms, levando-se em conta os efeitos da interpolação entre blocos. 54

4.1 Amplitudes das amostras da janela binomial usada para a suavização espectral. 59

4.2 Razões sinal-ruído segmentada perceptual em dB, para 4 sinais processados com suavização espectral, expansão dos CPL e nenhum dos métodos. 61
4.3 Característica do dicionário adaptativo utilizado. 66
4.4 Razões sinal-ruído segmentada perceptual em dB, para o sistema CELP com os dois tipos de dicionários: o sem subtração de amostras acima do limiar (tipo 1); e o com subtração de amostras acima do limiar (tipo 2). 69
4.5 Características de G_a e $\tilde{G}_f$ dentro das faixas $[0;2]$ e $[-0,05;0,05]$, respectivamente. 74
4.6 Alocação de bits para os parâmetros da excitação. 74
4.7 Resumo das características do sistema CELP implementado. 75
4.8 Sentenças usadas para a avaliação objetiva do sistema CELP implementado. 76
4.9 Medidas objetivas de qualidade para o sistema implementado. 79
4.10 Medidas objetivas de qualidade para o DoD-CELP 4,8 kbps. 79
4.11 Medidas objetivas de qualidade para o LD-CELP 16 kbps. 79
4.12 Medidas objetivas de qualidade para o sistema implementado com o procedimento de busca rápida no dicionário adaptativo. 80
4.13 Tempos de processamento (TP) e percentuais de tempo real (%TR) para os sistemas com e sem busca rápida no dicionário adaptativo. . . 81

5.1 Resposta ao impulso do filtro protótipo que gera o banco de filtros de análise uniforme. 93
5.2 GVE para cada medida de distância para 200 sentenças de um locutor masculino para segmentos de 10 ms, usando o banco de filtros uniforme. 95
5.3 VPR para cada medida de distância para 200 sentenças de um locutor masculino para segmentos de 10 ms, usando o banco uniforme. 95
5.4 VPR para 200 sentenças de cada locutor, usando a D_{EL} e segmentos fixos de 10 ms com o banco uniforme. 96

5.5 VPR para as 600 sentenças para vários tamanhos de segmentos, usando a D_{EL} e o banco uniforme. 97

5.6 VPR para as 600 sentenças para vários tamanhos de segmentos, usando a D_{EL} e o banco não-uniforme. 97

A.1 Parte das sentenças usadas para o projeto dos quantizadores dos parâmetros de $H(z)$ no Capítulo 3, com a indicação dos locutores. . . 109

A.2 Sentenças usadas para a avaliação dos quantizadores dos parâmetros de $H(z)$ no Capítulo 3, com a indicação dos locutores. 109

A.3 Parte das sentenças usadas para o projeto dos quantizadores dos ganhos G_a e $\tilde{G}_f$ no Capítulo 4, com a indicação dos locutores. 109

B.1 Intervalos de decisões (ID) e valores de saída (VS) para os quantizadores não-uniformes de 4 bits para os coeficientes Δw_1 e Δw_2. 111

B.2 Intervalos de decisões (ID) e valores de saída (VS) para os quantizadores não-uniformes de 3 bits para os coeficientes $\Delta w_3, \ldots, \Delta w_6$. . . 112

B.3 Intervalos de decisões (ID) e valores de saída (VS) para os quantizadores não-uniformes de 3 bits para os coeficientes $\Delta w_7, \ldots, \Delta w_{10}$. . . 113

B.4 Intervalos de decisões (ID) e valores de saída (VS) para o quantizador não-uniforme de 4 bits para o ganho G_a. 114

B.5 Intervalos de decisões (ID) e valores de saída (VS) para o quantizador não-uniforme de 5 bits para o ganho $\tilde{G}_f$. 115

Capítulo 1

Introdução

Antes da era das comunicações digitais, a fala era transmitida e armazenada como um sinal analógico. Hoje ela pode ser representada de uma maneira digital, o que permite um armazenamento e transmissão de forma bastante eficiente. O primeiro sistema para codificação da fala foi criado na década de 1930 e foi usado em um sistema de telefonia segura durante a II Guerra Mundial. Desde lá até o início da década de 1970, parecia que somente os militares estavam interessados em codificação da fala. Tudo isso foi mudando a partir da década de 1980 através de várias transformações, iniciando pela digitalização das redes telefônicas. O sistema de codificação da fala por modulação por códigos de pulsos (*pulse code modulation*, PCM), que opera a 64 kilobits por segundo (kbps) e foi projetado para transmitir fala na banda da rede telefônica, tornou possível manter qualidade uniforme para conexões de longa distância. Não demorou muito e logo depois foi descoberto que ao usar o sistema de codificação PCM diferencial adaptativo (*adaptive differential PCM*, ADPCM) a 32 kbps, poderia ser dobrada a capacidade de importantes enlaces de banda estreita, tais como cabos submarinos. O mundo entrou na era do computador pessoal e telefone celular digital. Aplicações como mensagem de voz, videofones, documentos multimídia e Internet passaram a ter necessidade de codificadores da fala digital, cada uma com seus requisitos próprios que dizem respeito à taxa de bits, qualidade, atraso de codificação, complexidade e proteção contra erros durante a transmissão. Conseqüentemente, muitos novos codificadores da fala foram

padronizados a partir de 1987, e também muitos outros foram desenvolvidos para a utilização em sistemas proprietários cujas aplicações não necessitam de padronização, tais como secretárias eletrônicas digitais. Atualmente uma das técnicas mais utilizadas para a codificação da fala é a predição linear com excitação por códigos de dicionários (*code-excited linear prediction*, CELP) porque consegue boa qualidade a baixas taxas de bits.

1.1 Proposta do trabalho

Este trabalho desenvolve um sistema CELP para a codificação da fala que opera a uma taxa de bits igual a 4,67 kbps. O sistema proposto incorpora quantização escalar dos parâmetros do filtro de síntese, escolhida dentre outras formas de quantização através de testes com medidas de distorção; dicionário adaptativo com atrasos fracionários; dicionário fixo obtido a partir de amostras de ruído branco gaussiano; e método de busca seqüencial com otimização dos ganhos. Além disso, são usados procedimentos de busca rápida nos dois dicionários tendo em vista acelerar o processo de codificação. As avaliações do sistema são feitas com base em medidas objetivas de qualidade e testes informais subjetivos, comparando-o com um algoritmo padrão existente e que opera à mesma taxa. Também é apresentada uma análise de sinais de fala em sub-bandas, com possíveis aplicações para codificação.

1.2 Organização da dissertação

O Capítulo 2 fornece uma idéia geral do campo de codificação digital da fala, chegando à parte mais restrita que é a codificação CELP. Para esta técnica são descritas detalhadamente as suas partes constituintes, bem como suas respectivas importâncias para a qualidade do sinal decodificado. São também abordados resumidamente alguns padrões de codificação.

O Capítulo 3 propõe uma quantização escalar eficiente dos parâmetros do filtro de síntese usado nos sistemas CELP. São mostrados os projetos de alguns quantizadores para dois tipos de parâmetros: logaritmos da razão das áreas (*log*

area ratio, LAR) e freqüências do espectro de linha (*line spectral frequencies*, LSF). Os quantizadores são comparados tendo em vista o funcionamento em um sistema CELP, com base em uma medida de distorção espectral, a fim de verificar o melhor deles levando-se em conta o compromisso entre taxa de bits e distorção introduzida pela quantização.

O Capítulo 4 descreve o projeto e implementação de um sistema CELP que opera a uma taxa de 4,67 kbps. São mostradas todas as suas características, tal como a análise LPC (*linear predictive coding*), os tipos de dicionários adaptativo e fixo utilizados e procedimentos de busca rápida nos mesmos para agilizar o processo de codificação. A avaliação é feita ao compará-lo com um algoritmo padrão de codificação CELP que opera à mesma taxa, com base em medidas de qualidade objetivas e testes informais de escuta.

O Capítulo 5 mostra uma análise de estacionaridade de sinais de fala em sub-bandas usando sete medidas de distâncias entre envoltórias de espectros de magnitude de curto termo. O objetivo é a verificação de características que possam ser aplicadas para a codificação em sub-bandas.

O Capítulo 6 mostra o resumo de toda a dissertação com comentários a respeito dos resultados obtidos e propostas para trabalhos futuros.

Capítulo 2

Codificação da fala

2.1 Introdução

Este capítulo fornece uma idéia geral do campo de codificação da fala, abordando desde os conceitos resumidos de sua forma de produção biológica até alguns sistemas padrões de codificação existentes atualmente. É dada uma ênfase bem maior à técnica CELP, que corresponde ao objetivo principal deste trabalho.

A Seção 2.2 mostra conceitos sobre a fala humana, tais como suas características, digitalização e modelamento elétrico do sistema de produção; a Seção 2.3 fala sobre os tipos de codificadores da fala; na Seção 2.4 é descrita detalhadamente a técnica de codificação CELP; na Seção 2.5 são descritas as características de alguns padrões de codificação; e na Seção 2.6 estão as conclusões do capítulo.

2.2 A fala humana

2.2.1 Características

A fala humana é produzida quando ar é forçado dos pulmões a passar pelas cordas vocais e trato vocal, que corresponde à região que vai da abertura das cordas vocais (ou glote) até os lábios. O fluxo de ar dos pulmões é modificado pela ação da glote ou do trato vocal, para ser transformado na *excitação da fala*, ou somente *excitação*, que pode tomar três formas possíveis [2, 3]:

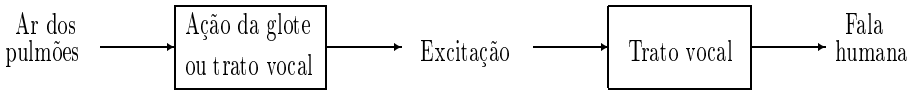


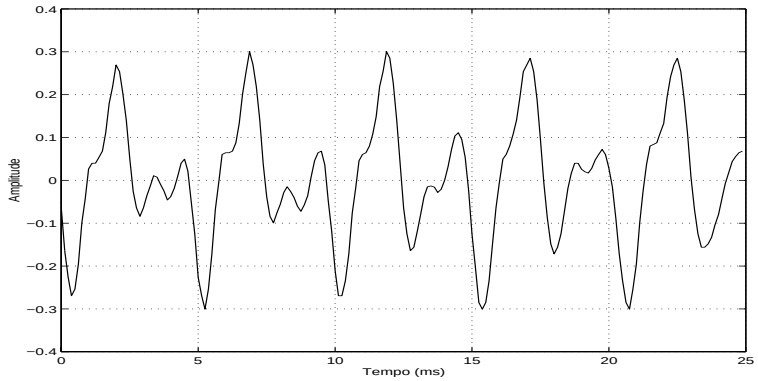
Figura 2.1: Diagrama de blocos ilustrando simplificadamente como a fala é produzida.

- pulsos de ar quase periódicos;
- onda sonora com característica ruidosa;
- um simples pulso de ar.

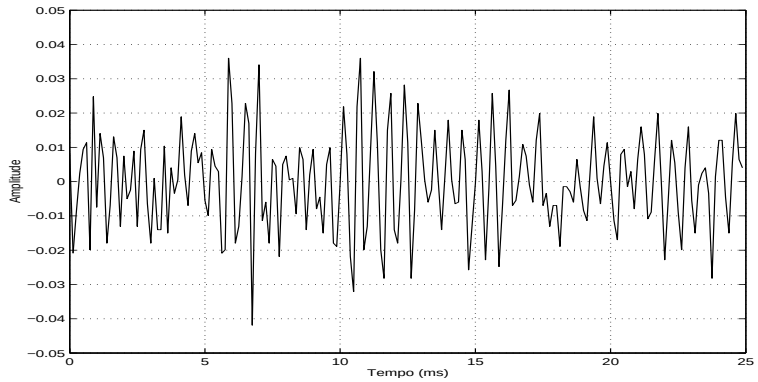
O sinal de excitação é depois modificado pelo trato vocal para gerar o sinal sonoro que representa a fala humana, conforme é mostrado na Figura 2.1.

A maior parte dos sons da fala pode ser associada a um dos três tipos de excitações mencionadas anteriormente. As excitações do tipo pulsos quase periódicos geram os sons *sonoros*, tal como a pronúncia da vogal "a" em "c**a**sa". As excitações do tipo ruidosas geram os sons *surdos*, tal como a pronúncia de "ch" em "**ch**á". Já as excitações do tipo pulso simples geram os sons plosivos, como o bilabial "p" em "**p**ai". Alguns sons, no entanto, são formados pela combinação de diferentes tipos de excitações [3], mas os mais considerados quando o assunto diz respeito à codificação são os sonoros e surdos. Os sons sonoros são quase periódicos no domínio do tempo, onde o período relativo à sua freqüência fundamental é geralmente chamado de período de *pitch*, sendo este parâmetro muito importante para os codificadores da fala atuais. Os sons surdos têm a característica de um ruído com uma alta taxa de cruzamentos por zeros. Mais informações sobre as características dos sons da fala podem ser obtidas em [2, 4]. A Figura 2.2 mostra exemplos de sons sonoro e surdo. Neste caso, pode-se perceber para o som sonoro um período de *pitch* de aproximadamente 5 ms.

O espectro de potência dos sinais de excitação cobrem uma ampla faixa de freqüência. Para produzir os diferentes sons para cada tipo de excitação, a onda acústica é filtrada pelo trato vocal, cuja resposta em freqüência depende das posições da língua, lábios e outros órgãos articulatórios. O espectro de potência do trato



(a)



(b)

Figura 2.2: Exemplos de tipos diferentes de sons para segmentos de 25 ms de um sinal de fala: (a) sonoro; (b) surdo.

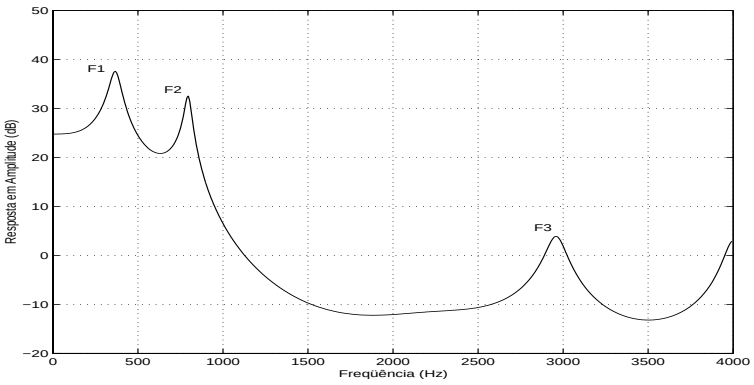


Figura 2.3: Espectro de potência de um modelo do trato vocal humano com a indicação dos formantes.

vocal é caracterizado por um conjunto de freqüências de ressonâncias chamadas de *formantes*, que molda o espectro do sinal de excitação e determina a característica de cada som da fala [3]. A Figura 2.3 mostra o espectro de potência de um modelo de trato vocal para um segmento sonoro de 25 ms de um sinal de fala. Nesta figura pode-se perceber as freqüências relativas aos formantes nos picos em aproximadamente 300, 750 e 2900 Hz.

Os sinais da fala são não estacionários, ou seja, suas características espectrais e estatísticas variam ao longo do tempo. Na melhor das hipóteses, curtos segmentos, tipicamente de 5 a 30 ms, podem ter suas características espectrais consideradas aproximadamente invariantes [4]. Portanto, para estes sinais, a análise deve ser feita em trechos de pequena duração, empregando-se uma janela temporal.

2.2.2 Digitalização

Na digitalização de sinais de fala para telefonia, é comum passar o sinal analógico por um filtro passa-baixas com freqüência de corte igual a aproximadamente 3,4 kHz e posteriormente amostrá-lo a 8 kHz. Em alguns casos é realizada depois uma filtragem passa-altas digital para remover ruídos de baixa freqüência [5].

A taxa de bits de um sinal de fala representado na forma digital é dada por

$$T_{bits} = R \times F_s \times N_{canais}, \tag{2.1}$$

onde T_{bits} é a taxa de bits, R é o número de bits por amostra ou resolução, F_s é a freqüência de amostragem, e N_{canais} o número de canais (mono ou estéreo). A unidade usada é geralmente o *kilobits por segundo* (kbps).

2.3 Técnicas de Codificação

As técnicas de codificação digital da fala podem ser divididas em três principais grupos: codificadores de forma de onda, codificadores fonte e codificadores híbridos. Há ainda alguns autores que preferem juntar os dois últimos grupos em um só e chamá-lo de codificadores paramétricos [5]. Nesta dissertação será usada a primeira classificação.

2.3.1 Codificadores de forma de onda

Os codificadores de forma de onda utilizam-se somente das características temporais ou espectrais dos sinais de fala [4]. Devido a este fato são também aplicáveis na codificação de qualquer sinal onde o interesse seja somente restituir a forma de onda original, sem levar em conta características particulares do sinal que está sendo processado. A essa família pertencem, dentre outras, as técnicas [4, 6]: PCM, ADPCM, codificação em sub-bandas (*sub-band coding*, SBC) e o padrão MPEG (*motion pictures experts group*) para áudio.

A técnica mais simples de codificação de forma de onda é a PCM, onde para cada amostra do sinal de fala é atribuído um código binário. Apesar de bastante redundante, essa técnica ainda é o padrão de telefonia fixa digital utilizado no Brasil e outros países. Para a regulamentação G.711 da União Internacional de Telecomunicações, setor de padronização para telecomunicações (*International Telecommunications Union, Telecommunications Standardization Sector*, ITU-T), que representa o antigo Comitê de Consultoria para Telefonia e Telégrafos Internacional (*Consultative Committee for International Telephone and Telegraph*, CCITT) são utilizados quantização logarítmica, que pode ser a μ-*law* (Estados Unidos) ou a *A-law* (Europa e Brasil) [6], com oito bits para cada amostra, resultando em uma

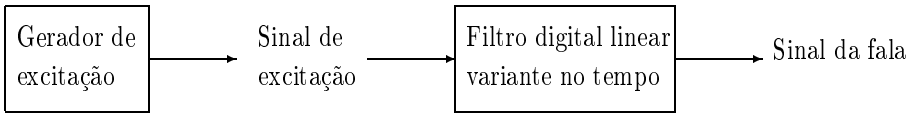


Figura 2.4: Modelo elétrico digital simplificado de produção da fala humana.

taxa de 64 kbps. Os graus de compressão da fala são tomados tendo esta taxa como referência. Assim, se uma determinada técnica codifica a uma taxa de 32 kbps, afirma-se que ela realiza uma compressão de 2 para 1 (2:1). Fala reconstruída pela codificação PCM G.711 possui qualidade quase indistinguível da original [1].

Para taxas abaixo de 16 kbps a qualidade dos codificadores de forma de onda degrada-se rapidamente. A solução para este fato é a utilização dos codificadores fonte e híbridos, que são discutidos a seguir.

2.3.2 Codificadores fonte

Os codificadores fonte conseguem os mais altos graus de compressão com média complexidade. Isso é feito através da extração de parâmetros no sinal original que modelam o sistema humano de produção da fala. A grande desvantagem desta família de codificadores é a qualidade sintética (ou robótica) do sinal reconstruído.

O modelo elétrico digital simplificado para produção da fala humana [4] é formado por um filtro linear variante no tempo e um gerador de excitação, como mostrado na Figura 2.4. O filtro variante representa o trato vocal, e a sua excitação representa o ar que é passado pelo mesmo para produzir os diversos sons da fala.

O codificador fonte mais conhecido atualmente é o *vocoder* LPC (*linear predictive coding*), que será tratado a seguir.

2.3.2.1 O *vocoder* LPC

O codificador da fala por predição linear, também conhecido como *vocoder LPC*, faz uso do modelo simplificado de produção da fala humana, e consegue obter fala inteligível mas de má qualidade a taxas entre 1 e 2,4 kbps [6].

A Figura 2.5 mostra o diagrama de blocos do *vocoder* LPC. O sinal de fala

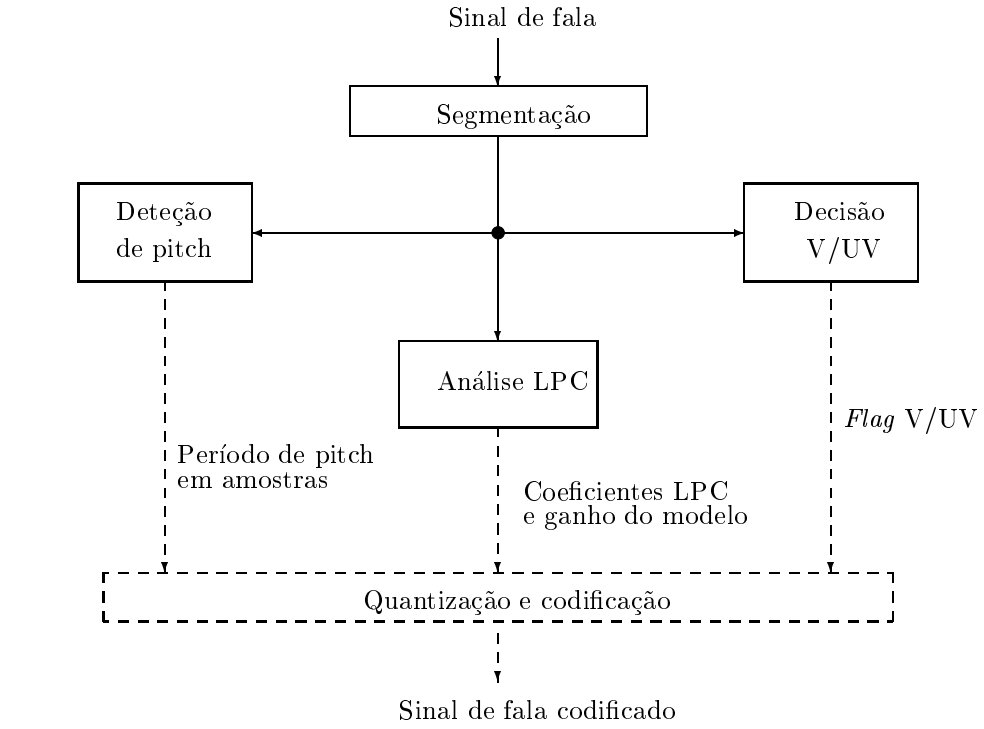
digitalizado é dividido em blocos, geralmente de 5 a 30 ms. Para cada bloco são determinados os parâmetros do filtro síntese, que representa o trato vocal, e um ganho. Também é feita uma classificação do bloco do sinal entre sonoro ou surdo. Se o sinal for sonoro é determinado o período de *pitch*. Dessa forma, as informações que são enviadas ao receptor para cada bloco são os parâmetros do filtro de síntese, o ganho, a classificação entre sonoro e surdo (*flag* U/UV), e o período de *pitch* se ocorrer o primeiro caso. No decodificador, estas informações são usadas para compor o modelo simplificado de produção da fala. O sinal decodificado é obtido quando uma excitação, escalada pelo ganho, é passada pelo filtro de síntese. Essa excitação poderá ser formada de duas formas: (1) para um bloco sonoro a excitação será um trem de impulsos discretos cujo período é igual ao período de *pitch* determinado; (2) para um bloco surdo a excitação será uma seqüência de ruído branco gaussiano, em geral de média zero e variância unitária.

A codificação LPC descrita acima é geralmente utilizada para fins militares e outros onde a boa qualidade da fala decodificada não é importante. O governo dos Estados Unidos padronizou em 1976 uma codificação LPC, conhecida como LPC-10, para comunicações seguras a uma taxa de 2,4 kbps [2, 4, 6]. Tal regulamentação ficou conhecida como padrão federal 1015 (*federal standard 1015*, FS-1015).

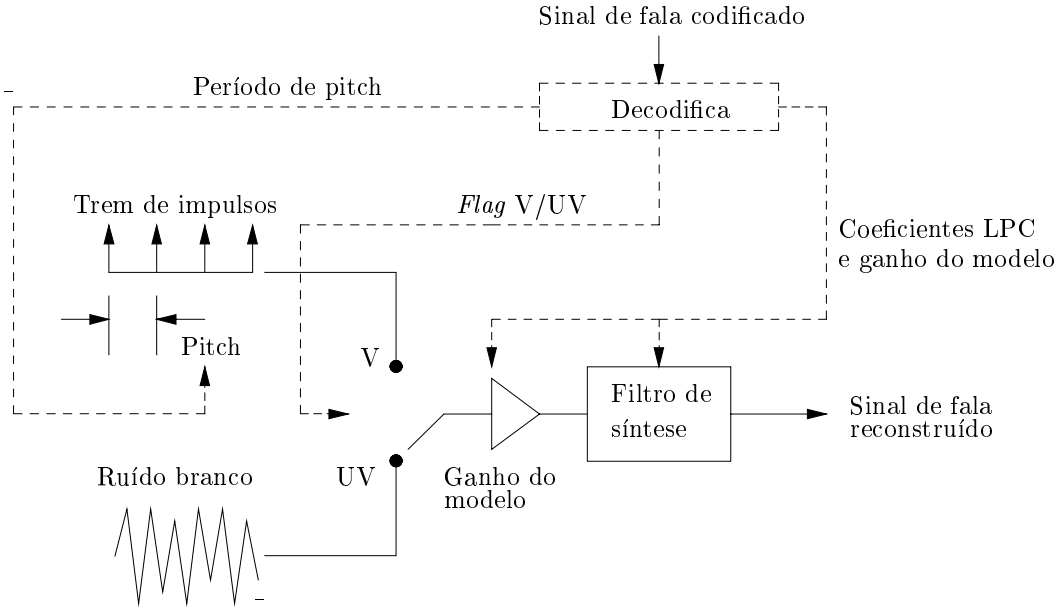
2.3.3 Codificadores híbridos

Os codificadores híbridos realizam a extração de parâmetros dos sinais da fala tal como os codificadores fonte, e ao mesmo tempo utilizam características temporais e espectrais como os codificadores de forma de onda. Dessa forma conseguem obter boa qualidade a taxas entre 2 e 16 kbps, a um custo de um considerável aumento na complexidade do processo de codificação.

A Figura 2.6 mostra o diagrama de blocos de uma classe de codificadores híbridos que utiliza o procedimento de análise-por-síntese [6, 7]. Neste procedimento, o gerador de excitação envia um sinal que é passado pelo filtro de síntese para resultar no sinal reconstruído. O erro entre o sinal reconstruído e o sinal original é passado por um filtro perceptual com o objetivo de atenuar o erro nas regiões do espectro



(*a*)



(*b*)

Figura 2.5: Diagrama de blocos do *vocoder* LPC: (a) codificador; (b) decodificador.

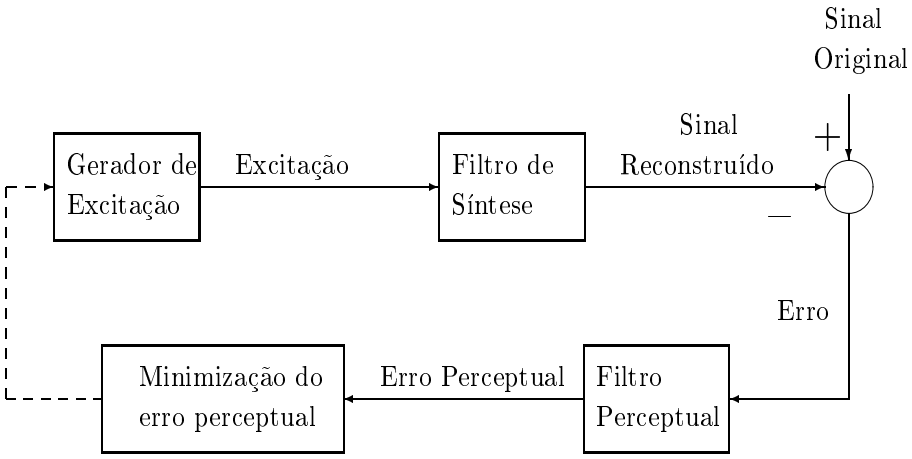


Figura 2.6: Diagrama de blocos dos codificadores híbridos que utilizam o procedimento de análise-por-síntese.

de freqüência onde o sinal da fala possui mais energia e enfatizar naquelas onde ele possui menos. Esse erro perceptual é usado pelo gerador de excitação para selecionar a melhor excitação. O filtro de síntese geralmente é o mesmo usado no *vocoder* LPC. A diferença entre os dois sistemas está na melhor excitação proveniente dos codificadores híbridos, que minimizam o erro entre os sinais original e reconstruído.

Existem vários tipos de codificadores híbridos. Basicamente todas eles diferem na forma como são geradas as excitações. As três principais técnicas pertencentes a este grupo são: codificação por excitação de multi-pulsos (*multi-pulse excitation*, MPE), codificação por excitação de pulsos regulares (*regular pulse excitation*, RPE) e CELP. As duas primeiras serão brevemente mostradas a seguir, enquanto que a técnica CELP, o foco principal desta dissertação, será abordada com mais detalhes na seção seguinte.

2.3.3.1 Codificação por excitação de multi-pulsos (MPE)

No codificador MPE, a excitação é formada por impulsos discretos com amplitudes e localizações variáveis [4, 6]. Em geral são utilizados de 4 a 6 pulsos por bloco de aproximadamente 20 ms de fala, com amplitudes e posições que minimizam o erro médio quadrático perceptual. Para o receptor são enviadas as informações

de tais posições e amplitudes, além dos parâmetros do filtro de síntese para cada bloco da fala. Os codificadores MPE obtêm boa qualidade de voz a taxas acima de 9,6 kbps [6].

2.3.3.2 Codificação por excitação de pulsos regulares (RPE)

Os codificadores RPE geram um sinal de excitação onde todos os pulsos são espaçados com a mesma distância. A informação que deve ser enviada ao receptor é constituída pela posição do primeiro pulso e as amplitudes de cada um deles. A técnica RPE obtém qualidade igual à técnica MPE, a taxas menores.

O sistema global para telecomunicação móvel (*global system for mobile telecommunication*, GSM) empregado na Europa utiliza em uma das suas versões a codificação RPE a uma taxa de 13 kbps [4, 6].

2.3.4 Comparação entre os grupos de codificadores

O gráfico da Figura 2.7 ilustra o comportamento qualidade×taxa para os codificadores de forma de onda, fonte e híbridos. Pode-se perceber que os codificadores de forma de onda não conseguem obter boa qualidade a taxas abaixo de 16 kbps. Enquanto isso, os codificadores fonte não obtêm boa qualidade mesmo aumentando-se indefinidamente sua taxa. Os codificadores híbridos aparecem como a melhor alternativa quando se deseja obter boa qualidade a baixas taxas.

2.4 A técnica CELP

A técnica CELP pertence ao grupo dos codificadores híbridos que empregam o procedimento de análise-por-síntese [7]. Atualmente ela consegue obter fala decodificada de boa qualidade a taxas entre 4 e 16 kbps.

A primeira aparição da técnica CELP foi no trabalho de Schroeder e Atal [8], onde se tratava de um sistema de difícil implementação em tempo real devido à enorme complexidade computacional. Somente com algumas melhorias no algoritmo [9, 10, 11, 12, 13, 14] e o desenvolvimento dos processadores de sinal digital mais

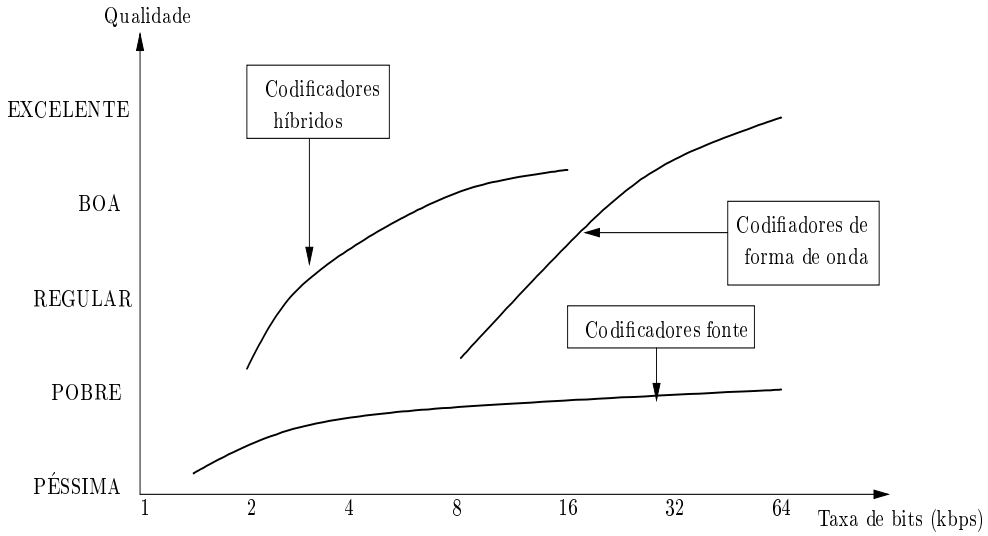


Figura 2.7: Comportamento taxa×qualidade dos principais grupos de codificadores da fala (figura reproduzida de [1]).

modernos é que foi possível obter codificação CELP em tempo real [7].

2.4.1 A idéia inicial

Em um segmento de sinal da fala digital existem correlações entre amostras vizinhas, devido ao efeito do trato vocal humano, que são geralmente chamadas de correlações de curto termo (do inglês *short-term correlations*). Além disso, se o segmento considerado for sonoro, existem também correlações entre amostras da ordem de períodos de *pitch* que são geralmente chamadas de correlações de longo termo (do inglês *long-term correlations*). Se o segmento for surdo, somente correlações de curto termo existem porque, conforme falado na Sub-seção 2.2.1, os sons surdos são produzidos por uma excitação ruidosa filtrada pelo trato vocal [3].

Atal mostrou em [15] que se forem removidas as correlações de longo e curto termo de determinado segmento de sinal de fala sonoro, o resultado será um sinal com características semelhantes a um ruído branco gaussiano. A Figura 2.8 mostra como pode ser feito este processo de "branqueamento". Primeiro é realizada a remoção das correlações de curto termo através da passagem do sinal $s(n)$ pelo filtro $A(z)$, que é o inverso do filtro que modela os efeitos do trato vocal, isto é, o inverso do

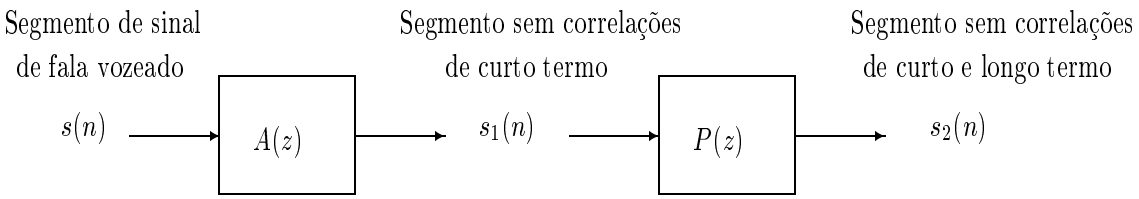
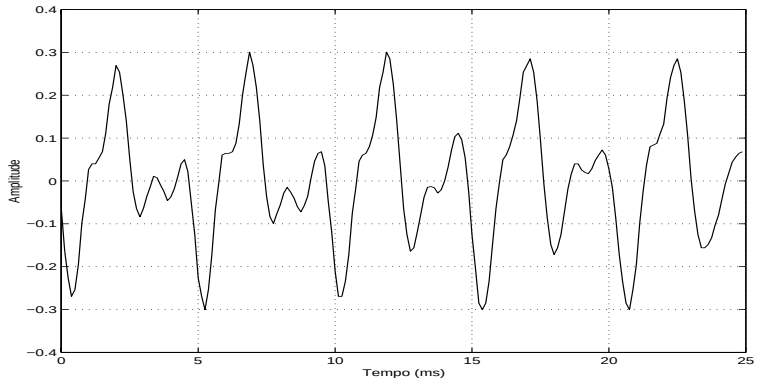


Figura 2.8: Processo de remoção de correlações de um segmento de sinal de fala.

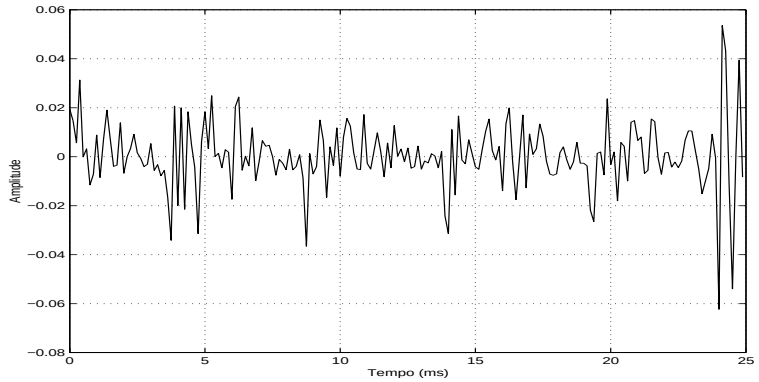
filtro de síntese presente no *vocoder* LPC. A remoção das correlações de longo termo é feita através da passagem do segmento de sinal da fala resultante $s_1(n)$ pelo filtro $P(z)$, que é o inverso do chamado filtro de *pitch* [16, 17]. O sinal sem correlações resultante $s_2(n)$ é freqüentemente chamado de *sinal residual*. As características dos filtros $P(z)$ e $A(z)$ serão vistas mais adiante. A Figura 2.9 mostra um segmento de sinal sonoro de 25 ms, amostrado a 8 kHz, e o resultante sinal residual.

O primeiro sistema CELP, publicado por Schroeder e Atal em [8], foi projetado baseando-se nesta característica. A Figura 2.10 mostra o diagrama de blocos deste sistema. Um dicionário armazena um conjunto de sinais residuais $x_{fI}(n)$, onde I é o índice do respectivo sinal no dicionário. O sinal reconstruído $\hat{s}(n)$ é obtido ao passar o sinal residual $x_{fI}(n)$ pelos filtros de *pitch* $G(z) = 1/P(z)$ e de síntese (trato vocal) $H(z) = 1/A(z)$, respectivamente. O primeiro introduz as correlações de longo termo no sinal residual de acordo com o período de *pitch* pré-calculado; o segundo introduz as correlações de curto termo. Para cada sinal residual armazenado no dicionário é determinado um erro perceptual $e_w(n)$, obtido ao passar o sinal de erro $e(n)$ pelo filtro perceptual $W(z) = A(z)/A(z/\gamma)$, cuja energia $\sum e_w^2(n)$ é usada para determinar o melhor sinal residual dentre os existentes no dicionário. A constante γ representa o fator perceptual do filtro $W(z)$, cujo significado será melhor explicado mais adiante. O índice relacionado ao melhor sinal é enviado ao decodificador, juntamente com os parâmetros dos filtros $P(z)$ e $A(z)$, e o ganho G_f.

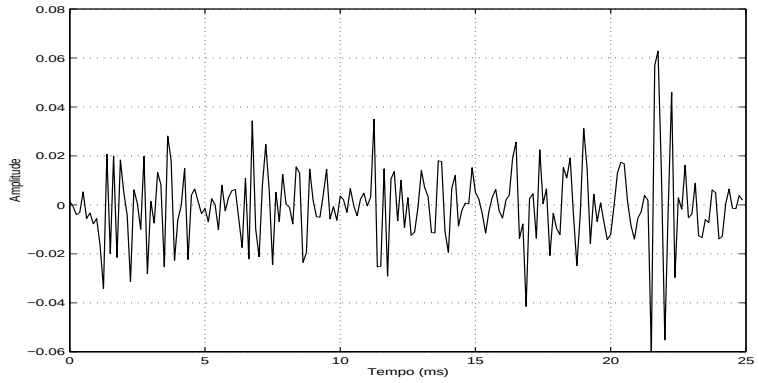
Os sistemas CELP processam os sinais da fala em blocos que posteriormente são divididos em sub-blocos, onde para cada um deles é obtida a melhor excitação para o filtro de síntese $H(z)$, cujos coeficientes são determinados a cada bloco, ou seja, valem para todos os sub-blocos contidos no respectivo bloco.



(a)



(b)



(c)

Figura 2.9: Exemplos de remoção de correlações: (a) segmento sonoro de sinal de fala; (b) segmento mostrado em (a) sem correlações de curto termo; (c) segmento mostrado em (a) sem correlações de curto e longo termo.

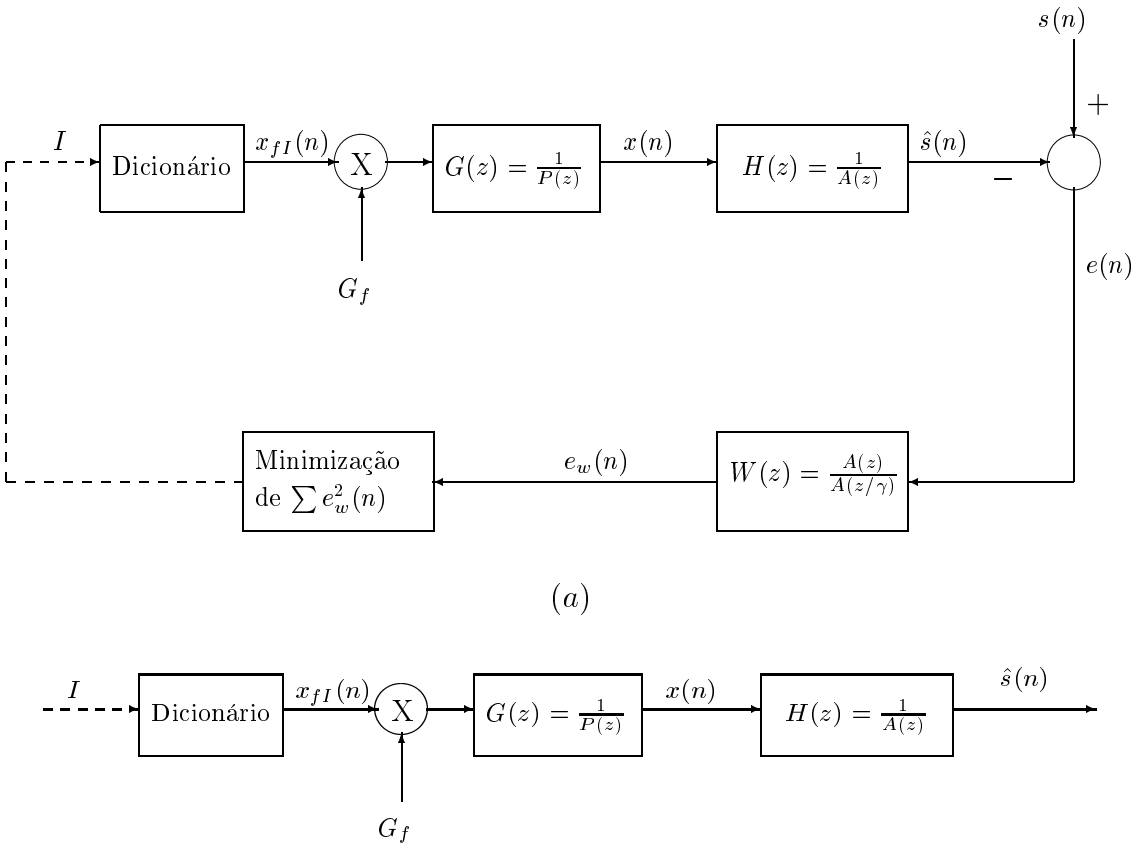


Figura 2.10: Primeiro sistema de codificação CELP: (a) codificador; (b) decodificador.

A qualidade deste sistema CELP inicial estava muito relacionada ao tamanho do dicionário usado. Devido a esse fato, para obter boa qualidade os codificadores CELP tomavam um enorme tempo para processar um sinal de fala. Muitas mudanças foram sendo feitas a partir do sistema mostrado na Figura 2.10 no intuito de diminuir a complexidade computacional e melhorar a qualidade da fala reconstruída. A seguir será tratado o sistema CELP com uma estrutura considerada como padrão atualmente.

2.4.2 CELP com dicionário adaptativo

O diagrama de um sistema de codificação CELP com dicionário adaptativo está mostrado na Figura 2.11. Pode-se perceber que o filtro de *pitch* $G(z)$ da Figura 2.10 foi substituído pelo dicionário adaptativo. A seguir será explicada cada uma das partes constituintes do sistema da Figura 2.11.

2.4.2.1 O filtro de síntese $H(z)$

O filtro de síntese $H(z)$ é da forma

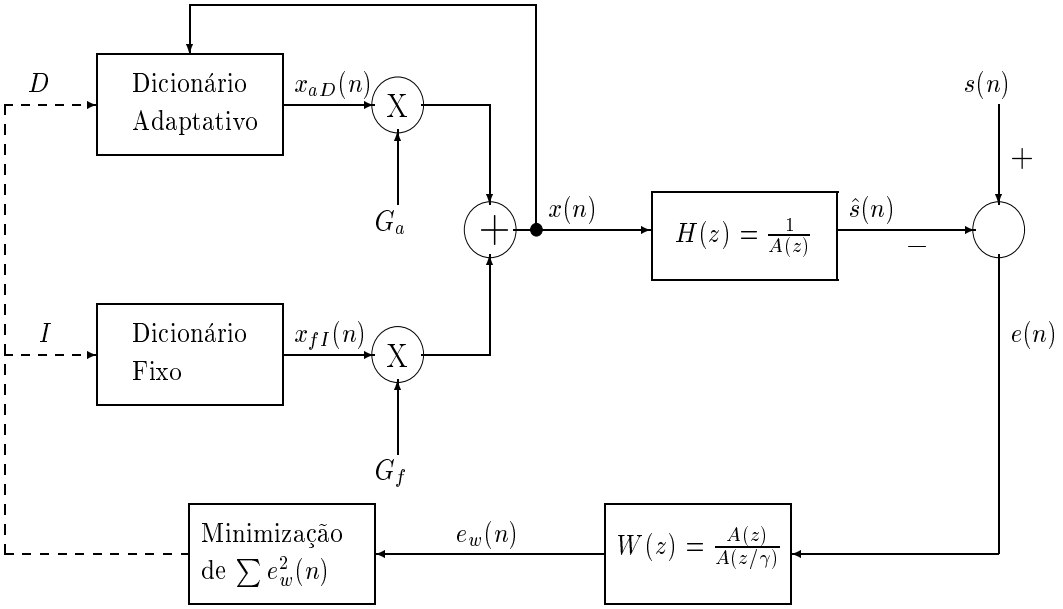
$$H(z) = \frac{1}{A(z)}, \tag{2.2}$$

onde o fitro inverso $A(z)$ é dado por

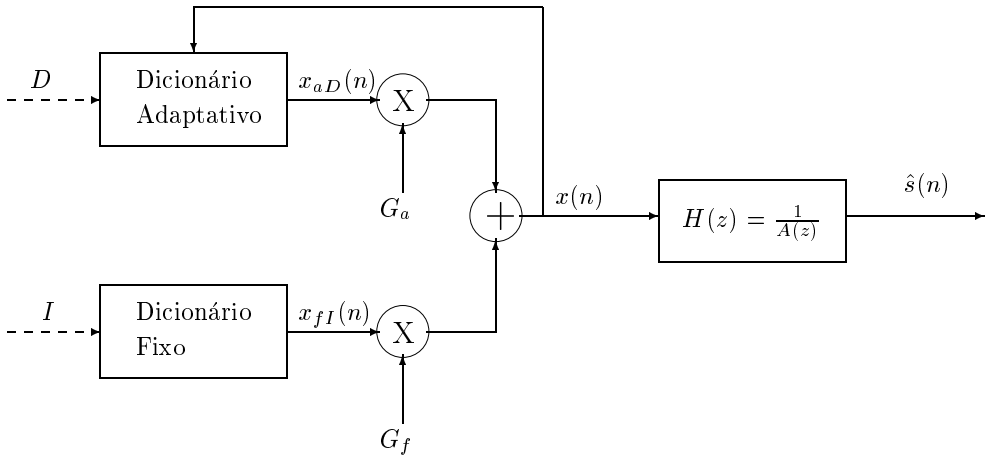
$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}. \tag{2.3}$$

Os coeficientes $\{a_1, a_2, \ldots, a_p\}$ são os coeficientes de predição linear (CPL), obtidos através de uma análise LPC [2, 4, 18, 19].

O filtro $H(z)$ representa os efeitos do trato vocal e é responsável pela introdução de correlações de curto termo. A constante p representa a ordem do modelo LPC e define a precisão na qual $H(z)$ modela os efeitos do trato vocal. Geralmente os codificadores CELP utilizam $p = 10$, onde se obtém um bom compromisso entre qualidade e taxa de bits.



(a)



(b)

Figura 2.11: Sistema de codificação CELP com dicionário adaptativo: (a) codificador; (b) decodificador.

2.4.2.2 O filtro de ponderação do erro $W(z)$

O filtro $W(z)$ serve para modificar o espectro de freqüência do sinal de erro $e(n)$ entre os sinais original $s(n)$ e reconstruído $\hat{s}(n)$. É geralmente nos vales dos espectros de amplitude dos sinais de fala que o erro produz maior distorção audível, segundo o fenômeno do "mascaramento auditivo" [15, 20]. O que o filtro $W(z)$ faz é acentuar a energia deste erro nos vales e reduzir nas regiões formânticas. A Figura 2.12 mostra os espectros de $H(z)$ e $W(z)$, onde se pode perceber o que foi afirmado. A função de transferência de $W(z)$ é dada por

$$W(z) = \frac{A(z)}{A(z/\gamma)}, \tag{2.4}$$

onde o fator perceptual $\gamma \in (0, 1)$ é quem indica o grau de mudança do espectro do erro $e(n)$. Segundo [8], ele pode ser determinado por

$$\gamma = e^{-2\pi 100/f_s}, \tag{2.5}$$

onde f_s corresponde à freqüência de amostragem do sinal da fala digital a ser processado. Em geral, boa parte dos sistemas CELP utilizam um fator perceprual fixo e igual a $\gamma = 0,8$ [10, 14, 21].

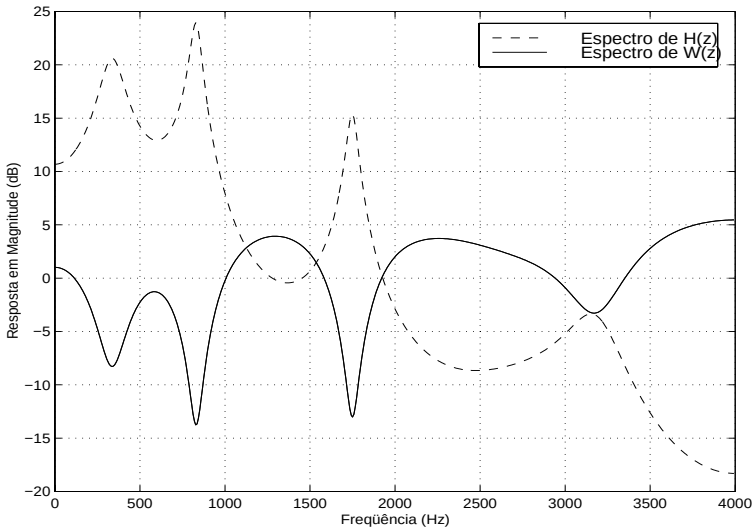


Figura 2.12: Espectros de amplitude dos filtros $H(z)$ e $W(z)$ para um segmento sonoro de um sinal de fala.

2.4.2.3 O dicionário fixo

O dicionário fixo $\mathbf{C_f} = \{\{x_{f0}(n)\}, \{x_{f1}(n)\}, \ldots, \{x_{fK_f-1}(n)\}\}$ representa o mesmo dicionário existente no sistema CELP da Figura 2.10. Ele armazena K_f seqüências $x_{fI}(n)$, onde I indica o índice da respectiva seqüência, que serão usadas para reconstruir várias versões do mesmo sub-bloco do sinal de fala.

Existem vários tipos de dicionário fixo. Nos primeiros sistemas CELP, ele era geralmente composto por um conjunto de seqüências de ruído branco gaussiano de média zero e variância unitária, devido à propriedade mostrada em [15]. Atualmente, dicionários cujas seqüências possuem grande quantidade de elementos nulos proporcionam uma melhor rapidez durante o processo de codificação, além de melhorar a qualidade. Segundo [10], dicionários cujas amostras são o resultado de ceifagem de seqüências gaussianas proporcionam melhor qualidade da fala reconstruída. Existem ainda os dicionários ternários cujas amostras possuem somente os valores -1, 0, e 1 para facilitar as operações de filtragem, como o existente no sistema DoD-CELP (*Department of Defense CELP*, CELP do Departamento de Defesa) [6], que será tratado mais adiante. Os dicionários treinados [22], cujas seqüências são projetadas tendo em vista minimizar os erros entre os sinais originais e reconstruídos de uma determinada base de dados, melhoram bastante a qualidade e representam uma das melhores opções quando se deseja utilizar um dicionário com poucas seqüências, tendo em vista diminuir a taxa de bits e a complexidade computacional.

2.4.2.4 O dicionário adaptativo

O dicionário adaptativo $\mathbf{C_a} = \{\{x_{a0}(n)\}, \{x_{a1}(n)\}, \ldots, \{x_{aK_a-1}(n)\}\}$, de um ponto de vista teórico, armazena K_a seqüências $x_{aD}(n)$, onde D representa o respectivo índice de cada uma delas. Tal conjunto de seqüências é renovado constantemente, conforme é indicado pela realimentação mostrada na Figura 2.11. Como no caso do dicionário fixo, cada seqüência é usada para determinar várias versões do mesmo sub-bloco do sinal de fala.

Conceitualmente, o dicionário adaptativo representa uma melhor forma de visualizar o filtro de *pitch* $G(z) = 1/P(z)$, existente nos primeiros sistemas CELP,

quando seus parâmetros são determinados através do procedimento de análise-por-síntese [6, 11]. O filtro $P(z)$, mostrado na Figura 2.10, é da forma

$$P(z) = 1 - \sum_{i=-(q-1)/2}^{(q-1)/2} b_i z^{-D-i}, \tag{2.6}$$

onde b_i são os coeficientes de predição de longo termo e D é o período de *pitch* em amostras. Usualmente se faz $q = 1$ ou $q = 3$ em (2.6), ou seja, usa-se um ou três coeficientes. A eficiência do filtro $G(z)$ quando $P(z)$ tem somente um coeficiente, ou seja, $P(z) = 1 - bz^{-D}$, é aumentada significantemente se seus parâmetros b e D forem escolhidos de maneira a minimizar a energia do sinal de erro perceptual $e_w(n)$. Neste caso, este preditor comporta-se como um dicionário que armazena um conjunto de seqüências candidatas [6], e por isso é substituído pelo dicionário adaptativo mostrado na Figura 2.11. Pode-se citar duas importantes vantagens de utilizar o dicionário adaptativo: a primeira é determinação dos parâmetros de maneira a minimizar o erro perceptual, e a segunda é que os ganhos G_a e G_f podem ser otimizados simultaneamente [5, 23].

Na discursão que segue, a distinção entre amostras de seqüências e seqüências será feita de acordo com o contexto. Por exemplo, em $f(n) = \{x(D), \ldots, x(D+N)\}$, $f(n)$ representa uma seqüência e não uma amostra.

Cada seqüência candidata do dicionário adaptativo é dada por

$$x_{aD}(n) = x(n - D), \quad \text{para} \quad 0 \leq n \leq N - 1, \tag{2.7}$$

onde N é o tamanho de cada sub-bloco em amostras. Como se pode perceber, cada seqüência candidata é obtida por versões atrasadas do sinal de excitação $x(n)$, onde durante o procedimento de codificação cada atraso D é testado dentro de uma faixa pré-fixada $[D_{min}, D_{max}]$ que irá depender do codificador. Portanto, na prática, o dicionário adaptativo corresponde a uma seqüência contendo as amostras da excitação passada, ou seja,

$$C_a(n) = \{x(-D_{max}), x(-D_{max} + 1), \ldots, x(-2), x(-1)\}, \tag{2.8}$$

e para um determinado atraso D, a seqüência candidata é

$$x_{aD}(n) = \{x(-D), \ldots, x(-D + N - 1)\}. \tag{2.9}$$

Neste instante deve surgir a questão: "E se o atraso D for menor que o tamanho de um subloco N?" Justamente é esta a única diferença entre o dicionário adaptativo e o filtro de *pitch*, porque se $D < N$, então

$$x_{aD}(n) = \{x(-D), \ldots, x(-1), x(0), \ldots, x(N-D-1)\}. \tag{2.10}$$

Mas durante o processo de codificação de um sub-bloco, a amostras de $x(n)$ para $n \geq 0$ não são conhecidas. Uma das formas encontradas para solucionar isto [11] é repetir as primeiras amostras de $x_{aD}(n)$ de forma a torná-la periódica (ou quase) com período D, ou seja, a seqüência candidata passa a ser dada por

$$x_{aD}(n) = x\left(n - \left\lfloor \frac{n+D}{D} \right\rfloor D\right), \quad \text{para} \quad 0 \leq n \leq N-1, \tag{2.11}$$

onde $\lfloor c \rfloor$ significa "maior inteiro menor ou igual a c". Conforme se pode notar, para $0 \leq n \leq (D-1)$ é usado um atraso de D amostras. Para $D \leq n \leq (2D-1)$, o atraso passa a ser de $2D$ amostras, e assim sucessivamente.

Os filtros de *pitch* com 3 coeficientes produzem fala reconstruída de melhor qualidade, principalmente para os sinais femininos [24]. O problema da utilização do filtro de terceira ordem é o aumento na quantidade de parâmetros que devem ser codificados, além também de um aumento na complexidade para a determinação dos coeficientes. Em [24] é proposto um filtro de *pitch* de primeira ordem que obtém performance semelhante ao filtro com três coeficientes, utilizando atrasos correspondentes a frações de uma amostra. Os codificadores CELP atuais geralmente utilizam um dicionário adaptativo com atrasos fracionários, que é baseado neste tipo de filtro de *pitch* [10, 21]. A idéia consiste em determinar atrasos do tipo $D\frac{1}{d}$, ou seja, com frações de amostras. A desvantagem deste tipo de dicionário é que existe um aumento no número de seqüências candidatas e uma maior complexidade durante o procedimento de busca da melhor excitação. No entanto, a melhora é bastante considerável, o que pode até permitir a redução do número de amostras contidas no dicionário fixo. Em [5] é descrito como é feito o procedimento de busca da melhor excitação em um dicionário adaptativo com atrasos fracionários.

2.5 Padrões de codificação CELP

Existem algumas padronizações da técnica CELP usadas principalmente para sistemas de comunicação móvel. Algumas delas são brevemente descritas a seguir.

2.5.1 DoD-CELP a 4,8 kbps

Em 1984 o Departamento de Defesa (*Department of Defense*, DoD) dos Estados Unidos lançou um programa para desenvolver uma terceira geração de telefonia segura, tendo em vista substituir o LPC-10 [6], cuja fala reconstruída soava sintética de forma a impossibilitar o reconhecimento do locutor. Em 1988 foi selecionado um sistema CELP conjuntamente desenvolvido pelos Laboratórios AT&T Bell e o DoD, que foi melhorado e regulamentado como padrão federal 1016 (*federal standard 1016*, FS-1016).

O DoD-CELP usa uma estrutura igual àquela mostrada na Figura 2.11, ou seja, um dicionário fixo e outro adaptativo produzem a excitação para o filtro de síntese. Neste sistema são utilizados blocos de 30 ms que são posteriormente divididos em sub-blocos de 7,5 ms. Para cada bloco são determinados 10 CPL que são transformados em freqüências do espectro de linha (*line spectral frequencies*, LSF), onde é realizada uma quantização escalar com um total de 34 bits. A excitação para o filtro de síntese é determinada para cada sub-bloco usando um dicionário adaptativo com 256 possíveis atrasos, sendo 128 inteiros e 128 não inteiros, e um dicionário fixo ternário, onde 77% de suas amostras são nulas. Ele é composto de 1082 amostras, das quais são obtidas 512 seqüências candidatas, cada uma com 60 amostras. A forma encontrada para isso é o uso de superposição, onde cada nova seqüência difere da anterior somente por 4 amostras: as duas primeiras e as duas últimas, ou seja, para a geração de cada seqüência existe um atraso de duas amostras na seqüência de tamanho 1082. A excitação é determinada através do procedimento de análise-por-síntese para minimizar o erro perceptual entre as falas original e reconstruída, e os respectivos ganhos são quantizados escalarmente com 5 bits cada um. Em sub-blocos ímpares o índice do dicionário adaptativo é codificado com 8

bits, mas para reduzir a complexidade e a taxa de bits, em sub-blocos pares ele é codificado de forma diferencial com 6 bits. Além disto, 1 bit por bloco é usado para sincronização, e 4 bits são usados para prover correção de erro para os bits mais sensitivos transmitidos. Finalmente mais 1 bit por bloco é usado para futura expansão, de forma que o sistema não se torne obsoleto conforme a tecnologia avança.

No decodificador o fluxo de bits recebido é usado para formar os coeficientes do filtro de síntese, e para selecionar a excitação (gerada a partir dos dicionários) para este filtro a fim de produzir a fala reconstruída. Um pós-filtro adaptativo é então aplicado para melhorar a qualidade.

A qualidade gerada pelo sistema DoD-CELP é considerada regular, onde no teste subjetivo de pontos de opinião média (*mean opinion score*, MOS) a nota conseguida é igual a 3,2 [25]. Neste teste a escala vai de 1,0 (qualidade inaceitável) a 5,0 (qualidade excelente).

2.5.2 LD-CELP a 16 kbps

Um dos problemas de codificação da fala é o atraso de codificação, que é basicamente o tempo decorrido desde o instante em que uma amostra do sinal chega na entrada do codificador até o instante que ela se encontra na saída no decodificador. Esta definição não inclui atrasos relativos ao modem ou ao canal de transmissão. De uma forma aproximada, o atraso de um sistema de codificação que processa o sinal de fala bloco a bloco, tal como os codificadores híbridos, fica entre 2 e 4 blocos.

Em 1988 o CCITT (recentemente renomeado para ITU-T) lançou uma proposta para a codificação a uma taxa de 16 kbps com baixo retardo, para uma possível regulamentação da série G. para aplicações universais. Dentre os requisitos exigidos, estava que a qualidade não fosse pior que o ADPCM a 32 kbps, que faz parte da regulamentação G.721 [4, 6], e que o atraso de codificação máximo fosse 5 ms. Dentre todos os candidatos, o CELP com baixo retardo (*low delay CELP*, LD-CELP) proposto por Chen *et al.* [26] foi selecionado.

O LD-CELP obtém baixo retardo ao usar um preditor retroativo-adaptativo (do inglês *backward-adaptive*) e vetores de excitação curtos (5 amostras). Na pre-

dição retroativa-adaptativa, os CPL são determinados a cada bloco ao operar nas amostras de fala anteriormente processadas que estão disponíveis no decodificador. O LD-CELP não utiliza dicionário adaptativo ou filtro de *pitch*; ao invés disso a ordem do preditor de curto termo é aumentada para 50. O tamanho de cada bloco é de 2,5 ms, que é dividido em sub-blocos de 0,625 ms. A excitação para o filtro de síntese é obtida a partir de um dicionário contendo 128 seqüências, escaladas por um ganho. A cada sub-bloco, 7 bits são usados para codificar o índice da melhor seqüência do dicionário e 3 bits para o ganho, que é determinado através da excitação quantizada anteriormente usando um preditor de ordem 10, cujos coeficientes são atualizados a cada bloco. O filtro perceptual é de ordem 10, e é atualizado a cada bloco também. Sua função de transferência é mais geral do que aquela mostrada em (2.4). A fala decodificada ainda é passada por um pós-filtro a fim de enfatizar as regiões formânticas.

O retardo de codificação do LD-CELP, que foi regulamentado como G.728, é menor que 2 ms, e sua qualidade foi julgada como equivalente ou melhor que o sistema ADPCM a 32 kbps. O teste de MOS atribuiu-lhe a nota de 4,0 [25], o que representa uma qualidade de boa para excelente.

2.5.3 CS-ACELP a 8 kbps

Em 1990 o CCITT começou a avaliar codificadores candidatos para um padrão de codificação da fala a 8 kbps com baixo retardo. Os requisitos especificados diziam respeito à qualidade da fala reconstruída, robustez a erros de canal e tamanho do bloco de fala processado. Até julho de 1991 nenhum sistema havia satisfeito tais requisitos, de forma que em novembro do mesmo ano o tamanho máximo do bloco passou de 5 ms para 16 ms. Em novembro de 1992 dois candidatos foram submetidos: um do Japão consistindo de um algoritmo CELP com estrutura conjugada (*conjugate structure CELP*, CS-CELP), cujos blocos tinham duração de 13 ms; e outro projetado pela France Telecom e University of Sherbrook no Canadá, e consistia de um CELP algébrico (*algebraic CELP*, ACELP), cujos blocos tinham 12 ms de duração. Foi decidido então que aplicações considerando blocos de 10 ms seriam

preferíveis, e assim ambos os grupos decidiram reduzir o tamanho dos blocos para 10 ms. Aspectos tanto do CS-CELP como do ACELP foram usados na versão final que foi padronizada, que consiste de um CELP algébrico de estrutura conjugada (*conjugate structure algebraic CELP*, CS-ACELP).

A estrutura do CS-ACELP é semelhante àquela mostrada na Figura 2.11. Os parâmetros do filtro de síntese são determinados a cada bloco de 10 ms através de uma análise LPC e depois convertidos para LSF, onde depois é realizada uma quantização vetorial de 2 estágios usando predição de coeficientes, com um total de 18 bits por conjunto. Cada bloco é dividido em sub-blocos de 5 ms, sendo a excitação para o filtro de síntese determinada a cada intervalo desta duração pelos dicionários adaptativo e fixo. O primeiro deles é constituído de 256 possíveis atrasos, contando os inteiros e fracionários, onde 8 bits são usados para representá-los no primeiro sub-bloco e 5 bits no segundo, de forma diferencial. Para melhorar a robustez a erro de canal, os 6 bits mais significativos que representam o índice do dicionário adaptativo no primeiro sub-bloco têm 1 bit de paridade adicionado. Isto permite que erros nestes bits sejam detectados no decodificador. O dicionário fixo é do tipo algébrico e contém 2^{17} seqüências possíveis, sendo portanto usados 17 bits para representar cada uma delas para cada sub-bloco, onde é usado um procedimento rápido de procura. Os ganhos dos dicionários são quantizados vetorialmente com 7 bits usando um dicionário estruturado conjugado de dois estágios, usando predição para o ganho do dicionário fixo através de um filtro FIR de quarta ordem, a fim de melhorar a eficiência da quantização. Os índices dos ganhos são determinados a cada sub-bloco usando o procedimento de análise-por-síntese para minimizar o erro perceptual entre as falas original e reconstruída, tal como são feitas as escolhas das melhores excitações.

No decodificador os parâmetros transmitidos são usados para compor o filtro de síntese, e para selecionar os índices dos dicionários e ganhos para representar a sua excitação. A fala reconstruída é então pós-processada para melhorar sua qualidade perceptual.

A padronização do CS-ACELP ficou conhecida como G.729. A qualidade

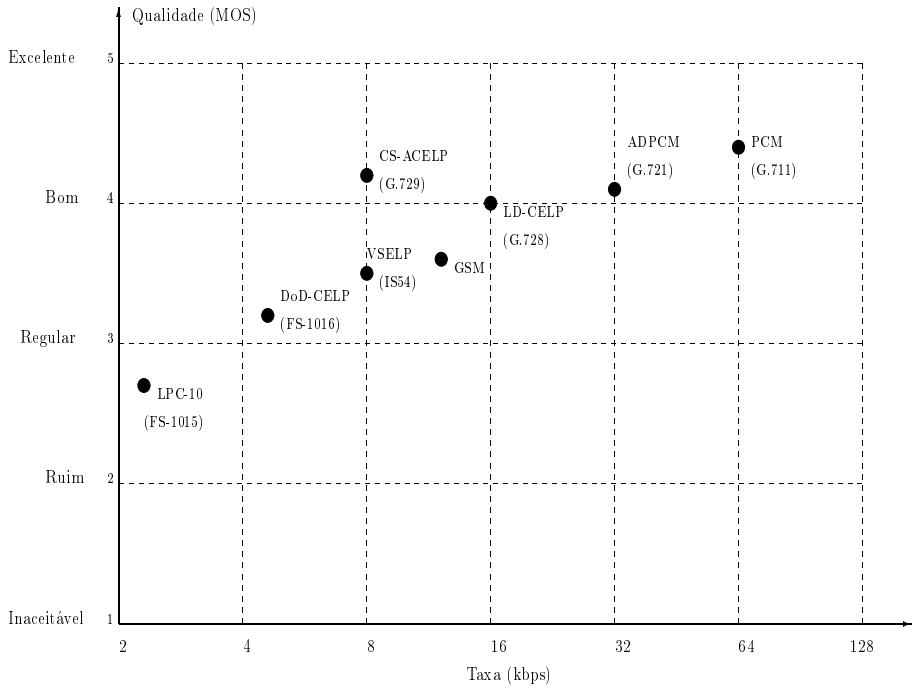


Figura 2.13: Comportamento taxa×qualidade de alguns algoritmos padrões de codificação da fala.

deste sistema, que na escala MOS a nota é 4,1 [25], ficou comparável ao LD-CELP a 16 kbps, perdendo somente no quesito atraso de codificação. Além da taxa de bits, outras vantagens do CS-ACELP sobre o LD-CELP são a robustez a erros de canal e menor complexidade.

2.5.4 Qualidade subjetiva dos padrões

O gráfico mostrado na Figura 2.13 mostra o comportamento taxa×qualidade para alguns padrões de codificação, onde a qualidade é baseada no teste de MOS. O padrão indicado por IS54 representa o algoritmo VSELP (*vector-sum excited linear prediction*), ou predição linear excitada por soma vetorial, que é um tipo de CELP onde existem mais de um dicionário fixo, cujas seqüências são combinações lineares de uma determinada. Este sistema é utilizado nos Estados Unidos como padrão para o sistema de telefonia móvel digital. Maiores informações e detalhes a respeito dele podem ser vistos em [27].

Pode-se reparar que o G.728 e G.729 possuem qualidade comparável ao G.711, que é tido como referência para a codificação da fala. Enquanto isso, o IS54, que opera na mesma taxa que o G.729, tem qualidade bem inferior. Já o FS-1016 possui qualidade apenas razoável.

2.6 Conclusão

Este capítulo apresentou alguns conceitos, cuja abordagem foi desde características da produção da fala humana a alguns sistemas padrões de codificação. Foi dada bastante ênfase à técnica CELP, onde foram explicadas detalhadamente suas partes constituintes e respectivas influências na qualidade do sinal decodificado.

Capítulo 3

Quantização dos parâmetros do filtro de síntese

3.1 Introdução

Este capítulo apresenta algumas formas de quantização escalar dos parâmetros do filtro de síntese presente em codificadores CELP e outros baseados no modelo excitação-filtro, e as respectivas avaliações feitas com base em uma medida de distorção espectral.

A Seção 3.2 trata do problema da quantização dos parâmetros do filtro de síntese; a Seção 3.3 mostra parâmetros alternativos para a sua representação; a Seção 3.4 trata do projeto de quantizadores baseados em alguns destes parâmetros; a Seção 3.5 diz respeito às avaliações; e na Seção 3.6 estão as conclusões.

3.2 O problema de quantização

O filtro de síntese

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \tag{3.1}$$

mostrado nas Figuras 2.10 e 2.11 é responsável pela envoltória espectral de curto termo do sinal de fala que está sendo codificado, modelando os efeitos do trato

vocal humano [6]. Os coeficientes $\{a_1, a_2, \ldots, a_p\}$ do filtro inverso $A(z)$ são obtidos através de uma análise LPC de ordem p [2, 4, 18, 19]. Para a codificação CELP, ou qualquer outra que utilize o filtro $H(z)$, é necessário que tais coeficientes sejam quantizados a fim de serem enviados ao decodificador.

Sabe-se da literatura que os coeficientes preditores a_i não são bons para quantização devido principalmente a dois motivos [28]:

- possuem uma ampla faixa dinâmica;
- são muito sensíveis a erros de quantização, podendo até mesmo tornar a versão quantizada de $H(z)$ um filtro instável.

Para resolver estes dois problemas, seria necessário alocar uma grande quantidade de bits (cerca de 16 bits para cada coeficiente [14]), o que é impraticável para sistemas de codificação a baixas taxas.

Atualmente os codificadores CELP utilizam de 18 a 40 bits para quantizar eficientemente um conjunto de 10 coeficientes que caracterizam o filtro $H(z)$ [10, 13]. A forma encontrada para esta redução é a utilização de outros tipos de parâmetros, que devem apresentar as seguintes propriedades [28]:

- a versão quantizada dos parâmetros deve corresponder a um filtro $H(z)$ estável, ou pelo menos permitir facilmente a verificação desta estabilidade;
- os parâmetros devem ter uma ordenação natural, ou seja, caso seja modificada a ordem ou permutados os coeficientes, o filtro produzido deve ser diferente de $H(z)$.

A primeira propriedade implica que os pólos de $H(z)$ devem continuar dentro do círculo de raio unitário mesmo depois da quantização dos coeficientes, ou então permitir que isto seja verificado de forma simples. A segunda propriedade implica que os parâmetros tenham uma ordenação natural. Por exemplo, se os parâmetros alternativos f_i são ordenados como $f_1, f_2, \ldots, f_p$, e depois f_1 e f_2 são permutados, então o filtro produzido com esta nova ordenação passa a ser diferente de $H(z)$.

3.3 Parametrização de $H(z)$

Os parâmetros que representam o filtro $H(z)$ mais utilizados pelos codificadores CELP são: coeficientes de reflexão, logaritmos da razão das áreas e freqüências do espectro de linha. Todos eles possuem as duas propriedades descritas anteriormente [28, 29].

3.3.1 Coeficientes de reflexão

Os coeficientes de reflexão (CR) k_i são obtidos como subproduto do algoritmo de Levinson-Durbin para a determinação dos coeficientes preditores a_i pelo método da autocorrelação. Os CR também podem ser obtidos a partir destes coeficientes através da seguinte recursão [28]:

$$\begin{aligned} k_i &= a_i^{(i)} \\ a_j^{(i-1)} &= \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \qquad \text{para} \quad 1 \leq j \leq i-1 \end{aligned} \tag{3.2}$$

onde o índice i assume os valores $p, p-1, \ldots, 1$, nessa ordem, e inicialmente $a_j^{(p)} = a_j$, para $1 \leq j \leq p$, sendo a_j^i o coeficiente a_j na recursão i.

Considerando que o trato vocal pode ser modelado pela conexão de vários tubos com áreas diferentes e mesmos comprimentos, como mostrado na Figura 3.1, o coeficiente k_i representa uma medida que indica a porção de ar refletida na conexão das seções i e $i+1$ [2, 4].

Os CR ainda são bastante utilizados [27] devido às suas limitadas faixas dinâmicas que se situam no intervalo $(-1; +1)$ para um filtro $H(z)$ estável, sendo esta uma maneira simples de verificar estabilidade. Além disso, os codificadores que utilizam os CR não precisam fazer a conversão destes para os coeficientes a_i porque eles implementam o filtro $H(z)$ na forma de treliça [2, 4]. No entanto, os principais problemas da utilização dos CR para representar o filtro $H(z)$ são:

- desigual sensibilidade de quantização, ou seja, os primeiros CR são mais importantes que os demais, devendo assim ser utilizada uma distribuição não uniforme de bits;

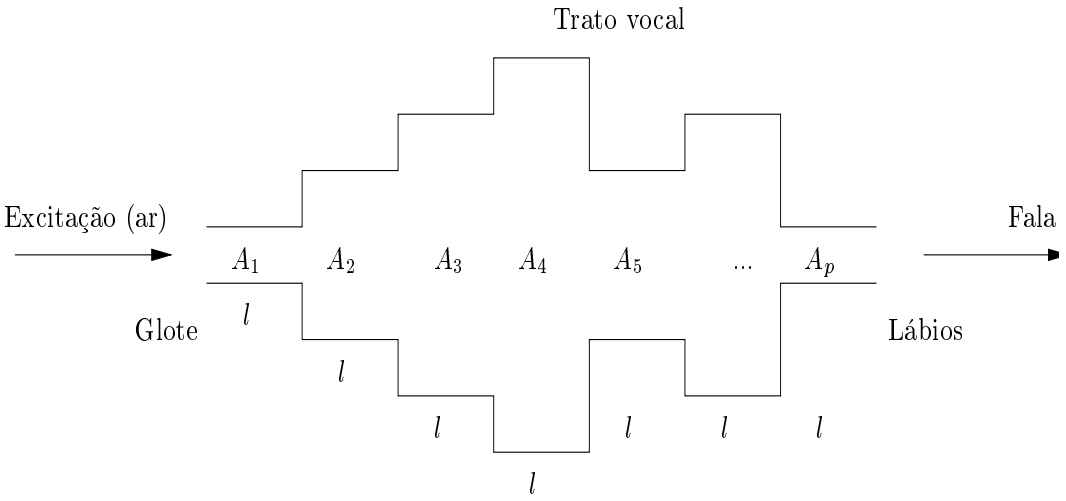


Figura 3.1: Modelo de tubos com iguais comprimentos para o trato vocal humano.

- maior sensibilidade a erros de quantização para valores próximos de +1 e -1, segundo análise de sensitividade mostrada em [28].

Apesar destes dois problemas, o algoritmo VSELP utiliza quantização dos CR. Neste sistema cada conjunto de 10 coeficientes é quantizado de maneira escalar não uniforme com um total de 38 bits [27].

3.3.2 Logaritmos da razão das áreas

Os logaritmos da razão da áreas (*log area ratio*, LAR) l_i são obtidos a partir dos CR k_i por [6]

$$l_i = \ln \frac{1 + k_i}{1 - k_i}. \tag{3.3}$$

Para o modelo de tubos na Figura 3.1 vale a seguinte relação [28]:

$$A_i = A_{i+1} \frac{1 + k_i}{1 - k_i}, \quad \text{com} \quad A_{p+1} = 1, \quad \text{e} \quad 1 \leq i \leq p, \tag{3.4}$$

onde A_i e A_{i+1} são as áreas dos tubos nas seções i e $i+1$, respectivamente. Portanto:

$$l_i = \ln \frac{A_i}{A_{i+1}}. \tag{3.5}$$

Daí surge a expressão *logaritmo da razão das áreas*.

Os LAR foram derivados a partir dos CR como uma forma de eliminar o problema destes últimos com relação à maior sensibilidade a erros de quantização para valores próximos de -1 e +1. Segundo é mostrado em [28], uma quantização uniforme dos LAR representa o mesmo que uma quantização não uniforme bastante eficiente dos CR.

As vantagens da utilização dos LAR para representar $H(z)$ em relação aos CR são:

- sensibilidade uniforme ao longo de toda a faixa dinâmica, permitindo assim uma quantização uniforme [28];
- sensibilidade igual para todos os coeficientes quanto a erros de quantização [28];
- tratando-se de sistemas CELP, possuem melhor qualidade para interpolação entre sub-blocos, segundo afirmado em [5].

As desvantagens da utilização dos LAR em relação aos CR são:

- diferentes faixas dinâmicas para cada l_i, sendo este fato o responsável pela diferente quantidade de bits atribuída para cada coeficiente;
- maior tempo de computação para determinar os LAR, e a partir destes obter os CR.

A verificação de estabilidade sempre é realizada com os CR antes de serem convertidos em LAR. No sistema CELP proposto por [10] são utilizados LAR para representação de $H(z)$, onde cada conjunto de coeficientes é eficientemente quantizado de maneira linear não uniforme com 40 bits. Esta forma de quantização é diferente da proposta em [28], onde cada l_i é quantizado de maneira uniforme com um passo de quantização ótimo.

3.3.3 Freqüências do espectro de linha

As freqüências do espectro de linha [29] (*line spectral frequencies*, LSF) w_i são bastante utilizadas atualmente para representar o filtro de síntese $H(z)$ [13, 14, 21].

A seguir será mostrado como elas são obtidas a partir dele. Um polinômio simétrico $P(z)$ e outro anti-simétrico $Q(z)$ podem ser obtidos a partir de $A(z)$ dado por (3.1), através de

$$P(z) = A(z) + z^{-p-1}A(z^{-1}), \tag{3.6}$$

$$Q(z) = A(z) - z^{-p-1}A(z^{-1}), \tag{3.7}$$

onde p é o número de CPL. De (3.7) e (3.7) tem-se que

$$A(z) = \frac{P(z) + Q(z)}{2}. \tag{3.8}$$

É provado que se $A(z)$ é de fase mínima (todas as suas raízes estão dentro do círculo de raio unitário) então todas as raízes de $P(z)$ e $Q(z)$ são distintas e alternam-se em cima da circunferência de raio unitário (uma após a outra), sendo a recíproca também verdadeira [29]. Além disso, como $P(z)$ e $Q(z)$ são simétrico e anti-simétrico, eles possuem uma raiz cada em -1 e +1, respectivamente. Estas raízes podem ser removidas porque são redundantes:

$$\begin{aligned} P_1(z) = \frac{P(z)}{1+z^{-1}} \quad &\text{e} \quad Q_1(z) = \frac{Q(z)}{1-z^{-1}}, \quad \text{para} \quad p \quad \text{par}, \\ P_1(z) = P(z) \quad &\text{e} \quad Q_1(z) = \frac{Q(z)}{1-z^{-2}}, \quad \text{para} \quad p \quad \text{ímpar}. \end{aligned} \tag{3.9}$$

Os polinômios $P_1(z)$ e $Q_1(z)$ são simétricos de ordem par. Como as raízes ocorrem em pares de números complexos conjugados, somente metade delas precisa ser determinada. Assim, para p par, $p/2$ raízes de $P_1(z)$ mais $p/2$ raízes de $Q_1(z)$, totalizando p, podem representar os polinômios $P(z)$ e $Q(z)$, e conseqüentemente o filtro inverso $A(z)$. Para o caso p ímpar, $(p+1)/2$ raízes de $P_1(z)$ e $(p-1)/2$ raízes de $Q_1(z)$ representam os p coeficientes necessários para reconstruir $A(z)$. Como as p raízes $\{jw_1, \ldots, jw_p\}$ estão sobre a circunferência de raio unitário, somente os ângulos (ou freqüências) são necessários para representar $A(z)$. Tais freqüências $\{w_1, \ldots, w_p\}$ correspondem às LSF. Portanto, os índices ímpares representam os ângulos da metade das raízes de $P_1(z)$ que ficam na parte superior da circunferência de raio unitário, e os índices pares representam o mesmo caso, só que para $Q_1(z)$. A Figura 3.2 mostra o diagrama com as raízes de $A(z)$, $P(z)$ e $Q(z)$, representadas

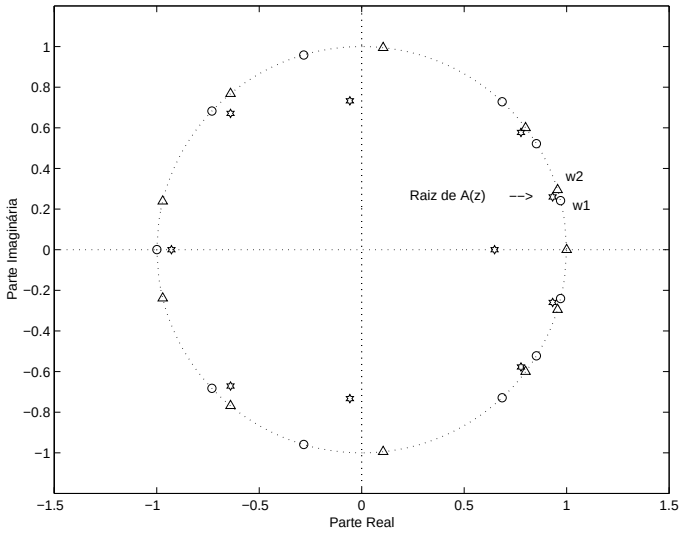


Figura 3.2: Gráfico com as raízes de $A(z)$, $P(z)$ e $Q(z)$ indicadas respectivamente por estrelas, círculos e triângulos, para um segmento sonoro de 25 ms.

respectivamente por estrelas, círculos e triângulos. Eles foram retirados de um segmento sonoro de 25 ms amostrado a 8 kHz através de uma análise LPC de ordem 10. Pode-se perceber a alternância das raízes de $P(z)$ e $Q(z)$, bem como os zeros em -1 e +1.

No modelo de tubos que representa o trato vocal humano, mostrado na Figura 3.1, a fonte de perdas é representada por um impedância casada na extremidade direita através de mais uma seção de área A_{p+1} infinita, correspondendo aos lábios. No trato vocal real, além deste último, existem perdas devido ao atrito viscoso entre o ar e as paredes do trato [4]. Se no local dos lábios for inserida mais uma seção com coeficiente de reflexão igual a +1, tem-se um tubo totalmente fechado. Caso a nova seção inserida tenha coeficiente de reflexão -1, então será um tubo totalmente aberto. Em ambos os casos a onda acústica (ar expelido pelos pulmões) é refletida totalmente sem perda de energia. As funções de transferências destes dois modelos considerados correspondem respectivamente a $P(z)$ e a $Q(z)$. O fato dos modelos não terem perdas causa o fato de todas as raízes estarem sobre a circunferência de raio unitário [4].

A propriedade na qual um conjunto ordenado de LSF, tal que $w_1 < w_2 <$

$w_3 < \ldots < w_p$, produz um filtro $A(z)$ cujas raízes encontram-se dentro do círculo de raio unitário, constitui um fator importante para a verificação da estabilidade do filtro $H(z)$. Além desta, outras vantagens da utilização das LSF em relação aos LAR são [5]:

- menor sensibilidade a erros de quantização;
- sensibilidade localizada, ou seja, erros de quantização na freqüência w_i somente afeta o espectro de $H(z)$ na vizinhança desta freqüência, diferente do que acontece para os LAR onde o erro em um dado l_i afeta toda a resposta em amplitude;
- são melhores para quantização vetorial.

A desvantagem do uso das LSF é o aumento na complexidade computacional. Tal problema, porém, está cada vez mais sendo atenuado através da obtenção de algoritmos rápidos para a determinação das LSF [29, 30].

Para p par, a proximidade de duas LSF indica a existência de uma raiz de $A(z)$ bem próxima ao referido agrupamento, e conseqüentemente à circunferência de raio unitário, indicando assim a existência de um formante [4]. Na Figura 3.2 pode-se perceber que os agrupamentos das freqüências w_1 com w_2, w_4 com w_5, e w_8 com w_9, indicam a presença de raízes de $A(z)$ bem próximas aos respectivos agrupamentos. Observando o espectro de potência do segmento em questão, mostrado na Figura 2.3, pode-se observar a existência dos formantes *F1*, *F2* e *F3* aproximadamente nas freqüências correspondentes.

3.4 Quantização escalar dos parâmetros de $H(z)$

Neste trabalho foram implementadas quantizações escalares dos parâmetros de $H(z)$ nas formas de LAR e LSF. A base de dados usada para o projeto dos quantizadores consistiu de aproximadamente 60 s de fala, obtidos de 23 frases, onde:

- 18 foram retiradas de [31] e pronunciadas por 2 locutores do sexo masculino e 1 do sexo feminino, onde cada um foi responsável por 6 frases;

- 5 da língua inglesa, sendo pronunciados por 3 locutores do sexo feminino e 2 do sexo masculino.

O sinais da língua portuguesa foram digitalizados com 16 bits por amostra a uma taxa de amostragem de 44,1 kHz em dois canais, e depois transformados para 8 kHz com um canal e os mesmos 16 bits por amostra. Já os sinais da língua inglesa foram digitalizados a 8 bits por amostra com taxa de amostragem de 8 kHz, usando a quantização logarítmica μ-*law*.

As análises LPC foram todas de ordem 10 feitas pelo método da autocorrelação, com janelas de Hamming de 25 ms centradas em cada segmento de fala correspondente. Foi utilizado o método de suavização espectral proposto por [32] para a expansão das bandas dos formantes. O objetivo deste procedimento é evitar que picos formânticos de bandas muito estreitas tornem o sinal de fala pouco natural. Melhor descrição do método de suavização espectral e a forma como foi utilizado são tratados no Capítulo 4.

3.4.1 Quantização dos LAR

A Figura 3.3 mostra histogramas com as distribuições dos valores dos coeficientes l_i, enquanto que a Tabela 3.1 mostra características como média, valor máximo, valor mínimo e variância, obtidos a partir da base de dados descrita acima.

A quantização dos LAR aqui realizada foi conforme a proposta em [28]. Segundo tal trabalho, uma forma de quantização bastante eficiente dos CR é realizar uma quantização uniforme dos LAR com passo de quantização ótimo δ dado por

$$\delta = \left\lfloor \frac{\prod_{i=1}^{p}(l_i^{sup} - l_i^{inf})}{2^M} \right\rfloor, \tag{3.10}$$

onde l_i^{sup} e l_i^{inf} são respectivamente os limites superior e inferior para o coeficiente l_i, e M é o número total de bits para a quantização de todos os coeficientes. O número ótimo de níveis de quantização para cada l_i é então dado por

$$N_i = \frac{l_i^{sup} - l_i^{inf}}{\delta}. \tag{3.11}$$

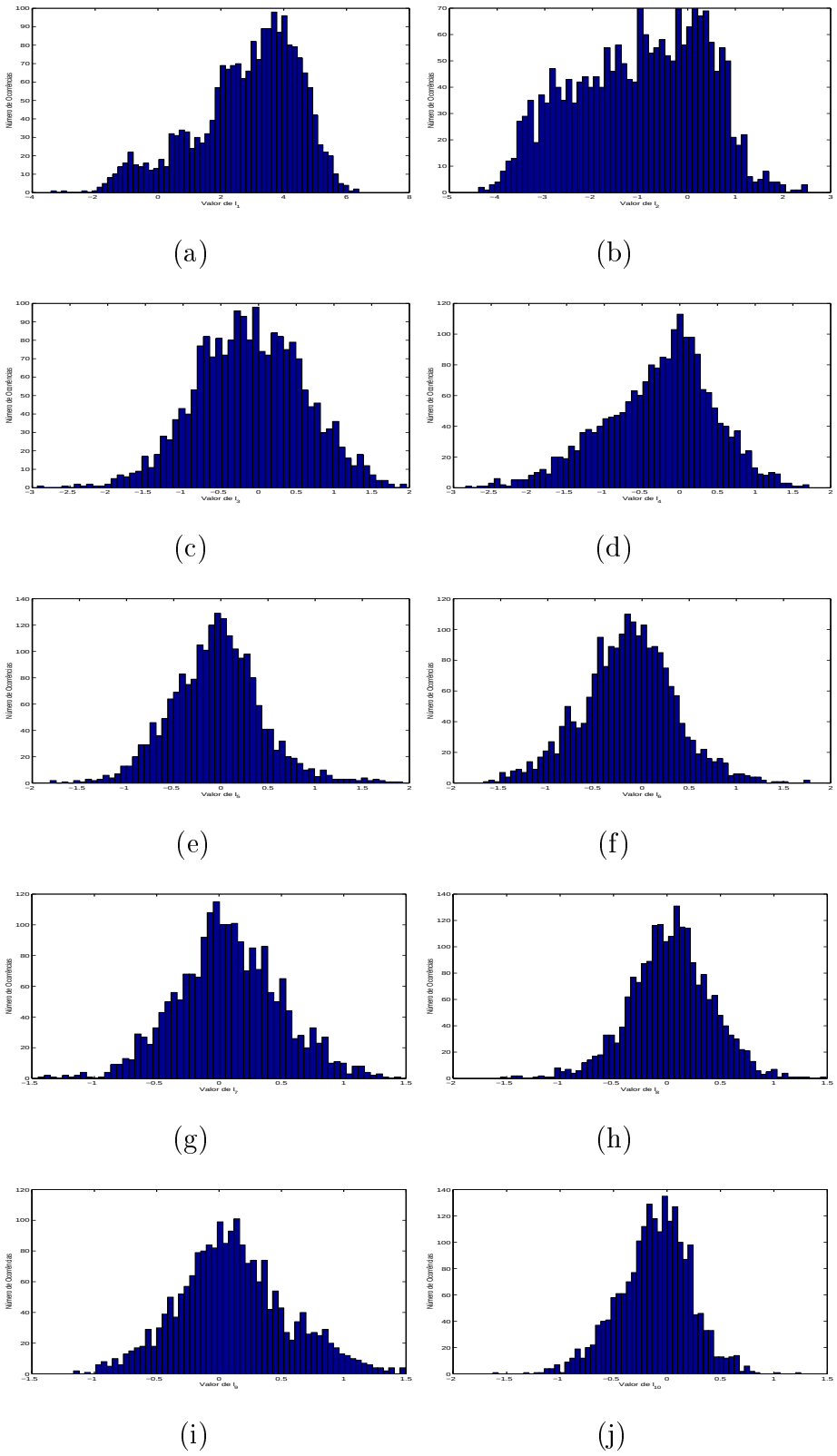


Figura 3.3: Histogramas dos valores assumidos por cada LAR para a base de dados considerada: de l_1 (a) a l_{10} (j).

Tabela 3.1: Características de cada LAR l_i para a base de dados considerada.

Coeficiente	Valor Máximo	Valor Mínimo	Média	Variância
l_1	6.3883	-3.3997	2.8388	2.7544
l_2	2.5089	-4.3889	-1.0193	1.8605
l_3	1.9610	-2.9320	-0.1131	0.5117
l_4	1.7104	-2.8377	-0.2724	0.5231
l_5	1.9275	-1.8089	-0.0453	0.2335
l_6	1.7778	-1.6819	-0.1460	0.2445
l_7	1.4556	-1.4503	0.0760	0.1747
l_8	1.4857	-1.5578	0.0324	0.1449
l_9	1.4943	-1.1661	0.1124	0.1943
l_{10}	1.2518	-1.6277	-0.1141	0.1122

O fato de quantizar todos os LAR com mesmo passo δ provém da premissa que todos eles são igualmente sensíveis a erros de quantização. Já a quantização uniforme é realizada porque é mostrado [28] que tais coeficientes são também igualmente sensíveis a erros de quantização ao longo de toda faixa dinâmica.

Os limites l_i^{sup} e l_i^{inf} para cada coeficiente l_i foram determinados através da inspeção dos histogramas mostrados na Figura 3.3. A alocação ótima, que está mostrada na Tabela 3.2, foi determinada por (3.11), para um número total de bits $M = 40$. O tamanho de passo ótimo obtido por (3.10) foi $\delta = 0,1864$. Daqui por diante esta quantização será referenciada como *QLAR-40*.

3.4.2 Quantização das LSF

A Figura 3.4 mostra os histogramas com as distribuições de cada uma das LSF w_i para a base de dados considerada, enquanto que a Tabela 3.3 mostra as suas características.

Três tipos de quantização foram feitas nas LSF: com 40, 36 e 32 bits por conjunto, que serão referenciadas daqui por diante como *QLSF-40*, *QLSF-36* e *QLSF-32*,

Tabela 3.2: Alocação de bits para a quantização uniforme dos LAR, com indicação dos limites superiores e inferiores inferidos pelos histogramas.

Coeficiente	No. de bits	l_i^{inf}	l_i^{sup}
l_1	6	-2	6
l_2	5	-4	2
l_3	4	-2	1,7
l_4	4	-2	1,3
l_5	4	-1,2	1,2
l_6	4	-1,5	1
l_7	3	-0,8	1
l_8	3	-1	1
l_9	4	-1	1,2
l_{10}	3	-1	0,7
Total de bits	40		

Tabela 3.3: Características de cada LSF w_i para a base de dados considerada.

Coeficiente	Valor Máximo	Valor Mínimo	Média	Variância
w_1	0.5123	0.0443	0.1912	0.0050
w_2	0.9706	0.1278	0.3474	0.0158
w_3	1.4171	0.2305	0.5710	0.0349
w_4	1.7090	0.4878	0.9107	0.0385
w_5	1.8835	0.6641	1.2603	0.0466
w_6	2.0587	0.9825	1.5743	0.0333
w_7	2.3629	1.2845	1.9249	0.0180
w_8	2.6296	1.5722	2.2042	0.0137
w_9	2.8296	2.0972	2.5243	0.0099
w_{10}	3.0056	2.3852	2.8013	0.0084

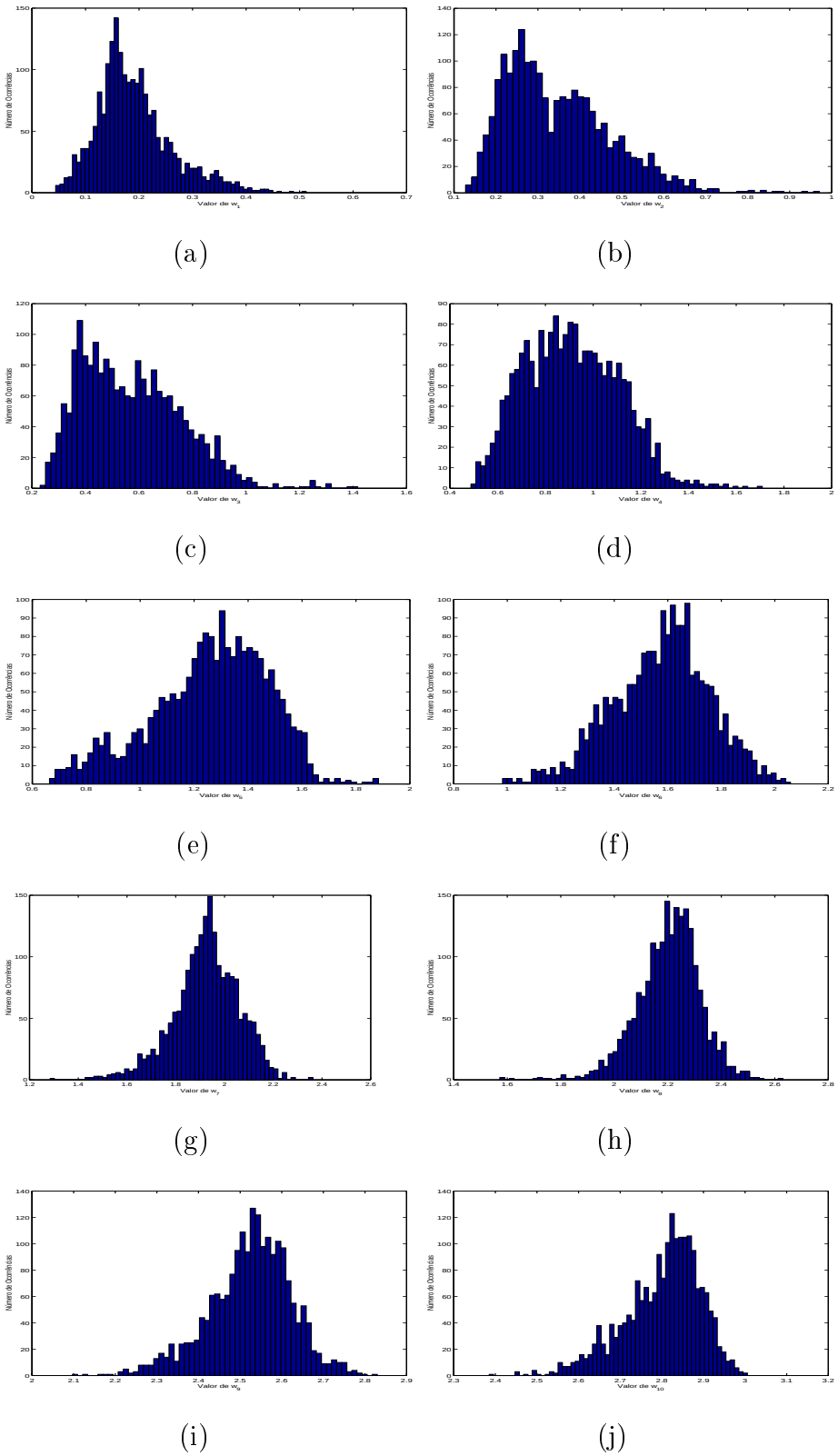


Figura 3.4: Histogramas dos valores assumidos por cada LSF para a base de dados considerada: de w_1 (a) a w_{10} (j).

Tabela 3.4: Alocações de bits para quantizações não-uniformes das LSF.

Quantizador	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	Total (bits)
QLSF-40	4	4	4	4	4	4	4	4	4	4	40
QLSF-36	4	4	4	4	4	4	3	3	3	3	36
QLSF-32	4	4	3	3	3	3	3	3	3	3	32

respectivamente. Em todos os casos foram determinados dicionários e partições para quantização não uniforme de cada coeficiente w_i através do treinamento da base de dados pelo algoritmo de Lloyd I [33].

Na quantização com 40 bits foram alocados 4 para cada w_i, partindo-se da premissa de que todas as LSF são igualmente sensíveis a erros de quantização [5]. Nas quantizações com 36 e 32 bits foram alocados 3 bits para os coeficientes que possuem as menores variâncias, com exceção de w_1 e w_2. Para eles foram alocados 4 bits em qualquer situação porque estes dois componentes são responsáveis pela reprodução do primeiro formante em sons sonoros, cuja boa representação é muito importante para a qualidade da fala reconstruída [4]. A Tabela 3.4 mostra as alocações de bits para os 3 quantizadores.

3.4.3 Quantização diferencial das LSF

O grande problema de quantizar individualmente cada w_i é que muitas vezes a versão quantizada perde a propriedade de ordenação das LSF, que é necessária para produzir o filtro $H(z) = 1/A(z)$ estável. Isso acontece principalmente em sons sonoros onde o primeiro formante possui banda muito estreita, ou seja, quando as freqüências w_1 e w_2 encontram-se muito próximas. Uma forma de evitar isso é quantizar as diferenças entre LSF consecutivas. Além disso, segundo é afirmado em [34] e comprovado experimentalmente, as diferenças entre LSF possuem menores faixas dinâmicas, o que pode proporcionar uma quantização mais eficiente. Isto é verificado pelos histogramas dos Δw_i, mostrados na Figura 3.5, e pelas suas características

Tabela 3.5: Características das diferenças entre LSF, Δw_i, para a base de dados considerada.

Coeficiente	Valor Máximo	Valor Mínimo	Média	Variância
Δw_1	0.5123	0.0443	0.1912	0.0050
Δw_2	0.8297	0.0219	0.1562	0.0105
Δw_3	0.6991	0.0326	0.2237	0.0132
Δw_4	0.9024	0.0416	0.3397	0.0205
Δw_5	1.0147	0.0299	0.3495	0.0313
Δw_6	1.0865	0.0269	0.3141	0.0266
Δw_7	1.1121	0.0571	0.3506	0.0224
Δw_8	0.7766	0.0314	0.2793	0.0125
Δw_9	0.8428	0.0694	0.3200	0.0124
Δw_{10}	0.6391	0.0507	0.2771	0.0099

mostradas na Tabela 3.5, onde

$$\begin{aligned} \Delta w_1 &= w_1 \\ \Delta w_i &= w_i - w_{i-1}, \quad \text{para} \quad 2 \leq i \leq p. \end{aligned} \tag{3.12}$$

Tal como no caso não diferencial, foram desenvolvidos quantizadores com 40, 36 e 32 bits, obtidos pelo treinamento da base de dados pelo algoritmo de Lloyd I, que serão referenciados como *QDLSF-40*, *QDLSF-36* e *QDLSF-32*, respectivamente. Para evitar a propagação de erros, são quantizadas as diferenças entre uma dada w_i e a freqüência anterior já quantizada $\hat{w}_{i-1}$. Em todos os casos foram atribuídos 4 bits para w_1 e $w_2 - \hat{w}_1$ pelo fato destes serem responsáveis pela reprodução do primeiro formante em sons sonoros. A Tabela 3.6 mostra as alocações de bits para o 3 quantizadores diferenciais.

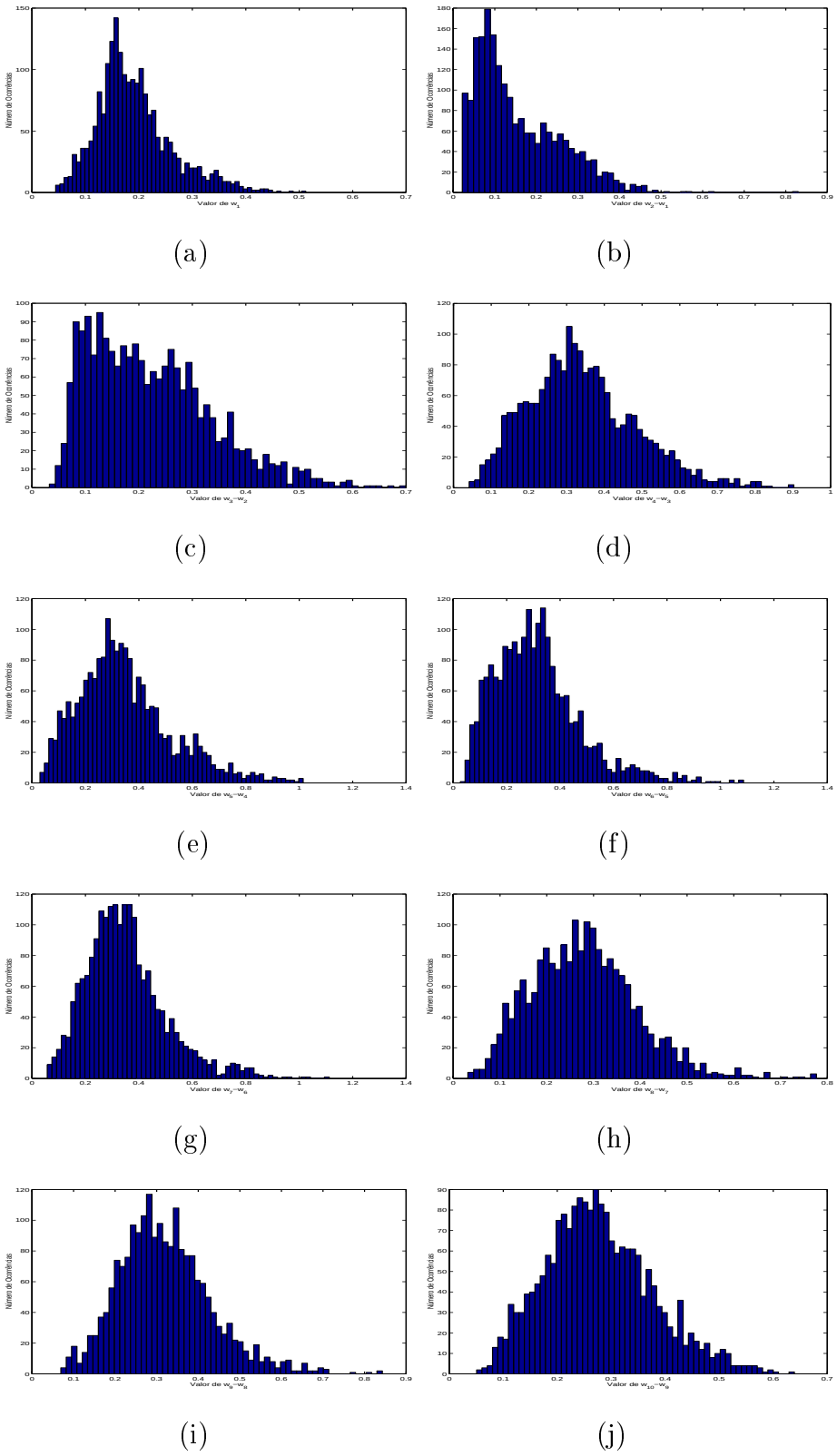


Figura 3.5: Histogramas dos valores assumidos pelas diferenças entre LSF consecutivas para a base de dados considerada: de Δw_1 (a) a Δw_{10} (j).

Tabela 3.6: Alocações de bits para quantizações diferenciais não-uniformes das LSF.

Quantizador	Δw_1	Δw_2	Δw_3	Δw_4	Δw_5	Δw_6	Δw_7	Δw_8	Δw_9	Δw_{10}	Total
QDLSF-40	4	4	4	4	4	4	4	4	4	4	40
QDLSF-36	4	4	4	4	4	4	3	3	3	3	36
QDLSF-32	4	4	3	3	3	3	3	3	3	3	32

3.5 Avaliação dos quantizadores

Os quantizadores descritos neste capítulo foram avaliados por uma base de dados diferente da que foi usada para projetá-los, tanto no que diz respeito às frases usadas quanto aos locutores. A base usada para a avaliação consistiu de aproximadamente 50 s de fala, gerados por um conjunto de 10 frases foneticamente balanceadas, retiradas de [31], e pronunciadas por 10 locutores, sendo 5 do sexo masculino e 5 do sexo feminino. Cada locutor do sexo masculino e do sexo feminino pronunciou um par de frases, gerando assim 20 sinais. Os sinais foram gravados com 16 bits por amostra a uma taxa de amostragem de 8 kHz, com filtragem passa-altas digital para remoção de ruídos de baixa freqüência.

3.5.1 Medida de distorção

A medida utilizada para avaliar o desempenho dos quantizadores foi a distância espectral [35]. Seja o modelo original dado por $H(z) = 1/A(z)$ e a versão quantizada dada por $\hat{H}(z) = 1/\hat{A}(z)$, a distância espectral entre esses dois modelos, em dB, é dada por

$$DE(n) = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} 10\log\left[\frac{H_n(e^{jw})}{\hat{H}_n(e^{jw})}\right]^2 dw} \qquad (dB), \tag{3.13}$$

onde

$$H_n(e^{jw}) = \left|\frac{1}{A_n(e^{jw})}\right| \quad \text{e} \quad \hat{H}_n(e^{jw}) = \left|\frac{1}{\hat{A}_n(e^{jw})}\right|, \tag{3.14}$$

sendo que $1/A_n(e^{jw})$ e $1/\hat{A}_n(e^{jw})$ são as respostas em freqüência dos modelos $1/A(z)$ e $1/\hat{A}(z)$ respectivamente para o segmento n. A distorção média $\bar{DE}$ é definida por

$$\bar{DE} = \frac{1}{M}\sum_{n=0}^{M-1} DE(n), \tag{3.15}$$

onde M é o número de medidas realizadas e $n = 0, \ldots, M-1$, indicam os segmentos aos quais estão associadas estas medidas.

3.5.2 Formas de avaliação

A avaliação dos quantizadores foi feita de duas formas, tendo em vista o sistema CELP: considerando e não considerando os efeitos da interpolação dos CPL entre blocos [10, 14]. Nos codificadores CELP os parâmetros de $H(z)$ são determinados a cada bloco. Enquanto isso o sinal de excitação é determinado a intervalos de tempo menores, que são os sub-blocos. Neste caso, os CPL de um determinado bloco valem para todos os sub-blocos que nele estão contidos. Para permitir que haja transição suave, os conjuntos de CPL entre dois blocos consecutivos (o corrente e o anterior) são interpolados para gerar os CPL efetivamente utilizados em cada sub-bloco.

A seguir serão descritos os procedimentos para avaliação dos codificadores considerando ou não os efeitos da interpolação e os respectivos resultados obtidos.

3.5.2.1 Avaliação sem o efeito da interpolação

A Figura 3.6 mostra como é determinada a distorção espectral DE_q usada para avaliar a quantização dos parâmetros de $H(z)$ sem levar em conta os efeitos da interpolação entre blocos. É determinada a diferença entre cada modelo $A_n(z)$ e sua versão quantizada $\hat{A}_n(z)$. F_t representa a freqüência na qual os parâmetros são transmitidos, que para o caso de um sistema CELP é a cada bloco. A Figura 3.7 mostra os detalhes de como são obtidos os modelos $A_n(z)$, tais como o posicionamento da janela de Hamming de 25 ms, que fica centrada no último sub-bloco de cada bloco.

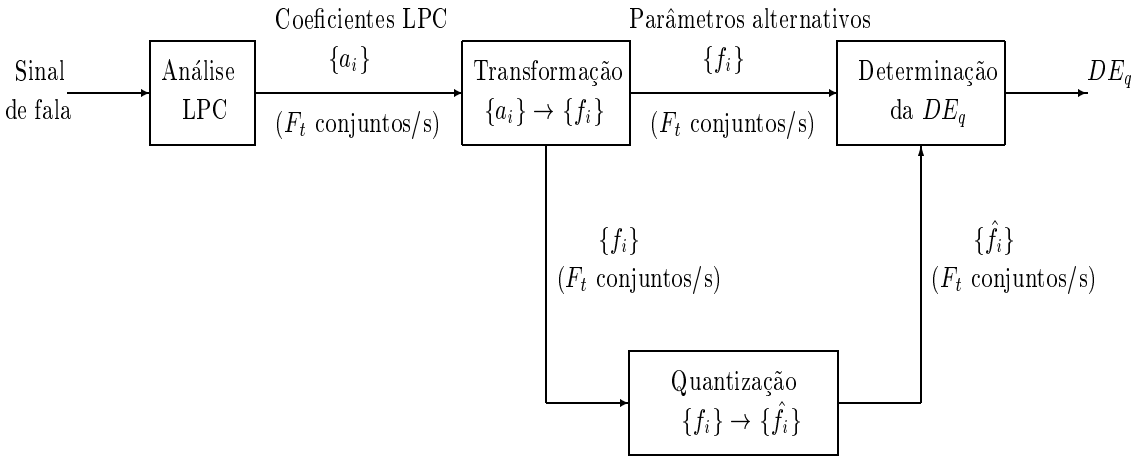


Figura 3.6: Procedimento de determinação da DE_q para a avaliação da quantização dos parâmetros de $H(z)$ sem levar em conta a interpolação entre blocos.

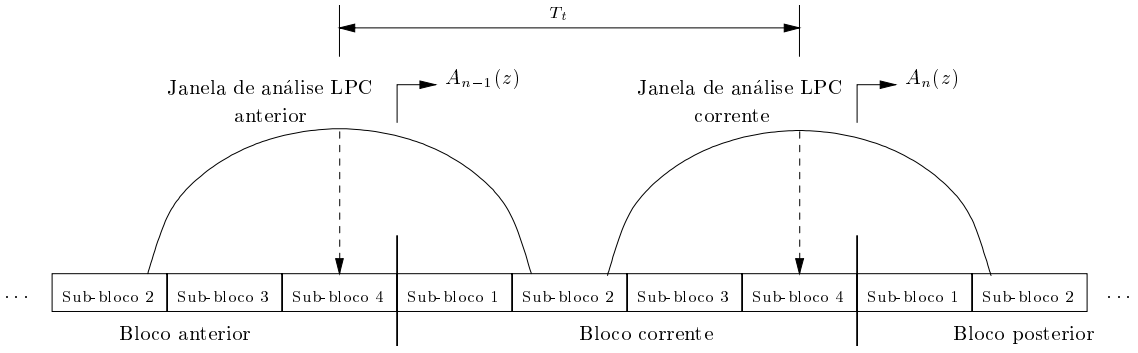


Figura 3.7: Detalhes da forma como são determinados os coeficientes LPC para cada bloco, para a avaliação dos quantizadores sem levar em conta a interpolação entre blocos.

O termo *quantização transparente* dos coeficientes de predição linear (CPL) é usado para indicar que uma dada quantização de um conjunto de CPL transmitidos não introduz na fala reconstruída qualquer distorção audível adicional, ou seja, as duas versões de fala reconstruídas - uma obtida com os CPL não quantizados e a outra com os CPL quantizados - são indistinguíveis auditivamente. Segundo [5], as condições suficientes para que uma quantização de CPL seja transparente é que a DE_q atenda às seguintes restrições:

- a distorção espectral média $\bar{DE}_q$ deve resultar aproximadamente igual ou menor que 1 dB;
- dentre todos os blocos n não deve ser observado nenhum valor $DE_q(n)$ acima de 4 dB;
- o número de conjuntos de CPL apresentando valores $DE_q(n)$ entre 2 e 4 dB deve ser menor que 2%.

A Tabela 3.7 mostra a média de DE_q e os percentuais de valores entre 2 e 4 dB e acima de 4 dB, para todos os quantizadores projetados. Neste caso foram usados blocos de $T_t = 20\ ms$ com sub-blocos de 5 ms ($F_t = 50\ Hz$). A quinta coluna indica o número de vezes que a quantização dos parâmetros gerou um filtro $\hat{H}(z)$ instável. As Tabelas 3.8 e 3.9 mostram os casos para blocos de 30 ms com sub-blocos de 6 ms, e blocos de 30 ms com sub-blocos de 7,5 ms, respectivamente.

Dos resultados mostrados nas tabelas pode-se concluir que:

- a condição de quantização transparente mostrada anteriormente é alcançada somente pelos quantizadores *QLSF-40*, *QDLSF-40* e *QDLSF-36* para todos os casos;
- a quantização diferencial das LSF foi melhor que a sua quantização individual e quantização dos LAR em todos os casos, tanto no que diz respeito à qualidade quanto ao número de vezes que o filtro $H(z)$ torna-se instável após a quantização.

Tabela 3.7: Desempenho dos quantizadores para blocos de 20 ms com sub-blocos de 5 ms, sem levar em conta os efeitos da interpolação entre blocos.

Quantizador	$\bar{DE}_q$ (dB)	$\%DE_q \in [2;4]$	$\%DE_q > 4$	No. inst.	No. bits
$QLAR-40$	0,86	5,01	0,89	-	40
$QLSF-40$	0,74	0,32	0,00	2	40
$QDLSF-40$	0,62	0,24	0,00	-	40
$QLSF-36$	0,97	2,30	0,00	15	36
$QDLSF-36$	0,85	1,62	0,00	-	36
$QLSF-32$	1,31	8,25	0,20	29	32
$QDLSF-32$	1,09	4,49	0,00	-	32

Tabela 3.8: Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 6 ms, sem levar em conta os efeitos da interpolação entre blocos.

Quantizador	$\bar{DE}_q$ (dB)	$\%DE_q \in [2;4]$	$\%DE_q > 4$	No. inst.	No. bits
$QLAR-40$	0,85	5,40	0,79	-	40
$QLSF-40$	0,74	0,36	0,00	-	40
$QDLSF-40$	0,63	0,36	0,00	-	40
$QLSF-36$	0,96	2,24	0,00	11	36
$QDLSF-36$	0,85	1,52	0,00	-	36
$QLSF-32$	1,31	8,13	0,18	20	32
$QDLSF-32$	1,10	5,4	0,06	-	32

Tabela 3.9: Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 7,5 ms, sem levar em conta os efeitos da interpolação entre blocos.

Quantizador	$\bar{DE}_q$ (dB)	$\%DE_q \in [2;4]$	$\%DE_q > 4$	No. inst.	No. bits
$QLAR-40$	0,86	5,03	0,79	-	40
$QLSF-40$	0,74	0,30	0,00	-	40
$QDLSF-40$	0,63	0,36	0,00	-	40
$QLSF-36$	0,96	2,00	0,12	9	36
$QDLSF-36$	0,85	1,82	0,00	-	36
$QLSF-32$	1,31	7,82	0,18	19	32
$QDLSF-32$	1,10	5,09	0,06	-	32

3.5.2.2 Avaliação com o efeito da interpolação

Para uma avaliação mais precisa dos quantizadores, é necessário que sejam levadas em conta os efeitos causados pela interpolação entre sub-blocos dos parâmetros de $H(z)$. A Figura 3.8 mostra como é determinada a distorção espectral DE_c usada para avaliar a quantização dos parâmetros de $H(z)$ levando-se em conta os efeitos da interpolação entre blocos. Os modelos são determinados a uma freqüência F_i correspondente à duração dos sub-blocos, sendo depois decimados para a freqüência F_t correspondentes à duração dos blocos. Nesta última freqüência eles são quantizados e depois interpolados de volta para a freqüência F_i, a fim de ser determinada a distorção espectral DE_c. Portanto, a cada sub-bloco são determinados os modelos, mas aqueles que são usados para a interpolação correspondem aos últimos sub-blocos de cada bloco, tal como em alguns sistemas CELP [5, 27]. A Figura 3.9 ilustra com detalhes o posicionamento das janelas de Hamming de 25 ms para a determinação dos modelos para cada sub-bloco. Este método de avaliação corresponde ao mesmo utilizado em [36]. Quando a quantização resulta em um filtro $H(z)$ instável são utilizados os CPL do sub-bloco precedente, tal como no sistema CELP proposto em [14].

A Tabela 3.10 mostra o desempenho de codificação dos quantizadores para

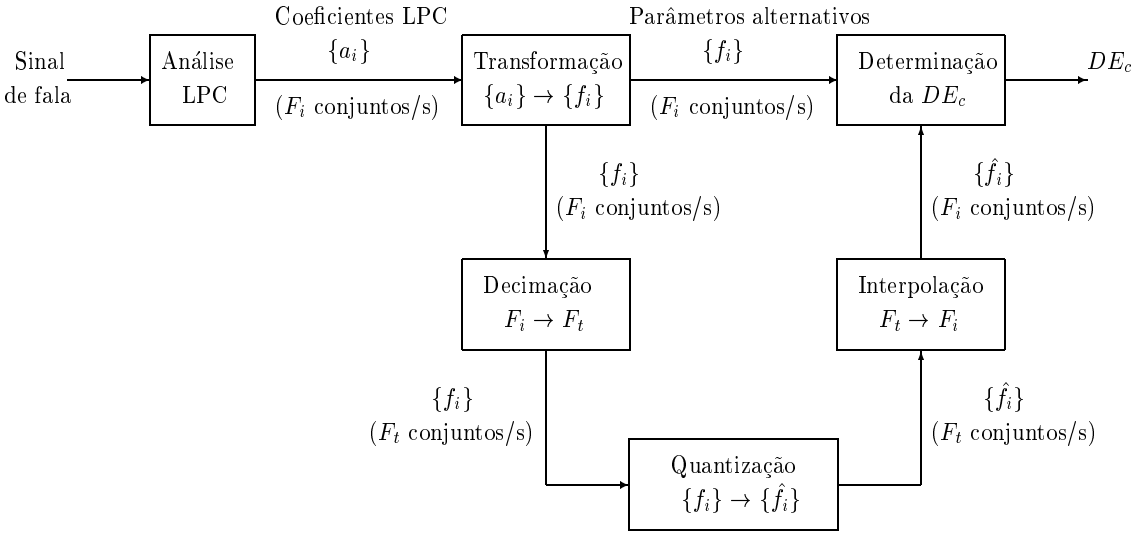


Figura 3.8: Procedimento de determinação da DE_c usada para a avaliação da quantização dos parâmetros de $H(z)$, levando-se em conta os efeitos da interpolação entre blocos.

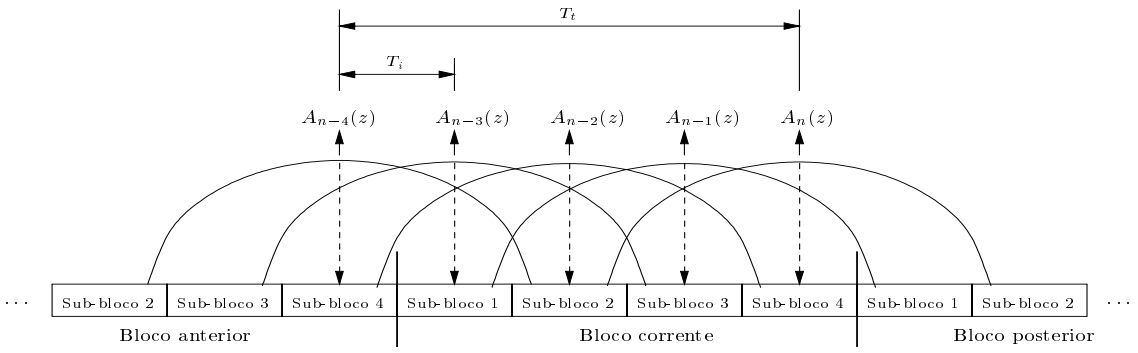


Figura 3.9: Detalhes da forma como são determinados os coeficientes LPC para cada sub-bloco, para a avaliação dos quantizadores levando-se em conta a interpolação entre blocos.

Tabela 3.10: Desempenho dos quantizadores para blocos de 20 ms com sub-blocos de 5 ms, levando-se em conta os efeitos da interpolação entre blocos.

Quantizador	$\bar{DE}_c$ (dB)	$\%DE_c \in [2;4]$	$\%DE_c > 4$	No. inst.	Taxa (kbps)
$QLAR-40$	1,59	20,93	3,96	-	2,000
$QLSF-40$	1,31	13,51	0,58	3	2,000
$QDLSF-40$	1,26	12,89	0,67	-	2,000
$QLSF-36$	1,44	15,79	0,79	28	1,800
$QDLSF-36$	1,38	14,77	0,68	-	1,800
$QLSF-32$	1,65	22,42	1,09	44	1,600
$QDLSF-32$	1,52	18,73	0,78	-	1,600

blocos de 20 ms com sub-blocos de 5 ms ($T_t = 20\ ms$ e $T_i = 5\ ms$), a Tabela 3.11 mostra os resultados para blocos de 30 ms com sub-blocos de 6 ms, enquanto que a Tabela 3.12 mostra para blocos de 30 ms com sub-blocos de 7,5 ms. A quinta coluna indica a taxa de bits consumida pelos CPL quando é utilizada a respectiva quantização.

Ainda não existe uma forma de avaliar a condição de quantização transparente quando considerados os efeitos da interpolação entre sub-blocos [5]. A condição mostrada na Sub-seção 3.5.2 diz respeito somente às diferenças entre os conjuntos de coeficientes e suas versões quantizadas. Para o caso de avaliação da quantização levando-se em conta os efeitos de interpolação, os resultados mostram que a quantização diferencial das LSF é superior em todos os casos. Nota-se para todos os quantizadores um considerável aumento nos valores de distorção espectral de pico, o que pode indicar que existem formas mais adequadas de realizar a interpolação (que para o presente caso foi linear).

Tabela 3.11: Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 6 ms, levando-se em conta os efeitos da interpolação entre blocos.

Quantizador	$\bar{DE}_c$ (dB)	$\%DE_c \in [2;4]$	$\%DE_c > 4$	No. inst.	Taxa (kbps)
$QLAR-40$	1,95	32,70	6,80	-	1,333
$QLSF-40$	1,64	25,33	2,62	-	1,333
$QDLSF-40$	1,60	24,81	2,53	-	1,333
$QLSF-36$	1,75	27,49	2,91	12	1,200
$QDLSF-36$	1,69	26,81	2,59	-	1,200
$QLSF-32$	1,94	34,44	3,42	35	1,067
$QDLSF-32$	1,82	31,06	2,70	-	1,067

Tabela 3.12: Desempenho dos quantizadores para blocos de 30 ms com sub-blocos de 7,5 ms, levando-se em conta os efeitos da interpolação entre blocos.

Quantizador	$\bar{DE}_c$ (dB)	$\%DE_c \in [2;4]$	$\%DE_c > 4$	No. inst.	Taxa (kbps)
$QLAR-40$	1,93	32,22	6,63	-	1,333
$QLSF-40$	1,62	25,29	2,78	-	1,333
$QDLSF-40$	1,58	25,00	2,73	-	1,333
$QLSF-36$	1,74	27,66	3,25	12	1,333
$QDLSF-36$	1,68	27,43	2,76	-	1,200
$QLSF-32$	1,94	34,94	3,46	27	1,067
$QDLSF-32$	1,81	30,90	2,91	-	1,067

3.6 Conclusão

Neste capítulo foram desenvolvidas quantizações escalares dos parâmetros de $H(z)$, representados por LAR e LSF. Em todas as situações verificou-se que a quantização diferencial das LSF foi mais eficiente. As vantagens deste método em relação à quantização individual de cada LSF e dos LAR são:

- menores distorções introduzidas nos coeficientes originais, para todos os casos de tamanhos de blocos e sub-blocos, quando são utilizadas iguais quantidades de bits para cada método;
- garantia de estabilidade da versão quantizada de $H(z)$, fato este problemático quando da quantização individual de cada LSF.

Além dos fatores acima, a quantização diferencial equivaleu de forma aproximada, em termos de distorção, à quantização individual das LSF com 4 bits a mais por conjunto, ou seja, os quantizadores *QDLSF-36* e *QDLSF-32* foram aproximadamente iguais aos quantizadores *QLSF-40* e *QLSF-36*, respectivamente.

Atualmente, a forma mais eficiente de quantização dos parâmetros de $H(z)$ é a quantização vetorial (QV) [36, 37], que chega a usar de 18 a 30 bits por conjunto, obtendo qualidade melhor que a quantização escalar. A grande vantagem da aplicação da QV nos LSF é que se pode utilizar uma medida de distorção que leve em conta a importância destes coeficientes na reprodução do sinal de fala. O grande problema, no entanto, é a complexidade do seu projeto, demandando um enorme tempo para o treinamento do dicionário, fato este que impossibilitou a sua implementação neste trabalho. A descrição do projeto de alguns QV para LSF pode ser vista em [5, 27].

Capítulo 4

Implementação de um sistema CELP

4.1 Introdução

Desde o início da década de 90 têm sido publicados muitos algoritmos CELP em jornais especializados ou conferências internacionais [10, 12, 13, 14, 21, 26]. Isto tem sido motivado pela crescente demanda, representada pelo aumento de aplicações com fala na Internet e propostas para padronizações de sistemas de telefonia móvel [38]. As diferenças entre os vários sistemas CELP vão desde a forma de determinação dos coeficientes de predição linear ao tipo de dicionário que é utilizado, bem como o método de seleção da melhor excitação. Este capítulo tem o objetivo de descrever detalhadamente o desenvolvimento e implementação de um sistema CELP que opera a uma taxa de 4,67 kbps.

Na Seção 4.2 são mostradas algumas considerações iniciais que dizem respeito à técnica CELP, particularmente à estrutura do sistema implementado; na Seção 4.3 são mostradas as características que dizem respeito ao filtro de síntese; na Seção 4.4 é descrito o procedimento de busca da melhor excitação; na Seção 4.5 são mostradas as características do dicionário adaptativo; na Seção 4.6 são tratadas as características do dicionário fixo; na Seção 4.7 é tratada a quantização dos ganhos e alocação de bits para os parâmetros da excitação; na Seção 4.8 é feito um resumo

das características do sistema implementado; a Seção 4.9 trata da sua avaliação; e finalmente na Seção 4.10 estão as conclusões do capítulo.

4.2 Considerações iniciais

A Figura 2.11 mostra o diagrama de blocos do sistema de codificação CELP cuja estrutura é tida como padrão. A fala reconstruída é obtida pela passagem do sinal de excitação $x(n)$ pelo filtro de síntese $H(z) = 1/A(z)$. O sinal de excitação $x(n)$ é formado pela soma de uma seqüência oriunda do dicionário adaptativo $x_{aD}(n)$ e outra do dicionário fixo $x_{fI}(n)$, escaladas pelos respectivos ganhos G_a e G_f. O dicionário adaptativo [11] é responsável pela componente relativa à periodicidade do sinal de fala, e substitui de maneira mais eficiente o filtro de *pitch* presente nos primeiros sistemas CELP [8]. O dicionário fixo contém um conjunto de seqüências que representam sinais residuais [15].

Além dos bits para sinalização e correção de erros, a informação enviada ao decodificador consiste nos coeficientes do filtro $H(z)$, os índices D e I, e os ganhos G_a e G_f dos dicionários adaptativo e fixo, respectivamente. Estes quatro últimos parâmetros são obtidos através de um procedimento de análise-por-síntese em que todas as combinações de seqüências candidatas são testadas, ou pelo menos uma parte delas, escolhendo-se aquela que minimiza a energia do erro perceptual $e_w(n)$.

A estrutura do CELP implementado corresponde àquela mostrada na Figura 2.11. Ele processa o sinal da fala em blocos de 30 ms, que é o intervalo de tempo entre duas análises LPC consecutivas. Cada bloco é dividido em 4 sub-blocos de 7,5 ms, onde para cada um deles é obtida a melhor seqüência de excitação $x(n)$. Estas durações foram escolhidas para que fosse possível o sistema operar a uma taxa em torno de 4 kbps, seguindo uma tendência dos codificadores CELP que são publicados atualmente tendo em vista a terceira geração de telefonia móvel. Nas seções seguintes são dados detalhes a respeito do sistema implementado.

4.3 A análise LPC

Em qualquer sistema de codificação que utilize o filtro de correlação de curto termo $H(z) = 1/A(z)$, a forma como é realizada a análise LPC tem uma grande influência sobre a qualidade do sinal reconstruído.

4.3.1 Características gerais

Para o sistema implementado é utilizado o método da autocorrelação de ordem 10 ($p = 10$), com uma janela de Hamming de 25 ms (200 amostras para uma taxa de amostragem de 8 kHz), centrada no último sub-bloco de cada bloco, conforme mostrado na Figura 3.7. Este artifício também é utilizado no VSELP [27] e facilita a formulação de uma maneira de interpolar os CPL para determinar os que são efetivamente utilizados em cada sub-bloco.

O tamanho da janela determina a resolução na freqüência para a determinação dos CPL; é utilizada a duração de 25 ms porque este tamanho permite uma boa resolução, e conseqüentemente uma melhor representação dos efeitos do trato vocal [2]. Além disso o tamanho entre 20 e 30 ms é escolhido pela maioria dos sistemas CELP [10, 21, 14]. O método para determinação dos CPL é o da autocorrelação porque garante que os CPL obtidos conduzem a um filtro $1/A(z)$ estável [4].

4.3.2 Suavização espectral

Durante a análise LPC é usada a técnica de suavização espectral [32]. Este procedimento evita que haja super resolução em freqüência causada pela janela de Hamming quando esta cobre muitos períodos de *pitch*. Tal fato produz picos espectrais bastante estreitos no espectro de curto termo de segmentos sonoros, produzindo som não natural [32]. O que o método de suavização espectral faz é alargar a banda dos formantes para evitar este efeito indesejado na fala reconstruída.

O método consiste em modificar a seqüência de autocorrelação do segmento de sinal de fala da forma

$$r_s(i) = r(i)w_s(i), \qquad \text{para} \qquad 0 \leq i \leq p, \tag{4.1}$$

Tabela 4.1: Amplitudes das amostras da janela binomial usada para a suavização espectral.

Coeficiente	Amplitude	Coeficiente	Amplitude
$w_s(0)$	1,000000000	$w_s(6)$	0,977758059
$w_s(1)$	0,999375390	$w_s(7)$	0,969848379
$w_s(2)$	0,997503900	$w_s(8)$	0,960801286
$w_s(3)$	0,994392535	$w_s(9)$	0,950649874
$w_s(4)$	0,990052911	$w_s(10)$	0,939431025
$w_s(5)$	0,984501218		

onde $r(n) = \{r(0), \ldots, r(p)\}$ é a seqüência de autocorrelação, $w_s(n) = \{w_s(0), \ldots, w_s(p)\}$ uma janela no tempo, p a ordem do modelo LPC, e $r_s(n) = \{r_s(0), \ldots, r_s(p)\}$ é a seqüência de autocorrelação *suavizada* que será usada para determinar os CPL através do algoritmo de Levinson-Durbin [2, 4].

No sistema implementado foi utilizada uma janela binomial, onde cada coeficiente é dado por [32]

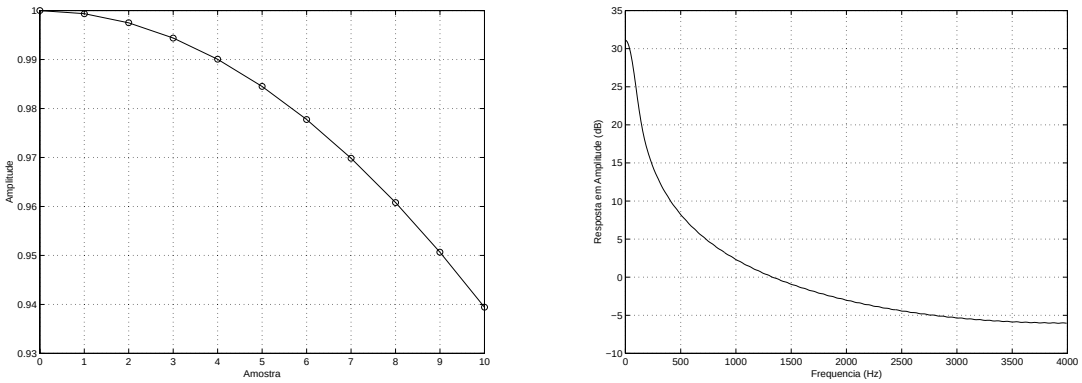
$$w_s(i) = \frac{\binom{2K}{K-i}}{\binom{2K}{K}}, \quad \text{para} \quad 0 \leq i \leq K-1, \tag{4.2}$$

onde K é o número de pontos da janela binomial e $\binom{n}{r} = \frac{n!}{(n-r)!r!}$. Conforme se pode perceber, para a suavização espectral somente os $p+1$ primeiros coeficientes de $w_s(n)$, que estão listados na Tabela 4.1 e mostrados na Figura 4.1(a), são utilizados. Esta janela foi determinada de maneira que a largura do lobo de sua resposta em amplitude fosse 130 Hz, como mostrado na Figura 4.1(b).

O método de suavização espectral aqui empregado substitui o bastante utilizado método de expansão dos coeficientes a_i, obtido por [18]

$$\alpha_i = a_i \rho^{i-1}, \quad \text{para} \quad 1 \leq i \leq p, \tag{4.3}$$

onde α_i são coeficientes expandidos, a_i os coeficientes originais, p a ordem do modelo, e ρ é o fator de expansão, que geralmente fica em torno de 0,988 a 0,996 [32]. Este



(a) (b)

Figura 4.1: Janela binomial usada para suavização espectral: (a) resposta ao impulso; (b) resposta em amplitude.

método tem também o objetivo de alargar as bandas dos formantes, sendo utilizado em alguns sistemas CELP [10, 14].

Foram feitos testes experimentais com os métodos de suavização espectral e expansão dos CPL para determinar qual o melhor, e que posteriormente seria utilizado no sistema CELP proposto. O testes consistiram em executar o sistema CELP descrito na Tabela 4.7 sem quantização dos ganhos usando suavização espectral, expansão dos coeficientes e nenhum dos métodos. Foram escolhidos 4 sinais de fala para executar os testes: dois pronunciados por locutores do sexo masculino (M1 e M2), e dois por locutores do sexo feminino (F1 e F2). Para a avaliação foi escolhida uma medida de qualidade objetiva, que correspondeu à razão sinal-ruído segmentada perceptual [5], que é definida em 4.9.1.1. O fator de expansão empregado foi $\rho = 0,994$, que corresponde ao mesmo utilizado em [14]. A Tabela 4.2 mostra os resultados dos testes. Apesar de apresentar em algumas situações melhores valores em termos de razão sinal-ruído, os sinais decodificados com nenhum dos métodos tiveram qualidade subjetiva ligeiramente inferior, principalmente para os sinais originados pelos locutores femininos. Comparando os dois métodos de alargamento das bandas dos formantes, pode-se perceber que a técnica de suavização espectral não obteve melhor desempenho somente em uma situação. Em avaliações subjetivas informais não foram percebidas diferenças significativas entre os sinais reconstruídos

Tabela 4.2: Razões sinal-ruído segmentada perceptual em dB, para 4 sinais processados com suavização espectral, expansão dos CPL e nenhum dos métodos.

	Locutor			
Método	M1	M2	F1	F2
Suavização Espectral	15,85	16,25	17,28	15,26
Expansão dos Coeficientes	15,68	16,18	17,08	15,29
Com nenhum método	15,86	16,36	17,21	15,22

das duas formas. Portanto, foi escolhida a suavização espectral para o sistema CELP pela combinação dos melhores resultados objetivos e subjetivos.

4.3.3 Quantização e interpolação

Os 10 CPL obtidos a cada bloco são transformados em freqüências do espectro de linha [29] $\{w_1, \ldots, w_p\}$. Tais parâmetros são quantizados de forma escalar diferencial com um total de 32 por conjunto, que corresponde ao quantizador *QDLSF-32* tratado no Capítulo 3. A alocação de bits está mostrada na Tabela 3.6. Vale lembrar que

$$\Delta w_i = w_i - \hat{w}_{i-1}, \qquad \text{para} \qquad 1 \leq i \leq p \tag{4.4}$$

onde $\hat{w}_i$ representa a versão quantizada de w_i, e $\hat{w}_0 = 0$.

Para produzir uma transição suave entre conjuntos adjacentes de CPL, eles são interpolados no formato de LSF, já devidamente quantizados. Para a interpolação, os sub-blocos de um determinado bloco são numerados de 1 a 4, conforme mostra a Figura 3.7. O i-ésimo coeficiente do n-ésimo sub-bloco é obtido por

$$w_i^n = (1 - q_n) w_i^a + q_n w_i^c, \tag{4.5}$$

onde w_i^c refere-se ao conjunto de coeficientes do bloco corrente e w_i^a ao do bloco anterior. O vetor de pesos usado foi $q_n = \{0,25; 0,5; 0,75; 1\}$ porque deve-se entender que são determinados os CPL para cada último sub-bloco, de acordo com

o posicionamento da janela de Hamming. Dessa forma os coeficientes dos demais sub-blocos são obtidos por interpolação linear.

4.4 Obtenção da melhor excitação

Os sistemas CELP determinam a melhor excitação para cada sub-bloco através de um procedimento conhecido como *análise-por-síntese* [7]. Este termo é empregado porque a análise da melhor excitação para o filtro de síntese é feita através da síntese de várias versões do sub-bloco do sinal de fala. A Figura 4.2 mostra o diagrama de blocos detalhado, na forma prática, do codificador CELP implementado. Nesta figura os procedimentos referentes à codificação não mostrados no diagrama da Figura 2.11 estão indicados pelas linhas pontilhadas.

Para a implementação prática do sistema CELP, duas mudanças principais são realizadas:

- o filtro perceptual $W(z)$ é deslocado para os dois ramos à esquerda do somador, transformando o filtro $H(z)$ em $H_w(z)$;
- a resposta do filtro $H_w(z)$ no intervalo $0 \leq n < N$, onde N é o número de amostras de cada sub-bloco, é decomposta em duas partes: a resposta com estado inicial zero $\hat{t}(n)$ e a resposta à entrada zero $\hat{s}_0(n)$.

A seqüência $\hat{s}_0(n)$ não depende da seqüência de excitação candidata $x(n)$ e, portanto, pode ser subtraída do sub-bloco de fala perceptual $s_w(n)$, resultado da filtragem de $s(n)$ por $W(z)$, antes de iniciar o procedimento de análise-por-síntese. O sinal resultante

$$t(n) = s_w(n) - \hat{s}_0(n), \tag{4.6}$$

é geramente denominado *sinal alvo*, sendo portanto o sinal que se deseja reproduzir da melhor maneira possível durante o processo de codificação. O filtro de síntese perceptual $H_w(z)$ passa então a ser da forma

$$H_w(z) = \frac{1}{A(z/\gamma)}. \tag{4.7}$$

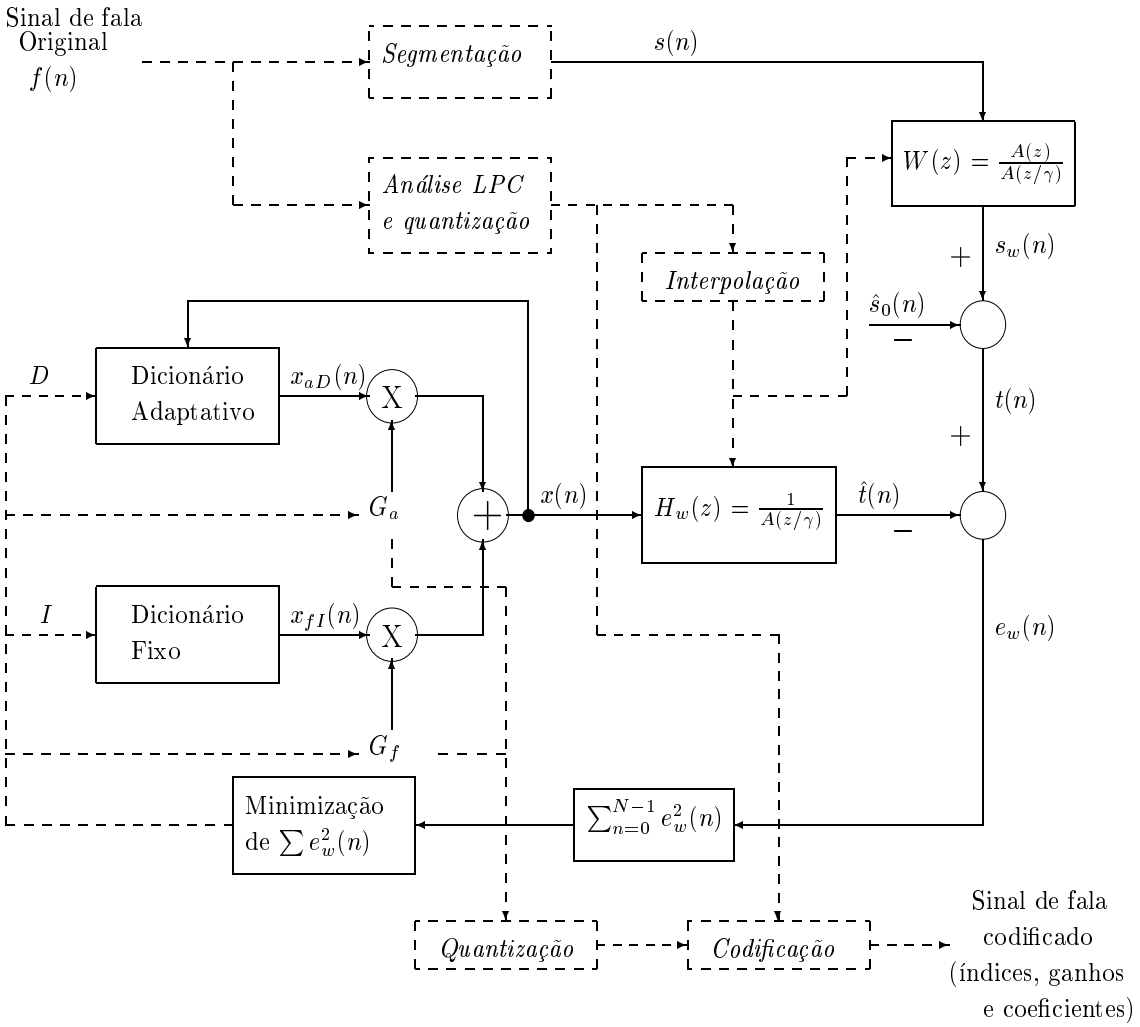


Figura 4.2: Diagrama de blocos detalhado do codificador CELP implementado.

Como pode ser visto na Figura 4.2, para testar cada seqüência candidata $x(n)$, é necessário apenas passá-la pelo filtro $H_w(z)$ com estado inicial zero. Esta estrutura propicia uma redução significativa no esforço computacional do processo de codificação.

Usando notação vetorial, onde por exemplo uma dada seqüência $f(n) = \{f(0), \ldots, f(N-1)\}$ passa a ser representada pelo vetor $\mathbf{f}$, da Figura 4.2 tem-se que o vetor reconstruído $\hat{\mathbf{t}}$ gerado pelos vetores $\mathbf{x_{aD}}$ e $\mathbf{x_{fI}}$, e pelos ganhos G_a e G_f, é dado por

$$\hat{\mathbf{t}} = G_a \mathbf{H_w x_{aD}} + G_f \mathbf{H_w x_{fI}}, \tag{4.8}$$

onde $\mathbf{H_w}$ é uma matriz Toeplitz triangular inferior cujos elementos são amostras da resposta ao impulso do filtro $H_w(z)$, ou seja,

$$\mathbf{H_w} = \begin{pmatrix} h_w(0) & 0 & \cdots & 0 \\ h_w(1) & h_w(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_w(N-1) & h_w(N-2) & \cdots & h_w(0) \end{pmatrix}. \tag{4.9}$$

Definindo a resposta de $H_w(z)$ à entrada $x_{aD}(n)$ como $y_{aD}(n)$, e à entrada $x_{fI}(n)$ como $y_{fI}(n)$, então

$$\mathbf{y_{aD}} = \mathbf{H_w x_{aD}} \quad \text{e} \quad \mathbf{y_{fI}} = \mathbf{H_w x_{fI}}. \tag{4.10}$$

A busca nos dicionários é feita de forma seqüencial: primeiro determina-se o melhor vetor no dicionário adaptativo considerando-o como único dicionário; depois é feita a procura pelo vetor no dicionário fixo considerando o vetor ótimo $\mathbf{x_{aD_{ot}}}$ e o ganho G_a determinados. Ao final do processo, quando também é determinada a melhor excitação do dicionário fixo, o ganho G_a é recalculado juntamente com o ganho G_f do dicionário fixo. Portanto, inicialmente o vetor alvo é $\mathbf{t}$, e deseja-se determinar o vetor $\mathbf{x_{aD}}$ que minimiza a energia do sinal de erro perceptual $\mathbf{e}_w$, dada por

$$\varepsilon = \mathbf{e_w}^T \mathbf{e_w} = (\mathbf{t} - G_a \mathbf{y_{aD}})^T (\mathbf{t} - G_a \mathbf{y_{aD}}), \tag{4.11}$$

onde T indica transposição. O ganho G_a que minimiza ε é obtido fazendo $\partial\varepsilon/\partial G_a = 0$, isto é

$$G_a = \frac{\mathbf{t}^T \mathbf{y_{aD}}}{\mathbf{y_{aD}}^T \mathbf{y_{aD}}}. \tag{4.12}$$

Substituindo (4.12) em (4.11), tem-se

$$\varepsilon = \mathbf{t}^T \mathbf{t} - \frac{(\mathbf{t}^T \mathbf{y_{aD}})^2}{\mathbf{y_{aD}}^T \mathbf{y_{aD}}}. \tag{4.13}$$

O erro ε dado por (4.13) é avaliado para cada vetor candidato $\mathbf{x_{aD}}$, sendo escolhido aquele que ocasionar o menor de todos os valores. Tendo feito isso, o ganho é calculado através de (4.12) com $\mathbf{y_{aD_{ot}}}$, que é a resposta do filtro $\mathbf{H_w}$ ao melhor vetor escolhido no dicionário adaptativo $\mathbf{x_{aD_{ot}}}$, no lugar de $\mathbf{y_{aD}}$.

Escolhido o melhor vetor do dicionário adaptativo $\mathbf{x_{aD_{ot}}}$ e determinado o ganho G_a, a busca do dicionário fixo é iniciada pela definição do novo sinal alvo $\mathbf{t_1}$ dado por

$$\mathbf{t_1} = \mathbf{t} - G_a \mathbf{y_{aD_{ot}}}. \tag{4.14}$$

A partir daí a busca é realizada da mesma maneira que no caso do dicionário adaptativo, ou seja, determina-se o vetor $\mathbf{x_{fI}}$ que minimize o erro

$$\varepsilon = \mathbf{t_1}^T \mathbf{t_1} - \frac{(\mathbf{t_1}^T \mathbf{y_{fI}})^2}{\mathbf{y_{If}}^T \mathbf{y_{fI}}}. \tag{4.15}$$

A determinação do ganho G_f é feita com a recomputação do ganho G_a, através de [23]

$$G_a = \frac{\mathbf{t}^T \mathbf{y_{aD}} - \mathbf{y_{aD}}^T \mathbf{y_{fI}} \mathbf{t}^T \mathbf{y_{fI}}}{\mathbf{y_{aD}}^T \mathbf{y_{aD}}}, \tag{4.16}$$

$$G_f = \frac{\mathbf{t}^T \mathbf{y_{fI}} - \mathbf{y_{aD}}^T \mathbf{y_{fI}} \mathbf{t}^T \mathbf{y_{aD}}}{\mathbf{y_{fI}}^T \mathbf{y_{fI}}}. \tag{4.17}$$

Esta forma de determinar os ganhos conduz a melhor qualidade que a maneira na qual o ganho G_f é determinado por uma expressão semelhante àquela mostrada em (4.12) [23].

O método de busca da melhor excitação descrito é sub-ótimo, pois não testa todas as combinações possíveis de excitações $x_{aD}(n)$ e $x_{fI}(n)$ para determinar o sinal mais próximo de $t(n)$. Para isto, no entanto, o esforço computacional seria demasiado grande.

4.5 O dicionário adaptativo

O dicionário adaptativo usado é do tipo com atrasos fracionários, cobrindo a faixa de 20 a 147 amostras com diferentes resoluções, conforme mostra a Tabela 4.3. No total 511 atrasos são usados, o que corresponde ao mesmo tipo de dicionário utilizado em [10].

Durante o processo de codificação a resposta do filtro de síntese perceptual $H_w(z)$ a cada seqüência candidata $x_{aD}(n)$, denotada por $y_{aD}(n)$, é correlacionada

Tabela 4.3: Característica do dicionário adaptativo utilizado.

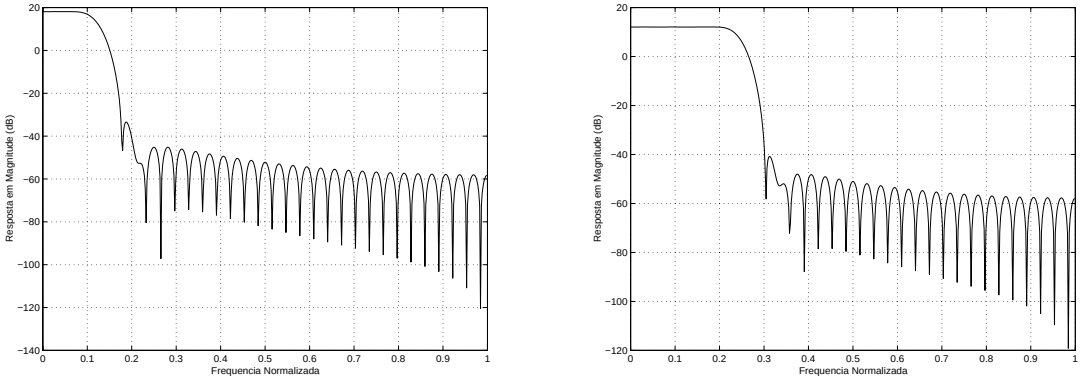
Faixa (atrasos)	Faixa (Hz)	Resolução (em amostras)	No. de atrasos
20 a 55	400 a 145	1/8	281
55 a 101	145 a 79	1/4	184
101 a 147	79 a 54	1	46
Total de atrasos	511		

com o sinal alvo $t(n)$. Esta correlação é obtida através de

$$C = \sum_{n=0}^{N-1} y_{aD}(n)t(n), \tag{4.18}$$

onde N é o número de amostras das seqüências $y_{aD}(n)$ e $t(n)$, e corresponde ao tamanho de cada sub-bloco em amostras. Somente seqüências candidatas que ocasionam $C > 0$ são levadas em conta. Caso $C < 0$ para todas as seqüências candidatas de um determinado sub-bloco, o índice D é ajustado para 0 para indicar que naquele sub-bloco a excitação será formada apenas pelo dicionário fixo. Segundo [9], este procedimento melhora a qualidade porque é obtida uma estimativa mais consistente do atraso. Outra vantagem é a de que o ganho G_a assume somente valores positivos, o que reduz a faixa dinâmica e ocasiona uma quantização mais eficiente.

A obtenção de cada seqüência candidata $x_{aD}(n)$ é feita filtrando o vetor de amostras das melhores excitações passadas, do qual é composto o dicionário adaptativo, pelas componentes polifásicas dos filtros interpoladores, conforme procedimento descrito em [5]. Para cada faixa de atrasos onde a resolução é diferente foi utilizado um diferente filtro interpolador de resposta ao impulso finita de ordem 64, projetados através do método de janelamento ao truncar a resposta ao impulso de um filtro ideal com a janela de Hamming [39]. A Figura 4.3(a) mostra a resposta em amplitude do filtro usado para a faixa de resolução em oitavas (20 a 55 amostras), enquanto que a Figura 4.3(b) representa o filtro usado para a faixa de resolução em quartas (55 a 101 amostras).



(a) (b)

Figura 4.3: Respostas em amplitude dos filtros interpoladores usados na busca da melhor excitação no dicionário adaptativo: (a) para a faixa de resolução em oitavas; (b) para a faixa de resolução em quartas.

4.5.1 Procedimento de busca rápida

O dicionário adaptativo corresponde a uma seqüência $C_a(n)$ que armazena amostras da melhor excitação passada, ou seja,

$$C_a(n) = \{x(-D_{max}), \ldots, x(-1)\}, \tag{4.19}$$

onde D_{max} é o comprimento do dicionário adaptativo. As seqüências candidatas $x_{aD}(n)$ são obtidas a partir de $C_a(n)$ de acordo com cada atraso D segundo procedimento mostrado no Capítulo 2. Quando o dicionário é do tipo fracionário, a busca pela melhor excitação exige grande esforço computacional, devido à filtragem de $C_a(n)$ pelas componentes polifásicas dos filtros interpoladores [10] e também pela filtragem de todas as seqüências candidatas por $H_w(z)$.

Em [40] é proposto um método de busca rápida no dicionário adaptativo através da modificação do modelo de síntese do codificador CELP, evitando que cada seqüência $x_{aD}(n)$ seja filtrada por $H_w(z)$. O método proposto aqui é ligeiramente diferente [41]. Trata-se de filtrar o conteúdo do dicionário adaptativo, a seqüência $C_a(n)$, por $H_w(z)$ antes da busca pela melhor excitação, formando um

novo dicionário adaptativo filtrado $C_{aF}(n)$, ou seja,

$$C_{aF}(n) = h_w(n) * C_a(n), \tag{4.20}$$

onde $h_w(n)$ é a resposta ao impulso de $H_w(z)$ com tamanho D_{max}. A busca pela melhor excitação passa então a ser feita em $C_{aF}(n)$, sem que haja a necessidade de filtrar as seqüências candidatas por $H_w(z)$ individualmente. Sendo determinado o melhor atraso D, o restante do procedimento é feito como mostrado na Seção 4.4, ou seja, faz-se a leitura da seqüência ótima $x_{aD_{ot}}(n)$ em $C_a(n)$ (dicionário não filtrado) e depois filtra-a por $H_w(z)$, para obter $y_{aD_{ot}}(n)$. De posse deste último, determina-se o ganho G_a através de (4.12), e o novo sinal alvo $t_1(n)$ através de (4.14) para ser realizada a busca no dicionário fixo da mesma forma que no método convencional.

4.6 O dicionário fixo

É utilizado um dicionário fixo composto de 1082 amostras, onde cerca de 90% delas são nulas (dicionário esparso). Ele foi gerado anulando-se todas as amostras de um conjunto obtido a partir de ruído branco gaussiano com média zero e variância unitária, que estivessem abaixo de um determinado limiar. O valor escolhido heuristicamente foi 1,645 para possibilitar o grau de esparsidade desejado.

A idéia de obter este tipo de dicionário partiu da afirmação [9] de que dicionários cujas seqüências são obtidas através da ceifagem de amostras de ruído branco gaussiano de média zero e variância unitária, de modo a resultar em 90 a 95% de suas amostras nulas, obtêm melhor qualidade que os dicionários estocásticos, como o utilizado em [8]. Além da qualidade, os dicionários deste tipo proporcionam a oportunidade de se agilizar a busca pela melhor excitação através de métodos de busca rápida em dicionários esparsos [14]. O dicionário usado aqui é um pouco diferente daquele descrito em [9] porque as amostras com valores acima do limiar de ceifagem não são subtraídas de tal valor.

Foram feitos testes experimentais com os dois tipos de dicionários com o objetivo de determinar qual o melhor deles em qualidade, ao utilizá-los no codificador CELP da Tabela 4.7 sem quantização dos ganhos G_a e G_f. Foi utilizada a razão

Tabela 4.4: Razões sinal-ruído segmentada perceptual em dB, para o sistema CELP com os dois tipos de dicionários: o sem subtração de amostras acima do limiar (tipo 1); e o com subtração de amostras acima do limiar (tipo 2).

Dicionário	Locutor			
	M1	M2	F1	F2
Tipo 1	13,98	10,81	10,79	11,18
Tipo 2	13,79	10,70	10,80	11,16

sinal-ruído segmentada perceptual como medida objetiva de qualidade para avaliar os sinais reconstruídos pelos dois casos. Foram escolhidos 4 sentenças de fala (M1, M2, F1 e F2) para comparar a qualidade. A Tabela 4.4 mostra os resultados obtidos, onde "tipo 1" representa o dicionário que não ceifa as amostras cujo valor está acima do limiar (dicionário proposto); e "tipo 2" corresponde ao tipo onde ocorre ceifagem de amostras, descrito em [9]. Pode-se perceber que para os sinais dos locutores do sexo masculino o Tipo 1 foi melhor, enquanto que para os sinais do sexo feminino houve uma certa equivalência. Testes subjetivos informais mostraram que a qualidade entre os sinais reconstruídos pelos dois tipos foi essencialmente indistinguível.

A partir das 1082 amostras que compõem o dicionário fixo são obtidas 512 seqüências candidatas através de um procedimento de superposição [14], onde cada seqüência candidata difere da anterior de 4 amostras: as duas primeiras e as duas últimas, ou seja, as seqüências candidatas são obtidas a cada deslocamento de duas amostras no vetor que compõe o dicionário fixo, conforme mostra a Figura 4.4. A vantagem de se usar superposição está na redução da quantidade de memória necessária para armazenar o dicionário fixo, e também na possibilidade do uso de uma forma mais rápida de busca, que será descrita a seguir.

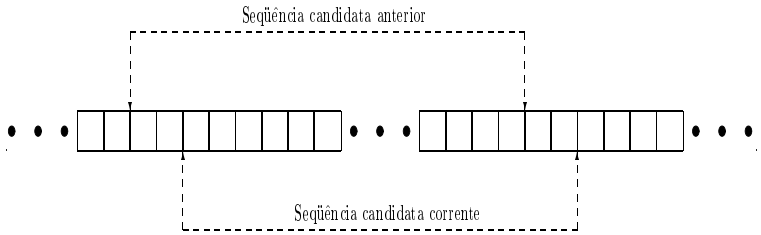


Figura 4.4: Forma de determinação das seqüências candidatas no dicionário fixo com superposição utilizado.

4.6.1 Procedimento de busca rápida

Para a obtenção das seqüências candidatas é utilizado um método de busca rápida que aproveita a esparsidade e superposição do dicionário fixo. Da Seção 4.4 sabe-se que a resposta do filtro de síntese perceptual $H_w(z)$ à seqüência candidata do dicionário fixo $x_{fI}(n) = \{x_{fI}(0), \ldots, x_{fI}(N-1)\}$ é a seqüência $y_{fI}(n) = \{y_{fI}(0), \ldots, y_{fI}(N-1)\}$, ou seja,

$$y_{fI}(n) = \sum_{j=0}^{n} h_w(n-j)x_{fI}(j), \quad \text{para} \quad 0 \leq n \leq N-1, \tag{4.21}$$

onde $h_w(n) = \{h_w(0), \ldots, h_w(N-1)\}$ é a resposta ao impulso do filtro $H_w(z)$ com N amostras, que corresponde ao tamanho de um sub-bloco.

A propriedade da esparsidade é aproveitada ao evitar que seja efetuado o produto $h_w(n-j)x_{fI}(j)$ quando a amostra $x_{fI}(j) = 0$, o que acontece para cerca de 90% dos casos.

Para entender como é aproveitada a propriedade da superposição, suponha que a seqüência candidata para $I = 0$ seja $x_{f0}(n) = \{x_{f0}(0), \ldots, x_{f0}(N-1)\}$, então

$$\begin{aligned}
y_{f0}(0) &= h_w(0)x_{f0}(0) \\
y_{f0}(1) &= h_w(1)x_{f0}(0) + h_w(0)x_{f0}(1) \\
y_{f0}(2) &= h_w(2)x_{f0}(0) + h_w(1)x_{f0}(1) + h_w(0)x_{f0}(2) \\
&\vdots \qquad \vdots \\
y_{f0}(N-1) &= h_w(N-1)x_{f0}(0) + h_w(N-2)x_{f0}(1) + \cdots + h_w(0)x_{f0}(N-1).
\end{aligned}$$

Suponha agora que a nova seqüência candidata seja

$$\begin{aligned} x_{f1}(n) &= \{x_{f1}(0), x_{f1}(1), \ldots, x_{f1}(N-3), x_{f1}(N-2), x_{f1}(N-1)\} \\ &= \{x_{f0}(2), x_{f0}(3), \ldots, x_{f0}(N-1), x_{f1}(N-2), x_{f1}(N-1)\}, \end{aligned}$$

então a sua resposta ao filtro de síntese será

$$\begin{aligned} y_{f1}(0) &= h_w(0)x_{f1}(0) \\ &= h_w(0)x_{f0}(2) \\ y_{f1}(1) &= h_w(1)x_{f1}(0) + h_w(0)x_{f1}(1) \\ &= h_w(1)x_{f0}(2) + h_w(0)x_{f0}(3) \\ y_{f1}(2) &= h_w(2)x_{f1}(0) + h_w(1)x_{f1}(1) + h_w(0)x_{f1}(2) \\ &= h_w(2)x_{f0}(2) + h_w(1)x_{f0}(3) + h_w(0)x_{f0}(4) \\ \vdots &\quad \vdots \\ y_{f1}(N-1) &= h_w(N-1)x_{f1}(0) + h_w(N-2)x_{f1}(1) + \cdots + h_w(0)x_{f1}(N-1) \\ &= h_w(N-1)x_{f0}(2) + \cdots + h_w(1)x_{f1}(N-2) + h_w(0)x_{f1}(N-1). \end{aligned}$$

Portanto, cada nova seqüência $y_{fI}(n)$ é obtida através de

$$y_{fI}(n) = y_{fI-1}(n+2) - \sum_{j=0}^{1} h_w(n+2-j)x_{fI-1}(j), \quad \text{para} \quad 0 \leq n \leq N-3, \tag{4.22}$$

$$y_{fI}(n) = \sum_{j=0}^{n} h_w(n-j)x_{fI}(j), \quad \text{para} \quad N-2 \leq n \leq N-1. \tag{4.23}$$

Esta forma de determinar a resposta do filtro de síntese à excitação do dicionário fixo reduz bastante o número de multiplicações, conforme se pode notar.

4.6.2 Excitação restrita

Para eliminar um certo grau de rouquidão introduzida pelo sistema CELP nos sinais decodificados, muitos sistemas empregam um procedimento onde a excitação do dicionário fixo é limitada de acordo com a eficiência do dicionário adaptativo na reprodução do sub-bloco considerado. O procedimento consiste primeiro em determinar uma medida que avalie a contribuição do dicionário adaptativo. De posse

de tal variável, o ganho G_f é atenuado para diminuir a componente do dicionário fixo.

Existem vários métodos para limitar a excitação do dicionário fixo [10, 21], sendo o utilizado no sistema implementado igual ao utilizado no DoD-CELP [42], no qual a contribuição do dicionário adaptativo é dada por

$$R = \frac{\sum_{i=0}^{N-1} t(n)t_1(n)}{\sum_{i=0}^{N-1} t(n)t(n)}, \tag{4.24}$$

onde $t(n)$ é o sinal alvo para a busca no dicionário adaptativo, $t_1(n)$ é o sinal alvo para o dicionário fixo e N é o tamanho de um sub-bloco. Partindo-se deste valor, o novo ganho do dicionário fixo $\tilde{G}_f$ é obtido a partir do ganho G_f por

$$\tilde{G}_f = \begin{cases} 0,2G_f, & \text{para} \quad |R| < 0,04, \\ 1,4G_f\sqrt{|R|}, & \text{para} \quad |R| > 0,81, \\ G_f\sqrt{|R|}, & \text{para outro} \ \ R. \end{cases} \tag{4.25}$$

Esta correção de ganho é feita no final de cada procedimento de análise-por-síntese, antes da quantização.

4.7 Quantização dos ganhos e alocação de bits

Para a quantização dos ganhos G_a e $\tilde{G}_f$, o sistema CELP da Tabela 4.7 foi executado em uma base de dados consistindo de 17 frases, onde:

- 12 foram retiradas das listas de frases foneticamente balanceadas para o português falado no Rio de Janeiro [31] e pronunciados por 3 locutores do sexo masculino e um do sexo feminino, onde cada um foi responsável por 3 frases;
- 5 frases da língua inglesa, onde duas foram pronunciadas por locutores do sexo masculino e 3 do sexo feminino.

As frases foram diferentes das utilizadas para o projeto dos quantizadores do Capítulo 3, com exceção dos sinais da língua inglesa. Os locutores foram os mesmos.

A Figura 4.5 mostra a distribuição do ganho G_a, enquanto que a Figura 4.6 mostra a distribuição para $\tilde{G}_f$. Os valores dos ganhos foram considerados somente

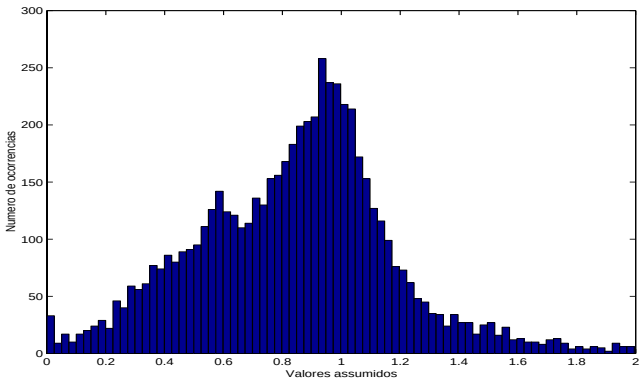


Figura 4.5: Distribuições do ganho G_a, considerando apenas valores na faixa $[0; 2]$.

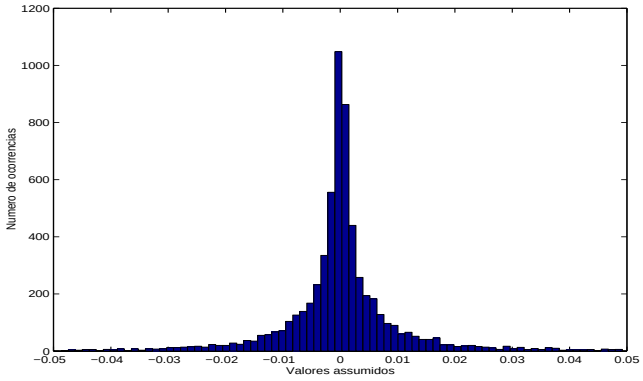


Figura 4.6: Distribuições do ganho $\tilde{G}_f$, considerando apenas valores na faixa $[-0,05; 0,05]$.

nas faixas de 0 a 2 para G_a e de -0,05 a 0,05 para $\tilde{G}_f$. Sobre estas faixas foram determinadas medidas importantes como média, variância, e valores máximos e mínimos, conforme mostra a Tabela 4.5. Além destas características, também são mostrados os percentuais de valores que situam-se fora delas. Pode-se perceber que o fato de estabelecer tais faixas não prejudica porque cerca de 96% dos valores assumidos pelos ganhos encontram-se nelas.

Os ganhos G_a e $\tilde{G}_f$ são quantizados de maneira escalar com 4 e 5 bits, respectivamente após serem determinados por (4.16) e (4.25). Os quantizadores não uniformes para cada um deles foram obtidos através de um treinamento pelo algoritmo de Lloyd I [33] da base de dados comentada acima, considerando apenas os

Tabela 4.5: Características de G_a e $\tilde{G}_f$ dentro das faixas $[0;2]$ e $[-0,05;0,05]$, respectivamente.

Ganho	Características				
	Valor Máx.	Valor Mín.	Média	Variância	% Fora da faixa
G_a	2	0	0,8564	0,8420	2,7%
$\tilde{G}_f$	0,0492	-0,0486	$8,1 \cdot 10^{-5}$	$1,03 \cdot 10^{-4}$	1,3%

Tabela 4.6: Alocação de bits para os parâmetros da excitação.

Parâmetro	Faixa	No. de bits
Ganho do dicionário fixo ($\tilde{G}_f$)	-0,05 a 0,05	5
Índice do dicionário fixo (I)	0 a 511	9
Ganho do dicionário adaptativo (G_a)	0 a 2	4
Índice do dicionário adaptativo (L)	0 a 511	9
Total	27 bits	

valores dentro das faixas idealizadas.

A alocação de bits para os parâmetros da excitação está mostrada na Tabela 4.6.

4.8 Resumo das características

Considerando apenas os bits necessários para a reprodução do sinal de fala no decodificador, a taxa do sistema implementado fica em torno de 4,67 kbps, conforme se vê na Tabela 4.7, que resume todas as suas características.

Tabela 4.7: Resumo das características do sistema CELP implementado.

Tamanho de cada bloco:	30 ms.
Tamanho de cada sub-bloco:	7,5 ms.
Análise LPC:	Método da autocorrelação com janela de Hamming de 25 ms (200 amostras), centrada no último sub-bloco de cada bloco.
Número de CPL:	10 ($p = 10$).
Quantização dos CPL:	Quantizador *QDLSF-32*.
Interpolação dos CPL:	Linear, no domínio das LSF.
Dicionário adaptativo:	Atrasos fracionários de 20 a 147 com diferentes resoluções em determinadas faixas (Tabela 4.3).
Dicionário fixo:	512 seqüências geradas a partir de um vetor esparso de 1082 amostras obtidas a partir de de ruído branco gaussiano.
Quantização de G_a	Não-uniforme de 0 a 2.
Quantização de $\tilde{G}_f$	Não-uniforme de -0,05 a 0,05
Parâmetro perceptual de $W(z)$:	$\gamma = 0,8$
Taxa de bits:	CPL: 32 bits/30 ms = 1,067 kbps. Excitação: 27 bits/7,5 ms = 3,6 kpbs. Total: 4,667 kbps.

Tabela 4.8: Sentenças usadas para a avaliação objetiva do sistema CELP implementado.

Locutor	Sentença	Duração (s)
M1	Eu vi logo a Iô-Iô e o Leo Um homem não caminha sem um fim	5,25
M2	Vi Zé fazer essas viagens seis vezes O atabaque do Tito é coberto com pele de gato	6,75
F1	Paira um ar de arara rara no Rio Real Foi muito difícil entender a canção	5
F2	Depois do almoço te encontro Esses são nossos filhos	6

4.9 Avaliação do sistema implementado

O desempenho do sistema implementado foi avaliado usando três medidas objetivas de qualidade e testes subjetivos informais. Os testes foram realizados em 4 sentenças, cada uma delas composta de duas frases pronunciadas por um locutor diferente, onde 2 foram do sexo masculino (M1 e M2), e 2 do sexo feminino (F1 e F2). As frases e os locutores foram diferentes dos utilizados nos projetos dos quantizadores dos CPL e ganhos dos dicionários. A Tabela 4.8 mostra as sentenças usadas, bem como a duração de cada uma delas. Os sinais foram digitalizados com 16 bits por amostra a 8 kHz, com posterior filtragem passa-altas para a remoção de ruídos de baixa freqüência.

A avaliação consistiu primeiro em comparar o sistema implementado com o algoritmo DoD-CELP [42] que faz parte da regulamentação FS-1016 e opera a uma taxa de 4,8 kbps. A segunda etapa consistiu em verificar o funcionamento do método de busca rápida no dicionário adaptativo, mostrado na Sub-seção 4.5.1.

4.9.1 Medidas objetivas de qualidade usadas

As medidas objetivas usadas para a avaliação do sistema foram a razão sinal-ruído segmentada perceptual (RSRSP) [5], a distância cepestral (DC) [43] e a distância de Itakura (DI) [4]. Cada uma delas é brevemente descrita a seguir, bem como a forma como foram usadas. A DC e a DI são explicadas de uma maneira mais detalhada no Capítulo 5.

4.9.1.1 Razão sinal-ruído segmentada perceptual

Sejam $v(n)$ e $\hat{v}(n)$ os sinais de fala original e reconstruído, respectivamente. A RSRSP entre eles é dada por

$$RSR_{SP} = \frac{1}{M}\sum_{i=0}^{M-1} 10\log\left\{\frac{\sum_{n=0}^{N-1} v^2(n+iN)}{\sum_{n=0}^{N-1} e_w^2(n+iN)}\right\} \quad (dB), \tag{4.26}$$

onde o sinal de erro perceptual $e_w(n)$ é obtido ao filtrar o sinal de erro $e(n) = v(n) - \hat{v}(n)$ pelo filtro perceptual $W(z) = A(z)/A(z/\gamma)$, M é o número de segmentos dos sinais, e N é o tamanho de cada segmento em amostras. A idéia da RSRSP é dividir os sinais em M segmentos iguais de comprimento N e determinar a razão sinal-ruído perceptual para cada um deles. O valor efetivo será a média aritmética de todas as medidas, conforme mostra (4.26). Os coeficientes do filtro $A(z)$ são obtidos a partir do sinal original.

Neste trabalho foi utilizado $N = 80$, ou seja, os sinais foram segmentados a 10 ms, e a análise LPC foi feita pelo método da autocorrelação de ordem 10 com uma janela de Hamming de 25 ms centrada em cada segmento de 10 ms correspondente.

Segundo [5], a RSRSP possui boa correlação com as medidas subjetivas de qualidade devido ao fato de ser utilizado o erro perceptual $e_w(n)$ ao invés do erro simples $e(n)$.

4.9.1.2 Distância cepestral

A DC entre um segmento $s(n)$ do sinal de fala original e o correspondente $\hat{s}(n)$ do sinal de fala reconstruído, é dada por [43]

$$D_C = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{p'}[c_i - \hat{c}_i]^2} \qquad (dB), \tag{4.27}$$

onde c_i e $\hat{c}_i$ correspondem aos coeficientes cepstrais de $s(n)$ e $\hat{s}(n)$, respectivamente.

Neste trabalho foi utilizado $N = 80$. As análises LPC foram de ordem 16 feitas pelo método da autocorrelação com janelas de Hamming de 25 ms centradas em cada segmento correspondente. O número de coeficientes cepstrais usado foi 32.

4.9.1.3 Distância de Itakura

A DI entre um determinado segmento $s(n)$ do sinal de fala original e o correspondente segmento $\hat{s}(n)$ do sinal de fala reconstruído é dada por [4]

$$D_I = \frac{1}{2}\left[10\log\left\{\frac{\hat{\alpha}\cdot\mathbf{R_s}\cdot\hat{\alpha}}{\alpha\cdot\mathbf{R_s}\cdot\alpha}\right\} + 10\log\left\{\frac{\alpha\cdot\mathbf{R_{\hat{s}}}\cdot\alpha}{\hat{\alpha}\cdot\mathbf{R_{\hat{s}}}\cdot\hat{\alpha}}\right\}\right] \qquad (dB), \tag{4.28}$$

onde $\alpha = [1 \quad -a_1 \cdots \quad -a_p]^T$ e $\hat{\alpha} = [1 \quad -\hat{a}_1 \cdots \quad -\hat{a}_p]^T$ são vetores obtidos, respectivamente, a partir dos CPL $\{a_1, \ldots, a_p\}$ e $\{\hat{a}_1, \ldots, \hat{a}_p\}$ dos segmentos $s(n)$ e $\hat{s}(n)$, sendo $\mathbf{R_s}$ e $\mathbf{R_{\hat{s}}}$ suas respectivas matrizes de autocorrelação.

Nos testes foi usado $N = 80$. As análises LPC de ordem 16 foram feitas através do método da autocorrelação com janelas de Hamming de 25 ms centradas nos segmentos correspondentes.

4.9.2 Resultados

A Tabela 4.9 mostra as medidas de qualidade para o sistema implementado, enquanto que a Tabela 4.10 mostra os resultados para o DoD-CELP. Considerando a RSRSP, o sistema proposto foi melhor para todas as sentenças; considerando a DC e a DI, o DoD-CELP obteve melhor qualidade para os sinais femininos (F1 e F2). Uma avaliação subjetiva informal mostrou que o sistema implementado foi melhor que o DoD-CELP para todas as sentenças, com exceção de M2, onde a qualidade

Tabela 4.9: Medidas objetivas de qualidade para o sistema implementado.

Locutor	RSRSP (dB)	DC (dB)	DI (dB)
M1	15,62	3,31	1,34
M2	16,28	3,27	1,32
F1	17,08	3,55	1,55
F2	15,03	3,52	1,54

Tabela 4.10: Medidas objetivas de qualidade para o DoD-CELP 4,8 kbps.

Locutor	RSRSP (dB)	DC (dB)	DI (dB)
M1	9,43	3,54	1,53
M2	9,92	3,51	1,50
F1	13,89	3,40	1,40
F2	11,45	3,51	1,49

foi equivalente. Como uma forma de ilustração, a Tabela 4.11 mostra os resultados para o LD-CELP, cuja qualidade é considerada de boa para excelente segundo o teste de MOS, e opera a uma taxa de 16 kbps.

Tabela 4.11: Medidas objetivas de qualidade para o LD-CELP 16 kbps.

Locutor	RSRSP (dB)	DC (dB)	DI (dB)
M1	20,76	2,38	0,73
M2	21,36	2,30	0,68
F1	22,69	2,20	0,63
F2	21,09	2,12	0,58

A Tabela 4.12 mostra os resultados para o sistema CELP implementado com o procedimento de busca rápida no dicionário adaptativo. Pode-se perceber que a qualidade não foi comprometida porque as diferenças foram bem pequenas. Consi-

Tabela 4.12: Medidas objetivas de qualidade para o sistema implementado com o procedimento de busca rápida no dicionário adaptativo.

Locutor	RSR-SP (dB)	DC (dB)	DI (dB)
M1	15,51	3,36	1,37
M2	16,02	3,27	1,31
F1	16,88	3,35	1,36
F2	14,94	3,50	1,50

derando a RSRSP, este procedimento piorou a qualidade para todas as sentenças. Considerando a DC e a DI, a qualidade foi equivalente para M2 e melhor para os sinais femininos (F1 e F2). Uma avaliação subjetiva informal mostrou que a qualidade dos sistemas com e sem o método é indistinguível.

A fim de comprovar a eficácia do procedimento de busca rápida no dicionário adaptativo em termos de agilidade de busca, os sistemas com e sem este método foram executados em uma estação de trabalho Sun Ultra-60. Os tempos de processamento (TP) para a codificação estão indicados em segundos, bem como o percentual de tempo real (TR), conforme mostra a Tabela 4.13. Este último fator, para uma determinada sentença, é dado pela relação entre o tempo que o sistema CELP leva para codificá-la e o seu tempo de duração, ou seja,

$$\%TR = \frac{\text{Tempo de Codificação da Sentença}}{\text{Tempo de Duração da Sentença}} \times 100. \tag{4.29}$$

Pode-se perceber que em todas as situações o sistema com busca rápida foi mais veloz, efeito este causado pela não filtragem de cada seqüência candidata do dicionário adaptativo pelo filtro de síntese perceptual $H_w(z)$.

Tabela 4.13: Tempos de processamento (TP) e percentuais de tempo real (%TR) para os sistemas com e sem busca rápida no dicionário adaptativo.

	Sem busca rápida		Com busca rápida		Redução do tempo
Locutor	TP (s)	$\%TR$	TP (s)	$\%TR$	de processamento (%)
M1	38	723	26	495	31,6
M2	48	711	33	488	31,3
F1	35	700	24	480	31,4
F2	42	700	30	500	28,6

4.10 Conclusão

Neste capítulo foram mostradas as características de um sistema de codificação CELP que funciona à taxa de 4,67 kbps. O sistema utiliza a quantização *QDSLF-32* descrita no Capítulo 2 para os coeficientes de predição linear, dicionário adaptativo com atrasos fracionários, dicionário fixo obtido a partir de seqüências estocásticas, e método seqüencial para determinação da melhor excitação.

Durante a fase de projeto do sistema foram comparados alguns fatores de forma experimental, tendo chegado aos seguintes resultados:

- na análise de predição linear, o método de suavização espectral mostrou melhor qualidade que o método de expansão dos coeficientes [32] para três das 4 sentenças quando considerada a razão sinal-ruído segmentada perceptual, enquanto que uma avaliação subjetiva informal mostrou qualidade indistinguível;

- o dicionário fixo empregado, que consistiu somente em anular amostras de ruído branco gaussiano abaixo do limiar 1,645, foi comparado com o tipo descrito em [10], que também subtrai amostras acima do limiar, e constatou-se que em termos de razão sinal-ruído o dicionário proposto foi ligeiramente melhor, apesar de uma avaliação subjetiva informal revelar qualidade indistinguível;

Quanto ao desempenho do sistema proposto, que foi avaliado através de três

medidas de qualidade objetivas e avaliações subjetivas informais, sendo comparado com o DoD-CELP, os resultados mostraram que:

- segundo a RSRSP, o sistema proposto obtém melhor qualidade para todas as sentenças testadas;
- para as medidas DC e DI a qualidade foi melhor para M1 e M2, e pior para as sentenças F2 (ligeiramente) e F1;
- segundo avaliações subjetivas informais, o sistema proposto é melhor em todas as sentenças, com exceção para M2, onde a qualidade entre os sinais reproduzidos pelos dois codificadores foi equivalente.

Quanto ao procedimento de busca rápida no dicionário adaptativo proposto na Sub-seção 4.5.1:

- o sistema com busca rápida foi ligeiramente pior em todos os casos considerando a RSRSP;
- considerando a DC e a DI, o sistema com o procedimento de busca rápida obteve qualidade pior em relação ao sistema sem o método para M1, equivalente para M2, e melhor para F1 e F2;
- uma avaliação subjetiva informal mostrou que os sinais processados pelo sistema com e sem o método de busca rápida possuem a mesma qualidade;
- o procedimento de busca rápida diminui o tempo de processamento de 28 a 32%.

Capítulo 5

Análise de sinais de fala em sub-bandas

5.1 Introdução

Para tornar possível uma maior eficiência dos sistemas de codificação da fala, no que diz respeito à qualidade do sinal reconstruído e taxa de bits consumida, é necessário que sejam exploradas cada vez mais características próprias destes sinais [12, 13, 21]. Este capítulo procura dar uma contribuição neste estudo através de uma análise de estacionaridade em sub-bandas. Em [44] é mostrada uma forma de determinar o grau de estacionaridade de um sinal digital através da sua decomposição em funções bases. O método usado aqui consiste em quantificar a variação da envoltória dos espectros de amplitude de curto termo em instantes consecutivos [45].

Na Seção 4.2 é descrito o procedimento utilizado para analisar a variação espectral em sub-bandas; na Seção 4.3 são abordadas algumas medidas de distâncias; na Seção 4.4 são tratados os experimentos realizados; e na Seção 4.5 estão as conclusões.

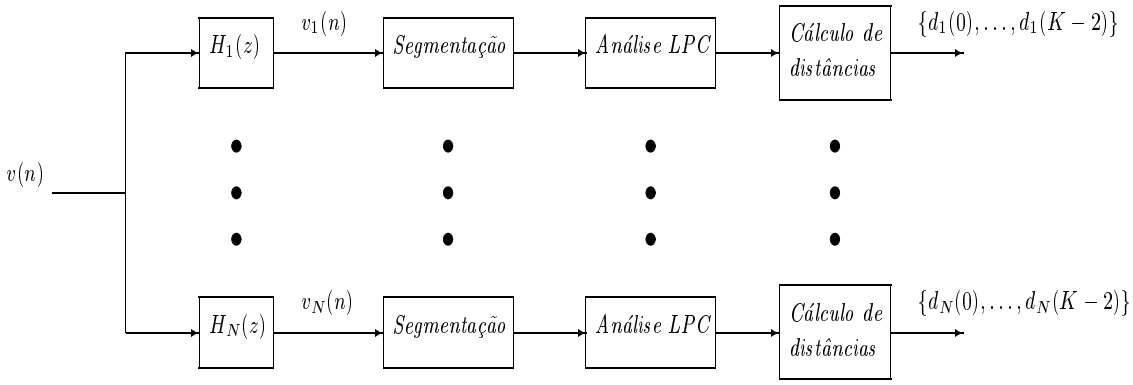


Figura 5.1: Procedimento utilizado para analisar a estacionaridade de sinais de fala em sub-bandas.

5.2 Método utilizado para a análise

O método utilizado para analisar a estacionaridade de sinais de fala em sub-bandas determina diferenças entre as envoltórias dos espectros de magnitude de segmentos adjacentes dos sinais sub-bandas, que são obtidos na saída de um banco de filtros de análise. A representação detalhada deste procedimento é vista na Figura 5.1. O sinal de fala $v(n)$ é passado por um banco de filtros de análise onde são obtidos os sinais $v_1(n), v_2(n), \ldots, v_N(n)$, sendo N o número de bandas do banco. Depois os sinais $v_i(n)$ são segmentados, e em cada segmento é feita uma análise LPC, que tem por finalidade determinar o modelo

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \tag{5.1}$$

cuja resposta em amplitude representa uma aproximação para a envoltória do espectro de magnitude do segmento de sinal de fala correspondente [4, 6]. A partir daí são determinadas, para cada sinal $v_i(n)$, distâncias entre as envoltórias espectrais de segmentos adjacentes, ou seja, é feita uma medida de variação destas envoltórias ao longo do tempo.

Para cada sub-banda i os vetores $\mathbf{d}_i$ armazenam as distâncias entre segmentos adjacentes. Tais vetores têm dimensões $K-1$, onde K é o número de segmentos dos sinais $s_i(n)$. Assim, por exemplo, $d_1(0)$ corresponde à distância entre as envoltórias espectrais do primeiro e segundo segmentos do sinal $v_1(n)$; $d_2(1)$ à distância entre as

envoltórias do segundo e terceiro segmentos do sinal $v_2(n)$, e assim sucessivamente.

Um determinado grau de variação espectral (GVE) para a sub-banda i, g_i, é dado pela média dos elementos do vetor $\mathbf{d}_i$, ou seja,

$$g_i = \frac{1}{K-1} \sum_{k=0}^{K-2} d_i(k), \tag{5.2}$$

onde $d_i(k)$ representa o k-ésimo elemento do vetor $\mathbf{d}_i$. Uma variação percentual da sub-banda i em relação às demais é definida por

$$\Delta_i\ (\%) = \frac{g_i}{\sum_{i=1}^{N} g_i} \times 100. \tag{5.3}$$

Esta medida de variação percentual relativa (VPR) fornece uma melhor visualização das variações espectrais nas sub-bandas.

5.3 Medidas de distância

As medidas de distância são responsáveis pela quantificação da diferença entre duas envoltórias espectrais de curto termo, para dois segmentos distintos de um sinal de fala.

5.3.1 Distância espectral logarítmica

Dados dois modelos $\sigma/A(z)$ e $\sigma'/A'(z)$, onde

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \quad \text{e} \quad A'(z) = 1 - \sum_{i=1}^{p} a'_i z^{-i}, \tag{5.4}$$

o erro ou diferença entre estes modelos em uma escala de amplitude logarítmica versus a escala de freqüência é dado por

$$V(\theta) = \ln \left[\frac{\sigma^2}{|A(e^{j\theta})|^2} \right] - \ln \left[\frac{(\sigma')^2}{|A'(e^{j\theta})|^2} \right], \tag{5.5}$$

onde θ é a freqüência normalizada, com $\theta = \pi$ representando a metade da freqüência de amostragem. Segundo [35], uma família de medidas de distância entre os modelos espectrais seria o conjunto de parâmetros D_p, onde

$$(D_p)^p = \int_{-\pi}^{\pi} |V(\theta)|^p \frac{d\theta}{2\pi}. \tag{5.6}$$

Para $p = 1$, tem-se a medida espectral logarítmica absoluta; para $p = 2$, a medida espectral logarítmica média quadrática (*melmq*); e finalmente para o caso limite quando $p \to \infty$, tem-se a medida espectral logarítmica de pico. Neste último as diferenças de pico entre os dois espectros, que para o caso de um sinal de fala seriam os formantes, têm maior peso. O caso $p = 2$ representa o valor médio da diferença entre os dois espectros. Segundo experimentos realizados em [35], as distâncias d_2 e d_∞ assemelham-se. Neste trabalho foi optado por uma D_2, que passará a ser chamada de D_{EL}, dada por

$$D_{EL} = \frac{10}{\ln 10}\sqrt{\frac{1}{L}\sum_{k=-L/2}^{L/2}\left\{\ln\left|\frac{\sigma}{A(e^{\frac{jk\pi}{L}})}\right|^2 - \ln\left|\frac{\sigma'}{A'(e^{\frac{jk\pi}{L}})}\right|^2\right\}^2} \quad (dB), \tag{5.7}$$

onde L é o número de pontos usados para aproximar a integração de (5.6).

Se os ganhos σ e σ' dos modelos forem normalizados pelas energias dos respectivos segmentos, obtém-se a distância espectral logarítmica normalizada pela energia:

$$D'_{EL} = \frac{10}{\ln 10}\sqrt{\frac{1}{L}\sum_{k=-L/2}^{L/2}\left\{\ln\left|\frac{\tau}{A(e^{\frac{jk\pi}{L}})}\right|^2 - \ln\left|\frac{\tau'}{A'(e^{\frac{jk\pi}{L}})}\right|^2\right\}^2} \quad (dB). \tag{5.8}$$

A diferença entre esta expressão e a mostrada em (5.7) é que os ganhos σ e σ' são substituídos respectivamente por τ e τ', onde

$$\tau = \frac{\sigma}{\varepsilon} \quad \text{e} \quad \tau' = \frac{\sigma'}{\varepsilon'}. \tag{5.9}$$

Os termos ε e ε' representam respectivamente as energias dos segmentos de sinal de fala que deram origem aos modelos $\sigma/A(z)$ e $\sigma'/A'(z)$, ou seja,

$$\varepsilon = \sum_{i=0}^{N_s-1} s^2(i) \quad \text{e} \quad \varepsilon' = \sum_{i=0}^{N_s-1} s'^2(i), \tag{5.10}$$

onde N_s é o tamanho de cada segmento.

5.3.2 Distância cepestral

Os coeficientes cepstrais podem ser obtidos a partir de amostras do sinal de fala ou a partir dos CPL, para um dado segmento. Neste último caso eles são

chamados coeficientes cepstrais LPC, e podem ser obtidos através de [35]

$$\begin{aligned} c_1 &= a_1 \\ c_i &= a_i + \sum_{k=1}^{i-1} \frac{k}{i} c_k a_{i-k}, \quad \text{para} \quad 1 \leq i \leq p \\ c_i &= \sum_{k=1}^{p} \left(1 - \frac{k}{i}\right) c_{i-k} a_k, \quad \text{para} \quad p < i \leq p' \end{aligned} \tag{5.11}$$

onde $\{c_1, \ldots, c_{p'}\}$ são os coeficientes cepstrais LPC, $\{a_1, \ldots, a_p\}$ os CPL do modelo em (5.1), e p' o número de coeficientes cepstrais.

Uma medida de distância cepestral (DC) entre os coeficientes cepstrais c_i e $\hat{c}_i$ pode ser obtida por [43]

$$D_C = \frac{10}{\ln 10} \sqrt{2 \sum_{i=1}^{M} [c_i - \hat{c}_i]^2} \quad (dB), \tag{5.12}$$

que é ligeiramente diferente daquela mostrada em [35]:

$$D'_C(dB) = \frac{10}{\ln 10} \sqrt{(c_0 - \hat{c}_0)^2 + 2 \sum_{i=1}^{L} [c_i - \hat{c}_i]^2}, \tag{5.13}$$

onde

$$c_0 = \ln\{\sigma^2\} \quad \text{e} \quad \hat{c}_0 = \ln\{\sigma'^2\}. \tag{5.14}$$

Os termos σ e σ' representam respectivamente os ganhos dos modelos espectrais $\sigma/A(z)$ e $\sigma'/A(z)$. De fato, a distância cepestral de (5.12) representa a mesma que (5.13) quando os ganhos σ e σ' são iguais.

Em [35], a distância cepestral de (5.13) é citada como uma forma mais simples para cálculo da *melmq*, devido à alta correlação que ocorre entre estas medidas.

5.3.3 Distância de Itakura

A distância de Itakura (DI) é um dos métodos mais utilizados para determinar distância entre conjuntos de CPL, segundo afirmado em [4]. Ela foi introduzida inicialmente na aplicação de reconhecimento da fala [46].

Para entender a DI, deve-se supor um modelo $H(z) = 1/A(z)$ tirado de um determinado segmento de sinal de fala $s(n) = \{s(0), s(1), \ldots, s(N-1)\}$. Se for realizada a filtragem inversa do segmento $s(n)$ pelo modelo, ou seja, se for passado o segmento $s(n)$ pelo filtro inverso não recursivo $A(z)$, tem-se na saída um sinal residual $\epsilon(n)$. Supondo agora que um outro modelo $\hat{H}(z) = 1/\hat{A}(z)$ seja formado pelos coeficientes $\{\hat{a}_1, \ldots, \hat{a}_p\}$, passando o segmento de sinal de fala $s(n)$ pelo filtro inverso $\hat{A}(z)$, obtém-se um outro sinal residual $\hat{\epsilon}(n)$. Partindo-se da premissa que os coeficientes a_i do modelo $H(z)$ correspondem aos coeficientes referência, ou seja, são os coeficientes realmente obtidos a partir de $s(n)$ através de uma análise LPC; e que os coeficientes $\hat{a}_i$ são os coeficientes teste, ou seja, não foram obtidos a partir de $s(n)$, tem-se que [4]

$$\hat{\varepsilon} = \sum_{n=0}^{N-1} \hat{\epsilon}^2(n) \quad \geq \quad \varepsilon = \sum_{n=0}^{N-1} \epsilon^2(n). \tag{5.15}$$

A expressão anterior indica que a energia ε do resíduo resultante da filtragem inversa de $s(n)$ pelo filtro $A(z)$, formado pelos coeficientes referência, é sempre menor ou igual que a energia $\hat{\varepsilon}$ do resíduo resultante da filtragem inversa do mesmo segmento pelo modelo $\hat{A}(z)$ formado pelos coeficientes de teste.

A DI é obtida pelo logaritmo da razão entre as energias destes resíduos, ou seja,

$$D_I = \ln \frac{\hat{\varepsilon}}{\varepsilon}. \tag{5.16}$$

Pode-se reparar que $D_I \geq 0$. Após algumas manipulações (veja [4] para detalhes), chega-se a

$$D_I(\alpha, \hat{\alpha}) = \ln \frac{\hat{\alpha}\mathbf{R}\hat{\alpha}}{\alpha\mathbf{R_s}\alpha}, \tag{5.17}$$

que em dB fica igual a

$$D_I(\alpha, \hat{\alpha}) = 10 \log \frac{\hat{\alpha}\mathbf{R}\hat{\alpha}}{\alpha\mathbf{R}\alpha} \qquad (dB), \tag{5.18}$$

onde α e $\hat{\alpha}$ são vetores obtidos a partir dos CPL de referência e de teste, respectivamente, através de

$$\alpha = \begin{pmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{pmatrix} \quad \text{e} \quad \hat{\alpha} = \begin{pmatrix} 1 \\ -\hat{a}_1 \\ -\hat{a}_2 \\ \vdots \\ -\hat{a}_p \end{pmatrix}. \tag{5.19}$$

A matriz de autocorrelação $\mathbf{R_s}$ do segmento $s(n)$ é da forma

$$\mathbf{R_s} = \begin{pmatrix} r_s(0) & r_s(1) & \cdots & r_s(N) \\ r_s(1) & r_s(0) & \cdots & r_s(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ r_s(N) & r_s(N-1) & \cdots & r_s(0) \end{pmatrix}, \tag{5.20}$$

onde

$$r_s(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} s(n)s(n-\tau), \tag{5.21}$$

que é uma estimativa polarizada da seqüência de autocorrelação para $s(n)$ [4].

Quando deseja-se determinar a diferença entre dois conjuntos de CPL provenientes de dois diferentes segmentos de sinal de fala $s(n)$ e $\hat{s}(n)$, surge um problema porque a DI não obedece à propriedade da simetria, ou seja, $D_I(\alpha, \hat{\alpha}) \neq D_I(\hat{\alpha}, \alpha)$. Para resolver isso, uma forma é obter uma média aritmética entre as duas medidas possíveis [4], ou seja,

$$\begin{aligned} D_I &= \frac{1}{2}[D_I(\alpha, \hat{\alpha}) + D_I(\hat{\alpha}, \alpha)] \\ &= \frac{1}{2}\left[10\log_{10}\left\{\frac{\hat{\alpha}\cdot\mathbf{R_s}\cdot\hat{\alpha}}{\alpha\cdot\mathbf{R_s}\cdot\alpha}\right\} + 10\log_{10}\left\{\frac{\alpha\cdot\mathbf{R_{\hat{s}}}\cdot\alpha}{\hat{\alpha}\cdot\mathbf{R_{\hat{s}}}\cdot\hat{\alpha}}\right\}\right] \quad (dB), \end{aligned} \tag{5.22}$$

onde $R_{\hat{s}}$ é a matriz de autocorrelação de curto termo para o segmento $\hat{s}(n)$.

A relação entre a DI e a *melmq* é altamente não linear, como mostrado em [35].

5.3.4 Distancia de Itakura-Saito

A DI representa uma interpretação mais simples a partir da fórmula original obtida por Itakura em [46], e que está mostrada a seguir:

$$D_{IS} = \frac{(\alpha_\mathbf{o} - \hat{\alpha})\mathbf{R}(\alpha_\mathbf{o} - \hat{\alpha})}{\alpha_\mathbf{o}\mathbf{R}\alpha_\mathbf{o}}. \tag{5.23}$$

Para chegar a tal expressão, deve-se interpretar que os possíveis CPL de um dado segmento $s(n)$ de sinal de fala são variáveis aleatórias conjuntamente gaussianas cuja média é dada pelo vetor α_o que contém os CPL ótimos para o segmento. A partir daí é realizado um teste de hipótese para se chegar à fórmula acima. Segundo [4] essa expresão ficou conhecida no campo de processamento de sinais de fala como distância de Itakura-Saito (DIS).

Para se utilizar a DIS deve-se supor que os coeficientes de referência representam os CPL ótimos, ou seja, $\alpha \approx \alpha_o$, o que não é verdade porque a técnica de análise de predição linear somente fornece uma estimativa dos CPL ótimos para um determinado segmento de um sinal de fala [4]. Em dB tem-se

$$D_{IS}(\alpha, \hat{\alpha}) = 10 \log \frac{(\alpha - \hat{\alpha})\mathbf{R_s}(\alpha - \hat{\alpha})}{\alpha \mathbf{R_s} \alpha} \quad (dB). \tag{5.24}$$

Tal como a DI, a distância de Itakura-Saito também não obedece à propriedade da simetria, ou seja, $D_{IS}(\alpha, \hat{\alpha}) \neq D_{IS}(\hat{\alpha}, \alpha)$. Isto é resolvido da mesma forma que no caso da DI, ou seja,

$$\begin{aligned} D_{IS} = & \quad \tfrac{1}{2}[D_{IS}(\alpha, \hat{\alpha}) + D_{IS}(\hat{\alpha}, \alpha)] \\ = & \quad \tfrac{1}{2}\left[\left\{10 \log \tfrac{(\alpha-\hat{\alpha})\mathbf{R_s}(\alpha-\hat{\alpha})}{\alpha \mathbf{R_s} \alpha}\right\} + \left\{10 \log \tfrac{(\hat{\alpha}-\alpha)\mathbf{R_{\hat{s}}}(\hat{\alpha}-\alpha)}{\hat{\alpha} \mathbf{R_{\hat{s}}} \hat{\alpha}}\right\}\right] \quad (dB). \end{aligned} \tag{5.25}$$

5.3.5 Distância entre CPL no formato LAR

Os logaritmos da razão das áreas (LAR), $\{l_1, \ldots, l_p\}$, são obtidos a partir dos coeficientes de reflexão $\{k_1, \ldots, k_p\}$, através de

$$l_i = \ln \frac{1 + k_i}{1 - k_i}. \tag{5.26}$$

Os coeficientes de reflexão, por sua vez são obtidos durante a recursão de Levinson-Durbin para determinação dos CPL $\{a_1, \ldots, a_p\}$ usando o método da autocorrelação [19].

Uma distância entre dois conjuntos de LAR de dois segmentos de sinal de fala $s(n)$ e $\hat{s}(n)$ é obtida ao simplesmente computar a distância euclidiana entre eles [4],

ou seja,

$$D_{LAR} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(l_i - \hat{l}_i)^2}, \tag{5.27}$$

que em dB fica

$$D_{LAR} = 10\log\left[\sqrt{\frac{1}{p}\sum_{i=1}^{p}(l_i - \hat{l}_i)^2}\right] \quad (dB), \tag{5.28}$$

onde l_i e $\hat{l}_i$ são respectivamente os LAR correspondentes a $s(n)$ e $\hat{s}(n)$, e p é o número de coeficientes.

5.4 Experimentos

5.4.1 Base de dados usada

O método proposto para analisar sinais de fala em sub-bandas foi aplicado em uma base de dados consistindo de 600 frases faladas em português, cada uma com duração de 2 a 3 segundos. As sentenças foram todas aquelas listadas em [31], que correspondem a frases foneticamente balanceadas para o idioma português falado na cidade do Rio de Janeiro. Todo o conjunto de frases foi pronunciado uma vez por cada um de três locutores, onde dois são do sexo masculino e um do sexo feminino.

Os sinais foram gravados em uma freqüência de amostragem igual a 44,1 kHz com 16 bits por amostra em dois canais (estéreo). Depois foram transformados para a freqüência de 8 kHz com 16 bits por amostra e reduzido o número de canais para um (mono). Neste último formato é que eles foram processados pelo método proposto.

5.4.2 Aplicação das medidas

Foram aplicadas a D_{EL} de (5.7) e a D'_{EL} de (5.8) usando $L = 512$; a D_C de (5.12) e a D'_C de (5.13) usando 32 coeficientes cepstrais ($p' = 32$); a D_I mostrada em (5.22); a D_{IS} em (5.25); e finalmente a D_{LAR} de (5.28).

As análises LPC foram todas de ordem 16, feitas pelo método da autocorrelação usando uma janela de Hamming centrada em cada segmento correspondente. Para as análises feitas com segmentos de 10 e 15 ms foram usadas janelas de 25 ms, enquanto que para analises com segmentos de 20, 25 e 30 ms foram usadas janelas de 30, 35 e 40 ms, respectivamente. Os tamanhos das janelas foram escolhidos de forma que elas fossem sempre 10 ms maiores que os segmentos correspondentes. Isto serve para compensar os efeitos da janela de Hamming que tende a minimizar as amostras de fronteira. A exceção ocorreu para o caso de 10 ms onde foi usada uma janela de 25 ms ao invés de 20 ms, para permitir uma melhor resolução da envoltória espectral [2].

5.4.3 Bancos de filtros usados

Para realizar o procedimento descrito na Seção 4.2 é interessante, para efeito de normalização, que o banco de filtros de análise seja complementar em potência, ou seja,

$$\sum_{i=1}^{N} |H_i(e^{j\theta})|^2 = 1, \qquad \forall\ \theta \in [0, \pi], \tag{5.29}$$

onde $|H_i(e^{j\theta})|$ é a resposta em amplitude do filtro na banda i, N é o número de bandas e θ é a freqüência normalizada, sendo π metade da freqüência de amostragem.

Foram utilizados dois bancos de filtros de $N = 4$ bandas nos experimentos: um uniforme e outro não-uniforme. O banco uniforme foi obtido a partir de um filtro protótipo $H_p(z)$, cujos coeficientes foram tirados de [47] e estão mostrados na Tabela 5.1. Os coeficientes dos filtros do banco foram obtidos a partir dos coeficientes deste filtro protótipo através da modulação por cossenos dada por [47]

$$h_k(n) = 2h_p(n) \cos\left[\frac{\pi}{N}\left(k + \frac{1}{2}\right)\left(n - \frac{M-1}{2}\right) + \frac{\pi}{4}(-1)^k\right], \tag{5.30}$$

onde $h_k(n)$ é a resposta ao impulso do filtro na banda k, $h_p(n)$ é a resposta ao impulso do filtro protótipo, e M é o comprimento das respostas ao impulso. A Figura 5.2 mostra as respostas em magnitude dos quatro filtros deste banco.

Tabela 5.1: Resposta ao impulso do filtro protótipo que gera o banco de filtros de análise uniforme.

Coeficientes de $H_p(z)$	
$h_p(0) = h_p(31) = 0,0001$	$h_p(8) = h_p(23) = -0,0098$
$h_p(1) = h_p(30) = -0,0006$	$h_p(9) = h_p(22) = -0,0011$
$h_p(2) = h_p(29) = 0,0013$	$h_p(10) = h_p(21) = 0,0174$
$h_p(3) = h_p(28) = -0,0004$	$h_p(11) = h_p(20) = 0,0447$
$h_p(4) = h_p(27) = -0,0022$	$h_p(12) = h_p(19) = 0,0772$
$h_p(5) = h_p(26) = -0,0043$	$h_p(13) = h_p(18) = 0,1101$
$h_p(6) = h_p(25) = -0,0096$	$h_p(14) = h_p(17) = 0,1368$
$h_p(7) = h_p(24) = -0,0117$	$h_p(15) = h_p(16) = 0,1518$

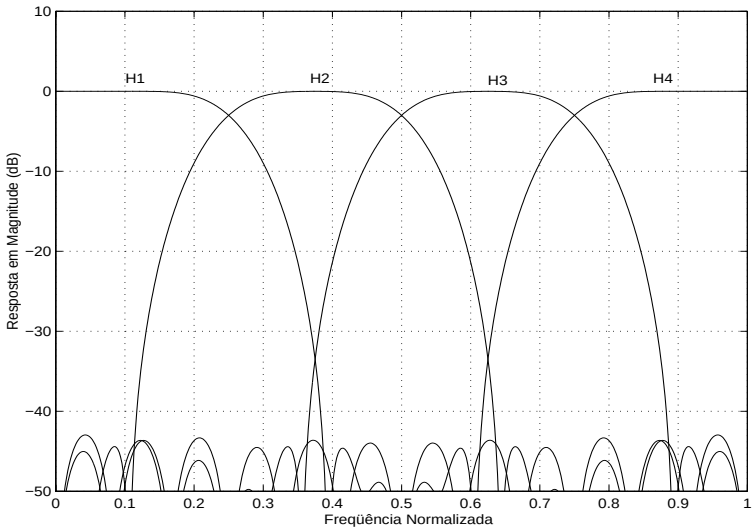


Figura 5.2: Respostas em magnitude dos filtros do banco uniforme.

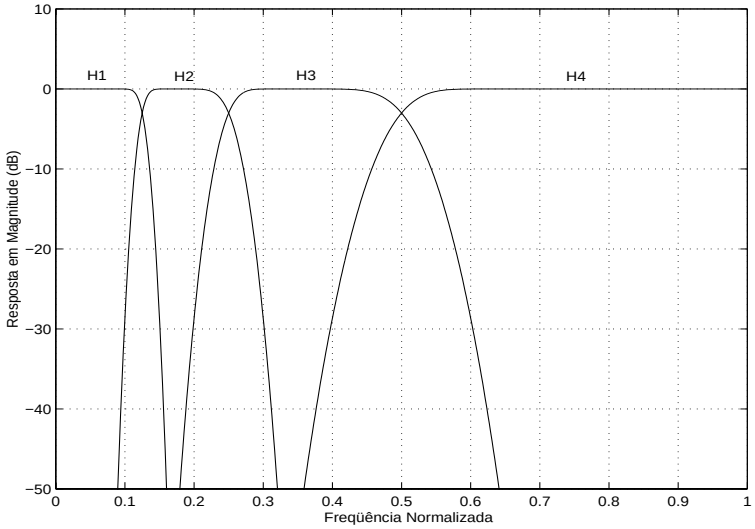


Figura 5.3: Respostas em magnitude dos filtros do banco não-uniforme.

O banco de filtros não-uniforme foi derivado a partir da wavelet de Daubechies [48] de ordem 90 [47]. A Figura 5.3 mostra as respostas em magnitude para os filtros deste segundo banco.

5.4.4 Resultados

A Tabela 5.2 mostra os GVE dados por (5.2) para cada sub-banda, para todas as medidas de distância usadas. Foram analisadas as 200 sentenças de um dos locutores masculinos. Os valores mostrados correspondem à média de todos os GVE. Foi usado o banco de filtros uniforme e segmentos de 10 ms. A Tabela 5.3 mostra o equivalente da Tabela 5.2 em VPR definidas em (5.3). Pode-se perceber ao observar esta última tabela que a maioria das medidas de distâncias conduziu a resultados similares, ou seja, indicaram que existe uma ligeira maior variação espectral na sub-banda de baixa freqüências, entre 0 e 1 KHz. A exceção ocorreu para a D_C e a D_{LAR}, que são exatamente as medidas que não levam em conta a energia dos espectros de curto termo dos sinais.

A Tabela 5.4 mostra as VPR considerando apenas a D_{EL} para os três locutores, dois do sexo masculino (M1 e M2) e um do sexo feminino (F1). Foram usadas as 200 sentenças de cada locutor. O banco de filtros e o tamanho dos segmentos

Tabela 5.2: GVE para cada medida de distância para 200 sentenças de um locutor masculino para segmentos de 10 ms, usando o banco de filtros uniforme.

	GVE			
	Sub-banda 1 0 - 1 kHz	Sub-banda 2 1 - 2 kHz	Sub-banda 3 2 - 3 kHz	Sub-banda 4 3 - 4 kHz
D_{EL}	5,55	4,66	4,39	4,38
D'_{EL}	6,87	5,97	5,63	5,78
D_C	3,40	3,56	3,36	3,32
D'_C	5,53	4,62	4,32	4,35
D_I	1,67	1,65	1,55	1,48
D_{IS}	-6,01	-4,58	-4,79	-5,11
D_{LAR}	-4,65	-5,40	-5,27	-4,51

Tabela 5.3: VPR para cada medida de distância para 200 sentenças de um locutor masculino para segmentos de 10 ms, usando o banco uniforme.

	VPR (%)			
	Sub-banda 1 0 - 1 kHz	Sub-banda 2 1 - 2 kHz	Sub-banda 3 2 - 3 kHz	Sub-banda 4 3 - 4 kHz
D_{EL}	29,26	24,54	23,13	23,07
D'_{EL}	28,32	24,62	23,22	23,84
D_C	24,91	26,10	24,66	24,33
D'_C	29,41	24,53	22,97	23,09
D_I	26,28	26,06	24,37	23,33
D_{IS}	29,30	22,36	23,40	24,94
D_{LAR}	23,43	27,26	26,58	22,73

Tabela 5.4: VPR para 200 sentenças de cada locutor, usando a D_{EL} e segmentos fixos de 10 ms com o banco uniforme.

	VPR (%)			
	Sub-banda 1 0 - 1 kHz	Sub-banda 2 1 - 2 kHz	Sub-banda 3 2 - 3 kHz	Sub-banda 4 3 - 4 kHz
H1	29,26	24,54	23,13	23,07
H2	29,14	24,00	22,49	24,37
M1	27,12	26,04	23,37	23,47

foram os mesmos do caso anterior. Pode-se notar que para os sinais masculinos os resultados foram praticamente os mesmos, onde a sub-banda 1 obteve consistentemente uma maior VPR. Para o locutor feminino houve um certo equilíbrio entre as sub-bandas 1 e 2.

A Tabela 5.5 mostra as VPR provenientes de testes com vários tamanhos de segmentos. Em cada situação foram usadas as 600 sentenças da base de dados, 200 de cada locutor. Pode-se observar que para segmentos de 10 ms a sub-banda 1 obteve maior VPR, enquanto que para os demais tamanhos os resultados indicam uma equivalência dos graus de estacionaridade.

A Tabela 5.6 mostra as VPR provenientes de testes também com vários tamanhos de segmentos, mas com o uso do banco não-uniforme. Em cada situação também foram usadas as 600 sentenças da base de dados, com 200 de cada locutor. Pode-se notar que os resultados diferem daqueles obtidos para o banco uniforme somente para segmentos de 10 ms, quando no caso do banco de filtros uniforme houve uma equivalência de variações espectrais.

Tabela 5.5: VPR para as 600 sentenças para vários tamanhos de segmentos, usando a D_{EL} e o banco uniforme.

	VPR (%)			
	Sub-banda 1 0 - 1 kHz	Sub-banda 2 1 - 2 kHz	Sub-banda 3 2 - 3 kHz	Sub-banda 4 3 - 4 kHz
10 ms	28,50	24,86	23,00	23,64
15 ms	26,69	25,05	24,02	24,24
20 ms	26,06	25,10	24,41	24,43
25 ms	25,65	25,12	24,64	24,59
30 ms	25,23	25,19	24,87	24,71

Tabela 5.6: VPR para as 600 sentenças para vários tamanhos de segmentos, usando a D_{EL} e o banco não-uniforme.

	VPR (%)			
	Sub-banda 1 0 - 500 Hz	Sub-banda 2 500 - 1000 Hz	Sub-banda 3 1000 - 2000 Hz	Sub-banda 4 2000 - 4000 Hz
10 ms	24,98	25,32	25,22	24,48
15 ms	24,31	25,25	25,40	25,04
20 ms	24,11	25,37	25,43	25,09
25 ms	24,14	25,39	25,32	25,15
30 ms	23,85	25,48	25,40	25,27

5.5 Conclusão

Neste capítulo foi apresentada uma análise de estacionaridade em sub-bandas para um conjunto de 600 sinais de fala gerados pela pronúncia das frases foneticamente balanceadas listadas em [31] e pronunciadas por três locutores, sendo dois do sexo masculino e um do sexo feminino, onde cada um foi responsável por 200 frases. O método empregado consiste em determinar as variações percentuais relativas entre as sub-bandas através de 7 medidas de distância entre envoltórias espectrais de curto termo. O número de sub-bandas analisadas foi no total de 4, com a faixa de freqüência de 0 a 4 kHz, e foram usados dois bancos de filtros: um uniforme obtido por modulação de cossenos e outro não-uniforme obtido a partir da wavelet de Daubechies. Os resultados experimentais mostraram que:

- cinco das sete medidas de distâncias empregadas conduziram a resultados semelhantes, conforme se vê na Tabela 5.3; as duas que diferiram não consideram a energia do espectro de curto termo do sinal de fala;
- as variações espectrais entre as sub-bandas são essencialmente independentes do locutor conforme se observa na Tabela 5.4, apesar do número de oradores analisados ter sido pequeno;
- a análise feita com o banco de filtros não-uniforme diferiu da análise com o banco uniforme somente para segmentos de 10 ms, onde para o primeiro caso a sub-banda de baixas freqüências foi a menos estacionária, enquanto que para o segundo caso as sub-bandas mostraram-se igualmente estacionárias;
- para segmentos acima de 10 ms, os sinais filtrados em sub-bandas foram igualmente estacionários para os dois bancos de filtros utilizados.

Em [49] é afirmado que a informação fonética crucial da fala está contida no ataque vocálico em um segmento de aproximadamente 10 ms. Talvez este fato seja a razão da diferença de estacionaridade que existe para segmentos deste tamanho quando feita a análise com o banco de filtros uniforme. Já para o caso da análise com o banco não-uniforme, o fato de haver maior estacionaridade para a faixa de 0 a 500 Hz

pode ser devido à alta presença de ruído de baixas freqüências existente nos sinais de fala, principalmente entre 0 e 200 Hz. Talvez este fato tenha mascarado a maior variação espectral que existe entre 0 e 1 kHz.

A análise em sub-bandas mostrada neste capítulo teve como objetivo procurar descobrir diferenças de estacionaridade para os sinais de fala ao longo do espectro de freqüência, característica esta com possíveis aplicações em codificação. A idéia é modificar a função de erro, usada para o cálculo dos CPL $\{a_i\}$ na análise LPC [19], através da decomposição deste sinal e a subseqüente utilização de diferentes resoluções. Em outras palavras, o erro efetivamente utilizado seria obtido por uma soma ponderada dos erros nas diferentes sub-bandas, onde os pesos seriam escolhidos de acordo com os graus de variações espectrais entre elas. Este procedimento tem sido aplicado com sucesso em funções de erros para algoritmos de processamento adaptativo de sinais baseados nos convencionais LMS (*least mean square*) e RLS (*recursive least squares*), como mostrado em [50, 51].

Capítulo 6

Conclusão

6.1 Contribuições do trabalho

Este trabalho apresentou o desenvolvimento de um sistema CELP para a codificação da fala que opera a uma taxa de 4,67 kbps. O sistema proposto utiliza quantização escalar dos parâmetros do filtro de síntese, dicionário adaptativo com atrasos fracionários, dicionário fixo esparso obtido a partir de amostras de ruído branco gaussiano, procedimentos de busca rápida e método seqüencial com otimização dos ganhos para a determinação da melhor excitação. Também foi apresentada uma análise de estacionaridade de sinais de fala em sub-bandas, com aplicações para codificação. A seguir são resumidas as contribuições de cada capítulo.

O Capítulo 2 forneceu uma introdução à área de codificação da fala. Foi abordada com detalhes a técnica CELP e características de alguns algoritmos que são padrões de codificação foram mostradas.

O Capítulo 3 apresentou quantizações dos parâmetros do filtro de síntese nas formas de LAR e LSF. Os quantizadores foram avaliados com base em uma medida de distorção espectral, tendo em vista a aplicação em um sistema CELP, onde foram consideradas avaliações com e sem os efeitos da interpolação entre sub-blocos. Os resultados alcançados fazem chegar às seguintes conclusões:

- a quantização individual de cada LSF mostrou-se superior à quantização dos LAR no quesito distorção espectral e inferior no quesito estabilidade do filtro

quantizado;

- a quantização diferencial das LSF mostrou-se bem melhor que a quantização individual porque obteve valores menores de distorção e eliminou o problema de instabilidade do filtro após a quantização.

O Capítulo 4 descreveu a implementação de um sistema CELP que opera a uma taxa de 4,67 kbps. O sistema foi avaliado através de medidas objetivas de qualidade e testes informais de escuta, sendo comparado ao DoD-CELP que opera a 4,8 kbps. Os resultados obtidos levam a concluir que de modo geral, o sistema proposto apresentou desempenho superior. Também foi mostrado um método de busca rápida no dicionário adaptativo muito parecido com o proposto em [40]. O sistema foi avaliado com e sem este procedimento, onde os resultados mostraram que:

- a avaliação subjetiva informal mostrou que a qualidade dos sinais processados pelo sistema com e sem o procedimento de busca rápida é indistinguível;

- o procedimento de busca rápida reduz entre 28 a 32% o tempo de codificação de um sinal de fala pelo sistema proposto.

O Capítulo 5 apresentou uma análise de estacionaridade de sinais de fala em sub-bandas. O método proposto foi aplicado em uma base de dados de 600 sinais gerados por um conjunto de 200 frases e 3 locutores diferentes, dentre os quais um do sexo feminino e dois do sexo masculino. Foram usadas 7 medidas de distância entre envoltórias de espectros de magnitude de curto termo e dois bancos de filtros: um uniforme e outro não-uniforme. Os resultados mostraram que:

- cinco das sete medidas empregadas conduziram a resultados semelhantes, onde as duas que diferiram das demais não consideraram a energia do espectro de curto termo;

- os graus de estacionaridade entre as sub-bandas são essencialmente independentes do locutor, apesar do número de oradores analisados ter sido pequeno;

- a análise feita com o banco de filtros não-uniforme diferiu da análise com o banco uniforme somente para segmentos de 10 ms, onde para o primeiro caso a sub-banda de baixas freqüências foi a menos estacionária, enquanto que para o segundo caso as sub-bandas mostraram graus equivalentes;

- para segmentos acima de 10 ms, as sub-bandas foram igualmente estacionárias para os dois bancos de filtros utilizados.

6.2 Propostas para trabalhos futuros

As seguintes propostas referem-se a uma possível continuação do presente trabalho:

- comparação dos vários métodos de quantização vetorial dos parâmetros do filtro de síntese $H(z)$ existentes atualmente;

- aproveitamento das características de estacionaridade em sub-bandas, desenvolvido neste trabalho, para a determinação de coeficientes de predição linear que ocasionem melhor qualidade;

- realizar um teste de MOS para o sistema, o que irá permitir uma melhor análise da qualidade, e uma comparação mais completa do mesmo com diversos outros sistemas apresentados na literatura;

- fazer o sistema operar em tempo real.

Referências Bibliográficas

[1] WOODARD, J. P., "Commonly speech codecs", 1995. http://www-mobile.ecs.soton.ac.uk/jason/speech_codecs/common_classes.html.

[2] RABINER, L. R., SCHAFER, R. W., *Digital Processing of Speech Signals*. Englewood Cliffs, N.J., Prentice-Hall, 1978.

[3] SCHROEDER, M. R., "Vocoders: analysis and synthesis of speech", *Proceedings of the IEEE*, v. 54, n. 5, May 1966.

[4] DELLER, J. R., PROAKIS, J. G., HANSEN, J. H. L., *Discrete-time Processing of Speech Signals*. New York, Macmillan, 1993.

[5] da SILVA, L. M., *Contribuições para a melhoria da codificação CELP a baixas taxas de bits*. Tese de D.Sc., PUC-RJ, Rio de Janeiro, RJ, Brasil, Fev. 1996.

[6] SPANIAS, A. S., "Speech coding: a tutorial review", *Proceedings of the IEEE*, v. 82, pp. 1541–1582, Oct. 1994.

[7] GERSHO, A., "Advances in speech and audio compression", *Proceedings of the IEEE*, v. 82, n. 6, pp. 900–918, Jun. 1994.

[8] SCHROEDER, M. R., ATAL, B. S., "Code-excited linear prediction (CELP): high-quality speech at very low bit rates". In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 937–940, 1985.

[9] KROON, P., DEPRETTERE, E. F., "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s", *IEEE Journal on Selected Areas in Communications*, v. 6, n. 2, Feb. 1988.

[10] KROON, P., SWAMINATHAN, K., "A high-quality multirate real-time CELP coder", *IEEE Journal on Selected Areas in Communications*, v. 10, n. 5, Jun. 1992.

[11] KLEIJN, W. B., KRASINSKI, D. J., KETCHUM, R. H., "Improved speech quality and efficient vector quantization in SELP". In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 155–158, 1988.

[12] KIM, H. K., "Adaptive encoding of fixed codebook in CELP coders". In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 149–152, 1998.

[13] KIM, H. K., LEE, M. S., LEE, H. S., "A 4 kbs adaptive fixed code-excited linear prediction speech coder". In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1999.

[14] LANGI, A., GRIEDER, W., KINSHER, W., "Fast CELP algorithm and implementation speech compression". In: *Proceedings of Digital Communications Conference*, 1994.

[15] ATAL, B. S., "Predictive coding of speech at low bit rates", *IEEE Transactions on Communications*, v. 30, n. 4, pp. 600–614, Apr. 1982.

[16] RAMACHANDRAN, R., KABAL, P., "Stability and performance analysis of pith filters in speech coders", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 35, n. 7, pp. 937–946, Jul. 1987.

[17] RAMACHANDRAN, R., KABAL, P., "Pitch prediction filters in speech coding", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 37, n. 4, pp. 467–478, Apr. 1989.

[18] MARKEL, J. D., GRAY, A. H., *Linear Prediction of Speech*. New York, Springer-Verlag, 1976.

[19] MAKHOUL, J., "Linear prediction: a tutorial review", *Proceedings of the IEEE*, v. 63, n. 4, pp. 561–579, Apr. 1975.

[20] ATAL, B. S., "Predictive coding of speech signals and subjective error criteria", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 27, n. 3, pp. 247–254, Jun. 1979.

[21] da SILVA, L. M., ALCAIM, A., "CELP with priority to critical segments". In: *Proceedings of EUSIPCO*, pp. 717–720, Sep. 1998.

[22] da SILVA, L. M., ALCAIM, A., "Multipulse stochastic codebook for CELP speech coding". In: *Proceedings of ICSP '97*, Seul, Corea, Aug. 1997.

[23] WOODARD, J. P., HANZO, L., "Improvements to the analysis-by-synthesis loop in CELP codecs". In: *Proceedings of the Radio Receivers and Associated Systems Conference*, pp. 114–118, Bath, UK, Sep. 1995.

[24] KROON, P., ATAL, B. S., "On the use the pitch predictors with high temporal resolution", *IEEE Transactions on Signal Processing*, pp. 733–735, 1991.

[25] GIBSON, J. D., BERGER, T., LOOKABAUGH, T., *et al.*, *Digital Compression for Multimedia*. Morgan Kaufmann, 1998.

[26] CHEN, J. H., COX, R. V., LIN, Y. C., *et al.*, "A low delay CELP coder for the CCITT 16 kb/s speech coding standard", *IEEE Journal on Selected Areas in Communications*, v. 10, n. 5, Jun. 1992.

[27] DALL'AGNOL, S. L. Q., *Quantização eficiente dos parâmetros do filtro em codificadores VSELP*. Tese de M.Sc., PUC-RJ, Rio de Janeiro, RJ, Brasil, Mar. 1993.

[28] VISWANATHAN, R., MARKHOUL, J., "Quantization properties of transmission parameters in linear predictive systems", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 23, pp. 309–321, Jun. 1976.

[29] KABAL, P., RAMACHANDRAN, R., "The computation of line spectral frequencies using Chebyshev polynomials", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 34, n. 6, pp. 1419–1426, Dec. 1986.

[30] ROTHWEILER, J., "On polynomial reduction in the computation of LSP frequencies", *IEEE Transactions on Speech and Audio Processing*, v. 7, n. 5, pp. 592–594, Sep. 1999.

[31] ALCAIM, A., SOLEMICZ, J. A., de MORAES, J. A., "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro", *Revista da Sociedade Brasileira de Telecomunicações*, v. 7, n. 1, Dez. 1992.

[32] TOHKURA, Y., ITAKURA, F., HASHIMOTO, S., "Spectral smoothing technique in PARCOR speech analysis-synthesis", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 26, n. 6, Dec. 1978.

[33] GERSHO, A., GRAY, R. M., *Vector Quantization and Signal Compression*. Boston, Kluwer Academic Publishers, 1992.

[34] SUGAMURA, N., FARVARDIN, N., "Quantizer design in LSP analysis-synthesis", *IEEE Journal on Selected Areas in Communications*, v. 6, n. 2, pp. 432–440, Feb. 1988.

[35] GRAY, A. H., MARKEL, J. D., "Distance measures for speech processing", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 24, pp. 380–391, Oct. 1976.

[36] da SILVA, L. M., ALCAIM, A., "Interpolation-based differential vector coding of speech LSF parameters". In: *Proceedings of GLOBECOM*, pp. 2049–2052, 1996.

[37] da SILVA, L. M., ALCAIM, A., "Sub-optimal quantization of line spectral frequencies". In: *Proceedings of ITS*, pp. 35–38, 1996.

[38] SALAMI, R., LAFLAMME, C., BASSETTE, B., *et al.*, "ITU-T G.729 annex A: reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data", *IEEE Communications Magazine*, , Sep. 1997.

[39] JACKSON, L. B., *Digital Filters and Signal Processing*. Kluwer Academic Publishers, 1996.

[40] da SILVA, L. M., ALCAIM, A., "A modified CELP model with computacionally efficient adaptive codebook search", *IEEE Signal Processing Letters*, v. 2, n. 3, pp. 44–45, Mar. 1995.

[41] RIBEIRO, C. B., Comunicação Pessoal.

[42] HANZO, L., BROOKS, F. C. A., WOODARD, J. P., *Voice compression and communications: principles and applications for fixed and wireless channels*. Book Proposed.

[43] KITAWAKI, N., NAGABUCHI, H., ITOH, K., "Objective quality evaluation for low-bit-rate speech coding systems", *IEEE Journal on Selected Areas in Communications*, v. 6, n. 2, pp. 242–248, Feb. 1988.

[44] THONET, G., VESIN, J. M., "Stationarity assessment with time-varying autoregressive modeling". In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1997.

[45] MAIA, R. da S., NETTO, S. L., RESENDE, F. G. V., Jr., "Subband stationarity analysis of speech signals". To be published at IEEE Symposium on Circuits and Systems (ISCAS) 2000.

[46] ITAKURA, F., "Minimum prediction residual principle applied to speech recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 23, pp. 66–72, Feb. 1975.

[47] STRANG, G., NGUYEN, T., *Wavelets and Filter Banks*. Wellesley,MA, Wellesley-Cambridge Press, 1996.

[48] DAUBECHIES, I., "Orthonormal bases of compactly supported wavelets", *Communications on Pure and Applied Mathematics*, v. XLI, pp. 909–996, 1988.

[49] FURUI, S., "On the role of spectral transition for speech perception", *Journal of Acoustics Society of America*, pp. 1016–1025, 1986.

[50] RESENDE, F. G., DINIZ, P. S. R., TOKUDA, K., *et al.*, "New adaptive algorithms based on multi-band decomposition of the error signal", *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, v. 45, n. 5, pp. 592–599, May 1998.

[51] RESENDE, F. G. V., Jr., *Multiband decomposition of error function in adaptive algorithms*. Ph.D. dissertation, Tokyo Institute of Technology, Tokio, Japan, 1996.

Apêndice A

Bases de dados utilizadas

Este apêndice aponta as sentenças utilizadas nos projetos e avaliações dos quantizadores do Capítulo 3, e dos quantizadores dos ganhos G_a e $\tilde{G}_f$ para o sistema CELP proposto no Capítulo 4.

Algumas sentenças usadas foram retiradas de [31], onde existem 20 listas, com cada uma delas contendo 10 frases foneticamente balanceadas. Cada sentença é indicada aqui através do número da lista onde ela se encontra, seguido pelo número da frase na lista, separados por um hífen. Por exemplo, "12-3" indica a terceira frase da décima segunda lista de frases pertencente ao considerado artigo. Cada frase é mostrada com a identificação do locutor que a pronunciou. Os locutores são identificados com a letra "M" ou "F", indicando respectivamente se trata-se do sexo masculino ou feminino, seguido por um número que indica a sua seqüência. Por exemplo, "M1" significa o primeiro locutor do sexo masculino, "M2" o segundo locutor do sexo masculino, "F1" o primeiro locutor do sexo feminino, e assim por diante. O título de cada tabela explica onde foi utilizada a base de dados.

Cinco frases de língua inglesa também foram usadas nos projetos dos quantizadores, sendo cada uma delas pronunciada por um locutor diferente, onde três foram do sexo masculino e dois do sexo feminino.

Tabela A.1: Parte das sentenças usadas para o projeto dos quantizadores dos parâmetros de $H(z)$ no Capítulo 3, com a indicação dos locutores.

Sentença	Locutor	Sentença	Locutor	Sentença	Locutor
3-8	M1	4-7	M1	5-7	M1
6-7	M1	7-7	M1	8-7	M1
3-1	M2	4-1	M2	5-1	M2
6-1	M2	7-1	M2	8-2	M2
3-2	F1	4-5	F1	5-2	F1
6-2	F1	7-2	F1	8-3	F1

Tabela A.2: Sentenças usadas para a avaliação dos quantizadores dos parâmetros de $H(z)$ no Capítulo 3, com a indicação dos locutores.

Sentença	Locutores	Sentença	Locutores
3-1	M1 e F1	3-2	M1 e F1
3-3	M2 e F2	3-4	M2 e F2
3-5	M3 e F3	3-6	M3 e F3
3-7	M4 e F4	3-8	M4 e F4
3-9	M5 e F5	3-10	M5 e F5

Tabela A.3: Parte das sentenças usadas para o projeto dos quantizadores dos ganhos G_a e $\tilde{G}_f$ no Capítulo 4, com a indicação dos locutores.

Sentença	Locutor	Sentença	Locutor	Sentença	Locutor
9-1	M1	10-1	M1	11-3	M1
9-2	M2	10-2	M2	11-4	M2
2-1	M3	2-5	M3	2-6	M3
9-4	F1	10-3	F1	11-6	F1

Apêndice B

Partições e dicionários dos quantizadores

Este apêndice lista os valores das partições e dicionários para o quantizador *QDLSF-32*, projetado no Captulo 3, e o quantizador para os ganhos G_a e $\tilde{G}_f$ no Capítulo 4. As partições indicam os níveis de decisão, enquanto que os dicionários correspondem aos valores de saída [33]. O título de cada tabela indica a qual quantizador pertencem as partições e dicionários listados.

Tabela B.1: Intervalos de decisões (ID) e valores de saída (VS) para os quantizadores não-uniformes de 4 bits para os coeficientes Δw_1 e Δw_2.

	Δw_1		Δw_2	
Índice	ID	VS	ID	VS
0	$(-\infty; 0.0865)$	0.0724	$(-\infty; 0.0518)$	0.0374
1	$[0.0865; 0.1123)$	0.1007	$[0.0518; 0.0793)$	0.0662
2	$[0.1123; 0.1339)$	0.1240	$[0.0793; 0.1068)$	0.0923
3	$[0.1339; 0.1519)$	0.1439	$[0.1068; 0.1360)$	0.1213
4	$[0.1519; 0.1700)$	0.1600	$[0.1360; 0.1655)$	0.1506
5	$[0.1700; 0.1903)$	0.1800	$[0.1655; 0.1958)$	0.1803
6	$[0.1903; 0.2107)$	0.2006	$[0.1958; 0.2246)$	0.2111
7	$[0.2107; 0.2317)$	0.2209	$[0.2246; 0.2522)$	0.2381
8	$[0.2317; 0.2539)$	0.2425	$[0.2522; 0.2813)$	0.2663
9	$[0.2539; 0.2793)$	0.2654	$[0.2813; 0.3111)$	0.2962
10	$[0.2793; 0.3043)$	0.2932	$[0.3111; 0.3424)$	0.3260
11	$[0.3043; 0.3292)$	0.3154	$[0.3424; 0.3745)$	0.3588
12	$[0.3292; 0.3552)$	0.3430	$[0.3745; 0.4115)$	0.3901
13	$[0.3552; 0.3802)$	0.3675	$[0.4115; 0.4520)$	0.4329
14	$[0.3802; 0.4185)$	0.3929	$[0.4520; 0.5576)$	0.4711
15	$[0.4185; +\infty)$	0.4441	$[0.5576; +\infty)$	0.6443

Tabela B.2: Intervalos de decisões (ID) e valores de saída (VS) para os quantizadores não-uniformes de 3 bits para os coeficientes $\Delta w_3, \ldots, \Delta w_6$.

	Δw_3		Δw_4		Δw_5		Δw_6	
Índice	ID	VS	ID	VS	ID	VS	ID	VS
0	$(-\infty;$ $0,1135)$	$0,0882$	$(-\infty;$ $0,1606)$	$0,1241$	$(-\infty;$ $0,1608)$	$0,1118$	$(-\infty;$ $0,1567)$	$0,1129$
1	$[0,1135;$ $0,1664)$	$0,1389$	$[0,1606;$ $0,2315)$	$0,1972$	$[0,1608;$ $0,2508)$	$0,2097$	$[0,1567;$ $0,2382)$	$0,2005$
2	$[0,1664;$ $0,2237)$	$0,1939$	$[0,2315;$ $0,2970)$	$0,2658$	$[0,2508;$ $0,3312)$	$0,2921$	$[0,2382;$ $0,3109)$	$0,2758$
3	$[0,2237;$ $0,2820)$	$0,2535$	$[0,2970;$ $0,3629)$	$0,3283$	$[0,3312;$ $0,4166)$	$0,3705$	$[0,3109;$ $0,3893)$	$0,3459$
4	$[0,2820;$ $0,3451)$	$0,3105$	$[0,3629;$ $0,4426)$	$0,3976$	$[0,4166;$ $0,5265)$	$0,4628$	$[0,3893;$ $0,4818)$	$0,4327$
5	$[0,3451;$ $0,4238)$	$0,3797$	$[0,4426;$ $0,5394)$	$0,4877$	$[0,5265;$ $0,6509)$	$0,5903$	$[0,4818;$ $0,6060)$	$0,5308$
6	$[0,4238;$ $0,5214)$	$0,4680$	$[0,5394;$ $0,6706)$	$0,5910$	$[0,6509;$ $0,7978)$	$0,7115$	$[0,6060;$ $0,7861)$	$0,6812$
7	$[0,5214;$ $+\infty)$	$0,5747$	$[0,6706;$ $+\infty)$	$0,7500$	$[0,7978;$ $+\infty)$	$0,8840$	$[0,7861;$ $+\infty)$	$0,8909$

Tabela B.3: Intervalos de decisões (ID) e valores de saída (VS) para os quantizadores não-uniformes de 3 bits para os coeficientes $\Delta w_7, \ldots, \Delta w_{10}$.

	Δw_7		Δw_8		Δw_9		Δw_{10}	
Índice	ID	VS	ID	VS	ID	VS	ID	VS
0	$(-\infty;$ $0,1757)$	$0,1345$	$(-\infty;$ $0,1374)$	$0,1055$	$(-\infty;$ $0,1748)$	$0,1372$	$(-\infty;$ $0,1494)$	$0,1180$
1	$[0,1757;$ $0,2526)$	$0,2168$	$[0,1374;$ $0,1982)$	$0,1693$	$[0,1748;$ $0,2430)$	$0,2123$	$[0,1494;$ $0,2064)$	$0,1808$
2	$[0,2526;$ $0,3240)$	$0,2884$	$[0,1982;$ $0,2554)$	$0,2270$	$[0,2430;$ $0,3040)$	$0,2736$	$[0,2064;$ $0,2558)$	$0,2320$
3	$[0,3240;$ $0,3970)$	$0,3597$	$[0,2554;$ $0,3118)$	$0,2837$	$[0,3040;$ $0,3643)$	$0,3344$	$[0,2558;$ $0,3050)$	$0,2795$
4	$[0,3970;$ $0,4785)$	$0,4344$	$[0,3118;$ $0,3694)$	$0,3399$	$[0,3643;$ $0,4297)$	$0,3943$	$[0,3050;$ $0,3556)$	$0,3304$
5	$[0,4785;$ $0,5778)$	$0,5227$	$[0,3694;$ $0,4405)$	$0,3988$	$[0,4297;$ $0,5126)$	$0,4650$	$[0,3556;$ $0,4121)$	$0,3807$
6	$[0,5778;$ $0,7202)$	$0,6329$	$[0,4405;$ $0,5616)$	$0,4821$	$[0,5126;$ $0,6261)$	$0,5601$	$[0,4121;$ $0,4854)$	$0,4436$
7	$[0,7202;$ $+\infty)$	$0,8076$	$[0,5616;$ $+\infty)$	$0,6411$	$[0,6261;$ $+\infty)$	$0,6921$	$[0,4854;$ $+\infty)$	$0,5273$

Tabela B.4: Intervalos de decisões (ID) e valores de saída (VS) para o quantizador não-uniforme de 4 bits para o ganho G_a.

Índice	ID	VS	índice	ID	VS
0	$(-\infty; 0,1662)$	0,0767	1	$[0,1662; 0,3196)$	0,2556
2	$[0,3196; 0,4375)$	0,3836	3	$[0,4375; 0,5394)$	0,4913
4	$[0,5394; 0,6368)$	0,5874	5	$[0,6368; 0,7338)$	0,6862
6	$[0,7338; 0,8243)$	0,7813	7	$[0,8243; 0,9077)$	0,8672
8	$[0,9077; 0,9888)$	0,9482	9	$[0,9888; 1,0762)$	1,0295
10	$[1,0762; 1,1798)$	1,1230	11	$[1,1798; 1,3063)$	1,2366
12	$[1,3063; 1,4518)$	1,3760	13	$[1,4518; 1,6174)$	1,5275
14	$[1,6174; 1,8083)$	1,7073	15	$[1,8083; +\infty)$	1,9093

Tabela B.5: Intervalos de decisões (ID) e valores de saída (VS) para o quantizador não-uniforme de 5 bits para o ganho $\tilde{G}_f$.

Índice	ID	VS	índice	ID	VS
0	$(-\infty; -0,0453)$	-0,0469	1	$[-0,0453; -0,0422)$	-0,0436
2	$[-0,0422; -0,0395)$	-0,0408	3	$[-0,0395; -0,0365)$	-0,0382
4	$[-0,0365; -0,0331)$	-0,0348	5	$[-0,0331; -0,0298)$	-0,0313
6	$[-0,0298; -0,0264)$	-0,0281	7	$[-0,0264; -0,0229)$	-0,0247
8	$[-0,0229; -0,0190)$	-0,0210	9	$[-0,0190; -0,0152)$	-0,0170
10	$[-0,0152; -0,0116)$	-0,0133	11	$[-0,0116; -0,0085)$	-0,0099
12	$[-0,0085; -0,0057)$	-0,0070	13	$[-0,0057; -0,0034)$	-0,0045
14	$[-0,0034; -0,0015)$	-0,0024	15	$[-0,0015; 0,0001)$	-0,0006
16	$[0,0001; 0,0019)$	0,0009	17	$[0,0019; 0,0044)$	0,0030
18	$[0,0044; 0,0072)$	0,0057	19	$[0,0072; 0,0104)$	0,0087
20	$[0,0104; 0,0138)$	0,0121	21	$[0,0138; 0,0173)$	0,0156
22	$[0,0173; 0,0207)$	0,0191	23	$[0,0207; 0,0238)$	0,0223
24	$[0,0238; 0,0270)$	0,0252	25	$[0,0270; 0,0301)$	0,0287
26	$[0,0301; 0,0332)$	0,0315	27	$[0,0332; 0,0363)$	0,0350
28	$[0,0363; 0,0398)$	0,0376	29	$[0,0398; 0,0437)$	0,0420
30	$[0,0398; 0,0469)$	0,0455	31	$[0,0469; \infty)$	0,0483