UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

Melhoria da Qualidade de Fala Através da Interpolação de Parâmetros do Codec GSM-AMR em Redes de Pacotes

Autor:
Paulo Henrique Marques Santos

Orientador:
Prof. Dr. Luís Geraldo Pedroso Meloni

Dissertação submetida à Faculdade de Engenharia Elétrica e de Computação da UNICAMP como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Banca Examinadora:

Prof. Dr. Luís Geraldo Pedroso Meloni (Orientador)	FEEC/UNICAMP
Prof. Dr. José Sindi Yamamoto	CPqD
Prof. Dr. Amauri Lopes	FEEC/UNICAMP
Prof. Dr. Leonardo de Souza Mendes	FEEC/UNICAMP

Campinas, outubro de 2004.

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

Sa59m

Santos, Paulo Henrique Marques
Melhoria da qualidade de fala através da interpolação de parâmetros do codec GSM-AMR em redes de pacotes / Paulo Henrique Marques Santos. --Campinas, SP: [s.n.], 2005.

Orientador: Luís Geraldo Pedroso Meloni
Dissertação (Mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Interpolação. 2. Codificador de voz. 3. Processamento de sinais. I. Meloni, Luís Geraldo Pedroso. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Resumo

As redes atuais da Internet não garantem transmissão sem perdas de pacotes. Ao contrário do tráfego de dados, o tráfego de voz admite alguma ocorrência de perda na comunicação entre os usuários. Entretanto, quando um ou mais pacotes são perdidos e nenhuma providência é tomada na tentativa de recuperá-los, a qualidade perceptual da fala fica significativamente deteriorada.

Um codificador de fala como o GSM-AMR, que emprega um esquema de codificação ACELP, comprime o sinal de fala e reduz a taxa de bits transmitindo apenas alguns parâmetros, ao invés de transmitir todas as amostras processadas. Assim, na tentativa de estimar os parâmetros de pacote(s) perdido(s), é proposto um esquema de interpolação dos parâmetros dos pacotes recebidos, adjacentes ao(s) pacote(s) perdido(s).

O codificador GSM-AMR já apresenta um esquema de substituição e silenciamento para quadros perdidos. Os dois esquemas são avaliados e comparados através de medidas de distorção. Apesar do codificador poder transmitir pacotes a taxas variadas de bits, as simulações restringiram-se à taxa de 7,95 kbps. Os resultados mostram que o esquema de interpolação proposto fornece melhorias na qualidade do sinal de fala decodificado em relação ao esquema já existente no codec GSM-AMR.

Abstract

The current Internet networks do not guarantee transmission without packet loss. Contrarily to data traffic, voice traffic still allows some packet loss, maintaining the information usefulness. However, when one or several packets are lost and no effort is made to recover those packets, the perceptual quality of the received speech can deteriorate significantly.

A speech coder such as the GSM-AMR, which uses the principle of ACELP, compresses speech signals and reduces bit rates transmitting only some parameters, instead of transmitting all processed samples. In this work, an interpolation scheme has been proposed trying to estimate the lost packet parameters. This interpolation is performed on the correctly received parameters, from packets adjacent to the lost ones.

The GSM-AMR already presents a scheme for substituting and muting lost packets. Both schemes are evaluated and compared according to a distortion measure. Although the coder can transmit packets in multi rates, simulations have been restricted to 7.95 kbps. Results show that the proposed interpolation scheme provides an enhancement in the decoded speech signal quality, compared to the already existing GSM-AMR scheme.

"Tudo o que quereis que os homens vos façam, fazei-o vós a eles." (Mt - 7:12)

À minha amada família, à qual dedico esse trabalho, por todo o amor, apoio e incentivo.

Agradecimentos

À Deus, pela força dada na realização deste trabalho.

À minha família, em especial, meus pais, José Paulo Araújo Santos e Rubenita Marques Santos, e minha irmã, Flávia Marques Santos.

Gostaria de agradecer à Ericsson Telecomunicações S.A. pelo suporte dado à este trabalho.

Ao Prof. Dr. Luís G. P. Meloni, pela oportunidade, orientação e paciência durante o desenvolvimento deste trabalho.

Aos grandes amigos Glauco, Márzio e Euler pela cooperação e ajuda.

Agradeço também ao Alexandre e ao Ginalber, bem como aos colegas do Laboratório de Processamento Digital de Sinais de Multimídia em Tempo Real pelo apoio e colaboração.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

ACELP - *Algebraic Code-Excited Linear Prediction*

ADM - *Adaptive Delta Modulation*

ADPCM - *Adaptive Differential Pulse Code Modulation*

APC - *Adaptive Predictive Coding*

APCM - *Adaptive Pulse Code Modulation*

ARMA - *Autoregressive Moving Average*

ATC - *Adaptive Transform Coding*

ATM - *Asynchronous Transfer Mode*

BFI - *Bad Frame Indication*

CELP - *Code-Excited Linear Prediction*

CRC - *Ciclic Redundancie Check*

DM - *Delta Modulation*

DPCM - *Differential Pulse Code Modulation*

ETSI - *European Telecommunications Standards Institute*

FFT - *Fast Fourier Transform*

FIR - *Finite Impulse Response*

GSM-AMR - *Global System Mobile - Adaptive Multi-Rate*

GSM-EFR - *Global System Mobile - Enhanced Full-Rate*

IP - *Internet Protocol*

ISDN - *Integrated Services Digital Network*

ITU-T - *International Telecommunication Union - Telecommunication Standardization Sector*

LP - *Linear Prediction*

LPC - *Linear Predictive Coding*

LSF - *Line Spectral Frequencies*

LSP - *Line Spectral Pairs*

LTP - *Long-Term Prediction*

MA - *Moving Average*

MOS - *Mean Opinion Score*

MR-ACELP - *Multi Rate - Algebraic Code-Excited Linear Prediction*

NTIMIT - *Network Texas Instruments Massachussets Institute of Technology*

PC - *Personal Computer*

PCM - *Pulse Code Modulation*

PESQ - *Perceptual Evaluation of Speech Quality*

PPP - *Point-to-Point Protocol*

PSQM - *Perceptual Speech-Quality Measure*

PSTN - *Public Switched Telephone Network*

QoS - *Quality of Service*

RELP - *Residual Excited Linear Prediction*

RTCP - *Real-Time Control Protocol*

RTP - *Real-Time Protocol*

SD - *Spectral Distortion*

SID - *Silence Insertion Descriptor*

SMQ - *Split Matrix Quantization*

SNR - *Signal to Noise Ratio*

SVQ - *Split Vector Quantization*

TCP - *Transmission Control Protocol*

TIPHON - *Telecommunications and Internet Protocol Harmonization over Networks*

UDP - *User Datagram Protocol*

VAD - *Voice Activity Detection*

VoIP - *Voice over Internet Protocol*

VSELP - *Vector-Sum Excited Linear Prediction*

WG5 - *Working Group* 5

Capítulo 1

Introdução

Os algoritmos de codificação de fala apresentam vários requisitos de desempenho a serem alcançados. Dentre eles, a taxa média de bits e a qualidade da fala são os aspectos mais importantes a serem considerados. Embora seja interessante diminuir a taxa de bits o máximo possível, torna-se muito difícil manter uma qualidade de fala aceitável à medida que esta taxa diminui. Almeja-se então alcançar um compromiso entre estes aspectos, de preferência com uma baixa complexidade computacional. Para conseguir uma fala de alta qualidade a taxas reduzidas de bits, os algoritmos de codificação aplicam métodos sofisticados para diminuir redundâncias inerentes à fala humana.

Uma menor taxa de bits implica que uma largura de banda menor será necessária para transmissão. Apesar da introdução da fibra ótica ter resultado em larguras de banda muito grandes para comunicações com fio, em comunicações por satélite e em sistemas de rádio móvel digital esta largura de banda ainda é limitada. Além do mais, em algumas aplicações relacionadas à fala existe a necessidade do armazenamento da voz digitalizada, que utilizará menos memória com a redução da taxa de bits. Estas aplicações de compressão de fala fazem da codificação deste sinal um campo de pesquisa bastante atraente.

1.1 Codificação de fala

A fala é uma forma de onda variante no tempo. O primeiro estágio na digitalização da fala envolve amostragem e quantização. O sinal de fala analógico $s(t)$ é primeiramente amostrado respeitando o critério de Nyquist, que define que amostras deste sinal, tomadas a intervalos regulares de T segundos, são suficientes para representar toda a informação do sinal original, desde que $T \leq 1/(2f_{max})$, ou seja, $f_s \geq 2f_{max}$, onde f_s e f_{max} são, respectivamente, a freqüência de amostragem e a freqüência máxima de $s(t)$.

O sinal discreto no tempo e quantizado é denotado por $s(n)$. Este sinal é codificado utilizando-se um ou vários esquemas de codificação. Ao longo das últimas décadas, uma variedade de técnicas de codificação de fala tem sido proposta, analisada, e desenvolvida. Os métodos mais importantes de codificação podem ser divididos em três categorias gerais: os codificadores de forma de onda, os codificadores paramétricos (*vocoders*) e os codificadores por transformadas (embora este último possa ser considerado um codificador de forma de onda).

Em codificação de forma de onda, explora-se as características temporais e/ou espectrais dos sinais de fala, através de codificação direta das formas de onda. Por outro lado, a codificação paramétrica envolve a representação do sinal de fala por um conjunto de parâmetros, pela estimativa dos parâmetros dos quadros de fala, e pela codificação eficiente destes parâmetros na forma digital para possível transmissão ou armazenamento [1].

Em várias técnicas de codificação de forma de onda, o processamento das amostras é realizado individualmente, quantizando-se e codificando-se uma amostra por vez separadamente. Este processo é chamado de codificação amostra a amostra. Por outro lado, em vários esquemas de codificação, como na paramétrica ou por transformada, há a possibilidade de se quantizar um bloco de amostras como uma única entidade. Este processo é chamado de codificação bloco a bloco.

Existem várias técnicas de codificação de forma de onda tanto no domínio do tempo quanto no domínio da freqüência. Estas técnicas têm sido amplamente utilizadas em

telecomunicações de fala desde de 1950. Como exemplos de técnicas de codificação de forma de onda no domínio do tempo têm-se a modulação por código de pulso (PCM) [2], a PCM diferencial (DPCM) [3], a DM (*Delta Modulation*) [4] e muitas versões adaptativas destes métodos, como APCM [5], ADPCM [6], ADM [7] [8], além da Codificação Preditiva Adaptativa (APC) [9], que pode ser vista como uma versão melhorada da ADPCM, na qual utiliza-se a periodicidade da fala sonora (*pitch*) para redução do erro. A ATC (*Adaptive Transform Coding*) [10] é um exemplo de codificação de forma de onda no domínio da freqüência.

Na codificação PCM, o sinal discreto no tempo é quantizado a um dos 2^R níveis, onde cada amostra $s(n)$ é representada por R bits. O quantizador pode ser uniforme ou não uniforme, escalar ou vetorial. Um quantizador uniforme típico usa de 8 a 16 bits por amostra. Um quantizador não uniforme pode empregar menos bits por amostra, pois o passo de quantização é ajustado à distribuição estatística das amostras do sinal. Por exemplo, os quantizadores logarítmicos com lei-μ ou lei-A usam 8 bits por amostra.

As técnicas de codificação paramétrica incluem o *vocoder* por banco de filtros [11], o homomórfico [12], o *vocoder* de fase [13], o *vocoder* de formantes [14] e o codificador preditivo linear [15]. Este último é o mais utilizado na prática hoje. Em codificação preditiva, o codificador processa um grupo de amostras em um determinado tempo, extrai os coeficientes que podem modelar estas amostras de uma forma compacta usando poucos bits, e os transmite. O decodificador reconstrói o sinal de fala a partir dos parâmetros transmitidos pelo codificador, os bits recebidos são convertidos de volta para a forma de coeficientes, e estes são filtrados para obtenção da fala codificada. Dentre os *vocoders* que utilizam esta técnica, pode-se citar o RELP (*Residual Excited Linear Prediction*) [16], o LPC Multipulso [17], o CELP (*Code-Excited Linear Prediction*) [18], o VSELP (*Vector-Sum Excited Linear Prediction*) [19], que é uma variação do CELP, e o ACELP (*Algebraic Code-Excited Linear Prediction*) [20].

A Codificação Preditiva Linear (LPC) é uma técnica que modela o sinal de fala como um filtro linear, excitado por um sinal chamado de sinal de excitação ou sinal residual.

O grupo de amostras processadas pelos codificadores é chamado de quadro ou segmento. Além dos coeficientes do filtro, o codificador encontra um sinal de excitação para cada quadro processado. Este filtro é chamado de filtro LPC de análise. No decodificador, o inverso do filtro LPC de análise age como um filtro LPC de síntese; e o sinal residual, que outrora era a resposta de saída do filtro LPC de análise, será o sinal de excitação do filtro LPC de síntese. Todo o processo é ilustrado na Figura 1-1.

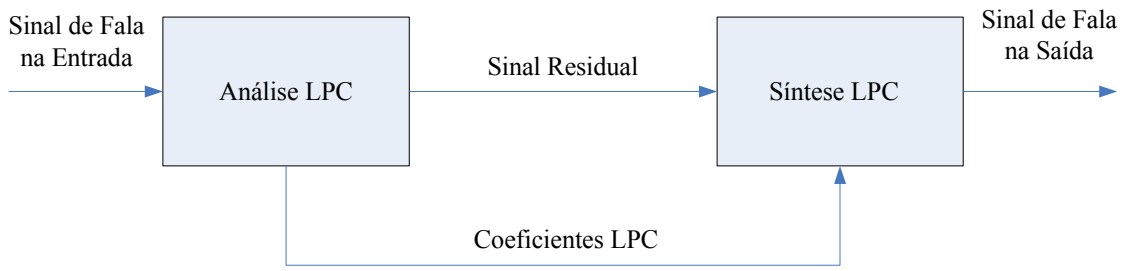


Figura 1-1: Análise e Síntese LPC.

Para reduzir a taxa total de bits, codificadores de fala tais como CELP não transmitem diretamente o sinal residual. Ao invés disso, eles utilizam uma técnica chamada de quantização vetorial, onde um vetor de um dicionário é usado para a codificação do sinal de excitação. Nesta técnica, o codificador seleciona um dos sinais de excitação de um dicionário pré-determinado, e o índice do sinal de excitação selecionado é transmitido. O dicionário é um conjunto finito de sinais de excitação conhecido tanto pelo codificador quanto pelo decodificador. O sinal de excitação é selecionado de tal forma que uma medida de distorção ponderada entre o segmento de fala original e o quadro reconstruído seja minimizada. O codificador transmite apenas o índice do sinal de excitação no dicionário, os coeficientes do filtro, informações do preditor de atraso longo (*pitch*), bem como os ganhos respectivos.

Tomando-se como exemplo o codificador GSM-AMR, que utiliza esta técnica de codificação, verifica-se que a taxa de amostragem do conversor A/D é de 8 kHz, e o comprimento do quadro é de 20 ms [21]. Isto implica que existem 160 amostras em cada

quadro. Para modelagem de um pacote de sinal com taxa de amostragem de 8 kHz, a utilização de um filtro LPC de décima ordem é suficiente. Isso significa que o codificador transmite somente os coeficientes LPC e os ganhos e índices de *pitch* e do dicionário, ao invés de 160 amostras de fala em um único quadro.

A motivação deste trabalho vem da necessidade de melhoria da Qualidade de Serviço (QoS) em redes IP. Uma característica da implementação atual do TCP/IP (IPv4) é que não há como garantir parâmetros como latência e perda de pacotes na rede, uma vez que as filas nos roteadores são dependentes fundamentalmente da carga da rede, da qual não se tem controle.

Normalmente os codificadores de fala já apresentam, na decodificação, procedimentos de silenciamento e substituição para quadros perdidos. O tratamento para quadros perdidos do codificador GSM-AMR será descrito em seções posteriores.

A recuperação de quadros perdidos na rede é o objetivo deste estudo. Um esquema de interpolação é executado entre os parâmetros dos quadros recebidos perfeitamente, adjacentes ao(s) quadro(s) perdido(s). Vale a pena salientar que este esquema de recuperação restringe-se à perda de até três quadros consecutivamente, devido à característica não estacionária dos sinais de fala. A interpolação é feita a partir dos parâmetros LPC, dos ganhos de *pitch* e do dicionário e do período de *pitch*. Os índices do dicionário são utilizados de acordo com a norma do codec GSM-AMR [37], na ocorrência de quadros perdidos .

1.2 Organização da tese

O objetivo deste trabalho é propor um esquema de interpolação de parâmetros que acarrete em melhorias na qualidade da fala codificada. A finalidade é a recuperação de quadros perdidos em uma rede de telefonia IP. A tese está organizada da seguinte maneira. O capítulo 2 introduz conceitos relativos a qualidade de serviço na transmissão de voz sobre pacotes. O Capítulo 3 trata dos princípios básicos de codificação de fala. O assun-

to é primeiramente abordado de uma forma geral e em seguida estuda-se o codificador GSM-AMR, cujo esquema de silenciamento e substituição de quadros perdidos é alvo desta pesquisa. No Capítulo 4 são apresentados o esquema de recuperação de quadros existente no codec GSM-AMR e o esquema de interpolação proposto. O Capítulo 5 trata da métrica utilizada para avaliação da qualidade dos sinais codificados, e apresenta os resultados obtidos durante as simulações. O Capítulo 6 apresenta as conclusões do trabalho e sugestões para trabalhos futuros.

Capítulo 2

Qualidade de Serviço na Transmissão de Voz sobre Pacotes

A Qualidade de Serviço (QoS) em redes de pacotes pode ser definida como a habilidade de um elemento de rede prover algum nível de garantia para entrega consistente de dados na rede. Os requisitos de qualidade a serem adotados dependem da aplicação. A tecnologia envolvida em telefonia IP impõe restrições severas devido à característica de tempo real que a mesma precisa atender.

Vários tópicos devem ser explorados para que haja um melhor entendimento do que está relacionado à transmissão de voz sobre pacotes. Isso envolve uma compreensão das características da audição humana, especialmente sua percepção de eco e atraso; das tecnologias de compressão e empacotamento de voz; de supressão de silêncio e geração de ruído de conforto; bem como de tecnologias de cancelamento de eco. Quanto as deficiências da Internet relacionadas a aplicações em tempo real: atraso, *jitter* e perda de pacotes devem ser analisados. É importante um conhecimento das estratégias para superar essas limitações, tais como *bufferização*, redundância, *timestamp* e serviços diferenciados; além de um estudo sobre as características do tráfego de voz empacotada, e como este se relaciona com fluxo de dados que não exigem tempo real. Este capítulo trata dos parâmetros que influenciam a qualidade da voz transmitida em uma rede de

pacotes. Também são apresentados elementos pertencentes à tecnologia empregada em uma solução IP.

2.1 Qualidade de Voz e Desempenho da Rede

A qualidade de voz obtida no sistema está diretamente ligada ao desempenho da rede, sendo mais diretamente afetada pelas seguintes características de desempenho:

- Atraso;
- Variação do Atraso (*jitter*);
- Perda de Pacotes;

A seguir descreve-se brevemente cada um desses aspectos.

2.1.1 Atraso

A latência ou atraso é um parâmetro importante para a qualidade de serviço das aplicações. Ambos os termos podem ser utilizados na especificação de QoS, embora o termo "latência" seja mais utilizado para equipamentos e o termo "atraso" seja mais utilizado com as transmissões de dados.

De uma forma geral, a latência da rede pode ser entendida como o somatório dos atrasos impostos pela rede e pelos equipamentos utilizados na comunicação. O atraso que ocorre nas redes IP é conseqüência do compartilhamento da largura de banda e do processamento nos roteadores e terminais. As aplicações de dados, para as quais as redes de pacotes foram desenvolvidas, são mais tolerantes ao atraso que as redes de voz.

O atraso pode ser classificado como fixo e variável. O atraso fixo corresponde ao atraso fim-a-fim para qualquer pacote de voz, independentemente de pontos de congestionamento na rede. Esse atraso está associado a fatores como compressão/descompressão

(codificação/decodificação), tamanho do *buffer* de *jitter*, tempo gasto para o empacotamento dos quadros, dentre outros. O atraso variável é causado por congestionamentos na rede ou nos *gateways*. Esse atraso corresponde essencialmente à soma dos atrasos na fila e na transmissão em cada roteador/*switch* intermediário na rede.

O atraso é o maior desafio da qualidade de serviço para voz em redes de pacotes, e corresponde ao tempo necessário para transmitir os pacotes de dados da origem ao destino. É o parâmetro que mais contribui para a perda da interatividade da conversação. Na Tabela 2.1 são apresentados alguns valores de tolerância ao atraso, citados na recomendação G.114 do ITU-T [22].

Atraso da Voz	**Tolerância**
até 150ms	Aceitável com boa interatividade
150ms - 400ms	Aceitável, mas o usuário já percebe alguma perda de interatividade
acima de 400ms	Inaceitável, com perda de interatividade

Tabela 2.1: Tolerância ao atraso em comunicações de voz.

As redes IP podem ser projetadas para minimizar o atraso acrescentando-se largura de banda e reduzindo-se as aplicações que competem entre si. O IP foi originalmente projetado como um protocolo para rede local com pouca ênfase na eficiência de banda. Os cabeçalhos dos pacotes IP são necessários para direcionar a transmissão da voz sobre a rede IP. A Figura 2-1 mostra um módulo em camadas de protocolos para uma aplicação voz sobre IP [23].

O protocolo *Real-Time Protocol* (RTP) [24] provê serviços fim-a-fim para aplicações de tráfego em tempo-real. Por isso, ele é utilizado para transportar pacotes de voz. As principais funcionalidades oferecidas pelo RTP são a identificação do tipo de tráfego, o número de seqüência de pacotes, o *timestamping* e, com o auxílio do *Real-Time Control Protocol* (RTCP), o monitoramento da entrega dos pacotes. Geralmente, o RTP é utilizado sobre o protocolo UDP (*User Datagram Protocol*), que provê um serviço de transporte que não prevê a retransmissão de pacotes perdidos. Para transmissão de pacotes contendo sinais de tempo-real, um pacote retransmitido provavelmente não chegará ao receptor a tempo de ser reproduzido. Além disso, o RTP faz uso da multiplexação e

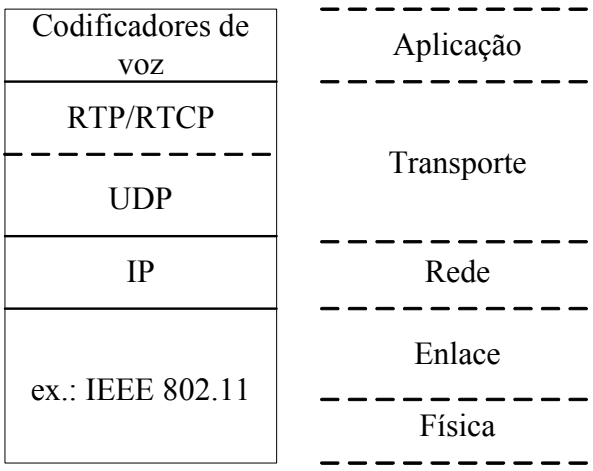


Figura 2-1: Módulo em Camadas de Protocolos para uma Aplicação VoIP.

do *checksum* provido pelo UDP. Em contrapartida, o RTP não provê nenhuma garantia de atraso ou de qualquer outro parâmetro de QoS. O RTP também não garante a entrega ordenada dos pacotes, mas o receptor pode utilizar o número de seqüência para ordená-los.

O RTCP é um protocolo de controle utilizado em conjunto com o RTP, e fornece informações sobre a qualidade de uma dada sessão RTP. Uma sessão RTP consiste em um conjunto de participantes que se comunicam através do protocolo RTP, sendo que para cada sessão são necessárias duas comunicações UDP (duas portas diferentes): uma utilizada pelo RTP e a outra pelo RTCP. Estas informações estão relacionadas a características da sessão, tais como: os participantes, a variação do atraso, a taxa de perdas, entre outras.

Do ponto de vista de sobrecarga de cabeçalhos e de processamento dos protocolos, deve-se enviar a maior quantidade possível de informação de voz em cada pacote para maximizar a utilização da capacidade da rede. No entanto, quanto maior a informação de voz, maior o tempo de espera para a geração do pacote e maior o tempo de transferência nó-a-nó na rede de comutação por pacotes. Assim, existe um compromisso entre a eficiência na utilização da rede e o atraso inserido pelo tamanho do pacote. O tama-

nho dos cabeçalhos dos pacotes IP demanda um nível significativo de largura de banda, porém é o mesmo para todos os pacotes. Desta forma, o tamanho dos dados de voz é o fator que determina o tamanho do pacote que será transmitido [25]. Como resultado, a transmissão de voz na rede IP será mais susceptível a congestionamento e atraso.

2.1.2 *Jitter*

O *jitter* é um outro parâmetro importante para a qualidade de serviço. O atraso de *jitter* é definido como a diferença entre o maior e o menor atraso sofrido pelos pacotes na conexão [39]. O *jitter* pode acontecer mesmo que a rede não esteja fortemente congestionada, uma vez que este é originado a partir dos diferentes atrasos nas filas, aos quais os pacotes são submetidos durante a transmissão. A Figura 2-2 ilustra a ocorrência do *jitter* em uma transmissão em uma rede de pacotes. Na transmissão (Tx), os pacotes são enviados a intervalos regulares de tempo; entretanto, essa regularidade não é mantida na recepção (Rx). O *playout time*, que é o tempo no qual o pacote de voz começa a ser executado, será explicado na seção 2.2.3.

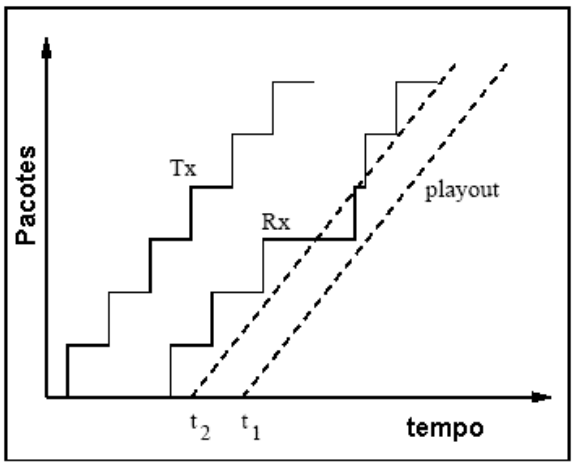


Figura 2-2: Conseqüência do Atraso de *Jitter*. (Adaptado de Kansal).

Os pacotes de voz precisam ser tocados a intervalos regulares para que uma qualidade

perceptual satisfatória possa ser alcançada. O agrupamento desses pacotes em *buffers* na recepção é a solução mais comum para tal problema. O instante de partida de cada pacote é determinado com o uso da informação de *timestamp* fornecida pelo RTP, e essa informação é utilizada para que os pacotes sejam armazenados e conseqüentemente tocados na ordem correta.

A especificação de requisitos de dimensionamento para uma rede VoIP do *European Telecommunications Standardization Institute* (ETSI), TR 101 329 V2.1.1 (1999-06) - *"Telecommunications and Internet Protocol Harmonization over Networks* (TIPHON); *General Aspects of Quality of Service (QoS)"* tem por escopo, o estabelecimento de requisitos mínimos de QoS para uma rede de pacotes com serviços de tempo real, como o de voz por exemplo.

O ETSI TIPHON foi um dos primeiros organismos de padrões a trabalhar em tópicos de QoS para telefonia IP. O TIPHON *Working Group* 5 (WG5) levou em conta parâmetros como a qualidade da voz percebida pelo usuário, atentando para degradação causada por codecs e pacotes perdidos, bem como interatividade entre outros. O objetivo do WG5 foi permitir alguma medida de desempenho da qualidade das redes de telefonia IP, em uma perspectiva fim-a-fim. Permitir que gerentes de rede determinem medidas para cada componente da rede de telefonia IP (*gateways*, terminais, *software*) para que seja possível algum planejamento.

A Tabela 2.2 relaciona os níveis de degradação da rede, usando valores da especificação TR 101 329 do ETSI TIPHON. A perda de pacotes observada na tabela é escopo da seção a seguir.

Categoria de Degradação da Rede	**Perda de Pacotes**	**Pico de Jitter**
Perfeita	0	0 ms
Boa	3 %	75 ms
Média	15 %	125 ms
Pobre	25 %	225 ms

Tabela 2.2: Níveis de degradação da rede.

2.1.3 Perda de Pacotes

As perdas de pacotes em redes IP ocorrem principalmente em função de fatores tais como:

- Descarte de pacotes nos roteadores e *switches* (erros, congestionamento) e
- Perda de pacotes devido a erros ocorridos na camada 2 (PPP - *Point-to-Point Protocol*, *Ethernet*, *frame relay*, ATM) durante o transporte dos mesmos.

Do ponto de vista da qualidade de serviço da rede, a preocupação é normalmente no sentido de especificar e garantir limites razoáveis (taxas de perdas) que permitam uma operação adequada da aplicação.

Protocolos de transporte de dados tais como o TCP, automaticamente retransmitem os pacotes perdidos. Devido a sua característica de tempo real, as aplicações de VoIP utilizam os protocolos UDP e RTP, que não efetuam a retransmissão em ocorrência de perda de pacotes.

Um fator relevante para a transmissão de voz é o número de pacotes perdidos consecutivamente em um mesmo fluxo. A perda consecutiva (em rajadas) de pacotes é mais prejudicial do que a perda esporádica.

Existem algumas técnicas para suavizar a perda de pacotes. A grande vantagem destas técnicas é não acrescentar sobrecarga na rede [23]. Dentre essas técnicas pode-se citar a substituição por silêncio, substituição por ruído, repetição, e outras como a interpolação, objeto de estudo deste trabalho.

2.2 Solução VoIP

A implementação de uma solução de voz sobre IP envolve uma série de requisitos e compromissos que precisam ser alcançados para que um sistema consistente possa ser concebido. Pode-se afirmar hoje que, a parte referente ao processamento digital dos sinais de voz está fundamentada nos seguintes itens:

- Determinação do codec mais adequado para aplicação;
- Implementação de um cancelador de eco elétrico (caso a solução envolva *gateway*);
- Implementação de um controlador para o *buffer* de *jitter*;

Foi reconhecido nos primeiros encontros do TIPHON que a faixa de aplicações da telefonia IP era tão ampla que apenas um tipo de nível de QoS seria insuficiente. Impor qualidade telefônica a todas as aplicações iria, na maior parte dos casos, inviabilizar a Internet pública como um *backbone* de transporte de voz [26]. Obviamente, as características ótimas esperadas de uma terminação diferem de acordo com a aplicação: telefonia IP sobre conexões por *modem* com pouca largura de banda, telefonia IP sobre conexões WAN corporativas com limitações de largura de banda e telefonia IP de alta qualidade sobre redes com grande largura de banda. A seguir são analisados os fatores a serem implementados nas soluções VoIP atuais.

2.2.1 Codificadores de Voz

A maioria dos codificadores de voz baseia-se em quadros. Isso significa que eles comprimem blocos contendo um número fixo de amostras quantizadas linearmente, em vez de comprimirem amostra por amostra. Dessa maneira, o fluxo de dados de áudio precisa ser acumulado até que ele atinja o tamanho do bloco antes de ser processado pelo codificador. Esse acúmulo de amostras leva tempo, portanto, soma-se ao atraso de fim-a-fim. Além disso, alguns codificadores precisam conhecer mais amostras do que aquelas contidas no quadro que eles vão codificar (*lookahead*).

O codificador de voz utilizado é um dos fatores que influencia a qualidade da voz em uma rede de transmissão. A escolha desse codificador depende da largura de banda disponível e dos requisitos de qualidade de serviço da aplicação.

Pesquisas e desenvolvimento de codificadores de voz com baixas taxas e boa qualidade de voz foram impulsionados, principalmente, pelo desenvolvimento de sistemas de comunicação sem fio. Esses sistemas utilizam faixas limitadas de freqüência e necessitam

acomodar o maior número possível de usuários. Para isso, torna-se necessário o uso de codificadores de voz que empreguem taxas reduzidas de bits para transmissão.

Atualmente, existe um grande número de codificadores de voz utilizando diferentes algoritmos para codificação. Os codificadores de voz têm por objetivo transmitir o sinal de voz com a maior qualidade possível utilizando a menor taxa de bits. Em geral, quanto maior a eficiência do codificador (alta qualidade e baixa taxa de bits), maior a complexidade do algoritmo empregado.

Para a escolha de um codificador de voz para uma determinada aplicação, algumas características dos mesmos devem ser analisadas, uma vez que essas características influenciam o desempenho da aplicação. Entre essas características estão a taxa de bits, a complexidade do algoritmo, a qualidade de voz e o atraso. A Tabela 2.3 apresenta valores de MOS[1] para alguns codificadores, sendo que os testes foram realizados em condições livres de erro [27] [28]. Em [29], estudos realizados sobre o codificador GSM-AMR mostraram que este apresenta valores de MOS variando entre 3,3 e 3,9 entre as taxas de 4,75 e 12,2 kb/s.

Codec	**Taxa de bits (kbps)**	**MOS**
G.711	64,0	4,43
G.729	8,0	4,18
G.723.1 (ACELP)	5,3	3,83
G.723.1 (MP-MLQ)	6,3	4,00
GSM-FR	13	3,6 - 3,8
GSM-HR	5,6	3,5 - 3,7
GSM-EFR	12,2	4,1
VRC (IS - 96)	8	3,3
EVRC (IS - 127)	8,5	4,1
SMV	8,55	4,1

Tabela 2.3: Valores MOS para codificadores de voz, condições livres de erro.

O atraso devido ao processo de codificação do sinal de voz é uma característica importante dos codificadores, sendo composto por dois fatores: atraso de *buffering* e atraso de

[1]Definido na seção 5.1.1.

codificação. O atraso de processamento depende do processador utilizado para a implementação do codificador. A Tabela 2.4 apresenta os atrasos de *buffering* e de codificação para alguns codecs. O atraso de codificação máximo foi estimado como sendo o atraso de *buffering* mais o tamanho do quadro utilizado pelo codificador.

Codificador	**Atraso de *buffering* (ms)**	**Atraso max. de codif. (ms)**
G.711	0	$< 0{,}125$
G.722	0	$< 0{,}125$
G.723.1	37,5	67,5
G.726	0	$< 0{,}125$
G.728	0,625	1,25
G.729	10	25
G.729-A	10	25
GSM-AMR	20	40
GSM-FR	20	40
GSM-HR	20	40
GSM-EFR	20	40
GSM-AMR-WB	25	45
VRC (IS-96)	27,5	47,5
VRC (IS-733)	27,5	47,5
EVRC	30	50
SMV	30 ou 33	50 ou 33

Tabela 2.4: Atraso de buffering e atraso de codificação para codificadores de voz.

2.2.2 Cancelador de Eco Elétrico

No caso de voz em tempo real em uma rede de pacotes, o eco elétrico deve ser cancelado caso o *gateway* IP venha a estabelecer uma chamada entre um terminal PSTN de dois fios e um terminal VoIP. Dentre as várias funções do *gateway* de telefonia IP está o cancelamento do eco elétrico gerado pela transformação de quatro fios para dois fios.

Basicamente, os canceladores de eco elétrico são filtros digitais adaptativos colocados na rede. O eco é modelado como uma soma de sinais, semelhantes ao sinal de entrada, porém atrasados e com menor amplitude. O eco surge no momento em que o sinal de entrada encontra um descasamento de impedância ao chegar na híbrida, que é onde ocorre

a conversão de dois para quatro fios.

Mesmo quando se usa esquemas de codificação de baixo atraso em sistemas de telefonia celular digital, VoIP, voz em *frame relay* ou mesmo voz em ATM, a utilização de canceladores de eco elétrico geralmente é obrigatória, e isso causa impacto sobre o custo do sistema como um todo.

2.2.3 Controlador do *Buffer* de *Jitter*

Controlar o *buffer* de *jitter* tem o objetivo explícito de minimizar o atraso causado por este *buffer* e, ao mesmo tempo, minimizar o impacto do *jitter* na transmissão.

Cada pacote armazenado no receptor deve aguardar o seu momento de reprodução (*playout time*), causando um aumento no atraso do pacote. Tal atraso poderá ser maior, quão maior for o *jitter* inserido pela rede, sendo limitado pelo tamanho do *buffer* utilizado. Caso um pacote chegue após o momento de sua reprodução, ele é automaticamente descartado. Por este motivo, na escolha do instante de *playout* existe um compromisso entre o atraso do pacote e a taxa de descarte. Um *playout time* pequeno pode diminuir o atraso do pacote, no entanto, se ele for muito pequeno em relação à média da variação do atraso, muitos pacotes serão descartados. Por outro lado, quanto maior for o tamanho do instante de *playout*, a fim de minimizar o descarte de pacotes, maior será o atraso do pacote. Vale a pena salientar que a perda de pacotes tolerável depende bastante do codec utilizado.

Desta forma, vários trabalhos procuram analisar mecanismos mais adequados para o armazenamento da voz. Para alguns terminais, a configuração do instante de *playout* é estática; entretanto esta pode não ser uma solução ótima quando as condições da rede não forem estáveis. Outros terminais podem redimensionar dinamicamente a configuração de tais instantes, através da utilização de algoritmos adaptativos. Normalmente esses algoritmos levam em consideração informações relacionadas a atraso, taxa de perda e *jitter*.

Capítulo 3

Análise e Modelagem de Sinais de Fala

3.1 Modelagem Preditiva Linear de Sinais de Fala

Atualmente, um dos métodos mais importantes de análise de fala é a Codificação Preditiva Linear, ou análise LPC, como é comumente referida. Na análise LPC, as correlações de curta duração entre amostras de fala são modeladas e removidas por um filtro de pequena ordem de forma muito eficiente. Os parâmetros provenientes dos formantes[1] são tratados pelos codificadores através dos coeficientes que modelam este filtro. O processo de produção da fala humana mostra que a geração de cada fonema é caracterizada basicamente pela excitação da fonte e pela forma do trato vocal. Assim, a modelagem da produção da fala implica, necessariamente, na modelagem desses dois fatores. O modelo do trato vocal é excitado por um sinal de excitação glótico discreto no tempo $u(n)$ para produzir um sinal de fala $s(n)$. Um diagrama de blocos simplificado deste modelo é mostrado na Figura 3-1. Neste modelo, a entrada de excitação, ou sinal de excitação, é modelada tanto por um trem de impulsos (para fonemas sonoros) quanto por um ruído aleatório (para fonemas surdos).

[1]Conjunto de picos observados na envoltória do espectro de um fonema sonoro. Ver Figura 3-2.

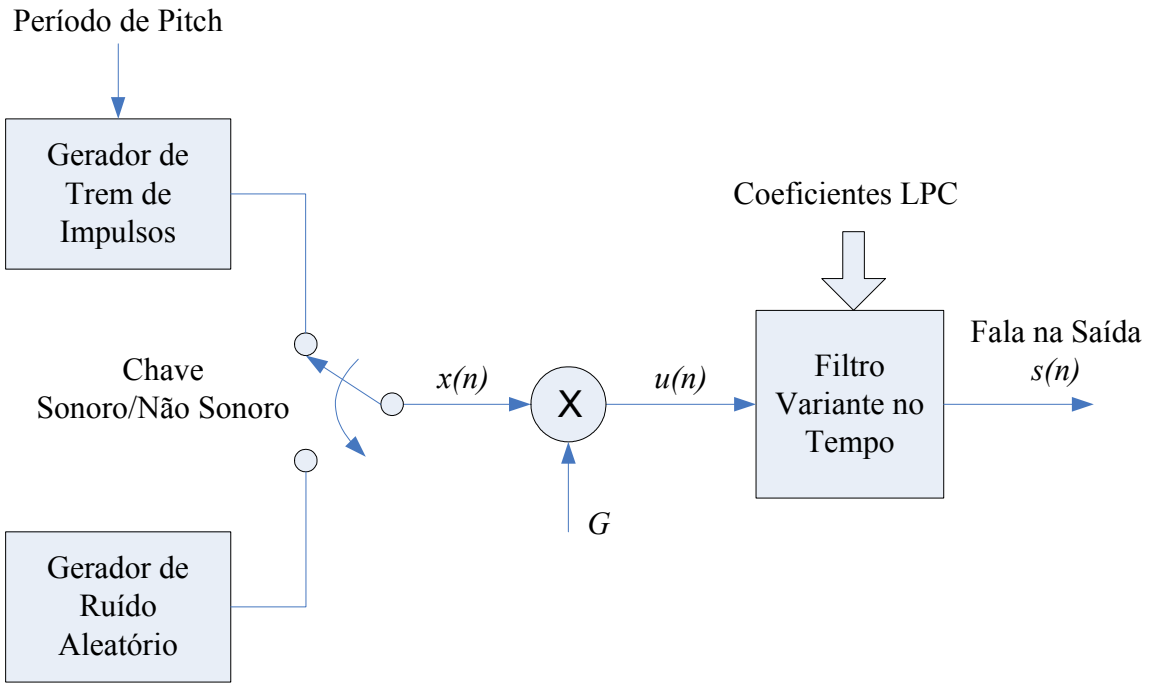


Figura 3-1: Diagrama de Blocos do Modelo Filtro Fonte Simplificado de Produção de Fala.

Dentre os vários recursos usados para modelar o trato o vocal[2], o ARMA (*autoregressive moving average*) é um modelo paramétrico geral, contendo pólos e zeros, e que pode ser usado na modelagem de tal sistema. Neste modelo, considera-se um sinal de fala $s(n)$ como a saída de um sistema cuja entrada é um sinal de excitação $u(n)$. A amostra de fala $s(n)$ pode ser expressa pela seguinte equação:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G \sum_{l=0}^{q} b_l u(n-l), \quad b_0 = 1 \tag{3.1}$$

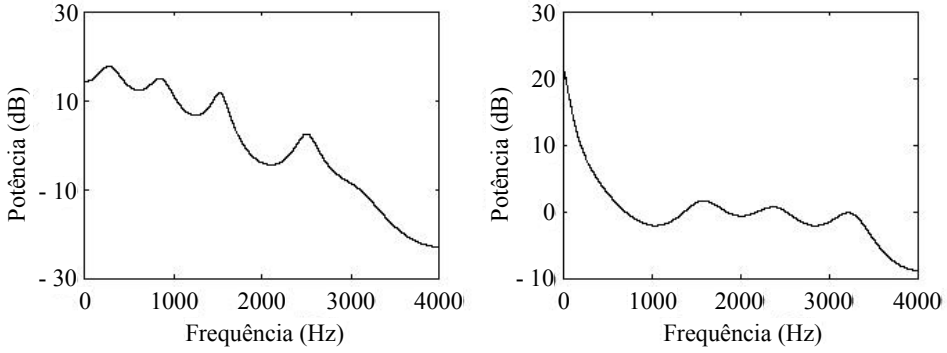
A equação (3.1) estabelece que o valor da saída atual, $s(n)$, pode ser determinado como uma combinação linear das saídas passadas e das entradas passadas e presente. Os

[2]O trato vocal é modelado como um filtro variante no tempo. Compreendem os articuladores do trato vocal: cordas vocais, língua, lábios, dentes, véu palatino e mandíbula. A produção da fala pode ser vista como uma operação de filtragem na qual uma fonte sonora excita o filtro do trato vocal [30].

parâmetros do sistema são o fator de ganho G e os coeficientes do filtro $\{a_k\}$ e $\{b_l\}$. O número p representa a ordem da predição linear. A função de transferência que modela o trato vocal é obtida aplicando-se a transformada z em (3.1):

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{3.2}$$

Os sons nasais são modelados pelos zeros, enquanto que os pólos da função de transferência modelam os formantes da fala. Na Figura 3-2, pode-se observar a contribuição dos pólos e zeros para a resposta em freqüência do filtro LPC. Os zeros são responsáveis pelos vales espectrais, enquanto que os picos espectrais são formados a partir dos pólos.



(a) Espectro de Potência de um sinal de fala sonoro

(b) Espectro de Potência de um sinal de fala não sonoro

Figura 3-2: Diferença entre os espectros de potência de sinais sonoros e não sonoros. (Adaptado de Tamanna Islam, 2000).

Há dois casos especiais do modelo ARMA. Quando todos os zeros de $H(z)$ são nulos, ou seja, $b_l = 0$ para $1 \leq l \leq q$, $H(z)$ reduz-se a um modelo só-pólos também conhecido como modelo autoregressivo. Quando todos os pólos de $H(z)$ são nulos, ou seja, $a_k = 0$ para $1 \leq k \leq p$, $H(z)$ torna-se um modelo só-zeros, também conhecido como média móvel.

Na obtenção de um modelo de pólos e zeros, torna-se necessário resolver um conjunto

de equações não-lineares de alta complexidade computacional. Uma maneira de se evitar tais cálculos é a utilização do modelo autoregressivo como uma aproximação do modelo do trato vocal, uma vez que a resolução do modelo autoregressivo conduz a um conjunto de equações lineares. Este modelo trata fonemas como vogais de uma maneira bastante satisfatória, além de apresentar uma boa eficiência computacional. Os zeros podem ser modelados aproximadamente pelo conjunto de pólos.

A função de transferência do modelo autoregressivo é:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{3.3}$$

Se o fator de ganho for $G = 1$, a função de transferência torna-se

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}, \tag{3.4}$$

onde $A(z)$ representa o polinômio $\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)$. Os coeficientes $\{a_k\}$ do filtro são chamados de coeficientes LPC (de predição linear).

O sinal de erro $e(n)$ é a diferença entre a fala de entrada e a fala estimada. Assim, a seguinte relação advém:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{3.5}$$

No domínio z, (3.5) é equivalente a

$$E(z) = S(z)A(z). \tag{3.6}$$

Agora, o modelo completo pode ser decomposto nas partes de análise e síntese como mostrado na Figura 3-3.

Observando-se a figura, nota-se que o sinal de erro é produzido durante a análise da fala, e que este sinal será parâmetro de entrada para a parte de síntese do sistema.

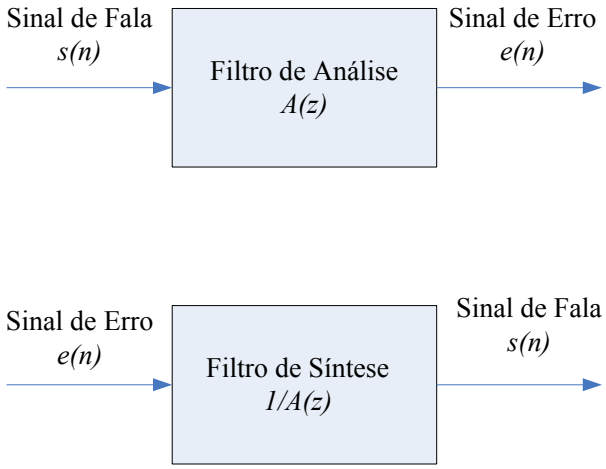


Figura 3-3: Modelo LPC de Análise e Síntese.

O sinal de erro, ou sinal residual, ou sinal de excitação, é filtrado pelo filtro de síntese originando, na saída deste, o sinal de fala reconstruído. Quando o sinal de erro originado na análise não é utilizado na síntese, o sinal de fala sintetizado não será igual ao sinal original. A mesma situação ocorre se o filtro de síntese não for exatamente o inverso do filtro de análise. Para diferenciar entre os sinais original e codificado, utiliza-se a notação $\hat{s}(n)$ para o sinal de fala sintetizado.

A análise de sinal assume que as propriedades deste, na maioria das vezes, mudam lentamente com o tempo. Isto permite que o sinal seja dividido em quadros sucessivos, sendo que a análise de predição linear é executada sobre esses quadros. A divisão em quadros está relacionada com a quase-estacionariedade do sinal de fala [30]. Assumir tal hipótese é razoável quando se trata de segmentos de curta duração, embora obviamente falsa para segmentos de fala de longa duração.

O sinal $s(n)$ é multiplicado por uma janela de análise de comprimento fixo $w(n)$, para selecionar um quadro particular em um determinado tempo. Este procedimento é chamado de janelamento. A escolha da forma correta de janela é muito importante, uma vez que a forma e o comprimento desta podem afetar a representação em freqüência do sinal. Vários tipos de janela têm sido estudados, produzindo formas e características

adequadamente utilizadas em várias aplicações. A seguir são apresentadas considerações sobre a implementação prática da análise LPC.

Janelamento

O tamanho da janela é dependente das características do sinal sob análise. Durante a análise da fala, é necessário que se tenha um comprimento que represente a estrutura harmônica precisamente, ou seja, que tenha mais que um ou dois períodos de *pitch* em cada janela. A duração de um período de *pitch* varia aproximadamente de 16 amostras, para um alto *pitch* feminino ou de uma criança, até 150 amostras, para um *pitch* muito baixo masculino. Assim, há um comprometimento para a determinação do tamanho da janela que, na prática, deve estar entre 120 e 240 amostras (de 15 a 30 ms para uma freqüência de amostragem de 8 kHz).

Determinação de Demais Fatores

Vários métodos podem ser usados no cálculo dos coeficientes de predição linear. Dentre os métodos que têm por característica minimizar a energia residual, com base na técnica clássica de mínimos quadrados, encontram-se os métodos da autocorrelação e da covariância [36]. A escolha da ordem do filtro também é um fator importante a ser analisado. Como em codificação de fala o sinal geralmente é amostrado a 8 kHz (dando um espectro de 4 kHz para análise), o número máximo de formantes observados nos espectros normalmente é igual a quatro. Isto implica que a ordem do filtro deve ser, pelo menos, igual a oito. Geralmente, um filtro de 10 pólos é usado para que ressonâncias de formantes e formas espectrais gerais sejam modeladas de forma mais precisa.

Atualmente, muitos codificadores de fala são baseados em modelagem preditiva linear, e o codificador GSM-AMR é parte integrante deste grupo. O objetivo deste trabalho é pesquisar um esquema de interpolação de parâmetros, que seja melhor do que o esquema de substituição e silenciamento de quadros perdidos, já existente no referido codificador. A seguir é apresentada uma visão geral do codificador GSM-AMR, definida nos padrões em [31] e em [32].

3.2 O Codificador GSM-AMR

Esta seção apresenta algumas informações sobre o codificador GSM-AMR, principalmente as mais relevantes para comparação com o esquema de interpolação de parâmetros, objeto de estudo deste trabalho.

3.2.1 Visão Geral

O codificador GSM-AMR [31] [32] emprega taxas variadas para transmissão dos sinais de voz, e apresenta um esquema de controle de taxas que inclue um detector de atividade de voz e um sistema de geração de ruído de conforto, além de um mecanismo de supressão de erro para combater os efeitos de erros de transmissão e de perda de pacotes.

O codificador de fala de taxas variadas é um único *codec* integrado com oito taxas que variam de 4,75 kb/s a 12,2 kb/s, e com um modo de codificação de ruído de fundo de baixa taxa de bits. O GSM-AMR é capaz de realizar o chaveamento entre as diferentes taxas a cada 20 ms de quadro de fala.

Uma configuração de referência onde as várias funções de processamento de fala são identificadas é dada na Figura 3-4. As partes de áudio são incluídas (incluindo as conversões analógica para digital e digital para analógica) para mostrar a trajetória completa da fala entre a entrada/saída de áudio no equipamento do usuário (UE) e a interface digital da rede. As partes de áudio são consideradas aqui apenas para mostrar que seu desempenho afeta o desempenho do *codec* de fala. Os módulos relevantes para cada função são listados abaixo:

1. PCM de 8 bits lei-A ou lei-μ (Recomendação G.711 ITU-T), 8000 amostras/s.
2. PCM uniforme de 13 bits, 8000 amostras/s.
3. *Flag* do Detector de Atividade de Voz (VAD).
4. Quadro de fala codificado, 50 quadros/s, o número de bits/quadro depende do modo do *codec* AMR.

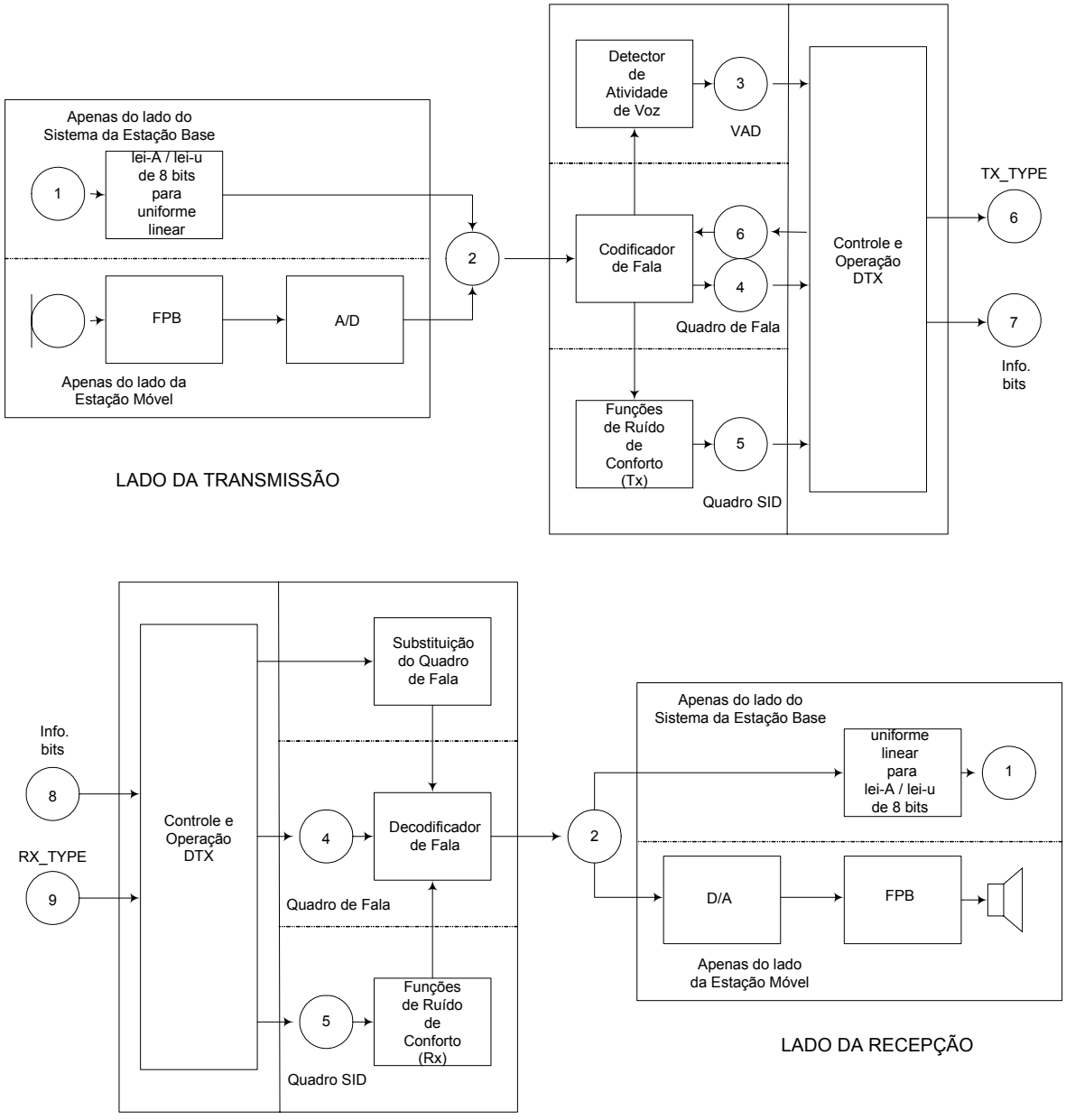


Figura 3-4: Visão Geral das Funções de Processamento de Áudio. (Adaptado de ETSI EN 301 703, 1999).

5. Quadro Descritor de Inserção de Silêncio (SID).

6. TX_TYPE, 2 bits, indica se as informações de bits estão disponíveis e se elas são informações de fala ou SID.

7. Bits de informação entregues a rede de acesso.

8. Bits de informação recebidos da rede de acesso.

9. RX_TYPE, o tipo de quadro recebido quantizado em três bits.

Como mostrado na Figura 3-4, o codificador de fala tem como entrada um sinal linear Modulado por Código de Pulsos (PCM) tanto da parte de áudio do equipamento do usuário quanto do lado da rede (da rede de telefonia pública chaveada (PSTN) via uma conversão PCM de 8 bits lei-A ou lei-μ para representação linear). A fala codificada é empacotada e entregue a interface de rede. As operações inversas acontecem no sentido da recepção.

O mapeamento detalhado de blocos de entrada de 160 amostras de fala em formato PCM linear para blocos codificados (nos quais o número de bits depende do modo usado pelo *codec*), e destes, para os blocos de saída de 160 amostras da fala reconstruída, é descrito na seção 3.2.2. As taxas de bits do *codec* fonte são listadas na Tabela 3.1.

Modos do Codec	**Taxa de Bits do Codec Fonte**
AMR_12.20	12,20 kb/s
AMR_10.20	10,20 kb/s
AMR_7.95	7,95 kb/s
AMR_7.40	7,40 kb/s
AMR_6.70	6,70 kb/s
AMR_5.90	5,90 kb/s
AMR_5.15	5,15 kb/s
AMR_4.75	4,75 kb/s
AMR_SID	1,80 kb/s

Tabela 3.1: Taxas de bits do codec fonte para o codec AMR.

3.2.2 Codec de Fala AMR: Funções de Transcodificação

Esta seção descreve o mapeamento detalhado dos blocos de entrada de 160 amostras de fala em formato PCM linear para os blocos codificados de 95, 103, 118, 134, 148, 159, 204 e 244 bits e vice-versa. A taxa de amostragem é de 8000 amostras/s, levando o fluxo codificado de bits a taxas de 4,75, 5,15, 5,90, 6,70, 7,40, 7,95, 10,2 ou 12,2 kb/s. O esquema de codificação para os modos de codificação a taxas variadas é o ACELP (*Algebraic Code-Excited Linear Predictive*). O codificador ACELP de taxa variada é referido como MR-ACELP.

Descrição Funcional das Partes de Áudio

A conversão analógica para digital e digital para analógica em geral compreenderá os seguintes elementos:

1. Analógica para PCM digital uniforme

- microfone;
- dispositivo de ajuste de nível de entrada;
- filtro *anti-aliasing* de entrada;
- dispositivo *sample-hold* amostrando a 8 kHz;
- conversão analógica para digital uniforme em representação de 13 bits.
- O formato uniforme será representado em complemento de dois.

2. PCM digital uniforme para analógica

- conversão de PCM uniforme 13 bits/8 kHz para analógica;
- um dispositivo *hold*;
- filtro de reconstrução incluindo correção sen(x)/x;

- dispositivo de ajuste de nível de saída;
- fone de ouvido ou alto-falante.

No equipamento terminal, a função A/D pode ser alcançada tanto

- por conversão direta para formato PCM uniforme de 13 bits;
- como por conversão para formato composto 8 bits lei-A ou lei-μ, baseada em um codec/filtro padrão lei-A ou lei-μ de acordo com as Recomendações ITU-T G.711 [2] e G.714 [33] (substituída pela G.712 [34]), seguida pela conversão de 8 bits para 13 bits, como especificado na seção *Conversão Formato PCM* a seguir.

Para a operação D/A ocorrem as operações inversas.

Preparação de Amostras de Fala

O codificador é alimentado com inclusão de dados de amostras com uma resolução de 13 bits justificadas à esquerda em uma palavra de 16 bits. Os três bits menos significativos são colocados em '0'. O decodificador tem dados no mesmo formato como saída.

Conversão Formato PCM

Na entrada do codificador de fala, a conversão de amostras codificadas (de 8 bits lei-A ou lei-μ) em amostras lineares (com resolução de 13 bits) segue a recomendação ITU-T G.711. Esta recomendação especifica, por meio de dados tabelados, a lei-A ou a lei-μ para conversão linear e vice-versa. Exemplos de como executar a conversão por aritmética de ponto fixo podem ser encontrados na recomendação ITU-T G.726 [35]. A seção 4.2.1 desta recomendação descreve a lei-A ou a lei-μ para expansão linear e, a seção 4.2.8 fornece uma solução linear para compressão lei-A ou lei-μ.

Princípios do Codificador de Fala AMR

O *codec* AMR consiste de oito *codecs* fonte com taxas de bit de 12,2, 10,2, 7,95, 7,40, 6,70, 5,90, 5,15 e 4,75 kb/s, e é um codificador da família CELP (*Code-Excited Linear Predictive*). É usado um filtro de síntese de predição linear (LPC) de décima ordem, também chamado de curta duração, o qual é dado por (3.4).

O filtro de síntese de longa duração ou filtro de *pitch*, é dado por:

$$\frac{1}{B(z)} = \frac{1}{1 - g_p z^{-T}}, \tag{3.7}$$

onde T é o atraso de *pitch* e g_p é o ganho de *pitch*.

O modelo de síntese de fala CELP é mostrado na Figura 3-5. Neste modelo, o sinal de excitação na entrada do filtro de síntese LPC de curta duração é construído adicionando-se dois vetores de excitação dos dicionários adaptativo e fixo. A fala é sintetizada alimentando-se o filtro de síntese de curta duração com os dois vetores escolhidos apropriadamente destes dicionários. A seqüência de excitação ótima em um dicionário é escolhida usando-se um procedimento de busca de análise por síntese, no qual o erro entre as falas original e sintetizada é minimizado de acordo com uma medida de distorção ponderada perceptualmente. O filtro perceptual ponderado usado na técnica de busca análise por síntese é dado por:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \tag{3.8}$$

onde $A(z)$ é o filtro LPC não quantizado e $0 < \gamma_2 < \gamma_1 \leq 1$ são os fatores ponderados perceptuais. Os valores $\gamma_1 = 0,9$ (para os modos AMR_12.20 e AMR_10.20) ou $\gamma_1 = 0,94$ (para todos os outros modos) e $\gamma_2 = 0,6$ são usados. O filtro ponderado usa os parâmetros LPC não quantizados.

O codificador opera com quadros de fala de 20 ms, correspondendo a 160 amostras na freqüência de amostragem de 8000 amostras/s. A cada 160 amostras de fala, o sinal de fala é analisado para extração dos parâmetros do modelo CELP (coeficientes LPC do filtro,

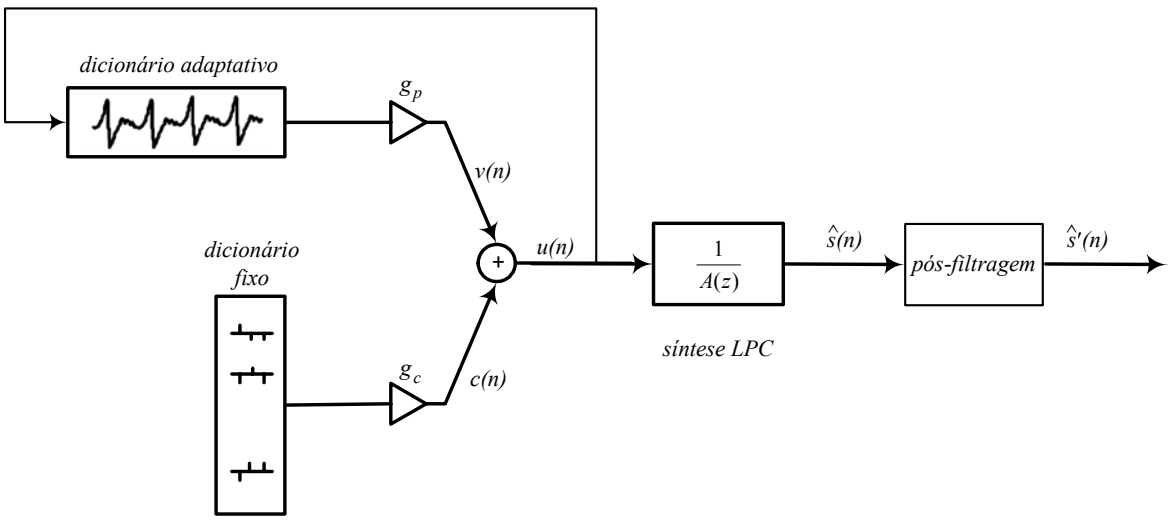


Figura 3-5: Diagrama de Blocos Simplificado do Modelo de Síntese CELP.

índices e ganhos dos dicionários adaptativo e fixo). Estes parâmetros são codificados e transmitidos. No decodificador, os parâmetros são decodificados e a fala é sintetizada filtrando-se, através do filtro LPC de síntese, o sinal de excitação reconstruído.

O fluxo completo de sinal no codificador é mostrado na Figura 3-6. A análise LPC é executada duas vezes por quadro para o modo AMR_12.20 e uma vez para os demais modos. Para o modo AMR_12.20, os dois conjuntos de parâmetros LPC são convertidos para pares de linhas espectrais (LSP - *Line Spectral Pairs*) e são juntamente quantizados usando-se quantização matricial partida (SMQ - *Split Matrix Quantization*) com 38 bits. Para os outros modos, o único conjunto de parâmetros LPC é convertido para pares de linhas espectrais (LSP) e é quantizado vetorialmente usando-se quantização vetorial partida (SVQ - *Split Vector Quantization*). Os pares de linhas espectrais (relacionados às Frequências de Linhas Espectrais (LSF - *Line Spectral Frequencies*)) são uma representação dos parâmetros LPC no domínio da frequência. Detalhes da conversão dos parâmetros LPC para a representação LSP/LSF podem ser encontrados em [36]. O quadro de fala é dividido em 4 sub-quadros de 5 ms cada (40 amostras). Os parâmetros do dicionário adaptativo e fixo são transmitidos na cadência de sub-quadro. Os parâmetros

LPC quantizados e não quantizados ou suas versões interpoladas são usados dependendo do sub-quadro. Um período de *pitch* em malha aberta é estimado em todos os sub-quadros (exceto para os modos AMR_5.15 e AMR_4.75, nos quais isto é feito uma vez por quadro), baseado no sinal de fala ponderado perceptualmente.

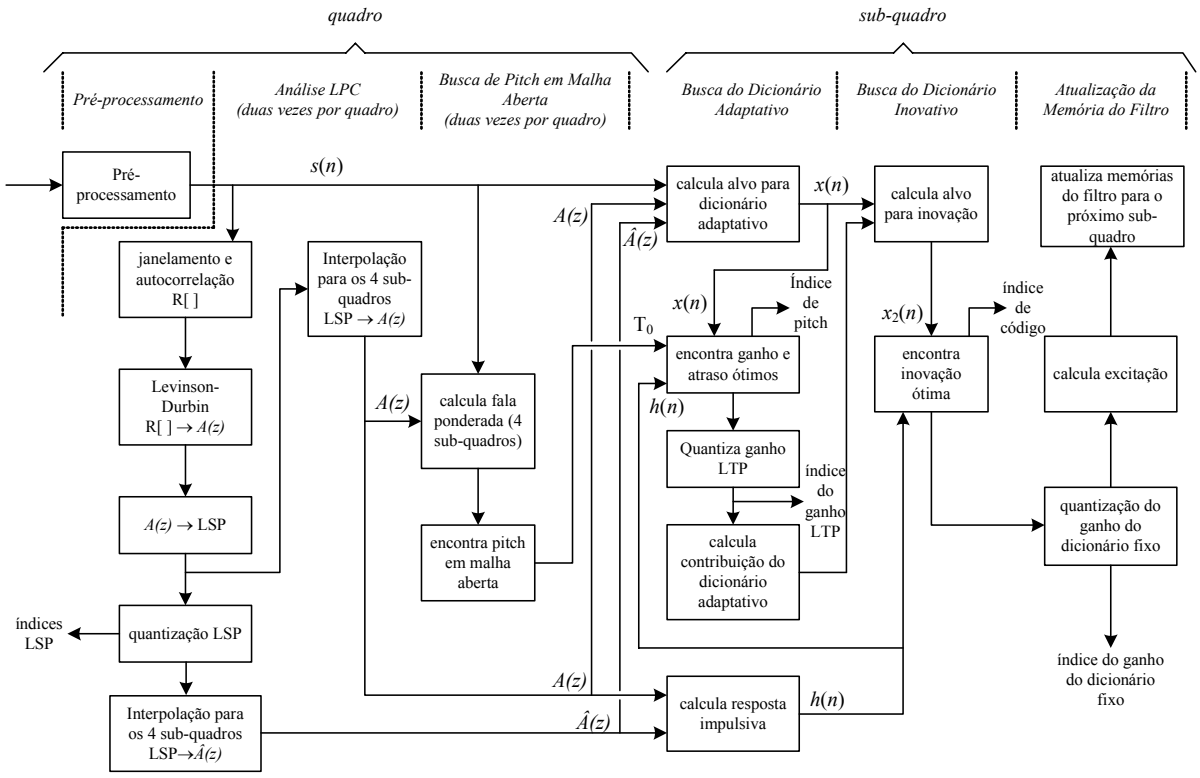


Figura 3-6: Diagrama de Blocos Simplificado do Codificador GSM-AMR.

As seguintes operações são repetidas para cada sub-quadro:

- O sinal alvo $x(n)$ é calculado filtrando-se o resíduo LPC através do filtro ponderado de síntese $W(z)H(z)$ com os estados iniciais do filtro tendo sido atualizados através da filtragem do erro entre o resíduo LPC e a excitação (isto é equivalente a subtrair do sinal ponderado de fala, a resposta à entrada zero do filtro ponderado de síntese).
- É calculada a resposta ao impulso $h(n)$ do filtro ponderado de síntese.

- A análise de *pitch* em malha fechada é então executada (para determinar o período e o ganho de *pitch*), usando-se o alvo $x(n)$ e a resposta ao impulso $h(n)$, procurando-se em torno do período de *pitch* em malha aberta. É usado um *pitch* fracionário com 1/6 ou 1/3 da resolução de uma amostra (dependendo do modo).

- O sinal alvo $x(n)$ é atualizado removendo-se a contribuição do dicionário adaptativo (vetor de código adaptativo filtrado), e este novo alvo, $x_2(n)$, é usado na busca do dicionário algébrico fixo (para encontrar a inovação ótima).

- Os ganhos dos dicionários adaptativo e fixo são quantificados de forma escalar com 4 e 5 bits respectivamente, ou quantificados de forma vetorial com 6-7 bits (com predição média móvel (MA) aplicada ao ganho do dicionário fixo).

- Finalmente, as memórias do filtro são atualizadas (usando-se o sinal de excitação determinado) para que se encontre o sinal alvo no próximo sub-quadro.

Princípios do Decodificador de Fala AMR

O fluxo do sinal no decodificador é mostrado na Figura 3-7. Os índices transmitidos são extraídos do fluxo de bits recebido e, então, decodificados para obtenção dos parâmetros do codificador em cada quadro transmitido. Estes parâmetros são os vetores LSP, os períodos de *pitch* fracionários, os vetores de código de inovação, e os respectivos ganhos de *pitch* e de inovação. Os vetores LSP são convertidos para os coeficientes LPC do filtro e interpolados para obter filtros LPC em cada sub-quadro. Então, a cada 40 amostras de sub-quadro:

- a excitação é construída adicionando-se os vetores de código adaptativo e de inovação escalados pelos seus respectivos ganhos;

- a fala é reconstruída filtrando-se a excitação através do filtro LPC de síntese.

Finalmente, o sinal de fala reconstruído passa por um pós-filtro adaptativo.

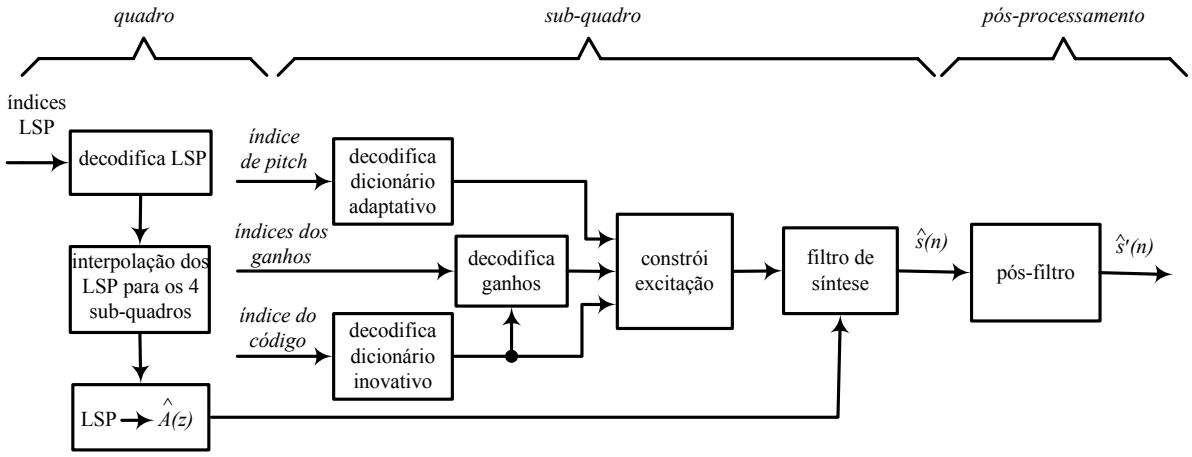


Figura 3-7: Diagrama de Blocos Simplificado do Decodificador GSM-AMR.

Análise de Predição Linear e Quantização

Modo AMR_12.20

A análise de predição de curta duração, ou predição linear (LP), é executada duas vezes por quadro de fala, usando-se a aproximação da auto-correlação com janelas assimétricas de 30 ms. Não são usadas amostras de quadros futuros no cálculo da auto-correlação.

As auto-correlações da fala janelada são convertidas para os coeficientes LPC usando-se o algoritmo de Levinson-Durbin, que é apresentado no final deste capítulo. Estes são então transformados para o domínio LSP por razões de quantização e interpolação. Os coeficientes interpolados do filtro, quantificados e não quantizados, são convertidos de volta aos coeficientes do filtro LPC (para construir os filtros de síntese e ponderado para cada sub-quadro).

Outros Modos

A análise de predição de curta duração, ou predição linear, é executada uma vez por quadro de fala, usando-se a aproximação da auto-correlação com janelas assimétricas de 30 ms. São usadas 40 amostras (5 ms) do quadro futuro no cálculo da auto-correlação.

As auto-correlações da fala janelada são convertidas para os coeficientes LPC usando-

se o algoritmo de Levinson-Durbin. Estes são então transformados para o domínio LSP por razões de quantização e interpolação. Os coeficientes interpolados do filtro, quantificados e não quantizados, são convertidos de volta aos coeficientes do filtro LPC (para construir os filtros de síntese e ponderado para cada sub-quadro).

Janelamento e Cálculo da Auto-Correlação

Modo AMR_12.20

A análise LP é executada duas vezes por quadro usando-se duas janelas assimétricas diferentes. A primeira janela tem seu peso concentrado no segundo sub-quadro e consiste de duas metades de janelas de Hamming com tamanhos diferentes. A janela é dada por:

$$w_I(n) = \begin{cases} 0,54 - 0,46\cos(\frac{n}{L_1^{(I)}-1})\,; & n = 0,\ ...,\ L_1^{(I)} - 1 \\ 0,54 + 0,46\cos(\frac{\left(n-L_1^{(I)}\right)}{L_2^{(I)}-1})\,; & n = L_1^{(I)},\ ...,\ L_1^{(I)} + L_2^{(I)} - 1 \end{cases} \tag{3.9}$$

Os valores $L_1^{(I)} = 160$ e $L_2^{(I)} = 80$ são usados. A segunda janela tem seu peso concentrado no quarto sub-quadro e consiste de duas partes: a primeira parte é a metade de uma janela de Hamming e a segunda parte é um quarto de um ciclo de função cosseno. A janela é dada por:

$$w_{II}(n) = \begin{cases} 0,54 - 0,46\cos(\frac{2n}{2L_1^{(II)}-1})\,; & n = 0,\ ...,\ L_1^{(II)} - 1 \\ \cos(\frac{2\left(n-L_1^{(II)}\right)}{4L_2^{(II)}-1})\,; & n = L_1^{(II)},\ ...,\ L_1^{(II)} + L_2^{(II)} - 1 \end{cases} \tag{3.10}$$

onde os valores $L_1^{(II)} = 232$ e $L_2^{(II)} = 8$ são usados.

Note que ambas análises LP são executadas no mesmo conjunto de amostras de fala. As janelas são aplicadas a 80 amostras do quadro de fala anterior e mais as 160 amostras do quadro de fala presente. Não são usadas amostras de quadros futuros. Um diagrama das duas janelas de análise LP é mostrado na Figura 3-8.

As auto-correlações da fala janelada $s\prime(n)$, $n = 0,\ ...,\ 239$, são calculadas por:

$$r_{ac}(k) = \sum_{n=k}^{239} s\prime(n)s\prime(n-k), \qquad k = 0,\ ...,\ 10, \tag{3.11}$$

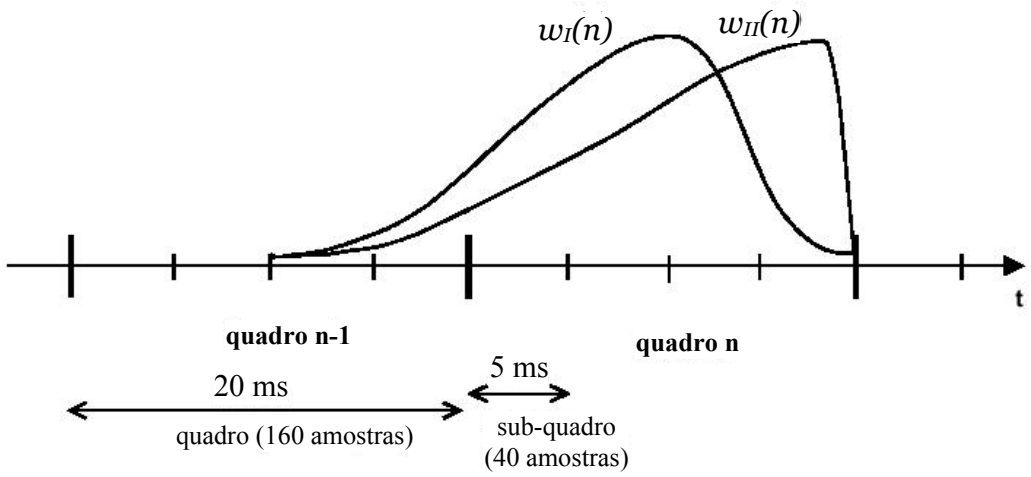


Figura 3-8: Janelas de análise LP.

e uma expansão de largura de banda de 60 Hz é usada janelando-se periodicamente as auto-correlações por meio da função:

$$w_{lag}(i) = \exp\left[-\frac{1}{2}\left(\frac{2f_0 i}{f_s}\right)^2\right], \qquad i = 1, ..., 10, \tag{3.12}$$

onde $f_0 = 60$ Hz é o fator de expansão da largura de banda e $f_s = 8000$ Hz é a freqüência de amostragem. Além disso, $r_{ac}(0)$ é multiplicado pelo fator de correção do ruído branco 1,0001, que é equivalente a somar um limiar inferior de ruído em -40 dB.

Outros Modos

A análise LP é executada uma vez por quadro usando-se uma janela assimétrica. A janela tem seu peso concentrado no quarto sub-quadro e consiste de duas partes: a primeira parte é a metade de uma janela de Hamming e a segunda parte é um quarto de um ciclo de função cosseno. A janela é dada por (3.10), onde os valores $L_1 = 200$ e $L_2 = 40$ são usados.

As auto-correlações da fala janelada $s\prime(n)$, $n = 0, ..., 239$, são calculadas por (3.11), e uma expansão de largura de banda de 60 Hz é usada janelando-se periodicamente as auto-correlações usando-se a janela em (3.12). Além disso, $r_{ac}(0)$ é multiplicado pelo

fator de correção do ruído branco 1,0001, que é equivalente a somar um limiar inferior de ruído em -40 dB.

Algoritmo Levinson-Durbin

As auto-correlações modificadas $r\prime_{ac}(0) = 1,0001 r_{ac}(0)$ e $r\prime_{ac}(k) = r_{ac}(k) w_{lag}(k)$, $k = 1, ..., 10$, são usadas para obter os coeficientes LPC do filtro na forma direta a_k, $k = 1, ..., 10$, resolvendo-se o seguinte conjunto de equações:

$$\sum_{k=1}^{10} a_k r\prime_{ac}(|i-k|) = -r\prime_{ac}(i), \qquad i = 1, ..., 10. \tag{3.13}$$

O conjunto de equações em (3.13) é resolvido através do algoritmo de Levinson-Durbin. Este algoritmo usa a seguinte recursão em pseudo código de programação:

$E_{LD}(0) = r\prime_{ac}(0)$

$for\, i = 1\, to\, 10\, do$

$\quad a_0^{(i-1)} = 1$

$\quad k_i = -\left[\sum_{j=0}^{i-1} a_j^{(i-1)} r\prime_{ac}(i-j)\right] / E_{LD}(i-1)$

$\quad a_i^{(i)} = k_i$

$\quad for\, j = 1\, to\, i-1\, do$

$\quad\quad a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}$

$\quad end$

$\quad E_{LD}(i) = (1 - k_i^2) E_{LD}(i-1)$

end

A solução final é dada como $a_j = a_j^{(10)}$, $j = 1, ..., 10$.

Capítulo 4

Recuperação de Quadros Perdidos

Este capítulo trata em detalhes do esquema executado pelo codificador GSM-AMR para substituição e silenciamento de quadros perdidos. É apresentado também o esquema de interpolação como uma proposta alternativa para o tratamento da perda de quadros. Este esquema consiste na interpolação de todos os parâmetros transmitidos pelo codec GSM-AMR, com exceção dos índices do dicionário. Estes são utilizados na decodificação de acordo com a descrição da norma do codec. Resultados comparativos entre os dois esquemas são apresentados no próximo capítulo.

4.1 Substituição e Silenciamento de Quadros Perdidos

Esta seção define o procedimento de silenciamento e substituição de quadros perdidos, usado pelo codificador GSM-AMR. A finalidade da substituição é atenuar e ocultar o efeito dos quadros perdidos. A finalidade de silenciar a saída, no caso de muitos quadros perdidos, é indicar a interrupção do canal ao usuário e evitar a geração de possíveis sons importunos como um resultado do procedimento de substituição de quadro.

A indicação de quadros perdidos de fala ou de quadros descritores de inserção de

silêncio (SID) é feita por um sub-sistema de rádio (codificador de canal) que configura um *flag* (BFI - *Bad Frame Indication*), baseando-se em controles de redundância cíclica (CRC) ou em outros mecanismos de detecção de erro, cujas possibilidades estão previstas em [37]. Se o *flag* BFI estiver levantado, o decodificador executará substituição de parâmetros para ocultação de erros.

A decodificação normal de quadros perdidos de fala resultaria em efeitos ruidosos muito desagradáveis. Para aumentar a qualidade subjetiva, quadros de fala perdidos são substituídos tanto por repetição como por extrapolação de bom(ns) quadro(s) de fala anterior(es). Esta substituição é feita para que o nível de saída diminua gradualmente, resultando em silêncio a partir de um determinado número de quadros perdidos.

4.1.1 Máquina de Estado do Esquema de Substituição e Silenciamento

O exemplo de solução por substituição e silenciamento, descrito em [37], é baseado em uma máquina de estado com sete estados (Figura 4-1). O sistema começa no estado 0. Cada vez que um quadro ruim é detectado, o contador de estado é incrementado de um e é saturado quando alcança 6. Cada vez que um quadro bom de fala é detectado, o contador de estado é levado a zero, exceto quando a máquina está no estado 6, quando o contador é levado ao estado 5. O estado implica na qualidade do canal: quanto maior for o valor do contador de estado, pior será a qualidade de canal. O controle de fluxo da máquina de estado pode ser descrito pelo seguinte código em linguagem C (**BFI** = indicador de quadro ruim, **Estado** = variável de estado):

```
if(BFI!=0)
    Estado=Estado+1;
else if(Estado==6)
    Estado=5;
else
    Estado=0;
```

```
if(Estado>6)
    Estado=6;
```

O *flag* BFI do quadro anterior (**prevBFI**) também é verificado na execução da máquina de estado. Desta forma, o processamento é dependente do valor da variável **Estado**, bem como dos *flags* **BFI** e **prevBFI**.

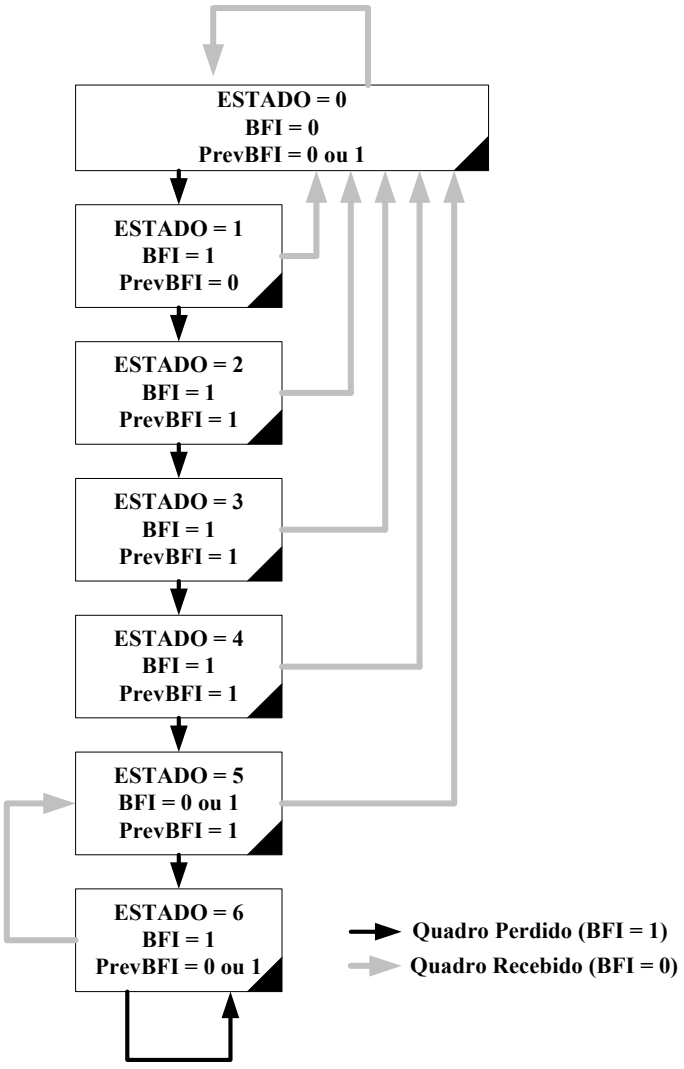


Figura 4-1: Máquina de Estado para Controlar a Substituição de um Quadro Perdido. (Adaptado de ETSI EN 301 705, 1998).

A seguir são apresentadas algumas configurações de estados e os respectivos procedimentos adotados no decodificador do GSM-AMR. Estas ações são descritas na norma como ações para um quadro com atividade de fala[1]. Existem também ações para quadros sem atividade[2] e para quadros descritores de silêncio, mas como estes fogem do escopo deste trabalho, somente uma visão geral sobre os mesmos será vista mais adiante.

4.1.2 Ações para Quadros com Atividade de Fala

BFI=0, prevBFI=0, Estado=0

Nesta condição, não é detectado erro no quadro de fala recebido ou no anteriormente recebido. Os parâmetros de fala recebidos são usados normalmente na síntese da fala. O quadro atual de parâmetros de fala é salvo.

BFI = 0, prevBFI = 1, Estado = 0 ou 5

Nesta condição, não é detectado erro no quadro de fala recebido, mas o quadro de fala recebido anteriormente foi ruim. O ganho LTP e o ganho do dicionário fixo são limitados a valores abaixo dos valores usados pelo último sub-quadro bom recebido.

$$g^p = \begin{cases} g^p, & g^p \leq g^p(-1) \\ g^p(-1), & g^p > g^p(-1) \end{cases} \tag{4.1}$$

onde g^p é o ganho atual LTP decodificado, $g^p(-1)$ é o ganho LTP usado para o último sub-quadro bom (BFI = 0), e

$$g^c = \begin{cases} g^c, & g^c \leq g^c(-1) \\ g^c(-1), & g^c > g^c(-1) \end{cases} \tag{4.2}$$

[1]Entende-se por quadro com atividade de fala, um quadro cujo conteúdo carrega informações relativas a fonemas sonoros.

[2]Entende-se por quadro sem atividade de fala, um quadro cujo conteúdo carrega informações relativas a fonemas surdos.

onde g^c é o ganho atual do dicionário fixo decodificado e $g^c(-1)$ é o ganho do dicionário fixo usado para o último sub-quadro bom (BFI = 0).

O resto dos parâmetros recebido é usado normalmente na síntese da fala. O quadro atual de parâmetros de fala é salvo.

BFI = 1, prevBFI = 0 ou 1, Estado = 1 ... 6

Nesta condição, um erro é detectado no quadro de fala recebido e inicia-se o procedimento de substituição e silenciamento. O ganho LTP e o ganho do dicionário fixo são substituídos por valores atenuados dos sub-quadros anteriores:

$$g^p = \begin{cases} P(estado)\, g^p(-1), & g^p(-1) \leq median5(g^p(-1), ..., g^p(-5)) \\ P(estado)\, median5(g^p(-1), ..., g^p(-5)), & g^p(-1) > median5(g^p(-1), ..., g^p(-5)) \end{cases} \tag{4.3}$$

onde g^p é o ganho atual LTP decodificado, $g^p(-1), ..., g^p(-n)$ são os ganhos LTP usados para os últimos n sub-quadros, $median5()$ representa a operação mediana de 5 valores, $P(estado)$ é o fator de atenuação ($P(1) = 0,98;\ P(2) = 0,98;\ P(3) = 0,8;\ P(4) = 0,3;\ P(5) = 0,2;\ P(6) = 0,2$), $estado$ é o número do estado, e

$$g^c = \begin{cases} C(estado)\, g^c(-1), & g^c(-1) \leq median5(g^c(-1), ..., g^c(-5)) \\ C(estado)\, median5(g^c(-1), ..., g^c(-5)), & g^c(-1) > median5(g^c(-1), ..., g^c(-5)) \end{cases} \tag{4.4}$$

onde g^c é o ganho atual do dicionário fixo decodificado, $g^c(-1), ..., g^c(-n)$ são os ganhos do dicionário fixo usados para os últimos n sub-quadros e $C(estado)$ é o fator de atenuação ($C(1) = 0,98;\ C(2) = 0,98;\ C(3) = 0,98;\ C(4) = 0,98;\ C(5) = 0,98;\ C(6) = 0,7$).

Quanto maior é o valor do estado, maior é a atenuação dos ganhos. A memória do ganho de predição do dicionário fixo também é atualizada usando o valor médio dos quatro últimos valores na memória:

$$ener(0) = \frac{1}{4}\sum_{i=1}^{4} ener(-i) \tag{4.5}$$

Os últimos LSFs são deslocados em direção à sua média pela relação

$$lsf\_q1(i) = lsf\_q2(i) = \alpha.past\_lsf\_q(i) + (1-\alpha)mean\_lsf(i), \quad i = 0 \ldots 9 \tag{4.6}$$

onde $\alpha = 0,95$ (valor descrito pela norma; porém, o valor usado no código de referência do ETSI é 0,90), lsf_q1 e lsf_q2 são dois conjuntos de vetores LSF para o quadro atual, $past\_lsf\_q$ é o lsf_q2 do quadro anterior, e $mean\_lsf$ é o vetor LSF médio. Dois conjuntos de vetores LSFs são disponibilizados apenas no modo AMR_12.20, para os demais modos há apenas um vetor de parâmetros LSF.

Atualização do Período de Pitch

Os valores do período de *pitch* são substituídos pelo valor do quarto sub-quadro do último quadro (modo AMR_12.20), ou por valores levemente modificados baseados no último valor recebido corretamente (todos os outros modos).

Sinal de Excitação

Os índices do dicionário carregam informação de amplitude e posição do sinal de excitação do filtro de síntese. Os pulsos de excitação do dicionário fixo são usados pelo decodificador, mesmo quando há ocorrência de dados corrompidos. No caso da perda desses dados, índices aleatórios do dicionário fixo são empregados.

4.1.3 Ações para Quadros Sem Atividade e Quadros Descritores de Silêncio

As ações realizadas para quadros sem atividade têm por objetivo reduzir a geração de espúrios sonoros. Isto ocorre quando o procedimento convencional para quadros com

atividade é executado de forma inadequada em sinais não sonoros.

Quando há ocorrência de dados corrompidos em um quadro descritor de inserção de silêncio (SID), a informação deste quadro é substituída pela última informação válida de quadros SID (armazenada em memória), sendo então aplicado o procedimento convencional de decodificação de quadros SID. Se o tempo entre as atualizações das informações dos quadros SID é maior do que um segundo, ocorre um procedimento de atenuação aplicado aos parâmetros armazenados [38].

4.2 Interpolação de Parâmetros para Construção de Quadro(s) que Substitua(m) Quadro(s) Perdido(s)

Com o desenvolvimento da tecnologia, cada vez mais se pensa em utilizar a estrutura da Internet para transmissão de mídias contínuas, como voz e vídeo. As redes IP, entretanto, apresentam graves problemas quando se trata de transmissão de mídias sensíveis a atraso. A transmissão de voz por pacotes em tempo-real, por exemplo, deve satisfazer alguns requisitos tais como garantir um atraso máximo para cada pacote, uma variação máxima do atraso dos pacotes e uma taxa máxima de perda de pacotes. Quando um ou mais pacotes são perdidos e nenhuma providência é tomada na tentativa de recuperá-los, a qualidade perceptual da fala recebida fica significativamente deteriorada.

Um dos desafios da telefonia IP está relacionado ao controle da variação de atraso (*jitter*) dos pacotes recebidos [39]. O problema de *jitter* precisa ser tratado até mesmo nas redes mais recentes, caracterizadas por apresentarem grande largura de banda e valores reduzidos de atraso. Isto se deve ao fato de que os pacotes normalmente são submetidos a valores de atraso diferentes durante o percurso de transmissão. Como a reprodução da fala precisa ser executada a intervalos regulares para que uma qualidade perceptual aceitável seja alcançada, torna-se essencial o armazenamento de alguns pacotes em um

buffer na recepção. Tal armazenamento permite que processos de recuperação de pacotes, como o de interpolação por exemplo, sejam executados sobre os pacotes recebidos.

Os codecs têm por característica transmitir apenas alguns parâmetros para a síntese do sinal de fala. Parâmetros, tais como coeficientes LPC e período de *pitch*, apresentam como propriedade uma diferença suave entre valores de quadros subseqüentes [40]. A partir desta informação, surge a idéia do desenvolvimento de um estudo baseado na interpolação dos parâmetros de um dos codecs estudado atualmente para telefonia IP: o GSM-AMR. Tal estudo visa avaliar a qualidade da fala provida pelo esquema de recuperação de quadros descrito na seção anterior, e a qualidade deste sinal mediante a utilização do esquema de interpolação.

O esquema foi aplicado sobre os coeficientes LPC, período e ganho de *pitch*, bem como para o ganho do dicionário fixo. A Figura 4-2 mostra a interpolação linear empregada sobre os parâmetros dos quadros perfeitamente recebidos. Os índices do dicionário fixo foram utilizados de acordo com o esquema proposto pela norma do codificador GSM-AMR, visto no sub-item *Sinal de Excitação* da seção 4.1.2. O algoritmo de interpolação foi restrito a taxa de transmissão de 7,95 kbps. Devido à não-estacionariedade do sinal de fala, o esquema foi configurado de forma a permitir um número máximo de três quadros perdidos consecutivamente. A partir deste número, o esquema de tratamento do codificador GSM-AMR é assumido para construção dos quadros. Aplica-se ao modelo de simulação a restrição de que o primeiro quadro sempre será um quadro bom. E no caso da perda do último quadro, o programa empregará parâmetros reconstruídos pelo decodificador do codec GSM-AMR.

O esquema aplicado, no caso dos parâmetros LPC, é uma interpolação linear simples entre os vetores dos parâmetros LSP extraídos dos quadros adjacentes ao(s) quadro(s) perdido(s). Os parâmetros LSP são uma representação dos parâmetros LPC no domínio da freqüência. O codificador GSM-AMR já faz a conversão para tal domínio, pelo fato de já realizar uma interpolação no momento da quantização dos parâmetros. Os vetores dos parâmetros LSP possuem dez elementos, haja vista que o filtro de predição linear do

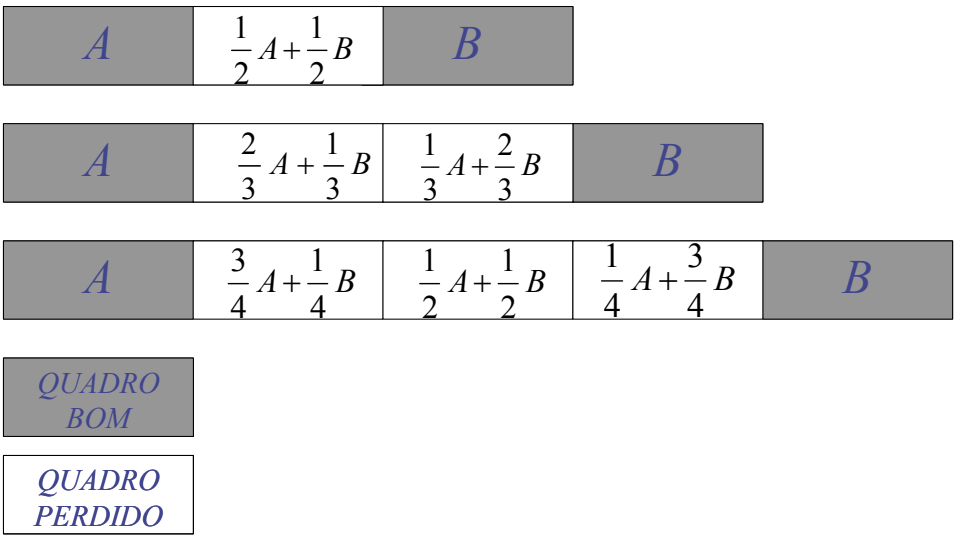


Figura 4-2: Esquema de Interpolação.

codec é de décima ordem.

Na decodificação dos parâmetros LPC ocorre um problema inerente à codificação dos parâmetros LSFs entre quadros. É que tal codificação causa erro de propagação devido à mudança dos estados do codificador na ocorrência de quadros perdidos [41]. O diagrama de blocos da Figura 4-3 ilustra o problema de propagação do erro. A atualização dos estados é função dos parâmetros transmitidos pelos quadros; quando há perda de quadros na rede, uma atualização com dados incorretos irá gerar estados atualizados também incorretos. Como os parâmetros LSPs do quadro seguinte são função dos estados atualizados do quadro atual, mesmo que os parâmetros daquele quadro sejam recebidos corretamente, o erro de atualização dos estados do quadro perdido acaba se refletindo nos parâmetros LSPs gerados pelo próximo quadro. Vale a pena salientar que, para o codec GSM-AMR, este erro é refletido somente no quadro subseqüente ao quadro perdido; o esquema de desvanecimento do erro não permite propagação para quadros posteriores.

Este problema interfere diretamente no resultado dos parâmetros LPC interpolados. Para efeito de simulação, os estados gerados na codificação foram transmitidos para que, na decodificação, fossem usados corretamente quando fosse detectada a perda de um

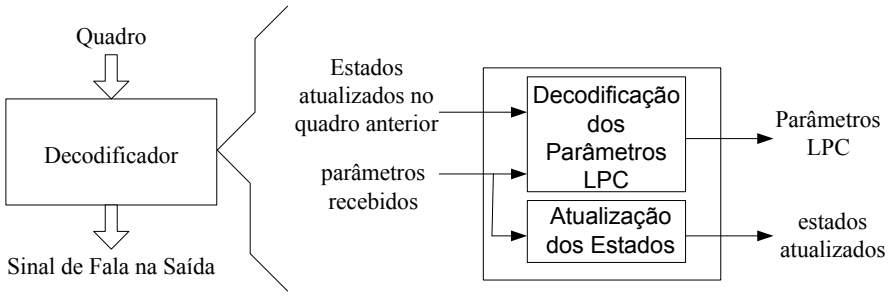


Figura 4-3: Diagrama de Blocos para Erro de Propagação.

quadro. Na prática, esta transmissão seria responsável por um acréscimo na taxa de transmissão de bits. Devido a norma do codificador GSM-AMR especificar somente o número de bits dos parâmetros transmitidos, não há como calcular de quanto seria o acréscimo causado pelo envio dos estados. O mesmo problema também ocorre durante a interpolação do ganho do dicionário.

4.2.1 Simulações

A rede de voz sobre pacote é simulada de tal forma que cada pacote contenha um quadro. A perda de pacote é aproximada por um processo aleatório de Markov que enfatiza a natureza em rajadas desta perda na rede. A Figura 4-4 ilustra o diagrama de estados do processo aleatório de Markov.

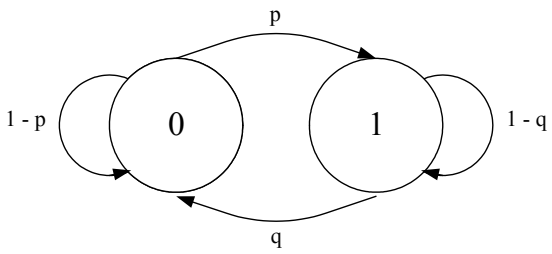


Figura 4-4: Perda de Pacotes Modelada por um Processo Aleatório de Markov.

No diagrama de estados, o estado "0" representa um pacote recebido corretamente

e o estado "1" representa a perda de um pacote. Supõe-se que p seja a probabilidade de transição de "0" para "1" e que q seja a probabilidade de "1" para "0" [41]. Três taxas diferentes de perda são simuladas como listado na Tabela 4.1. O sinal comparativo utilizado na métrica de distorção é o sinal decodificado sem perda de pacote.

Taxa(%)	**p**	**q**
10	0,1	0,85
20	0,2	0,70
30	0,3	0,65

Tabela 4.1: Taxas Simuladas de Perda.

Os sinais de fala usados nas simulações são da base *"Telephone Network Acoustic-Phonetic Continuous Speech Corpus"* - NTIMIT [42]. Esta base de dados é constituída por um conjunto de 2680 arquivos de sinais de voz, gravados após transmissão em linhas telefônicas reais. A base é dividida em oito grupos (DR1-DR8), onde cada grupo corresponde a uma região de gravação diferente [43]. Para um levantamento estatístico da medida de distorção empregada, dez frases da base foram utilizadas: cinco com vozes masculinas e cinco com vozes femininas. A Tabela 4.2 apresenta os sinais de fala usados nas simulações.

Todos os arquivos da base NTIMIT estão gravados com taxa de amostragem de 16 kHz e quantização de 16 bits por amostra. Como o codificador GSM-AMR toma sinais amostrados a 8 kHz em sua entrada, os sinais foram filtrados e dizimados para esta freqüência de amostragem, usando-se o seguinte programa para um filtro FIR passa-baixa, projetado no MATLAB^{TM}:

```
f=[0 0.425 0.5 1];
m=[1 1 0 0];
b=remez(96,f,m);
```

A resposta em frequência deste filtro pode ser observada na Figura 4-5. O filtro tem 97 coeficientes, uma freqüência de corte de aproximadamente 3400 Hz e oferece 64 dB de atenuação em 4000 Hz. A relação sinal ruído entre os sinais original e filtrado é no mínimo 30 dB para a base empregada nos testes, mostrando que quase toda energia

Região	Arquivo	Frase	Gênero	Duração (s)
DR1	SI923.WAV	"To many experts, this trend was inevitable."	Masculino	2,961
DR1	SX113.WAV	"A muscular abdomen is good for your back."	Masculino	3,327
DR2	SX134.WAV	"December and January are nice months to spend in Miami."	Masculino	3,084
DR2	SX275.WAV	"Steve wore a bright red cashmere sweater."	Masculino	2,957
DR2	SX95.WAV	"Iguanas and alligators are tropical reptiles."	Masculino	2,930
DR1	SI1894.WAV	"My father ran him off here six years ago."	Feminino	2,943
DR1	SX4.WAV	"Jane may earn more money by working hard."	Feminino	3,014
DR2	SX115.WAV	"The emblem depicts the Acropolis all aglow."	Feminino	2,962
DR2	SX50.WAV	"Catastrophic economic cutbacks neglect the poor."	Feminino	3,102
DR2	SX284.WAV	"Jeff thought you argued in favor of a centrifuge purchase."	Feminino	3,257

Tabela 4.2: Sinais de Fala Utilizados nas Simulações.

do sinal está dentro da banda de freqüência do canal de telefonia. Aos sinais da base NTIMIT ainda foi aplicado um ganho de dois, como exige os esquemas dos codificadores de voz.

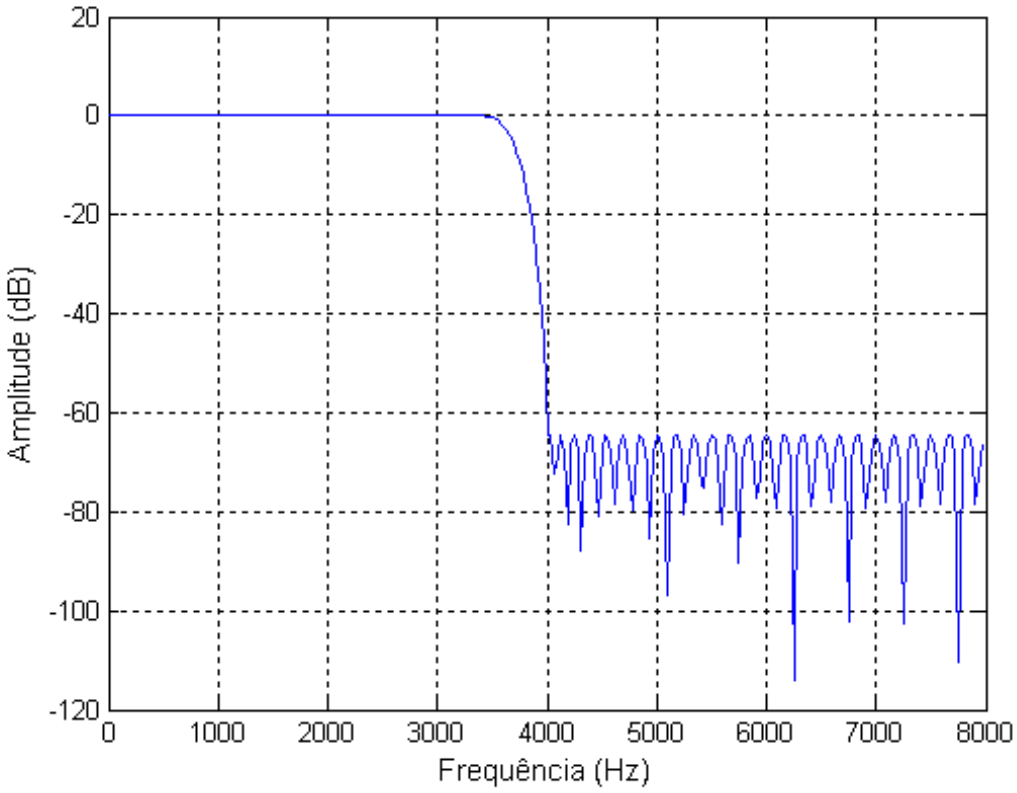


Figura 4-5: Resposta em Frequência do Filtro FIR Passa-Baixa.

Capítulo 5

Resultados

5.1 Métricas para Avaliação da Qualidade de Fala

Em geral, os métodos usados para avaliar qualidade de voz podem ser divididos em duas categorias:

· Medidas Subjetivas de Distorção.

· Medidas Objetivas de Distorção.

Os testes subjetivos têm uma confiabilidade maior devido à qualidade perceptual a ser observada no sinal de voz. Normalmente um grupo de pessoas analisa e compara os sinais de fala codificado e original, e uma nota é atribuída ao sinal codificado de acordo com o grau de semelhança deste com o original. Entretanto, as medidas subjetivas além de custosas e trabalhosas demandam bastante tempo para serem realizadas. Neste contexto, as medidas objetivas de distorção surgem como uma alternativa interessante para avaliação da qualidade do sinal de fala. Estas duas medidas são discutidas abaixo em detalhes.

5.1.1 Medidas Subjetivas de Distorção

Em testes subjetivos de distorção, geralmente a qualidade da fala é medida através da *inteligibilidade*, que é definida tipicamente como a porcentagem de palavras ou fonemas

corretamente ouvidos. Juntamente com a inteligibilidade, a naturalidade é um dos aspectos perceptuais mais significantes para um sinal de fala. Os testes subjetivos podem ser executados de várias formas diferentes. Um dos mais conhecidos é o MOS (*Mean Opinion Score*), descrito pela recomendação P.800 do ITU-T [44][45].

MOS - *Mean Opinion Score*

A qualidade de voz é comumente caracterizada por pontuações subjetivas geradas a partir de testes controlados. Nas pontuações MOS os resultados para um sistema em teste são sempre comparados com uma referência bem estabelecida. O método recomendado para testes do tipo "somente ouvir" é o ACR: *Absolute Category Rating*. Vários tipos de escalas de cinco pontos são usados nos testes ACR. A escala mostrada na Tabela 5.1 é usada freqüentemente para aplicações do ITU-T, sendo também recomendada para testes em sistemas.

Pontuação	**Qualidade de Fala**	**Nível de Distorção**
5	Excelente	Imperceptível
4	Boa	Apenas perceptível mas não importuna
3	Razoável	Perceptível e levemente importuna
2	Pobre	Importuna mas não condenável
1	Insatisfatória	Muito Importuna e condenável

Tabela 5.1: Escala MOS de Qualidade de Fala.

Testes MOS são utilizados na avaliação da qualidade dos diferentes codecs existentes na telefonia IP. O valor numérico desta métrica permite uma comparação direta com medidas objetivas, mas não ajuda a entender a causa da distorção. Nesta escala, um índice de 4.0 é considerado de ótima qualidade, igual à qualidade da voz ouvida em uma linha telefônica convencional.

5.1.2 Medidas Objetivas de Distorção

Às vezes torna-se necessária uma avaliação imediata e confiável da qualidade de um sinal de fala. Durante a fase de desenvolvimento de um novo algoritmo, por exemplo,

não seria viável a utilização de uma medida subjetiva devido às desvantagens que estas apresentam, principalmente em relação ao tempo demandado para sua execução. Assim, o uso de medidas objetivas de distorção surge como uma alternativa interessante para avaliação da qualidade de voz. Estas medidas podem ser calculadas tanto no domínio do tempo quanto no domínio da freqüência.

Dentre as principais medidas objetivas de distorção no domínio do tempo, pode-se citar a relação sinal ruído (SNR) e a SNR em segmentos. No domínio da freqüência, o espectro LPC do sinal original e o espectro LPC do sinal interpolado ou quantizado são comparados. A distorção ou diferença entre os dois espectros afeta a percepção do som. Exemplos de medidas de distorção no domínio da freqüência são a distorção log espectral e a distância euclidiana ponderada.

Atualmente existem métodos objetivos que apresentam resultados mais satisfatórios do que os convencionais descritos anteriormente. Tais métodos levam em consideração as diversas características peculiares a audição humana, como os efeitos de mascaramento, e avaliam perceptualmente a qualidade do sinal [46]. Como exemplo dessas medidas, pode-se citar o PESQ (*Perceptual evaluation of speech quality*) [47] e o PSQM (*Perceptual Speech-Quality Measure*) [48].

Neste trabalho, utilizou-se uma medida objetiva de distorção para calcular a qualidade do algoritmo de codificação. A distorção espectral foi utilizada tanto para calcular a diferença entre os espectros dos parâmetros LPC originais e interpolados, quanto para calcular a diferença entre os espectros dos quadros construídos pela interpolação dos demais parâmetros e os quadros construídos pelo esquema de recuperação do codec GSM-AMR.

Além de trabalhar em uma cadência de quadros, a distorção espectral é uma medida bastante utilizada em literaturas referentes à qualidade de fala. Assim, esta medida foi adotada para que se tivesse uma comparação mais precisa com os trabalhos já publicados pela comunidade. A seguir é mostrada a definição de distorção espectral.

Distorção Espectral

A distorção espectral, para um dado quadro, é definida como a raiz quadrada média da diferença entre o log espectro de potência LPC original, e o log espectro de potência LPC quantizado ou interpolado [49]. Geralmente, é calculada a média da distorção espectral sobre um grande número de quadros, e esta é usada como uma medida de desempenho de quantização ou interpolação. A definição matemática da distorção espectral comum para um quadro i é a seguinte:

$$SD_i = \sqrt{\frac{1}{F_s}\int_0^{F_s}\left[10\log_{10}\frac{S_i(f)}{\hat{S}_i(f)}\right]^2 df} \quad (dB), \tag{5.1}$$

onde, F_s é a freqüência de amostragem, $S_i(f)$ e $\hat{S}_i(f)$ são os espectros de potência do i-ésimo quadro dado por,

$$S_i(f) = \frac{1}{\left|A_i(e^{j2\pi f/F_s})\right|^2} \tag{5.2}$$

$$\hat{S}_i(f) = \frac{1}{\left|\hat{A}_i(e^{j2\pi f/F_s})\right|^2} \tag{5.3}$$

onde $A_i(z)$, $\hat{A}_i(z)$ são os polinômios LPC original e interpolado (definidos em (3.4)), respectivamente, para o i-ésimo quadro.

Para este trabalho, utilizou-se uma versão discreta da distorção espectral. O cálculo do espectro de energia usa uma FFT de 512 pontos. As freqüências inferior e superior usadas foram, respectivamente, 125 Hz e 3400 Hz. A SD_i é calculada como um somatório sobre pontos uniformemente espaçados entre estes limites de freqüência. Isto pode ser expresso como [50]

$$SD_i = \sqrt{\frac{1}{n_1 - n_0}\sum_{n=n_0}^{n_1-1}\left[10\log_{10}\frac{S(e^{\frac{j2\pi n}{N}})}{\hat{S}(e^{\frac{j2\pi n}{N}})}\right]^2} \quad (dB) \tag{5.4}$$

Como a distorção espectral foi utilizada para medir qualidade tanto para os espectros

LPC, quanto diretamente para os espectros dos sinais; convencionou-se como SD1 a distorção espectral para os parâmetros LPC, e como SD2 a distorção espectral para o outro caso. A diferença é que o algoritmo que calcula a SD1 recebe os coeficientes LSP como parâmetros de entrada, e a SD2 é calculada tendo o sinal de fala decodificado como parâmetro de entrada. Além disso, observou-se a porcentagem de ocorrência de quadros *outliers* (quadros com distorção espectral grande), que também afetam a qualidade dos sinais de fala. Há dois tipos de quadros *outliers*:

· Os quadros tendo SD na faixa de 2-4dB (*outlier* tipo 1).

· Os quadros tendo SD maior que 4dB (*outlier* tipo 2).

Os resultados mostrados a seguir estão divididos da seguinte forma: resultados somente para os parâmetros LPC sem e com utilização dos estados de atualização; e resultados para interpolação de todos os parâmetros, com exceção dos índices do dicionário, sem e com utilização dos estados de atualização. Em todos os casos há uma comparação entre o método de interpolação proposto e o esquema de recuperação já existente no codec GSM-AMR.

5.2 Resultados

As simulações foram realizadas em dois microcomputadores PC Dell com sistema operacional *Windows 2000*, e com processadores Intel Pentium 4 e Pentium III com frequências de 1,7 GHz e 1 GHz respectivamente. A plataforma de simulação utilizada é composta pelo compilador da Texas Instruments^{TM}, Code Composer Studio versão 2.00, configurado no modo *Simulator* para DSPs das famílias C62xx e C67xx. O tempo médio para execução de uma simulação[1] foi de cerca de duas horas. O código fonte do codec GSM-AMR foi adquirido junto ao ETSI (*European Telecommunications Standards Ins-*

[1]Durante a simulação são executados, para todas as taxas de perda: a codificação do sinal, a decodificação do sinal sem a utilização do esquema de interpolação, o tratamento do sinal pelo esquema de interpolação, a decodificação do sinal tratado pelo esquema de interpolação e o cálculo da distorção espectral para os dois sinais decodificados.

titute). Tal código sofreu algumas alterações para que fossem inseridos os parâmetros interpolados, e para que atualizações de estado fossem executadas adequadamente.

As tabelas a seguir apresentam as médias aritméticas bem como todos os resultados obtidos sobre os dez sinais de fala utilizados nas simulações. As médias são relativas a quatro situações de interpolação: somente dos parâmetros LPC sem e com utilização dos estados gerados na codificação; e dos demais parâmetros sem e com utilização dos estados gerados na codificação. No caso dos parâmetros LPC, os sinais comparativos utilizados para calcular a distorção espectral, nas diferentes taxas de perda, foram os coeficientes LSP decodificados dos sinais recebidos sem perda de pacotes. Os sinais decodificados com os demais parâmetros interpolados tiveram como sinal alvo, para o cálculo da distorção espectral, os sinais decodificados de fala sem perda de pacotes.

A Figura 5-1 mostra o gráfico da distorção espectral em termos das taxas de perda de pacotes. Este resultado é referente à interpolação dos parâmetros LPC, sem a utilização dos estados atualizados gerados na codificação. O gráfico representa a média dos resultados dos dez sinais de fala codificados nas simulações. Observa-se que o esquema de interpolação alcança uma distorção espectral média 0,03 - 0,16 dB menor.

A Tabela 5.2 apresenta os valores da distorção espectral relacionados ao gráfico da Figura 5-1, assim como a porcentagem de ocorrência de quadros *outliers* para as respectivas taxas de perda.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
10	0,289976	8,37	2,10	0,263640	9,09	1,45
20	0,645217	18,22	5,45	0,556355	17,87	4,09
30	0,936867	21,73	9,77	0,780929	22,29	7,34

Tabela 5.2: Outliers da Distorção Espectral LPC (sem utilização de estados).

Os resultados para os dez sinais de fala simulados são apresentados nas Tabelas 5.3 a 5.12.

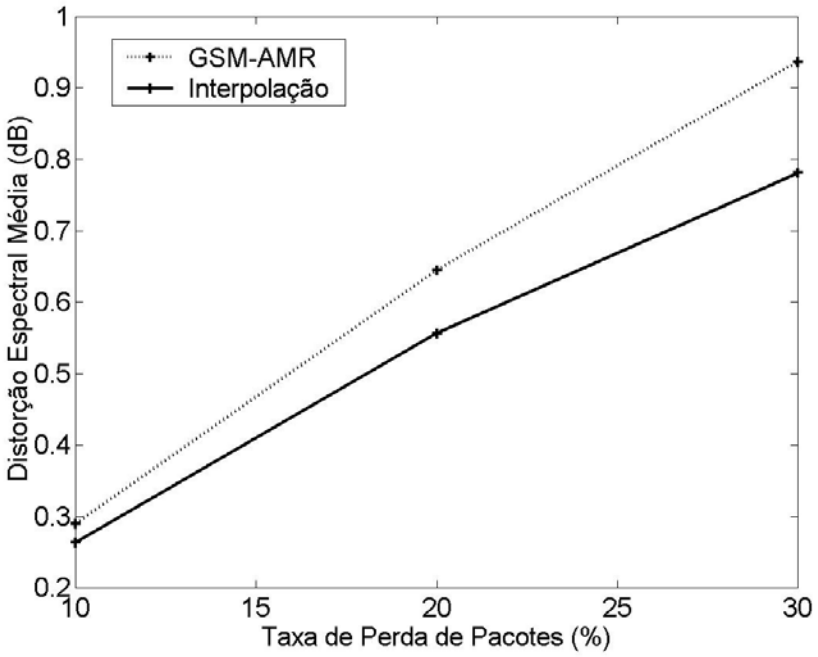


Figura 5-1: Distorção Espectral Média SD1 para Parâmetros LPC (sem utilização de estados).

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,064491	0,68	0,00	1,064491	0,68	0,00
10	1,333964	10,14	1,35	1,326690	11,49	0,68
20	1,636968	21,62	4,05	1,610265	22,30	2,70
30	1,774434	25,00	5,41	1,735092	27,03	4,73

Tabela 5.3: Sinal: Arquivo SI923.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,022501	0,60	0,00	1,022501	0,60	0,00
10	1,260803	8,43	1,81	1,206487	8,43	0,00
20	1,581212	17,47	4,82	1,468369	17,47	1,20
30	1,794565	23,49	6,63	1,596785	21,69	2,41

Tabela 5.4: Sinal: Arquivo SX113.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,109080	3,40	0,00	1,109080	3,40	0,00
10	1,417631	10,88	2,04	1,374952	12,24	1,36
20	1,863073	19,05	7,48	1,670508	17,69	4,76
30	2,203539	22,45	14,29	2,036094	22,45	10,88

Tabela 5.5: Sinal: Arquivo SI1894.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,132523	1,33	0,00	1,132523	1,33	0,00
10	1,451875	11,33	2,00	1,400984	11,33	1,33
20	1,778616	19,33	5,33	1,649447	18,67	3,33
30	2,088335	24,67	9,33	1,821936	21,33	6,67

Tabela 5.6: Sinal: Arquivo SX4.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	0,941216	0,68	0,00	0,941216	0,68	0,00
10	1,238542	9,46	2,03	1,217721	8,11	2,03
20	1,702459	20,27	6,08	1,574523	18,24	5,41
30	2,205524	22,30	13,51	1,962321	18,24	13,51

Tabela 5.7: Sinal: Arquivo SX115.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	0,940445	0,00	0,00	0,940445	0,00	0,00
10	1,211904	5,81	2,58	1,249754	8,39	2,58
20	1,511967	13,55	5,81	1,516262	17,42	4,52
30	1,797830	18,06	8,39	1,757215	23,87	7,74

Tabela 5.8: Sinal: Arquivo SX50.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
0	1,067241	1,95	0,00	1,067241	1,95	0,00
10	1,395458	9,74	2,60	1,376060	11,69	1,95
20	1,759246	19,48	7,14	1,710117	22,73	4,55
30	1,961362	21,43	9,74	1,816626	26,62	5,19

Tabela 5.9: Sinal: Arquivo SX134.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
0	1,094919	1,36	0,00	1,094919	1,36	0,00
10	1,373326	12,93	1,36	1,311786	11,56	1,36
20	1,655994	23,81	2,04	1,596635	19,73	4,08
30	1,969320	25,17	7,48	1,832615	27,21	6,12

Tabela 5.10: Sinal: Arquivo SX275.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
0	1,034027	1,23	0,00	1,034027	1,23	0,00
10	1,340082	8,64	2,47	1,324104	9,26	1,85
20	1,743976	20,99	5,56	1,625538	17,90	5,56
30	1,970695	24,07	9,88	1,873106	20,37	8,64

Tabela 5.11: Sinal: Arquivo SX284.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
0	1,060635	0,68	0,00	1,060635	0,68	0,00
10	1,343252	8,22	2,74	1,314943	10,27	1,37
20	1,685734	18,49	6,16	1,608967	18,49	4,79
30	2,070143	22,60	13,01	1,844583	26,03	7,53

Tabela 5.12: Sinal: Arquivo SX95.WAV da Base NTIMIT.

A Figura 5-2 mostra os resultados referentes à interpolação dos parâmetros LPC, com a utilização dos estados atualizados gerados na codificação. O gráfico representa a média dos resultados dos dez sinais de fala codificados nas simulações. Observa-se que o esquema de interpolação alcança uma distorção espectral média 0,09 - 0,31 dB menor.

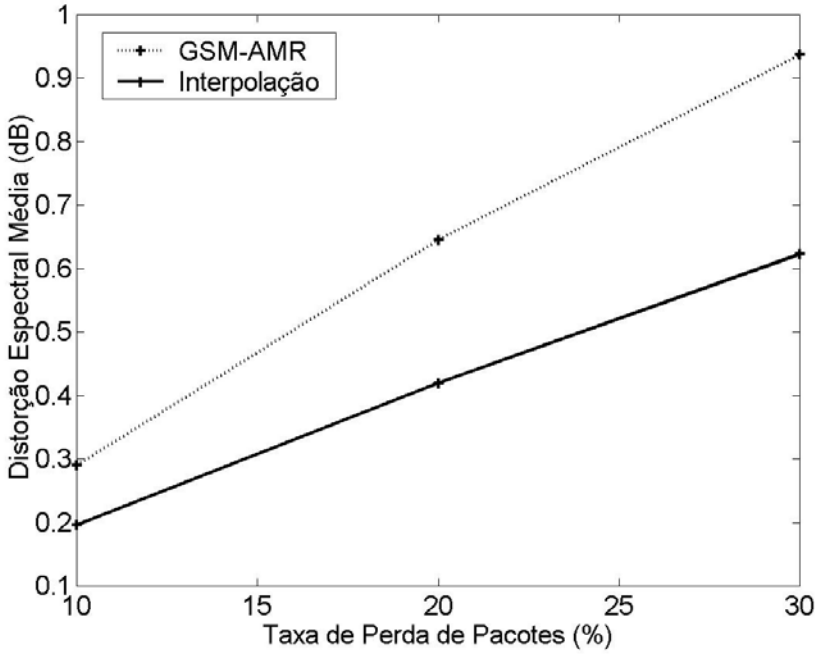


Figura 5-2: Distorção Espectral Média SD1 para Parâmetros LPC (com utilização de estados).

A Tabela 5.13 apresenta os valores da distorção espectral relacionados ao gráfico da Figura 5-2, assim como a porcentagem de ocorrência de quadros *outliers* para as respectivas taxas de perda.

Os resultados para os dez sinais de fala simulados são apresentados nas Tabelas 5.14 a 5.23.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
10	0,289976	8,37	2,10	0,196080	7,15	0,73
20	0,645217	18,22	5,45	0,419485	13,79	2,60
30	0,936867	21,73	9,77	0,623021	19,26	4,78

Tabela 5.13: Outliers da Distorção Espectral LPC (com utilização de estados).

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,064491	0,68	0,00	1,064491	0,68	0,00
10	1,333964	10,14	1,35	1,254608	8,11	0,68
20	1,636968	21,62	4,05	1,484615	16,22	3,38
30	1,774434	25,00	5,41	1,615598	22,97	4,05

Tabela 5.14: Sinal: Arquivo SI923.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,022501	0,60	0,00	1,022501	0,60	0,00
10	1,260803	8,43	1,81	1,145516	6,63	0,00
20	1,581212	17,47	4,82	1,335761	12,65	0,00
30	1,794565	23,49	6,63	1,452217	18,67	0,60

Tabela 5.15: Sinal: Arquivo SX113.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,109080	3,40	0,00	1,109080	3,40	0,00
10	1,417631	10,88	2,04	1,287923	8,16	1,36
20	1,863073	19,05	7,48	1,492923	12,93	3,40
30	2,203539	22,45	14,29	1,825344	18,37	8,84

Tabela 5.16: Sinal: Arquivo SI1894.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,132523	1,33	0,00	1,132523	1,33	0,00
10	1,451875	11,33	2,00	1,331593	10,00	0,67
20	1,778616	19,33	5,33	1,523067	16,00	2,00
30	2,088335	24,67	9,33	1,658000	19,33	4,00

Tabela 5.17: Sinal: Arquivo SX4.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	0,941216	0,68	0,00	0,941216	0,68	0,00
10	1,238542	9,46	2,03	1,165326	5,41	1,35
20	1,702459	20,27	6,08	1,425200	12,16	4,05
30	2,205524	22,30	13,51	1,766666	18,92	7,43

Tabela 5.18: Sinal: Arquivo SX115.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	0,940445	0,00	0,00	0,940445	0,00	0,00
10	1,211904	5,81	2,58	1,182354	9,03	1,29
20	1,511967	13,55	5,81	1,392803	16,13	2,58
30	1,797830	18,06	8,39	1,605685	21,94	3,87

Tabela 5.19: Sinal: Arquivo SX50.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,067241	1,95	0,00	1,067241	1,95	0,00
10	1,395458	9,74	2,60	1,290375	10,39	0,00
20	1,759246	19,48	7,14	1,555590	16,88	1,95
30	1,961362	21,43	9,74	1,667494	21,43	2,60

Tabela 5.20: Sinal: Arquivo SX134.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,094919	1,36	0,00	1,094919	1,36	0,00
10	1,373326	12,93	1,36	1,255163	8,16	0,00
20	1,655994	23,81	2,04	1,479759	14,97	2,04
30	1,969320	25,17	7,48	1,691542	20,41	4,08

Tabela 5.21: Sinal: Arquivo SX275.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,034027	1,23	0,00	1,034027	1,23	0,00
10	1,340082	8,64	2,47	1,247359	8,64	0,62
20	1,743976	20,99	5,56	1,471850	15,43	1,85
30	1,970695	24,07	9,88	1,713334	18,52	6,17

Tabela 5.22: Sinal: Arquivo SX284.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
0	1,060635	0,68	0,00	1,060635	0,68	0,00
10	1,343252	8,22	2,74	1,267658	8,90	1,37
20	1,685734	18,49	6,16	1,500362	16,44	4,79
30	2,070143	22,60	13,01	1,701407	23,97	6,16

Tabela 5.23: Sinal: Arquivo SX95.WAV da Base NTIMIT.

A Figura 5-3 mostra os resultados referentes à interpolação de todos os parâmetros, com exceção dos índices do dicionário fixo, sem a utilização dos estados atualizados gerados na codificação. O gráfico representa a média dos resultados dos dez sinais de fala codificados nas simulações. Observa-se que o esquema de interpolação alcança uma distorção espectral média 0,52 - 2,01 dB menor.

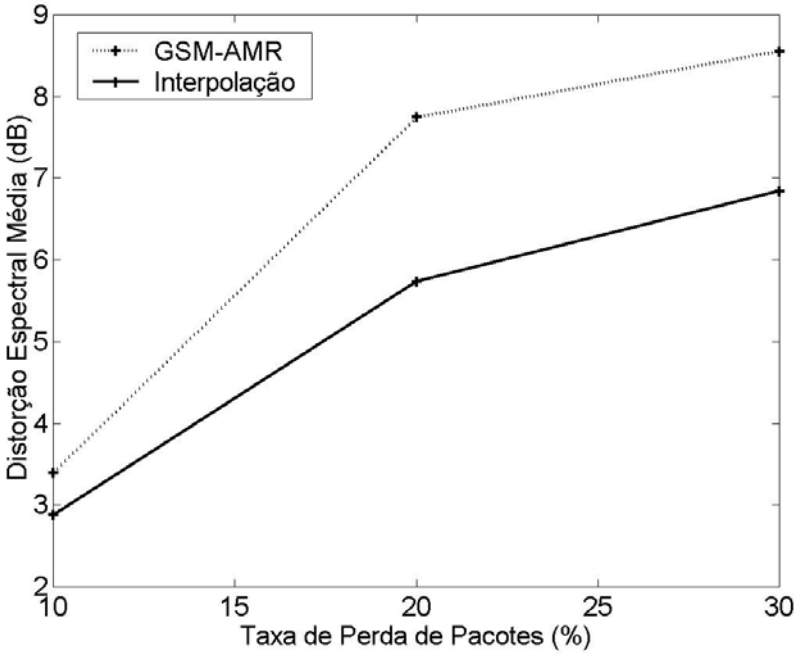


Figura 5-3: Distorção Espectral Média SD2 para Todos os Parâmetros Interpolados (exceto índices do dicionário fixo), sem Utilização de Estados.

A Tabela 5.24 apresenta os valores da distorção espectral relacionados ao gráfico da Figura 5-3, assim como a porcentagem de ocorrência de quadros *outliers* para as respectivas taxas de perda.

Os resultados para os dez sinais de fala simulados são apresentados nas Tabelas 5.25 a 5.34.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,393036	7,43	26,54	2,876574	7,29	26,42
20	7,747129	7,54	47,07	5,734861	7,47	47,47
30	8,552753	9,42	61,04	6,842916	9,23	61,10

Tabela 5.24: Outliers da Distorção Espectral para Todos os Parâmetros Interpolados (exceto índices do dicionário fixo), sem Utilização de Estados.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	2,880366	5,41	22,97	2,407330	4,05	23,65
20	7,809893	9,46	50,00	5,908055	8,78	50,68
30	7,535169	10,81	56,08	6,114266	10,81	56,08

Tabela 5.25: Sinal: Arquivo SI923.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	2,633641	6,63	24,70	2,344020	6,63	24,70
20	6,594780	8,43	51,20	5,383007	9,04	51,20
30	7,450713	4,82	63,86	6,266271	4,22	65,06

Tabela 5.26: Sinal: Arquivo SX113.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,542271	6,80	27,21	2,879048	6,12	27,21
20	7,314561	5,44	42,86	5,371230	6,80	42,86
30	8,836039	8,16	63,27	7,287371	6,80	62,59

Tabela 5.27: Sinal: Arquivo SI1894.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	4,170348	8,67	30,00	3,508656	8,00	30,00
20	8,979136	13,33	50,67	6,621398	12,00	50,67
30	9,527634	14,67	62,67	7,319121	14,00	62,67

Tabela 5.28: Sinal: Arquivo SX4.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	4,259584	8,78	29,73	3,538330	4,05	29,73
20	8,575676	7,43	43,24	6,379532	8,78	44,59
30	10,159557	10,81	63,51	8,075859	10,81	63,51

Tabela 5.29: Sinal: Arquivo SX115.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,072086	11,61	27,74	2,808678	12,90	27,10
20	6,692660	7,10	45,16	5,214644	7,10	45,81
30	7,322660	5,81	60,65	6,244142	7,10	59,35

Tabela 5.30: Sinal: Arquivo SX50.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,253745	7,14	24,03	2,644742	5,84	24,03
20	7,658407	9,09	50,65	5,612233	7,14	51,30
30	8,197700	13,64	59,74	6,501066	13,64	59,74

Tabela 5.31: Sinal: Arquivo SX134.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,668352	4,08	26,53	2,978771	4,76	25,85
20	8,087284	2,72	44,90	5,437062	2,04	45,58
30	9,305765	6,80	57,14	6,894791	8,16	57,14

Tabela 5.32: Sinal: Arquivo SX275.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	2,974972	11,11	27,16	2,652430	12,35	26,54
20	6,962189	5,56	48,15	5,337113	6,17	48,15
30	7,901691	11,11	61,11	6,539002	9,88	61,11

Tabela 5.33: Sinal: Arquivo SX284.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%) 2-4dB	>4dB	SD média (dB)	Outliers (%) 2-4dB	>4dB
10	3,474994	8,22	25,34	3,003738	8,22	25,34
20	8,796704	6,85	43,84	6,084337	6,85	43,84
30	9,290604	7,53	62,33	7,187274	6,85	63,70

Tabela 5.34: Sinal: Arquivo SX95.WAV da Base NTIMIT.

A Figura 5-4 mostra os resultados referentes à interpolação de todos os parâmetros, com exceção dos índices do dicionário fixo, com a utilização dos estados atualizados gerados na codificação. O gráfico representa a média dos resultados dos dez sinais de fala codificados nas simulações. Observa-se que o esquema de interpolação alcança uma distorção espectral média 0,55 - 2,07 dB menor.

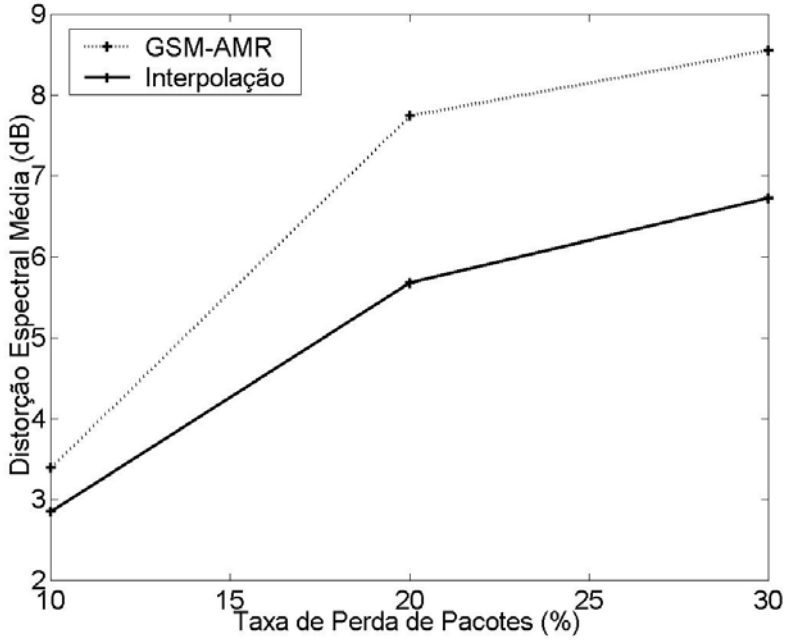


Figura 5-4: Distorção Espectral Média SD2 para Todos os Parâmetros Interpolados (exceto índices do dicionário fixo), com Utilização de Estados.

A Tabela 5.35 apresenta os valores da distorção espectral relacionados ao gráfico da Figura 5-4, assim como a porcentagem de ocorrência de quadros *outliers* para as respectivas taxas de perda.

Os resultados para os dez sinais de fala simulados são apresentados nas Tabelas 5.36 a 5.45.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,393036	7,43	26,54	2,845352	7,81	26,09
20	7,747129	7,54	47,07	5,679373	7,62	47,53
30	8,552753	9,42	61,04	6,728236	8,71	61,10

Tabela 5.35: Outliers da Distorção Espectral para Todos os Parâmetros Interpolados (exceto índices do dicionário fixo), com Utilização de Estados.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	2,880366	5,41	22,97	2,330600	4,73	22,97
20	7,809893	9,46	50,00	5,864106	8,78	50,68
30	7,535169	10,81	56,08	6,114189	10,81	56,08

Tabela 5.36: Sinal: Arquivo SI923.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	2,633641	6,63	24,70	2,328323	7,83	23,49
20	6,594780	8,43	51,20	5,351800	7,83	51,81
30	7,450713	4,82	63,86	6,212799	4,22	64,46

Tabela 5.37: Sinal: Arquivo SX113.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,542271	6,80	27,21	2,761413	5,44	26,53
20	7,314561	5,44	42,86	5,124688	6,12	42,86
30	8,836039	8,16	63,27	7,137861	8,84	62,59

Tabela 5.38: Sinal: Arquivo SI1894.WAV da Base NTIMIT.

Taxa de Perda(%)	GSM-AMR			Interpolação		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	4,170348	8,67	30,00	3,434392	8,00	30,00
20	8,979136	13,33	50,67	6,519492	12,67	50,67
30	9,527634	14,67	62,67	7,097731	14,00	62,67

Tabela 5.39: Sinal: Arquivo SX4.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	4,259584	8,78	29,73	3,549693	4,73	29,05
20	8,575676	7,43	43,24	6,306300	8,78	44,59
30	10,159557	10,81	63,51	7,869767	6,76	63,51

Tabela 5.40: Sinal: Arquivo SX115.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,072086	11,61	27,74	2,780138	12,90	27,10
20	6,692660	7,10	45,16	5,212856	5,81	46,45
30	7,322660	5,81	60,65	6,266201	5,81	59,35

Tabela 5.41: Sinal: Arquivo SX50.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,253745	7,14	24,03	2,628709	9,74	23,38
20	7,658407	9,09	50,65	5,513830	8,44	50,65
30	8,197700	13,64	59,74	6,331890	14,29	59,74

Tabela 5.42: Sinal: Arquivo SX134.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,668352	4,08	26,53	2,987858	4,76	25,85
20	8,087284	2,72	44,90	5,506907	2,04	45,58
30	9,305765	6,80	57,14	6,780736	8,84	56,46

Tabela 5.43: Sinal: Arquivo SX275.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	2,974972	11,11	27,16	2,674268	11,73	27,16
20	6,962189	5,56	48,15	5,318126	6,79	48,15
30	7,901691	11,11	61,11	6,261313	8,02	61,11

Tabela 5.44: Sinal: Arquivo SX284.WAV da Base NTIMIT.

Taxa de Perda(%)	**GSM-AMR**			**Interpolação**		
	SD média (dB)	Outliers (%)		SD média (dB)	Outliers (%)	
		2-4dB	>4dB		2-4dB	>4dB
10	3,474994	8,22	25,34	2,978125	8,22	25,34
20	8,796704	6,85	43,84	6,075627	8,90	43,84
30	9,290604	7,53	62,33	7,209870	5,48	65,07

Tabela 5.45: Sinal: Arquivo SX95.WAV da Base NTIMIT.

Capítulo 6

Conclusão

Apesar do tráfego de voz sobre IP admitir alguma ocorrência de perda de pacotes, existe um certo limite para esta perda de modo a não prejudicar a inteligibilidade do sinal de fala. Existem algumas técnicas para suavizar a perda de pacotes. Este trabalho propôs estudar uma destas técnicas e comparar a qualidade dos sinais codificados com e sem a sua utilização. A técnica utilizada foi a de interpolar parâmetros de quadros adjacentes ao(s) quadro(s) perdido(s). O codificador usado durante as simulações foi o GSM-AMR, transmitindo a uma taxa fixa de 7,95 kbps. Vale a pena salientar que o codificador em questão já apresenta um esquema de substituição e silenciamento para quadros perdidos.

Os resultados mostraram que o esquema de interpolação atinge valores de distorção espectral (SD) inferiores aos valores alcançados pelo esquema do codec GSM-AMR. Quatro situações foram analisadas:

1. SD relacionada aos parâmetros LPC sem a utilização dos estados gerados na codificação.

2. SD relacionada aos parâmetros LPC com a utilização dos estados gerados na codificação.

3. SD relacionada aos sinais decodificados com todos os parâmetros interpolados, com exceção dos índices do dicionário fixo, e sem a utilização dos estados gerados na codificação.

4. SD relacionada aos sinais decodificados com todos os parâmetros interpolados, com exceção dos índices do dicionário fixo, e com a utilização dos estados gerados na codificação.

As simulações mostraram que no primeiro caso o esquema de interpolação atingiu valores de distorção espectral 0,03 - 0,16 dB menores do que a distorção inserida pelo esquema do codificador GSM-AMR. Para o segundo caso, a distorção espectral para o esquema de interpolação atingiu valores 0,09 - 0,31 dB menores que o esquema do codificador GSM-AMR. Também foram alcançados valores inferiores para as situações relacionadas aos sinais decodificados com todos os parâmetros interpolados. Nas situações 3 e 4, foram atingidos valores de distorção espectral 0,52 - 2,01 dB e 0,55 - 2,07 dB menores, respectivamente.

Talvez a distorção espectral possa não ser a melhor métrica a ser utilizada para avaliação do sinal de fala com todos os parâmetros interpolados. Uma sugestão para um trabalho futuro seria a procura por uma métrica que explore de maneira mais adequada a variação da qualidade perceptual provocada pela utilização dos estados gerados na codificação. Acredita-se que esta contribuição seja percebida de uma forma mais eficiente, a partir da utilização de medidas que levem em conta as diversas características peculiares à audição humana, tais como as citadas na seção 5.1.2.

Uma outra sugestão para trabalhos futuros seria a expansão da utilização do esquema de interpolação para as demais taxas de transmissão de bits do codec GSM-AMR. Acredita-se que os bons resultados obtidos neste trabalho permanecerão para quase todas as demais taxas de transmissão, haja vista que não há grandes diferenças entre os algoritmos utilizados pelas outras taxas, excetuando-se a taxa de 12,2 kbps, na qual o codificador GSM-AMR se assemelha ao codificador GSM-EFR.

Um outro bom alvo de pesquisa seria tentar encontrar uma forma eficiente para suprimir os efeitos causados pela perda de pacotes sobre a atualização dos estados dos parâmetros LPC e dos ganhos do dicionário fixo. Em [40], foram utilizados esquemas diferentes de quantização para redução da necessidade de transmissão de informação

lateral, e bons resultados foram conseguidos com um aumento de apenas 0,4 kbps na taxa de transmissão.

Por fim, ainda no campo da interpolação, sugere-se a utilização dos quadros anteriores ao quadro perdido para a estimação dos parâmetros deste [51]. O princípio seria o mesmo deste trabalho, explorando a característica estacionária existente entre quadros de fala de curta duração.

Referências Bibliográficas

[1] Discrete-time processing of speech signals, John R. Deller, John G. Proakis, John H. L. Hansen. Prentice-Hall, c1987.

[2] ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies,"Nov. 1988.

[3] E. Janardhanan, "Differential PCM systems," IEEE Trans. on Commun., vol. COM-27, pp. 82-93 Jan., 1979.

[4] H. R. Schindler, "Delta Modulation," IEEE Spectrum, Vol. 7, pp. 69-78, October 1970.

[5] N. I. Pilipchouk and V. P. Jacovlev, "Adaptive Pulse Code Modulation," - Moscow, Radio i svjaz, 1986 (in Russian).

[6] W. R. Daumer, X. Mitre, P. Mermelstein and I. Tokizawa, "Overview of the ADPCM Coding Algorithm," IEEE Proc. of GLOBECOM, pp.774-777,1984.

[7] J. E. Abate, "Linear and Adaptive Delta Modulation," Proc. IEEE, Vol. 55, pp. 298-308, March 1967.

[8] N. S. Yayant, "Adaptive Delta Modulation With a One-Bit Memory," Bell System Tech. J., pp. 321-342, March 1970.

[9] B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals", Bell Syst. Tech. J., Vol. 49, pp. 1973-1986, October 1970.

[10] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 299-309, Aug. 1977.

[11] Gold, Bernard and Rader, Charles M. *The Channel Vocoder*. IEEE Transactions on Audio and Electroacoustics, Vol. AU-15, No. 4, December 1967.

[12] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Am.*, vol.45, pp.458-465, Feb. 1969.

[13] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol.45, pp. 1493-1509, Nov 1966.

[14] L. R. Rabiner, "Digital-Formant Synthesizer for Speech Synthesis Studies," *J. Acoust. Soc. Am.*, Vol. 43, No. 4, pp. 822-828 (1968).

[15] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in *Proc. 1967 IEEE Conf. Communication Processing*, 1967, pp. 360-361.

[16] Chong Un and Magill, D. "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbits/s". *Communications, IEEE Transactions on [legacy, pre - 1988], Volume: 23, Issue: 12*, Dec 1975. Pages: 1466 - 1474.

[17] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Conf. Rec., 1982 IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1982, pp. 614-617.

[18] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," Proc. IEEE ICASSP'85, 25.1.1, pp.937-940, Apr.1985.

[19] I. A. Gerson, and M. A. Jasuik, "Vector Sum Excited Linear Prediction (VSELP)," *Advances in Speech Coding*, Kluwer Academic Publisher pp.69-79, 1991.

[20] J.-P. Adoul et al, "Fast CELP coding based on algebraic codes," *Proc. Intl. Conference on Acoustics, Speech and Signal Processing*, pp. 1957-1960, 1987.

[21] 3GPP TS 26.093 V4.0.0 (2000-12), AMR Speech Codec; Source Controlled Rate operation.

[22] ITU-T Recommendation G.114, *One-way transmission time*. Mar. 1993.

[23] Velloso, P. B. "Transmissão de voz em redes ad hoc". Tese de Mestrado. COPPE/PEE/UFRJ - Agosto 2003.

[24] Schlzrinne, H., Casner, S., Frederick, R., e Jacobson, V. RTP: A transport protocol for real-time applications. *Internet RFC 1889* (janeiro de 1996).

[25] Hui Dong, Ian D. Chakeres, C.-H. Lin, Allen Gersho, Elizabeth Belding-Royer, U. Madhow and Jerry Gibson. "Speech Coding for Mobile Ad hoc Networks." *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 2003.

[26] Oliver Hersent, David Gurle e Jean-Pierre Petit, IP Telephony - Packet-based multimedia communications systems, Pearson Education Limited, 2000.

[27] ITU-R Recommendation M.1309: "Digitally Coded Speech In The Land Mobile Service", ITU-R, 1997.

[28] Markopoulou, F. Tobagi, M. Karam, "Assessing the quality of Voice Communications over Internet Backbones", *IEEE Transactions on Networking*, Vol. 11 No. 5, October 2003.

[29] C. Hoene, H. Karl, and A. Wolisz, "A Perceptual Quality Model for Adaptive VoIP Applications", In *Proceedings of International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'04)*, San Jose, California, USA, July 2004, Paper won the Best Paper Award of the conference.

[30] Speech Communications: Human and Machine, 2nd Edition. Douglas O'Shaughnessy. November 1999, Wiley-IEEE Press.

[31] ETSI EN 301 703, "Digital cellular telecommunications system (Phase 2+) (GSM); Adaptive Multi-Rate (AMR); Speech processing functions"; General description (GSM 06.71 version 7.0.2 Release 1998).

[32] ETSI EN 301 704, "Digital cellular telecommunications system (Phase 2+) (GSM); Adaptive Multi-Rate (AMR) speech transcoding"; (GSM 06.90 version 7.2.1 Release 1998).

[33] ITU-T Rec. G.714, "Separate performance characteristics for the encoding and decoding sides of PCM channels applicable to 4-wire voice-frequency interfaces," ITU-T, Nov. 1988.

[34] ITU-T Rec. G.712, "Transmission performance characteristics of pulse code modulation channels," ITU-T, Nov. 2001.

[35] ITU-T Rec. G.726, "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)," ITU-T, Dec. 1990.

[36] Kondoz, A. M.; *Digital Speech. Coding for Low Bit Rate Communication Systems.* John Wiley & Sons, Ltd. Chichester. UK. 1999.

[37] ETSI EN 301 705, "Digital cellular telecommunications system (Phase 2+); Substitution and muting of lost frames for Adaptive Multi Rate (AMR) speech traffic channels"; (GSM 06.91 version 7.1.1 Release 1998).

[38] ETSI EN 301 707, "Digital cellular telecommunications system (Phase 2+); Discontinuous transmission (DTX) for Adaptive Multi-Rate (AMR) speech traffic channels"; (GSM 06.93 version 7.5.0 Release 2000).

[39] Kansal A. and Karandikar A., "An Overview of Delay Jitter Control for Packet Audio in IP Telephony". Indian Institute of Technology, Bombay, India.

[40] Jian Wang and Jerry D. Gibson, *"Parameter Interpolation to Enhance the Frame Erasure Robustness of CELP Coders in Packet Networks". IEEE,* 2001.

[41] Jian Wang and Jerry D. Gibson, *"Performance Comparison of Intraframe and Interframe LSF Quantization in Packet Networks". Proc. 2000 IEEE Workshop on Speech Coding,* Delavan, WI, USA, September 2000.

[42] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database", Proc. ICASSP, Albuquerque, Apr. 1990.

[43] Barbosa, L. M. J. "Algoritmos de Busca em Codificadores ACELP,". Tese de Mestrado. Universidade Estadual de Campinas - UNICAMP. Brasil. Novembro 2002.

[44] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," ITU-T, Aug. 1996.

[45] ITU-T Rec. P.800.1, "Mean Opinion Score (MOS) terminology," ITU-T, Mar. 2003.

[46] Barbedo, J.G.A; Lopes, A. "Avaliação Objetiva de Qualidade de Sinais de Áudio e Voz". Tese de Doutorado. Universidade Estadual de Campinas - UNICAMP. Brasil. 2004.

[47] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[48] Beerends, J.G.; Stemerdink, J.A. *A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation*, J. Audio Eng. Soc., Vol. 42, No. 3, pp. 115-123, March 1994.

[49] Tamanna Islam, "Interpolation of Linear Prediction Coefficients for Speech Coding". McGill University, Montreal, Canada, April 2000.

[50] J. H. Y. Loo, "Intraframe and interframe coding of speech spectral parameters," Master's thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada, Sept. 1996.

[51] Perkins, C., Hodson, O., Hardman, V. "A survey of packet-loss recovery techniques for streaming audio". *IEEE Network*. September 1998, 40-48.