UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

ESCOLA POLITÉCNICA

DEPARTAMENTO DE ELETRÔNICA E DE COMPUTAÇÃO

ANÁLISE PRELIMINAR DA ROBUSTEZ DO CODIFICADOR DE VOZ CELP

Autora: Natasha da Rocha Moura

Orientador: Prof. Sérgio Lima Netto, Ph. D.

Examinador: Prof. Gelson Vieira Mendonça, Ph. D.

Examinador: Filipe Castello da Costa Beltrão Diniz, M. Sc.

DEL
Dezembro de 2005

Aos meus pais, João Pedro e Tânia Regina.

Agradecimentos

Agradeço:

- Ao professor Sérgio Lima Netto pela orientação acadêmica ao longo do curso;
- Ao professor Gelson Vieira Mendonça e a Filipe Castello da Costa Beltrão Diniz pela participação na banca avaliadora;
- A todos os professores do Departamento de Engenharia Eletrônica e de Computação da UFRJ que contribuíram para minha formação acadêmica e me incentivaram ao longo da graduação;
- Aos colaboradores que participaram da análise subjetiva realizada neste trabalho;
- Aos amigos do curso de Engenharia Eletrônica e de Computação da UFRJ, por sua amizade e companheirismo;
- Aos meus pais pelo apoio que sempre me ofereceram;
- Ao Eduardo Brasil e ao Rodrigo Lamas, pessoas que sempre estiveram ao meu lado durante a graduação.

Natasha da Rocha Moura

Resumo

Neste trabalho, é realizada uma análise bit-a-bit para avaliar a relevância de cada bit do *bitstream* sobre a qualidade e inteligibilidade do sinal de voz codificado. Como todo o trabalho foi desenvolvido sobre a implementação CELP proposta por [**1,2**], detalhes sobre esta são apresentados.

A análise dos resultados foi realizada em duas etapas. Na primeira, foram utilizadas as medidas objetivas: Razão Sinal-Ruído Segmentada Perceptual, Distância de Itakura, Distância Cepstral e o algoritmo PESQ (*Perceptual Evaluation of Speech Quality*). Na segunda, foi realizada uma análise subjetiva através do MOS (*Mean Opinion Score*).

Ao final deste trabalho verificou-se que foi possível classificar o *bitstream* em grupos de bits. Sendo possível identificar que cada um destes apresenta uma "importância" diferente no processo de codificação/ decodificação do sinal de voz quando se leva em consideração a qualidade e a inteligibilidade do sinal codificado.

As conclusões aqui alcançadas permitem explorar a questão da segurança dos codificadores de voz. Ao serem identificados os bits que mais contribuem com informações sobre a mensagem transmitida, torna-se possível aplicar algoritmos de segurança diretamente sobre eles e, assim, impedir que pessoas sem permissão possam decodificar a informação. Embora os resultados obtidos sejam aplicáveis a uma implementação específica de codificador, toda a metodologia de análise utilizada pode ser adaptada para a avaliação da robustez de outros codificadores.

Índice

Capítulo 1

Introdução

1.1 Proposta do Trabalho

O objetivo deste trabalho é realizar um estudo sobre a robustez de codificadores de voz, possibilitando assim, a proposição de soluções para reduzir suas fragilidades. Este estudo será feito através da análise da importância de cada bit do *bitstream* sobre a qualidade e inteligibilidade do sinal codificado. Assim, será possível classificar o *bitstream* em grupos que "contribuem" com diferentes níveis de informação. Procuramos com isto identificar aqueles bits que carregam a informação de inteligibilidade do sinal de voz.

1.2 Motivação

Atualmente, a maior parte dos sistemas de informação utiliza a tecnologia digital. Assim, a transformação dos sinais analógicos de áudio em sinais digitais é algo rotineiramente utilizado na telefonia, nos equipamentos de som, nos sistemas de televisão e nas comunicações VoIP. O advento da digitalização facilitou e ampliou a circulação da informação de uma forma geral. Paralelamente, aumentou a preocupação com a segurança da informação transmitida.

A segurança da informação é caracterizada por um conjunto de medidas tomadas para preservar: a confidencialidade, a integridade e a disponibilidade da informação. A confidencialidade consiste em garantir que a informação só está acessível a que tem autorização. A integridade consiste em garantir a acurácia e preservação da informação e dos métodos de processamento. A disponibilidade consiste em proteger os serviços prestados pelo sistema, de forma que não sejam degradados e tornem-se indisponíveis, garantindo que usuários autorizados tenham acesso à informação sempre que desejado.

O foco deste trabalho é em sistemas de comunicação que transmitem sinais de voz. A substituição da tecnologia analógica pela digital já representou um aumento na segurança dos sistemas de transmissão de voz, uma vez que dificulta a realização de "escutas clandestinas". Nos sistemas analógicos, o sinal de voz é, simplesmente, transmitido através de ondas de rádio. Já os sistemas digitais realizam uma codificação do som para uma representação binária, sendo então, transmitido por algum meio, por exemplo, por ondas de rádio, e depois decodificado no receptor. Dessa forma, além de aumentar a qualidade do sinal transmitido, por reduzir ruídos do meio, a digitalização também aumenta a segurança. Entretanto, as técnicas para burlar esta relativa segurança também se desenvolveram. E, também, são utilizados novos meios para transmissão da informação, como, por exemplo, a Internet, que trazem novas fragilidades ao sistema.

Neste trabalho, será estudada a robustez do codificador CELP. Para tal, será utilizada uma implementação desenvolvida por [**1,2**]. Neste codificador, o sinal de voz do emissor é representado por várias seqüências de bits, os *bitstreams*. Cada seqüência é transmitida através do meio até atingir o receptor, onde é então decodificada

para recompor o sinal original. No entanto, durante o processo de transmissão, o *bitstream* pode ser interceptado por uma pessoa mal intencionada. Essa pode decodificar os pacotes e, assim, ter acesso às informações. Os maiores alvos desses ataques são as empresas, já que é por onde trafegam informações comercialmente úteis. Dessa forma, sendo possível identificar os bits do *bitstream* que "carregam" a maior parte da informação, é possível protegê-los contra ataques e dessa forma, fortalecer a segurança do sistema de codificação/ decodificação de voz.

1.3 Organização do Trabalho

Para realizarmos este estudo, este projeto foi organizado da seguinte forma:

No Capítulo 2, são abordados aspectos gerais da codificação do sinal de voz, sendo apresentados conceitos sobre a modelagem da mesma. São também vistas as principais características dos diferentes tipos de codificadores de voz, bem como, seus padrões mais conhecidos. É dado maior destaque para o codificador CELP, já que toda a análise, em termos de robustez, é realizada a partir dele.

Como este trabalho realiza uma análise sobre a robustez do codificador implementado em [**1**,**2**], é necessário conhecer este codificador. Assim, no Capítulo 3, são apresentados detalhes específicos desta implementação. Os resultados desta análise estarão diretamente relacionados às características da implementação CELP (*Code-Excited Linear Predictive*). No entanto, a mesma metodologia pode ser utilizada para a análise de outro codificador.

Foram utilizados dois tipos de análise de dados, a objetiva e a subjetiva.

No capítulo 4, é apresentada em detalhes a metodologia para realização do procedimento experimental. Em seguida, é realizada a análise objetiva para verificar a qualidade do sinal reconstituído. Para isto, foram utilizadas as medidas: Razão Sinal-Ruído Segmentada Perceptual, Distância de Itakura, Distância Cepstral e o algoritmo PESQ (*Perceptual Evaluation of Speech Quality*). Ao fim do capítulo são também apresentados os resultados encontrados.

No Capítulo 5, dá-se continuidade à análise de dados. Sendo que, neste capítulo, é apresentada a metodologia para realização da análise subjetiva do sinal. Esta análise é realizada através do MOS (*Mean Opinion Score*) e tem como principal objetivo avaliar a importância de cada bit sobre a inteligibilidade e naturalidade do sinal de voz codificado. No final do capítulo, são apresentados os novos resultados, além de ser feita uma discussão sobre os mesmos.

No Capítulo 6, é realizada a conclusão sobre todo o trabalho desenvolvido e propõem-se idéias para trabalhos futuros.

Capítulo 2

Codificadores de Voz

2.1 Introdução

Neste capítulo, é apresentada uma visão geral sobre a codificação da voz. São abordados conceitos sobre a geração e a modelagem da mesma. Também, faz-se um apanhado sobre os tipos de codificadores e as técnicas mais conhecidas. Como o foco deste trabalho é a codificação CELP, é dado maior destaque a ela.

A Seção 2.2 aborda importantes características do sinal de voz que são exploradas pela codificação; a Seção 2.3 descreve, de forma geral, os diferentes tipos de codificadores; na Seção 2.4, é detalhada a técnica de codificação CELP e são apresentados alguns padrões existentes; já na Seção 2.5, é feita uma breve conclusão.

2.2 Propriedades da Voz

Na produção da voz, ar dos pulmões é impulsionado através da traquéia. Parte desse fluxo de ar segue através do trato vocal, o qual se estende da abertura das cordas vocais até a boca. Outra parte deste fluxo segue através da cavidade nasal. O trato vocal apresenta algumas características de ressonância, que podem ser alteradas devido à variação no seu formato, por exemplo, através da movimentação da língua. Tais características associadas a vibrações das cordas vocais modulam o ar, dando origem a diferentes sons. De forma geral, os sons podem ser classificados em duas categorias, vozeados e não-vozeados, quando consideradas diferenças perceptuais e espectrais. Na Figura 2.1, pode ser observado o aparelho fonador humano.

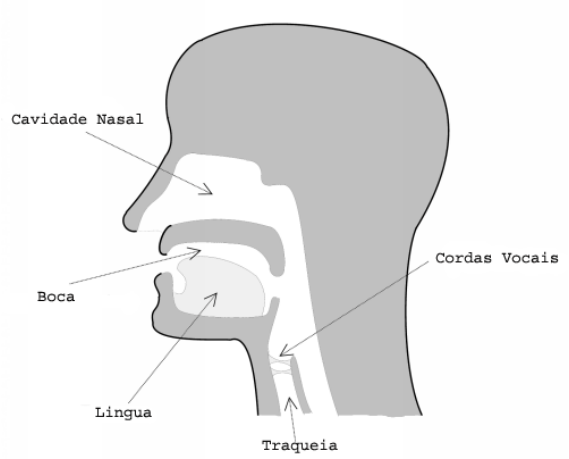


Figura 2. 1: Aparelho fonador humano conforme visto em [3].

O som vozeado é formado quando o fluxo de ar dos pulmões pressiona as cordas vocais, forçando-as a abrir e fechar, gerando uma excitação periódica, ou quase, no trato vocal. A freqüência com que ocorre a abertura e o fechamento das cordas é que determina a freqüência fundamental da excitação. Essa pode assumir valores que variam de 50 a 400 Hz, dependendo do tamanho das cordas vocais, o que está diretamente relacionado à idade e ao sexo da pessoa. Já a freqüência fundamental percebida pelo nosso cérebro é denominada *pitch*. A ação do trato vocal sobre os pulsos periódicos irá, finalmente, modificar a distribuição de energia no espectro através da introdução de ressonâncias, chamadas formantes.

Como o período de *pitch* e o formato do trato vocal se alteram ao longo do tempo, não se pode dizer que o som vozeado seja verdadeiramente periódico. Suas características espectrais e estatísticas variam ao longo do tempo. Porém, pode-se dizer que esse sinal é aproximadamente estacionário quando é considerado um trecho de 10 a 30 ms. Dessa forma, para explorar essa característica do sinal da fala, a análise do mesmo deve ser feita em intervalos periódicos de pequena duração. Na Figura 2.2, pode-se observar o espectro de potência de um trecho de sinal vozeado, no qual se evidenciam os harmônicos espaçados em um período de *pitch* e a concentração de energia em baixas freqüências.

Os sons não-vozeados, chamados fricativos, não são gerados pela vibração das cordas vocais e, portanto, não apresentam a mesma periodicidade encontrada na estrutura dos sinais vozeados. As fricativas são formadas quando o ar dos pulmões é forçado através das cordas vocais abertas e de uma "brecha" do trato vocal, gerando um som semelhante ao ruído. No domínio do tempo, esses sons perdem periodicidade e seu espectro de potência é aproximadamente "achatado", não apresentando tão claramente os picos de ressonância encontrados em sinais vozeados (Figura 2.2).

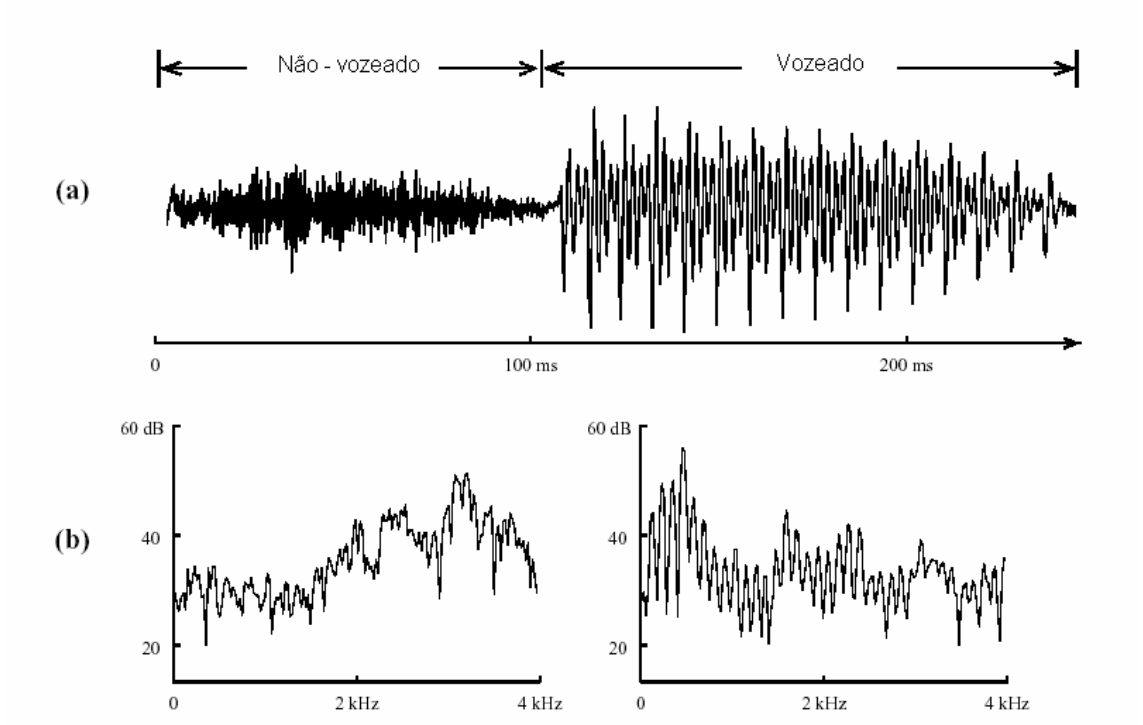


Figura 2. 2: Exemplos de (a) sons não-vozeados e vozeados e seus (b) respectivos espectros de potência.

Para o sucesso de qualquer sistema de processamento em que a voz é o principal meio de transmissão da informação é fundamental a representação da mesma no meio digital. Isto é denominado digitalização e envolve as seguintes etapas: amostragem, que consiste em tomar amostras do sinal periodicamente; quantização, que fixa o valor representado pelos bits que codificam cada uma das amostras; codificação, que determina um código com número de bits que pode ser variável para representar cada uma das amostras obtidas.

2.3 Codificadores

Os codificadores de voz são, normalmente, divididos em três classes: codificadores de forma de onda, codificadores paramétricos e codificadores híbridos, sendo que este último combina características dos dois primeiros. Cada uma das três classes possui um comprometimento diferente com relação à qualidade da voz e com a taxa de bits requeridos como visto na Figura 2.3.

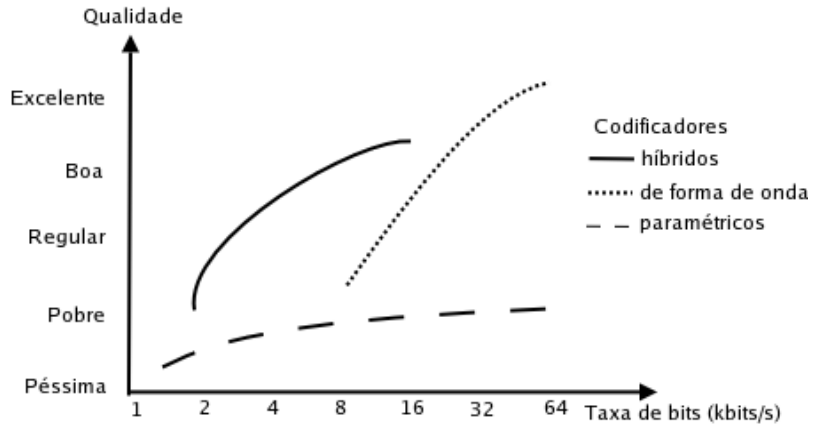


Figura 2. 3: Comportamento taxa de bits x qualidade dos diferentes tipos de codificadores de voz.

A codificação por forma de onda tenta reproduzir o sinal de voz analógico, amostra por amostra, com base nas suas características temporais e espectrais. Esse método é, em geral, de baixa complexidade e possibilita a produção de um sinal de voz com alta qualidade uma vez que a curva do sinal de voz obtida no receptor é uma cópia fiel da curva do sinal de voz original. No entanto, para taxas de transmissão inferiores a 16 kbits/s a qualidade do sinal reconstituído degrada-se rapidamente. O codificador de forma de onda mais simples é o PCM (*Pulse Code Modulation*), seguido pelo DPCM (*Differential Pulse Code Modulation*) e pelo ADPCM (*Adaptive Differential Pulse Code Modulation*).

A codificação paramétrica utiliza um modelo que descreve a forma como o sinal de voz original é gerado. A informação efetivamente transmitida é constituída pelos parâmetros obtidos desse modelo, que são atualizados periodicamente e determinados em intervalos periódicos denominados quadros, onde o sinal de voz pode ser considerado estacionário. Essa codificação gera um sinal de voz no receptor com baixa qualidade, soando de forma sintética, além de apresentar atraso e complexidade elevados. No entanto, opera com baixas taxas de transmissão, normalmente inferiores a 4 kbits/s. O codificador paramétrico mais conhecido é o LPC (*Linear Predictive Coding*).

A codificação híbrida une as boas características dos codificadores de forma de onda e paramétricos. Por apresentar um compromisso razoável entre a qualidade da voz reconstituída e a eficiência na codificação, esta é a codificação mais usada atualmente em sistemas de telefonia digital. O codificador híbrido mais conhecido é o CELP (*Code-Excited Linear Predictive*), o qual será abordado com maior detalhamento mais adiante neste trabalho.

2.3.1 Família PCM

Na codificação PCM, o sinal de voz original é digitalizado no tempo e na amplitude sendo desconsideradas informações redundantes de amostras adjacentes. Suponhamos que um sinal é amostrado a uma freqüência de 8000 Hz. Em seguida, é quantizado, sendo que a cada amostra é associado um 1 de 256 possíveis valores. Cada uma dessas amostras quantizadas é então codificada, sendo expressa por um código binário de 8 bits. Este código binário de 8 bits é chamado de palavra PCM. Dessa forma, se 8000 palavras PCM são geradas por segundo, o *bit stream* será de 8 x 8000 = 64000 bits/s em um *link* digital. A ITU-T (*Iternational Telecommunications Union, Telecommunications Standardizantion Sector*) chama a este tipo de codificação de voz de "64 kbits/s PCM", sendo este considerado o padrão de referência para codificação de voz em redes telefônicas. Sua recomendação para o PCM, encontrada em [**4**], é chamada G.711 que regulamenta que a quantização pode ser do tipo *μ-law* ou *A-law*, com 8 bits para cada amostra, resultando em uma taxa de 64 kbits/s.

Novos métodos de codificação, com a capacidade de dobrar ou quadruplicar a capacidade de transmissão de voz, foram desenvolvidos. Um deles é o DPCM. Como um sinal amostrado, de acordo com o teorema da amostragem, apresenta elevada correlação entre amostras adjacentes, pode-se codificar a diferença entre estas amostras utilizando um menor número de bits e dessa forma obter um elevado ganho em termos de banda. Uma desvantagem do DPCM é que se o sinal de voz original apresentar muitas variações entre amostras adjacentes, torna-se impossível a representação do mesmo com menos de 8 bits.

O ADPCM combina o DPCM com técnicas de codificação e predição adaptativas e assim possibilita a redução da taxa de bits de 64 kbits/s para 32 kbits/s mantendo a fidelidade e qualidade do sinal de voz reconstituído. A recomendação da ITU-T para o ADPCM é chamada G.726 e pode ser encontrada em [**4**].

2.3.2 Codificação Paramétrica LPC

Este codificador consegue operar a taxas baixas, enquanto preserva a qualidade do sinal, através da exploração de redundâncias do sinal de voz e de limitações perceptuais do ouvido humano. As redundâncias ocorrem devido a: (i) em geral o espectro do sinal de voz se altera relativamente devagar permanecendo estável ao longo de intervalos de 10 a 30 ms; (ii) sucessivos períodos de *pitch* são, geralmente, similares em trechos vozeados; (iii) a envoltória espectral é relativamente suave, com a maior parte da energia concentrada em baixas freqüências. Já as limitações perceptuais estão relacionadas ao fato de o ouvido humano ser, nos sinais de fala, insensível à fase e mais sensível a baixas freqüências do que a altas.

A redundância no sinal de voz leva à conclusão de que suas amostras são correlacionadas. A envoltória espectral corresponde às correlações de curto-termo (*short-term correlations*) e a estrutura de harmônicos

corresponde às correlações de longo-termo (*long-term correlations*). Essas correlações podem ser exploradas através da técnica de predição linear para resultar em um codificador com baixa taxa de bits.

Para a análise deste codificador de predição linear é fundamental a escolha da representação do modelo de geração da voz. De forma a estabelecer uma representação dos sinais de voz, é necessário considerar um modelo para o aparelho fonador humano. Uma modelagem física, muito utilizada, é aquela em que o trato vocal é representado por um conjunto de tubos acústicos, que contêm em uma de suas extremidades as cordas vocais e na outra os lábios. A partir deste modelo, uma representação digital para a geração dos sinais de voz pode ser equacionada, por meio de um filtro digital variável no tempo controlado por coeficientes que modelam os parâmetros do trato vocal (Figura 2.4).

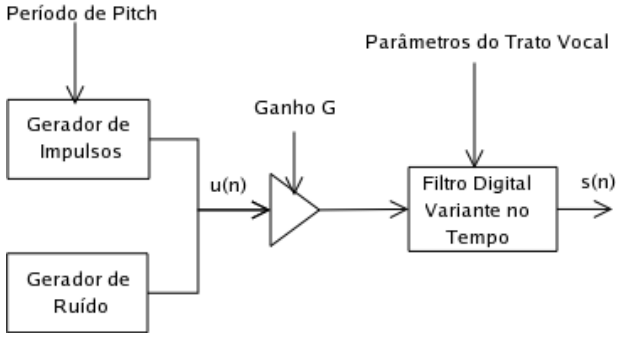


Figura 2. 4: Diagrama simplificado de um modelo de produção de voz.

A excitação deste filtro é feita através de um trem de impulsos quase periódico representando os sons vozeados ou através de uma fonte de ruído branco responsável pelos sons não-vozeados. Este modelo é a base para uma série de representações dos sinais de voz e sua função de transferência é representada na Equação (2.1). A metodologia para obtenção do $H(z)$ será vista, em detalhes, na Sub-seção 2.4.1.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad (2.1)$$

O modelo do trato vocal representado por essa equação é simplificado (modelo só pólos), de forma que sua solução pode ser feita por meio de um sistema de equações lineares.

O funcionamento do codificador LPC baseia-se nesse modelo do trato vocal. O sinal de voz original é dividido em segmentos de 10 a 30 ms (Figura 2.5). A partir da análise de cada segmento, o codificador LPC estima os parâmetros $a_1,...,a_N$ e o ganho G a serem usados no filtro que modela o trato vocal. O inverso deste filtro é aplicado a cada segmento, resultando em um resíduo que basicamente descreve qual deve ser a excitação utilizada para modelar o sinal daquele segmento. Também, através do sinal residual determina-se se o sinal original é vozeado ou não-vozeado e, caso seja do primeiro tipo, determina-se o período de *pitch*.

Para obter os parâmetros do filtro, o algoritmo LPC determina, basicamente, os formantes do sinal original. Este problema é resolvido através de uma equação de diferenças que descreve cada amostra como sendo uma combinação linear das amostras anteriores.

O processo de decodificação utiliza os parâmetros e o ganho do filtro, o período de *pitch* e a classificação de cada segmento para recompor o sinal de voz (Figura 2.5). Com este método são produzidos sinais inteligíveis a taxas de 2,4 kbit/s. Entretanto, o som reconstituído apresenta um aspecto sintético e ligeiramente "metálico".

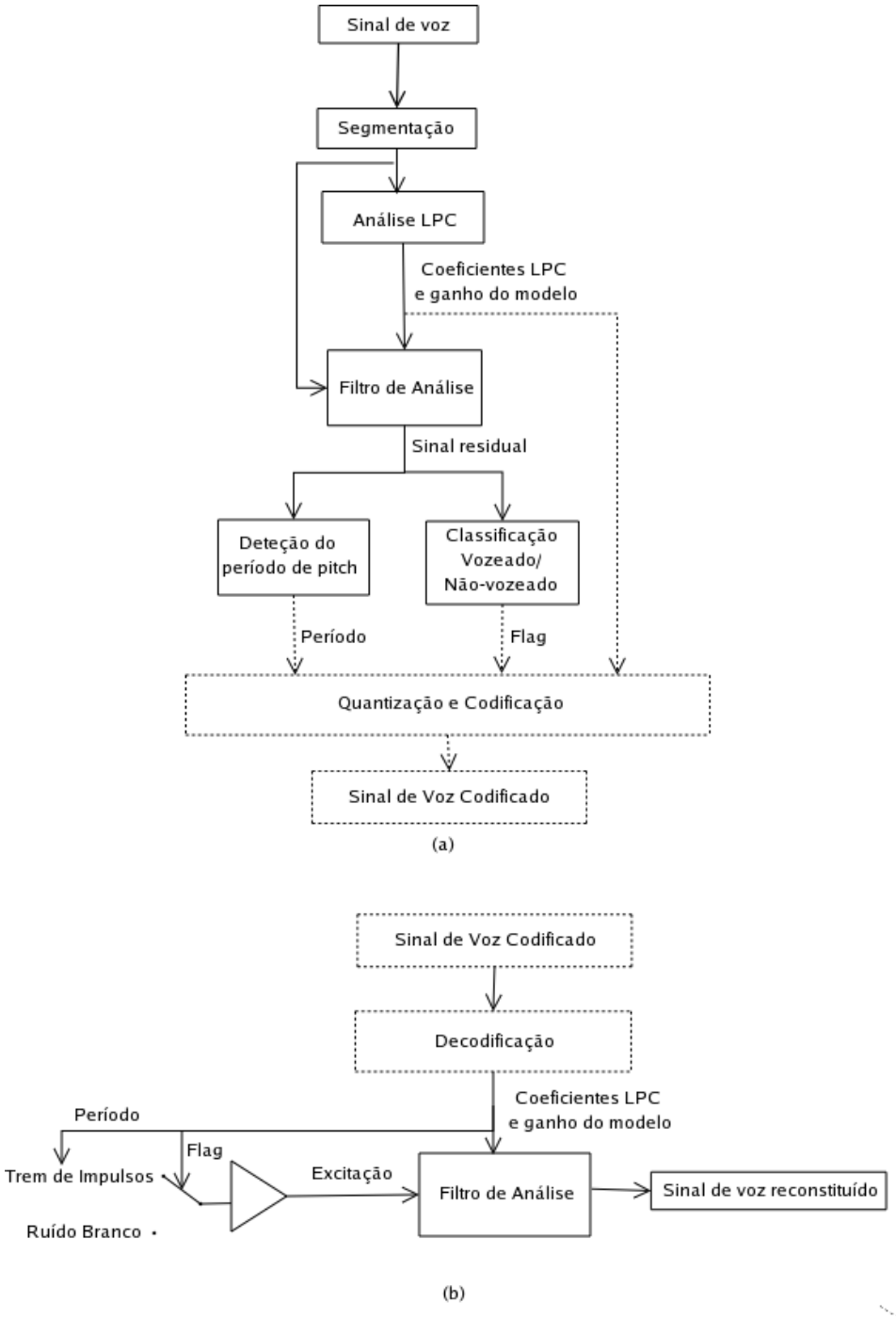


Figura 2. 5: Sistema de codificação LPC como visto em [2]: (a) codificador e (b) decodificador.

2.4 Análise-por-Síntese e a Codificação CELP

Existem vários tipos de codificadores híbridos. No entanto, os mais comumente usados são os de Análise-por-Síntese (*Analysis-by-Synthesis* - *AbS*) no domínio do tempo. Tais codificadores usam o mesmo filtro de predição linear para modelar o trato vocal como encontrado em codificadores LPC. Porém, eles não utilizam a aplicação de apenas dois tipos de excitação, vozeado ou não-vozeado, para encontrar a entrada necessária a esse filtro. Ao invés disto, o sinal de excitação é escolhido tentando-se produzir um sinal reconstituído que tenha a maior proximidade possível com o sinal original. Os codificadores de AbS foram introduzidos, primeiramente, em 1982 por Atal e Schroeder com o codificador que viria a ser conhecido como MPE (*Multi-Pulse Excited*). Mais tarde foram introduzidos o RPE (*Regular-Pulse Excited*) e o CELP (*Code-Excited Linear Predictive*).

O modelo geral para codificadores AbS pode ser visto na Figura 2.6.

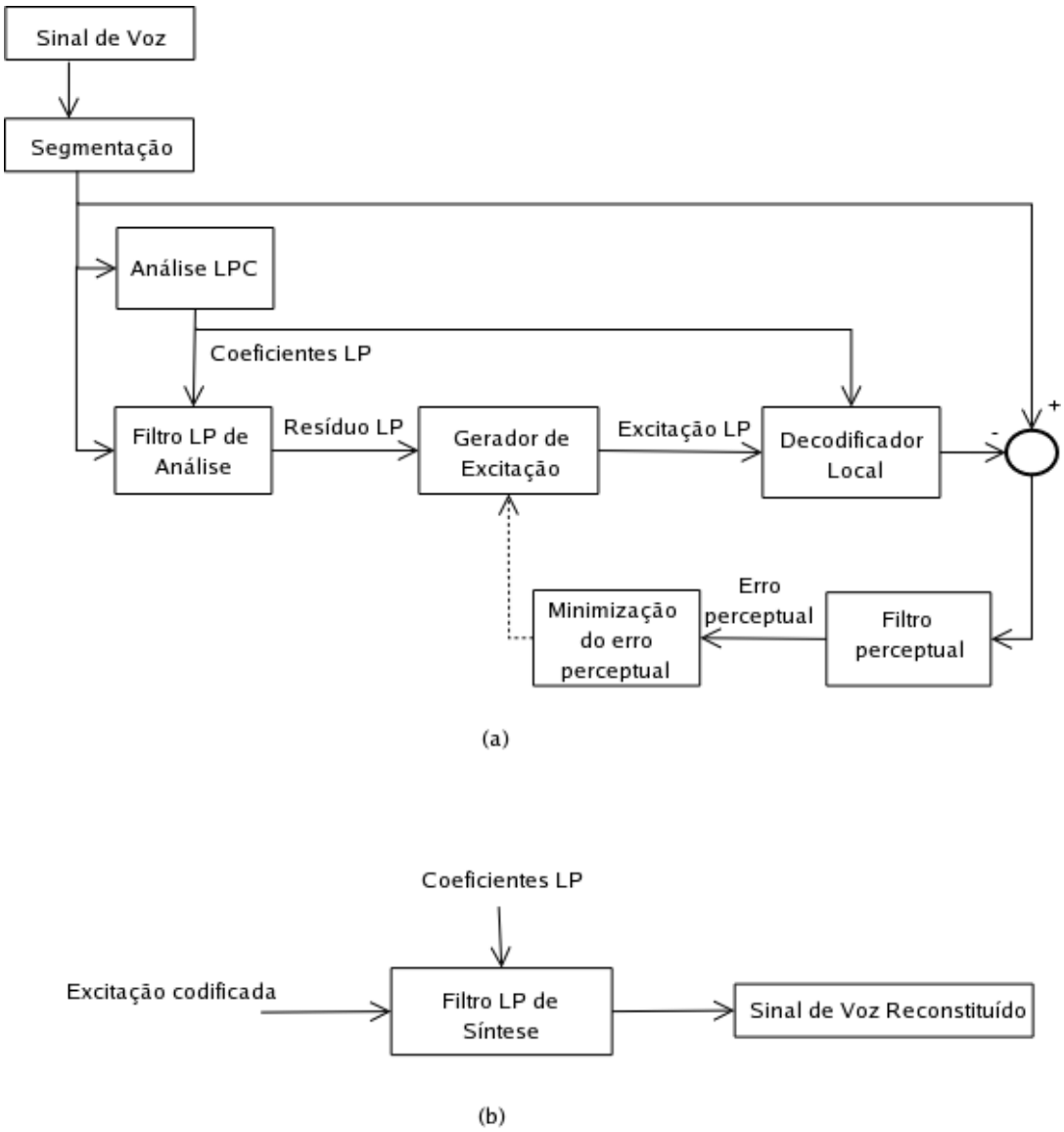


Figura 2. 6: Modelo do (a) codificador e (b) decodificador de análise-por-síntese por predição linear.

Os codificadores AbS dividem o sinal original em segmentos, de aproximadamente 20 ms de duração. Para cada segmento são determinados os parâmetros para o filtro de síntese e a excitação a este filtro. Isto é feito encontrando o sinal de excitação que quando passado pelo filtro de síntese minimize o erro entre o sinal original e o sinal reconstituído. Finalmente, para cada segmento, o codificador transmite a informação que representa os parâmetros do filtro de síntese e a excitação do decodificador. No decodificador, a excitação dada é passada através do filtro de síntese para se obter o sinal reconstituído.

Este método é normalmente conhecido como análise em malha fechada *(closed-loop analysis)* em oposição à análise em malha aberta *(open-loop analysis)*, utilizada em codificadores LPC, na qual os parâmetros são determinados sem que ocorra a reconstituição do sinal de voz. Em geral as duas análises são aplicadas conjuntamente. A malha aberta é útil para a determinação de candidatos iniciais para o sinal de excitação, já a malha fechada seleciona a melhor excitação.

2.4.1 Análise inicial do codificador CELP

A estrutura mais básica de um codificador CELP pode ser vista na Figura 2.7. Ela se baseia em modelos de predição linear que exploram as correlações de curto e longo termos. As correlações de curto-termo são aquelas presentes entre amostras adjacentes de um sinal de voz. Já as correlações de longo-termo são encontradas apenas em sinais vozeados e estão presentes entre amostras da ordem de períodos de *pitch*.

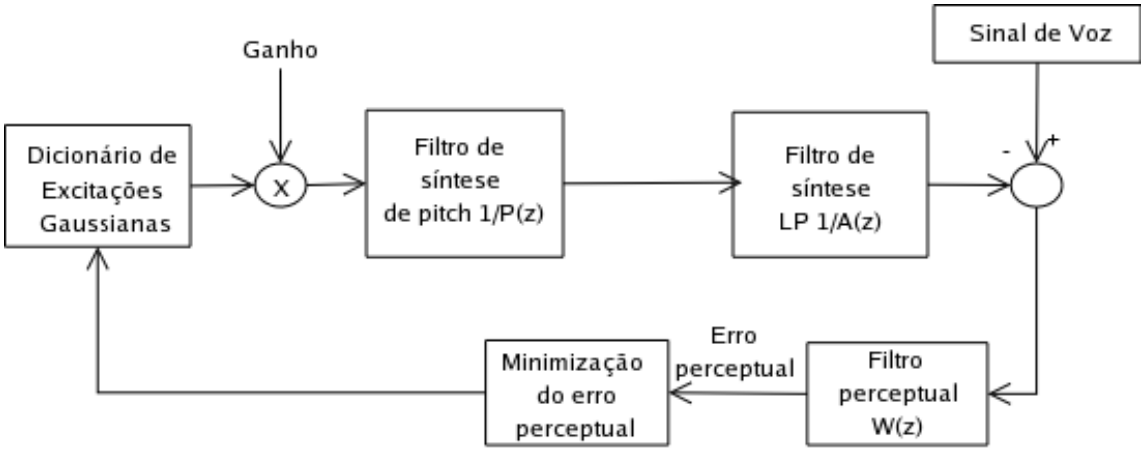


Figura 2. 7: Estrutura básica de um codificador CELP.

Dessa forma, a utilização da predição linear de curto-termo reduz a correlação de curto-termo no sinal residual, e a predição linear de longo-termo elimina a correlação de longo-termo do sinal residual. Este processo possibilita a utilização de um ruído gaussiano como sinal de excitação do codificador.

Filtro de Predição Linear

O modelo de produção da voz, apresentado na figura 2.4, permite a utilização da predição linear para remoção das redundâncias de curto-termo. Considerando, inicialmente, uma janela do sinal de voz ($s[n]$) com p amostras, pode-se assumir que $s[n]$ é a saída de um sistema que tem como entrada a excitação $u[n]$:

$$s[n] = \sum_{i=1}^{p} a_i s[n-i] + G\sum_{k=0}^{Q} b_k u[n-k] \ , \ b_0 = 1 \tag{2. 2}$$

onde a_i, b_k e o ganho G são parâmetros do sistema. Na Equação (2.2), cada amostra do sinal de voz é representada como a combinação linear de saídas e entradas anteriores e da entrada atual. A transformada Z desse sistema é dada por

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1+\sum_{k=1}^{Q} b_k z^{-1}}{1-\sum_{i=1}^{p} a_i z^{-i}} \tag{2. 3}$$

A Equação (2.3) é o modelo geral de pólos e zeros. Na maior parte das aplicações em tempo real, o modelo só pólos é preferido, porque ele é computacionalmente mais eficiente além de ser uma boa representação dos efeitos do trato vocal. Como o ouvido humano é mais sensível aos pólos do espectro do que aos zeros [**5**], torna-se possível a simplificação de $H(z)$. Além disso, é conhecido que o efeito dos zeros na função de transferência pode ser alcançado pela inclusão de mais pólos.

Baseando-se no modelo só pólos, uma amostra do sinal de voz é predita pela combinação linear de p amostras passadas.

$$s[n] = \sum_{i=1}^{p} a_i s[n-i] \tag{2. 4}$$

Define-se, então, o filtro de análise de predição linear como

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{2. 5}$$

e o filtro de síntese de predição linear como

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2. 6}$$

o qual modela a correlação de curto termo do sinal de voz produzida pelo trato vocal. Os coeficientes LP $a_1,...,a_p$ são estimados através das técnicas de autocorrelação ou covariância, de modo a formar um conjunto de parâmetros que apresentem maior robustez a erro. O método de autocorrelação costuma ser o preferido já que ele garante que o filtro de análise de predição linear $A(z)$ é de fase mínima, o que significa que o filtro de síntese $1/A(z)$ é sempre estável.

O número p de coeficientes LP utilizados está relacionado ao número de formantes presentes no espectro do sinal de voz. Em geral, para uma freqüência de amostragem de 8000 Hz, são usados de 10 a 12 coeficientes para modelar o formato do espectro. Antes da codificação, esses coeficientes são transformados em outro conjunto de parâmetros de forma a obter uma quantização mais eficiente. Segundo [**5**], uma das representações mais usadas é o LSF (*Line Spectral Frequencies*), também conhecido como LSP (*Line Spectral Pairs*).

A excitação, que quando aplicada ao filtro de síntese $1/A(z)$ irá gerar um sinal de voz, é modelada através do resíduo LP resultante da filtragem do sinal de voz original pelo filtro de análise de predição linear.

Sons vozeados e não-vozeados que foram aplicados ao filtro de análise $A(z)$ e o resíduo LP correspondente podem ser vistos na Figura 2.8. Pode ser observado que o resíduo correspondente ao sinal não-vozeado é de natureza aleatória. Já no resíduo do sinal vozeado podem-se observar picos de energia bem definidos, que correspondem aos pulsos de *pitch* presentes na excitação da voz.

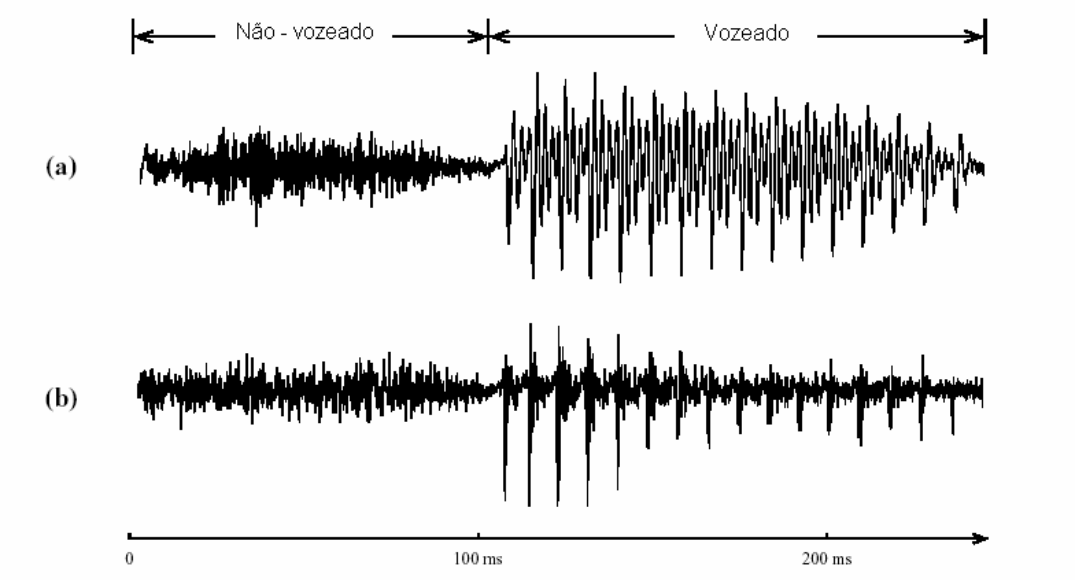


Figura 2. 8: (a) Trecho de sinal vozeado e não-vozeado aplicados ao filtro de análise A(z) e o (b) resíduo LP correspondente.

O modelo de predição linear é capaz de produzir sinais reconstituídos de elevada qualidade quando são codificados trechos não-vozeados do sinal original. A excitação utilizada neste caso costuma ser um ruído gaussiano. A dificuldade deste modelo em codificar os trechos vozeados está no fato de o ouvido humano ser muito sensível a variações de periodicidade do sinal de voz. Dessa forma, pequenas variações em pulsos consecutivos de *pitch* dificultam a modelagem dos sinais vozeados por este método.

Filtro de *Pitch*

A periodicidade encontrada no sinal residual LP devido à correlação de longo-termo pode ser modelada pelo filtro de síntese de *pitch*, com forma

$$\frac{1}{P(z)} = \frac{1}{1-\beta z^{-M}}, \tag{2. 7}$$

onde β e M são respectivamente o coeficiente de predição e o período de *pitch* estimado em amostras (*pitch lag)*. O coeficiente de predição β pode ser interpretado como um indicativo do nível de periodicidade do sinal e assume valores entre 0 e 1. Assim, β aproxima-se de 1 quanto mais periódico for o sinal e de 0 quanto menos periódico for, sendo que nesse caso o valor de M é irrelevante. Embora os parâmetros do filtro de *pitch* sejam determinados através do método de AbS (método da malha fechada), a estimativa inicial dos mesmos é realizada pelo método da malha aberta.

Quando o resíduo LP é aplicado ao filtro de análise de *pitch* $P(z)$, pode-se observar que os picos de energia correspondentes aos pulsos de *pitch* são removidos do espectro de potência, resultando em um sinal aleatório. Já as propriedades do resíduo LP correspondentes ao sinal não-vozeado não se alteram, já que este é uma excitação aleatória (sem estrutura de harmônicos). Assim, a utilização de uma excitação gaussiana aplicada ao filtro de síntese de predição linear em série com o filtro de síntese de *pitch* representa uma boa modelagem das propriedades do sinal de voz.

Filtro de Ponderação Perceptual

No codificador, os parâmetros que descrevem a excitação de entrada do filtro de síntese são determinados através da minimização do erro médio quadrático ponderado perceptual (*perceptually weighted mean square error*) entre o sinal original e o reconstituído. A ponderação perceptual (*perceptual weighting*) explora a propriedade de mascaramento do sistema auditivo humano. O mascaramento faz com que o ruído localizado em bandas de freqüência de elevada energia seja menos audível do que o ruído localizado em freqüências que correspondem a vales de energia. O filtro de ponderação perceptual enfatiza o erro nesse vales do espectro do sinal original e atenua o erro em regiões de pico do espectro. Como efeito, o ruído quantizado dos vales é

reduzido. Este ruído enfatizado nos picos é mascarado pelo sistema auditivo humano. O filtro de ponderação perceptual é especificado como

$$W(z) = \frac{A(z)}{A\left(\frac{z}{\gamma}\right)}, \quad 0 < \gamma \leq 1 \tag{2. 8}$$

onde $A(z)$ é o filtro de análise de predição linear dado pela Equação (2.5) e γ é um valor fixo ou adaptativo. Na Figura 2.9 pode-se observar seu comportamento.

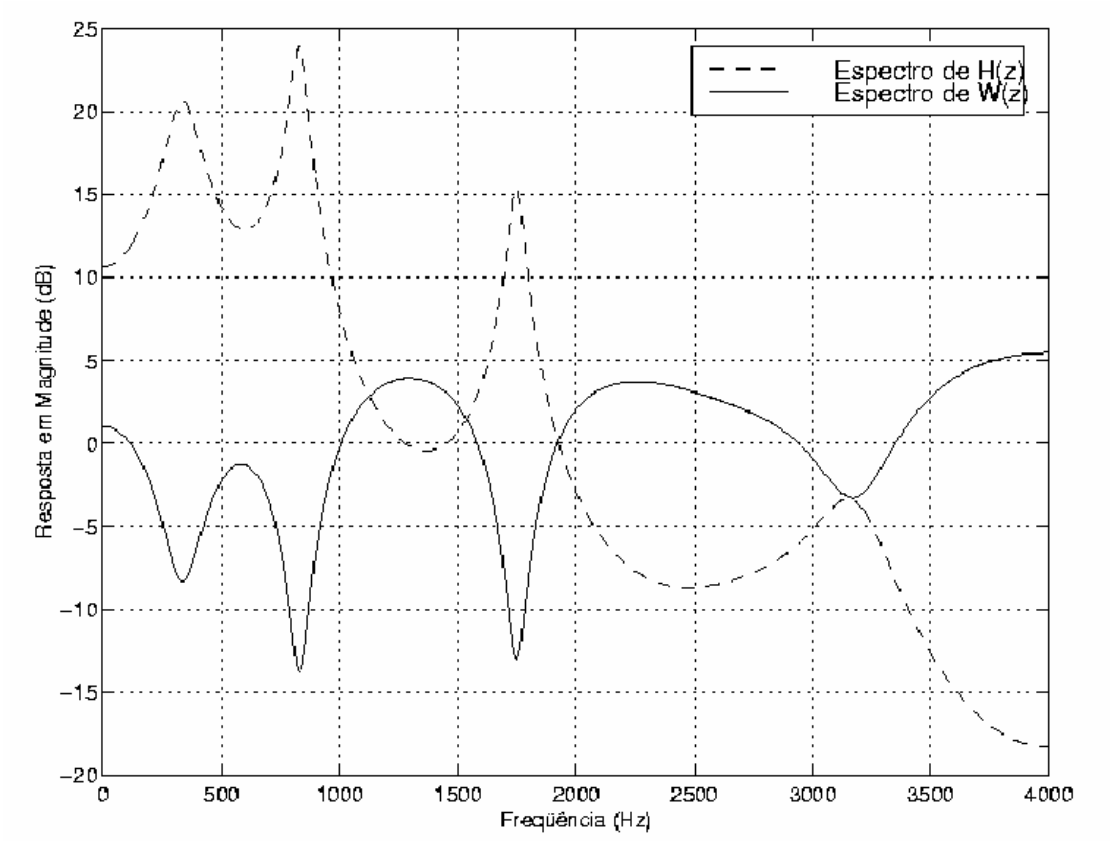


Figura 2. 9: Comportamento do filtro de ponderação perceptual como visto em [2].

2.4.2 CELP com dicionário adaptativo

O sinal de excitação LP pode ser modelado através da quantização vetorial. O vetor que representa a excitação LP é selecionando, através do método de análise por síntese, de um banco de vetores de excitação denominado dicionário (*codebook*). Os dicionários podem ser fixos ou adaptativos como veremos a seguir.

Dicionário Fixo

O trecho não-vozeado do sinal de excitação pode ser modelado pelo dicionário fixo. Este dicionário pode ser de dois tipos: (i) estocástico ou (ii) determinístico. O primeiro é populado por números gaussianos aleatórios e independentes, enquanto que o segundo é populado por sinais derivados da voz obtidos iterativamente durante a busca em malha fechada.

Em uma versão inicial do CELP, eram utilizados dicionários fixos estocásticos. A realização de uma busca em um dicionário tão desestruturado causa um grande aumento da complexidade computacional do sistema na busca pelo vetor de excitação ótima. Para reduzir esta complexidade e o espaço de armazenamento algumas variações estruturais são implementadas. As modificações propostas incluem dicionário com superposição (*overlapped codebook*), dicionário esparso (*sparse codebook*), dicionário algébrico (*algebraic codebook*), dicionário *lattice* (l*attice codebook*), dicionário ternário (*ternary codebook*) e dicionário **"**treinado**"** (*trained codebooks*).

O dicionário fixo é usado também para modelar o início e as mudanças da excitação da voz. O mesmo dicionário fixo é utilizado no codificador e no decodificador, dessa forma, apenas o índice do dicionário relativo à excitação selecionada é transmitido.

Dicionário Adaptativo

Em implementações mais modernas do codificador CELP, o filtro de síntese de *pitch* $1/P(z)$ é modelado por um dicionário adaptativo. A estrutura do codificador CELP com dois dicionários, o fixo e o adaptativo, pode ser observada na Figura 2.10.

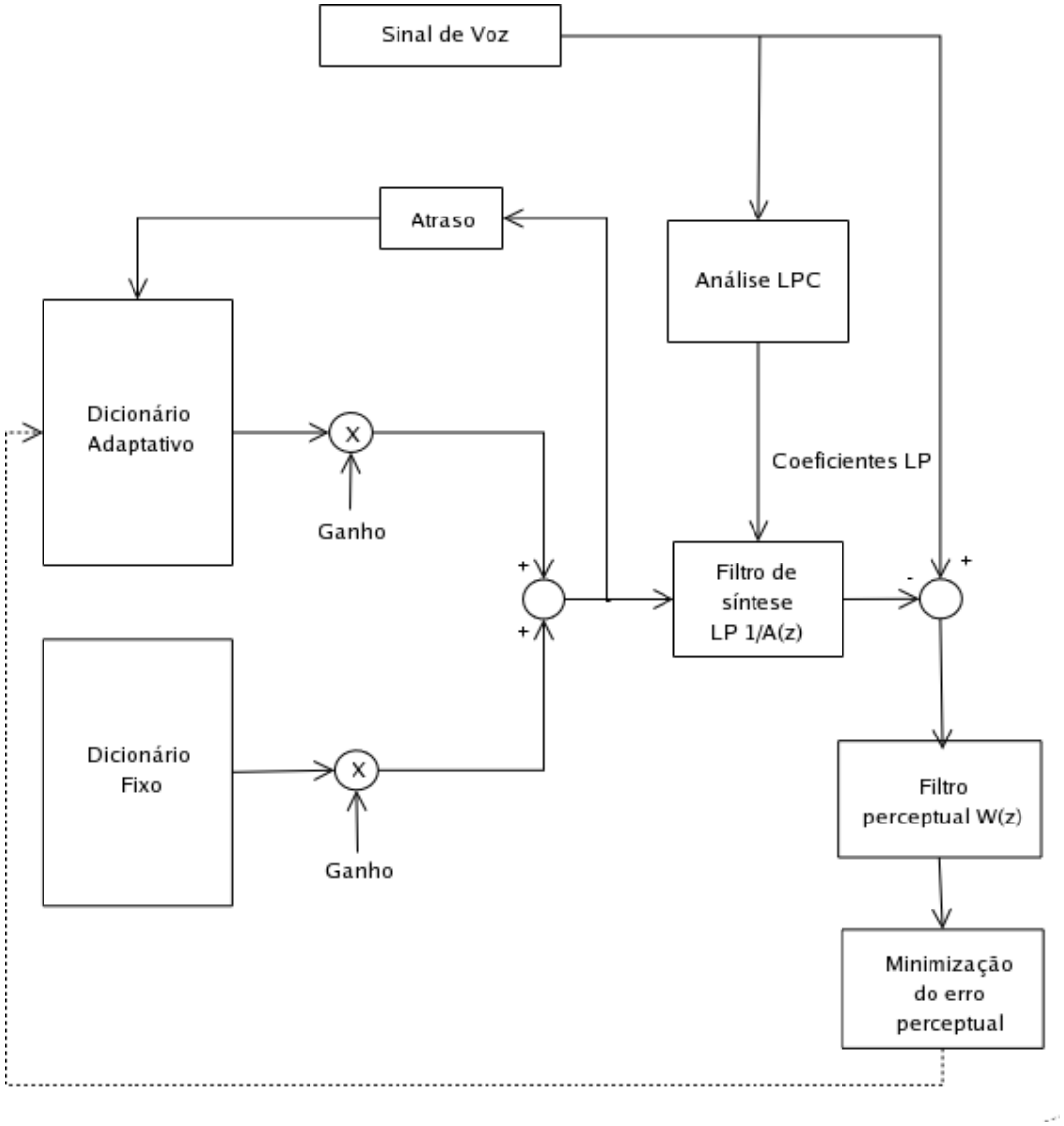


Figura 2. 10: Codificador CELP implementado com um dicionário fixo e um adaptativo.

Inicialmente, é realizada uma busca seqüencial em que todas as entradas do dicionário adaptativo são testadas e o vetor que minimiza o erro perceptual é selecionado. O ganho ótimo relativo a este vetor é também calculado. A diferença entre o sinal original e o vetor do dicionário adaptativo multiplicado pelo ganho ótimo é utilizada para realizar a busca no dicionário fixo. Novamente, o vetor que minimiza o erro perceptual é selecionado e o ganho ótimo relativo a ele é calculado.

O dicionário adaptativo pode ser interpretado como uma generalização do filtro de *pitch* e suas entradas podem ser formadas pela resposta do filtro de análise *pitch* quando uma excitação LP é aplicada na sua entrada.

Na estrutura CELP com dois dicionários, várias técnicas são aplicadas para obter a melhor representação da excitação LP, que quando aplicada ao filtro de síntese de predição linear $1/A(z)$ resulta no sinal reconstituído. Para tal, o codificador CELP realiza as seguintes operações:

(i) As entradas dos dicionários são selecionadas e os ganhos referentes às mesmas são calculados. Este cálculo será visto, em detalhes, na Seção 3.6;
(ii) As entradas selecionadas são multiplicadas pelos ganhos e combinadas para formar a excitação LP. O sinal codificado é então sintetizado. Podem ser utilizados os ganhos calculados em (i) ou estes podem ser atualizados;
(iii) O dicionário adaptativo é atualizado.
As operações (i) e (iii) são realizadas no codificador, tendo como objetivo calcular os parâmetros que serão usados em (ii). As operações (ii) e (iii) são realizadas no decodificador. A realização da operação (iii) garante que os dicionários adaptativos do codificador e do decodificador permanecerão exatamente iguais.

2.4.3 Padrões CELP

O padrão de codificação CELP produz sinais de voz reconstituídos com elevada qualidade a taxas de bits que variam de 4,8 a 16 kbit/s (Figura 2.11). No entanto, eles, geralmente, apresentam um elevado atraso causado pela utilização do filtro de curto-termo. Este atraso é definido como o tempo gasto quando dada uma amostra na entrada do codificador gera-se a correspondente na saída do decodificador. Para os codificadores híbridos este atraso pode ser de 50 a 100 ms. Portanto, em 1988 a ITU-T definiu uma série de exigências para o padrão de codificação a 16 kbits/s. A principal exigência era de que este novo codificador apresentasse qualidade comparável à recomendação G.721 32 kbits/s (ADPCM) para canais ruidosos, e com atraso inferior a 5ms, sendo que em condições ideais este atraso deveria ser inferior a 2ms.

Todas as especificações da ITU-T foram atendidas pelo *backward adaptive* CELP desenvolvido por AT\&T Bell Labs, o qual foi padronizado em 1992 como LD-CELP (*Low Delay CELP*).

LD-CELP

A recomendação da ITU-T para o LD-CELP é chamada G.728 e pode ser encontrada em [**4**]. Este codificador utiliza adaptação retroativa (*backward adaption*) para calcular os coeficientes do filtro de curto-termo, o que significa que eles são calculados a partir de amostras anteriores do sinal. Assim, não é mais necessário armazenar 20 ms do sinal para este cálculo, o que permite a utilização de segmentos menores dos que os tradicionalmente usados. O G.728 utiliza segmentos com apenas 20 amostras resultando em um atraso de 2,5ms.

O preditor de longo-termo é eliminado, sendo ao invés utilizado um preditor de curto-termo de elevada ordem (N = 50). Dessa forma, os 10 bits disponíveis, referentes a cada sub-segmento de 5 amostras, são utilizados para representar a excitação e o ganho do dicionário fixo. Desses 10 bits, 7 são usados para codificar o índice do dicionário fixo e 3 para codificar o ganho a esta excitação. Amostrando-se o sinal de voz original a 8000 Hz e utilizando-se 10 bits para codificar cada sub-segmento de 5 amostras obtém-se um codificador de 16 kbits/s, que apresenta qualidade igual ou superior ao G.721 e é bastante robusto a canais ruidosos.

CS-ACELP

A recomendação da ITU-T para o CS-ACELP (*Conjugate-Structured Algebraic* CELP) é chamada G.729 e pode ser encontrada em [**4**]. Seu princípio de funcionamento é bastante semelhante ao do DoD-CELP.

Neste codificador, o sinal de voz é dividido em segmentos de 10 ms (onde são analisadas 80 amostras de 8 bits), sendo novamente dividido em sub-segmentos de 5 ms. Ele utiliza quantização vetorial em dois estágios para codificar os parâmetros LSF e os ganhos. Os 80 bits utilizados na codificação são distribuídos da seguinte forma: 18 para os coeficientes LSF, 14 para o filtro de predição de *pitch*, 34 para os índices do dicionário e 14 para os ganhos.

VSELP

Como já visto, codificadores CELP apresentam boa performance a baixas taxa de bits. Entretanto, apresentam uma grande desvantagem devido ao peso computacional do algoritmo. Este problema motivou o desenvolvimento de um codificador com dicionários estruturados e com rotinas mais rápidas e eficientes de busca pela excitação ótima.

Gerson e Jasiuk propuseram o codificador VSELP (*Vector Sum Excited Linear Predictive*) que apresenta uma rápida busca nos dicionários e é robusto a erros de canal. O VSELP com taxa de 8 kbits/s foi adotado pela TIA (*Telecommunications Industry Association*) como o padrão para telefonia digital celular da América do Norte.

Neste codificador, os dicionários são organizados com uma estrutura pré-definida que reduz significativamente o tempo gasto para buscar a excitação ótima. São utilizadas 3 fontes de excitação. A primeira é um filtro de longo-termo (dicionário adaptativo). A segunda e a terceira são dicionários fixos. Cada um dos dicionários contém 128 vetores de excitação. As três fontes de excitação, multiplicadas pelos seus respectivos ganhos, são somadas para formar a excitação final. É utilizado um filtro LP de síntese de 10ª ordem. Os coeficientes LPC são codificados a cada 20 ms e são atualizados em cada sub-segmento de 5 ms através de interpolação. Os parâmetros da excitação também são atualizados em cada sub-segmento de 5 ms. À uma freqüência de amostragem de 8000 Hz, cada sub-segmento conterá 40 amostras Os detalhes da implementação podem ser encontrados na recomendação IS-54.

DoD-CELP

Em 1991, o Departamento de Defesa Americano (*American Department of Defense, DoD*) padronizou o codificador CELP a 4,8 kbits/s como o Padrão federal 1016. No DoD-CELP, o sinal original é dividido em segmentos de 30 ms, sendo cada um novamente dividido em sub-segmentos de 7,5 ms. Para cada segmento o codificador calcula um conjunto de 10 coeficientes do filtro de síntese de curto-termo usado para modelar o trato vocal. A excitação a ser aplicada a este filtro é estimada para cada sub-segmento e é dada pela soma das entradas escaladas de 2 dicionário, 1 fixo e 1 adaptativo. Essa é uma aplicação direta dos conceitos do CELP descritos na Sub-seção 2.4.2.

2.5 Conclusão

Como foi visto, o codificador utilizado influencia na qualidade do sistema de várias maneiras. Ele próprio introduz uma distorção no sinal de voz. Essa distorção pode ser comparada entre vários codificadores utilizando-se a técnica MOS (*Mean Opinion Score*) de avaliação da qualidade da voz. Na Figura 2.11 observamos um gráfico com esta comparação. O segundo aspecto a ser observado é a largura de banda ocupada. Em uma rede digital saturada, com um grande número de canais de voz, isto pode ser crítico. Devido a esta preocupação, a maior parte das pesquisas nesta área está concentrada na região marcada no gráfico em que garantem-se baixas taxas de bits e elevada qualidade da voz.

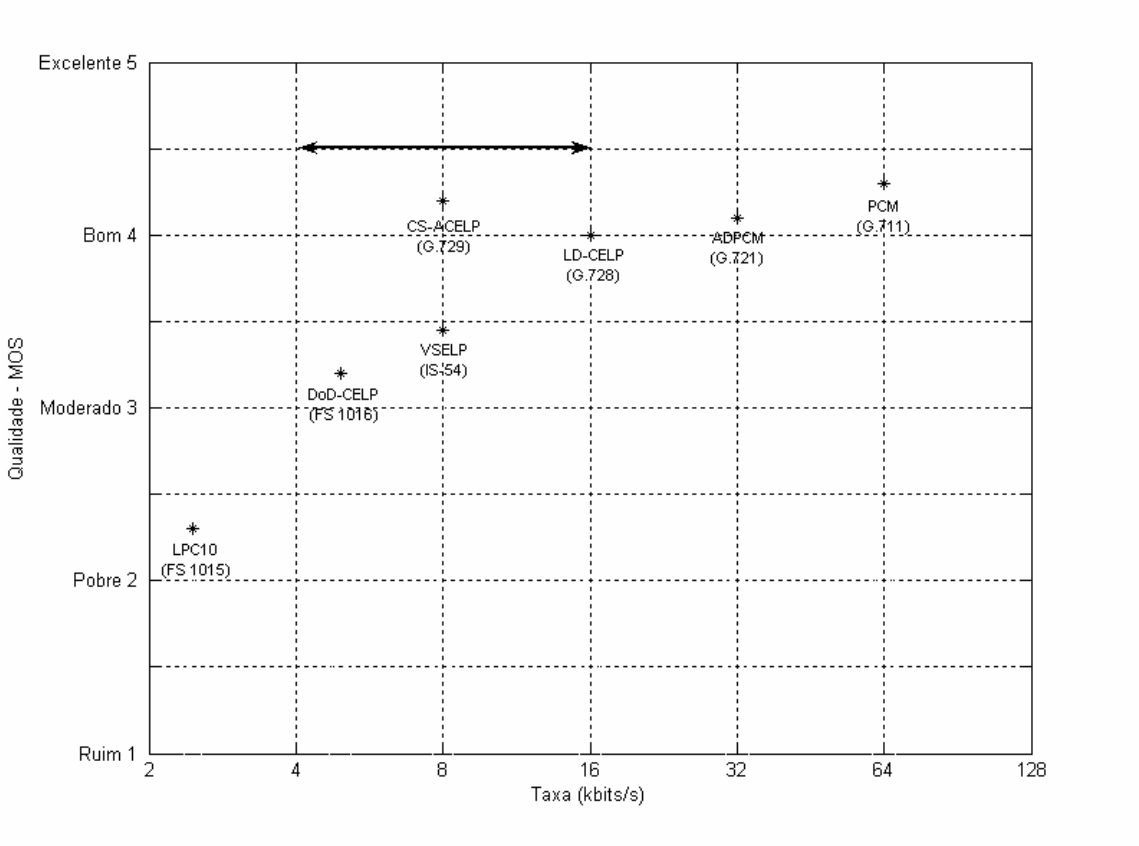


Figura 2. 11: Taxa de bits x qualidade de alguns algoritmos de codificação da voz. A seta evidencia a região onde as pesquisas estão concentradas atualmente: codificadores com baixas taxas de bits e elevada qualidade da voz codificada.

Outro aspecto é o atraso introduzido pela formação de um pacote. Isto depende do número de amostras que são tomadas e da taxa de compressão. Um último fator a ser analisado é a complexidade do algoritmo de compressão. Esse se torna um fator crítico em sistemas de processamento de voz em tempo real, uma vez que é necessário dimensionar processadores que suportem o algoritmo sem imposição de atrasos adicionais.

Capítulo 3

Codificador CELP

3.1 Introdução

Neste capítulo, são apresentados detalhes da implementação do codificador CELP descrito em [1,2]. De um modo geral, ele é modelado conforme mostra a Figura 2.9. É sobre esta implementação que será desenvolvido todo o restante do trabalho.

A Seção 3.2 aborda o tipo de janela utilizada; as Seções 3.3 e 3.4 apresentam aspectos do filtro de síntese e alterações realizadas nele; na Seção 3.5, são apresentadas particularidades da estrutura dos dicionários e a Seção 3.6 apresenta detalhes sobre o cálculo de seus ganhos; a Seção 3.7 apresentada detalhes da preparação do sinal de entrada; na Seção 3.8, é apresentada a estrutura do *bitstream;* nas Seções 3.9 e 3.10, o funcionamento do algoritmo é detalhado; já na Seção 3.11, é realizada uma conclusão sobre esta implementação CELP.

3.2 Segmentação do sinal de voz

Como visto na Seção 2.2, o sinal de voz não é verdadeiramente estacionário, pois suas características espectrais e estatísticas variam ao longo do tempo. Porém, observa-se que, em trechos de 10 a 30 ms, este sinal é aproximadamente estacionário. Na implementação CELP proposta em [1], são processados 20 ms de sinal por vez. Como é utilizada uma freqüência de amostragem de 8000 Hz, cada segmento é formado por 160 amostras.

O conjunto de 160 amostras será chamado janela. A divisão do sinal de voz em janelas não pode ser feita pelo simples truncamento de amostras. O resultado seria a inserção de componentes de alta freqüência que não estavam presentes no sinal original, efeito chamado “vazamento” espectral (*spectral leakage*). Isto ocorre, pois o período de amostragem não contém um número inteiro de períodos de *pitch*.

Em [1] foi utilizada a janela de *Hamming* que é dada por:

$$w_H[n] = \begin{cases} \alpha + (1-\alpha)\cos\left(\dfrac{2\pi n}{M}\right); & |n| \le \dfrac{M}{2} \\ 0; & |n| > \dfrac{M}{2} \end{cases} \qquad (3.1)$$

onde $\alpha = 0{,}54$, $M = N - 1$ e $N = 160$ (número de amostras de uma janela do sinal de voz).

3.3 Filtro de síntese de predição linear

O filtro de síntese de predição linear de forma

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{3. 2}$$

é de 10ª ordem. Para o cálculo dos coeficientes LP $a_1,...,a_p$ foi escolhido o método de autocorrelação. Este método utiliza a janela de *hamming* de tamanho finito descrita na Seção 3.2. Após a aplicação da janela ($w_H[n]$) ao sinal de voz original ($s[n]$), obtém-se o sinal:

$$s_w[n] = w_H[n]s[n] \tag{3. 3}$$

A predição linear seleciona os coeficientes LP que minimizam a energia E_p do erro de predição linear ($e_w[n]$), dada por:

$$E_p = E\left[e_w^2[n]\right] = E\left[\left(s_w[n] - \sum_{i=1}^{p} a_i s_w[n-i]\right)^2\right] \tag{3. 4}$$

Igualando a zero a derivada parcial de E_p com relação aos coeficientes LP a_i obtém-se a equação linear:

$$\sum_{i=1}^{N} a_i r_s[k-i] = r_s[k] \tag{3. 5}$$

onde $r_s[k]$ é a autocorrelação do sinal $s_w[n]$.

A Equação (3.5) pode ser representada na forma matricial como:

$$\begin{bmatrix} r_s(0) & r_s(1) & \cdots & r_s(p-1) \\ r_s(1) & r_s(0) & \cdots & r_s(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_s(p-1) & r_s(p-2) & \cdots & r_s(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_s(1) \\ r_s(2) \\ \vdots \\ r_s(p) \end{bmatrix} \tag{3. 6}$$

$$R_s a = r_s$$

Dessa forma, os coeficientes LP são obtidos através da solução da equação matricial

$$a = R_s^{-1} r_s \tag{3. 7}$$

onde a é o vetor com os coeficientes LP ótimos de dimensão p por 1, r_s é o vetor de autocorrelação de dimensão p por 1 e R_s^{-1} é a inversa da matriz de autocorrelação de dimensão p por p. Observando a matriz de autocorrelação R_s pode-se notar que

(i) todos os elementos da diagonal principal da matriz de autocorrelação são iguais;
(ii) os elementos de qualquer outra diagonal paralela à diagonal principal também são iguais.

Uma matriz quadrada que apresenta essas propriedades é chamada *Toeplitz*. Além disto, esta matriz também é simétrica. Dessa forma, para solucionar a Equação (3.7) de forma eficiente é usado o algoritmo recursivo de *Levinson-Durbin* [2]. Ele foi desenvolvido para resolver, especificamente, o problema da inversão de uma matriz *Toeplitz* simétrica.

Coeficientes LSF

Como abordado no Capítulo 2, os coeficientes LP são extremamente sensíveis a erros de quantização. Assim, em [2] eles são transformados em coeficientes LSF (*Line Spectral Frequencies*), apresentados por Itakura [3]. Considerando os polinômios $P(z)$ e $Q(z)$ dados por:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned} \tag{3. 8}$$

Obtém-se:

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \tag{3. 9}$$

O filtro $A(z)$ será estável apenas se todos os zeros dos polinômios LSF $P(z)$ e $Q(z)$ estiverem intercalados sobre o círculo unitário. Os coeficientes LSF são dados pela posição angular destes zeros, a qual corresponde, aproximadamente, à posição angular das raízes de $A(z)$ (freqüências dos formantes). Já a separação entre uma raiz de $P(z)$ e a sua adjacente de $Q(z)$ fornece uma estimativa da largura de banda de ressonância. É necessário calcular apenas $p/2$ zeros, já que os polinômios apresentam pares de zeros complexos conjugados.

Os coeficientes LSF apresentam a importante propriedade de robustez à distorção. De acordo com ela, qualquer alteração sofrida por um desses coeficientes não terá um efeito global. Apenas será afetada a região do espectro próxima a esta freqüência. Esta propriedade pode ser explorada em sistemas de codificação da voz, uma vez que o ouvido humano não é muito sensível a variações em freqüências elevadas. Nesses sistemas, é possível representar os coeficientes LSF de elevadas freqüências com um menor número de bits, o que possibilita uma diminuição da taxa de bits do sistema.

De acordo com as propriedades acima, pode-se concluir que a utilização de coeficientes LSF apresenta vantagens em relação aos coeficientes LP em termos de transmissão, quantização e interpolação. No entanto, o cálculo direto dos coeficientes LSF exige uma elevada capacidade computacional. A alternativa é calcular os coeficientes LP e depois transformá-los em LSF.

A quantização é realizada sobre as diferenças entre os coeficientes LSF, chamadas DLSF. Por apresentarem menor amplitude, a quantização escalar das diferenças terá um melhor resultado. Foi utilizada a quantização QDLSF-32, cuja distribuição de bits pode ser vista na Tabela 3.1.

Tabela 3. 1: Distribuição de bits na quantização QDLSF-32 dos coeficientes DLSF.

Coeficientes LSF	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
Nº de bits por coeficiente	4	4	3	3	3	3	3	3	3	3
Total de bits da quantização	32									

Interpolação

Os coeficientes LP $a_1,...,a_p$ do filtro de síntese, que modela o trato vocal, são calculados para cada janela. No entanto, a utilização dos mesmos 10 coeficientes para modelar a produção de um sinal de voz de 20 ms não acompanha o comportamento contínuo de funcionamento do trato vocal. Como alternativa, cada janela é dividida em 4 sub-janelas de 40 amostras. Em cada sub-janela, é realizada a interpolação dos 10 coeficientes da janela atual com os da sub-janela anterior, o que garante uma evolução suave dos mesmos.

A interpolação é realizada sobre os coeficientes LSF $w_1,...,w_p$ por razões citadas anteriormente. Os 10 coeficientes LP, calculados em cada janela, são transformados em LSF. É realizada a quantização e em seguida a interpolação, que resulta em um conjunto de 10 coeficientes LSF por cada sub-janela. Eles são, novamente, convertidos em coeficientes LP. Dessa forma, cada sub-janela terá um filtro de síntese com 10 coeficientes LP para modelar o sinal de voz.

Em [**1**], a interpolação é realizada da seguinte forma:

$$w_i^n = (1-q_n)w_i^a + q_n w_i \tag{3. 10}$$

onde w_i representa o coeficiente LSF i da janela atual; w_i^a representa o coeficiente LSF i da sub-janela anterior; q_n representa o peso dos coeficientes na sub-janela n; e w_i^n representa o coeficiente i da sub-janela n que se deseja obter. Temos ainda que $n = \{1;2;3;4\}$ e $q_n = \{0{,}25;0{,}50;0{,}75;1\}$.

Após este processo, em cada sub-janela é formado um filtro de síntese de forma

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} \hat{a}_i \, z^{-i}} \tag{3. 11}$$

onde $\hat{a}_1, \ldots, \hat{a}_p$ correspondem aos coeficientes LP recalculados.

3.4 Filtro de síntese de predição linear resultante

No Capítulo 2, foi vista a razão pela qual o filtro de ponderação perceptual de forma

$$W(z) = \frac{A(z)}{A\left(\frac{z}{\gamma}\right)} \tag{3. 12}$$

é aplicado. Em [1], este filtro foi deslocado para a esquerda do somador de forma a aumentar a velocidade de processamento do algoritmo de codificação. Como resultado, obteve-se um filtro de síntese que é formado pela aglutinação dos filtros das Equações (3.11) e (3.12). Desta forma, não há aumento na complexidade computacional do sistema e o filtro passa a ser dado por

$$H(z) \times W(z) = \frac{1}{A(z)} \times \frac{A(z)}{A\left(\frac{z}{\gamma}\right)} = \frac{1}{A\left(\frac{z}{\gamma}\right)} \tag{3. 13}$$

onde o valor de γ, que controla a energia do erro na região dos formantes, é 0,8.

3.5 Dicionários

Como é obtido um filtro de síntese para cada sub-janela, as excitações dos dicionários terão um comprimento de 40 amostras.

Dicionário Fixo

O dicionário fixo é utilizado para modelar sons surdos do sinal de voz. Ele é formado por excitações gaussianas com média zero. Seu tamanho pode ser alterado de acordo com as necessidades de implementação. Para simplificar os cálculos, em [**1**], foi utilizado um processo de *clippling*. De acordo com este processo, qualquer amostra da excitação com módulo menor que um determinado valor é zerada. Foi escolhido o valor de 1,645 de acordo com [**6**].

Dicionário Adaptativo

O dicionário adaptativo é responsável pela introdução da periodicidade no sinal de voz reconstituído. Assim como o dicionário fixo, seu tamanho pode ser alterado de acordo com as necessidades de implementação. No início do processo de codificação, todas as suas amostras tem valor zero. Com o decorrer do processamento, ele é atualizado com o sinal resultante da soma das excitações ótimas, dos dois dicionários, já multiplicadas pelos respectivos ganhos.

Estrutura dos dicionários

Tendo em vista a redução do tempo de processamento, foi proposto por [**2**] uma alteração de acordo com a qual todas as excitações dos dicionários são filtradas de uma única vez pelo filtro mostrado na Equação (2.13). Esta filtragem é realizada em todo o dicionário no início do processamento de cada janela antes de se dar início à busca pela excitação ótima.

Além disso, também foi proposta a utilização de dicionários com estrutura em vetor linha. Assim, as excitações são extraídas dos dicionários da seguinte forma: a primeira excitação é composta pelas amostras de 1 a 40, a segunda pelas amostras de $(1+t)$ a $(40+t)$, a terceira pelas amostras de $(1+2t)$ a $(40+2t)$ e assim por diante. O passo t utilizado varia de acordo com o tipo do dicionário.

O trabalho será desenvolvido sobre a implementação CELP que apresenta as seguintes características:

1. É utilizado um passo (t) igual a 2 para percorrer o dicionário fixo. O número total de possíveis excitações a serem geradas é 256;
2. É utilizado um passo (t) igual a 1 para percorrer o dicionário adaptativo. O número total de possíveis excitações a serem geradas é 1024.

3.6 Ganhos dos dicionários

Como pode ser observado na Equação (3.2), o filtro de síntese de predição linear $1/A(z)$ utilizado é mônico. Isto resulta no surgimento de uma diferença de energia entre o sinal de voz original ($s[n]$) e o reconstituído ($\hat{s}[n]$).

O método de ajuste da energia do sinal reconstituído é aplicado a cada sub-janela e baseia-se na estimativa de um ganho G para reduzir essa distorção da seguinte forma: (i) A excitação ótima é selecionada do dicionário (fixo ou adaptativo). A ela é aplicado o filtro de síntese para compor o sinal reconstituído; (ii) É calculada a energia de ambos os sinais, original e reconstituído; (iii) O ganho é então calculado por:

$$G^2 = \frac{\sum_{n=0}^{\frac{N}{4}-1} s^2[n]}{\sum_{n=0}^{\frac{N}{4}-1} \hat{s}^2[n]} \qquad (3.14)$$

onde $N/4$ é o número de amostras em cada sub-janela do sinal. Este processo é realizado para estimar o ganho correspondente à excitação de ambos os dicionário.

3.7 Resposta à entrada zero

Durante o processamento de cada sub-janela de 5 ms, o sinal reconstituído é comparado a um sinal alvo. Ele é obtido quando o sinal original é “janelado”, aplicado ao filtro perceptual $W(z)$ e do resultado é subtraída a resposta à entrada zero relativa à sub-janela anterior. Isto é feito, pois durante o processo de decodificação, as primeiras amostras de cada nova sub-janela são preditas a partir de algumas amostras da sub-janela anterior como mostrado na Equação (2.4). Desse modo, no codificador, durante a determinação dos parâmetros do sinal de cada sub-janela, a informação relativa à sub-janela anterior deve ser extraída.

3.8 Bitstream

A seqüência de bits que contém informações sobre o sinal de voz original será chamada *bitstream.* O *bitstream* é utilizado para transmitir as informações do sinal de voz do codificador para o decodificador. Ele é composto pelos coeficientes DLSF quantizados, pelos índices das excitações ótimas e pelos ganhos quantizados dos dois dicionários. Na Tabela 3.2, observa-se a distribuição de bits no *bitstream*.

Tabela 3. 2: Composição do *bitstream*.

	Distribuição de bits	**Número total de bits gerados em cada janela de 20 ms**
Dicionário Adaptativo (1024 excitações)	10 bits por sub-janela	40
Ganho referente ao Dicionário Adaptativo	6 bits por sub-janela	24
Dicionário Fixo (256 excitações)	8 bits por sub-janela	32
Ganho referente ao Dicionário Fixo	6 bits por sub-janela	24
Coeficientes DLSF (10 coeficientes)	32 bits por janela	32
Bitstream		**152**

Como é gerado um *bitstream* composto de 152 bits quando uma janela de 20 ms é processada, obtém-se uma taxa de 7,6 kbits/s para este codificador.

3.9 Visão geral do processo de codificação

Na Figura 3.1, pode-se observar a estrutura CELP final proposta por [**1,2**]. De forma resumida, o processo funciona do seguinte modo:

1.O sinal de voz original $s[n]$ é dividido em janelas por $w_H[n]$. É gerado o sinal $s_w[n]$ de 20 ms;

2. São calculados os coeficientes LP $a_1,...,a_p$ referentes ao sinal $s_w[n]$. Eles são transformados em coeficientes LSF $w_1,...,w_p$, os quais são então quantizados;

3. Para cada sub-janela de 5 ms, os coeficientes LSF são interpolados e transformados novamente em coeficientes LP $\hat{a}_1,...,\hat{a}_p$;

4. Os filtros das Equações (3.12) e (3.13) são obtidos;

5. Os dois dicionários são filtrados pelo filtro de síntese da Equação (3.13);

6. O sinal $s_w[n]$ é dividido em sub-janelas e cada uma é filtrada pelo filtro de ponderação perceptual da Equação (3.12). De cada sub-janela filtrada é subtraída a resposta à entrada zero relativa à sub-janela anterior. O sinal de 5 ms resultante é chamado de sinal alvo ($s_J[n]$).
A resposta à entrada zero pode ser entendida como o resultado obtido na saída do filtro de síntese no momento em que não há mais excitação presente em sua entrada;

7. É realizada uma busca no dicionário adaptativo pelo índice da excitação ótima, através da minimização do erro perceptual entre o sinal reconstituído $\hat{s}_a[n]$ e o sinal alvo $s_J[n]$; Neste momento, G_a é utilizado com valor 1;

8. Quando a excitação ótima é encontrada, o ganho G_a correspondente a ela é calculado pela Equação (3.14). O sinal $\hat{s}_a[n]$ é então multiplicado por este ganho;

9. O sinal $s_J[n]$ é atualizado, através da subtração do sinal $\hat{s}_a[n]$ obtido em 8;

10. Com o sinal $s_J[n]$ atualizado é realizada a busca pelo índice da excitação ótima no dicionário fixo. Ela é feita através da minimização do erro perceptual entre o sinal reconstituído $\hat{s}_f[n]$ e o sinal $s_J[n]$ atualizado; Neste momento, G_f é utilizado com valor 1;

11. Quando a excitação ótima é encontrada, o ganho G_f correspondente a ela é calculado pela Equação (3.14);
O sinal $\hat{s}_f[n]$ é então multiplicado por este ganho;

12. O sinal reconstituído resultante $\hat{s}[n]$ é calculado por: $\hat{s}[n] = \hat{s}_a[n] + \hat{s}_f[n]$.

É importante notar que o dicionário adaptativo do decodificador também apresenta todas as amostras com valor zero no início do processamento. Assim como no codificador, ele é atualizada a cada iteração.

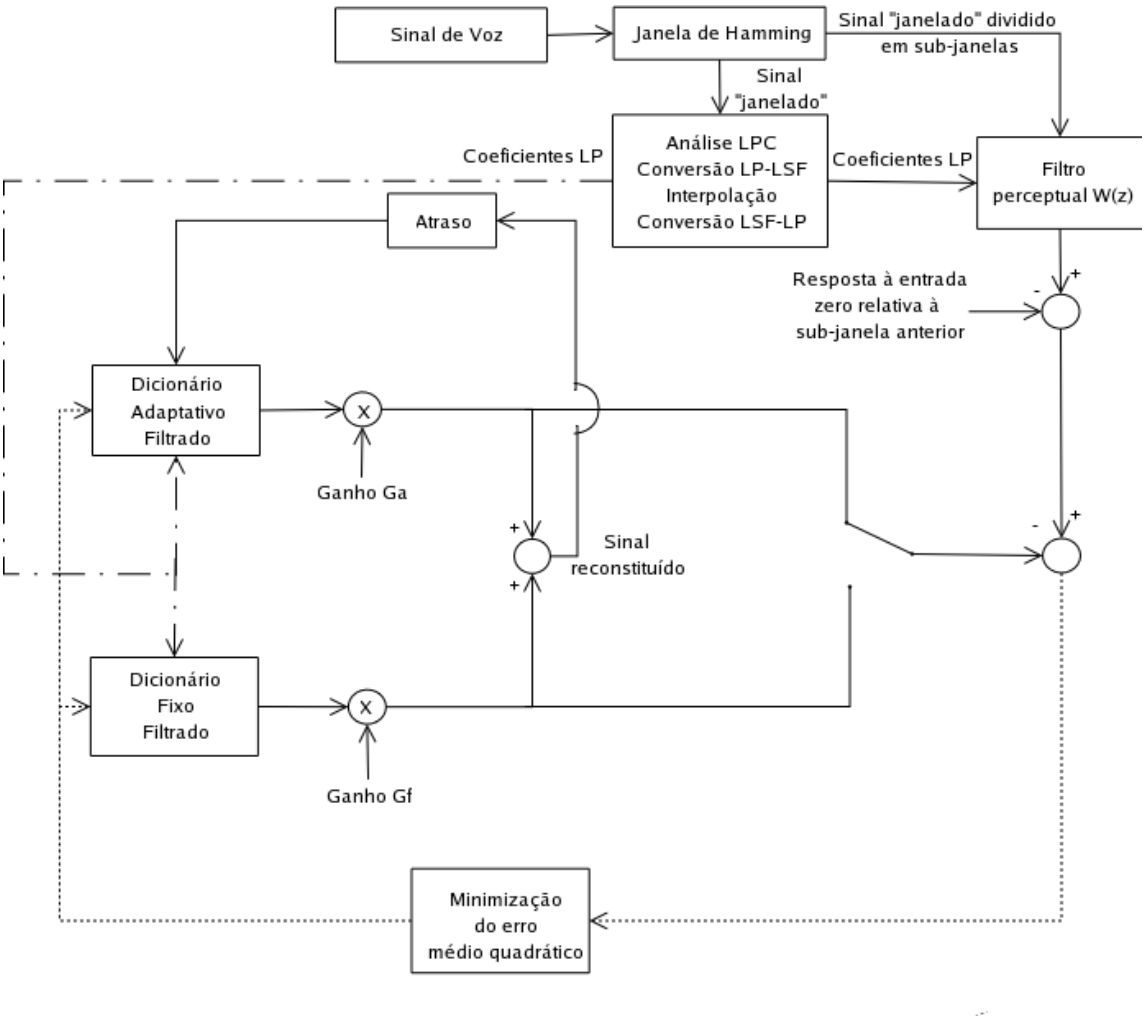


Figura 3. 1: Codificador CELP implementado em [1,2].

3.10 Comparação entre sinal original e reconstituído

A frase “A casa foi vendida sem pressa” foi gravada e processada por este codificador CELP para possibilitar uma comparação entre o sinal de voz original e o reconstituído. Na Figura 3.2, pode-se observar a evolução temporal e o espectro de potência de uma janela deste sinal de voz. Esta é composta por um trecho de sinal sonoro, sendo possível observar o comportamento periódico do mesmo na evolução temporal.

Através do espectro de potência observa-se que, em baixas freqüências, o espectro dos dois sinais, original e reconstituído, coincidem. Conforme a freqüência se eleva, observa-se que o espectro do sinal reconstituído (linha pontilhada) não acompanha tão perfeitamente o do sinal original (linha cheia). Isto ocorre, em parte, devido à forma como foi realizada a quantização dos coeficientes LSF. As freqüências mais baixas foram quantizadas com um maior número de bits como colocado na Tabela 3.1. No entanto, como foi visto, o sistema auditivo humano não é muito sensível a variações em freqüências elevadas, o que faz com que este sinal reconstituído represente uma boa aproximação do sinal original.

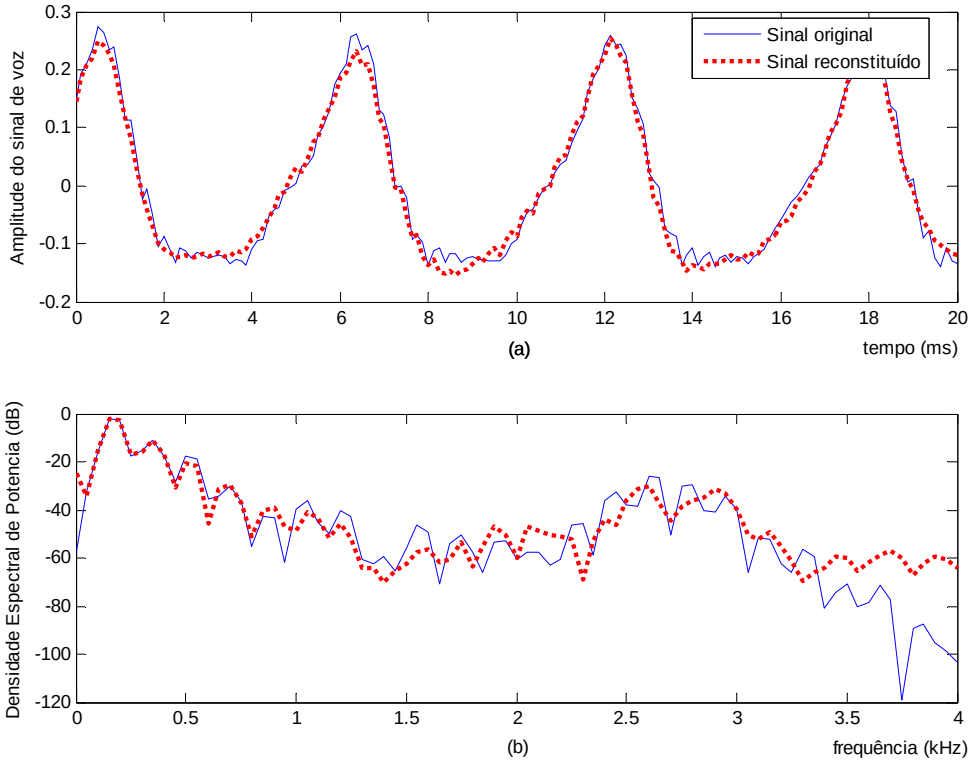


Figura 3. 2: (a) Trecho de 20 ms de um sinal de voz codificado pela implementação CELP de [1,2] e sua (b) densidade espectral de potência.

3.11 Conclusão

O codificador CELP apresentado neste capítulo possui um elevado ganho quando considerada a taxa de bits. Como pode ser visto na Figura 3.3, ele mantém um bom compromisso entre a taxa de bits e a qualidade do sinal reconstituído quando comparado aos padrões de codificação também apresentados no gráfico.

Podemos também observar a grande contribuição de [2] para esta implementação. Em [2], é realizada uma implementação em tempo real do codificador. Para tanto, como foi visto, nela são propostas alterações na estrutura do codificador CELP com o objetivo de reduzir o tempo de processamento. Esse, quando elevado,

automaticamente introduz um atraso no sinal de voz reconstituído. Isto é muito prejudicial para a comunicação em tempo real, uma vez que um atraso significativo afeta a interatividade dos locutores.

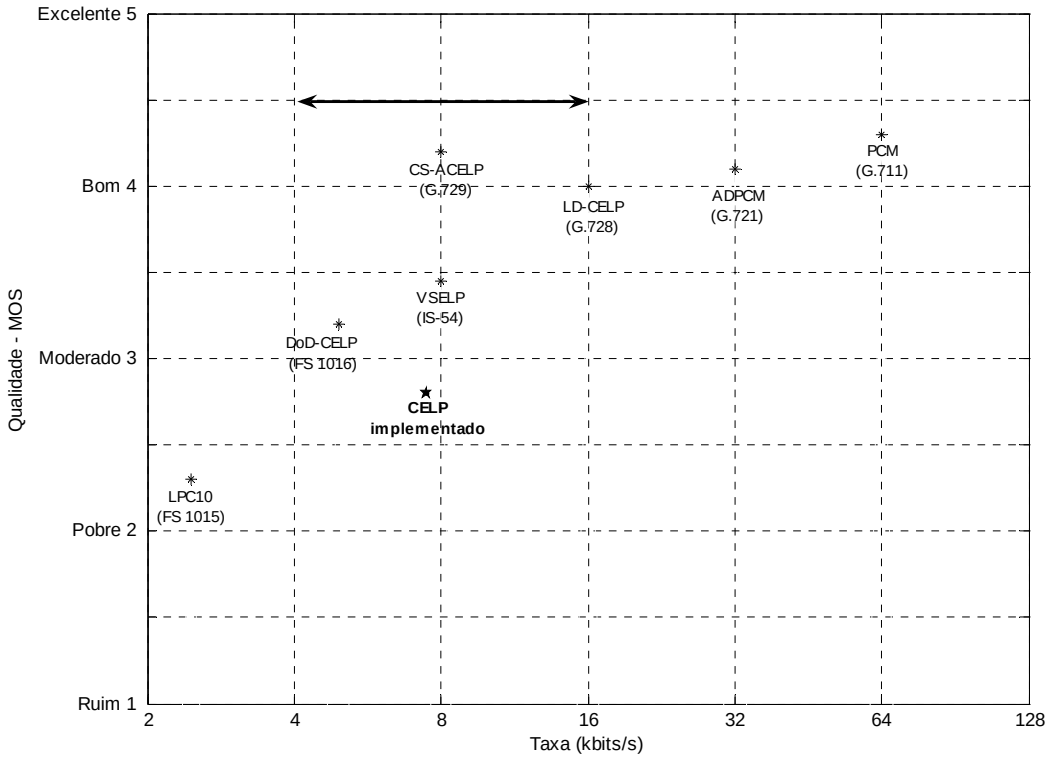


Figura 3. 3: Taxa de bits x qualidade de alguns algoritmos de codificação da voz (algoritmos padronizados e algoritmo de [**1,2**]). A seta evidencia a região onde as pesquisas estão concentradas atualmente: codificadores com baixas taxas de bits e elevada qualidade da voz codificada.

Capítulo 4

Qualidade Objetiva do Sinal de Voz Codificado

4.1 Introdução

Neste capítulo, será realizada uma análise bit-a-bit do *bitstream* obtido com o codificador CELP descrito no capítulo anterior. Dessa forma, será possível analisar a importância de cada bit na qualidade do sinal reconstituído e identificar aqueles que são mais importantes. Esta análise inicial será feita de forma objetiva.

Na Seção 4.2, são apresentadas as medidas objetivas de qualidades utilizadas na análise de dados; as Seções 4.3 e 4.4 abordam o procedimento experimental adotado para identificar o grupo crítico de bits do *bitstream* quando considerada a qualidade do sinal de voz e também mostram resultados; a Seção 4.5 apresenta a sub-divisão do grupo de bits restantes em função da sua relevância na qualidade do sinal de voz; na Seção 4.6 é realizada uma verificação sobre o comportamento de cada grupo identificado à medida que o sinal é degradado; já na Seção 4.7 é feita uma breve conclusão.

4.2 Medidas objetivas de qualidade

O emprego da análise subjetiva para avaliar a qualidade de codificadores de voz fornece resultados mais autênticos. Porém, este é um procedimento muito custoso para ser utilizado no início da análise de dados. Assim, neste capítulo, será dado início à análise objetiva. Com isto, torna-se possível repetir o procedimento experimental tantas vezes quanto necessário sem haver preocupação com a análise dos dados, uma vez que esta, envolve apenas a re-execução de algoritmos implementados computacionalmente. No próximo capítulo, será realizada uma análise subjetiva para comprovar os resultados aqui alcançados.

Para a análise objetiva da qualidade foram utilizadas as medidas: Razão Sinal-Ruído Segmentada Perceptual, Distância de Itakura, Distância Cepstral e o algoritmo PESQ. A seguir serão apresentados mais detalhes de cada uma.

4.2.1 Razão Sinal-Ruído Segmentada Perceptual

A medida da razão sinal-ruído (*Signa-to-Noise Ratio - SNR*) é pobre em informação. E por isso, em geral, não é muito usada na medição da qualidade do sinal reconstituído. Como alternativa, utiliza-se a Razão Sinal-Ruído Segmentada Perceptual (RSRSP). Ela baseia-se na divisão do sinal em janelas e é calculada por:

$$RSRSP = \frac{1}{M}\sum_{i=0}^{M-1} 10\log\left(\frac{\sum_{n=0}^{N-1} s^2[n+iN]}{\sum_{n=0}^{N-1} e_w^2[n+iN]}\right) \quad (4.1)$$

onde $s[n]$ e $\hat{s}[n]$ são, respectivamente, os sinais de voz original e reconstituído; $e_w[n]$ é o sinal de erro perceptual obtido quando o sinal de erro $e[n] = s[n] - \hat{s}[n]$ é aplicado ao filtro de ponderação perceptual $W(z) = A(z)/A(z/\gamma)$; N é o número de amostras que constitui cada janela do sinal e M é o número de janelas do sinal. Como especificado no capítulo 2, o cálculo dos coeficientes LP de $A(z)$ foi feito pelo método de autocorrelação de 10ª ordem e a janela usada tem tamanho de 20 ms resultando em $N = 160$.

A RSRSP apresenta uma boa correlação com a qualidade subjetiva do sinal reconstituído devido ao fato de utilizar nos cálculos o erro perceptual $e_w[n]$ ao invés do erro $e[n]$ utilizado no SNR.

4.2.2 Distância de Itakura

A distância de Itakura (D_I) procura analisar quão similar são dois espectros de potência. Ela é um dos métodos mais popularmente usados para determinar a distância entre coeficientes LP, a_i e $\hat{a}_i$. Esses coeficientes são obtidos através da análise LPC do sinais de voz original ($s[n]$) e reconstituído ($\hat{s}[n]$), respectivamente.

Quando um sinal de voz $s[n]$ é aplicado ao filtro de análise $A(z)$, pode-se calcular a energia E do erro de predição linear por:

$$E = E[e^2[n]] = E\left[\left(s[n] - \sum_{i=1}^{p} a_i s[n-i]\right)^2\right] \quad (4.2)$$

onde a_i representa os coeficientes LP. A partir do desenvolvimento da Equação (4.2) obtém-se:

$$E = aR_s a^T \quad (4.3)$$

onde R_s é a matriz de autocorrelação do sinal $s[n]$ e $a = [1 \quad -a_1 \quad \ldots \quad -a_p]$. Uma equação equivalente a (4.3) pode ser obtida se considerarmos que o sinal de voz $s[n]$ é aplicado ao filtro de análise $\hat{A}(z)$ formado pelos coeficientes LP $\hat{a}_i$:

$$\hat{E} = \hat{a} R_s \hat{a}^T \tag{4. 4}$$

onde R_s é a matriz de autocorrelação do sinal $s[n]$ e $\hat{a} = \begin{bmatrix} 1 & -\hat{a}_1 & \dots & -\hat{a}_p \end{bmatrix}$.

Através da comparação dos dois escalares E e $\hat{E}$ com relação ao erro de predição, obtém-se:

$$\hat{E} \geq E \tag{4. 5}$$

pois E representa o erro de predição quadrático mínimo, já que corresponde aos coeficientes LP realmente obtidos a partir de $s[n]$. Finalmente, a medida de quão "distante" estão os coeficientes LP $\hat{a}_i$ e a_i é dada por:

$$D_I(a, \hat{a}) = 10 \log \frac{\hat{a} R_s \hat{a}^T}{a R_s a^T} \tag{4. 6}$$

Esta medida será sempre positiva em função da condição definida pela Equação (4.5). No entanto, ela não apresenta a desejada propriedade de simetria já que,

$$D_I(a, \hat{a}) \neq D_I(\hat{a}, a) \tag{4. 7}$$

Esta simetria pode ser obtida através da seguinte combinação:

$$\begin{aligned} D_I(a, \hat{a}) &= \frac{1}{2}\left[D_I(a, \hat{a}) + D_I(\hat{a}, a)\right] \\ &= \frac{1}{2}\left[10 \log \frac{\hat{a} R_s \hat{a}^T}{a R_s a^T} + 10 \log \frac{a R_{\hat{s}} a^T}{\hat{a} R_{\hat{s}} \hat{a}^T}\right] \end{aligned} \tag{4. 8}$$

onde $R_{\hat{s}}$ é a matriz de autocorrelação do sinal $\hat{s}[n]$.

Pode-se observar que a medida dada pela Equação (4.8) também apresenta a seguinte propriedade: se o espectro de potência do sinal $\hat{s}[n]$ for igual ao do sinal $s[n]$, a distância resultante será zero.

Neste trabalho, os coeficientes LP de $A(z)$ e $\hat{A}(z)$ foram obtidos pelo método de autocorrelação de 10ª ordem.

4.2.3 Distância Cepstral

A distância Cepstral (D_C) calcula a diferença entre o formato do espectro de sinais de voz original $s[n]$ e reconstituído $\hat{s}[n]$. Ela é dada por:

$$D_C = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{q}(c_i - \hat{c}_i)^2} \tag{4. 9}$$

onde c_i são os coeficientes cepstrais obtidos a partir do sinal $s[n]$ e $\hat{c}_i$ são os coeficientes cepstrais obtidos a partir do sinal $\hat{s}[n]$. No entanto, é possível obter os coeficientes cepstrais a partir dos coeficientes LP da seguinte forma:

$$\begin{aligned} c_1 &= a_1 \\ c_i &= a_i + \sum_{k-1}^{i-1}\frac{k}{i}c_k a_{i-k} \quad ,1 \le i \le p \\ c_i &= \sum_{k=1}^{p}\left(1-\frac{k}{i}\right)c_{i-k}a_k \quad , p < i \le p' \end{aligned} \tag{4. 10}$$

onde $a_1,...,a_p$ são os coeficientes LP e $c_1,...,c_{p'}$ são os coeficientes cepstrais.

4.2.4 PESQ

O PESQ (*Perceptual Evaluation of Speech Quality*) é um método objetivo de medida perceptual da qualidade da voz em sistemas de telecomunicações. A recomendação da ITU-T para este algoritmo é o P.862 de Fevereiro de 2001. Este algoritmo se transformou no padrão mais utilizado para medir a qualidade da voz em redes VoIP (*Voice over Internet Protocol*). No entanto, a utilização do PESQ é mais abrangente. Ele pode ser aplicado para testar, por exemplo, a qualidade do sinal reconstituído em um codificador.

O algoritmo do PESQ faz uma comparação, através de um modelo perceptual, entre os sinais de voz original e reconstituído do sistema sob avaliação. Inicialmente, os dois sinais são ajustados para um limiar de audição padrão. Em seguida, são aplicados a um filtro que formata os sinais de acordo com as características da banda telefônica. São então, alinhados um com relação ao outro no tempo e representados com relação à percepção do sistema auditivo humano. É tomada a diferença entre os sinais, a qual é chama perturbação. Essa é processada de forma a extrair a pontuação PESQ. Em [**8**], obtém-se uma descrição mais detalhada do algoritmo.

O resultado do PESQ é apresentado em uma escala similar a do MOS. Ela varia de -0,5 a 4,5. No entanto, a maior parte dos resultados encontra-se entre 1,0 e 4,5. Valores próximos a 4.5 indicam que a qualidade do sinal de voz é muito boa. À medida que a pontuação do PESQ reduz-se a qualidade do sinal torna-se mais baixa. Uma pontuação inferior a 2,0 já corresponde a um elevado nível de degradação do sinal e torna-se difícil a compreensão do mesmo. Segundo [**7**], o PESQ pode ser mapeado na escala MOS através da seguinte equação:

$$y = 0{,}999 + \frac{4{,}999 - 0{,}999}{1 + e^{-1{,}4945x + 4{,}6607}} \tag{4. 11}$$

onde x é o valor da pontuação PESQ e y o da pontuação MOS. A Equação (4.11) mapeia a escala PESQ -0,5 a 4,5 em 1,02 a 4,56, que está bastante próximo dos valores da escala MOS 1,0 a 5,0.

4.3 Primeira Etapa do Procedimento Experimental

Foram utilizados 200 sinais de voz na análise de dados, os quais serão denominados sinais originais. Todos foram gravados pelo mesmo locutor no formato WAV a uma freqüência de amostragem de 8000 Hz e com 16 bits por amostra.

Inicialmente, os 200 sinais originais foram codificados e decodificados pelo codificador CELP de [1,2]. Eles foram utilizados como a base de dados de referência para as medidas objetivas e serão chamados sinais reconstituídos. Em seguida, os 200 sinais originais foram, novamente, codificados e decodificados. Nesta execução, o *bitstream* foi corrompido da seguinte forma:

1. Define-se a porcentagem x do número de janelas em que o *bitstream* terá um bit b corrompido;
2. Define-se qual bit b do *bitstream* que será corrompido. Onde b assume valores de 1 a 152;
3. Inicia-se a codificação/decodificação do sinal original, sendo que o bit b escolhido é corrompido em x % das janelas.

Este processo foi executado até que os 152 bits do *bitstream* fossem corrompidos em todos os 200 sinais originais, o que resultou em 152x200 = 30400 sinais corrompidos.

Na análise objetiva de dados, foram utilizadas: a Razão Sinal-Ruído Segmentada Perceptual, a Distância de Itakura, a Distância Cepstral e o algoritmo PESQ. Como cada sinal original deu origem a 152 sinais corrompidos, o mesmo sinal reconstituído será utilizado como sinal de referência para as medidas objetivas desses 152 sinais corrompidos. Nas medidas, o sinal reconstituído foi usado como o sinal de referência $s[n]$ e o sinal corrompido foi utilizado como o sinal $\hat{s}[n]$. Na Figura 4.1, observa-se um diagrama do procedimento descrito.

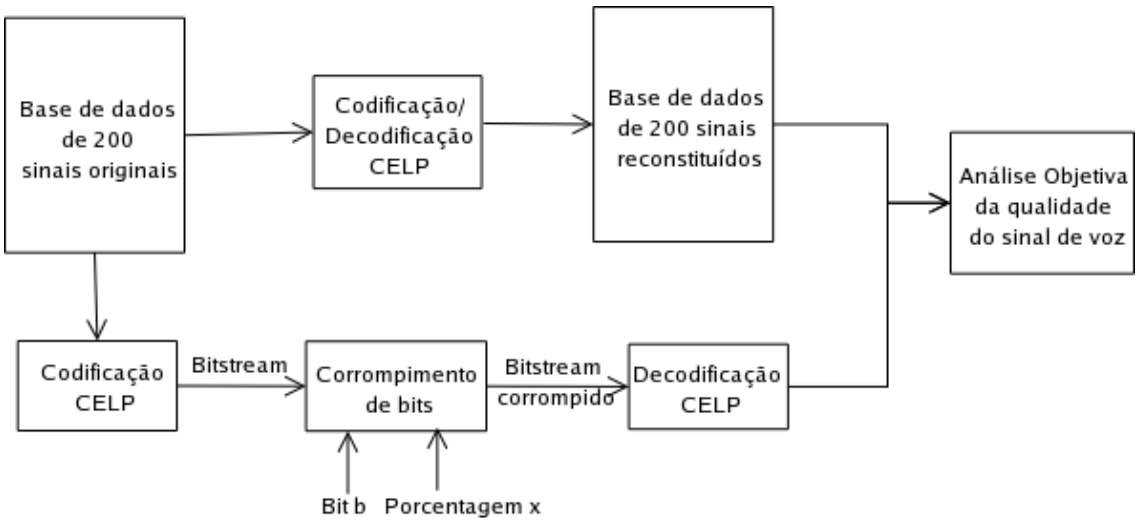


Figura 4. 1: Diagrama da primeira etapa do procedimento experimental.

4.3.1 Resultados

Os resultados serão apresentados em histogramas. Em cada um será representada uma medida de distância. No eixo vertical, estão apresentadas as médias dos valores obtidos em cada medida objetiva para cada uma das 200 frases. No eixo horizontal estão representados os 152 bits do *bitstream* organizados conforme descrito na Tabela 4.1.

Tabela 4. 1: Distribuição de bits no eixo horizontal do histograma.

Eixo horizontal do histograma	***Bitstream***				
	10 Coeficientes DLSF	4 Índices referentes ao Dicionário Adaptativo	4 Índices referentes ao Dicionário Fixo	4 Ganhos referentes ao Dicionário Adaptativo	4 Ganhos referentes ao Dicionário Fixo
	32 bits	40 bits	32 bits	24 bits	24 bits
	Índice de 1 a 32	Índice de 33 a 72	Índice de 73 a 104	Índice de 105 a 128	Índice de 129 a 152

A escolha da porcentagem x de janelas a serem corrompidas foi realizada através de testes preliminares. Para esta primeira etapa foi escolhida a porcentagem de 50%, pois com esta obtêm-se sinais de qualidade intermediária. Esta etapa teve como objetivo identificar conjuntos iniciais de bits considerados críticos para a qualidade do sinal de voz.

Razão Sinal-Ruído Segmentada Perceptual

Através da Equação (4.1) observa-se que, quanto menor o valor fornecido pela RSRSP, pior é a qualidade do sinal reconstituído. Dessa forma, foram considerados bits críticos aqueles que quando corrompidos em 50% das janelas apresentaram uma RSRSP negativa. Esta classificação pode ser observada na Figura 4.2.

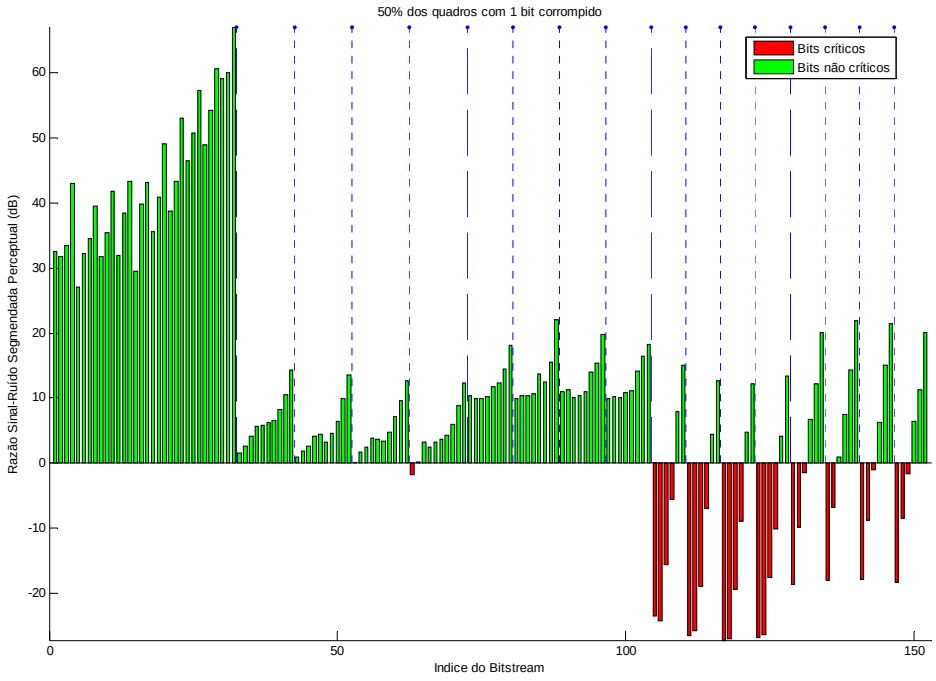


Figura 4. 2: Razão Sinal-Ruído Segmentada Perceptual calculada quando um bit do *bitstream* é corrompido em 50% das janelas.

Distância de Itakura

Quanto maior o valor fornecido pela D_I , pior é a qualidade do sinal reconstituído. Dessa forma, calculou-se a média dos resultados obtidos quando cada um dos 152 bits foi corrompido. Foram considerados bits críticos aqueles que quando corrompidos em 50% das janelas apresentaram uma D_I superior a média. Esta classificação pode ser observada na Figura 4.3.

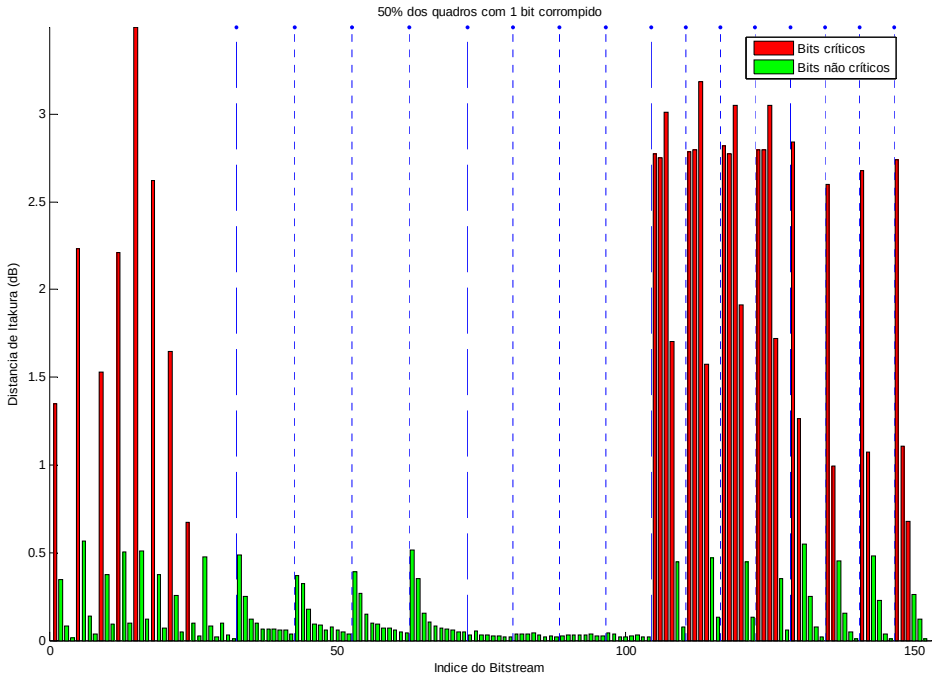


Figura 4. 3: Distância de Itakura calculada quando um bit do *bitstream* é corrompido em 50% das janelas.

Distância Cepstral

Quanto maior o valor fornecido pela D_C, pior é a qualidade do sinal reconstituído como pode ser visto na Sub-seção 4.2.3. Dessa forma, calculou-se a média dos resultados obtidos quando cada um dos 152 bits foi corrompido. Foram considerados bits críticos aqueles que quando corrompidos em 50% das janelas apresentaram uma D_C superior a média acrescida de a. O valor de a foi escolhido de forma empírica de modo que, após definido o limiar, não fosse considerado crítico um grande grupo de bits. Existem bits que possuem grande importância para a qualidade do sinal de voz, o que não significa que sejam críticos, podendo-se classificá-los em outro grupo. Isso será visto em maiores detalhes mais adiante neste capítulo. A classificação dos bits críticos pode ser observada na Figura 4.4.

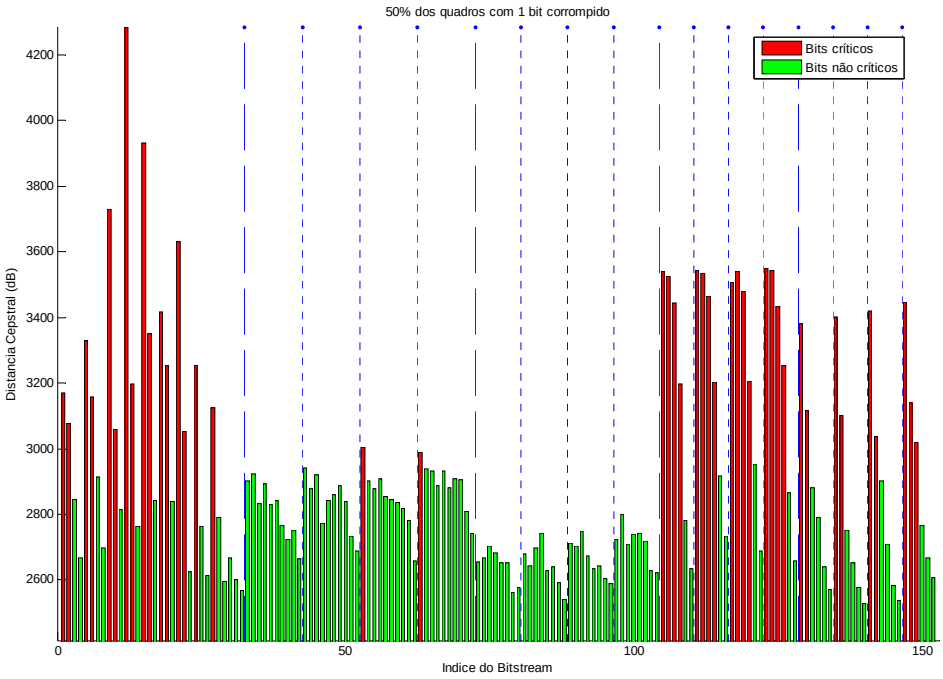


Figura 4. 4: Distância Cepstral calculada quando um bit do *bitstream* é corrompido em 50% das janelas.

PESQ

Conforme visto na Sub-seção 4.2.4, um sinal que resulta em uma pontuação PESQ inferior a 2 já apresenta um elevado nível de degradação. Dessa forma, foram considerados bits críticos aqueles que quando corrompidos em 50% das janelas apresentaram uma pontuação inferior a essa. Esta classificação pode ser observada na Figura 4.5.

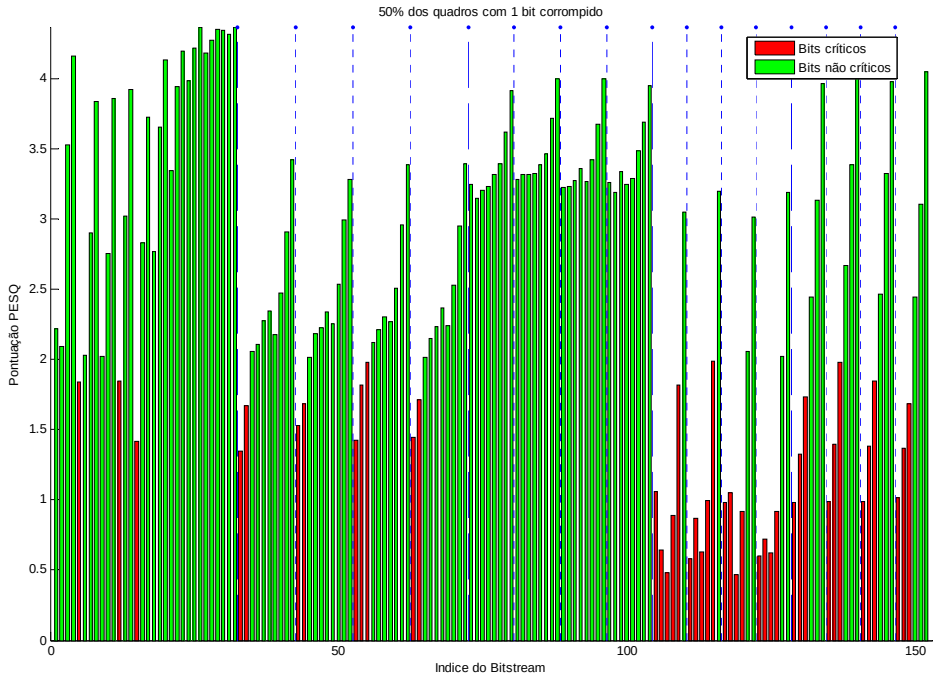


Figura 4. 5: Pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 50% das janelas.

Após esta análise inicial, fica claro porque não foi utilizada uma porcentagem x elevada. Esta prejudicaria os objetivos desta análise uma vez que, com a corrupção elevada, a tendência seria a qualidade cair de uma forma geral, dificultando a identificação de bits críticos para a qualidade do sinal de voz. O mesmo vale para a corrupção baixa. Esta pode resultar em sinais de elevada qualidade tornando a identificação de bits mais críticos bastante difícil devido ao comportamento homogêneo de todos eles. Por isto, optou-se pela escolha de uma porcentagem de corrupção intermediária.

4.4 Segunda Etapa do Procedimento Experimental

Foi realizada a união dos resultados fornecidos pelas quatro medidas de distância durante a primeira etapa do procedimento experimental. Dessa forma, foi possível obter um conjunto inicial de bits críticos quando considera a qualidade do sinal reconstituído. Nesta segunda etapa, foi realizada a validação deste conjunto identificado. O procedimento adotado foi semelhante ao da primeira etapa, sendo que, nesta, os bits foram corrompidos também em 10, 25 e 70% das janelas. A escolha destas porcentagens foi realizada através de testes preliminares. O objetivo era obter-se uma evolução do comportamento da qualidade dos sinais de voz à medida que os mesmos fossem corrompidos. Buscou-se, então, partir de uma situação com baixa corrupção de janelas, até uma situação de grande perda da qualidade, porém não extrema. Durante estes teste preliminares foram utilizadas as medidas descritas na Seção 4.2.

4.4.1 Resultados

Nas Figuras 4.6, 4.7, 4.8, 4.9, podem ser observados os resultados obtidos. Foram excluídos do conjunto de bits críticos aqueles que mantiveram uma elevada pontuação PESQ mesmo quando a porcentagem de janelas corrompidas aumentou.

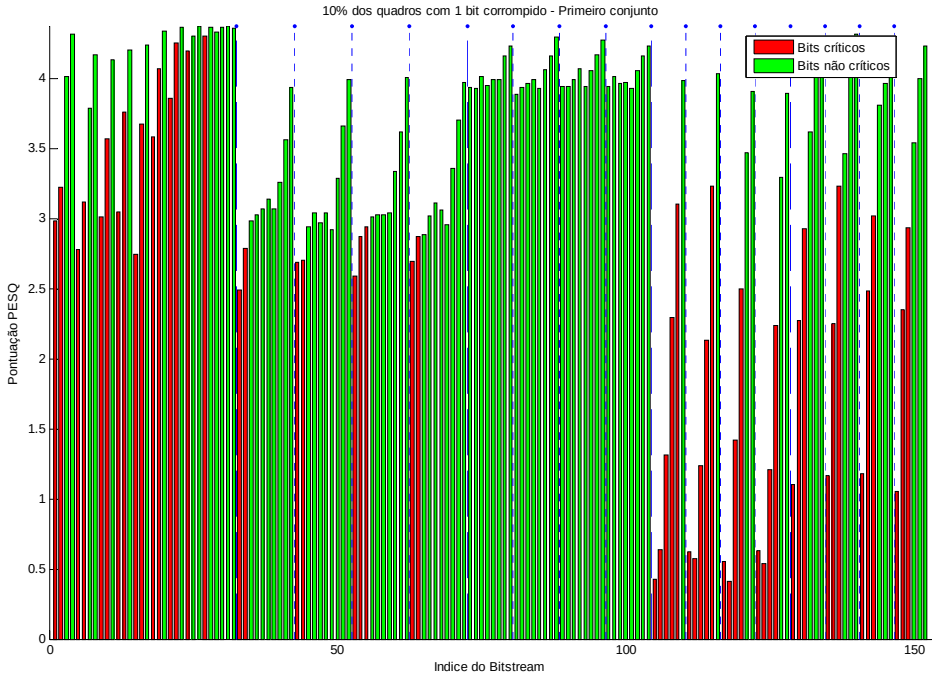


Figura 4. 6: Pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 10% das janelas.

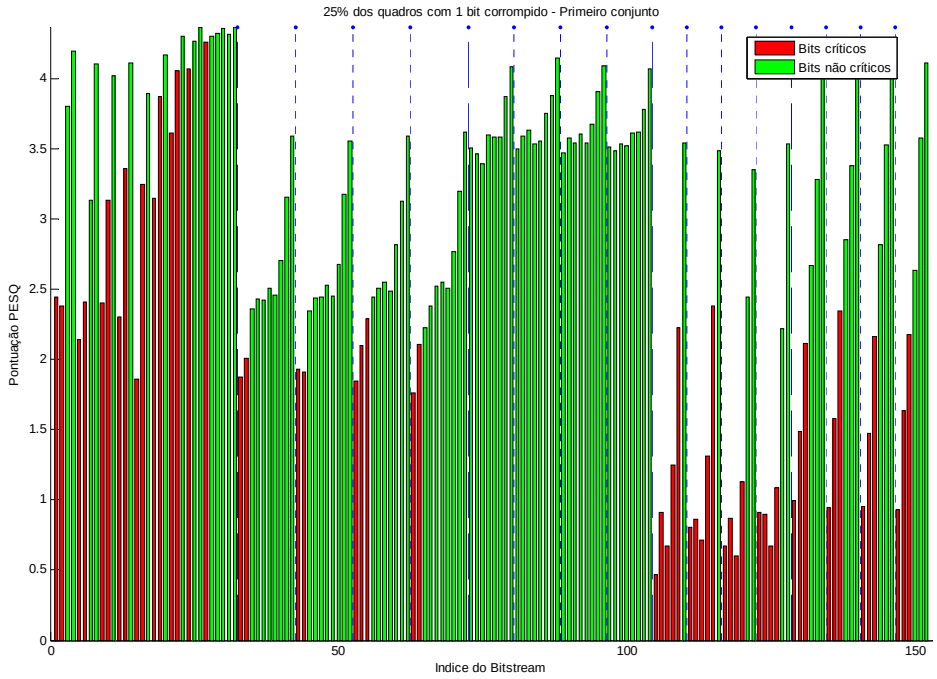


Figura 4. 7: Pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 25% das janelas.

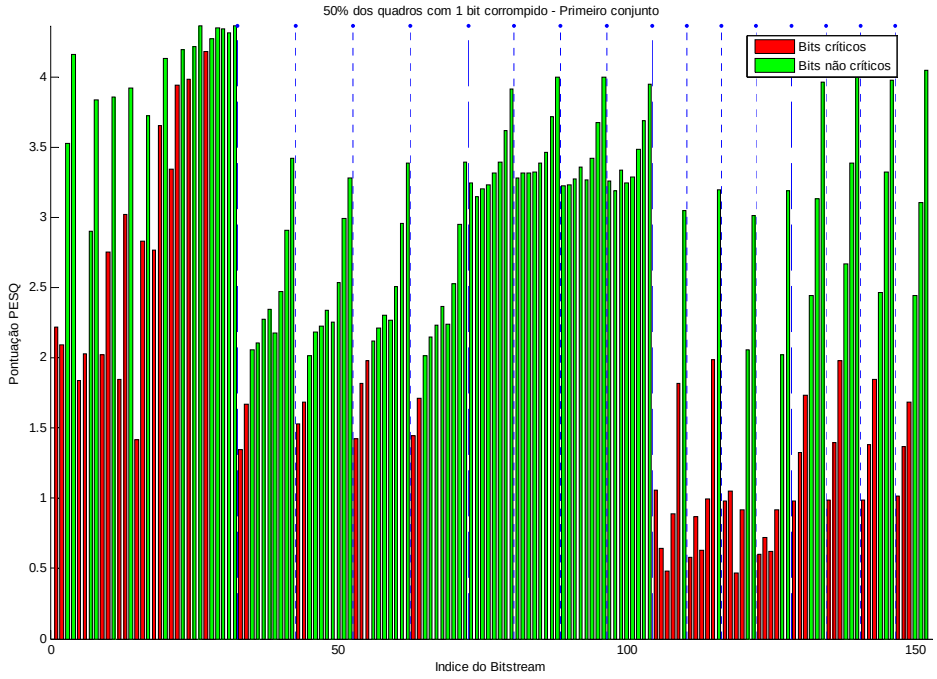


Figura 4. 8: Pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 50% das janelas.

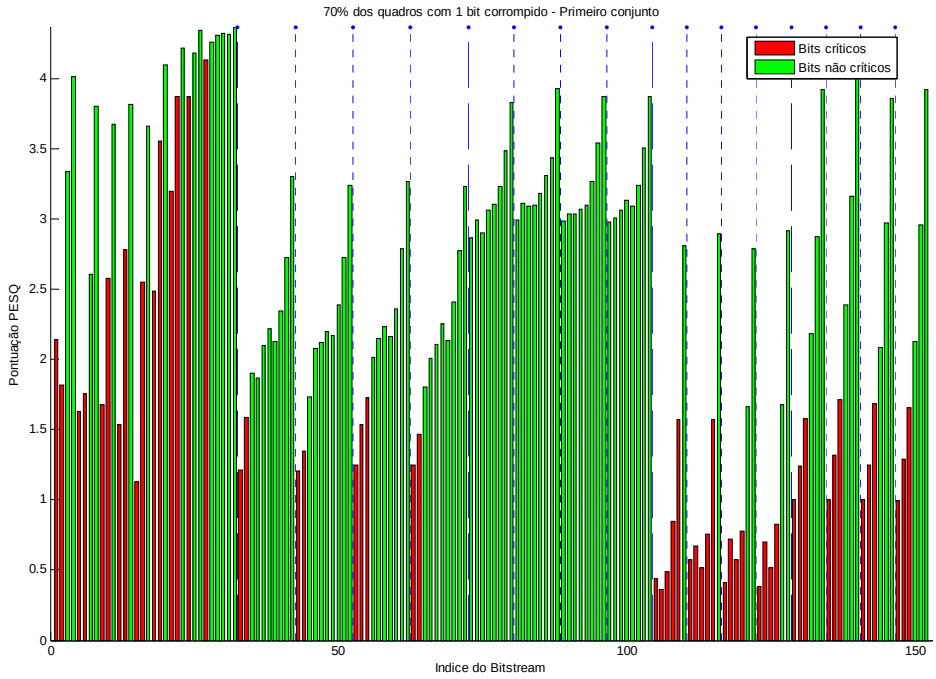


Figura 4. 9: Pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 70% das janelas.

Através desta análise, o conjunto final de bits críticos foi reduzido a 49 bits. Estes bits estão evidenciados, em vermelho, na Figura 4.10.

4.5 Terceira Etapa do Procedimento Experimental

Ao final da segunda etapa obteve-se um conjunto de bits que não é crítico para a qualidade do sinal de voz, mas que também contém importantes informações sobre o mesmo. Nesta terceira etapa do procedimento experimental, foi realizada uma divisão deste conjunto de bits não críticos. Este procedimento teve como objetivo identificar sub-conjuntos de bits quando considerada sua relevância na qualidade do sinal reconstituído e identificar as diferenças no comportamento deles. O procedimento adotado foi semelhante ao da primeira etapa, sendo que, nesta os bits foram corrompidos em 70% das janelas. Utilizou-se uma porcentagem de corrupção maior, pois como estamos analisando bits não críticos espera-se que estes sejam mais robustos à corrupção e assim, apenas uma corrupção elevada permite identificar o comportamento dos mesmos.

4.5.1 Resultados

A análise da terceira etapa foi realizada através do PESQ, pois essa medida fornecer um resultado mais fácil de ser interpretado. Consideraram-se diferentes limiares para o agrupamento dos bits. Estes limiares foram definidos empiricamente, de forma que os bits com pontuação PESQ entre dois limiares apresentassem um

comportamento semelhante. Dessa forma, os bits foram divididos em tantos grupos quanto foi possível identificar.

Após esta etapa, foram identificados 4 grupos distintos de bits:

1. Grupo 1: É o conjunto de bits críticos identificados através da primeira e segunda etapas;

2. Grupo 2: É formado pelos bits não críticos que quando corrompidos em 70% das janelas apresentaram um sinal reconstituído com pontuação PESQ inferior a 3;

3. Grupo 3: É formado pelos bits não críticos que quando corrompidos em 70% das janelas apresentaram um sinal reconstituído com pontuação PESQ entre 3 e 3,5;

4. Grupo 4: É formado pelos bits não críticos que quando corrompidos em 70% das janelas apresentaram um sinal reconstituído com pontuação PESQ superior a 3,5.

Esses grupos do *bitstream* podem ser observados na Figura 4.10 e na Tabela 4.2.

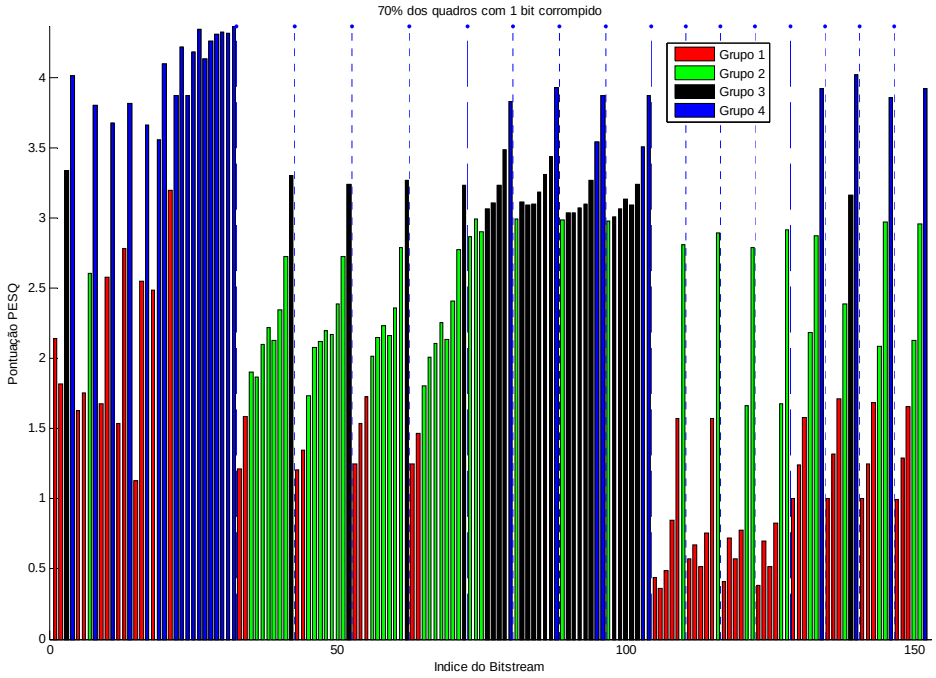


Figura 4. 10: Identificação dos 4 grupos de bits através da pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 70% das janelas.

Tabela 4. 2: Identificação dos bits de cada grupo.

		Bits
Grupo 1	Coeficientes DLSF	1,2,5,6,9,10,12,13,15,16,18,21
	Índices do Dicionário Adaptativo	33,34,43,44,53,54,55,63,64
	Ganhos do Dicionário Adaptativo	105,106,107,108,109,111,112,113,114,115,117,118,119,120,123,124,125,126
	Ganhos do Dicionário Fixo	129,130,131, 135,136,137, 141,142,143,147,148,149
Grupo 2	Coeficientes DLSF	7
	Índices do Dicionário Adaptativo	35,36,37,38,39,40,41,45,46,47,48,49,50,51,56,57,58,59,60,61,65,66,67,68,69,70,71
	Índices do Dicionário Fixo	73,74,75,81,89,97
	Ganhos do Dicionário Adaptativo	110,116,121,122,127,128
	Ganhos do Dicionário Fixo	132,133,138,144,145,150,151
Grupo 3	Coeficientes DLSF	3
	Índices do Dicionário Adaptativo	42,52,62,72
	Índices do Dicionário Fixo	76,77,78,79,82,83,84,85,86,87,90,91,92,93,94,98,99,100,101,102
	Ganhos do Dicionário Fixo	134
Grupo 4	Coeficientes DLSF	4,8,11,14,17,19,20,22,23,24,25,26,27,28,29,30,31,32
	Índices do Dicionário Fixo	80,88,95,96,103
	Ganhos do Dicionário Fixo	134,140,146,152

4.6 Quarta Etapa do Procedimento Experimental

Através desta quarta etapa, teve-se como objetivo analisar o comportamento da qualidade do sinal reconstituído quando cada grupo de bits é corrompido. Para isso, o seguinte procedimento foi utilizado:

1. Define-se o grupo de bits a ser corrompido;

2. Define-se a probabilidade x de cada bit do grupo ser corrompido. Foram utilizadas probabilidades de 0 a 12%;

3. Inicia-se a codificação/decodificação do sinal original, sendo que, para cada janela todos os bits do grupo têm a mesma probabilidade x de ser corrompido.

Foram escolhidas probabilidades de 0 a 12%, para que fosse possível acompanhar a evolução de cada grupo de bits em função da degradação do sinal de voz. Além disso, em cada iteração da codificação/ decodificação desta etapa um maior porcentagem do sinal será corrompida. Assim, a utilização de probabilidades mais elevadas resultaria em sinais totalmente corrompidos e sem nenhuma "informação", o que somente mostraria que todos os resultados iriam convergir para o mesmo ponto em termos de qualidade.

Em um primeiro momento, os grupos 3 e 4 foram corrompidos conjuntamente, já que essa união resulta no conjunto de bits menos críticos para a qualidade do sinal de voz. Isto foi feito de forma a facilitar a análise dos resultados através do equilíbrio da quantidade de bits em cada grupo. A consideração destes grupos separadamente invalidaria a comparação com os resultados dos grupos 1 e 2. Estes últimos possuem um maior número de bits e automaticamente resultariam em uma maior degradação do sinal reconstituído quando consideradas as mesmas probabilidades de corrompimento dos bits para todos os grupos.

Posteriormente, o procedimento acima descrito foi repetido. Neste segundo momento, os grupo 3 e 4 foram analisados separadamente, afim de se extrair informações sobre a relevância deles para a qualidade do sinal.

4.6.1 Resultados

A análise de dados foi feita através do algoritmo do PESQ. Nas Figuras 4.11 e 4.12, podem-se observar os resultados obtidos com o mapeamento da pontuação PESQ para a pontuação MOS através da Equação (4.11). Na Figura 4.11, observa-se que o grupo 1 já atinge uma pontuação MOS_{PESQ} muito baixa quando cada bit deste tem uma probabilidade de 1% de ser corrompido em cada janela do sinal. Isto já evidencia que estes bits são críticos para a qualidade do sinal. Entretanto, para probabilidades superiores, observa-se que, o algoritmo do PESQ retorna uma pontuação superior, representando uma melhora da qualidade mesmo quando o sinal está mais corrompido. Este é um fator a ser explorado na análise subjetiva que será realizada no próximo capítulo.

O grupo 2 apresenta um comportamento intermediário como já era esperado. O conjunto constituído pelos grupos 3 e 4 apresenta os melhores resultados quando comparado aos outros dois. Mesmo quando cada bit deste conjunto tem uma probabilidade de 6% de ser corrompido, a pontuação MOS_{PESQ} permanece superior a 2. Isto confirma que este é o conjunto dos bits menos críticos para a qualidade do sinal de voz.

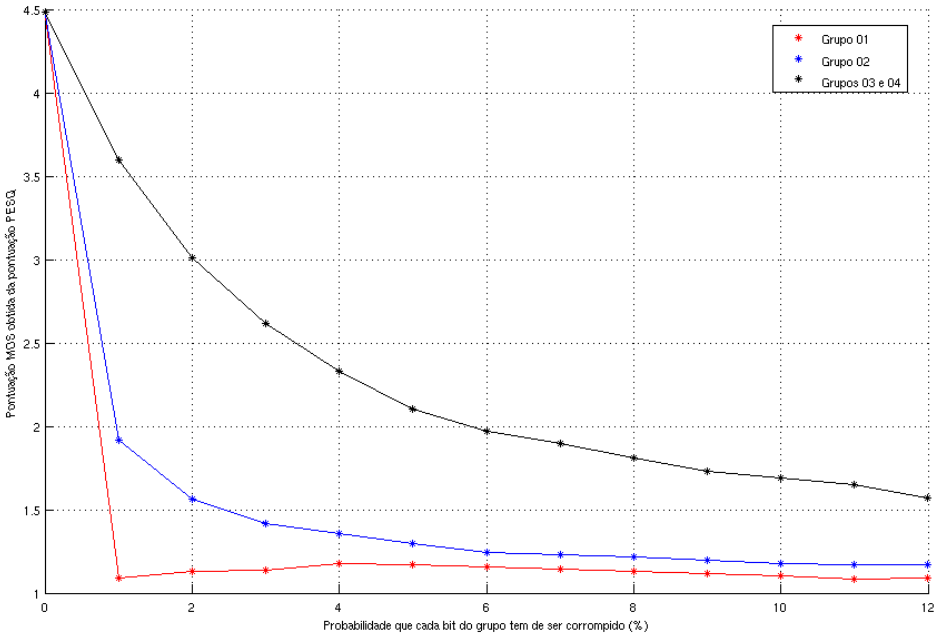


Figura 4.11: Relevância de cada grupo de bits para a qualidade do sinal de voz.

Através da Figura 4.12, pode-se observar melhor o comportamento dos grupos 3 e 4. Até o momento em que a probabilidade atinge 25%, o comportamento segue de acordo com o esperado, ou seja, o grupo 4 apresenta-se menos relevante para a qualidade do sinal. No entanto, a partir desta probabilidade, a situação se inverte. Provavelmente, isto ocorre, pois, como observado na Figura 4.11, o grupo 4 é constituído basicamente pelos bits que compõem os coeficientes DLSF. Estes coeficientes são os responsáveis por reproduzir o comportamento dos formantes do sinal. Já o grupo 3 é constituído basicamente pelos bits menos significativos dos índices do dicionário fixo.

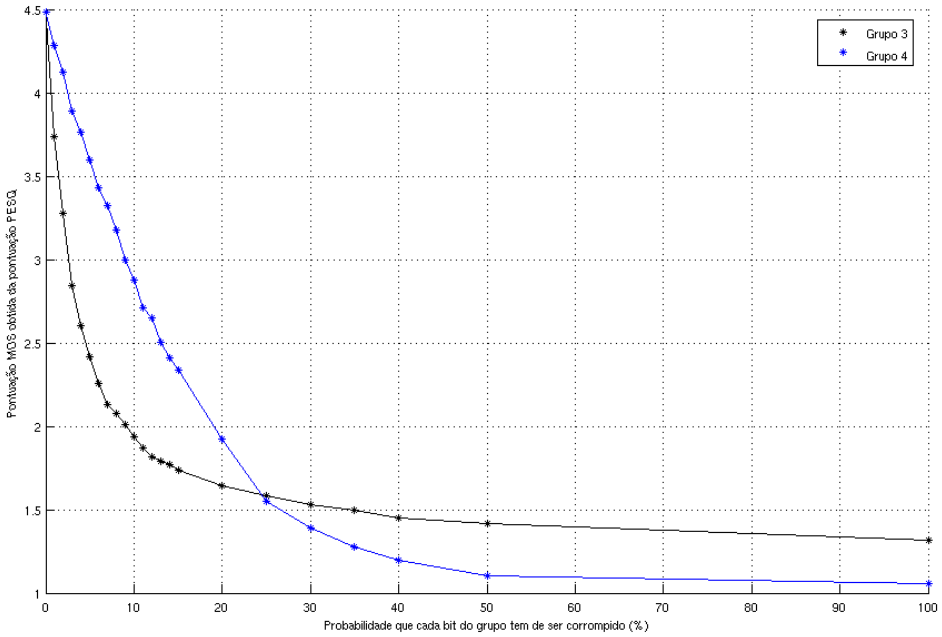


Figura 4. 12: Relevância dos grupos 3 e 4 para a qualidade do sinal de voz.

4.7 Conclusão

Com esta análise de dados, foi possível identificar a relevância de diferentes grupos de bits do *bitstream* quando considerada a qualidade do sinal reconstituído. Embora, estes resultados tenham sido apresentados apenas segundo aspectos objetivos, eles já evidenciam a possibilidade de realizar um mapeamento da importância de cada bit do *bitstream*. Com este conhecimento, torna-se possível melhorar, em um trabalho futuro, o desempenho do codificador CELP.

No próximo capítulo, serão feitas a comprovação e complementação destes resultados através da análise subjetiva da “importância” de cada bit.

Capítulo 5

Inteligibilidade do Sinal de Voz Codificado

5.1 Introdução

Neste capítulo, são apresentados os detalhes da análise subjetiva utilizada para avaliar a relevância de bits do *bitstream* quando se trata da inteligibilidade do sinal de voz reconstituído. Os resultados obtidos no capítulo anterior, através da análise objetiva, foram utilizados como ponto de partida.

A Seção 5.2 apresentada um resumo da metodologia de trabalho empregada; na Seção 5.3, é descrito todo o procedimento experimental utilizado na análise subjetiva da qualidade do sinal de voz; os resultados obtidos são apresentados na Seção 5.4 e a análise dos mesmos é realizada na Seção 5.5; já na Seção 5.6 é feita uma breve conclusão.

5.2 Medidas subjetivas de qualidade

A avaliação subjetiva da qualidade do sinal reconstituído é realizada segundo dois aspectos: a inteligibilidade e a naturalidade do mesmo. Esta forma de análise é a que conduz ao resultado mais "autêntico" sobre a qualidade do sinal reconstituído. No entanto, como este processo é muito custoso, justifica-se a realização de uma análise objetiva inicialmente. Dessa forma, após identificado o comportamento esperado, através da análise objetiva, procede-se à análise subjetiva.

Neste trabalho, foi utilizado o MOS (*Mean Opinion Score*) na avaliação subjetiva da qualidade do sinal de voz.

5.3 Procedimento Experimental

Assim como a análise objetiva, a subjetiva foi dividida em etapas. O objetivo da primeira etapa era identificar quais e quantos sinais deveriam ser utilizados nesta avaliação. Já a segunda etapa teve como objetivo obter um resultado final sobre a pontuação MOS para os diferentes grupos identificados no Capítulo 4.

5.4 Primeira Etapa da Análise Subjetiva

Durante a quarta etapa do procedimento experimental, descrita no Capítulo 4, obteve-se um conjunto de sinais reconstituídos com comprimento médio de 2 segundos. Em cada um deles, diferentes grupos de bits do *bitstream* foram corrompidos em porcentagens variadas durante a codificação/ decodificação. De forma a obter uma boa amostragem da relevância de cada bit do *bitstream* na qualidade do sinal de voz reconstituído, 92 destes

sinais corrompidos foram selecionados para a análise subjetiva. Como este tipo de análise envolve razoável tempo de preparação dos dados e aplicação dos testes, torna-se desejável que o conjunto total dos sinais não seja muito longo.

Os 92 sinais foram avaliados por 5 pessoas. Inicialmente, todas elas foram treinadas para poderem classificar os sinais em excelente, muito bom, razoável, pobre e ruim de acordo com o padrão definido na Tabela 5.1. Caso isto não fosse feito, cada pessoa utilizaria critérios próprios na classificação da qualidade dos sinais, o que invalidaria o resultado final.

Tabela 5. 1: Pontuação MOS.

Pontuação	Definição	Descrição
5	Excelente	Sinal de voz perfeito, gravado em local silencioso.
4	Muito Bom	Sinal de qualidade de chamada telefônica.
3	Razoável	Sinal com algum ruído. Requer algum esforço para o entendimento da mensagem.
2	Pobre	Sinal de baixa qualidade e difícil de entender.
1	Ruim	Sinal de qualidade inaceitável..

Dessa forma, mais 5 sinais, contendo a mesma frase, foram selecionados. Cada um, classificado em um dos 5 tipos de qualidade definidos acima. Esses sinais foram, então, tocados para todas as pessoas.

A partir do momento em que todas as pessoas conhecem o padrão de qualidade dos sinais da Tabela 5.1, a base de 92 sinais foi tocada, aleatoriamente, para cada uma delas. Neste momento, foi pedido que cada um atribuísse uma pontuação de 1 a 5 a cada sinal. A média das 5 pontuações de um dado sinal representa a naturalidade MOS do mesmo.

Seleção dos Sinais

Os 92 sinais são formados por 4 sinais diferentes em que os 4 grupos de bits do *bitstream* foram corrompidos em várias porcentagens durante a codificação/ decodificação. Esta seleção de sinais está identificada nas Figuras 5.1 e 5.2 através de quadrados.

Nesta etapa, toda a escolha dos sinais foi realizada de forma empírica. A maior preocupação foi com a escolha do número de sinais utilizados de forma que a massa de dados não se tornasse grande e exaustiva para análise MOS.

5.4.1 Resultados

Nas Figuras 5.1 e 5.2, estão apresentados os resultados obtidos através da primeira etapa da análise subjetiva. As faixas horizontais representam a escala MOS. No mesmo gráfico estão apresentados resultados das

duas análises, objetiva e subjetiva. Com esta representação é possível comparar as pontuações MOS obtidas através do mapeamento PESQ e através da análise subjetiva dos sinais de voz.

As Tabelas 5.2 e 5.3 mostram a pontuação MOS obtida para a análise dos 92 sinais.

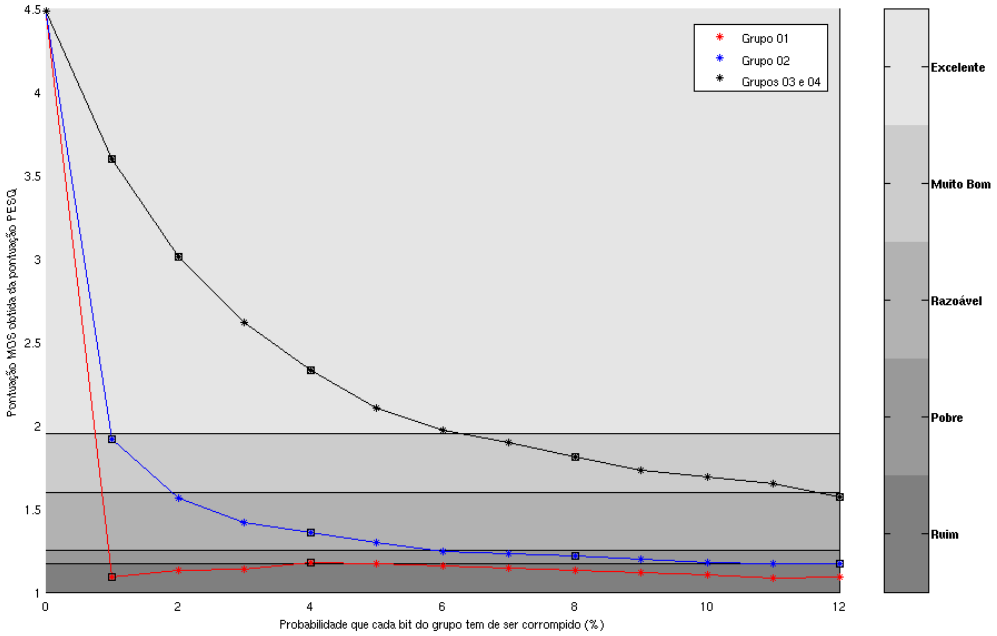


Figura 5. 1: Resultados das analises objetiva e subjetiva quando varia-se a probabilidade de corrompimento de bits do *bitstream*.

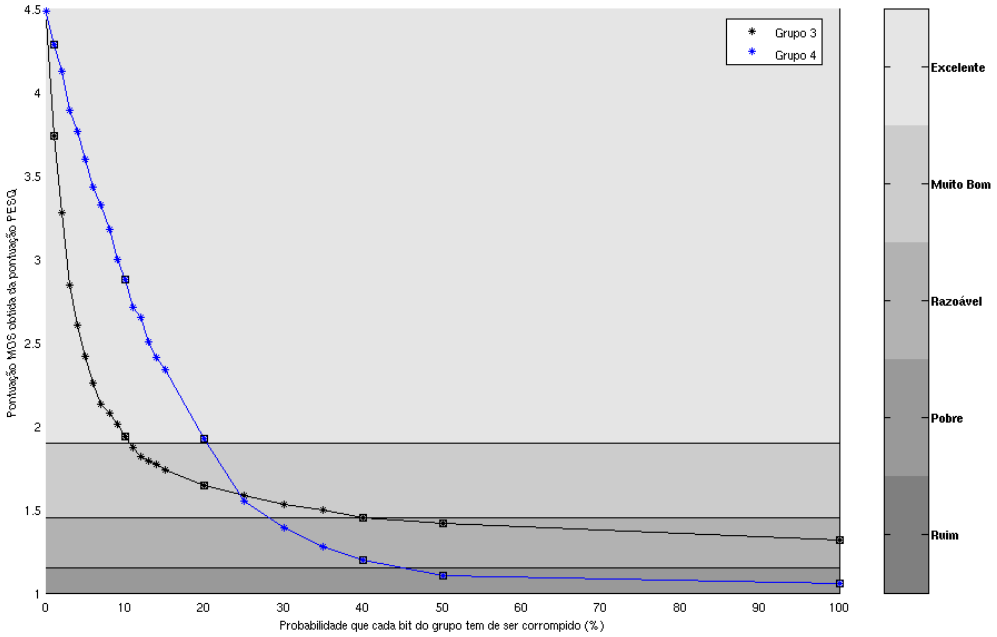


Figura 5. 2: Resultados das analises objetiva e subjetiva quando varia-se a probabilidade de corrompimento de bits dos grupos 3 e 4.

A Figura 5.3 é uma reprodução da Figura 4.10, com o objetivo de identificar quais bits compõem os conjuntos mostrados nas Figuras 5.1 e 5.2.

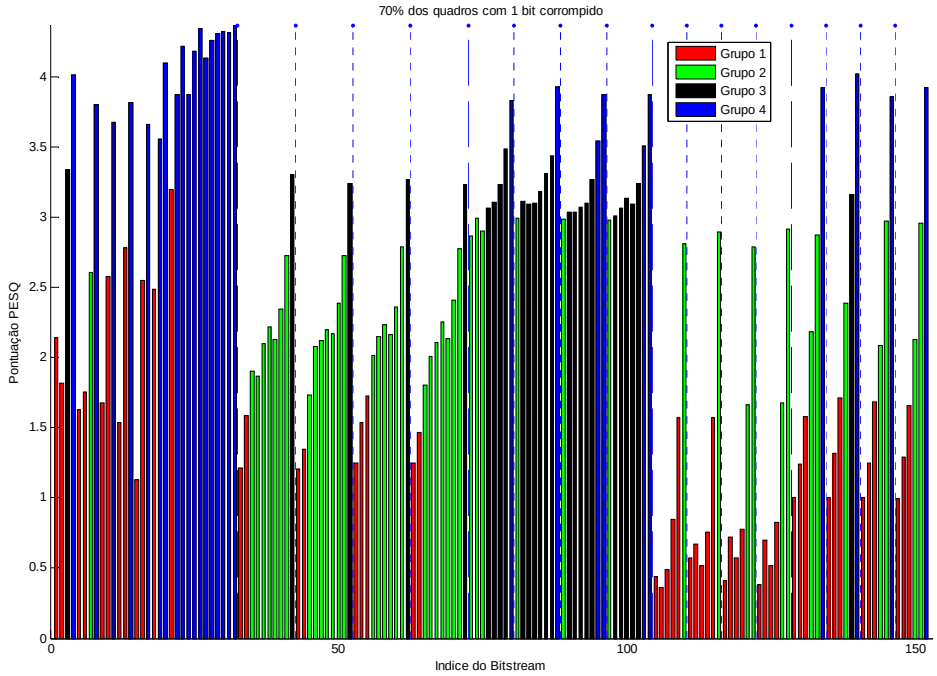


Figura 5. 3: Identificação dos 4 grupos de bits através da pontuação PESQ calculada quando um bit do *bitstream* é corrompido em 70% das janelas.

Tabela 5. 2: Pontuação MOS e desvio padrão (s) obtidos para frases dos grupos 1, 2, 3 e 4.

	Grupo 3 e 4		Grupo 2		Grupo 1			Grupo 3		Grupo 4	
Prob.	**média**	**s**	**média**	**s**	**média**	**s**	**Prob.**	**média**	**s**	**média**	**s**
1	4,9	0,1	3,7	0,8	1,0	0,0	**1**	4,7	0,2	4,7	0,3
2	4,6	0,2	-	-	-	-	**10**	4,0	0,2	4,2	0,4
4	4,6	0,5	2,4	0,5	1,0	0,0	**20**	3,2	0,2	3,0	0,3
8	3,8	0,5	2,0	0,5	-	-	**40**	3,0	0,4	2,2	0,3
12	2,9	0,5	2,2	0,3	-	-	**50**	3,3	0,5	1,8	0,1
-	-	-	-	-	-	-	**100**	2,5	0,3	1,1	0,1

Através da Tabela 5.2 observa-se que os bits do grupo 1 apresentam uma pontuação MOS 1,0 e um desvio padrão 0, o que confirma que este grupo é crítico quando leva-se em consideração a inteligibilidade do sinal de voz reconstituído.

A Tabela 5.2 também mostra a pontuação MOS para os demais grupos de bits. Através dos dados dessa tabela e das Figuras 5.1 e 5.2 nota-se que a escolha inicial das frases foi satisfatória para a primeira etapa da

análise subjetiva. No entanto, estes resultados evidenciam a necessidade de aprimoramento, pois alguns dos pontos selecionados encontram-se no limite entre duas classificações e apresentam um desvio padrão elevado. Como exemplo, há a pontuação MOS de 3,8 obtida para sinais do grupo 3 e 4, que indica uma classificação razoável, porém, apresenta um desvio padrão elevado (s = 0,5) indicativo de que sinais desse grupo pudessem ser classificados como muito bom. Com estes resultados iniciais torna-se possível melhorar a análise na segunda etapa.

5.5 Segunda Etapa da Análise Subjetiva

A segunda etapa tem como objetivo contornar os problemas evidenciados na etapa anterior. A primeira medida foi aumentar o número de pessoas que avalia os sinais de voz. Isto foi feito com o intuito de obter resultados com um desvio padrão menor. Assim, foram utilizadas 15 pessoas.

O procedimento experimental utilizado foi o mesmo da etapa anterior. As 15 pessoas foram inicialmente treinadas com as mesmas 5 frases, de forma que pudessem identificar os padrões definidos na Tabela 5.1, e em seguida procederam à avaliação do novo conjunto de sinais. Foram selecionados sinais dos tipos daqueles da primeira etapa, que apresentaram um elevado desvio padrão.

Seleção de sinais

Foi possível observar que sinais do mesmo grupo, em que os bits possuíam a mesma probabilidade de serem corrompidos, atingiram pontuações MOS bastante distintas. Isso leva a conclusão de que a freqüência de cada fonema no conjunto de sinais analisados influencia a pontuação MOS final, podendo causar um aumento do desvio padrão. Dessa forma, verifica-se o interesse em utilizar uma base de sinais que reflita a real ocorrência dos sons da fala. Nesta etapa, foram escolhidas 8 frases foneticamente balanceadas segundo [**9**], o que significa que nestas frases a freqüência de ocorrência dos fones reflete a ocorrência dos mesmos na língua portuguesa.

As frases escolhidas foram:

O cenário da história é um subúrbio do Rio
Eu tenho ótima razão para festejar.
A pequena nave medirá o campo magnético
O prêmio será entregue sem sessão solene
Ela e o namorado vão a Portugal de navio
O adiamento surpreendeu a mim e a todos
A gente sempre colhe o que plantou
A corrida de inverno aconteceu com vibração

5.5.1 Resultados

Na Tabela 5.3, são apresentadas as novas pontuações MOS obtidas. Pode-se observar que, de uma forma geral, o desvio padrão das medidas diminuiu. Além disso, as pontuações MOS se mostram mais consistentes, principalmente para o grupo 2, para o qual se observa uma queda da pontuação à medida que aumenta a probabilidade dos bits serem corrompidos.

Um dos fatores que influencia diretamente esta análise é a presença de fricativas na base de sinais de vozes utilizada. Frases que possuem fricativas, em geral, são classificadas com uma pontuação MOS inferior a de frases sem fricativas, mesmo quando as duas são corrompidas da mesma forma.

Tabela 5. 3: Pontuação MOS e desvio padrão (s) obtidos para frases dos grupos 2, 3 e 4.

	Grupo 3 e 4		Grupo 2			Grupo 3		Grupo 4	
Prob.	**média**	**s**	**média**	**s**	**Prob.**	**média**	**s**	**média**	**s**
1	-	-	3,8	0,5	**1**	-	-	-	-
2	-	-	-	-	**10**	3,9	0,2	4,3	0,1
4	4,4	0,2	2,7	0,4	**20**	3,5	0,3	3,7	0,3
8	3,8	0,3	1,9	0,5	**40**	3,3	0,3	2,3	0,4
12	3,3	0,5	1,7	0,3	**50**	-	-	-	-
-	**-**	**-**	**-**	**-**	**100**	-	-	-	-

Nas Figuras 5.4 e 5.5, verifica-se o resultado obtido com esta nova análise. Embora a Figura 5.4 mostre que a curva relativa ao grupo 1 atinge classificação pobre, através da análise subjetiva realizada, sabe-se que este grupo foi classificado como ruim com um desvio padrão zero. No entanto, a apresentação dos resultados no gráfico foi mantida desta forma para que os resultados obtidos para o grupo 2 não fossem erradamente apresentados.

A pontuação MOS_{PESQ}, obtida durante a análise objetiva, evidência uma melhora na qualidade do sinal quando cresce a probabilidade de corrompimento dos bits do grupo 1, o que agora se pode confirmar como um resultado inconsistente. Ao ouvir os sinais do grupo 1 para probabilidade superiores a 2% obtêm-se grandes trechos de silêncio, os quais, provavelmente, o algoritmo PESQ considera de qualidade superior ao grande ruído encontrado para a probabilidade de 1%.

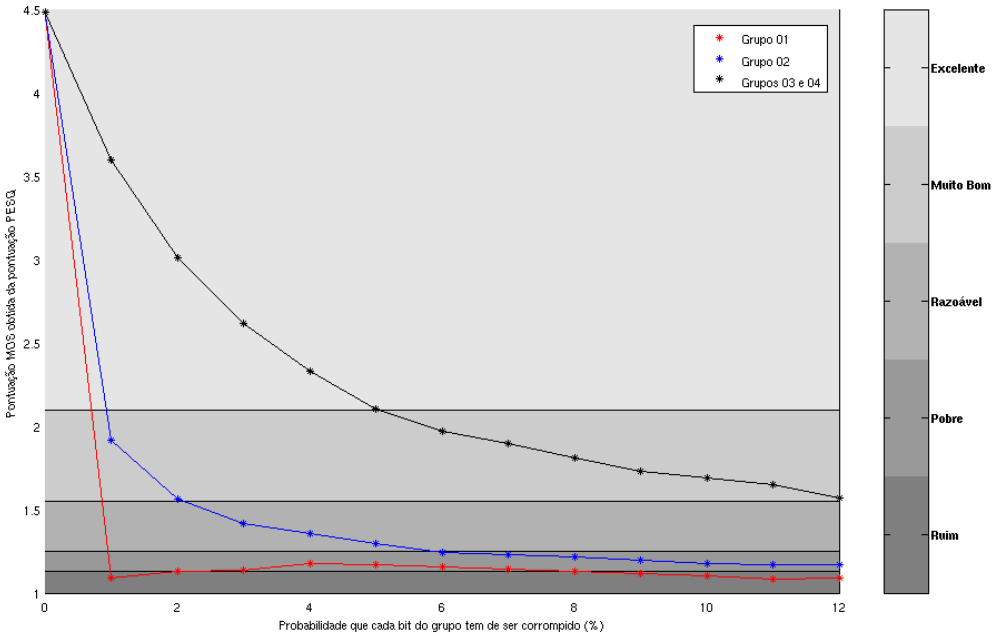


Figura 5. 4: Resultados das analises objetiva e subjetiva, obtidos durante a segunda etapa da análise subjetiva, quando varia-se a probabilidade de corrompimento de bits do *bitstream*.

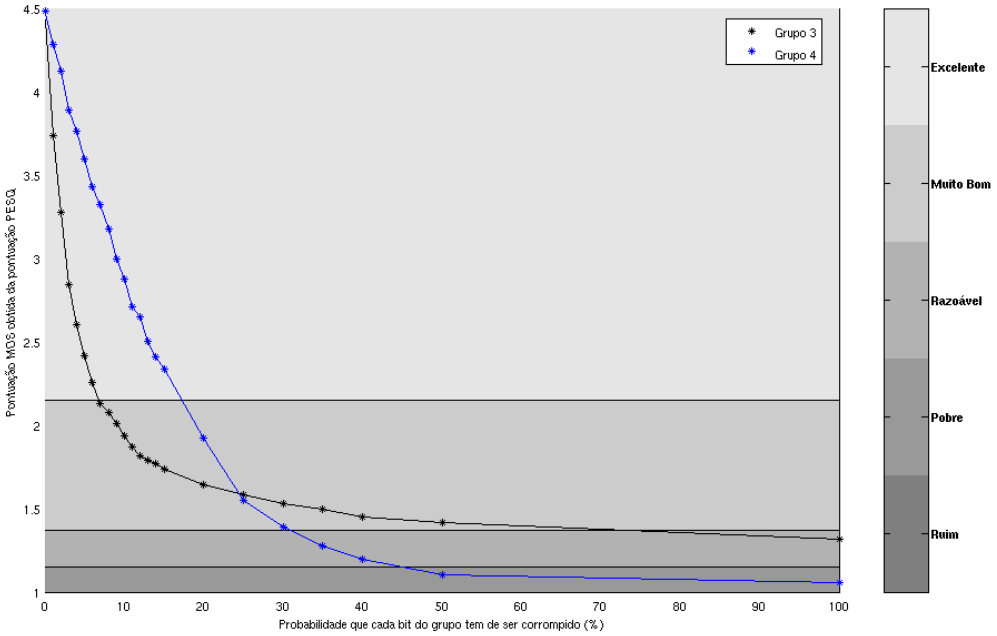


Figura 5. 5: Resultados das analises objetiva e subjetiva, obtidos durante a segunda etapa da análise subjetiva, quando varia-se a probabilidade de corrompimento de bits dos grupos 3 e 4.

Os resultados obtidos com esta análise poderiam ser mais precisos se o grupo de pessoas utilizado na avaliação fosse maior. Simplesmente, ao aumentar o número de pessoas de 5 para 10, pode-se obter uma pontuação MOS mais precisa. No entanto, para esta análise, os resultados foram satisfatórios.

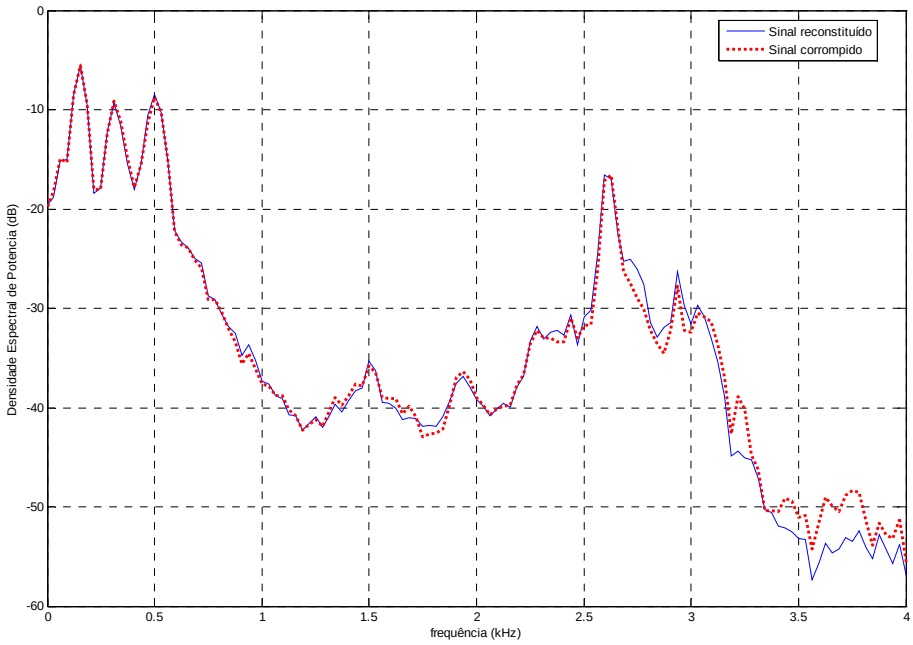
5.6 Análise de dados

Grupos 3 e 4

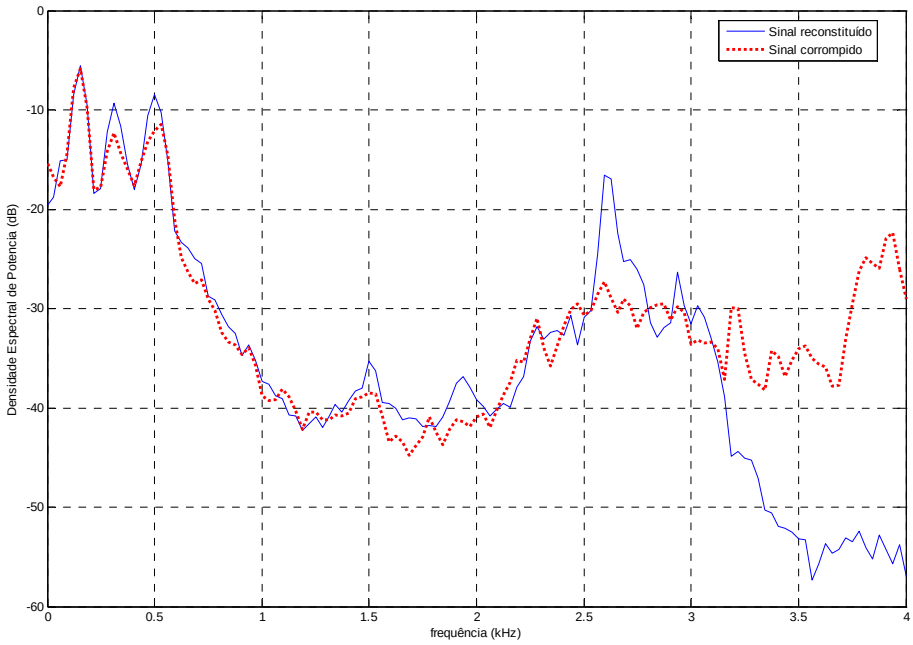
Os grupos 3 e 4 representam os bits de menor relevância para a inteligibilidade do sinal de voz como pode ser observado na Figura 5.4. Como foi visto, o grupo 3 é formado pelos bits menos significativos dos índices do dicionário fixo. Já o grupo 4 é formado pelos bits dos coeficientes DLSF de mais alta ordem. Dessa forma, como os dois grupos têm formações diferentes, conclui-se que a forma como cada um influencia na inteligibilidade do sinal, também será diferente. Isto pode ser observado na Figura 5.5. Embora, a princípio, o grupo 4 apresente uma qualidade superior, a situação se inverte para probabilidades maiores. Mesmo quando 100% dos bits do grupo 3 são corrompidos obtém-se uma classificação razoável, enquanto que o grupo 4 atinge uma classificação ruim.

Pelo fato de a excitação do dicionário fixo ser utilizada para modelar trechos não-vozeados do sinal, o início e as mudanças de excitação da voz, o corrompimento do grupo 3 gera sinais de elevada qualidade. O principal efeito é o aumento do ruído de fundo, o que não prejudica tanto a naturalidade do sinal de voz reconstituído.

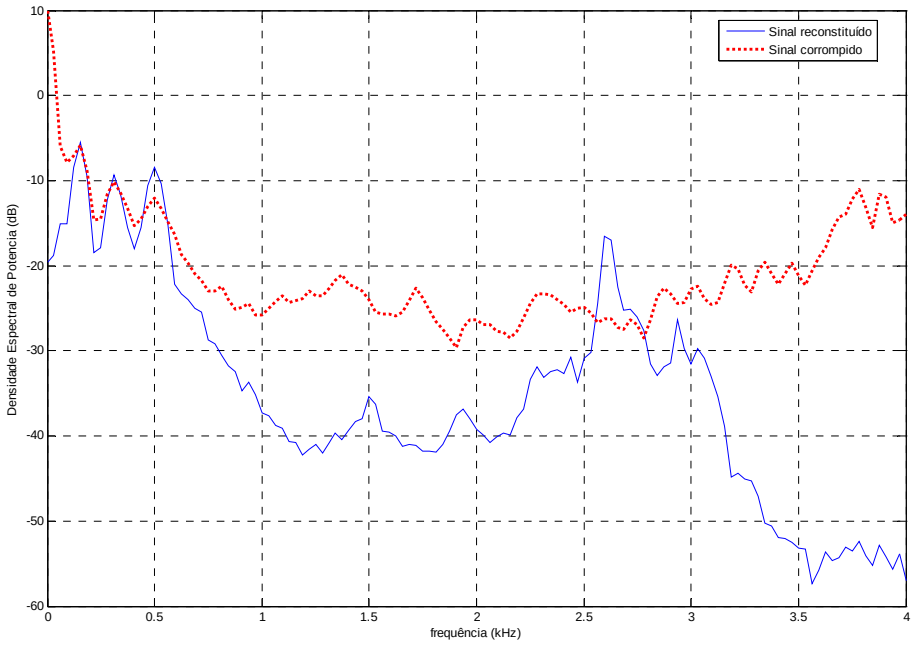
Quando o grupo 4 é corrompido, pode-se obter máxima degradação do sinal, pois o espectro de potência do mesmo em altas freqüências se distancia da situação ideal, o que pode ser observado na Figura 5.4. Isso corresponde a introdução de sons “metálicos” no sinal de voz e dessa forma, reduz a inteligibilidade do mesmo. É por isso que o corrompimento destes bits atinge uma classificação tão baixa na escala MOS para probabilidades elevadas. No entanto, este resultado não é considerado crítico quando comparado aos grupos 1 e 2. De acordo com o MOS, os grupos 3 e 4 apresentam boa classificação mesmo quando a probabilidade de cada bit ser corrompido é de aproximadamente 40%, ou seja, bastante elevada.


Sinal reconstituído
Sinal corrompido
Densidade Espectral de Potencia (dB)
frequência (kHz)
0
-10
-20
-30
-40
-50
-60
0
0.5
1
1.5
2
2.5
3
3.5
4

(a)


Sinal reconstituído
Sinal corrompido
Densidade Espectral de Potencia (dB)
frequência (kHz)
0
-10
-20
-30
-40
-50
-60
0
0.5
1
1.5
2
2.5
3
3.5
4

(b)



(c)

Figura 5. 6: Densidade espectral de potência de um sinal de voz em que os bits do grupo 4 tinham probabilidades de (a) 1%, (b) 20% e (c) 40% de serem corrompidos.

Grupo 2

O grupo 2 é composto, de uma forma geral, por bits menos significativos do índice do dicionário adaptativo e bits menos significativos dos ganhos relativos aos dois dicionários. Tanto a análise objetiva quanto a subjetiva evidenciam que este grupo tem grande relevância para a qualidade e a inteligibilidade do sinal resultante. Como pode ser visto na Figura 5.3, a queda de qualidade dos sinais quando esse grupo é corrompido é rápida. Isto se deve ao fato de uma alteração na excitação ótima, e nos ganhos resultarem ou em uma saturação do sinal, ou em uma redução da intensidade do mesmo. Logo, um dos maiores problemas está na alteração destes parâmetros relativos aos dicionários.

Outro fator relevante surge do fato de o dicionário adaptativo ser atualizado a toda iteração do algoritmo. Esta atualização é feita com o sinal resultante da soma das excitações ótimas, dos dois dicionários, já multiplicadas pelos respectivos ganhos. Assim, uma alteração no *bitstream* referente a esses parâmetros implica em que os dicionários adaptativos, do codificador e do decodificador, sejam atualizados com excitações distintas. Além disso, a alteração dos ganhos pode causar a atualização do dicionário com amostras saturadas. Desse modo, mesmo que o corrompimento dos bits ocorra em apenas algumas janelas, ele irá se propagar, e a redução da qualidade do sinal será rápida.

Grupo 1

O grupo 1 representa os bits mais críticos do *bitstream.* Ele é composto pelos bits mais significativos:

1. dos sete primeiros coeficientes DLSF, os quais representam os formantes do sinal original;
2. do índice referente à excitação ótima do dicionário adaptativo, o qual está diretamente relacionado à correlação de longo-termo presente nos trechos vozeados do sinal;
3. dos ganhos referentes aos dicionários, adaptativo e fixo.

É importante observar que, pela sua composição, esse grupo contém as principais informações das características do sinal de voz. De uma forma geral, o corrompimento destes bits irá gerar os mesmos problemas relatados anteriormente para os grupos 2, 3 e 4, sendo que agora estarão intensificados por este grupo ser formado pelos bits mais significativos.

A menor alteração nos coeficientes DLSF afeta, diretamente, a naturalidade do sinal de voz, o que já torna difícil a identificação do locutor. Além disso, grandes alterações nos ganhos dos dicionários causam uma saturação ou atenuação do sinal da janela corrente. E, o fator mais crítico é a atualização incorreta do dicionário adaptativo, que causa uma propagação de erros na codificação. Por isso, mesmo quando cada bit deste conjunto tem uma probabilidade de 1% de ser corrompido, o sinal resultante já apresenta uma classificação ruim.

É importante lembrar que, embora não tenha sido abordado anteriormente, independente do bit a ser corrompido, o erro causado por ele sempre se propaga para outras janelas em função da atualização do dicionário adaptativo. Ao serem utilizados sinais curtos (duração de 2 segundos), isto se torna crítico apenas para bits mais relevantes como visto através das análises objetiva e subjetiva. No entanto, é provável que para sinais mais longos a inteligibilidade do sinal se degrade ainda mais, tornando a atualização do dicionário crítica, também, para os demais bits.

5.7 Conclusão

Através das análises objetiva e subjetiva foi possível identificar o grupo 1 como sendo o conjunto de bits mais críticos tanto em termos da qualidade do codificador como da inteligibilidade e naturalidade do sinal de voz reconstituído. Estes são os bits que contém mais “informação” sobre o sinal de voz. O grupo 2 também apresentou-se bastante relevante em termos de informação. Embora a queda de qualidade não tenha sido brusca quando estes bits foram corrompidos, obtém-se uma classificação razoável quando considerada uma probabilidade de corrompimento de 2%.

Além disto, estas análises apontaram para o problema causado pela atualização dos dicionários adaptativos. Em um trabalho futuro, este é um importante ponto inicial de atuação para aumentar a robustez deste codificador CELP.

Capítulo 6

Conclusão

6.1 Resumo do Trabalho

Neste trabalho, foi apresentada uma análise sobre a avaliação da relevância de bits do *bitstream* na qualidade do sinal de voz codificado. Tal análise foi realizada sobre o codificador CELP implementado em [1,2]. Ao longo do trabalho, foi apresentada toda a metodologia de análise utilizada, bem como os resultados e a análise dos mesmos.

Os seguintes conteúdos foram abordados:

- No Capítulo 2, apresentou-se de uma visão geral sobre codificação. Apresentou-se, de forma simplificada, a geração e a modelagem do sinal de voz. Em seguida, foram apresentadas as principais características dos 3 tipos de codificadores – híbridos, paramétricos e de forma de onda – e também foram abordados os principais padrões de cada tipo encontrados atualmente. Como o foco deste trabalho está na avaliação da relevância bit-a-bit do codificador CELP, a ele foi dado maior destaque, tendo sido apresentadas características dos principais padrões existentes;

- No Capítulo 3, foram apresentados detalhes sobre a implementação CELP proposta em [1] bem como as modificações implementadas por [2]. Detalhes da implementação foram apresentados de forma que a análise dos resultados alcançados neste trabalho pudesse ser feita de forma mais criticada. Foram abordados: o tipo de janela aplicada ao sinal original, aspectos sobre o filtro de síntese e alterações realizadas no mesmo, particularidades estruturais dos dicionários (fixo e adaptativo), além de uma breve avaliação sobre o desempenho do codificador;

- No Capítulo 4, foi iniciada a análise de dados para avaliação da "contribuição" de bits do *bitstream* sobre a qualidade do sinal codificado. Esta primeira parte da análise foi feita de forma objetiva através das seguintes medidas: Razão Sinal-Ruído Segmentada Perceptual, Distância de Itakura, Distância Cepstral e o algoritmo PESQ. Detalhes sobre cada uma foram apresentados, já que este conhecimento foi necessário para a realização de escolhas durante a análise de dados. A metodologia de trabalho envolveu o corrompimento do *bitstream* e posterior avaliação sobre o impacto disto na qualidade do sinal reconstituído. Além de ter sido apresentada a metodologia de trabalho, apresentou-se também os primeiros resultados obtidos.

- No Capítulo 5, a análise de dados foi complementada através da realização de uma avaliação subjetiva da inteligibilidade e naturalidade dos sinais de voz corrompidos no capítulo anterior. Para isto foi utilizado o MOS no qual contou-se com a colaboração de pessoas para classificar os sinais em:

excelente, muito bom, razoável, pobre e ruim segundo padrões previamente determinados. Toda a metodologia de trabalho foi apresentada, bem como os resultados e discussões sobre os mesmos.

6.2 Contribuições

Este trabalho buscou propor uma metodologia de análise que permitisse verificar as vulnerabilidades de codificadores em termos de robustez. Embora toda a análise tenha sido realizada sobre uma implementação específica da codificação CELP, ela pode ser adaptada para outros codificadores.

Dessa forma, as principais contribuições foram:

- A estruturação de uma metodologia objetiva e subjetiva de avaliação da relevância de cada bit do *bitstream* sobre a qualidade do sinal de voz codificado;

- A idendificação de grupamentos no *bitstream* em função da "quantidade" de informação que eles contém sobre o sinal de voz transmitido;

- A apresentação de alguns padrões de codificação presentes hoje no mercado.

6.3 Propostas para Trabalhos Futuros

A seguir, são listadas algumas propostas para trabalhos futuros que tenham o objetivo de dar continuidade a este ou que venham a aproveitá-lo como referência:

- Estudos sobre implementações de dicionários adaptativos, uma vez que se verificou que estes são críticos para a qualidade do sinal quando o *bitstream* é corrompido. Uma possível solução seria a reinicialização do dicionário de tempos em tempos quando fossem detectados trechos de silêncio no sinal de voz;

- Estudos sobre técnicas para aumentar a segurança de codificadores em geral, baseando-se na análise de relevância dos bits realizada neste trabalho;

- A utilização de outras medidas de distância na análise objetiva deste trabalho;

- Estudos sobre códigos corretores de erro;

- Estudos sobre codificação de canal.

Referências Bibliográficas

[1] MAIA, R. da S., *Codificação CELP e Análise Espectral de Voz*, Tese de M.Sc., PEE-COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2000

[2] DINIZ, F. C. da C. B., *Implementação de um Codificador de Voz CELP em Tempo Real*, Projeto Final., DEL/UFRJ, Rio de Janeiro, RJ, Brasil, 2003

[3] http://research.edm.luc.ac.be/jori/thesis/onlinethesis/chapter4.html, *Chapter 4: Compression Techniques*, Setembro de 2005

[4] http://www.itu.int/ITU-T/recommendations/index.html, *ITU-T Recommendations*, Setembro de 2005

[5] http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/, *Processamento Digital da Fala*, Setembro de 2005

[6] B. de B. Oliveira, *Análise e Testes de um Codificador CELP*, Projeto Final., DEL/UFRJ, Rio de Janeiro, RJ, Brasil, 2001

[7] http://www.itu.int/rec/recommendation.asp?type=folders&lang=e&parent=T-REC-P.862.1,*P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO*, Outubro de 2005

[8] FERNANDES, N. L. L., *Relação entre a Qualidade das Respostas das Recomendações G.723.1 e G.729, e o Comportamento da Rede IP de Suporte*, Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2003

[9] Alcaim A., Solewicz J., Moraes J., *Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro*, Revista da Sociedade Brasileira de Telecomunicações, Vol. 7, Nº 1, Dezembro 1992

Apêndice A

Este apêndice lista as 200 frases utilizadas na análise de dados deste trabalho.

A questão foi retomada no congresso.
Leila tem um lindo jardim.
O analfabetismo é a vergonha do país.
A casa foi vendida sem pressa.
Trabalhando com união rende muito mais.
Recebi nosso amigo para almoçar.
A justiça é a única vencedora.
Isso se resolverá de forma tranqüila.
Os pesquisadores acreditam nessa teoria.
Sei que atingiremos o objetivo.
Nosso telefone quebrou.
Desculpe se magoei o velho.
Queremos discutir o orçamento.
Ela tem muita fome.
Uma índia andava na mata.
Zé , vá mais rápido!
Hoje dormirei bem.
João deu pouco dinheiro.
Ainda são seis horas.
Ela saía discretamente.
Eu vi logo a Iôiô e o Léo.
Um homem não caminha sem um fim.
Vi Zé fazer essas viagens seis vezes.
O atabaque do Tito é coberto com pele de gato.
Ele lê no leito de palha.
Paira um ar de arara rara no Rio Real.
Foi muito difícil entender a canção.
Depois do almoço te encontro.
Esses são nossos times.
Procurei Maria na copa.
A pesca é proibida nesse lago.
Espero te achar bem quando voltar.
Temos muito orgulho da nossa gente.
O inspetor fez a vistoria completa.
Ainda não se sabe o dia da maratona.

Será muito difícil conseguir que eu venha.
A paixão dele é a natureza.
Você quer me dizer a data?
Desculpe, mas me atrasei no casamento.
Faz um desvio em direção ao mar.
A velha leoa ainda aceita combater.
É hora do homem se humanizar mais.
Ela ficou na fazenda por uma hora.
Seu crime foi totalmente encoberto.
A escuridão da garagem assustou a criança.
Ontem não pude fazer minha ginástica.
Comer quindim é sempre uma boa pedida.
Hoje eu irei precisar de você.
Sem ele o tempo flui num ritmo suave.
A sujeira lançada no rio contamina os peixes.
O jogo será transmitido bem tarde.
É possível que ele já esteja fora de perigo.
A explicação pode ser encontrada na tese.
Meu vôo tinha sido marcado para as cinco.
Daqui a pouco a gente irá pousar.
Estou certo que mereço a atenção dela.
Era um belo enfeite todo de palha.
O comércio aqui tem funcionado bem.
É a minha chance de esclarecer a notícia.
A visita transformou-se numa reunião íntima.
O cenário da história é um subúrbio do Rio.
Eu tenho ótima razão para festejar.
A pequena nave medirá o campo magnético.
O prêmio será entregue em sessão solene.
A ação se passa numa cidade calma.
Ela e o namorado vão a Portugal de navio.
O adiamento surpreendeu a mim e a todos.
A gente sempre colhe o que plantou.
Aqui é onde existem as flores mais interessantes.
A corrida de inverno aconteceu com vibração.
Esse empreendimento será de enorme sucesso.
As feiras livres não funcionam amanhã.
Fumar é muito prejudicial à saúde.
Entre com seu código e o número da conta.
Reflita antes e discuta depois.

As aulas dele são bastante agradáveis.
Usar aditivos pode ser desastroso.
O clima não é mau em Calcutá.
A locomotiva vem sem muita carga.
Ainda é uma boa temporada para o cinema.
Os maiores picos da Terra ficam debaixo d'agua.
A inauguração da vila é quarta-feira.
Só vota quem tiver o título de eleitor.
É fundamental buscar a razão da existência.
A temperatura só é boa mais cedo.
Em muitas regiões a população está diminuindo.
Nunca se pode ficar em cima do muro.
Para quem vê de fora o panorama é desolador.
É bom te ver colhendo flores.
Eu me banho no lago ao amanhecer.
É fundamental chegar a uma solução comum.
Há previsão de muito nevoeiro no Rio.
Muitos móveis virão às cinco da tarde.
A casa pode desabar em algumas horas.
O candidato falou como se estivesse eleito.
A idéia é falha, mas interessa.
O dia esta' bom para passear no quintal.
Minhas correspondências não estão em casa.
A saída para a crise dele é o diálogo.
Finalmente o mau tempo deixou o continente.
Um casal de gatos come no telhado.
A cantora foi apresentar seu grande sucesso.
Lá é um lugar ótimo para tomar uns chopinhos.
O musical consumiu sete meses de ensaio.
Nosso baile inicia após as nove.
Apesar desses resultados tomarei uma decisão.
A verdade não poupa nem as celebridades.
As queimadas devem diminuir este ano.
O vão entre o trem e a plataforma é muito grande.
Infelizmente não compareci ao encontro.
As crianças conheceram o filhote de ema.
A bolsa de valores ficou em baixa.
O congresso volta atrás em sua palavra.
A médica receitou que eles mudassem de clima.
Não é permitido fumar no interior do ônibus.

A apresentação foi cancelada por causa do som.
Uma garota foi presa ontem à noite.
O prato do dia é couve com atum.
Eu viajarei ao Canadá amanhã.
A balsa é o meio de transporte daqui.
O grêmio ganhou a quadra de esportes.
Hoje irei à vila sem meu filho.
Essa magia não acontece todo dia.
Será bom que você estude esse assunto.
O menu incluía pratos bem saborosos.
Podia dizer as horas, por favor?
A casa é ornamentada com flores do campo.
A terra é farta, mas não infinita.
O sinal emitido é captado por receptores.
A mensalidade aumentou mais que a inflação.
O tele-jornal termina às sete da noite.
A cabine telefônica fica na próxima rua.
Defender a ecologia é manter a vida.
Nesse verão o calor está insuportável.
Um jardim exige muito trabalho.
O mamão que eu comprei estava ótimo.
Meu primo falará com a gerência amanhã.
De dia apague a luz sempre.
A sociedade uruguaia tem que se mobilizar.
Suas atitudes são bem calmas.
Dezenas de cabos eleitorais buscavam apoio.
A vitória foi paga com muito sangue.
Nossa filha tem amor por animais.
Esse peixe é mais fatal que certas cobras.
O time continua lutando pelo sucesso.
Essa medida foi devidamente alterada.
O estilete é uma arma perigosa.
Aguarde , quinta eu venho jantar em casa.
A mudança é lenta, porém duradoura.
O clima não é mais seco no interior.
A sensibilidade indicará a escolha.
A Amazônia é a reserva ecológica do globo.
O ministério mudou demais com a eleição.
Novos rumos se abrem para a informática.
O capital de uma empresa depende da produção.

Se não fosse ela tudo teria sido contido.
A principal personagem no filme é uma gueixa.
Receba seu jornal em sua casa.
A juventude tinha que revolucionar a escola.
A atriz terá quatro meses para ensaiar seu canto.
Muito prazer em conhecê-lo.
Eles estavam sem um bom equipamento.
O Sol ilumina a fachada de tarde.
A correção do exame está coerente.
As portas são antigas.
Sobrevoamos Natal acima das nuvens.
Trabalhei mais do que podia.
Hoje eu acordei muito calmo.
Esse canal é pouco informativo.
Parece que nascemos ontem.
Receba meus parabéns pela apresentação.
Eu planejo uma viagem no feriado.
No lado de cá do rio há uma boa sombra.
A maioria dos visitantes gosta desse monumento.
Minha filha é especialista em música sacra.
A casa só tem um quarto.
A duração do simpósio é de cinco dias.
Ao contrário de nossa expectativa, correu tranqüilo.
A intenção é obter apoio do governante.
A fila aumentou ao longo do dia.
À noite a temperatura deve ir a zero.
A proposta foi inspecionada pela gerência.
O quadro mostra uma face do cotidiano.
Já era bem tarde quando ele me abordou.
O canário canta ao amanhecer.
A lojinha fica bem na esquina de casa.
Meu time se consagrou como o melhor.
Um instituto deve servir a sua meta.
Ele entende quando se fala pausadamente.
Seu saldo bancário está baixo.
O termômetro marcava um grau.
O discurso de abertura é bem longo.
Eu precisei de microfone na conferência.
Joyce esticou sua temporada até quinta.
Nada como um almoço ao ar livre.

Nossa filha é a primeira aluna da classe.

Gostaria de deitar um pouco.

Não fizemos uma viagem muito cansativa.

Ainda tenho cinco telefonemas para dar.

Os hotéis do sudoeste são fantásticos.

Apêndice B

Este apêndice mostra o formulário utilizado na análise subjetiva realizada neste trabalho.

	Classificação				
frase	Excelente	Muito Bom	Razoável	Pobre	Ruim
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					
41					
42					