PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE ENGENHARIA

RICARDO BECKER

ANÁLISE QUALITATIVA / QUANTITATIVA DE ALGORITMOS PARA A COMPRESSÃO DE VOZ APLICADOS A REDES DE PACOTES

Porto Alegre
2009

RICARDO BECKER

ANÁLISE QUALITATIVA / QUANTITATIVA DE ALGORITMOS PARA A COMPRESSÃO DE VOZ APLICADOS A REDES DE PACOTES

Dissertação apresentada como requisito para obtenção do grau de Mestre pelo Programa de Pós-graduação da Faculdade Engenharia da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Dr. Rubem Dutra Ribeiro Fagundes

Porto Alegre

2009

RICARDO BECKER

ANÁLISE QUALITATIVA / QUANTITATIVA DE ALGORITMOS PARA A COMPRESSÃO DE VOZ APLICADOS A REDES DE PACOTES

Dissertação apresentada como requisito para obtenção do grau de Mestre pelo Programa de Pós-graduação da Faculdade Engenharia da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em _____ de _____ de _____ .

BANCA EXAMINADORA:

Prof. Dr. Rubem Dutra Ribeiro Fagundes

_____

Prof. Dr. Eduardo Augusto Bezerra

_____

Prof. Dr. Dario Francisco Guimarães de Azevedo

_____

Dedico esta obra a todas as pessoas que de alguma forma contribuíram para a minha formação até este momento de minha vida, em especial a minha família e meus amigos.

AGRADECIMENTOS

Ao final deste trabalho agradeço à minha família por todo o suporte, com carinho e afeto dado a mim em todos estes anos a conclusão deste curso.

Agradeço em especial a meu irmão Gabriel pelo companheirismo nestes últimos meses.

A meus amigos e “irmãos de fé” pelo companheirismo e amizade incondicional de tantos anos.

Ao Professor Doutor Engenheiro Rubem Dutra Fagundes, orientador deste trabalho, pela paciência, entendimento e orientação.

Aos colegas de mestrado pelo apoio e companheirismo.

Aos funcionários do programa de Pós-Graduação em Engenharia Elétrica, em especial para as colegas Inelve Colognese e Maria Helena Maciel de Almeida pela paciência e dedicação.

Estar no Presente significa desligar as distrações
e prestar atenção ao que é importante agora.
Você cria o seu próprio Presente com aquilo em
que você presta atenção hoje.

Spencer Johnson

RESUMO

Este trabalho tem por objetivo o estudo, implementação e avaliação de técnicas de compressão de voz, baseadas na detecção de períodos de silêncio, aplicadas a redes de pacotes. Para tanto, foram estudados os conceitos fundamentais de processamento digital de sinais, incluindo aplicações e modelos matemáticos. Posteriormente, estudou-se a estrutura dos sistemas de transmissão de sinais de voz via redes de pacotes, em essência, sistemas de Voz sobre IP (VoIP). Nestes sistemas, foram vistas a aplicabilidade e princípios de funcionamento dos componentes de DSP, desde a própria compressão da voz, baseada nos períodos de silêncio, bem como padrões de codificação, cancelamento de eco, controle automático de ganho e geração de ruído de conforto. Posteriormente é proposta então a implementação de seis técnicas de compressão de voz baseadas na combinação de diferentes algoritmos aplicados na detecção de períodos de silêncio ou não da fala. Dentre os algoritmos aplicados, está a análise no tempo e em frequência do conteúdo de energia do sinal de voz, a análise do sinal na busca dos sons fricativos da fala, e ainda aplicação de recobrimento e compensação por ruído de conforto. Para a implementação das técnicas foram desenvolvidas ferramentas computacionais de testes, e para fins de validação e comparação dos resultados foram utilizadas, com as devidas adaptações, e descritas no trabalho, as recomendações P.800 (MOS) e P.862 (PESQ) do ITU-T, sendo estas entre as mais reconhecidas em termos de avaliação da qualidade do sinal de áudio percebido em sistemas de telecomunicações. Por fim, são apresentados os resultados e as conclusões, onde nos mesmos buscava-se um compromisso das implementações entre percentual estimado de economia de banda proporcionada a redes de pacotes, e nível de degradação do sinal de voz proporcionado pela aplicação da compressão, ao mesmo tempo em que sem comprometimento com alta demanda computacional do sistema. Neste sentido, se verificou que em termos de economia de banda proporcionada e qualidade do áudio, as técnicas LSED, SFD e CVAD, todas implementadas no domínio frequência, apresentaram resultados bastante satisfatórios, assim como a LED e ALED, implementadas no domínio do tempo, que também não ficaram muito atrás em termos de resultados gerais. Também ficou claro o efeito da aplicação do recobrimento e da compensação por ruído de conforto amostrado do próprio microfone do locutor. Por fim, sugere-se um número de possibilidades para a continuidade do trabalho, bem como evolução dos mesmos, tanto em termos de melhorias quanto na diversificação das aplicações dos resultados.

ABSTRACT

This work aims at the study, implementation and evaluation of techniques for voice compression, based on detection of periods of silence, applied to packet networks. For that, were studied the fundamental concepts of digital signal processing, including applications and mathematical models. After that, were studied the transmission systems of voice signals by packet networks, in essence, Voice over IP (VoIP) systems. In this context, this work proposed the implementation of six techniques for compression of voice based on the combination of different algorithms using the detection of periods of silence in speech. Among the algorithms used, were done analysis of voice signal in time and frequency domain considering the analysis of energy content on voice signal, and also, the analysis of fricative sounds in speech, and the application of techniques for coating and for compensation by comfort noise. To implement the techniques, it was developed computational tools for testing, those were also used for evaluation and to compare the results using P.800 (MOS) and P.862 (PESQ) recommendations of ITU-T. Both, MOS and PESQ are accepted as techniques for assessing the quality of the voice signal perceived in telecommunications systems. Finally, the results and conclusions, as we can see results of bandwidth economy provided to networks, and the level of degradation of voice signal provided by the application of techniques, at the same time without compromising the system with high computational demands. We verify that in terms of bandwidth economy and quality of the audio provided by LSED, SFD and CVAD techniques, all of them, implemented in frequency domain with satisfactory results, as well as LED and ALED techniques, implemented in time domain, which also were not far behind in terms of overall results. Finally, it is suggested a number of possibilities for continuing the work and also improvements to applications in different subjects.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE SIGLAS

ACELP - *Algebraic Code Excited Linear Prediction*

ACR - *Absolute Category Rating*

AD - *Auditory Distance*

ADC - *Analog to Digital Converter*

ADPCM - *Adaptative Differential Pulse Code Modulation*

ALED - *Adaptative Linear Energy Based Detector*

BPSK - *Binary Phase-Shift Keying*

CCR - *Comparison Category Rating*

CDMA - *Code Division Multiple Access*

CELP - *Code Excited Linear Prediction*

CMOS - *Comparison Mean Opinion Score*

CNG - *Comfort Noise Generator*

CS-ACELP - *Conjugate Structure Algebraic Excited Linear Prediction*

CVAD - *Compreensive Voice Active Detector*

DAC - *Digital to Analog Converter*

DCR - *Degradation Category Rating*

DCT - *Discrete Cosine Transform*

DFT - *Discrete Fourier Transform*

DMOS - *Degradation Mean Opinion Score*

DPCM - *Differential Pulse Code Modulation*

DSP - *Digital Signal Processing*

DST – *Discrete Sine Transform*

DTA - *Digital Tape Audio*

DWT – *Discrete Wavelet Transform*

FFT - *Fast Fourier Transform*

GSM - *Global System for Mobile Communications*

Hi-fi - *High fidelity*

IP - *Internet Protocol*

ISDN - *Integrated Services Digital Network*

ISP - *Internet Service Provider*

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

JPEG - *Joint Photographic Experts Group*

LD-CELP - *Low Delay Code Excited Linear Prediction*

LED - *Linear Energy Based Detector*

LPC – *Linear Prediction Coefficients*

LPF - *Low Pass Filter*

LSED - *Linear Sub-band Energy Detector*

MLP - *Multilayer Perceptron*

MNB - *Measuring Normalizing Blocks*

MNRU - *Modulated Noise Reference Unit*

MOS - *Mean Opinion Score*

MOSc - *Mean conversation-Opinion Score*

MOSle - *Mean listening-effort Opinion Score*

MS-MLQ - *Multipulse Maximum Likelihood Quantization*

PAMS - *Perceptual Analysis Measurement System*

PC - *Personal Computer*

PCM - *Pulse Code Modulation*

PESQ - *Perceptual Evaluation of Speech Quality*

PSQM - *Perceptual Speech Quality Measure*

PSTN - *Public Switched Telephone Network*

QV - *Vetorial Quantization*

RAM - *Random Access Memory*

RELP - *Residual-Excited Linear Predictive*

RFC - *Request for Comments*

RNA - Redes Neurais Artificiais

ROM - *Read-Only Memory*

RTCP - *Real Time Transport Control Protocol*

RTP - *Real Time Transport Protocol*

RVoIP - *Robust Voice over IP system*

RX - Receptor

S/H - *Sample and Hold*

SFD - *Spectral Flatness Detector*

SNR - *Signal-to-Noise Ratio*

SQNR - *Signal-to-Quantization Noise Ratio*

STD - *Static Threshold Detector*

TCP - *Transmission Control Protocol*

TDM - *Time Division Multiplexing*

TV - *Television*

TX - Transmissor

UDP - *User Datagram Protocol*

VAD - *Voice Activity Detection*

VoIP - *Voice over IP*

WFD - *Weak Fricatives Detector*

WTSC - *World Telecommunication Standardization Conference*

XR - *Extended Report*

SUMÁRIO

1 INTRODUÇÃO

As aplicações com comunicação de voz sobre redes IP (IP - *Internet Protocol*) e os sistemas de Telefonia IP[1] já vem se consolidando tanto sobre os *backbones*[2] das grandes redes corporativas quanto sobre a Internet de um modo geral. Essa tecnologia se apresenta como uma forma de contornar os atuais sistemas de tarifação das operadoras tradicionais de serviços de telefonia, impactando de forma econômica, tanto na indústria técnica já instalada, bem como fazendo com que as operadoras de telecomunicações acelerem o processo para oferecimento de serviços de telefonia sobre IP, de modo a se manterem competitivas no mercado (VENDRUSCULO, 2005).

Neste cenário, a necessidade de economia de banda em redes de transmissão de voz se faz cada vez mais necessária, considerando o grande aumento da demanda por troca de informações entre as pessoas e os custos associados à implantação dos sistemas de transmissão de voz.

Quando se faz referência especificamente a sistemas de voz sobre IP (VoIP – *Voice over Internet Protocol*), a necessidade é a mesma. Foi constatado que durante uma conversação média entre duas pessoas, cada uma fala apenas de 35% a 50% do tempo total do canal ativo (HERSENT, 2002) (KONDOZ, 2000) (MONTEIRO, 2002) (DAVIS, 2002), sendo o canal passivo o do ouvinte que apenas escuta. Isto leva a pensar no desperdício de banda que ocorre na rede proporcionado pelo constante tráfego de pacotes que contém apenas silêncio. Visto isso, encontra-se uma oportunidade para o desenvolvimento e aplicação de técnicas de detecção e supressão de silêncio em chamadas telefônicas que utilizem a infra-estrutura da Internet, no contexto de comutação por pacotes.

Faz-se necessária a economia de banda em sistemas de voz sobre IP, especialmente quando consideradas as aplicações em tempo real e em redes de capacidade de transmissão

[1] Voz sobre IP não é telefonia IP, mas sim um conjunto de técnicas e questões mais amplo, dentro do quais a telefonia IP se situa. VoIP, nessa ótica, refere-se a toda utilização de voz aplicada em redes IP (em sistemas de audioconferência, videoconferência, telefonia IP, entre outros) (BALBINOT, 2004).

[2] No contexto de redes de computadores, o *backbone* (traduzindo para português, espinha dorsal) designa o esquema de ligações centrais de um sistema mais amplo. Por exemplo, os operadores de telecomunicações mantêm sistemas internos de elevado desempenho para comutar os diferentes tipos e fluxos de dados (voz, imagem, texto). Na Internet podem-se encontrar, hierarquicamente divididos, vários *backbones*: os de ligação intercontinental, que derivam nos *backbones* internacionais, que por sua vez derivam nos *backbones* nacionais, e assim sucessivamente (ROSE, 2007).

limitada, principalmente, devido à sobrecarga que a transmissão de voz sobre datagramas[3] causa (*overhead* do cabeçalho IP). O foco nessa situação é fazer com que o detector de silêncio identifique a inatividade de voz, e evite a transmissão de um datagrama que não possui informação significativa (BECKER, 2005).

1.1 OBJETIVO

Sendo apresentado o contexto ao qual o trabalho foca, o principal objetivo desta dissertação é implementar e comparar a eficiência da combinação de algoritmos para a compressão de voz, baseados no princípio da supressão dos períodos de silêncio ocorridos ao longo da fala. O comparativo de resultados focaliza questões referentes a validação da aplicabilidade das técnicas junto à tecnologias que tenham por princípio a transmissão de voz por comutação de pacotes, neste caso VoIP, considerando para isto questões como degradação da voz ocasionada, economia de banda proporcionada à rede IP, e o não comprometimento da técnica quanto a demanda computacional do algoritmo utilizado.

1.2 OBJETIVOS ESPECÍFICOS

Para validação dos algoritmos são utilizadas as recomendações P.800 (MOS – *Mean Opinion Square*) (ITU, 1996a) e P.862 (PESQ - *Perceptual Evaluation of Speech Quality*) (ITU, 2001) definidas pelo ITU-T (ITU-T - *Telecommunication Standardization Sector of International Telecommunication Union*) objetivando a avaliação do nível de degradação da qualidade do áudio quando da aplicação dos algoritmos de compressão de áudio, que serão apresentadas neste trabalho. A utilização de duas diferentes recomendações do ITU-T, vem proporcionar que seja realizado um comparativo de resultados, isso porque a recomendação P.800 (ITU, 1996a) visa uma avaliação subjetiva dos resultados, baseada na opinião de um grupo de ouvintes. Já a recomendação P.862 (ITU, 2001) visa uma avaliação objetiva da

[3] Datagrama é a unidade básica de dados no nível IP. O datagrama está dividido em duas unidades básicas, um é o campo de cabeçalho e o outro o campo de dados (KUROSE, 2003).

qualidade do áudio baseado na comparação entre o sinal original, e um sinal degradado, via um modelo matemático de percepção.

O sinal aqui dito original é o áudio do locutor na sua origem, sem qualquer degradação ocasionada pelo sistema a ser avaliado. Já o sinal degradado, é o sinal supostamente de pior qualidade por já ter sido submetido a alguma intempérie do sistema de transmissão, ou no caso deste trabalho, da compressão.

Também é apresentado um detalhamento das características de cada algoritmo de detecção de atividade de voz (VAD – *Voice Activity Detection*) implementado bem como cada parâmetro manipulado durante todo o processo de construção e testes dos mesmos, além dos resultados obtidos e a comparação entre os mesmos.

Dentre os parâmetros e aspectos a serem considerados, no que diz respeito ao desenvolvimento das técnicas de detecção e supressão de silêncio, e que se encontram detalhados ao longo do trabalho, citam-se, por exemplo: a base de cálculo, energia do sinal para a determinação da atividade ou não de voz em comunicações via redes IP, características físicas da voz, domínio de trabalho (tempo ou freqüência) e regra de decisão para a determinação do limiar de decisão para a indicação pelo algoritmo se o sinal em questão é voz ou apenas ruído. Quanto aos aspectos relevantes à validação das implementações, são necessários levar-se em conta itens como a aplicabilidade dos algoritmos a sistemas de tempo real (velocidade computacional), qualidade subjetiva e objetiva da voz após a mesma ser submetida à atuação do algoritmo e economia de banda proporcionada.

Como aspectos complementares à detecção e supressão de silêncio, são também apresentados nesta dissertação referências e ou tópicos que dizem respeito a outros itens que compõem um sistema de VoIP e que também estão relacionadas com a área de processamento digital de sinais (DSP – *Digital Signal Processing*). Dentre os tópicos abordados estão alguns conceitos básicos de DSP em termos de aplicabilidade, bem como tecnologias envolvidas. Posteriormente, é feita uma abordagem sobre pontos básicos da tecnologia de voz sobre IP, focando em especial as tecnologias que envolvem DSP dentro deste contexto, sendo elas canceladores de eco, codificadores de voz, geração de ruído de conforto, controle automático de ganho e mais detalhadamente, algoritmos para avaliação do sinal de voz. Por fim, são abordadas algumas formas de avaliação de qualidade de áudio aplicadas a sistemas de telecomunicações, incluindo seus princípios de funcionamento e parâmetros de referência.

1.3 ESTRUTURA DO TRABALHO

Quanto à apresentação desta dissertação, em específico a sua construção documental, o segundo capítulo relata os conceitos mínimos necessários para a fundamentação da proposta de trabalho e por consequência auxílio ao desenvolvimento das implementações e testes realizados. Nele é apresentado o resultado da revisão bibliográfica realizada, a qual contém desde a conceituação sobre processamento digital de sinais como um todo e também voltado para sistemas de VoIP, até detalhes específicos referentes à detecção e supressão de silêncio em sinais de voz. Por fim, são apresentados métodos, para a avaliação do nível de degradação da qualidade do áudio, aplicados em sistemas de telecomunicações, sendo a maioria deles baseados nas principais recomendações do ITU-T para tanto.

No terceiro e quarto capítulos são detalhadas as combinações de algoritmos e a proposta de desenvolvimento de um mecanismo para a verificação de conteúdo significativo de mídia, no caso, voz, para a decisão de gerar ou não tráfego de dados para as redes de pacotes. Para tanto, é proposta a implementação de técnicas para a compressão da voz baseadas na supressão de períodos de silêncio, bem como a verificação da necessidade da utilização de metodologias de avaliação da qualidade desse áudio, após a aplicação das mesmas. Esse detalhamento é baseado no levantamento bibliográfico apresentado no capítulo dois, vislumbrando resultados aplicáveis, embasados em justificativas próprias via a geração de um conteúdo estatístico resultado de testes que neste ponto do trabalho são propostos.

A apresentação da metodologia aplicada na construção desta dissertação de mestrado é feita no quinto capítulo. Nessa apresentação faz-se o processo de desenvolvimento iniciando com a revisão bibliográfica, passando pela definição da proposta de trabalho, implementação – modelagem e programação – das técnicas de detecção e supressão, definição da ferramenta de programação a ser utilizada, o detalhamento das ferramentas construídas para auxílio dos testes das implementações e a discussão dos mesmos.

No sexto capítulo é feita a apresentação, comparação e avaliação dos resultados obtidos com os testes realizados e a especificação das alterações que se observaram necessárias ao longo das implementações.

No sétimo capítulo são apresentadas as conclusões tiradas pelo autor com relação aos resultados obtidos baseado nas experimentações feitas e nas bibliografias consultadas. Também são abordados os possíveis trabalhos futuros a serem realizados dentro do tema proposto de forma que se possa dar continuidade ao mesmo.

Por fim, são listadas todas as bibliografias referenciadas ao longo do texto, apresentados os apêndices e anexos gerados.

2 FUNDAMENTAÇÃO TEÓRICA

Detectores de atividade de voz (VAD) se enquadram em um campo de aplicação cada vez mais abrangente. Sejam nas redes *wireless* com tecnologia GSM (GSM - *Global System for Mobile Communications*) ou CDMA (CDMA - *Code Division Multiple Access*), ambas com aplicação direta nas redes celulares, seja para codificação de voz ou sistemas de reconhecimento, equipamento *hands-free* para os telefones, áudio conferência, cancelamento de eco, bem como em sistemas de voz sobre redes de pacote. Em grande parte das aplicações destinadas a sistemas de voz onde a mesma é processada, ocorre a presença e a possibilidade de identificação dos períodos ditos *voiced* e *unvoiced* (vozeados e não vozeados) (TANYER, 1998).

Dada a conhecida evolução das redes de telecomunicações em um sentido de convergência das mesmas, além do exponencial crescimento no tráfego de dados nessas redes, soluções que venham a contribuir com a confiabilidade das mesmas são cada vez mais bem vindas. Do ponto de vista de aplicações de rede (KUROSE, 2003), essas contribuições seriam no sentido de garantia de largura de banda, na garantia da integridade da informação enviada, e na garantia de disponibilidade da informação, em termos de temporização, tudo isso conforme a demanda. Para tanto, novas tecnologias, detentoras de novos algoritmos e novos métodos, são de certa forma "empilhadas" na estrutura das redes, sempre com o objetivo de atender as demandas de novas e antigas aplicações.

Dentro do contexto da Internet, em específico, isso fica muito claro. Por exemplo, quando se passa a ter telefonia via redes de datagramas, redes essas originalmente estruturadas puramente para dados. Neste sentido, uma série de tecnologias emergem como forma de agregar qualidade a esse tipo de sistema. Mas como consequência, há um significativo aumento de tráfego nas redes, seja da própria mídia, sejam dos protocolos de mídia e sinalização da aplicação geradora desse tipo de dado. Como solução paliativa e até certo ponto barata, quando do funcionamento normal da rede, é o aumento da largura de banda dos enlaces como forma de garantir os serviços. Mas, diante de qualquer anormalidade no funcionamento da rede, a economia de qualquer recurso nas mesmas é bem vinda, especialmente quando proporciona a não geração de tráfego. Por esse caminho, as técnicas de VAD se inserem como forma de propiciar uma agilidade maior nas ferramentas de VoIP, isso porque os pacotes de áudio considerados silêncio são codificados, mas não comprimidos por codificadores específicos, já que os mesmos são descartados, gerando uma economia de

tempo de processamento junto aos codificadores e consequentemente, o objetivo maior proposto por esse trabalho, menos tráfego para a rede.

Neste cenário, esse capítulo dois se propõe a abordar, os tópicos necessários para o desenvolvimento de técnicas de VAD, incluindo como seguem, conceitos de DSP, bem como definição de ruído que se apresenta ao sistema em questão. Ainda aqui, é tratada da aplicabilidade de conceitos de DSP em sistemas de VoIP, bem como algoritmos utilizados para a distinção entre períodos de silêncio e períodos não só de fala, mas que possuem conteúdo significativo para a conversão quando se faz uso do tráfego de voz sobre IP. Por fim, são listados métodos aplicados em sistemas de telecomunicação com o objetivo de mensuração do nível de degradação proporcionado pelo sistema de telecomunicação quando da transmissão de voz pelo mesmo.

2.1 PROCESSAMENTO DIGITAL DE SINAIS

A área de Processamento Digital de Sinais se distingue de outras áreas da computação e da ciência por tratar de um único item: sinais. Na maioria dos casos estes sinais têm origem sensorial provenientes do mundo real: vibrações sísmicas, imagens visuais, ondas sonoras, etc. DSP é a matemática, os algoritmos, e as técnicas para a manipulação destes sinais depois deles serem convertidos para a forma digital (SMITH, 1997).

Processamento de sinais foca no modelamento, detecção, identificação e utilização de padrões e estruturas aplicados aos sinais processados. Aplicações de métodos em processamento de sinais incluem áudio de alta fidelidade, TV e rádio digital, telefonia celular, reconhecimento de voz, visão, radar, sonar, exploração geofísica, eletrônica médica, e ainda qualquer sistema que tenha relação com comunicações ou processamento de informação (VASEGUI, 2000).

Durante as últimas décadas, com a disponibilidade de computadores digitais compactos, com poder computacional cada vez maior e relativamente baratos, aumentaram significativamente as aplicações de processamento digital de sinais. Esta tendência tem sido reforçada pelo desenvolvimento simultâneo de procedimentos numéricos eficientes (algoritmos) para o processamento de sinais digitais. DSP tornou-se uma aplicação primária para a tecnologia de circuitos integrados com *chips* programáveis de alta velocidade capazes

de realizar as operações requeridas. Por isso é natural encontrar DSP aplicado em diversas áreas (STRUM, 1988).

Processamento digital de sinais diz respeito à representação dos sinais por seqüências de números e ou símbolos e a representação dos mesmos. A proposta deste processamento deve ser estimar parâmetros para obter características do sinal e ou transformar o mesmo a fim de deixá-lo compatível com o meio desejável (OPPENHEIM, 1975).

A teoria de processamento de sinais é o ponto central no desenvolvimento de sistemas de comunicação digital e automação de sistemas, na eficiência dos sistemas de transmissão de sinais, bem como na recepção e decodificação da informação.

Já a teoria estatística aplicada a processamento de sinais prove uma fundamentação para o modelamento da distribuição de sinais aleatórios e o ambiente no qual esses sinais se propagam. Modelos estatísticos são aplicados em processamento de sinais, em específico em sistemas decisórios, para a extração de informação do sinal, que pode ser ruidoso, distorcido ou ainda incompleto.

2.1.1 Sinais e informação

Um sinal pode ser definido como uma variação de quantidade no qual informação é transmitida considerando o estado, as características, a composição, a trajetória, e o curso da ação ou intenção da fonte do sinal. O sinal é o meio para se transmitir informação. A informação transmitida em um sinal pode ser usada por humanos ou máquinas para comunicação, prognósticos, tomada de decisão, controle, exploração. A Figura 1 ilustra uma fonte de informação seguida por um sistema para sinalização desta informação, o canal de comunicação onde ocorre a propagação do sinal a ser transmitido até o receptor, onde está a unidade de processamento do sinal necessária para tratamento e extração da informação do sinal (VASEGHI, 2000).

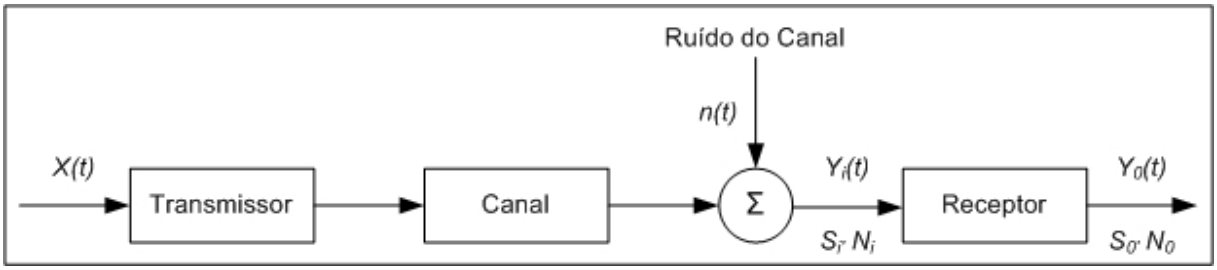


Figura 1 - Generalização em blocos de sistemas de telecomunicações
Fonte: Hsu (2006) e Vasegui (2000).

Métodos aplicados ao processamento de sinais são elementos centrais para a eficiência dos sistemas de comunicação, e para o desenvolvimento de sistemas inteligentes com interface entre homem e máquina, assim como no reconhecimento de padrões de fala e visual para sistemas multimídia. Em geral, processamento digital de sinais é focado em duas áreas da teoria da informação (VASEGHI, 2000):

- Eficiência e confiabilidade na codificação, transmissão, recepção, armazenamento e representação dos sinais em sistemas de comunicação;
- Extração da informação de sinais ruidosos para reconhecimento, detecção, prognósticos, auxílio à decisão, melhoria do sinal, controle e automação.

2.1.2 Métodos para processamento de sinais

Métodos de processamento de sinais envolvem algoritmos de alta complexidade objetivando melhorar a utilização da informação com o melhor desempenho possível. Em geral, a demanda computacional dos métodos aplicados a processamento de sinais aumenta exponencialmente, com a complexidade do algoritmo aplicado. No entanto, o custo de implementação de métodos avançados de processamento de sinais tem compensado e tido um preço acessível, considerando a tendência dos últimos anos do contínuo aumento de desempenho, casado ao simultâneo declínio do custo dos processadores de sinais implementados em *hardware* (LYONS, 2004).

Segundo Vasegui (2000), dependendo do método utilizado, os algoritmos aplicados a processamento de sinais podem ser divididos em quatro categorias: processamento de sinais não-paramétricos, processamento de sinais baseados em modelos, processamento de sinais via estatística e redes neurais.

2.1.2.1 Processamento de sinais não-paramétricos

Métodos não-paramétricos, como o próprio nome diz, não utiliza um modelo paramétrico de geração de sinal ou um modelo de distribuição estatística de sinal. O sinal é processado como uma forma de onda, ou uma sequência de dígitos. Métodos não-

paramétricos não são especializados em nenhuma classe específica de sinais. Eles são aplicados de um modo geral a qualquer sinal, sem levar em consideração a fonte do mesmo. A desvantagem deste método é que ele não faz distinção das características do sinal a ser processado, o que poderia vir a ser o primeiro fator a ser considerado para uma substancial melhoria de desempenho. Alguns exemplos de métodos não-paramétricos incluem filtros digitais e técnicas de transformação de base como Fourier, por exemplo, onde é possível aplicar técnicas para a estimação espectral de potência do sinal, interpolação e restauração deste sinal.

2.1.2.2 Processamento de sinais baseados em modelos

Métodos de processamento de sinais baseados em modelos utilizam um modelo paramétrico no processo de geração do sinal. O modelo paramétrico normalmente descreve a previsível estrutura e o padrão esperado para o sinal processado, e que possa ser usado para prever futuros valores do sinal baseado numa trajetória passada. No entanto, esses sinais podem ser sensíveis a desvios do sinal da classe de sinais característicos do modelo. Modelos paramétricos são largamente utilizados em preditores lineares. A predição linear tem facilitado o desenvolvimento de técnicas avançadas para o processamento de sinais, tais como codificação de sinais de voz aplicados a telefonia celular, codificação de vídeo, análise espectral de alta resolução, processamento de sinais de radares e sistemas de reconhecimento de fala.

2.1.2.3 Processamento de sinais via estatística

As flutuações de um sinal puramente aleatório, ou a distribuição da classe de sinais aleatórios no espaço, não podem ser modeladas por uma equação preditora, mas pode ser descrita em termos de valores estatísticos médios, e modelado pela função de distribuição de probabilidade do sinal multidimensional. Por exemplo, um preditor linear direcionado a sinais aleatórios pode modelar a formatação de uma palavra falada. No entanto, a entrada de sinais aleatórios do preditor linear, ou as variações nas características de diferentes formatos

acústicos da mesma palavra dentre toda a população, não podem ser apenas descritas em termos estatísticos e funções de probabilidade. A teoria para dedução Bayesiana[4], por exemplo, fornece uma base para generalização do processamento estatístico de sinais aleatórios, e para formular e resolver problemas de estimação e tomada de decisão.

2.1.2.4 Redes neurais

Redes neurais artificiais (RNA) são combinações de unidades de processamento de sinais não lineares, arranjados em estruturas construídas para a transmissão e processamento de sinais sob modelos de estruturas baseados em neurônios biológicos. Em redes neurais, diferentes camadas de elementos em paralelo são interconectadas com uma estrutura hierárquica de conexões da rede. Os pesos das conexões são treinados para executar a função de processamento do sinal como um preditor ou classificador (CASTRO, 2001).

A Figura 2 apresenta uma rede neural MLP (MLP - *Multilayer Perceptron*) com duas camadas escondidas e uma camada de saída.

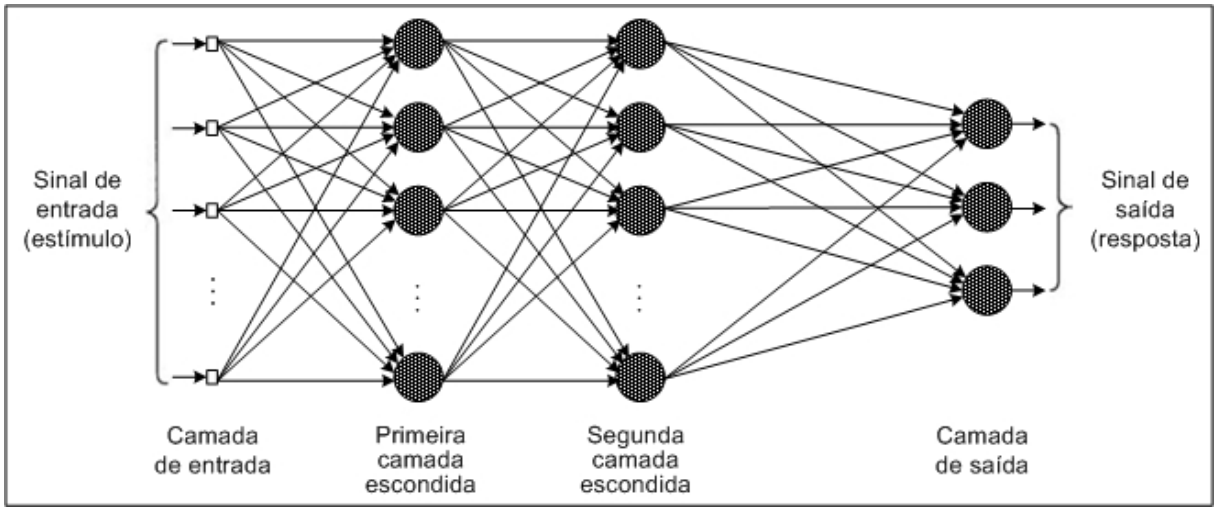


Figura 2 - Rede neural MLP com duas camadas escondidas e uma camada de saída
Fonte: Castro (2001).

Redes neurais são particularmente usadas no particionamento do sinal espacial, na extração de características e padrões de reconhecimento, e em sistemas de tomada de decisão.

[4] A inferência bayesiana é um tipo de inferência estatística que descreve as incertezas sobre quantidades invisíveis de forma probabilística. Incertezas são modificadas periodicamente após observações de novos dados ou resultados. A operação que calibra a medida das incertezas é conhecida como operação bayesiana e é baseada na fórmula de Bayes (EHLERS, 2003).

Alguns sistemas híbridos de reconhecimento de padrões baseados em redes neurais são usados para complementar métodos de inferência Bayesiana.

2.1.3 Aplicação de processamento digital de sinais

Nos últimos anos, o desenvolvimento e a disponibilidade comercial do incremento de potência e a barateamento dos computadores tem sido acompanhados pelo desenvolvimento de avançados algoritmos aplicados a processamento de sinais, tendo esses algoritmos uma larga variedade de aplicações como: redução de ruído, telecomunicações, radar, sonar, processamento de áudio e vídeo, reconhecimento de padrões, exploração geofísica, previsão de dados, e uma série de outras possibilidades. Dentre estas, alguns são tratadas de forma mais específica na sequência.

2.1.3.1 Cancelamento adaptativo de ruído e redução de ruído

Em comunicações de voz em um ambiente acusticamente ruidoso, como um carro ou um trem em movimento, ou em uma chamada telefônica ruidosa, o sinal de voz é constituído da adição de um ruído aleatório, em geral, proveniente de diferentes fontes. Quando medido esse sinal, o sinal que contém a informação está contaminado, ou degradado, pelo ruído deste ambiente. Esse ruído pode ser modelado como na equação 1:

$$y(m) = x(m) + n(m) \qquad ...(1)$$

onde $x(m)$ e $n(m)$ são o sinal de interesse e o ruído respectivamente, e m é o índice do tempo discretizado. Em alguns casos, como por exemplo, o de um telefone celular em um carro em movimento, ou um rádio de comunicação na cabine de um avião, é possível medir e estimar a amplitude instantânea do ruído ambiente usando um microfone direcional. O sinal $x(m)$ pode ser reconstruído pela estimação do ruído retirado do sinal ruidoso.

A Figura 3 apresenta um cancelador adaptativo de ruído para sinais de voz ruidosos composto de dois microfones para captação do áudio. Neste sistema o microfone direcional pega a entrada do sinal ruidoso $x(m)+n(m)$, e a o segundo microfone direcional, posicionado a uma certa distância do primeiro, mede o ruído $\alpha n(m+\tau)$. O fator de atenuação α e o atraso τ

proporcionam um modelo simplificado dos efeitos da propagação do ruído para diferentes posições no espaço onde os microfones estão colocados. O ruído do segundo microfone é processado por um filtro digital adaptativo para que o ruído seja proporcional ao ruído contido no sinal do microfone que contém voz para que esse ruído então, seja subtraído e tenha-se na saída do sistema apenas a voz do locutor. O cancelador adaptativo de sinal é mais efetivo no cancelamento de ruído em baixas frequências, mas geralmente sofre com características não estacionárias do sinal, e uma simplificação da suposição de que o filtro linear pode modelar a dispersão e propagação do ruído no espaço.

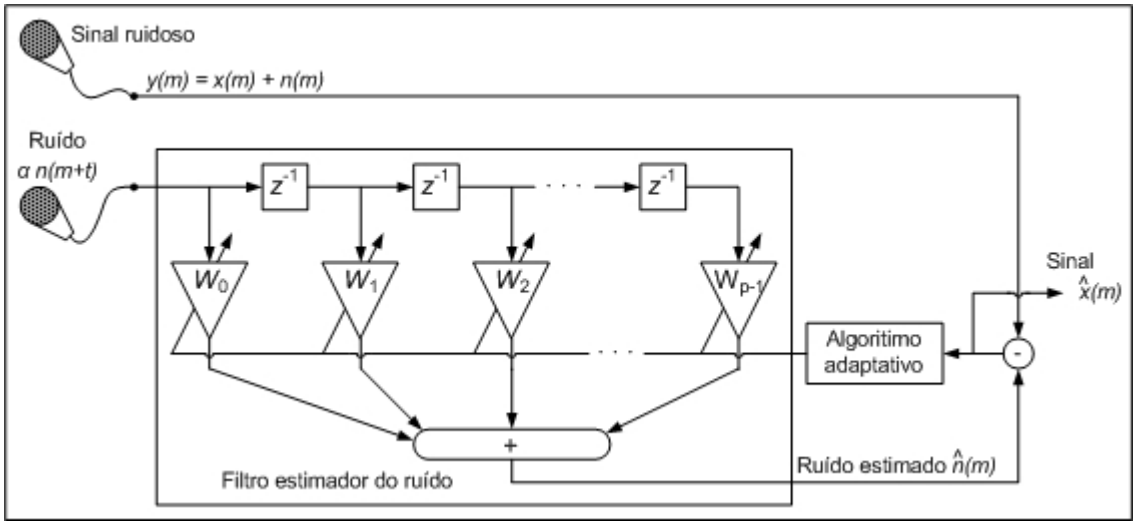


Figura 3 - Configuração de dois microfones junto a um cancelador adaptativo de ruído
Fonte: Vasegui (2000).

Em muitas aplicações, por exemplo, em receptores de sistemas de telecomunicações, não existe o acesso aos valores instantâneos de contaminação do sinal por ruído, tendo-se apenas o sinal ruidoso. Nesses casos o ruído não pode ser cancelado, mas apenas reduzido, baseado em uma média, usando características estatísticas do sinal e do ruído. A Figura 4 apresenta um banco de filtros Wierner[5] para a redução de ruído aditivo quando apenas o sinal ruidoso está disponível. Cada coeficiente do banco de filtros atenua o sinal ruidoso na proporção inversa a relação sinal-ruído - SNR[6] (SNR – *Signal-to-Noise Ratio*) - em cada frequência. Os coeficientes do banco de filtros Wierner são calculados a partir da estimativa da potência espectral do sinal ruído e do ruído processado.

[5] Na década de 1940, Norbert Wiener fio pioneiro na pesquisa para a elaboração de um filtro que produziria a estimativa de um sinal ruidoso, daí o nome deste tipo de filtro (NAKASHIMA, 2003).
[6] SNR é a relação sinal-ruído, que é uma forma quantitativa de se contabilizar o efeito do ruído. Define-se SNR como a razão entre a potência média do sinal e a potência média do ruído, sendo ambas medidas no mesmo ponto do sistema. SNR é expressa em decibéis (dBs), definidos como 10 vezes o logarítmo (na base 10) da relação de potência (SHENOI, 1995).

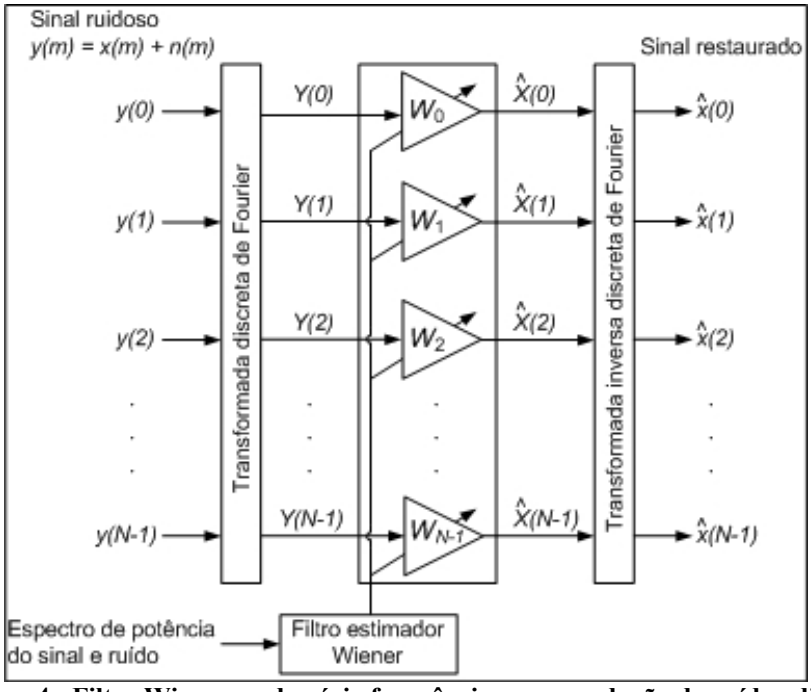


Figura 4 - Filtro Wiener no domínio frequência para a redução do ruído aditivo
Fonte: Vasegui (2000).

2.1.3.2 Equalizador de canal

O equalizador de canal recupera o sinal distorcido na transmissão ao longo do canal de comunicação, com atenuação de magnitude e não linearidade na resposta em fase. Quando a resposta do canal é desconhecida, o processo de recuperação do sinal é dito equalização as cegas. Esse tipo de equalização do sinal possui uma larga faixa de aplicação, especialmente em sistemas de comunicação digital para remoção de interferência intersimbólica devido à propagação em multipercurso, em sistemas de reconhecimento de voz para remoção do efeito de microfonia, na correção de imagens distorcidas, análise de dados sísmicos, entre outras possibilidades.

Na prática, a equalização as cegas do canal é factível, segundo Vasegui (2000), apenas se algumas estatísticas da entrada do canal estiverem disponíveis. O sucesso do uso do método de equalização as cegas depende do quanto se sabe a respeito das características do sinal de entrada e o quanto se pode usar esse conhecimento para o processo de equalização do canal. A Figura 5 apresenta a configuração de um equalizador as cegas de decisão direta.

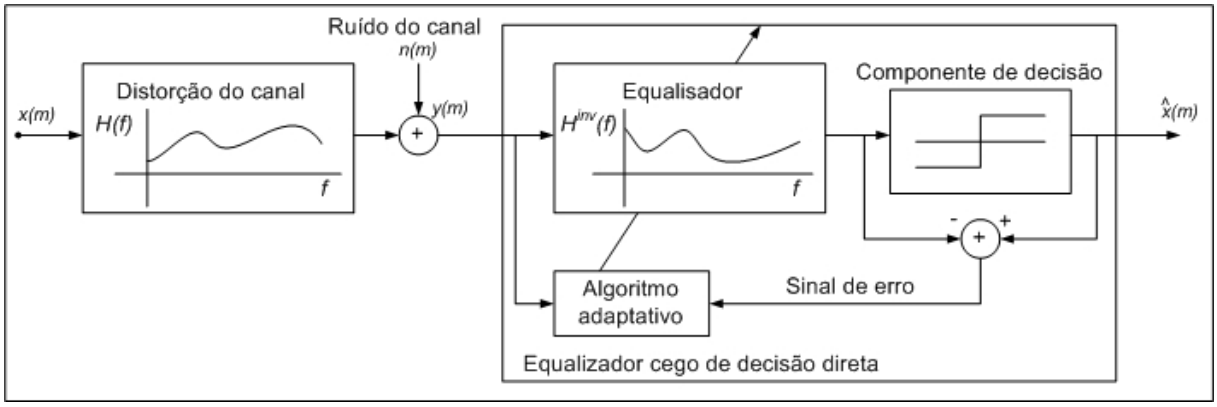


Figura 5 - Configuração de um equalizador de canal "as cegas"
Fonte: Vasegui (2000).

Esse tipo de equalizador de canal (Figura 5) é composto por duas partes: um equalizador adaptativo que remove grande parte da distorção do canal, seguido de um componente de decisão não linear, para uma melhora da estimativa do sinal de entrada do canal. A saída do componente de decisão é a estimativa final da entrada do canal, e é usada como o sinal desejado para direcionar o processo de adaptação do equalizador.

2.1.3.3 Classificação do sinal e padrão de reconhecimento

A classificação do sinal é usada na detecção, reconhecimento de padrões e tomada de decisão em sistemas. Por exemplo, um detector pode classificar de forma binária, quanto a presença ou não de uma forma de onda conhecida quando esta mesma onda está degradada por um ruído. Em classificação de sinais, objetiva-se um sistema que com um mínimo de erro, possa rotular um sinal como este sendo um provável sinal de uma determinada classe de sinais.

Para programar um classificador, classes de sinais que são de interesse para a aplicação são treinadas com base em um conjunto de sinais modelo. A forma mais simples que os modelos podem assumir está em um banco, ou em um livro de códigos (*codebook*), de formas de onda, cada uma representando um tipo de classe de sinais. Na fase de classificação, o sinal é "etiquetado" com a classe mais próxima ou mais parecida. Por exemplo, usando a comunicação binária, onde diante de um *stream* de bits que transitam pela banda passante de um canal, o padrão de codificação BPSK (BPSK – *Binary Phase-Shift Keying*) classifica como bit "1" quando da identificação da forma de onda $+A_c sin\omega_c t$ e o bit "0" quando da identificação da forma de onda $-A_c sin\omega_c t$. No receptor, apresentado na Figura 6, o

decodificador tem a tarefa de classificar e etiquetar o sinal de ruído recebido como “1” ou “0”. O receptor tem dois algoritmos de correlação, cada um programado com a representação de um dos dois símbolos binários. O decodificador correlaciona o sinal de entrada não etiquetada com cada um dos dois símbolos candidatos, e seleciona qual dos dois candidatos tem mais correlação com a entrada.

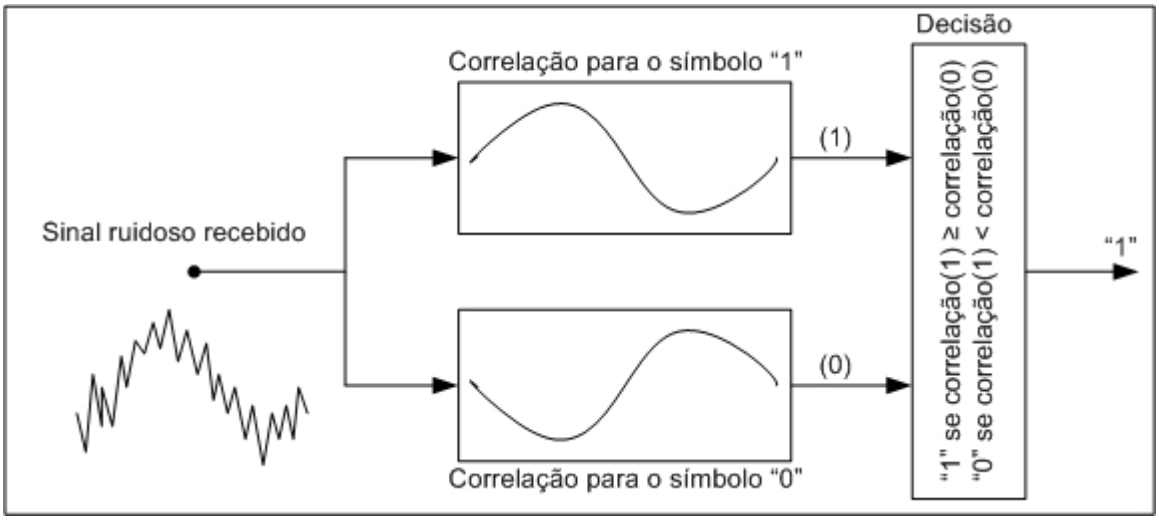


Figura 6 - Diagrama de blocos de um demodulador BPSK
Fonte: Vasegui (2000).

2.1.3.4 Modelagem de um preditor linear para voz

Modelos de predição linear são largamente utilizados em aplicações que envolvem o processamento de sinais de voz, como nas codificações de voz aplicadas a telefonia celular e nos sistemas de reconhecimento de fala. A voz é gerada pela constrição do ar junto aos pulmões, e exalando o mesmo gerando vibrações das cordas vocais e do trato vocal. A partir do efeito da vibração das cordas vocais e do trato vocal é possível introduzir uma medida de correlação e previsibilidade sobre as variações aleatórias do ar expelido dos pulmões. A Figura 7 apresenta um modelo para a reprodução da voz.

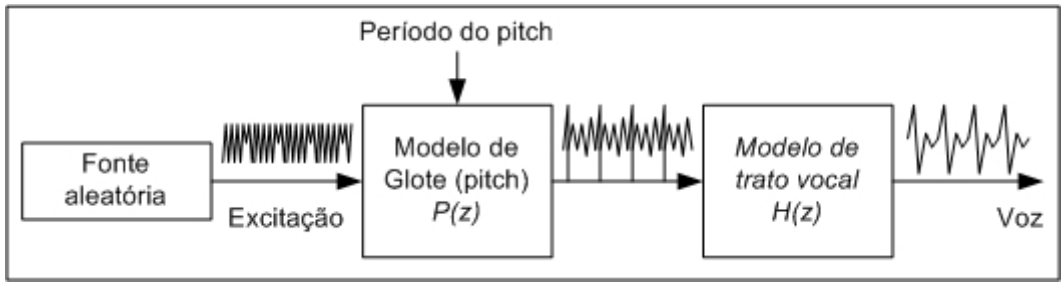


Figura 7 - Modelo de predição linear da voz
Fonte: Vasegui (2000)

A principal fonte da correlação na voz é o modelamento do trato vocal pelo preditor linear. O preditor linear faz uma previsão da amplitude do sinal no tempo *m*, usando uma combinação linear de *P* anteriores amostras *[x(m-1),...,x(m-P)]* como mostra a equação 2:

$$\bar{x}(m) = \sum_{k=1}^{P} a_k x(m-k) \qquad ...(2)$$

onde $\bar{x}(m)$ é a predição do sinal *x(m)*, e o vetor $a=[a_1,...,a_P]$ são os coeficientes do vetor do preditor de ordem *P*, também chamados de LPC (LPC – *Linear Prediction Coefficients*) (HERSENT, 2005). O erro do preditor *e(m)* é a diferença entre a amostra atual *x(m)* e o valor predito $\bar{x}(m)$, como é definido na equação 3:

$$e(m) = x(m) - \sum_{k=1}^{P} a_k x(m-k) \qquad ...(3)$$

A equação 4 descreve a voz sintetizada pelo modelo do preditor utilizado, sendo esta um arranjo da equação 3.

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + e(m) \qquad ...(4)$$

A Figura 8 ilustra um modelo do preditor usado para sintetizar o sinal de voz.

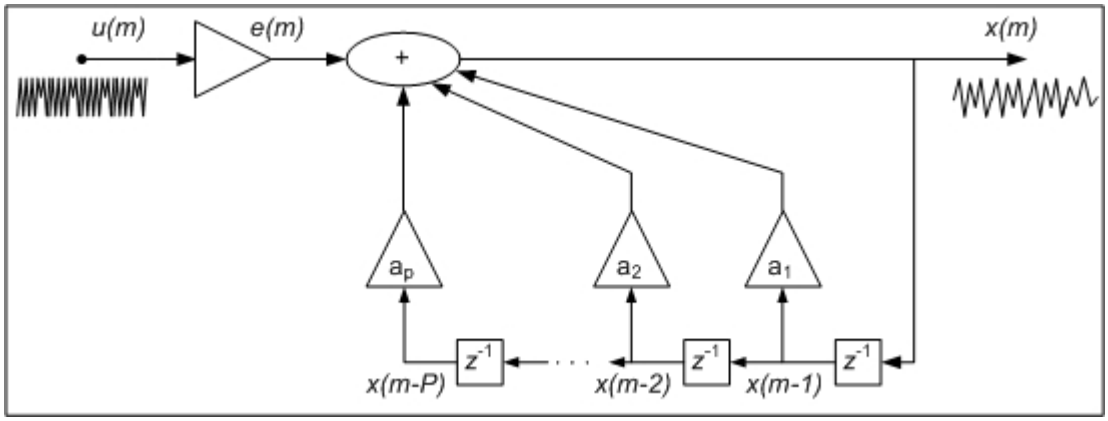


Figura 8 - Sinal de voz *x(m)* predito pelo preditor linear
Fonte: Vasegui (2000).

2.1.3.5 Codificação digital do sinal de áudio

No áudio digital, é necessária memória para armazenar esse sinal, largura de banda para a transmissão desse sinal e uma taxa, na relação sinal-ruído de quantização, proporcional ao número de bits por amostra. O objetivo de implementar a codificação é obter uma alta

fidelidade com o menor número de bits por amostra o possível, e um aceitável custo quanto a implementação. Algoritmos de codificação de áudio utilizam a estrutura estatística do sinal, e o modelo do sinal gerado. Em geral, existem duas principais categorias de codificadores de áudio: codificadores baseados em modelos, usados para codificação com alta taxa de compressão do sinal de voz e aplicados, por exemplo, na telefonia celular; e codificadores de transformação de base usados em codificação de alta definição (hi-fi[7] - *High fidelity*), aplicado em áudio digital.

A Figura 9 apresenta um diagrama de blocos simplificado da configuração de um codificador de voz do tipo usado na telefonia celular digital. No transmissor o sinal de voz é segmentado em pacotes de 20 ms a 30 ms de duração, quando os parâmetros da voz são assumidos como estacionários. Cada pacote de voz é analisado para extrair os parâmetros de excitação e filtro que podem ser utilizados na síntese da voz. No receptor, o modelo de parâmetros de excitação é usado para reconstruir a fala.

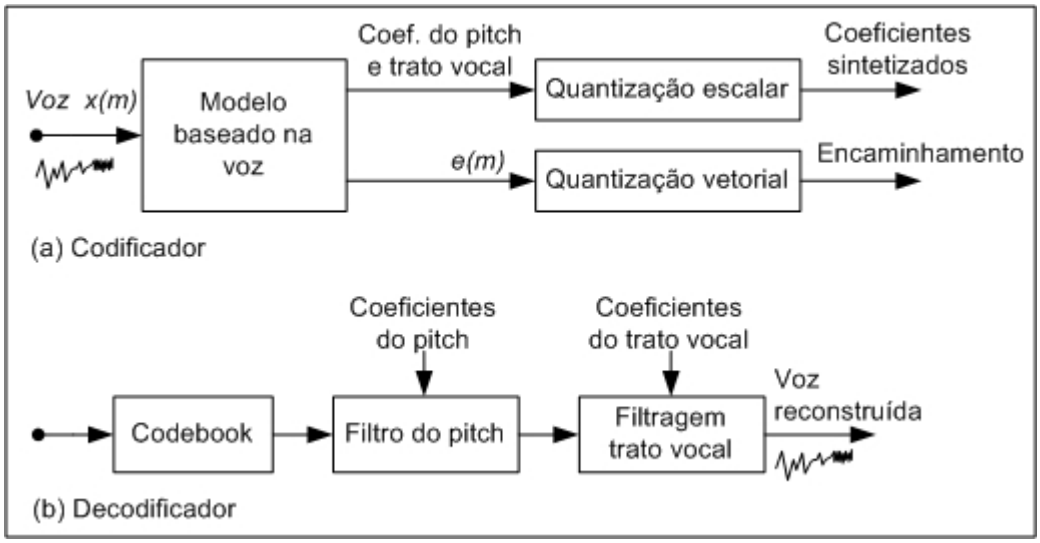


Figura 9 - Diagrama de bloco simplificado de um modelo de codificação de voz
Fonte: Vasegui (2000).

Um codificador baseado na transformação de domínio é apresentado na Figura 10. O objetivo da transformação é a de converter o sinal em um formulário em que se presta a uma interpretação mais conveniente para a manipulação. Na Figura 10 o sinal de entrada é transformado para o domínio da freqüência usando um banco de filtros, ou algum algoritmo específico tipo a transformada discreta de Fourier (DFT – *Discrete Fourier Transform*), ou a transformada discreta do cosseno (DCT – *Discrete Cosine Transform*), transformada discreta do seno (DST – *Discrete Sine Transform),* transformada discreta *wavelet* (DWT – *Discrete*

[7] hi-fi (hi-fi – *High fidelity*) é o termo usado para referenciar alta qualidade de reprodução de som e imagem que são altamente fieis aos sons originais.

Wavelet Transform), ou ainda outras possibilidades (DAVIS, 2002). Três principais vantagens da codificação de um sinal no domínio da freqüência são:

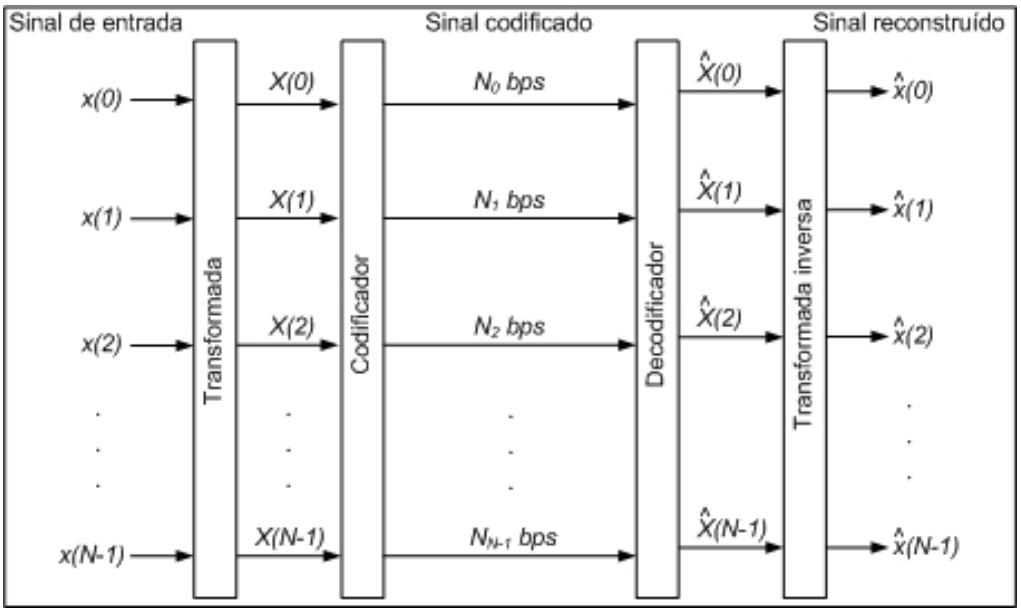


Figura 10 - Codificador baseado na transformação de domínio
Fonte: Vasegui (2000).

a) O espectro de frequências de um sinal de áudio, por exemplo, tem uma estrutura relativamente bem definida, por exemplo, a maior parte da potência do sinal é geralmente concentrada nas regiões mais baixas do espectro;

b) Freqüências com amplitude relativamente baixa são mascaradas nas proximidades de frequências com grande amplitude, o que pode ser codificado sem qualquer degradação audível;

c) As amostras de freqüência são ortogonais e podem ser codificadas de forma independente com diferentes precisões.

O número de bits atribuído a cada frequência de um sinal é uma variável que reflete a contribuição da freqüência para na percepção do nível de qualidade do sinal no momento da reprodução. Em um codificador adaptativo, a alocação de bits para diferentes frequências é feita de forma a variar com o tempo, e também as variações do espectro de potência do sinal.

2.1.3.6 Detecção do sinal no ruído

Na detecção de sinais ruidosos, o objetivo é determinar se a observação é a de apenas ruído, ou se contém algum sinal com informação significativa de fato. A observação do ruído *y(m)* pode ser modelado como na equação 5:

$$y(m) = b(m) \cdot x(m) + n(m) \quad ...(5)$$

onde *x(m)* é o sinal a ser detectado, *n(m)* é o ruído e *b(m)* é a referência binária para indicar a presença de sinal junto ao ruído, caso *b(m)=1*, ou não, caso *b(m)=0*. Se o sinal *x(m)* tem um formato conhecido, então o componente de correlação ou o filtro pode ser usado para detectar o sinal, conforme apresentado na Figura 11.

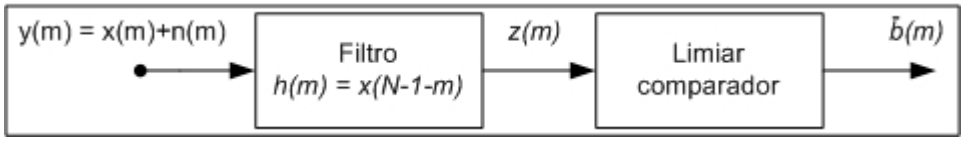


Figura 11 - Configuração do filtro seguido de um comparador para a detecção de sinais ruidosos
Fonte: Vasegui (2000)

A resposta ao impulso do filtro *h(m)* é dada pela equação 6:

$$h(m) = x(N-1-m) \qquad \text{para } 0 \le m \le N-1 \quad ...(6)$$

onde *N* é o tamanho de *x(m)*. A saída do filtro é dada pela equação 7:

$$z(m) = \sum_{m=0}^{N-1} h(m-k) \cdot y(m) \quad ...(7)$$

e a saída do filtro é comparada com o limiar e é feita uma decisão binária dada pela condição:

$$\bar{b}(m) = \begin{cases} 1 \ldots se \ldots z(m) \ge limiar \\ 0 \ldots se \ldots z(m) < limiar \end{cases}$$

onde $\bar{b}(m)$ é uma estimação do indicador seqüência binária de estado *b(m)*, e pode ser errada, em particular, se a SNR for baixa. A Tabela 1 enumera quatro possíveis resultados que *b(m)* e sua estimativa $\bar{b}(m)$ podem assumir.

Tabela 1 - Saídas possíveis do bloco de comparação pelo limiar

$\bar{b}(m)$	$b(m)$	Decisão do detector	
0	0	Ausência de sinal	Correto
0	1	Ausência de sinal	Falha
1	0	Presença de sinal	Falso alarme
1	1	Presença de sinal	Correto

Fonte: Vasegui (2000).

A escolha do limiar afeta diretamente a sensibilidade do detector. Quanto maior o limiar, menor a probabilidade de que o ruído venha a ser classificado como sinal, de modo que a taxa de alarme falso cai, mas a probabilidade de erro de sinal como ruído aumenta.

2.1.4 Amostragem e conversão analógico para digital

O sinal digital é uma seqüência de valores reais e / ou complexos, que representam a variação da quantidade de informação ao longo do tempo, espaço, ou qualquer outra variável. A base do sinal no tempo discreto é a unidade de amostra deste sinal *δ(m)* definido como:

$$\delta(m) = \begin{cases} 1 \ldots m = 0 \\ 0 \ldots m \neq 0 \end{cases}$$

onde *m* é o índice discreto do tempo. Um sinal digital *x(m)* pode ser expresso como a soma de uma série de amostras com determinada amplitude deslocadas no tempo (Equação 8).

$$x(m) = \sum_{k=-\infty}^{\infty} x(k) \cdot \delta(m-k) \qquad \ldots(8)$$

Muitos processos aleatórios, tais como voz, música, radar e sonar geram sinais que são contínuos no tempo. Sinais contínuos são ditos analógicos porque suas flutuações ao longo do tempo são análogas às variações do sinal fonte. Para o processamento digital, sinais analógicos são amostrados, e cada amostra é convertida em um determinado número de bits. O processo de digitalização deve ser realizado tal que o sinal original possa ser recuperado a partir de sua versão digital sem perda significativa de informação, e com uma fidelidade tão alta quanto a exigida pela aplicação a qual se destina o sinal. A Figura 12 apresenta um diagrama de blocos que ilustra a configuração de um processador de sinal digital com uma entrada analógica. O filtro passa-baixas (LPF – *Low Pass Filter*) remove as frequências do

sinal que estejam fora do intervalo de interesse. O amostrador (S/H – *Sample and Hold*) amostra o sinal com base em uma unidade periódica para converter o sinal em tempo contínuo em um sinal de tempo discreto (DAVIS, 2002).

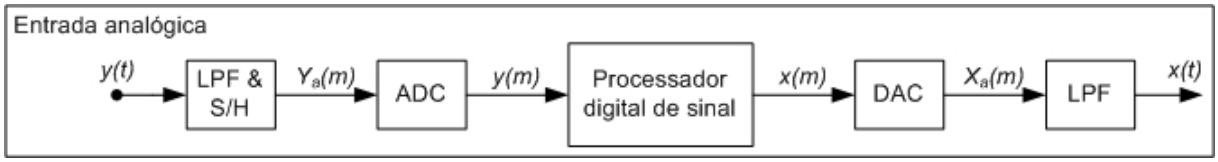


Figura 12 - Generalizando em blocos um sistema de processamento digital de sinais
Fonte: Davis (2002) e Shenoi (1995).

O conversor que transforma o sinal analógico em um sinal digital (ADC – A*nalogic to Digital Converter*) mapeia a amplitude de cada amostra em códigos de *n* bits. Após o processamento, a saída digital do processador pode ser convertido novamente em um sinal analógico com um conversor digital para analógico (DAC – *Digital to Analogic Converter*) e um filtro passa-baixas (LPF), como ilustrado na Figura 12.

2.1.4.1 Amostragem e reconstrução do sinal

A conversão do sinal analógico para o sinal digital consiste na amostragem e na quantização do mesmo. A amostragem do sinal pode ser modelada como o produto do sinal no tempo contínuo *x(t)* e um trem de impulsos periódicos *p(t)* como mostrado nas equações (9) e (10).

$$x_{amostrado}(t) = x(t) \cdot p(t) \quad ...(9)$$

$$x_{amostrado}(t) = x(t) \cdot \sum_{m=-\infty}^{\infty} \delta(t - m \cdot T_s) = \sum_{m=-\infty}^{\infty} x(m \cdot T_S) \cdot \delta(t - m \cdot T_s) \quad ...(10)$$

onde T_s é o período da amostragem e *p(t)* é a função da amostragem definida por (equação 11):

$$p(t) = \sum_{m=-\infty}^{\infty} \delta(t - m \cdot T_s) \quad ...(11)$$

O espectro *P(f)* da função de amostragem *p(t)* é também um trem de impulsos periódicos dado pela equação 12:

$$P(f) = \sum_{k=-\infty}^{\infty} \delta(f - k \cdot F_s) \quad ...(12)$$

onde $Fs=1/T_S$ é a frequência de amostragem. Assim, tendo a multiplicação dos dois sinais no domínio do tempo, temos o equivalente do domínio da frequência como sendo a convolução dos dois sinais. Desta forma, tem-se como espectro de frequência resultante a equação 13.

$$X_{amostrado}(f) = FT[x(t) \cdot p(t)] = X(f) * P(f) = \sum_{k=-\infty}^{\infty} X(f - k \cdot F_s) \quad ...(13)$$

onde *FT[.]* significa a aplicação da transformada de Fourier. Na equação 13 a convolução do espectro do sinal *X(f)* com cada impulso $\delta(f-kF_s)$, desloca *X(f)* centrando o mesmo em kF_s. Assim como expresso na equação 13, a amostragem do sinal *x(t)* resulta na repetição periódica do sinal *X(f)* no espectro, centrado nas frequências *0*, $\pm Fs$, $\pm 2Fs$,...(VASEGUI, 2000) (SHENOI, 1995) (DAVIS, 2002).

Para que esse modelo matemático seja aplicado na reconstrução "perfeita" do sinal (desconsideradas as perdas entre os períodos amostrados), a taxa de amostragem deve ser maior do que o dobro do componente de frequência F_s mais elevado do sinal da mensagem. Isso se deve a aplicação de um filtro *anti-alising* passa-baixas na entrada do amostrador para excluir frequências maiores do que F_s antes da amostragem. Dessa forma, a aplicação da amostragem permite a redução do sinal da mensagem (de duração finita) continuamente variável até um número limitado de valores discretos por segundo (HAYKIN, 2004).

O processo de amostragem pode ser visualizado na Figura 13.

2.1.4.2 Quantização

Para o processamento digital dos sinais, os sinais amostrados precisam ser quantizados e mapeados para códigos binários de *n* bits. Para a quantização do sinal, o mesmo é dividido em amplitude em 2^n níveis discretos, e cada amostra é quantizada via aproximação para o nível mais próximo. Na sequência, esse sinal é mapeado para um código binário, atribuído ao nível. A Figura 14 ilustra a quantização de um sinal discreto em 4 níveis.

O mapeamento entre os valores de amplitude analógicos das amostras $x_a(m)$ e o seu valor quantizado pode ser expresso pela equação 14.

$$x(m) = Q[x_a(m)] \quad ...(14)$$

onde *Q[.]* é a função de quantização.

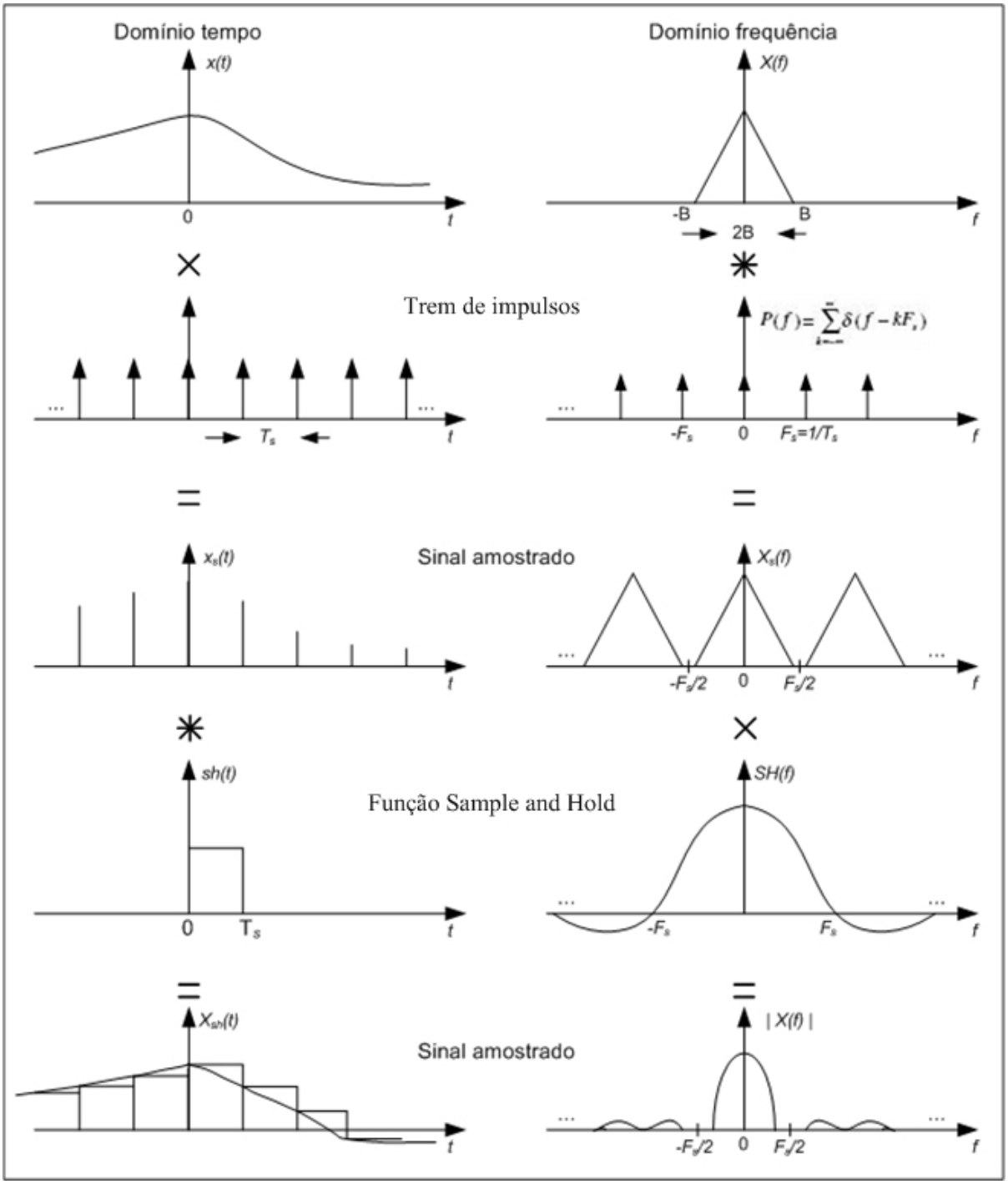


Figura 13 - Processo de amostragem: análise do sinal nos domínios do tempo e em freqüência. Fonte: Vasegui (2000).

O desempenho do quantizador é mensurado pela relação sinal ruído de quantização (SQNR – *Signal-to-Quantization Noise Ratio*) por bit. O ruído de quantização é dado pela equação 15.

$$e(m) = x(m) - x_a(m) \quad ...(15)$$

E por fim, a SQNR é dada pela equação 16.

$$SQNR(n) = 10 \cdot log_{10}\left(\frac{E[x^2(m)]}{E[e^2(m)]}\right) \quad ...(16)$$

onde $E[x^2(m)]$ é a potência do sinal, $E[e^2(m)]$, é a potência do ruído.

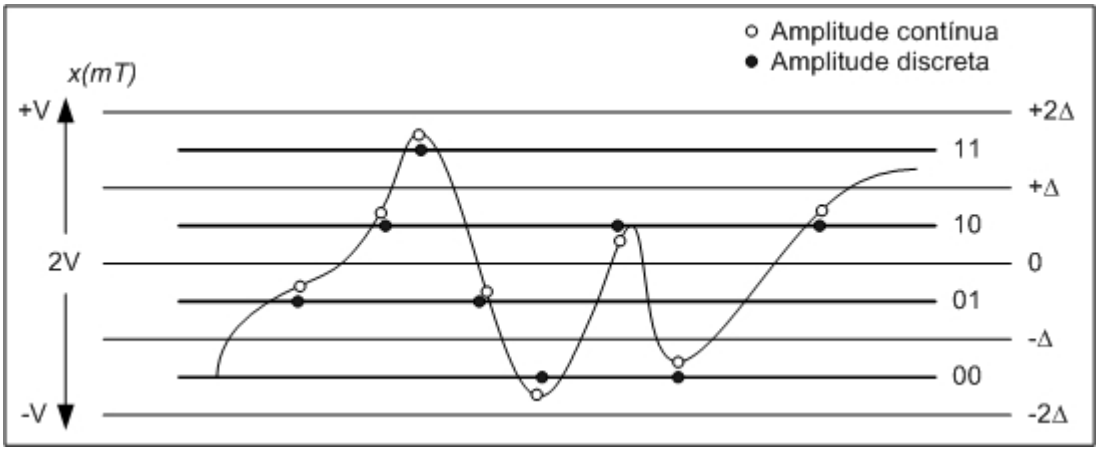


Figura 14 - Processo de quantização e codificação de um sinal.
Fonte: Vasegui (2000).

2.2 DSP EM VOIP

DSP apresenta-se nos dias de hoje empregado nos mais variados campos. Entre estes estão os sistemas de comunicação de voz via redes IP, ou mais comumente chamado, VoIP.

Sistemas de VoIP e técnicas de DSP estão diretamente relacionados no que diz respeito ao trato matemático do sinal de voz, tanto no que está para o locutor como para o ouvinte. Alguns aplicativos, dentre eles, os supressores de silêncio, os canceladores de eco, os codificadores de voz, os geradores de ruído Gaussiano e o controle automático de ganho compõem a infra-estrutura mínima no que diz respeito às técnicas de DSP para um sistema de VoIP. Todos estes aplicativos serão apresentados nas seções seguintes, dando tratamento específico, às técnicas de detecção e supressão de silêncio, foco deste trabalho.

Considerando toda a complexidade da estrutura dos sistemas de voz sobre redes IP, muitas vezes pode ficar um pouco obscuro onde está sendo empregado DSP nesta tecnologia. Como o foco deste trabalho é a análise dos algoritmos, não haverá aqui aprofundamento a fim de descrever cada componente do sistema de VoIP e a localização de cada técnica de DSP dentro destes componentes. Apenas como uma forma ilustrativa, é apresentada na Figura 15 a seqüência dos eventos para a realização de uma comunicação de voz via rede IP, e por conseqüência a localização das técnicas de DSP dentro da rotina do sistema.

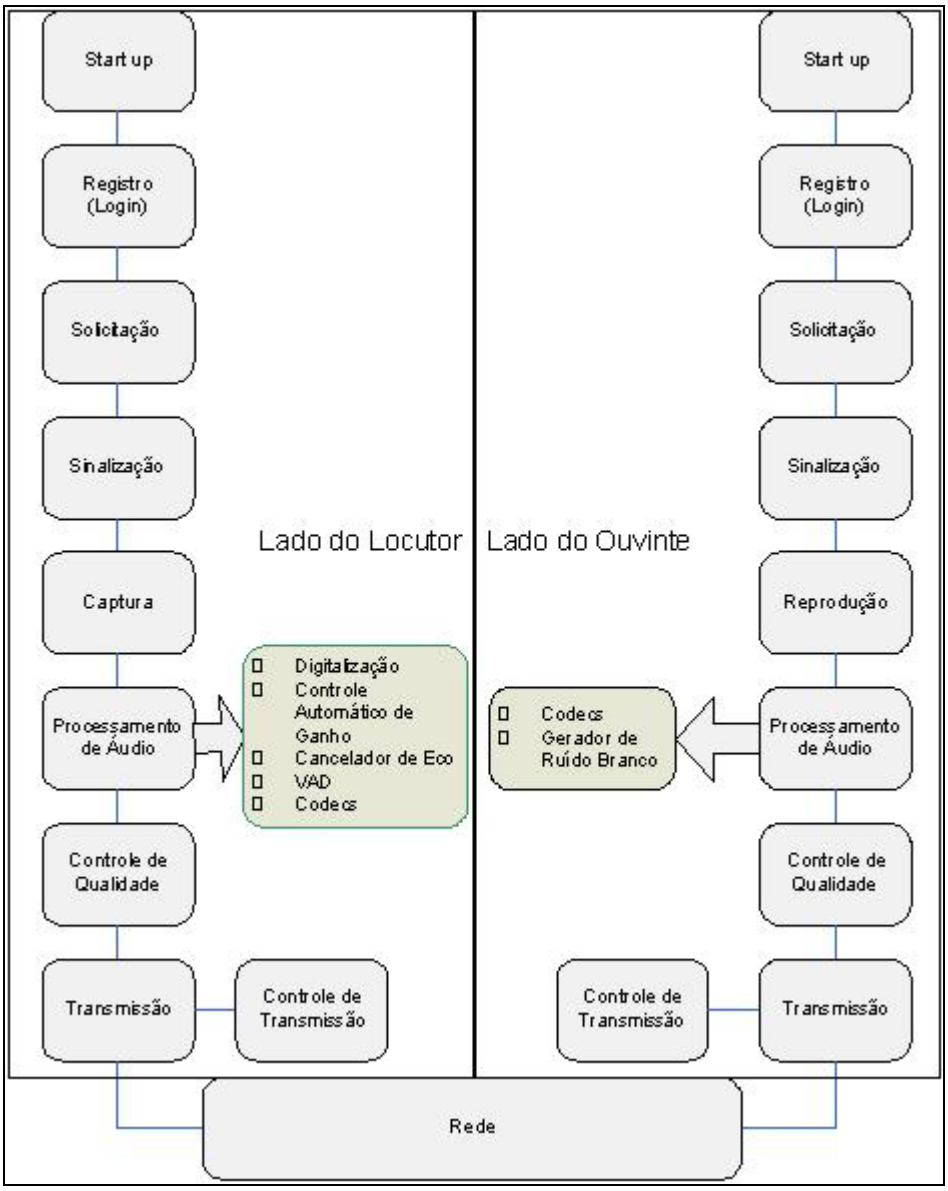


Figura 15 - Seqüência das macro-tarefas para a consolidação de um sistema de VoIP.
Fonte: O autor (2009).

Como citado, todos os sistemas que envolvem processamento de voz possuem muitas complicações quando expostos a extremas condições de ruído. Em sistemas de VoIP, não é muito diferente, já que o mesmo utiliza algoritmos de VAD para a detecção de períodos de silêncio durante as conversações que são feitas, objetivando não transmitir, via as redes IP, informações com conteúdo irrelevante (ruído captado pelo microfone do locutor) para o bom entendimento da conversação. Normalmente, nos ambientes dos locutores, o ruído é inconstante e variável. Verificando este problema, abre-se espaço para a pesquisa e desenvolvimento de algoritmos adaptativos ao ruído ambiente, objetivando a detecção de algum sinal relevante, no caso aqui, voz.

Algoritmos adaptativos utilizados no processamento de voz buscam em muitos casos, o treinamento do mesmo através da leitura dos dados passados pelo próprio ambiente a fim de

supor um próximo valor bastante próximo do real. No caso do ruído, em específico, poder cancelá-lo, desconsiderá-lo ou ainda conseguir uma harmonia de trabalho sem que o mesmo interfira no processamento da voz. Estes algoritmos adaptativos, no caso específico de aplicações de VoIP, são diretamente utilizados, de forma que neste trabalho isto está sendo considerado a ponto das implementações serem técnicas de detecção e supressão de silêncio adaptativas às condições de variabilidade do ruído ambiente.

2.2.1 Canceladores de eco

Canceladores de eco tem como função eliminar o eco gerado em uma chamada. Normalmente o eco é gerado quando há conexão com a rede pública de telefonia comutada, no *hardware* do sistema, chamado híbrida (que converte a chamada de quatro para dois fios) e este eco pode ser eliminado já no *gateway*[8] de entrada da rede IP. O eco acústico também existe na rede IP. Ele é gerado da mesma forma que na rede de telefonia convencional (originado das reflexões do sinal de voz do locutor no seu ambiente) e é agravado quando se utiliza um computador como meio de comunicação devido à realimentação do alto falante para o microfone (Figura 16). O eco é variável de chamada para chamada e até mesmo em uma mesma chamada, ou seja, é necessário que se tenha um mecanismo que se ajuste de acordo com a variação do eco e os filtros adaptativos possuem essa capacidade (CORSETTI, 2004).

Nas comunicações de VoIP, são necessários dois canceladores de eco. Um cancelador de eco localizado no telefone IP, ou junto a aplicação instalada em um PC (para eliminar o eco acústico quando da comunicação por viva-voz) e o outro no *gateway* de interface com a rede pública de telefonia (para eliminar o eco híbrido) (Figura 17) (CASTELLO, 2004).

Um detalhamento maior sobre as questões referentes à cancelamento de eco são encontradas disponíveis nas recomendações do *Telecommunication Standardization Sector of International Telecommunication Union* (ITU-T), G.164 (ITU 1988a) que trata dos supressores de eco, G.165 (ITU 1993a) que trata dos canceladores de eco, G.167 (ITU 1993b)

[8] *Gateway* é a entidade de interconexão do sistema IP com a rede pública de telefonia comutada (PSTN - *Public switched telephone network*). Seu principal objetivo é fornecer um link de comunicação entre as redes (BALBINOT, 2002) (CASTELLO, 2004) (CONWAY, 2000).

que descreve os controladores de eco acústico e a G.168 (ITU 1997a) que descreve os canceladores de eco para redes digitais.

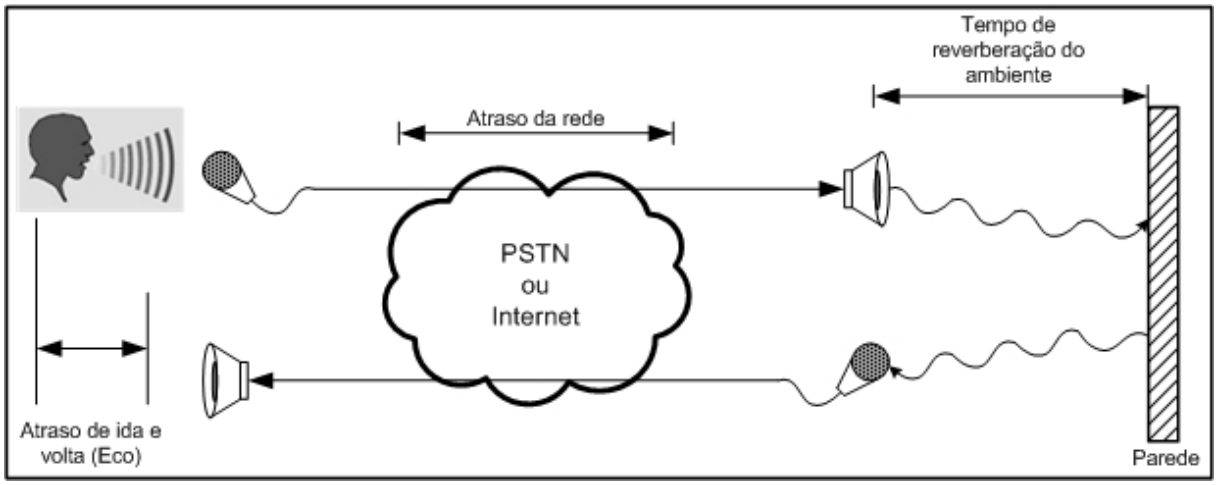


Figura 16 - Formação do eco acústico.
Fonte: Adaptado de HERSENT (2005).

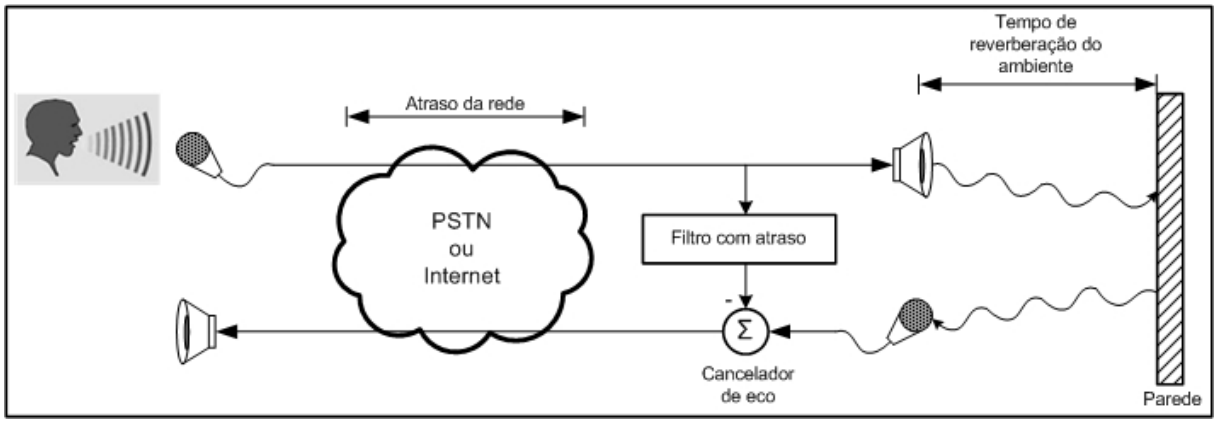


Figura 17 - Exemplo de um sistema sofrendo eco acústico com a adição de um cancelador de eco acústico no receptor.
Fonte: Adaptado de BALBINOT (2004).

2.2.2 Codecs

Antes de fazer qualquer referência aos codificadores de voz propriamente ditos, é preciso digitalizar o sinal de voz. O modelo genérico de um sistema digitalizador pode ser representado por um conjunto de quatro estágios seqüenciais (FERNANDES, 2003):

- Filtro: O sinal analógico é submetido a um filtro passa baixa de forma a limitar o sinal analógico segundo o critério de Nyquist que estabelece que a freqüência a ser amostrada deve ser no mínimo duas vezes a freqüência máxima desejada;

- Amostrador: Nesse estágio, um sinal contínuo no tempo é transformado em um sinal discreto;
- Quantizador: Processo de mapeamento do sinal discreto, para um número contável, representando os vários níveis amostrados;
- Codificador: Responsável pela representação binária do sinal digital, com o compromisso de manter a menor taxa de codificação possível e a melhor qualidade do sinal sintetizado.

Existem três tipos de codificadores de voz: Codificação por forma de onda, paramétricos e híbridos. Balbinot (2004) detalha cada um dos tipos de codificadores como segue.

Codificadores de forma de onda têm uma abordagem no domínio do tempo e são os mais intuitivos. Eles têm como objetivo codificar o sinal considerando apenas a sua forma de onda, sem considerar nenhuma outra característica. Esse tipo de codificação se dá por meio simplesmente das operações de amostragem e quantização. A codificação pode ser a PCM (*Pulse Code Modulation*), a DPCM (*Differential Pulse Code Modulation*), onde o que é codificado é a diferença entre as amostras consecutivas, ou ADPCM (*Adaptative Differential Pulse Code Modulation*), que é a versão adaptativa desta última.

Codificadores de fonte ou paramétricos têm uma abordagem no domínio da freqüência. Eles têm como objetivo codificar o sinal considerando apenas o modo através do qual este foi gerado, ou seja, sua fonte. No caso da voz, a fonte é o próprio trato vocal da pessoa que fala. É feita uma parametrização das características da fonte em várias janelas ao longo da produção do sinal em questão. No caso da voz, essas características são: se o som é vozeado (faz as cordas vocais vibrarem), se é não vozeado (não faz as cordas vocais vibrarem), o *pitch* do sinal e, finalmente, o filtro digital que modela o trato vocal. Esta última característica é obtida através da análise LPC aplicada a uma janela do sinal. Exemplos de codificadores de fonte são os ditos Vocoder LPC, o RELP (*Residual-Excited Linear Predictive*) e o QV (*Vetorial Quantization*).

Detalhes de implementação e aplicação de codificadores RELP e QV podem ser vistos em Taguchi (2003) para o RELP e em Fleury (2005) para o QV.

Contudo, codificadores de forma de onda têm uma relação de "qualidade x taxa de transmissão" quase unitária, ou seja, para a qualidade aumentar, deve-se aumentar igualmente a taxa de transmissão. No entanto, isso não é desejável em sistemas de voz sobre IP. Codificadores de fonte, por sua vez, possuem taxas de transmissão muito baixas, mas, por mais que a mesma seja ampliada, a qualidade não melhora significativamente. Assim,

codificadores de forma de onda possuem uma qualidade muito boa, mas uma taxa de transmissão muito alta; e codificadores de fonte possuem uma qualidade ruim, mas uma taxa de transmissão muito baixa.

Para resolver este problema, são utilizados os codificadores híbridos, que reúnem características de ambos os codificadores citados. Dessa maneira, pode-se ter uma qualidade muito boa com baixas taxas de transmissão. Um exemplo para esse tipo de codificador é o CELP (*Code Excited Linear Prediction*). Os padrões mais recentes para codificadores de voz da ITU são os G.728 (LD-CELP - *Low-Delay Code Excited Linear Prediction*) (ITU, 1992), G.729 (ITU, 1996c), G.729A (CS-ACELP - *Conjugate-Structure Algebraic-Code-Excited Linear-Prediction*) (ITU 1996e) e o G.723.1 (ACELP - *Algebraic-Code-Excited Linear-Prediction*) (ITU, 1996b). Os mesmos padrões também são detalhados em Rosenberg (1998) Ohrtman (2004) e Herseng (2005), além do G.722 (ITU, 1988c). Estes padrões diferem pelo custo e pela qualidade, mas a tendência é que todos estes se unifiquem em um único padrão. Devido à menor capacidade das redes, os algoritmos tendem a ser cada vez mais complexos, para gerar taxas de transmissão mais baixas.

Os fatores que devem ser levados em conta, quando comparamos diferentes técnicas de Vocoding, ou Vocoders, são:

a) taxa de bits (*Bit Rate*): na tecnologia VoIP, o meio de transmissão é compartilhado entre os dados e a voz, porém muitos Vocoders ainda operam com taxas fixas de transmissão, independente do sinal de voz que é transmitido, quando a idéia é evoluir-se para o uso de taxas variáveis de transmissão;

b) atraso: os atrasos se devem, basicamente, a dois componentes importantes que são o atraso de quadro, onde é preciso esperar o número de bits do quadro para poder processá-lo, e o atraso de processamento da voz, que se deve ao tempo necessário para codificação e decodificação;

c) complexidade do algoritmo: geralmente medida em termos da velocidade de computação da quantidade de RAM (*Random Access Memory*) e ROM (*Read-Only Memory*) que são exigidos. Uma complexidade maior do algoritmo resulta em custo maior de processamento e de consumo de energia (importante em aplicações portáteis);

d) qualidade: medida relativa da qualidade com que soa a voz sob condições ideais, ou seja, voz clara, sem erros de transmissão e com somente um processamento de codificação.

Alguns dos codificadores aqui citados possuem recomendações anexas. Estas recomendações anexas são funcionalidades específicas dos codificadores. Das recomendações anexas a que pode ser de maior interesse com relação ao foco deste trabalho é o Anexo B (ITU 1996d) do G.729. O Anexo B do G.729 descreve o detector de voz ativa e gerador de ruído de conforto. Ambos são usados na compressão de silêncio, tanto no G.729 como no G.729 Anexo A (FERNANDES, 2003) (ITU 1996e).

Conforme Balbinot (2002), a adoção de mecanismos de detecção e supressão de silêncio, agregada as técnicas adequadas de compressão podem possibilitar aos sistemas de voz sobre IP a redução da banda utilizada na ordem de até vinte vezes.

2.2.3 Ruído de conforto

Geradores de ruído de conforto nada mais são do que geradores de ruído branco[9], ou ruído Gaussiano. Este tipo de ruído é amplamente utilizado em aplicações DSP, seja para mitigar os efeitos causados por perdas de pacotes de voz ou apenas para gerar um ruído de fundo de forma a indicar ao usuário ouvinte que o canal de comunicação ainda está aberto durante o silêncio inerente a uma comunicação.

O ruído branco tem esse adjetivo "branco" atribuído no sentido de que a luz branca contém intensidades iguais em todas as frequências dentro da banda visível de radiação eletromagnética. Isso se atribui ao fato de que esse ruído se caracteriza por ter densidade espectral de potência independente de frequência de operação. A Figura 18 apresenta um gráfico característico do sinal do ruído branco (a), o espectro densidade de potência (b) e a função de autocorrelação (c). Na Figura 18 N_0 faz referência à densidade espectral de ruído. O fato de ser dividido por 2, representa a divisão da densidade no espectro, sendo metade para frequências negativas e a outra metade para frequências positivas.

[9] Conforme Haykin (2004), o ruído branco, cuja densidade espectral de potência é independente da frequência de operação, tem o adjetivo branco atribuído no sentido de que a luz branca contém intensidades iguais de todas as frequências dentro da banda visível de radiação eletromagnética.

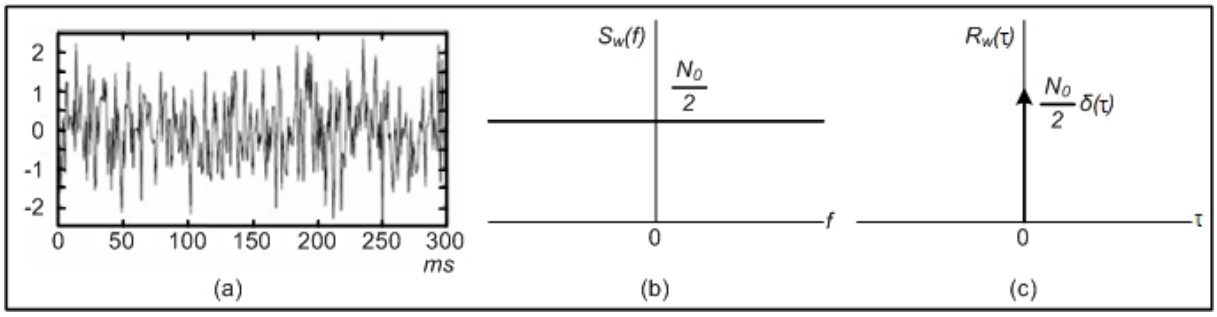


Figura 18 - Características do ruído branco. a) Ruído branco. b) Densidade espectral de potência. c) Função de autocorrelação.
Fonte: a) VASEGUI (2000). b) e c) HAYKIN (2004).

Para sistemas de VoIP, a geração de ruído de conforto tem um único objetivo, a manutenção da qualidade percebida pelo usuário deste tipo de aplicações. A utilização de ruído pode ocorrer de duas formas. Uma que é quando ocorrem significativos atrasos, na chegada do lado do ouvinte, de pacotes de áudio e este atraso extrapole a latência do *buffer* de reordenamento[10] de pacotes para a reprodução. Desta forma o ruído de conforto é produzido e inserido no lugar deste pacote de áudio que não chegou a tempo de ser reproduzido. Esse efeito da introdução de ruído é preferível a simples inserção de silêncio, conforme Balbinot (2002).

O outro caso para a utilização de ruído de conforto é o de maior interesse para este trabalho, é quando da utilização de supressores de silêncio nas aplicações de VoIP. Conforme Benyassine (1997) e Herseng (2005), algoritmos de VAD são usualmente combinados com um *confort noise generator* (CNG), gerador de ruído de conforto, o qual tenta gerar um ruído equivalente ao do ambiente para o ouvinte durante os períodos de silêncio do locutor. A presença de ruído de conforto no áudio do ouvinte evita que o mesmo pense que a chamada deu-se por encerrada em um momento de silêncio do locutor.

O funcionamento do CNG ocorre no lado do ouvinte, ficando geralmente atrelado à existência do algoritmo no codificador do mesmo, ou no lado do locutor que pode transmitir via um *Payload* de *Confort Noise* (IETF, 2002), quando da utilização do *Real Time Transport Protocol* (RTP) (IETF, 2003a). O CNG é acionado no receptor quando o mesmo recebe um pacote de dados provenientes do locutor informando a detecção de silêncio junto ao mesmo, de forma que o ruído será gerado até o momento do recebimento de um novo pacote ativo de voz.

Para o caso de alguns codificadores desenvolvidos, como no caso do G.723.1 (anexo A) (ITU, 1996b) ou o G.729 (anexo B) (ITU, 1996d), os mesmos já possuem opções para

[10] Um detalhamento da implementação e do funcionamento do *buffer* de *dejitter* são apresentados no Apêndice C deste trabalho.

enviar informações suficientes que permitirão ao decodificador remoto reconstruir para o ouvinte um ruído do ambiente do locutor próximo do ruído de fundo original (HERSENT, 2002).

2.2.4 Controle automático de ganho

O controle automático de ganho é um compensador para diferentes intensidades de som no microfone devido às diferentes distâncias que pode haver entre o microfone e a boca do locutor. Um controle rápido e robusto permite uma compensação instantânea de diferentes intensidades de som proporcionadas pelo ambiente onde se encontra o locutor.

Em sistemas de VoIP os algoritmos de controle automático de ganho desempenham uma função muito importante na manutenção da estabilidade do nível sonoro que é processado pelo sistema. Isto porque, por exemplo, os algoritmos de VAD, podem perder sua funcionalidade quando os níveis sonoros tornam-se muito altos, e em caso de níveis sonoros muito baixos, os algoritmos podem degradar completamente a voz.

2.3 DETECÇÃO E SUPRESSÃO DE SILÊNCIO

Sendo as técnicas de detecção e supressão de silêncio o foco deste trabalho, neste item é apresentado um embasamento teórico mínimo necessário para o entendimento do assunto e para a realização das implementações documentadas na seqüência desta dissertação. São abordadas aqui questões ligadas aos métodos empregados para a detecção de períodos de silêncio em comunicações de voz, aspectos desejáveis para os algoritmos de VAD além de um detalhamento dos mesmos. Também são apresentadas descrições dos parâmetros utilizados para a construção dos pacotes de voz tomados como padrão para o trabalho, determinação do limiar de silêncio, técnica de recobrimento e transformadas para a mudança de domínio (tempo – frequência).

2.3.1 Métodos empregados para a detecção de silêncio

Geralmente o método empregado para a detecção de silêncio, consiste na análise da quantidade de energia em cada pacote de voz como forma de estimativa para a detecção de silêncio (BALBINOT, 2002) (RENEVEY, 2001). Com este método, todo o pacote com uma determinada quantidade de energia abaixo de um determinado valor (limiar de silêncio) será, em alguns casos, reavaliado e se ainda não estiver dentro dos parâmetros relevantes para a aplicação, será considerado silêncio e conseqüentemente, o pacote não será transmitido. Desta forma o método estará proporcionando uma ocupação menor de banda da rede liberando assim, espaço para outros tipos de dados, seja da aplicação de voz ou de outras aplicações, que possam estar sendo também transmitidos pela mesma rede.

A detecção do silêncio e a sua remoção não se referem apenas ao silêncio que ocorre quando uma das partes em uma conversação (locutor e ouvinte) está em silêncio ou entre pequenos intervalos de fala, mas inclusive pausas entre palavras e sílabas. A remoção destes períodos de silêncio, especialmente pausas entre palavras e sílabas, merece especial cuidado na implementação das técnicas de detecção e supressão de silêncio. Algoritmos de VAD necessitam ser bastante específicos em suas operações. Isto se faz indispensável para que não sejam causados problemas de perda de qualidade da voz, dentre estes problemas, o efeito de *clipping* (corte) ocasionado na reprodução da voz e a não detecção de sons fricativos.

O efeito de *clipping* causa interrupções abruptas da fala, como o corte súbito do áudio no meio de uma sílaba ou letra, por exemplo, degradando de forma drástica a qualidade da voz a ser transmitida.

No referente aos sons fricativos, os mesmos podem ser os fonemas /s/, /f/ e /sh/ (HERSENT, 2002). São produzidos pela formação de uma constrição em um ponto do trato vocal e pela expulsão de ar por esta constrição, criando uma turbulência que produz uma fonte de ruído para excitar o trato vocal (RODRIGUES, 1988). Em termos simples, isto ocasiona uma grande quantidade de inversões no sinal da magnitude da amostra de voz devido à turbulência citada.

Outro ponto interessante relacionado à supressão do silêncio, é que ao referenciar períodos de silêncio na fala, não se está mencionando apenas pacotes com conteúdo de energia igual a zero ou que não estejam dentro dos parâmetros estabelecidos para se confirmar à presença de voz em um pacote. Também se está referenciando sons incompreensíveis ou ruídos ambientes ou não, que estão sempre presentes nas comunicações de voz.

2.3.2 VAD

VAD é o processo de separação da conversação por voz em segmentos de voz ativa ou inativa (BECKER, 2005) (RENEVEY, 2001) (TANYER, 1998). Desta forma são discriminados quais pacotes de voz serão transmitidos via rede IP e quais serão suprimidos pelo sistema.

Os algoritmos de VAD possuem uma série de aplicações dentro da área de processamento digital de sinais. A maioria destas necessita de soluções para a redução do ruído ambiente, de forma que trabalham de maneira combinada com algoritmos de VAD (RAMIREZ, 2002).

Algoritmos de VAD formam um componente inserido junto ao gateway, ou junto ao terminal ou aplicação do usuário que atua suprimindo os períodos de silêncio em conversações de voz. Os algoritmos de VAD operam do lado do lado do locutor, e podem frequentemente se adaptar as variações do nível de ruído em relação ao sinal de voz (DAVIS, 2002).

Conforme Tanyer (2000), a função básica dos algoritmos de VAD é extrair características e parâmetros do sinal de entrada para comparar com um limiar de silêncio, usualmente extraído das características do ruído e do próprio sinal de voz. Na seqüência é feita a decisão por voz ativa caso os valores mensurados sejam superiores ao limiar.

A inserção de algoritmos de VAD nas aplicações específicas de VoIP tem um importante papel no que diz respeito a todo o sistema de comunicação, deste o locutor até o receptor. Podendo as técnicas de detecção e supressão de silêncio, dentro das quais estão inseridos os algoritmos de VAD, serem o maior benefício para comunicações em tempo real via Internet quando se faz referência à limitação física das redes de pacotes de dados. Isto tudo devido à economia de banda que as mesmas técnicas de supressão podem vir a proporcionar.

2.3.3 Aspectos desejáveis para os algoritmos de VAD

Objetivando a detecção e supressão de silêncio em sistemas de comunicação de tempo real via redes IP, existem alguns aspectos desejáveis que precisam ser considerados e

respeitados quando do uso de algoritmos de VAD neste tipo de aplicação, em especial, voz sobre IP. Citam-se alguns destes aspectos (PRASAD, 2002) (SANGWAN, 2002a):

- Boa regra de decisão: diz respeito ao fato de se explorar algumas propriedades físicas da voz a fim de dar maior consistência ao julgamento e classificação dos segmentos do sinal de voz em silêncio ou voz;
- Adaptabilidade ao ruído variável: adaptabilidade ao ruído ambiente não estacionário torna o algoritmo mais robusto permitindo a aplicação do mesmo em ambientes mais específicos;
- Baixa complexidade computacional: baixa complexidade computacional exige rapidez e simplicidade do algoritmo a fim de garantir a aplicabilidade em sistemas de tempo real;
- Baixa perda de qualidade da voz: o algoritmo necessita acrescentar um mínimo de perdas quando da supressão dos segmentos considerados silêncio, do contrário o sistema se tornará inviável e não comercializável;
- Economia de banda maximizada: objetivo principal do VAD, economizar banda com o fim de tornar facilitada a comunicação de voz em tempo real, via mínima ocupação da rede IP.

É importante salientar a relação direta que existe entre alguns dos aspectos desejáveis para os algoritmos de VAD citados acima. Por exemplo, a relação que se pode fazer entre boa regra de decisão e baixa complexidade computacional, ou baixa perda de qualidade da voz e economia de banda maximizada.

Para a primeira relação, boa regra de decisão e baixa complexidade computacional, faz-se necessário um meticuloso trabalho para que o algoritmo de VAD possa fazer a análise de características físicas da voz e ainda assim, ter uma baixa complexidade. Isto porque a maioria das técnicas referenciadas nas bibliografias relacionadas utilizam uma combinação de algoritmos.

Já na segunda relação feita, baixa perda de qualidade da voz e economia de banda maximizada, uma tende a ser o inverso da outra, por um lado, e compatíveis por outro. Geralmente quanto maior for a economia de banda proporcionada pela detecção do silêncio feita pelo VAD, teoricamente mais facilitado será o tráfico de pacotes de voz pela rede IP. Mas em outro caso, um percentual muito elevado de supressão de silêncio pode degradar a qualidade da voz suprimindo partes da fala do locutor que seriam importantes para o bom entendimento da conversação por parte do ouvinte.

2.3.4 Construção dos pacotes de voz

Para a determinação do tamanho dos pacotes de voz, precisam ser levados em consideração alguns aspectos. Como se está falando de comunicações em tempo real, os pacotes precisam ser pequenos, em torno de 10 ms (com no caso do Vocoder G.729) a 30 ms (com no caso do Vocoder G.723) (HERSENG, 2005), de forma a facilitar o tráfego dos mesmos pela rede evitando assim constantes atrasos e grandes perdas de qualidade quando da perda de pacotes no transporte pela rede IP. Outro aspecto relevante é a freqüência de amostragem utilizada pelo sistema. Diferentes freqüências de amostragem irão alterar o número de amostras por pacotes, o que certamente eleva o tempo de processamento do mesmo no caso do aumento da freqüência.

A equação 17 apresenta a fórmula para a determinação da quantidade de amostras do pacote de voz a ser processada pelo algoritmo de VAD. Na mesma equação 17, N representa a quantidade de amostras do pacote, t_{pacote} representa o tempo determinado para o pacote e $f_{amostragem}$, a freqüência de amostragem utilizada pelo sistema para a discretização do sinal de voz.

$$N = \frac{t_{pacote}}{\frac{1}{f_{amostragem}}} \qquad ...(17)$$

Considerando ainda a equação 17, e uma frequência de 8 kHz, quando aplicada uma codificação PCM a 8 bits por amostra, pode-se estimar que cada pacote de áudio de 20 ms transporta 160 amostras do mesmo. Assim, tem-se 160 bytes de dados por pacote, ou 1280 bits de dados por pacote.

2.3.5 Parâmetros para a determinação da presença de voz no pacote de áudio

Todas as técnicas de detecção e supressão de silêncio necessitam de parâmetros para que possam mensurar a existência ou não de voz em um determinado segmento de fala. Os parâmetros relevantes devem ser extraídos do sinal de áudio de forma que possam fazer uma boa distinção entre os segmentos de voz ativa e voz inativa (RENEVEY, 2001). Como cita Tanyer (1998), para o reconhecimento dos segmentos de voz, em voz ativa ou inativa, as

propriedades usualmente utilizadas são o nível de energia, o "*pitch*" da voz, a taxa de cruzamentos do zero, propriedades estatísticas e análise espectral.

O parâmetro mais comum para a avaliação e determinação da existência ou não de voz é o nível de energia das amostras (RENEVEY, 2001) (BALBINOT, 2002). As amostras são analisadas uma a uma e no final é tirada a energia média do pacote de voz, assim como mostra a equação 18. Nesta Equação, E_m é a energia média do pacote de voz, $E_{amostra}$ é a energia da amostra de voz e N o número total de amostras do pacote determinado pela equação 18.

$$E_m = \frac{1}{N}\sum_{0}^{N-1} E^2_{amostra} \qquad ...(18)$$

O "*pitch*" é a freqüência fundamental de vibração das cordas vocais. Medidas de "*picth*" têm recebido especial atenção na pesquisa de voz. O "*pitch*" pode ser determinado no domínio freqüência pelo cálculo do espaçamento espectral entre picos do espectro ou, no domínio tempo pela medida direta do período da forma de onda da voz (RODRIGUES, 1988).

A taxa de cruzamentos do zero se refere à quantidade de vezes que o sinal de voz tem o sinal da sua magnitude invertido. Apesar de ser uma estimativa grosseira, em alguns casos pode ser muito eficiente. Como se sabe, a energia dos sons vozeados (ativos) tende a se concentrar abaixo de 3 kHz, enquanto que a energia dos fricativos geralmente está concentrada acima de 3 kHz (RODRIGUES, 1988). Deste modo a medida do número de cruzamentos de zero pode ser utilizada para se decidir se um determinado sinal de voz é ativo ou inativo. Já Prasad (2002) cita que o número de cruzamento do zero para um pacote de 10 ms de voz, por exemplo, varia dentro de uma faixa fixa, sendo o valor entre 5 e 15 cruzamentos.

Para uma avaliação estatística, diferentes algoritmos podem ser empregados, dentre eles os que utilizam o cálculo da variância, equação 19, do sinal de voz como citado por Prasad (2002). Como exemplo, o caso do sinal de voz avaliado no domínio freqüência, a verificação da variância deste sinal pode indicar a presença ou não de voz ativa ou inativa. O ruído ambiente tem normalmente uma variância bastante baixa, diferentemente do sinal ativo de voz, sendo possível desta forma fazer a distinção. Este algoritmo possui recomendação, na mesma referência (PRASAD, 2002), de emprego para sistemas executados em ambientes com baixa relação sinal ruído.

$$VAR^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \left(x_3 - \bar{x}\right)^2 + ... + \left(x_n - \bar{x}\right)^2}{(n-1)} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{(n-1)} \qquad ...(19)$$

Na equação 19, os termos x_n são as amostras distintas do sinal de voz, $\bar{x}$ é a média das amostras e n é o número de amostras.

Para a análise espectral existem diferentes formas de serem realizadas avaliações dos sinais de voz para fins de detecção de atividade de voz. Uma destas formas é a avaliação da energia por sub-bandas. Isto ocorre de forma que as faixas de maior concentração das freqüências da voz humana, no caso até 4 kHz, são avaliadas de maneira separada de modo a garantir que as principais componentes da voz estão de forma ativa em um determinado pacote.

2.3.6 Transformação do domínio tempo para freqüência

Como se pode perceber pelos itens anteriores deste trabalho, algumas das técnicas para a detecção do silêncio e sua supressão são executadas no domínio da freqüência. Partindo do fato de que o sistema recebe um sinal de voz a ser processado no domínio do tempo, precisa-se então transformá-lo para o domínio freqüência de forma a executar tais manipulações.

Para a transformação de domínio, no caso, do tempo para a freqüência, são muitos os algoritmos disponíveis para este tipo de tarefa. O que há de se considerar são as questões acerca dos benefícios que o algoritmo oferece ao sistema, no caso de uso do mesmo, e/ou impossibilidades para a aplicação que possam vir a ser causadas pelo uso de uma técnica de transformação de domínio não apropriada. Neste contexto, o que se busca é velocidade de transformação de um domínio para outro e forma de representação do sinal no domínio freqüência.

No que tange a velocidade de processamento, pode-se considerar todas as questões já citadas quanto à necessidade de se gerar o mínimo atraso possível em uma comunicação via VoIP.

Já no que se refere à forma de representação do sinal transformado para o domínio freqüência, refere-se à necessidade de um sinal discreto para a aplicação dos algoritmos.

Considerando estes dois pontos, velocidade do algoritmo e forma do sinal, para esta dissertação, citam-se dois algoritmos de transformação do domínio tempo para o domínio freqüência, a Transformada Discreta do Cosseno (DCT) e a Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*).

A DCT (Equação 20 e Equação 21) (GONZALES, 1993) é uma transformada que possui muitas aplicações para a eletrônica, de filtros de áudio à compressão de vídeo. Ela é a base do padrão JPEG (JPEG - *Joint Photographic Experts Group*) de compressão de imagens (MELLO, 2003). A DCT transforma a informação do domínio espacial ou temporal para o domínio freqüência, sobre o qual fica mais adequada a aplicação de algumas ferramentas.

$$C(0) = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} f(x) \qquad ...(20)$$

$$C(u) = \sqrt{\frac{2}{N}} \sum_{x=0}^{N-1} f(x) \cos \frac{(2x+1)u\pi}{2N} \qquad ...(21)$$

A utilização da DCT para sistemas de transmissão de voz em tempo real sobre redes IP, é sugerida por Prasad (2002), devido ao fato da mesma, conforme o autor, possuir baixa complexidade computacional. Outra característica válida é a facilidade de compreensão e uso desta transformada por se trabalhar apenas com valores reais.

Já a FFT é hoje o método mais citado para sistemas que necessitem de processamento em tempo real. A analogia feita por Agyei-Kodie (2003) com relação á FFT e a Tranformada Discreta de Fourier (DFT), método ou algoritmo antecessor a FFT, é de que seria o equivalente a comparação entre a velocidade de um jato e uma simples caminhada de um humano. Algo em torno de centenas de vezes mais rápido.

Para se chegar ao algoritmo da FFT, primeiro é necessário observar a DFT, sua antecessora. A equação 22 define a DFT:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{\frac{-j2\pi}{N} kn} \qquad ...(22)$$

A partir da definição da DFT pela equação 22, pode-se definir a FFT. A primeira operação a ser realizada é a separação da equação 22 em duas, partes pares e ímpares como apresentado na equação 23. Importante frisar que, para a separação da expressão em partes iguais, *N* terá que ser expresso em potência de *2*. Caso isso não ocorra e *x(n)* tenha um número qualquer de valores diferente de 2^n, o restante será preenchido com zeros até o múltiplo mais próximo (CORRÊA, 1996).

$$X(k) = \sum_{npar} x(n) \cdot e^{\frac{-j2\pi}{N} kn} + \sum_{nimpar} x(n) \cdot e^{\frac{-j2\pi}{N} kn} \qquad ...(23)$$

É necessário fazer a seguinte substituição no algoritmo:

$$n = 2r \ : \ n \ par$$
$$n = 2r + 1 \ ; \ n \ ímpar$$

de forma que a equação 23 possa ser representada pela equação 24, e ainda de forma mais simplificada, pela equação 25.

$$X(k) = \sum_{r=0}^{(N/2)-1} x(2r) \cdot e^{\frac{-j2\pi k 2r}{N}} + \sum_{r=0}^{(N/2)-1} x(2r+1) \cdot e^{\frac{-j2\pi k(2r+1)}{N}} \quad ...(24)$$

$$X(k) = G(k) + H(k) \cdot e^{-j\frac{2\pi k}{N}} \quad ...(25)$$

Nota-se, portanto que restam apenas duas parcelas independentes, mas periódicas em relação à *k*. Assim, serão calculadas duas DFT´s distintas entre si, mas aproveitando-se a simetria ocasionada por *k*. Este tipo de processo garante a economia de $Nlog_2N$ multiplicações, fazendo assim com que haja um aumento da velocidade do processamento (CORRÊA, 1996).

2.3.7 Determinação do limiar de silêncio

A escolha do limiar (*threshold*) de silêncio do ambiente é ponto crucial para a construção de um detector de voz (AGYEI-KODIE, 2003). Uma adequada escolha do limiar de silêncio é a indicação de um eficiente detector. Já o contrário, quando no caso de ocorrerem incorretas detecções de segmentos que não seriam voz, ou que seriam voz, mas não foram detectados, surtirá efeito direto na qualidade do sinal de voz percebida pelo ouvinte. A tendência para este caso é uma voz deteriorada, sem uma qualidade mínima para o bom entendimento da mensagem (YAMADA, 2000).

O limiar de silêncio é aplicado, baseado em algum parâmetro anteriormente determinado, para dividir o sinal de voz em períodos de voz ativa e voz inativa. Este limiar pode ser fixo (Figura 19) ou ter um valor variável, ou melhor dizendo, adaptativo (RENEVEY, 2001) (TANYER, 2000). Segundo cita Benyassine (1997), o ruído ambiente pode mudar consideravelmente entre diferentes conversações ou gravações, bem como a duração da conversa, sendo desde uma sala silenciosa até o ruído das ruas ou de um carro em movimento.

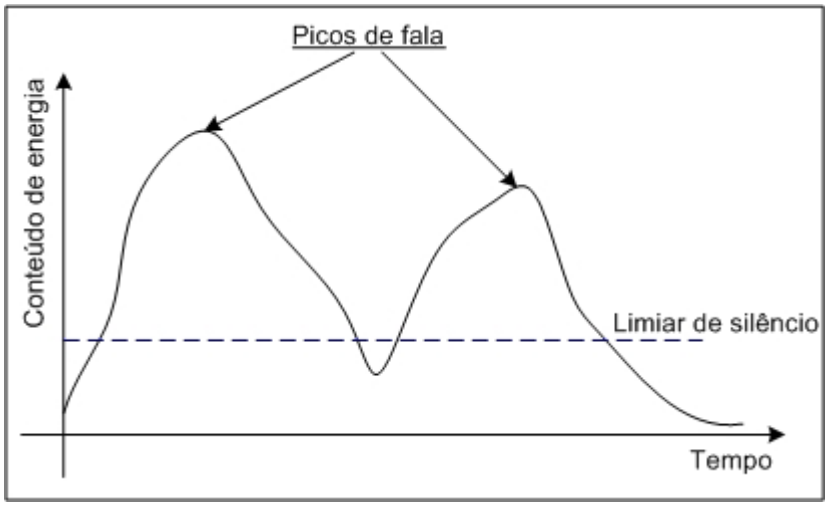


Figura 19 - Representação do limiar de silêncio fixo em relação à energia do sinal de fala
Fonte: O autor (2005).

Para o caso do limiar de silêncio com valor fixo, um valor pré-determinado, ou pelo usuário ou pelo sistema, fica sendo o limite entre o que é considerada voz ativa ou inativa em uma conversação. Sendo os algoritmos de VAD, em grande parte, baseados no cálculo da energia do sinal de voz, a determinação do limiar de silêncio inicial pode ser feita através da utilização da energia de amostras que contenham apenas ruído ambiente. Para a obtenção da energia do sinal de ruído podem ser usados dois métodos: por treinamento e ou amostragem inicial (PRASAD, 2002) (RENEVEY, 2001) (SANGWAN, 2002a) (SANGWAN, 2002b).

O primeiro método consiste no treinamento do algoritmo de VAD, de forma que ele calcule a energia de pacotes que contenham apenas sons de ruído ambiente. O valor de energia encontrado para estes pacotes será o utilizado como limiar de silêncio para o restante dos pacotes no caso de um limiar fixo, já no caso de um limiar adaptável, este será apenas o valor inicial. Para ambos os casos, um outro método assume que os primeiros 200 ms de uma comunicação via um sistema de VoIP, são apenas ruído ambiente, de maneira que a energia encontrada nestes primeiros 200 ms será utilizado da mesma forma que no método anterior.

Já para o caso do limiar de silêncio adaptativo, além de utilizar os métodos anteriores para mensurar o ruído ambiente, o mesmo é o mais indicado quando da utilização de algoritmos de VAD em ambientes de ruído não estacionário, onde há a necessidade do valor do limiar de silêncio ser variável ao longo do tempo. O valor deste limiar é usualmente obtido a partir de segmentos das conversações que foram considerados como sendo voz inativa (TANYER, 2000).

O que tem se verificado (BENYASSINE, 1997) (PRASAD, 2002) (RAMIREZ, 2002) (SANGWAN, 2002a) (SANGWAN, 2002b) (TANYER, 1998) (ZHANG, 2002) é que a forma mais eficiente para a determinação do valor do limiar de silêncio, especialmente em

ambientes com grande variação do ruído, é quando da utilização de algoritmos adaptativos. Os procedimentos adaptativos consistem principalmente de dois passos (ZHANG, 2002): decisão e adaptação. A decisão, no caso de algoritmos de VAD baseados no cálculo da energia, consiste na comparação entre a energia do sinal de voz e a energia do limiar de silêncio. Já a adaptação do algoritmo pode ser feita por diversas formas, algumas destas formas a serem abordadas no terceiro capítulo deste trabalho, junto às técnicas propostas para serem implementadas.

2.3.8 Técnica de recobrimento

Recobrimento é uma técnica para evitar o efeito de corte súbito (*clipping*) da voz, quando da utilização de técnicas de detecção e supressão de silêncio. Esta técnica funciona como uma pequena ligação entre trechos da fala, especialmente consoantes que seriam suprimidas. Dentro do "tempo" de recobrimento, mesmo que haja um pacote de voz considerado silêncio, este será considerado parte do último trecho de fala, e se dentro deste mesmo "tempo" de recobrimento um pacote ativo de voz é detectado, o "tempo" de recobrimento é renovado (JIANG, 2000). Esse tempo de recobrimento pode ser observado na Figura 20.

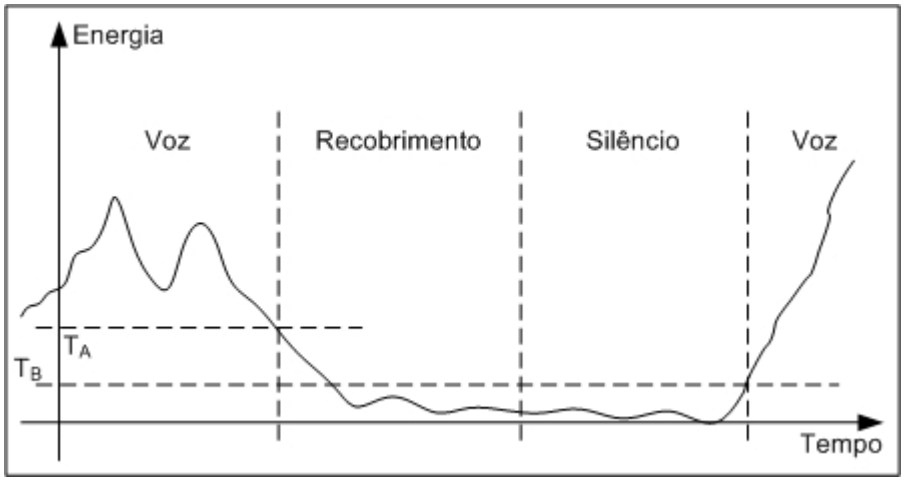


Figura 20 - Períodos de voz ativa, inativa com recobrimento e inativa
Fonte: Shenoi (1995).

Jiang (2000) cita que o "tempo" de recobrimento pode ser determinado por um limitado número de pacotes que no total irão perfazer um determinado tempo. Por exemplo, para o caso de um "tempo" de recobrimento com três pacotes de voz, sendo os pacotes de 20 ms, tem-se um tempo de 60 ms de tempo de recobrimento.

2.4 AVALIAÇÃO QUALIDADE DO ÁUDIO

Vários fatores podem influenciar a qualidade da voz. Esses problemas são bem conhecidos e comuns em redes de voz tradicionais. Essas redes são projetadas de forma a minimizar esses problemas e torná-los imperceptíveis para os usuários. Entre os fatores causadores de perda de qualidade podemos citar o atraso, *jitter*, eco, e ocorrência de erros. Para combater esses problemas diversas técnicas podem ser utilizadas tais como: utilização de supressores ou canceladores de eco, utilização de redundância e algoritmos de correção de erros, etc (SANTOS, 2006).

2.4.1 Medidas da Qualidade de Voz

A maioria dos codificadores de voz tenta oferecer a melhor qualidade possível dentro dos limites impostos pela taxa de transmissão, entretanto vários fatores podem influenciar na qualidade final percebida pelo usuário. A medida da qualidade de voz em uma rede de telefonia pode ser realizada por métodos subjetivos, baseados na avaliação feita por um conjunto de ouvintes, e objetivos baseados em uma série de medidas sobre o sinal ou parâmetros da rede. Para medir o nível de qualidade, medidas objetivas não são confiáveis para um novo codificador. Para a utilização de uma ferramenta de medição automatizada, é necessário ter conhecimento dos valores de qualidade já sabidos e medidos através de métodos subjetivos. As ferramentas utilizadas para medição de qualidade em redes tradicionais também não oferecem precisão quando utilizadas para medir a qualidade em redes VoIP, pois as propriedades do meio de transmissão são diferentes nas duas redes, por exemplo as degradações que numa rede TDM (TDM – *Time Division Multiplexing)* são medidas em termos de erros de bit, em uma rede IP são medidas em termos de perda de pacotes, além de outros novos fatores tais como o atraso e a variação do atraso causado pela utilização de filas (HERSENG, 2005) (SANTOS, 2006).

Medidas subjetivas são desse modo indispensáveis, pois uma medição através da percepção do ouvido humano é o meio mais preciso. Entretanto, medidas subjetivas precisas exigem um grande esforço para sua realização, pois devem seguir várias normas, como as que seguem, por exemplo:

- Assegurar que o número total de ouvintes é suficiente para um resultado estatisticamente confiável;
- Assegurar que a percepção auditiva dos ouvintes é normal;
- Instruir corretamente os ouvintes sobre a metodologia dos testes;
- Assegurar que o material utilizado é diversificado;
- Assegurar que os testes são realizados em diferentes línguas;
- Assegurar que todas as condições de uso do codificador foram testadas;
- Escolha adequada das condições em que são realizados os testes.

Estes testes são especificados nas recomendações P.800 (ITU, 1996a) e P830 (ITU, 1996f) do ITU-T, e alguns dos métodos são descritos na sequência.

Já medidas objetivas são baseadas em modelos que emulam características perceptivas via modelamento matemático (KAHRS, 1998). Em geral são utilizadas técnicas de característica intrusiva quando da medição da qualidade em sistemas de telecomunicações. Neste tipo de medição, um sinal de referência, não degradado pelo sistema, precisa ser injetado na rede e depois, via uma comparação do sinal de saída degradado com o sinal de entrada intacto, é emitida uma nota. Neste sentido, são apresentadas recomendações do ITU-T e mais parâmetros de análise do sinal na sequência deste capítulo, sendo a recomendação mais citada atualmente a P.862 (ITU, 2001).

2.4.2 Medição Subjetiva da Qualidade

Para codecs de taxas mais baixas, entre 4 kbit/s e 32 kbit/s, o teste ACR (ACR - *Absolute Category Rate*) (ITU, 1996a) é o mais utilizado método de medição subjetiva. Nos testes ACR, os ouvintes são questionados para que classifiquem a qualidade absoluta do sinal de voz, sem saber qual é a referência de áudio que está sendo utilizada. Este método produz o *Mean Opinion Score* (MOS) (ITU, 1996a), que é uma escala utilizada para medir a qualidade, mostrada na Tabela 2.

O MOS (ITU, 1996a) é um julgamento absoluto sem referências, mas para que haja uma coerência e calibração entre testes sucessivos, é necessária alguma referência. Para isso, um sinal de áudio de referência é inserido entre os sinais que estão sendo julgados pelos ouvintes. Freqüentemente é utilizada a um sinal modulado de referência (MNRU - *Modulated*

Noise Reference Unit), que é um equipamento que simula a degradação de voz e nível de ruído equivalente ao produzido pela codificação PCM.

Tabela 2 - Escala de classificação do MOS

Qualidade da fala	Escala
Excelente	5
Boa	4
Fraca	3
Pobre	2
Ruim	1

Fonte: ITU (1996a)

Além do ACR (ITU, 1996a), também são usados a categorização por taxa de degradação (DCR - *Degradation Category Rating*) e a categorização pela taxa de comparação (CCR - *Comparison Category Rating*). O método DCR é utilizado quando sinais de boa qualidade estão sendo comparados. Este método produz um nível de opinião médio de degradação (DMOS - *Degradation Mean Opinion Score*). A metodologia é semelhante ao ACR, exceto pelo fato do sinal de referência ser conhecida pelos ouvintes, e apresentada em primeiro lugar. O CCR é similar ao DCR, mas a ordem do sinal de referência e do codificador avaliado é escolhida aleatoriamente. O resultado é o nível médio de comparação por opinião (CMOS - *Comparison Mean Opinion Score*) (SANTOS, 2006).

Tabela 3 - Nível de classificação MOS para codificadores de voz

Codec	Tipo de codec	Tamanho do frame (ms)	Atraso total (ms)	Taxa de transmissão (kbit/s)	MOS
G.711	PCM	0,125	0,125	64	4.2
G.721	ADPCM	0,125	0,125	32	4.0
G.726	ADPCM	0,125	0,125	16/24/32/40	4,0 (32 kbit/s)
G.728	LD-CELP	0,625	5	16	4.0
G.729A	CS-ACELP	10	15	8	3.7
G.723	ACELP	30	37.5	6.3/5.3	3.9/3,7

Fonte: Ohrtman (2004), Santos (2006).

Para sistemas de comunicação interativos, especialmente VoIP, testes de conversação também são bastante úteis, pois procuram reproduzir as mesmas condições do serviço que é prestado aos usuários. As degradações introduzidas pelo eco e atrasos, e que não estão presentes nos testes MOS (ITU, 1996a), podem ser então levadas em consideração. Os

resultados também são dados em uma escala de 1 a 5, e produzem o MOSc (MOSc - *Mean Conversation-Opinion Score*) (ITU, 1996a). Estes testes apresentam grande dificuldade de realização, e a consistência e repetibilidade são difíceis de serem obtidas, devido a todo o cenário necessário, em termos de pessoas, ambiente e estrutura de rede. A Tabela 3 mostra os valores de MOS (ITU, 1996a) para diferentes codificadores de voz.

2.4.3 Medição Objetiva da Qualidade

Conforme Zha (2005) os métodos para medição subjetiva são dispendiosos e demandam tempo, sendo assim necessárias outras formas de avaliação que possam ser utilizadas mais amplamente, através da medição das características físicas dos terminais e redes. A medição objetiva permite aos provedores de serviços de Internet (ISP – *Internet Service Providers*) um rápido provisionamento, ou redirecionamento, de redes ou serviços de voz. A mensuração objetiva é o único meio de se avaliar a qualidade da voz *on-line* em tempo real, com fins de monitoramento e controle da qualidade da rede e da aplicação.

Algoritmos para a medição objetiva da qualidade vocal podem ser divididos em dois tipos: *single-ended* e *double-ended* (Figura 21). *Double-ended* são algoritmos que para a avaliação da qualidade do sistema a ser avaliado precisam como referência do sinal de entrada original (sinal não degradado) e do sinal de saída (sinal supostamente degradado). Algoritmos s*ingle-ended* podem ser utilizados para a "passividade" de acompanhamento, ou seja, o que se chama de não intrusivo a conexão de voz. Para esses algoritmos basta passar para ele o sinal de saída do sistema, supostamente degradado para que ele possa avaliar o mesmo. Já os algoritmos *double-ended* são os chamdos intrusivos, pois um sinal de voz de referência precisa ser injetado no meio de transmissão a ponto que no receptor o mesmo possa ser avaliado pela comparação da entrada com a saída (ZHA, 2005).

As metodologias de medição objetiva de qualidade podem ser classificadas em vários grupos a partir do ponto de vista do objetivo, procedimento de medida, informação de entrada e MOS. As metodologias que exploram aos parâmetros de qualidade da rede e terminais, e produzem uma estimativa do MOS conversacional (MOSc - *Mean conversation-Opinion Score*) (ITU, 1996a), podem ser classificadas como modelos de opinião. As metodologias que necessitam de um sinal de voz como entrada e produzem uma estimativa do MOS, podem ser classificadas como *speech-layer objective models*. Já aquelas que exploram as características

do pacote IP e produzem uma estimativa do ouvinte MOS podem ser chamadas de *packet-layer objetive models*.

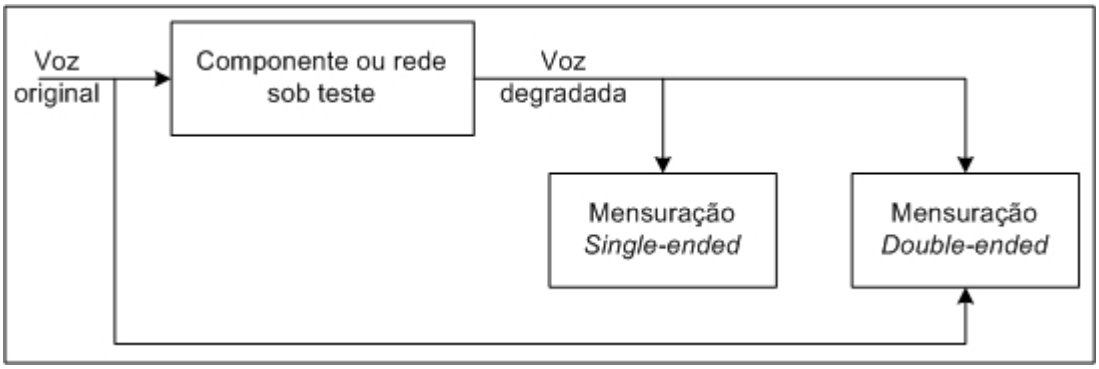


Figura 21 - Mensuração da qualidade da voz via métodos intrusivo e não intrusivo
Fonte: Zha (2005).

2.4.4 Modelos Objetivos *Speech Layer*

O estudo dos modelos objetivos do tipo *speech-layer* começaram com o uso da relação sinal-ruído (SNR) como meio de avaliar os sinais de voz codificados em PCM. Na segunda metade dos anos 80, vários modelos objetivos que exploravam a distorção espectral em vez da distorção da forma de onda foram propostos como métodos objetivos de qualidade mais aplicáveis para a avaliação dos codificadores de baixa taxa de compressão. Entretanto, devido à falta de precisão na estimação, nenhum deles foi padronizado como uma recomendação do ITU-T. Mais tarde, um modelo baseado na distorção espectral Bark, ofereceu uma precisão adequada, e formou a base da recomendação P.861 (PSQM - *Perceptual Speech Quality Mesure*) (ITU, 1998).

Entretanto, devido às características de perda de pacotes das redes IP, esta aproximação não é apropriada para as medidas em VoIP. Um algoritmo mais sofisticado foi criado, e padronizado na recomendação P.862 (PESQ - *Perceptual Evaluation of Speech Quality*) (ITU, 2001).

a) PSQM

O PSQM (ITU, 1998) é um método para estimar a qualidade de codificadores de voz que faz parte da recomendação ITU-T P.861, que foi posteriormente substituída pela recomendação P.862 (ITU, 2001). Ele foi desenvolvido pela empresa de telecomunicações holandesa KPN em 1997. Neste algoritmo, a medida da qualidade é feita através de um

modelo psico-acústico que procura reproduzir a percepção do ouvido humano. A Figura 22 mostra um diagrama da filosofia utilizada no desenvolvimento do PSQM onde um modelo da percepção auditiva humana é construído e é feita uma comparação entre os sinais codificados.

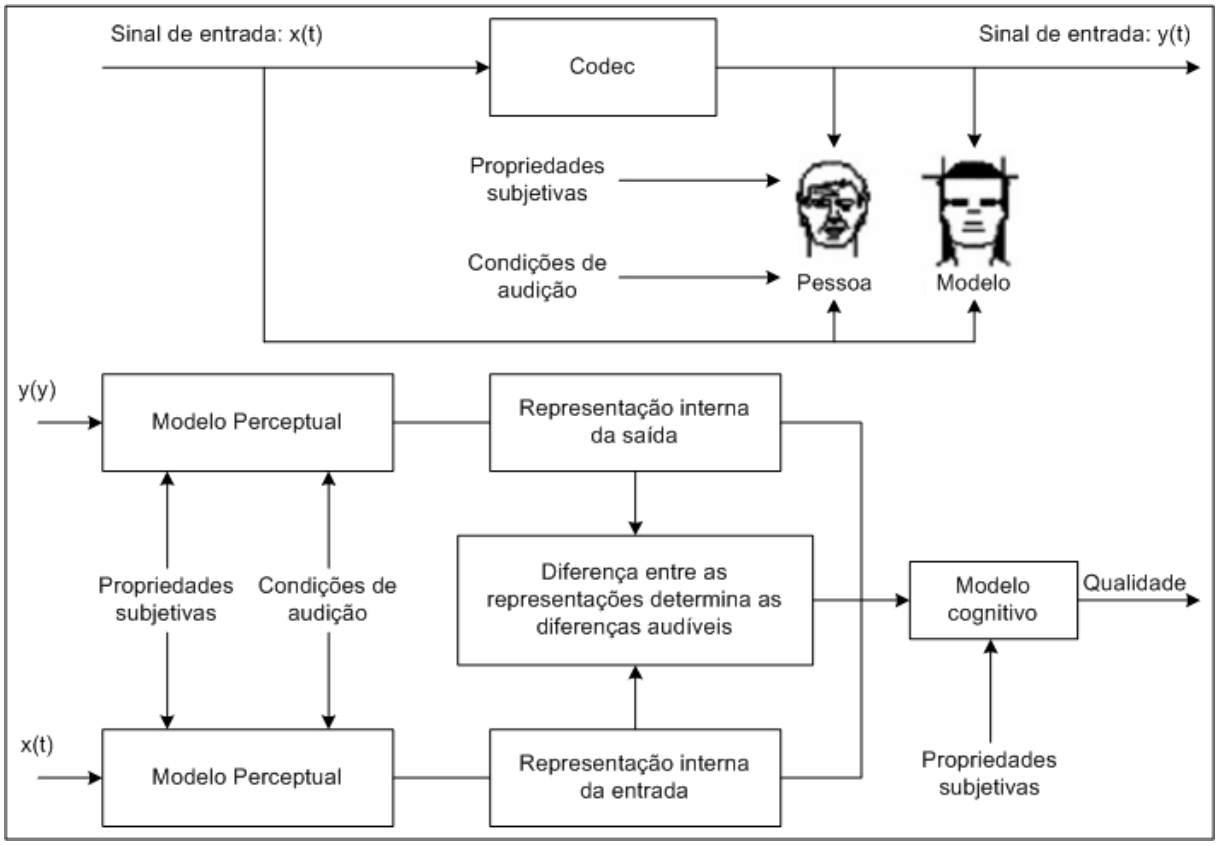


Figura 22 - Modelo funcional do PSQM
Fonte: ITU (1998).

O sinal é convertido para o domínio psico-acústico através de três operações:

- Mapeamento tempo-frequência - é implementado através de uma FFT com uma janela de Hanning;
- Mudança na escala de freqüência - é realizada através da conversão da escala em Hertz para uma escala em Bark[11];
- Mudança na escala de amplitude - é feita uma compressão de amplitude de acordo com a sensibilidade auditiva (*loudness*).

Através da comparação entre o sinal original e o distorcido, ou degradado, é obtido um fator chamado de perturbação de ruído. A distorção é computada para cada quadro com tamanho de 256 amostras e 50% de sobreposição. O resultado mostra a perturbação de ruído

[11] A escala de Bark é uma escala psico-acústica proposta por Eberhard Zwicker em 1961. Foi nomeada depois que Heinrich Barkhausen propôs a primeira forma subjetiva de mensuração do áudio. A escala cobre de 1 a 24, correspondendo as 24 primeiras bandas audíveis, sendo as mesmas (in Hz) 20, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500 (ZWICKER, 1961).

com relação ao tempo e à freqüência, e a média desses valores é relacionada com a qualidade do codificador. O resultado obtido pelo algoritmo é denominado valor PSQM. Ele indica o grau de degradação através de uma escala que varia de 0 a 6.5, sendo que 0 corresponde a um sinal sem degradação, e 6.5 a degradação máxima. O valor PSQM pode ser convertido para a escala MOS de acordo com a equação 26 a seguir:

$$MOS = \frac{4}{1 + e^{0,66 \cdot PSQM - 2,2}} + 1 \qquad ...(26)$$

O PSQM+ foi proposto como uma melhora ao algoritmo PSQM (ITU, 1998). Ele leva em consideração o valor PSQM e a potência do sinal, de forma que variações no volume do sinal são contabilizadas no valor final. O maior problema do PSQM é o alinhamento do sinal original com o sinal medido, pois o atraso da rede é desconhecido e pode sofrer variações, mas para que o algoritmo compare os dois sinais, é necessário o perfeito alinhamento. Desta forma, o algoritmo também não considera a degradação devida ao atraso e a variação do atraso.

b) MNB

O MNB (MNB - *Measuring Normalizing Blocks*) (ITU, 1998) foi desenvolvido pelo instituto de ciências de telecomunicações do departamento de comércio americano em 1997. Ele modela o julgamento humano da qualidade através da análise no tempo e na freqüência. Essas duas análises são combinadas de forma que seja obtido um valor chamado *Auditory Distance* (AD - *Auditory Distance*), que mede a "distância", em termos de qualidade, entre o sinal de entrada e o sinal de saída (REDDING, 2001).

A medida é feita através do alinhamento entre o sinal original e o sinal a ser medido. São então eliminados os componentes DC, e posteriormente é estimada a potência média dos dois sinais, para que sejam normalizados. O passo seguinte é a transformação do sinal para o domínio da freqüência utilizando uma FFT com janela Hamming de 128 amostras (16 ms) e sobreposição de 50 %. Os quadros resultantes são comparados entre os dois sinais, de forma que são eliminados os quadros cujas diferenças estão abaixo de um determinado valor, e também os quadros com componentes de freqüência com potência zero. Os quadros que não foram eliminados são transformados de acordo com uma escala de sonoridade, e comparados no domínio do tempo e da freqüência. São então obtidos valores que representam as diferenças entre os sinais para diversos intervalos de freqüências e é feita uma combinação desses valores, que representa o valor medido (SANTOS, 2006).

c) PAMS

Em 1998, a British Telecom desenvolveu o PAMS (PAMS - *Perceptual Analysis Measurement System*). Este algoritmo também faz uma análise no tempo e na freqüência e inclui o sincronismo no tempo, que era um problema que não havia sido resolvido nos algoritmos apresentados na recomendação P.861 (ITU, 1998).

d) PESQ

O algoritmo PESQ (PESQ - *Perceptual Evaluation of Speech Quality*) (ITU, 2001) é um padrão do ITU-T para medida de qualidade descrito na recomendação P.862 (ITU, 2001). Ele foi criado em conjunto pela KPN e Britsh Telecom através da combinação do PSQM+ e do PAMS, e foi desenvolvido especificamente para poder ser utilizado em redes como VoIP e ISDN[12] (*Integrated Services Digital Network*), já que os algoritmos da recomendação P.861 (ITU, 1998) não eram eficientes para tratar dos problemas específicos destas redes. O PESQ (ITU, 2001) apresenta a medida da qualidade diretamente na escala MOS (ITU, 1996a).

O algoritmo PESQ (ITU, 2001) basicamente segue os mesmos passos usados no PSQM (ITU, 1998), entretanto são introduzidas modificações de forma a melhorar sua performance. Primeiramente há uma compensação do ganho nos sinais original e degradado, de forma que apresentem o mesmo nível de potência, essa calibração é feita tanto no domínio do tempo quanto na freqüência. Após essa etapa, é realizada uma filtragem nos sinais, de forma que o sinal a ser analisado tenha a mesma característica de um sinal ouvido após atravessar uma rede telefônica, os sinais são então alinhados no tempo de forma que é definido o intervalo em que vai ser feita a análise. A conversão para o domínio da freqüência é feita utilizando uma janela Hanning sobre quadros de 32 ms e 50% de sobreposição. A análise é feita sobre as diferenças no domínio da freqüência na escala Bark, após serem realizadas as compensações de *loudness*[13], e de variações de ganho.

Devido ao seu melhor desempenho em relação aos outros modelos apresentados para a medição da qualidade dos sinais na faixa de voz, o PESQ tornou-se uma nova recomendação P.862, substituindo à anterior P861 (SANTOS, 2006).

---

[12] ISDN é a Rede Digital de Serviços Integrados, padrão usado em linhas telefônicas digitais que evolui das redes telefônicas convencionais (YOUNG, 2006).

[13] *Loudness* é a forma de percepção psicológica do som correlacionado com a sua intensidade em termos de amplitude. A percepção do *loudness* está relacionada com a pressão ocasionada por deterinado som e a sua duração.

2.4.5 Modelos Objetivos *Packet-Layer*

A ITU está também tentando padronizar uma metodologia de medição de qualidade baseada apenas nas informações do pacote IP. Esta metodologia está sendo provisoriamente chamada de P.VTQ. Este procedimento inicia com a estimação dos parâmetros de qualidade intermediários como a taxa de perda de pacotes, padrão da perda e variação do atraso obtidos a partir do cabeçalho RTP (IETF, 2003a) e das informações do RTCP (IETF, 2003b). O MOS é então estimado. Estes parâmetros intermediários formam um subconjunto de entidades definidas no RTCP-XR (XR - *extended report*) (IETF, 2003c) proposto pelo IETF. A qualidade é estimada através da aplicação de um segundo estágio do algoritmo P-VTQ, utilizando as informações do RTCP-XR (SANTOS, 2006).

3 TÉCNICAS DE DETECÇÃO E SUPRESSÃO DE SILÊNCIO

Como visto anteriormente, podem existir vários algoritmos de VAD que combinados formam as técnicas de detecção e supressão de silêncio. Esses algoritmos, não apenas podem formar as técnicas, mas também podem melhorar os respectivos desempenhos, como cita Rabiner (1978), por exemplo. Para tanto, neste capítulo três da dissertação, é levantada a hipótese de que a combinação de alguns dos algoritmos apresentados ao longo da revisão teórica possam contribuir de forma significativa para a formação das técnicas, visando à compressão do sinal de voz baseado na detecção dos períodos de silêncio, focando a aplicação em redes de pacotes.

Assim sendo, neste capítulo três, são propostas e descritas sete técnicas de detecção e supressão de silêncio que derivam da combinação dos algoritmos anteriormente apresentados e como já citado anteriormente. Também é apresentada uma proposta de avaliação das técnicas implementadas via aplicação de recomendações do ITU destinadas à mensuração da qualidade da voz proporcionada ao ouvinte do sistema, baseando-se em mensuração subjetiva e objetiva.

Como visto anteriormente, a utilização de algoritmos que buscam mensurar parâmetros da voz no domínio da freqüência são tão ou mais comuns que no domínio do tempo. De forma a caracterizar melhor as técnicas e facilitar o entendimento das mesmas, é feita uma separação em dois tópicos: técnicas de detecção e supressão de silêncio no domínio do tempo e técnicas de detecção e supressão de silêncio no domínio da freqüência.

Para as técnicas de detecção e supressão de silêncio aqui propostas, com exceção feita à técnica STD, com limiar de silêncio fixo, propõe-se adotar um método de mensuração de ruído ambiente para a determinação dinâmica dos limiares de silêncio. Este método é aplicado por Prasad (2002), Sangwan (2002a)(2002b) e consiste em assumir que os 200 ms iniciais de comunicação não possuem segmentos de fala relevantes à conversação, e sim, apenas ruído ambiente. Isto possibilita que o sistema mensure um limiar de silêncio apropriado ao ambiente com base no próprio ruído do local onde se encontra o locutor. Isto ocorre via verificação da intensidade de energia dos 200 ms iniciais que foram considerados silêncio.

3.1 TÉCNICAS DE DETECÇÃO E SUPRESSÃO DE SILÊNCIO NO DOMÍNIO DO TEMPO

Para as técnicas de detecção e supressão de silêncio que operam no domínio do tempo e aqui apresentadas, é usualmente utilizada uma constante de "segurança[14]" *k*, conforme citam (PRASAD, 2002) (SANGWAN, 2002a). Esta constante é aplicada no momento em que se faz a decisão entre a energia do pacote de voz e o limiar de silêncio, conforme a equação 27 O valor de *k* deve ser maior que 1 de forma que permita tornar o sistema menos "instável" com a adaptação do limiar de silêncio.

$$Energia\ Atual > k \cdot Energia\ Ruído \quad ...(27)$$

3.1.1 Detector de limiar fixo (STD)

A técnica de supressão de silêncio com limar fixo, é a mais antiga, pode-se dizer assim, e comum (RENEVEY, 2001) (TANYER, 2000) dentre todas as técnicas de detecção e supressão de silêncio. Para esta técnica, a energia de cada pacote de voz é monitorado e depois comparado com um limiar de silêncio fixo (TANYER, 2000). No entanto, este tipo de técnica é bastante sensível a variações de ruído. Em condições de ruído elevado, esta técnica tende a obter resultados pobres na detecção de voz em pacotes com baixo conteúdo de energia, como por exemplo, para o caso de segmentos não vozeados de fala, sendo estes mesmo segmentos mascarados pelo ruído (TANYER, 2000).

Esta técnica acaba por ser apresentada neste trabalho apenas por ter ser o ponto de partida para a construção das demais técnicas. O fato desta técnica não possuir um algoritmo adaptativo associado, praticamente descarta qualquer possibilidade de emprego da mesma em um sistema de comunicação em tempo real. Caso esta técnica STD fosse disponibilizada em uma ferramenta de comunicação de VoIP, o usuário do sistema necessitaria de prévio conhecimento sobre detecção e supressão de silêncio de forma que pudesse fazer um ajuste manual do limiar de silêncio. Por este motivo esta técnica não será testada junto com as demais.

3.1.2 Detector linear baseado na energia (LED – *Linear Energy Based Detector)*

A LED é uma técnica adaptativa que tem como base principal de cálculo a energia dos pacotes de voz, mesmo princípio da STD, mas com diferença no ajuste do limiar. A técnica LED possui uma característica adaptativa dada pela equação 28, onde a *Energia Atual* é a energia do pacote de voz atual, *Energia Anterior* é a energia do último pacote de voz considerado silêncio, antes do atual. O parâmetro *p* é o índice que determina o passo de adaptação, podendo variar de 0 a 1. Nesta técnica o parâmetro *p* é determinado pelo usuário conforme Prasad (2002) e Sangwan (2002a).

$$Limiar\ de\ Silêncio = (1 - p) \cdot Energia\ Anterior + p \cdot Energia\ Atual \quad ...(28)$$

Posteriormente nesse trabalho é discutida a especificação do parâmetro *p*, que aqui pode-se chamar de passo de adaptação, pois o mesmo acaba por determinar a velocidade de adaptação do limiar de silêncio diante do nível de energia de ruído apresentado. Mas considerando os objetivos de aplicação e usabilidade futura das técnicas, inclusive por outros usuários e por vezes leigos, o ajuste manual do passo de adaptação não se vislumbra como algo prático. Para tanto, testes são sugeridos para se chegar a um valor mais adequado, baseado no melhor desempenho do algoritmo quanto ao nível de supressão e qualidade do áudio. Também foram considerados parâmetros utilizados por outros autores para tanto como os utilizados por Prasad (2002) e Sangwan (2002b).

A equação 28 determina o valor do limiar de silêncio e o quão rápido o mesmo irá se adaptar às variações de ruído ambiente. O algoritmo de adaptação (Equação 28) faz a soma do percentual do pacote atual com do último pacote considerado inativo para fins de atualização do limiar de silêncio (OPPENHEIM, 1975).

3.1.3 Detector linear adaptativo baseado na energia (ALED - *Adaptative Linear Energy-Based Detector*)

O ALED é uma técnica que utiliza os mesmos algoritmos do LED, com exceção à determinação do índice *p*. Antes determinado pelo usuário, agora passa a ser variável, dado

[14] Segurança é colocada entre aspas, por na verdade se tratar de uma forma de se ter um controle praticamente

pelo resultado da comparação da razão da variância do pacote de voz atual com a variância[15] do último pacote processado considerado silêncio, e tem o seu valor obtido a partir da Tabela 4 (PRASAD, 2002) (SANGWAN, 2002a).

Tabela 4 – Valores do passo de adaptação *p* dependentes de *τNEW /τOLD*

Variância	*p*
$\frac{\tau_{NEW}}{\tau_{OLD}} \geq 1,25$	0,25
$1,25 \geq \frac{\tau_{NEW}}{\tau_{OLD}} \geq 1,10$	0,20
$1,10 \geq \frac{\tau_{NEW}}{\tau_{OLD}} \geq 1,00$	0,15
$1,00 \geq \frac{\tau_{NEW}}{\tau_{OLD}}$	0,10

Fonte: Sangwan (2002a).

3.1.4 Detector de fracos fricativos (WFD – *Weak Fricatives Detector*)

O WFD se diferencia das técnicas anteriormente apresentadas por não ser exclusivamente baseada na energia do sinal de voz. Esta técnica utiliza um artifício estatístico, após a verificação de energia, para a detecção de fonemas que possuam uma reduzida energia através da contagem do número de cruzamentos do zero que o sinal apresenta em um determinado tempo. Para o caso deste trabalho este tempo será de 20 ms, com um sinal sendo amostrado a 8 kHz (RABINER, 1978) (SANGWAN, 2002a).

Prasad (2002) considera que para 10 ms de sinal amostrado, seriam considerados ativos os pacotes que apresentassem entre 5 e 15 cruzamentos. Mas ao mesmo tempo, Prasad (2002), mesmo mostrando resultados em seus experimentos, não faz citação a referência que justifique o porquê da escolha desta faixa, entre 5 e 15 cruzamentos, para determinação dos períodos de voz ativa. Essa referência seria importante porque, para esta dissertação, são pretende-se analisar pacotes de áudio de 20 ms, o que acaba por tornar necessária uma adaptação na faixa de valores referente ao número de cruzamentos. Como solução paliativa, propõe-se que sejam considerados de dez a trinta cruzamentos para pacotes de 20 ms. A

---

manual sobre o limiar de silêncio e a consequente economia de banda da rede IP proporcionada pela técnica.

adaptação desta faixa de valores precisa ser verificada em termos de desempenho e também, como a definição desta faixa de valores deve vir a ter uma origem estatística, o que não foi possível determinar, assim a simples multiplicação dos valores por dois pode ser questionada, a ponto de ser necessária uma investigação maior sobre o mesmo.

3.2 TÉCNICAS DE DETECÇÃO E SUPRESSÃO DE SILÊNCIO NO DOMÍNIO DA FREQÜÊNCIA

Nesta seção, são descritas três técnicas de detecção e supressão de silêncio no domínio da freqüência. Para as técnicas aqui apresentadas, é utilizada a DCT como algoritmo de transformação do domínio tempo para o domínio frequência.

Para a definição da DCT como algoritmo a ser utilizado nas técnicas propostas no domínio freqüência, levou-se em consideração o fato de sua representação ser baseada somente por números reais. Isto porque a multiplicação de números complexos em um processador digital requer quatro multiplicações e duas adições em ponto fixo (ou flutuante), devido à própria natureza das operações complexas. A quantidade de memória requerida e o número de operações necessárias fazem com que a transformada de Fourier não seja a mais apropriada, no que diz respeito à eficiência computacional, para transformar seqüências reais para o domínio das freqüências, conforme Nascimento (2004).

3.2.1 Detector linear de energia por sub-banda (LSED – *Linear Sub-Band Energy Detector*)

A LSED tem sua tomada de decisão para a determinação do pacote ativo ou não, baseado na comparação da energia de quatro sub-bandas, de 1 kHz cada, do sinal de voz com um determinado limiar de silêncio (Figura 23). Esta operação ocorre após a transformação de domínio via DCT (SANGWAN, 2002a).

---

[15] Na teoria da probabilidade e na estatística, a variância de uma variável aleatória é uma medida da sua dispersão estatística, indicando quão longe em geral os seus valores se encontram do valor esperado.

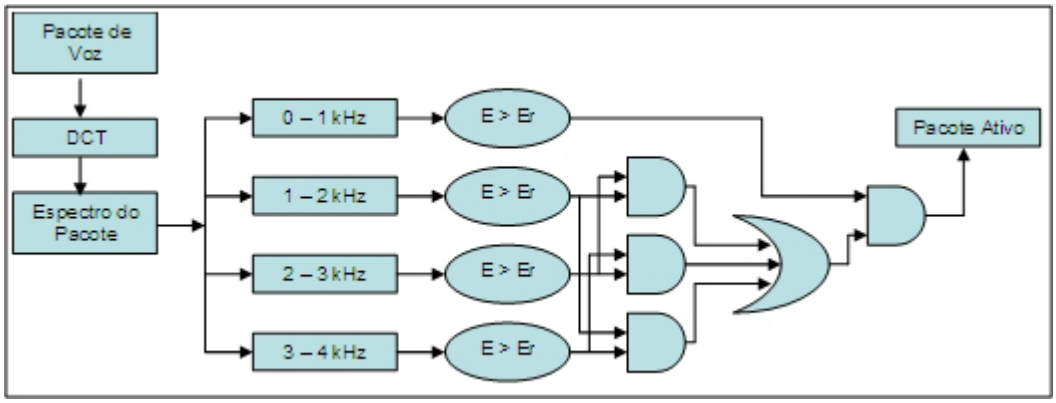


Figura 23 - Avaliação do sinal de voz por sub-bandas.
Fonte: Sangwan (2002a).

Como apresentado na Figura 23, a decisão sobre se foi detectada voz ou não pelo algoritmo, é a resultante da combinação lógica que prioriza a banda de 1 kHz, faixa de maior concentração das freqüências da voz (RABINER, 1978), e também, faixa de maior concentração das componentes em frequência da DCT.

Outra característica desta técnica de detecção e supressão de silêncio é a adaptabilidade ao ruído ambiente, fazendo uso da mesma Equação 28 (LED) para a atualização do limiar de silêncio. O algoritmo de adaptação funciona associado a cada sub-banda separadamente.

3.2.2 Detector de atenuação espectral (SFD – *Spectral Flatness Detector*)

A técnica SFD é utilizada com a finalidade de garantir que pacotes de voz com baixo SNR não sejam suprimidos. Seu funcionamento é baseado na comparação da variância do pacote atual de voz (*VarVoz*) com a variância do ruído ambiente (*VarSilêncio*) (Equação 29). Sua adaptabilidade ao ruído é dada pela Equação 30. Já o passo de adaptação *p* da mesma, que determina a velocidade de adaptação do algoritmo, é determinado pelo usuário do sistema. O valor de *p* será maior que zero e menor que um.

$$VarVoz \geq VarSilêncio \quad ...(29)$$

$$VarSilêncio = (1 - p) \cdot VarSilêncio + p \cdot VarVoz \quad ...(30)$$

3.2.3 CVAD (CVAD – *Compreensive VAD*)

CVAD é uma combinação de três algoritmos de VAD. A Figura 24 apresenta o fluxograma de funcionamento desta técnica. Como se observa, a técnica utiliza a avaliação da energia do pacote por sub-banda no domínio da freqüência com o uso da DCT. Caso seja verificada uma energia maior que o limiar de silêncio o pacote de voz é considerado ativo (mesmo funcionamento da LSED). Caso o pacote seja considerado inativo, o pacote de voz é encaminhado para uma segunda análise, sendo então verificada a quantidade de cruzamentos do zero ocorrida durante os 20 ms. Para este estágio da técnica, se o pacote de voz for considerado inativo o mesmo não é transmitido, mas caso seja considerado ativo, ele ainda passa pela avaliação da atenuação espectral do sinal. Se a variância do sinal no domínio freqüência for baixa, ocorre a indicação de inexistência de conteúdo significativo para a conversação em curso e o pacote é considerado inativo, sendo desta forma não transmitido, caso contrário, é voz ativa (SANGWAN, 2002a).

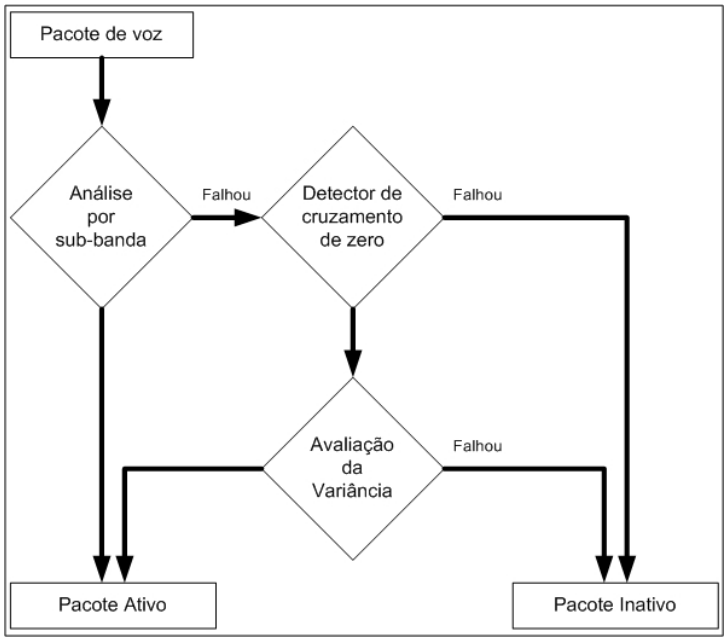


Figura 24 - Fluxo de decisão do CVAD.
Fonte: Sangwan (2002a)

4 PROPOSTA

Dentro ainda do contexto, esse trabalho se propõe a validar e sugerir melhorias para as implementações propostas neste capítulo. Para tanto, inicialmente é necessário se definir os parâmetros a serem avaliados. Assim, como elementos fundamentais na validação da implementação das técnicas de compressão de voz baseadas na detecção de períodos de silêncio, o que se vislumbra é um alto percentual de supressão de silêncio com um mínimo nível de degradação da qualidade do sinal da voz proporcionada. Ou seja, quanto mais próxima estiver a amostra "degrada", amostra essa que já tenha sido as técnicas de compressão, da amostra original, melhor a qualidade e melhor o desempenho da técnica, em termos de qualidade do sinal de voz. Lembrando que ainda precisa-se avaliar o percentual de compressão, ainda não considerado aqui.

Mas voltando a sequência, tendo sido realizada a revisão bibliográfica, e definidas as técnicas a serem implementadas, ainda faltava definir a metodologia para a condução dos testes de avaliação. O que se pode observar é a inexistência de uma metodologia de avaliação da voz, dentro da bibliografia consultada, específica para a avaliação isolada de técnicas de detecção e supressão de silêncio. Desta forma, decidiu-se então se propor uma adaptação para os objetivos do trabalho. O que se apresentava mais conveniente, e também por ter sido o método apresentado em outros trabalhos relacionados (PRASAD, 2002) (SANGWAN, 2002a) (SANGWAN, 2002b) e tantos outros, é a recomendação P.800 (MOS) (ITU 1996a) do ITU-T. A recomendação P.800 (ITU 1996a) baseia-se em um método de avaliação subjetiva da qualidade do sinal de voz em sistemas de telecomunicações. Essa recomendação é baseada no método de classificação por categoria absoluta (ACR) (ITU 1996a) para obter-se a pontuação média de opinião (MOS), como já detalhado anteriormente. A descrição dos parâmetros recomendados pela P.800 para este tipo de teste estão descritos no Anexo A deste trabalho.

Mesmo a recomendação P.800 (ITU, 1996a) sendo aplicada em diversos trabalhos, por exemplo os de Zha (2005), Cai (2004), Sangwan (2002a) e muitos outros, o que se sabe, é que essa mesma recomendação é bastante custosa, especialmente no público necessário para avaliar o áudio aplicado as técnicas. E também pensando na execução de mais rotinas de testes, que não avaliem apenas o resultado final do áudio, mas também o processo de construção das técnicas, com testes intermediários. Neste sentido verifica-se a possibilidade de uso da recomendação P.862 (PESQ) (ITU-T, 2001). A recomendação P.862 (ITU-T, 2001) baseia-se em resultados objetivos obtidos a partir de algoritmos matemáticos que analisam de

forma comparativa os sinais de áudio de entrada e saída do sistema. A partir disso, o PESQ (ITU-T, 2001) emite uma nota referente ao nível de entendimento do sinal, que aplicado aqui, pode mensurar o nível de degradação proporcionada pelo pelas técnicas ao áudio.

Outro ponto favorável a escolha de se usar o PESQ (ITU-T, 2001), conforme afirma Fernandez (2003) e é apresentado na própria recomendação P.862 (ITU-T, 2001), é o fato de a mesma possui condições de avaliar o *clipping* da voz, por exemplo.

Dentro deste cenário, a Figura 25 apresenta um diagrama de blocos com o sistema a ser construído, e os sinais, em pontos intermediários, a serem testados. Neste cenário o que se vislumbra é poder fazer um comparativo do processo evolutivo de implementação das técnicas e verificar o quanto as técnicas de recobrimento e a adição de ruído de conforto podem contribuir para a melhoria da qualidade do sinal de voz na saída do sistema. Destaca-se que em cada ponto do sistema avaliado em termos de qualidade da voz, irá se verificar também o percentual de supressão de silêncio proporcionado pela aplicação de cada técnica.

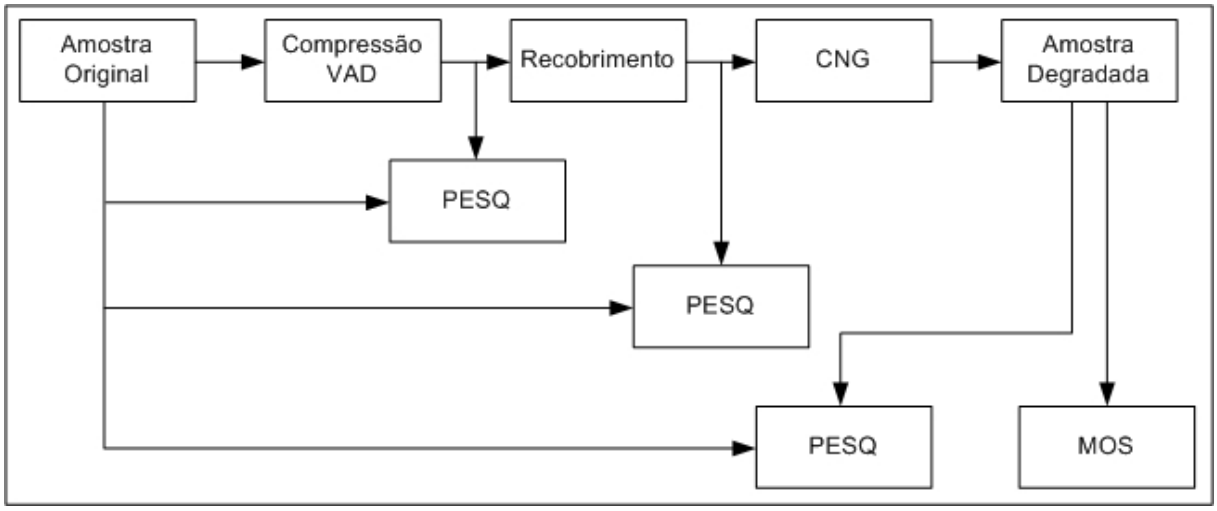


Figura 25 - Diagrama de blocos do sistema a ser implementado e sinais a serem avaliados pelo PESQ e pelo MOS
Fonte: O autor (2009).

A verificação do percentual de supressão de silêncio proporcionado possibilita que seja feita uma estimativa de redução de informação injetada na rede de pacotes considerando ainda não apenas o conteúdo de voz, mas o *overhead* dos cabeçalhos dos protocolos aplicados na transmissão da mídia.

Outro ponto que vale ressaltar, e que já foi citado na revisão deste trabalho, é quando da avaliação da qualidade do sinal da voz, existir a necessidade de uma avaliação subjetiva de referência quando da avaliação de codificadores, antes mesmo de uma avaliação objetiva. Aqui isso só reforça o procedimento adotado. Mesmo as técnicas de detecção de períodos de silêncio não serem padrões de codificação, as mesmas podem ser enquadradas na necessidade

de uma avaliação de referência. Nesse sentido a nota MOS certamente será de grande valia para o trabalho.

5 PROJETO, IMPLEMENTAÇÃO E VALIDAÇÃO DOS ALGORITMOS DE COMPRESSÃO DE VOZ

Para esta dissertação de mestrado, quanto a sua construção, a metodologia adotada foi a de sequenciar o texto de acordo com o desenvolvimento do trabalho. Assim, podem-se discriminar algumas partes, que podem ser ditas como fazes do desenvolvimento: levantamento bibliográfico, definição da proposta, implementação dos algoritmos matemáticos, testes de validação e avaliação, e construção do manuscrito.

Na sequência são detalhados os métodos aplicados para a construção do trabalho, considerando o detalhamento de cada uma das fazes citadas acima.

5.1 LEVANTAMENTO BIBLOGRÁFICO E DEFINIÇÃO DA PROPOSTA

Com a definição do objetivo do trabalho de avaliar técnicas de compressão de voz baseadas na detecção de períodos de silêncio, deu-se início à realização da primeira parte, ou primeira tarefa, a revisão bibliográfica. Foi feito um levantamento das bibliografias relacionadas e que pudessem contribuir com este trabalho. Os pontos relevantes deste material para o trabalho são apresentados no capítulo dois desta dissertação, bem como junto à definição da proposta no capítulo três, fundamentando a mesma. Por fim, as referências se fazem presente junto aos resultados e conclusões quando da análise e comparação dos resultados, buscando fundamentar os mesmos.

Ainda quanto ao levantamento bibliográfico, o que pode-se enfatizar é o foco tido nesse trabalho de pesquisa, onde quatro grandes campos foram mais abrangidos na pesquisa:

a) Uma linha de bibliografias associada a fundamentação matemática e teorias associadas a processamento digital de sinais. Para tanto foram utilizados materiais de autores reconhecidos no meio como Rabiner & Schafer (1978), Oppenhein (1975), Strum (1988), Gonzales (1993), Smith (1997), Vasegui (2000), Haykin (2004), entre outros;

b) uma segunda linha, em especial *papers* e outros trabalhos científicos, associados a aplicabilidade direta dos algoritmos implementados em sistemas não apenas em

sistemas de voz sobre IP, mas também para outros tipos de transmissão de áudio, bem como o tratamento de mídia estática quando degrada pelo tempo. Entre os autores que podem ser destacados aqui estão todos os trabalhos do ITU e IETF (*International Engineering Task Force*), além de Hersent (2002)(2005), Schulzrine (1996) e junto com Rosemberg (1998), Davis (2002), Sangwan (2002a) (2002b), Prasad (2002), Balbinot (2002) (2004), entre outros;

c) Uma terceira linha de material verificado quando a construção, aplicação e comparação de meios para a avaliação da qualidade de sinal de áudio quando do uso em meios de comunicação. Dentre os autores aqui trabalhados citam-se como mais relevantes, os trabalhos do ITU, Kagrs (1998), Barbedo (2001) e (2004), Fernandes (2003) Zha (2004), Santos (2006), Becvar (2007) e ainda alguns dos já citados no item *b*;

d) Por fim, ainda há algumas bibliografias utilizadas de forma mais específica na definição de termos técnicas citados e complementação de definições.

Com base nas referências, foi definida uma proposta de trabalho. A mesma proposta é apresentada no capítulo três. Para tanto foram consideradas algumas bibliografias como referência para determinação nas hipóteses de resultados esperados para a implementação das técnicas. Nesse sentido está se falando nos trabalhos de (citando apenas os autores) Sangwan, Rabiner, Oppenheim, Prasad, Tanyer, Davis e Shenoi. Esses trabalhos sugerem claramente a combinação de técnicas para detecção de períodos de silêncio, com inclusive sugestão de algoritmos direcionados ao assunto. Quanto ao método de avaliação, levou em consideração os diversos trabalhos realizados na área e o maior número de citações das recomendações P.800 (ITU, 1996a) e P.862 (ITU, 2001), sendo portanto as de maior campo de aplicação.

Trabalho fundamentado, e proposta elaborada, partiu-se para a implementação, conforme segue o detalhamento.

5.2 ETAPAS DA MODELAGEM

Em um momento inicial, a definição de quais técnicas de compressão que viriam a ser construídas, foi tomada pensando no interesse de se usar as técnicas de compressão de voz,

como classes bases de supressão de silêncio, a serem implementadas, para um comunicador de voz sobre IP, o chamado Locutus, Figura 26, do qual o autor deste trabalho participou do desenvolvimento.

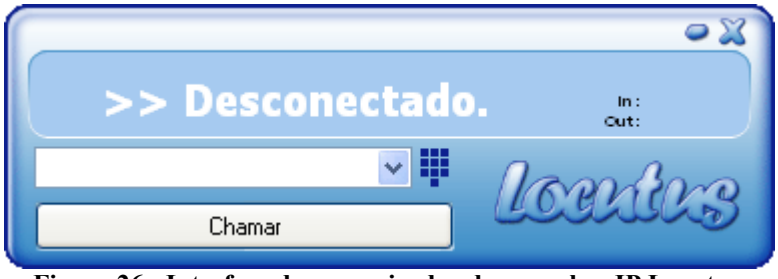


Figura 26 - Interface do comunicador de voz sobre IP Locutus.
Fonte: O autor 2005.

O desenvolvimento do comunicador Locutus estava associado ao objetivo principal do projeto RVoIP (*Robust Voice over IP system*) desenvolvido junto a Faculdade de Engenharia da PUCRS. O mesmo projeto era financiado pelo CNPq. Assim, o seu desenvolvimento se deu até o encerramento do aporte financeiro. Como consequência, não foram possíveis testes funcionais, usando a estrutura da rede de pacotes, no que diz respeito a integração do VAD ao comunicador, mas o que não impediu a continuidade do desenvolvimento do componente de supressão de silêncio, dado vasto campo de aplicação e as tecnologias envolvidas.

Depois de terminada a fase de revisão bibliográfica e definidas as técnicas para a detecção e supressão de silêncio em períodos de fala a serem implementadas, foi dado início à construção das mesmas.

Mas antes disso foi feita à escolha da ferramenta de programação a ser utilizada. A opção inicial feita foi pelo Delphi na versão 7.0 a ser programado em linguagem Pascal.

A escolha do Delphi como ambiente de desenvolvimento se deu considerando que o comunicador Locutus teve o seu desenvolvimento, feito no mesmo. A justificativa para o uso desse ambiente seria pelo fato da aplicação ser voltada para o Windows, e em dado momento o Delphi ter sido a ferramenta mais adequada. Isto porque se considerou o Delphi um ambiente de desenvolvimento de aplicações, orientado a objeto, que permite o desenvolvimento de aplicações baseadas no Microsoft Windows com o mínimo de codificação. O Delphi também oferece ferramentas de desenvolvimento, tais como *templates* de aplicações e *forms*, que permitem criar e testar o protótipo de aplicações (MADUREIRA, 2003).

Como dito, inicialmente o Delphi foi considerado o ambiente mais interessante, mas o que se observa, e algo que não é de agora, e sempre questionada, que é uma descontinuidade no desenvolvimento do ambiente de desenvolvimento. Isso acabou por fazer com que partes

finais do desenvolvimento do trabalho tenham sido já implementas em Linguagem C++, a qual sem necessidade de fundamentação, atinge um universo muito maior de usuários e aplicações.

Considerando alguns conceitos de engenharia de software, a implementação das classes foi construída com uma estrutura orientada a objetos. Para tanto, foi realizada a modelagem das classes utilizando o Model Maker. O Model Maker é um ambiente destinado a modelagem de dados, em específico para implementações em Delphi. O mesmo tem por base toda a diagramação UML[16] (UML – *Unified Model Language*), e possibilita, a partir dos diagramas de classe, a geração automática da estruturação do código. Assim, foram construídos os diagramas de casos de uso, diagramas de estado e diagramas de classes. Esses diagramas são apresentados no Apêndice A deste trabalho, para as técnicas de detecção e supressão de silêncio, e ainda, no Apêndice B o diagrama de classes das técnicas de geração de ruído de conforto.

5.3 IMPLEMENTAÇÕES

Quanto a construção dos algoritmos e associação dos mesmos para constituir técnicas mais elaboradas para a detecção de silêncio em sistemas de voz sobre IP, a implementação das mesmas foi realizada em duas partes. Na primeira parte das implementações foram feitas a construção dos algoritmos e os testes que podem-se dizer "estáticos". Isso porque os testes funcionais realizados com áudio foram todos via o uso de arquivos de voz gravados em formato *wave*[17]. Já na segunda parte, houve a construção das classes de compressão e os testes funcionais com o uso de áudio em tempo real, com captura, processamento e reprodução instantânea sem armazenamento.

[16] UML é uma linguagem para especificar, visualizar, construir e documentar os artefatos de sistemas de software, bem como para modelar negócios e outros sistemas que não sejam de software (LARMAN, 2004).
[17] *Wave* é o formato padrão de arquivo de áudio da Microsoft e da IBM.

5.3.1 *Wave Silence Suppressor*

Para a primeira fase das implementações foi construída uma ferramenta denominada para o trabalho como *Wave Silence Suppressor*.

A *Wave Silence Suppressor* (Figura 27) recebeu esse nome sugestivo pelo fato do software trabalhar executando simulações dos algoritmos de detecção de silêncio sobre amostras de voz no formato de arquivos *wave* de forma que apresenta, através de uma interface visual, trechos das amostras que foram suprimidos, bem como o percentual de supressão de silêncio atingido sobre cada amostra.

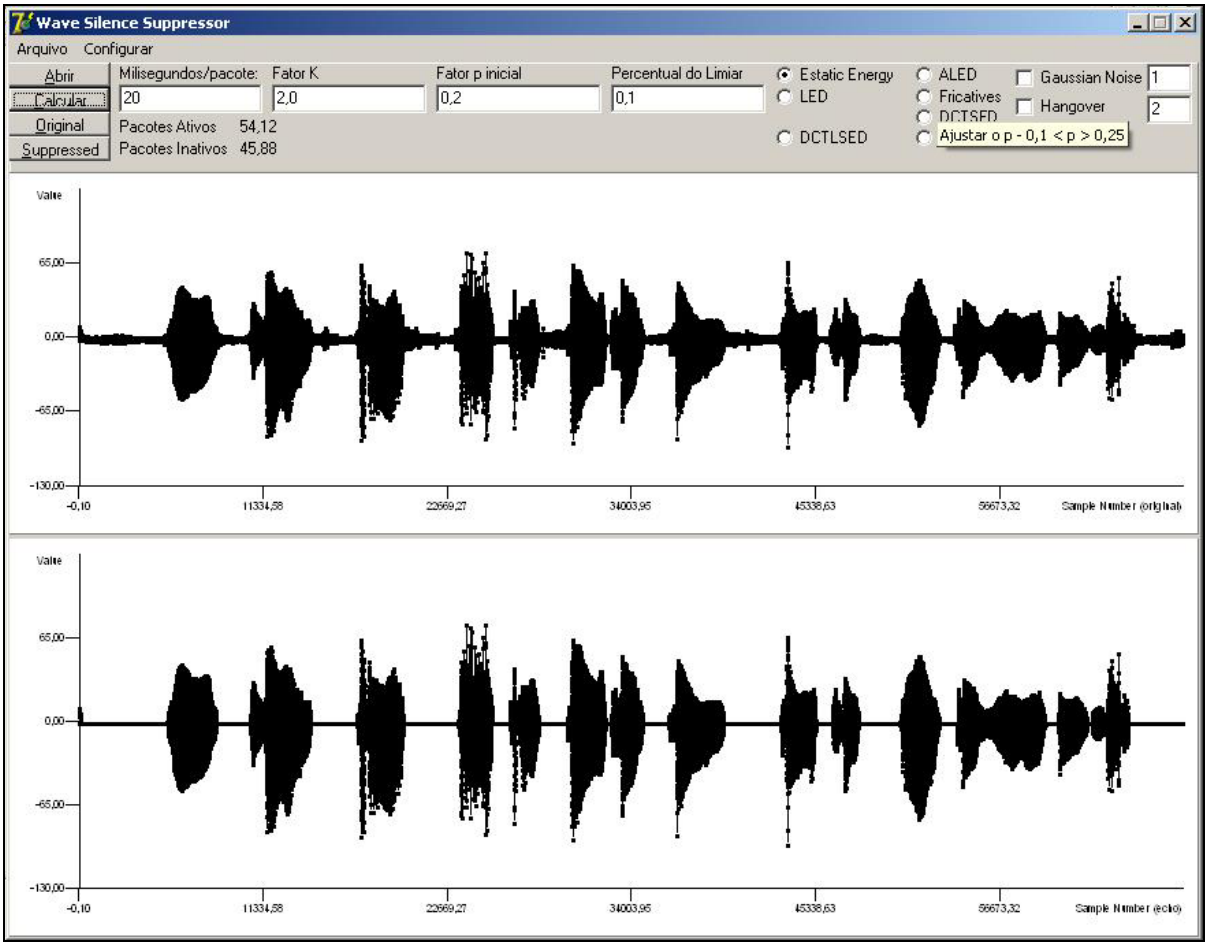


Figura 27 - *Wave Silence Suppression*.
Fonte: O autor 2009.

Na Figura 27 é apresentado o visual do processamento de uma amostra de voz que constitui uma a contagem de um a dez por uma mulher. Na representação mais acima é apresentada a amostra original, com a presença do ruído ambiente. Na representação mais abaixo, já se tem a mesma amostra, com os períodos de silêncio suprimidos. Nesta segunda

representação não foram aplicados recobrimento e nem ruído de conforto objetivando deixar bem clara, de forma visual, a capacidade de supressão dos algoritmos aplicados.

Por meio desta ferramenta foi possível selecionar a técnica de detecção e supressão de silêncio a ser testada, configurar todos os fatores relevantes à técnica, além da inserção de ruído gerado por um algoritmo de geração de ruído de conforto para fins de melhoria da qualidade subjetiva da voz. Também é possível a aplicação da técnica de recobrimento e fazer a determinação da quantidade de pacotes a ser aplicado, a voz suprimida, por esta técnica.

5.3.2 Classes de supressão de silêncio

A segunda fase de implementações foi a construção das classes bases de supressão de silêncio. Inicialmente foi feita uma modelagem destas classes, utilizando-se uma ferramenta específica para isto, já descrita, e posteriormente a sua construção. A codificação das técnicas foi baseada no que já havia sido feito anteriormente para a *Wave Silence Suppression*.

Depois de feita a modelagem e a codificação das classes, era necessário fazer alguns testes de validação e ajuste das mesmas. Focando agora em testes a serem realizados com amostras de voz capturadas e reproduzidas em tempo real e não mais arquivos de voz tipo *wave*, como no primeiro momento, foi construída a ferramenta, denominada para este trabalho como *Silence Suppression Tester* (Figura 28).

Esta ferramenta utiliza em sua base classes de captura e reprodução de áudio. A estas classes de captura e reprodução foram adicionadas além das classes de detecção e supressão de silêncio, classes de geração de ruído gaussiano, e ainda uma interface para leitura e ajuste de parâmetros.

Essas classes de geração de ruído também fazem parte do processo de desenvolvimento das técnicas de compressão de áudio desenvolvidas. No Apêndice B do trabalho são apresentadas as três classes de geração de ruído branco que foram modeladas e implementadas. Para integração com as ferramentas de teste descritas, foi usado apenas a classe de geração de ruído branco baseado em valores aleatórios com energia média proporcional ao período de silêncio inicial de 200 ms de cada “conversação”.

Como citado na revisão teórica e também como é apresentado nos resultados deste trabalho, a inserção de ruído de conforto, em substituição aos períodos de silêncio suprimidos

do áudio original, podem exercer forte influência na melhoria da qualidade do áudio percebido pelo ouvinte.

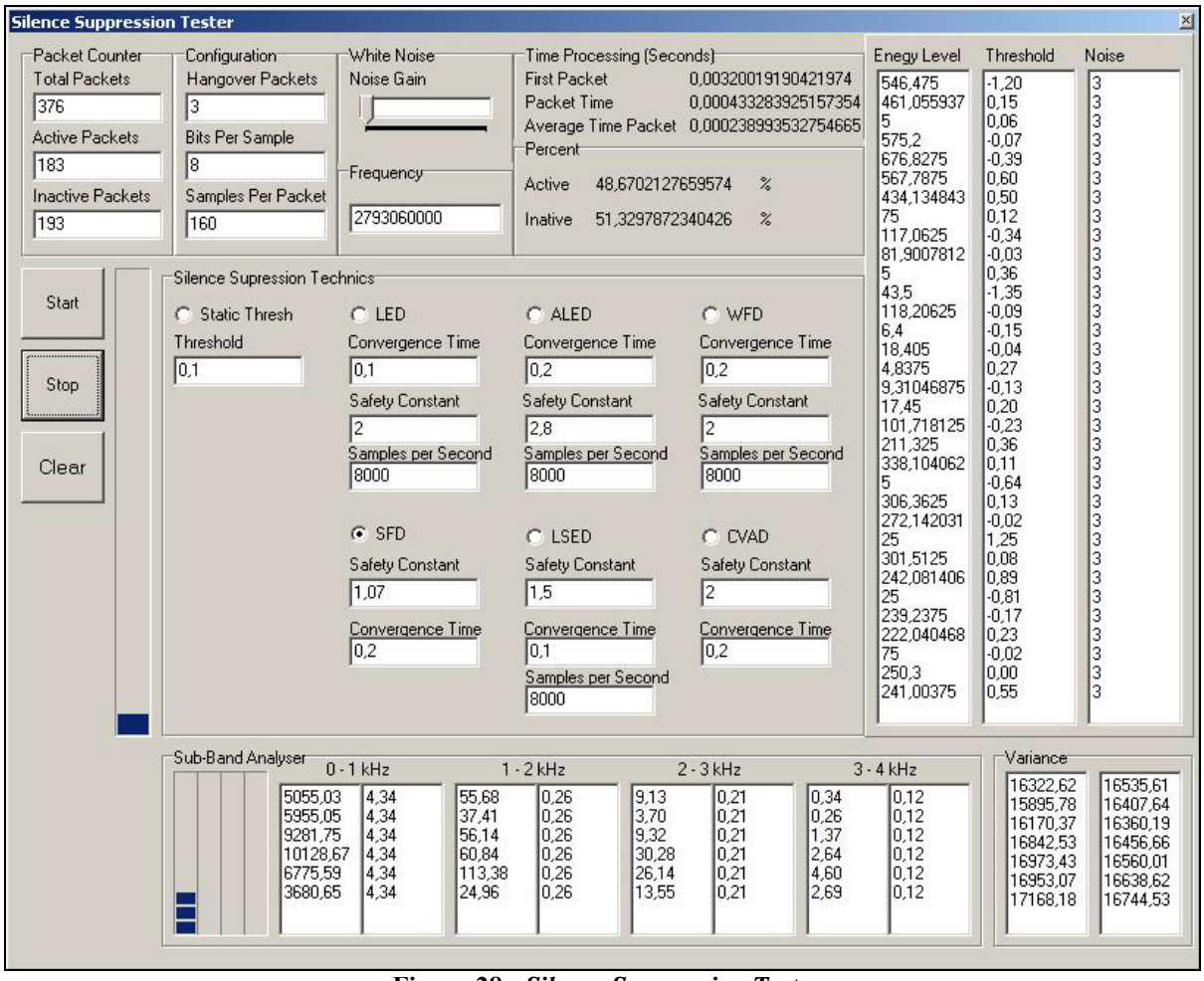


Figura 28 - *Silence Suppression Tester*
Fonte: O autor 2009.

A *Silence Suppression Tester*, por meio de sua interface permite não só fazer ajuste de todos os parâmetros necessários às classes, bem como parâmetros que devem ser genéricos a todo o sistema, como número de bits por amostra e número de amostras por segundo. Também é possível fazer a leitura de parâmetros durante o processamento de cada pacote de áudio em tempo real, tais como o total de pacotes ativos e inativos, o tempo de processamento de cada um e o tempo médio de processamento dos pacotes, além da evolução do limiar de silêncio ao longo do tempo.

Importante salientar que a *Silence Suppression Tester* é uma ferramenta de teste local. A mesma apenas captura o sinal do microfone do computador e o reproduz nas caixas de som ou fones de ouvido do mesmo, não tendo a finalidade de estabelecer de qualquer tipo de comunicação.

5.3.3 Outras implementações

Dentro do contexto de desenvolvimento de componentes para sistemas de voz sobre IP, ao longo do trabalho foram vislumbradas outras possibilidades de implementações, tanto como a construção de outros algoritmos para este tipo de aplicação, bem como melhorias das técnicas já detalhadas neste capítulo. Em parte, isso fica apresentado como trabalhos futuros. Mas dentre os mesmos, houve a implementação de um *buffer* de *dejitter*. Como o mesmo não está diretamente ligado ao objetivo do trabalho, um detalhamento da implementação do mesmo é apresentado no Apêndice C deste trabalho.

5.4 TESTES DE VALIDAÇÃO

Assim como citado anteriormente, existem alguns parâmetros desejáveis para os algoritmos de VAD. Questões como boa regra de decisão do algoritmo entre o que é voz e o que não é, complexidade computacional, perda de qualidade da voz e economia de banda proporcionada são pontos fundamentais e que precisavam de alguma forma ser avaliados.

Para a realização destes testes foram utilizadas as duas ferramentas implementadas. Através da *Silence Suppression Tester* foram realizados os testes de tempo de processamento de cada técnica de detecção e supressão de silêncio e via a *Wave Silence Suppressor* foram realizados os testes para mensuração do percentual de supressão proporcionada por cada técnica. Ficava assim, faltando um último teste, que seria o de verificar o nível de degradação da qualidade do áudio proporcionado pela aplicação das técnicas implementas e também em função do percentual de supressão de silêncio aplicado.

Desta forma, realizada a revisão bibliográfica, a construção das técnicas e definidos os testes finais a serem feitos, faltava ainda definir uma metodologia para a condução dos mesmos, em especial a mensuração da qualidade do sinal de voz. O que se pode observar é a inexistência de uma metodologia de avaliação da voz, dentro da bibliografia consultada, específica para a avaliação isolada[18] de técnicas de detecção e supressão de silêncio. Desta

[18] A expressão "[...]avaliação isolada[...]" especifica o fato de que existem técnicas que avaliam questões que são abordadas junto a supressão de silêncio, mas que não são específicas ao assunto, como é o caso do PESQ que possui condições de avaliar o *clipping* da voz, por exemplo (FERNANDEZ, 03) (ITU, 2001).

forma, decidiu-se então buscar algo que pudesse vir a ser adaptado para os objetivos do trabalho. O que se apresentou mais conveniente, foi a realização de uma adaptação dos dois métodos mais utilizados para avaliação de qualidade de áudio em sistemas de telecomunicações. O primeiro método foi a recomendação P.800 (MOS) (ITU 1996a) do ITU-T, e também por ter sido o método apresentado em outros trabalhos relacionados (PRASAD, 2002) (SANGWAN, 2002a) (SANGWAN, 2002b). A recomendação P.800 (ITU 1996a) determinada a metodologia para uma avaliação subjetiva da qualidade do sinal de voz em sistemas de telecomunicações. Já a recomendação P.862 (ITU, 2001) sugere a metodologia para uma avaliação objetiva, feita de uma forma artificial, baseada em modelos matemáticos, sendo esta recomendação também mundialmente aplicada na avaliação de sistemas de telecomunicações.

Definidas as metodologias e feita a adaptação (essa adaptação é descrita na sequência) para o cenário deste trabalho, utilizou-se da ferramenta *Wave Silence Suppression* para o preparo das amostras de voz para os testes de avaliação da perda de qualidade da voz proporcionado.

Quanto aos testes de tempo de processamento, os mesmos foram realizados mais com o objetivo de otimizar os algoritmos em termos de uso de funções e variáveis que pudessem agilizar o processamento. Esses podem ser considerados testes funcionais que objetivam a melhoria da codificação. Como não houve um aprofundamento nesse sentido, como por exemplo a verificação do número de ciclos de *clock* necessários para cada operação, os mesmos não tem os seus resultados apresentados aqui.

Na sequência são apresentadas as rotinas aplicadas para cada um dos testes de avaliação de qualidade do áudio realizadas, considerando a avaliação subjetiva e a objetiva.

5.4.1 Preparação das amostras

Para a realização dos testes de níveis de compressão de silêncio aplicado as amostras de áudio e de opinião por escuta, foram gravadas quatro sentenças, ou quatro amostras de voz em uma sala silenciosa fechada, não acusticamente isolada e com a utilização de um microfone dinâmico Coby CM-P24 de alta performance. As sentenças foram gravadas no formato *wave*, modulação PCM (*Pulse-Code Modulation*), com freqüência de amostragem de 44,1 kHz e filtro *anti-alising* da placa de som *on-board* do microcomputador. Esse sinal foi

codificado a 16 bits. Para os testes de interesse do trabalho, as amostras foram tanto gravadas, como posteriormente convertidas, com a utilização do gravador de som do Microsoft Windows, para 8 kHz e 8 bits, também se utilizando do filtro *anti-alising* do próprio gravador, tendo sido estes os parâmetros utilizados para todos os testes.

Como material a ser apresentado aos ouvintes (especificados na sequência do trabalho) no caso da avaliação subjetiva e ao PESQ, quando da avaliação objetiva, foram gravadas quatro sentença curtas, de 2 a 3 segundos, escolhidas ao acaso e consideradas de fácil entendimento e ainda sem conexão óbvia entre as mesmas. As frases gravadas para as amostras foram:

Choveu muito neste fim de semana;

Ela precisa esperar na fila;

O banco fechou sua conta;

Guardei o livro na primeira gaveta.

É importante observar que a escolha deste tipo de sentenças que foram gravadas, são baseadas na recomendação P.800 (ITU, 1996a), onde é feita referência ao tipo de amostra a ser utilizada nos tipos de teste ao qual este trabalho se propõe.

Estas amostras de voz foram submetidas às técnicas de detecção e supressão de silêncio via utilização da ferramenta *Wave Silence Suppression* desenvolvida ao longo do trabalho com específico fim. Esta ferramenta possibilitou a realização de ajustes nos algoritmos, em especial nos parâmetros de ajuste e da velocidade de adaptação do limiar de silêncio dos mesmos, diante das variações dos períodos de silêncio ou não, gerando assim, as amostras que seriam utilizadas nos testes.

Para os testes realizados foram geradas um total de 76 amostras de voz com o auxílio da *Wave Silence Suppression*. Do total de amostras:

- 4 amostras originais com as frases listadas acima;
- 4 amostras degradadas, para cada uma das 6 técnicas (total de 24), aplicando aqui apenas as técnicas de compressão;
- 4 amostras degradadas, para cada uma das 6 técnicas (total de 24), aplicando as técnicas de compressão e mais o recobrimento equivalente a 40 ms de áudio. Ou seja, sempre que detectado um pacote de voz inativo após um ativo, o atual, inativo, e o seguinte serão necessariamente ativos. Como explicado anteriormente, isso tem o objetivo de evitar o efeito de *clipping* na voz do locutor;

- 4 amostras degradadas, para cada uma das 6 técnicas (total de 24), aplicando as técnicas de compressão e mais o recobrimento equivalente a 40 ms de áudio, e ainda o ruído de conforto.

É importante salientar que para a formação dos arquivos no formato *wave*, com as amostras já sob a ação das técnicas de compressão, os parâmetros das técnicas, detalhados anteriormente, como a constante de segurança *k*, o fator de convergência *p* do algoritmo adaptativo e o limiar de silêncio inicial, foram mantidos iguais para todas as técnicas. A constante de segurança *k* era igual a 2, o fator de convergência *p* igual a 0,2 e o limiar de silêncio inicial era baseado na média da energia dos primeiros 200 ms de áudio das amostras gravadas.

Para a determinação destes valores, em específico para a constante de segurança e o fator de convergência, foram baseados nos trabalhos de Prasad (2002) e Sangwan (2002b), onde os mesmos aplicam estes mesmos valores. Para verificação da eficiência destes parâmetros, se utilizou como base os testes de nível máximo de supressão de períodos de silêncio, já tendo sido relatado neste capítulo os seus resultados.

É determinante a consideração também, da aplicação do ruído de conforto (ruído branco) dos períodos considerados silêncio pelas técnicas de compressão. O nível de ruído de conforto aplicado aos períodos de silêncio foi proporcional a média do nível de energia dos primeiros 200 ms de áudio das amostras gravadas. Lembrando que no caso da uma ferramenta de voz sobre IP real, o ruído de conforto é produzido no lado do ouvinte, não gerando assim, constante tráfego de dados para o transporte desse ruído na rede. O que pode haver é a transmissão do lado do locutor para o lado do ouvinte de valores correspondentes a energia média de períodos considerados silêncio pela técnica de compressão utilizada. Isso para que o algoritmo gerador de ruído de conforto, do lado do ouvinte, possa ser ajustado automaticamente a ponto de acompanhar as variações de intensidade do ruído do ambiente do locutor.

5.4.2 Avaliação subjetiva

Para a avaliação subjetiva da voz buscou-se seguir a recomendação P.800 (ITU, 1996a) pelo método de classificação por categoria absoluta (ACR – *Absolute Category Rating*) (ITU, 1996a) para obter-se a pontuação média de opinião (MOS – *Mean Opinion*

Score) de ouvintes. A descrição dos parâmetros recomendados pela P.800 (ITU 1996a) para este tipo de teste estão descritos no Anexo A deste trabalho, e os parâmetros utilizados para os testes realizados para este trabalho são descritos juntamente com os resultados obtidos.

Importante salientar que esta recomendação P.800 (ITU 1996a) visa geralmente a avaliação de codificadores de áudio e ou do sistema de telefonia como um todo. Para o caso deste trabalho, os testes foram feitos de forma isolada para as técnicas de detecção e supressão de silêncio aqui descritas e implementadas, visto que não foi verificada nenhuma outra forma de avaliação subjetiva reconhecida que pudesse ser aplicada ou adaptada para o tipo de aplicação em questão.

Para a avaliação das amostras pelo público foi elaborado um guia de instruções (Figura 29) que continha a forma que iria transcorrer a avaliação das sentenças pré-gravadas. Na mesma Figura 29, estavam descritos cinco níveis possíveis com respectivas pontuações, referentes ao nível de esforço necessário para o entendimento das sentenças.

Teste de Avaliação Subjetiva de Qualidade da Voz

Neste experimento você irá ouvir dois grupo de sentenças curtas, através do *headphone* conectado ao PC.

Você deve escutar os dois grupos de duas sentenças e após as mesmas, marcar a pontuação referente a necessidade de esforço empregado para entendê-las.

Esforço necessário para entender o significado das sentenças	**pontos**
Completamente relaxado; sem necessidade de esforço	5
Necessidade de atenção; pequeno esforço	4
Necessidade de esforço moderado	3
Necessidade de esforço considerável	2
Não existe entendimento, mesmo com todo esforço possível	1

Haverá uma pausa entre os grupos de sentenças. Sendo pronunciados ao todo dois grupos neste experimento.

Sexo do ouvinte:
Idade:
Você já participou de algum outro tipo de teste de escuta? (SIM – NÃO)

OBRIDADO PELA SUA AJUDA NESTA EXPERIÊNCIA.

Figura 29 - Guia de instruções para o teste de avaliação subjetiva
Fonte: O autor 2009, adaptado de ITU (1996a).

Para a realização dos testes, noventa pessoas do público em geral foram convidadas a participar. Todos os participantes, no momento da avaliação, encontravam-se sentados e usando fones de ouvido. Assim como no guia de instruções (Figura 29), os mesmos ouviam as quatro sentenças e depois faziam uma marcação na pontuação referente ao esforço médio necessário para entender o significado das quatro sentenças.

Cada participante do público foi submetido às sentenças que estavam apenas sob o efeito de uma técnica de supressão de silêncio. Por exemplo, o primeiro participante só ouviu

sentenças que estavam sob a ação da técnica LED e o segundo, apenas sentenças sob a ação da técnica ALED. De forma que ao final do experimento estavam contabilizadas quinze avaliações por técnica.

Como resultado dos testes, foram contabilizados os valores referentes ao nível de esforço necessário para o entendimento das quatro sentenças apresentadas a um público aleatório total de noventa pessoas, sendo oitenta e dois homens e oito mulheres com média de idade em torno de 21 anos e um mês. Deste total de pessoas, apenas treze já haviam participado de algum tipo de teste de escuta anteriormente.

Por fim, haveriam ainda os testes a serem feitos junto ao comunicador Locutus. Estes testes foram realizados apenas em caráter inicial, sem maiores formalidades quanto a medição dos resultados. Mesmo assim, as observações verificadas são comentadas no capítulo de resultados, já que mesmo sem uma metodologia específica, esses testes levaram a alguns questionamentos relevantes.

5.4.3 Avaliação objetiva

Para a avaliação objetiva aplicou a recomendação P.862 (PESQ) (ITU, 2001) de uma forma adaptada ao contexto do trabalho. O PESQ (ITU, 2001) é geralmente aplicado na avaliação objetiva de sistemas de telecomunicações, fazendo uma avaliação da qualidade do sinal de voz baseado em modelos matemáticos, comparando o sinal original, com o sinal que foi transmitido.

Essa recomendação é direcionada para medidas de sinais compreendidos na faixa de 300 Hz a 3400 Hz, aplicável a sistemas com fala codificada, atraso variável, filtragem, perda de pacotes, corte no tempo e erros de canal (FERNANDES, 2003).

Para o trabalho realizado junto a esta dissertação, a avaliação de qualidade do sinal de voz foi feito de forma isolada. Não houve transmissão pela rede IP do sinal de voz. A avaliação foi feita comparando-se o sinal original com o sinal processado pelas técnicas de detecção e supressão de silêncio.

5.4.4 Correlação dos resultados

A proximidade entre o PESQ e a pontuação subjetiva do MOS, pode ser medida calculando-se o coeficiente de correlação. Normalmente isto é feito com a pontuação média, depois do mapeamento dos valores objetivos para os subjetivos. O coeficiente de correlação *r* é calculado pela fórmula de Person (equação 31) (BECVAR, 2007):

$$r = \frac{\sum\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \cdot \sum\left(y_i - \bar{y}\right)^2}} \qquad ...(31)$$

Para esse trabalho foi adotado na fórmula que, x_i é o valor MOS para a condição *i*, e $\bar{x}$ é a média dos valores MOS x_i. y_i é a pontuação PESQ para a condição *i*, e $\bar{y}$ é a média dos valores PESQ y_i.

Para vinte e dois experimentos de comparação realizados pelo ITU-T entre valores PESQ e MOS, o coeficiente de correlação médio foi de 0,935 (ITU, 2001), o que é um resultado bastante bastante bom, sendo que na correção os valores podem ir de menos um, para o pior resultado mais um para o melhor resultado.

Para esse trabalho, tomou-se como referência o valor do MOS, e com ele foram correlacionados os valores de três medições do PESQ.

Talvez o mais indicado para a verificação da correlação dos sinais, seria o de correlacionar medidas de PESQ e MOS obtidas no mesmo ponto do sistema, como realiza Becvar (2007) em seu trabalho, por exemplo. Mas isso ficou inviabilizado aqui devido ao alto custo de se fazer o levantamento das medidas do MOS não só para as amostras degradadas na saída do sistema, já com a compressão, recobrimento e ruído de conforto, mas sim com medidas intermediárias, avaliando passo a passo a influência de cada técnica empregada. Nesse sentido, fazendo um contraponto ao trabalho de Becvar (2007), onde no mesmo é avaliada apenas uma técnica em diferentes níveis de configuração, mas seguindo a recomendação de outros autores, citados anteriormente na revisão bibliográfica, onde se descreve a importância de uma medida MOS de referência para os testes comparativos usando o PESQ. Aqui são avaliadas seis técnicas (LED, ALED, WFD, LSED, SFD e CVAD), também com diferentes níveis de configuração, o que leva a uma pequena diferença, tendo sido quatro medições MOS feitas por Becvar (2007), e aqui, se todas fossem feitas, seriam

necessárias dezoito medições, sendo que foram feitas seis, uma por técnica, considerando o sinal degradado na saída do sistema. Em termos de público entrevistado, isso significaria 270 pessoas, o que já eleva de forma drástica o custo destes testes.

Os resultados e análise das correlações são apresentados no capítulo seis como segue.

6 RESULTADOS

No capítulo seis desta dissertação de mestrado, serão apresentados e avaliados os resultados obtidos ao longo dos testes propostos e realizados. Também serão detalhados os ajustes empregados nas técnicas de detecção e supressão de silêncio implementadas, bem como os parâmetros utilizados, sempre tendo tido como objetivo a melhora do desempenho das implementações. Quanto a ordem de realização dos testes, utilizando a avaliação subjetiva e a avaliação objetiva, e ainda a preparação das amostras utilizadas, tem seu detalhamento apresentado no capítulo quatro, no item referente aos testes de validação. Por fim são apresentados comparativos de correlação dos resultados com objetivo de verificar a real contribuição de cada técnica na implementação em termos de degradação do áudio e validação das próprias metodologias de teste adotadas quando comparados os resultados finais entre as medidas do PESQ (ITU, 2001) e do MOS (ITU, 1996a).

Considerando os percentuais de compressão obtidos em duas diferentes etapas do processo de avaliação das técnicas, são levantadas estimativas da quantidade de dados que deixa de ser injetada na rede, considerando ainda o *overhead* dos cabeçalhos dos protocolos utilizados para o transporte da mídia.

6.1 ANÁLISE NA SAÍDA DO BLOCO DE COMPRESSÃO

A primeira análise do sistema foi com base nos testes realizados considerando-se o sinal de saída do bloco de compressão. Como elemento de referência em termos do sinal que se está analisando, é apresentada a Figura 30.

Como se observa na Figura 30, a medição realizada em termos de qualidade foi com a utilização do PESQ. Para tanto o mesmo compara o sinal de saída do bloco de compressão com o sinal de entrada do sistema, identificado no bloco amostra original, ou seja, sem degradação. O resultado dessa medição é apresentado no Gráfico 1.

Os resultados apresentados no Gráfico 1 apresentam o valor PESQ para cada uma das seis técnicas de supressão de silêncio. A escala vertical do gráfico vai até 4,5 por ser o valor máximo possível para o PESQ e no eixo horizontal são apresentadas o resultado das seis implementações. Vale ressaltar que valor três, tanto para o PESQ, quanto para o MOS, é o

valor considerado mínimo aceitável para os sistemas sob teste. Apesar do valor PESQ sempre se apresentar em torno de meio ponto abaixo do valor MOS.

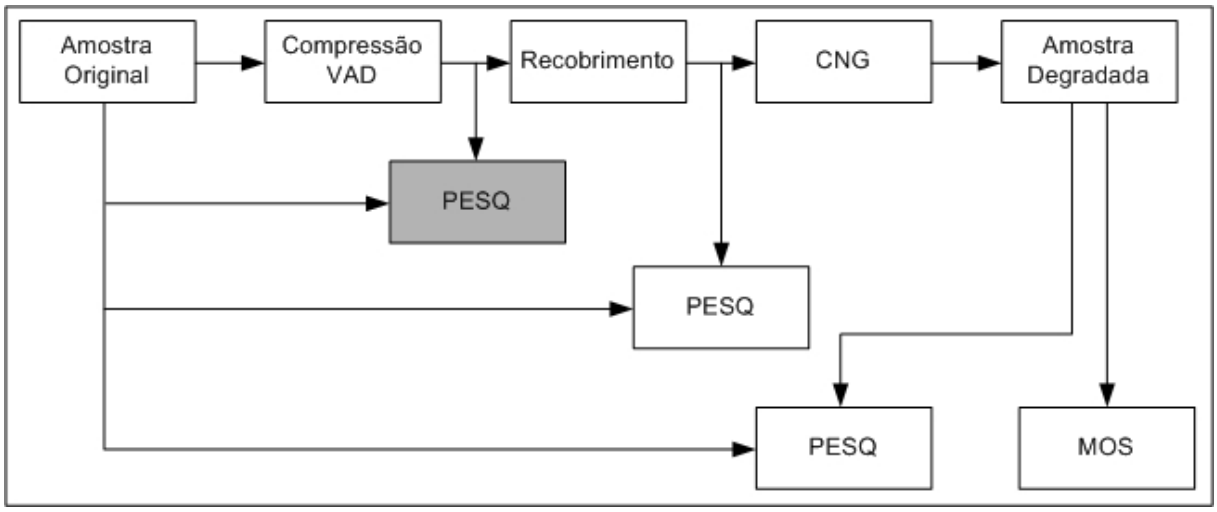


Figura 30 - Diagrama de blocos do cenário com destaque para o sinal avaliado pelo PESQ
Fonte: O autor 2009.

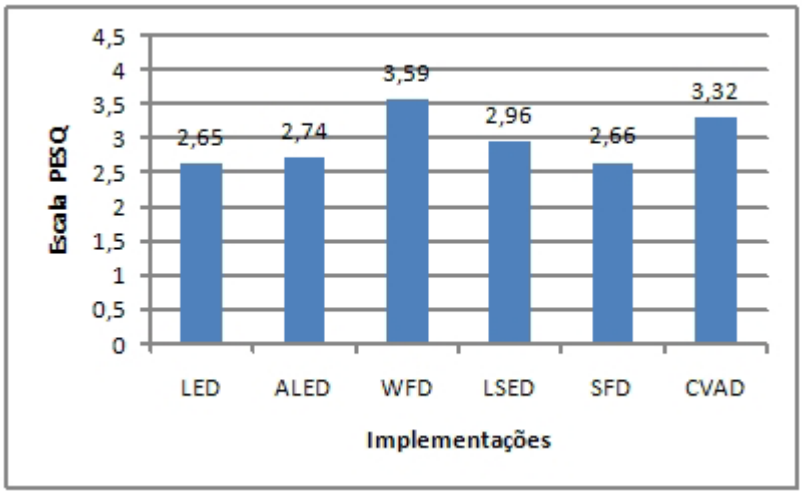


Gráfico 1 - Valor do PESQ obtido para as técnicas implementadas.
Fonte: O autor 2009.

Como se observa nesta etapa, apenas duas (WFD e CVAD) técnicas obtiveram nota acima de 3, o que pode parecer satisfatório em termos de sistemas de comunicação. O mau desempenho das demais técnicas pode significar muita coisa. Só que pouco se pode afirmar até agora em termos de desempenho, por não haver outro elemento de comparação. Só que, o que se pode dizer, conforme afirma Davis (2002), é que as técnicas de VAD não devem prejudicar em nada a qualidade do sinal da voz. Caso isso aconteça quer dizer que o sistema não está bem ajustado.

Como forma de facilitar um pouco a observação dos resultados, pode-se analisar os percentuais de compressão do sinal avaliado pelo PESQ no Gráfico 1. Para tanto, é apresentada a Tabela 5 com os percentuais de compressão aplicados a cada amostra por cada técnica implementada e avaliada.

Tabela 5 - Valores percentuais de compressão obtidos para as sentenças avaliadas.

	LED	ALED	WFD	LSED	SFD	CVAD
Choveu muito neste fim de semana.	44.7%	42,00%	12,10%	33,00%	51,70%	38,00%
Ela precisa esperar na fila.	38,90%	41,30%	19,56%	52,30%	45,10%	38,02%
O banco fechou sua conta.	42,00%	43,50%	22,40%	52,30%	52,50%	44,90%
Guardei o livro na primeira gaveta.	36,70%	38,80%	17,70%	35,40%	52,00%	39,00%
Percentual médio de supressão de silêncio	**39,20%**	**41,40%**	**17,94%**	**43,25%**	**50,33%**	**39,98%**

Fonte: O autor 2009.

Como cada amostra possuía um tamanho diferente (entre 2 e 3 segundos), foram tiradas as médias das quatro amostras para cada técnica aplicada. Estas médias são apresentadas no Gráfico 2 de forma a se obter um comparativo visual dos resultados.

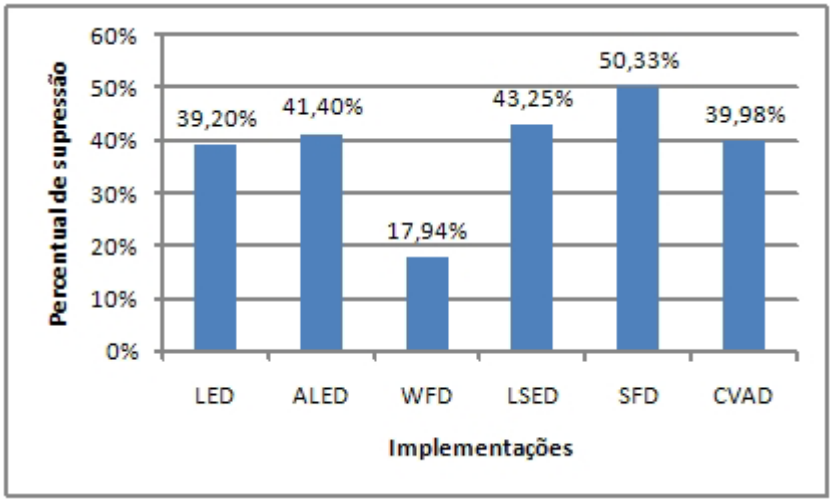


Gráfico 2 - Nível percentual de supressão de silêncio pelas técnicas de compressão
Fonte: O autor 2009.

Para esse primeiro teste, se faz ênfase mais uma vez, conforme apresentado na Figura 30, ao fato de que neste ponto as amostras utilizadas estavam apenas sob o efeito das técnicas de supressão dos períodos considerados silêncio. Assim, nem a técnica de recobrimento e nem a inserção de ruído de conforme foram utilizados.

Ainda quanto aos resultados já apresentados aqui, não basta avaliar apenas o valor do PESQ. Como visto no Gráfico 2, os percentuais de supressão aplicados as amostras neste ponto foram bastante consideráveis. Fazendo uma média das médias das seis técnicas apresentadas no Gráfico 2 de supressão, obteve-se quase 39% de compactação das amostras originais, o que é um resultado bastante expressivo. Verificando-se as citações de Hersent (2002), Davis (2002), Monteiro (2002) e Kondoz (2000), em que em uma conversação, cada pessoa faça uso de apenas 35% a 50% do tempo do canal ativo, esse resultado parece bastante

interessante. Mas há que se considerar que nas amostras utilizadas não existem pausas alongadas na fala, que sejam maiores do que os tempos normais entre uma sílaba e outra, ou entre uma palavra e outra.

Também se observa na análise comparativa do Gráfico 1com o Gráfico 2, é que existe quase uma proporcionalidade inversa entre os resultados das mesmas técnicas. Isso quer dizer que quanto maior a qualidade do sinal medido, menor o percentual de compressão. Pode-se dizer que isso seria um tanto óbvio. Mas do ponto de vista da aplicabilidade de cada técnica isso é bastante relevante, pois de nada adianta alta qualidade do sinal, e praticamente nenhuma contribuição em termos de supressão dos períodos de silêncio. Como descrito anteriormente, o que se busca é a combinação de um melhor resultado nos dois sentidos.

Outra análise que pode ser feita com os dados obtidos até aqui, é o da estimativa de economia de banda do canal ativo, em termos de volume de dados que não seriam injetados na rede. Para esta análise foi estimado um valor referente ao volume de dados produzido, considerando o conteúdo de ruído e mais o *overhead* dos cabeçalhos dos protocolos IP (20 bytes), UDP (*User Datagram Protocol*) (8 bytes) e RTP[19] (12 bytes), totalizando 40 bytes para cada datagrama injetado na rede (PERCY, 2005). Quanto ao volume de dados gerado, caso não fossem aplicadas as técnicas de compressão propostas nesse trabalho, e utilizando uma taxa de geração de dados de 64 kbit/s, considerando uma codificação PCM, frequência de amostragem de 8 kHz e 8 bits por amostra, e pacotes de 20 ms de áudio. No total tem-se 40 bytes de cabeçalho e mais 160 bytes de dados (isso sem a aplicação de outra técnica de compressão), totalizando 200 bytes por datagrama.

Isso significa que a cada pacote de áudio suprimido, evita-se que 200 bytes sejam injetados na rede. Sendo o tamanho dos pacotes de áudio de 20 ms, isso significa que a cada segundo teria-se 50 pacotes sendo injetados na rede, perfazendo um total de 10 kB/s. Assim, levando em consideração o percentual de compressão (Gráfico 2) obtido com a aplicação das técnicas, fez-se uma estimativa da quantidade de dados que não seriam injetados na rede pela aplicação de voz sobre IP. Esse resultado é apresentado no Gráfico 3 abaixo, em quantidade de bytes por segundo.

Observando ainda, e como citado, esse valor de economia de banda proporcionada pelas técnicas, é apenas do canal ativo de voz, o que equivale ao canal onde o locutor está falando. Ainda tem-se o canal passivo, onde o ouvinte, em um primeiro momento, estaria apenas escutando. Ou seja, o ouvinte estaria gerando apenas ruído. No caso no canal passivo,

[19] Mais informações sobre o protocolo RTP e a formação dos datagramas, considerando o *overhead* dos cabeçalhos, é apresentado no Anexo B deste trabalho.

não foram feitas medições específicas do percentual de supressão quando da presença de apenas ruído, mas estimasse algo próximo de 100 % de supressão do sinal. Não se afirma que o nível de supressão seria de 100 % diante de um sinal puramente ruidoso, devido a possibilidade de variação desse ruído. Além disso, nas técnicas implementadas, expostas a um sinal puramente ruidoso e razoavelmente constante em termos de amplitude do sinal, o algoritmo de adaptação do limiar de silêncio tente para um valor bastante baixo, o que pode fazer com que qualquer pequena alteração no sinal ruidoso leve a técnica a considerar que existe sinal ativo, especialmente a LED, na qual o algoritmo é puramente baseado na verificação do conteúdo de energia do sinal. Nesse caso, o que garante que o algoritmo de adaptação não venha a convergir para um valor igual ao do próprio ruído é a constante de segurança *k*. A constante de segurança acaba por dar certa estabilidade no algoritmo em termos de classificação do que é pacote ativo e inativo.

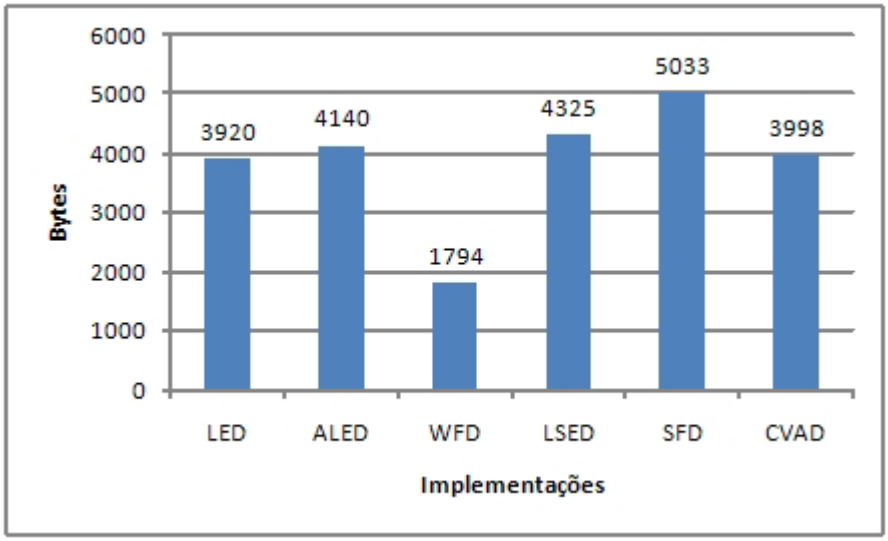


Gráfico 3 - Total de bytes suprimidos considerando o percentual de supressão obtido.
Fonte: O autor 2009.

Outro ponto é que a aplicação de técnicas de compressão bem mais eficientes que a taxa de dados proporcionada pela codificação PCM, melhoram e muito a compactação dos dados. Nesse sentido é óbvio que o uso de Vocoder como os das recomendações do ITU como G.723 e G.729, por exemplo, são muito mais eficientes. Mas esse poder de compactação dos Vocoder só se aplica aos dados de mídia e não ao cabeçalho do protocolo, o que remete ao que foi citado na introdução do trabalho como sobrecarga da transmissão de voz sobre redes de datagramas, a qual subentende-se como rede IP.

6.2 ANÁLISE NA SAÍDA DO BLOCO DE RECOBRIMENTO

Para a análise dos próximos resultados apresentados já foi considerado o uso da técnica de recobrimento, considerando assim o sinal de saída do bloco de recobrimento como mostrado na Figura 31.

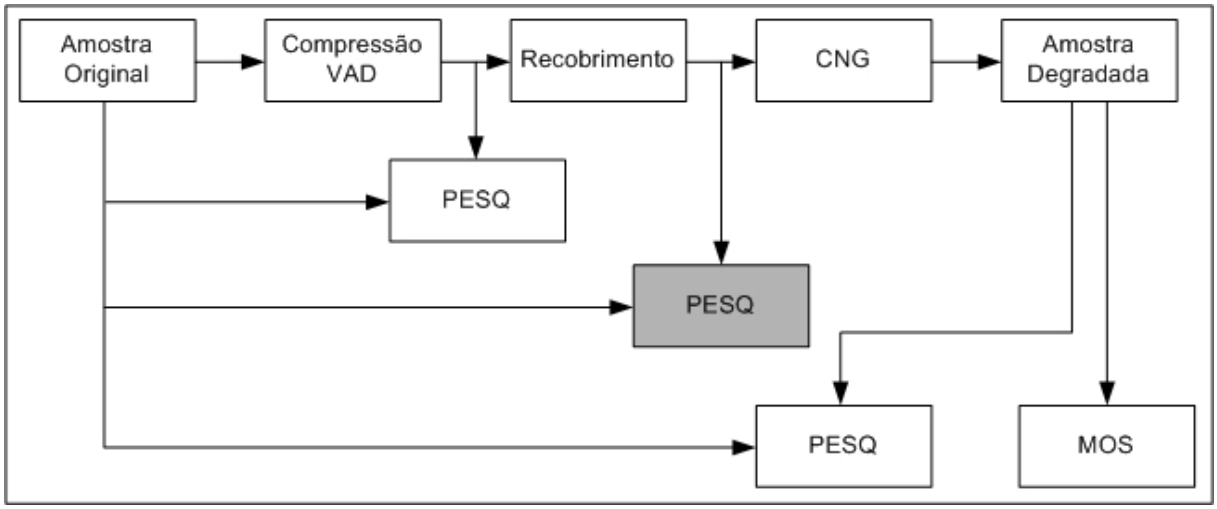


Figura 31 - Diagrama de blocos do cenário com destaque para o sinal avaliado pelo PESQ - com recobrimento
Fonte: O autor 2009.

Lembrando que a técnica de recobrimento objetiva evitar o efeito de *clipping* na voz, o que gera um efeito de cortes súbitos na fala.

Salienta-se que das técnicas de avaliação objetiva que se verificaram na bibliografia, o PESQ é o que apresenta melhores resultados em termos de avaliação do efeito de *clipping* segundo Fernandes (2003) e ITU (2001).

Quanto a medição realizada, mais uma vez se utilizou o PESQ, comparando o sinal degradado na saída do bloco de recobrimento com o sinal de entrada do sistema, sem degradação. O resultado do valor PESQ para as seis técnicas analisadas é apresentado no Gráfico 4.

Como se observa, os resultados são bem mais satisfatórios quando do uso do recobrimento, em termos de valores PESQ. O que mostra uma importante contribuição do recobrimento para melhoria da qualidade do áudio percebido. Mas como abordado no capítulo dois, o recobrimento acaba por complementar as técnicas de supressão com a consideração de um número de pacotes avaliados como inativos, como ativos, logo após um ativo. Nesse caso aqui, foram considerados dois pacotes de recobrimento, ou seja, 40 ms de áudio.

Com o uso do recobrimento, mais uma vez precisa-se avaliar a contribuição das técnicas em termos de percentual de supressão de silêncio, já que houve a consideração de

mais pacotes ativos e por consequência menor compactação. Esse resultado pode ser visto na Tabela 6 com o percentual de compactação aplicado a cada técnica.

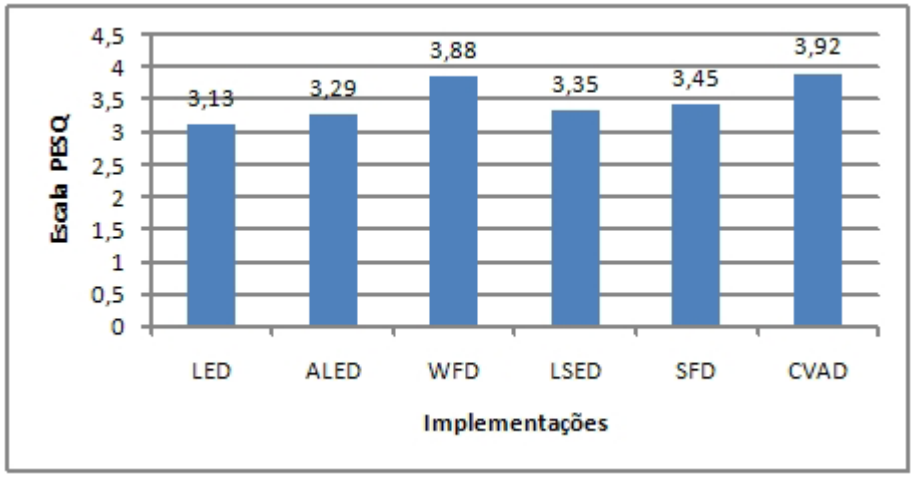


Gráfico 4 - Valor do PESQ obtido para as técnicas aplicadas após o uso do recobrimento
Fonte: O autor 2009.

Tabela 6 - Valores percentuais de compressão após o recobrimento.

	LED	ALED	WFD	LSED	SFD	CVAD
Choveu muito neste fim de semana.	24,70%	24,40%	0,84%	25,20%	43,70%	24,40%
Ela precisa esperar na fila.	23,90%	23,90%	13,30%	27,40%	34,50%	23,90%
O banco fechou sua conta.	38,50%	40,60%	15,40%	47,60%	42,70%	38,50%
Guardei o livro na primeira gaveta.	24,40%	24,20%	6,82%	26,50%	32,60%	26,50%
Percentual médio de supressão de silêncio	**27,88%**	**28,28%**	**9,09%**	**31,68%**	**38,38%**	**28,33%**

Fonte: O autor 2009.

As médias de compactação obtidas por cada técnica são visualizadas no Gráfico 5.

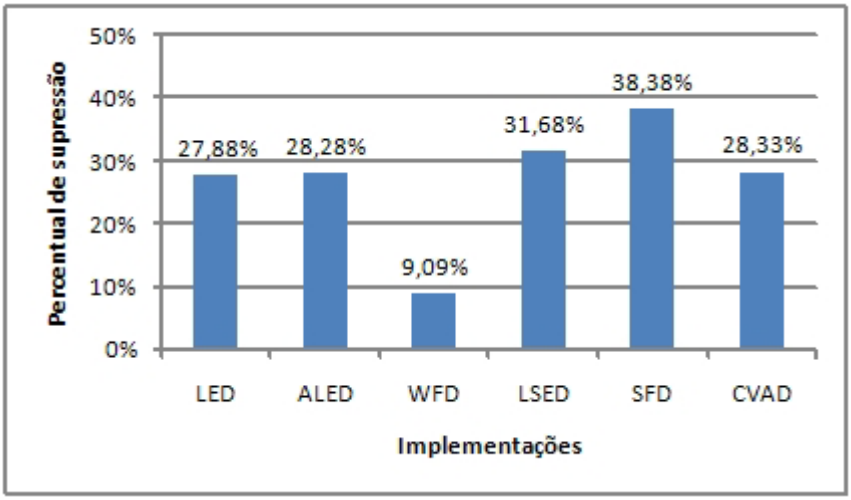


Gráfico 5 - Nível de supressão de silêncio após a aplicação da técnica de recobrimento
Fonte: O autor 2009.

Analisando o Gráfico 5 fica claro que quando comparado com o Gráfico 2, a contribuição das técnicas em termos de percentual de compressão é bem menor depois do uso do recobrimento. Mas em termos de qualidade do áudio medido pelo PESQ, comparando o Gráfico 1 com o Gráfico 4, a melhora é significativa.

No Gráfico 4 as técnicas WFD e CVAD continuam a mostrar os melhores resultados. Mas por outro lado, no Gráfico 5 há uma distinção clara entre as duas técnicas. Enquanto a CVAD apresenta um percentual de supressão acima da média, a WFD é a pior em termos de resultado. Com menos de 10% de supressão do sinal original, a qualidade proporcionado só pode ser alta.

Como na etapa anterior de análise dos resultados proporcionados, aqui também pode-se fazer uma estimativa de economia de banda proporcionada em termos de quantidade de dados não injetados na rede. Seguindo as mesmas especificações em termos de composição dos dados apresentado no item 5.1 deste trabalho, apresenta-se no Gráfico 6 abaixo a estimativa de bytes por segundo não colocados na rede para os percentuais de compressão apresentados agora no Gráfico 5 logo acima.

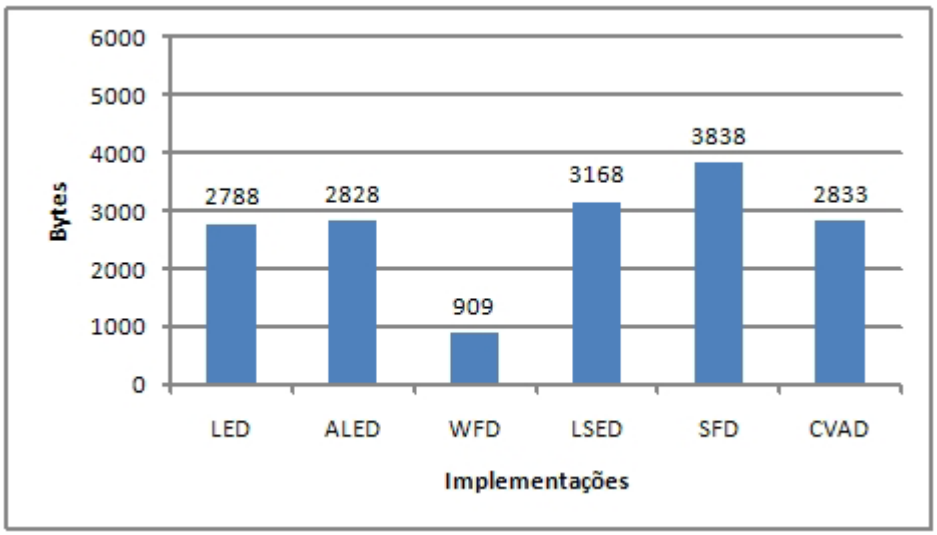


Gráfico 6 – Estimativa do total de bytes suprimidos após o uso do recobrimento
Fonte: O autor 2009.

Comparando o Gráfico 3 com o Gráfico 6, verifica-se que o resultado em termos de economia proporcionada piorou, como de certa forma já se esperava, considerando os percentuais de compressão. Mas ratifica-se que, como observado nos resultados apresentados que quanto menos compressão, menos degradação na qualidade do áudio.

6.3 ANÁLISE DA AMOSTRA DEGRADADA

Como relatado no capítulo quatro, para a análise da qualidade da amostra degradada foram aplicados dois tipos de métodos para a avaliação. O método objetivo usando o PESQ, dando sequência aos testes que tiverem seus resultados já apresentados, e o método subjetivo usando o MOS, com o objetivo de validar o processo, tanto de implementação das técnicas de compressão, bem como validar a metodologia aplicada.

Destaca-se que para essa etapa de análise dos resultados, foram utilizadas as mesmas amostras degradas, tanto para os testes de avaliação objetiva quanto para os testes de avaliação subjetiva. Isso significa, mesmos percentuais de supressão analisados pelos dois métodos, mesma quantidade de pacotes de recobrimento e mesmo ruído.

Para a análise dos resultados aqui apresentados, tanto as medições do PESQ, quando do MOS em termos de percentual de supressão de silêncio, os valores são os mesmos apresentados no Gráfico 5 e estimados em volume de dados no Gráfico 6, apresentados anteriormente. Não há alteração dos percentuais de compressão porque, na aplicação real, o ruído de conforto, como explicado anteriormente, só é aplicado junto ao lado do ouvinte. Isso faz com que esse ruído não trafegue pela rede, exceção feita a um pacote esporádico contendo o *Payload* de *Confort Noise* que pode transportar uma amostra do ruído do lado do locutor para o lado do ouvinte.

6.3.1 Avaliação objetiva

Para a avaliação objetiva da amostra degradada, mais uma vez se compara o sinal de saída com o sinal de entrada, como apresentado na Figura 32.

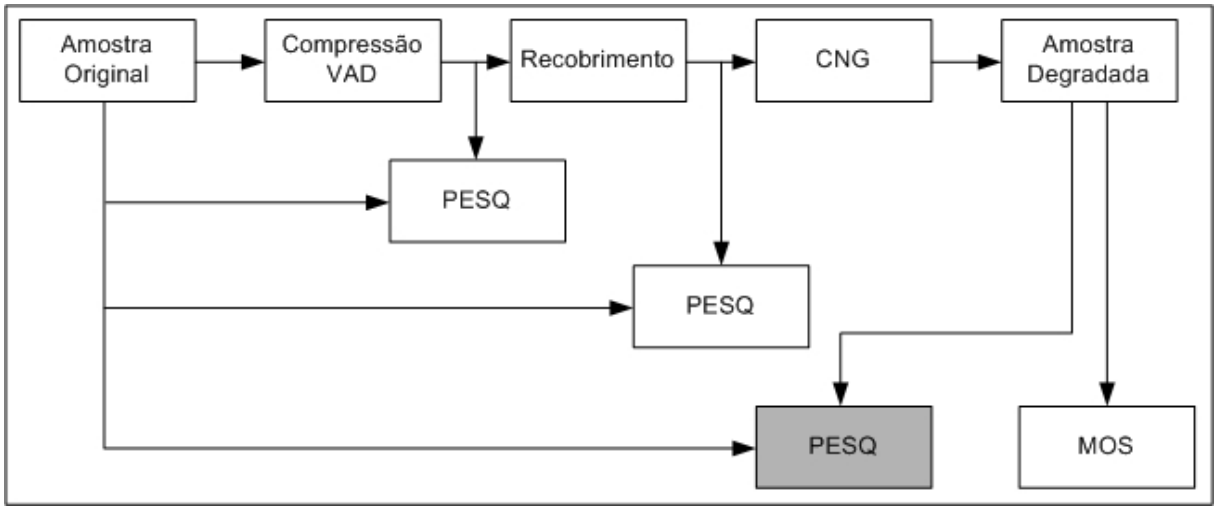


Figura 32 - Diagrama de blocos do cenário com destaque para o sinal avaliado pelo PESQ - com ruído de conforto
Fonte: O autor 2009.

Para esta análise, o sinal degradado já com a sobreposição de ruído de conforto sobre os períodos considerados silêncio até então.

O resultado do valor PESQ medido para este cenário é apresentado no Gráfico 7.

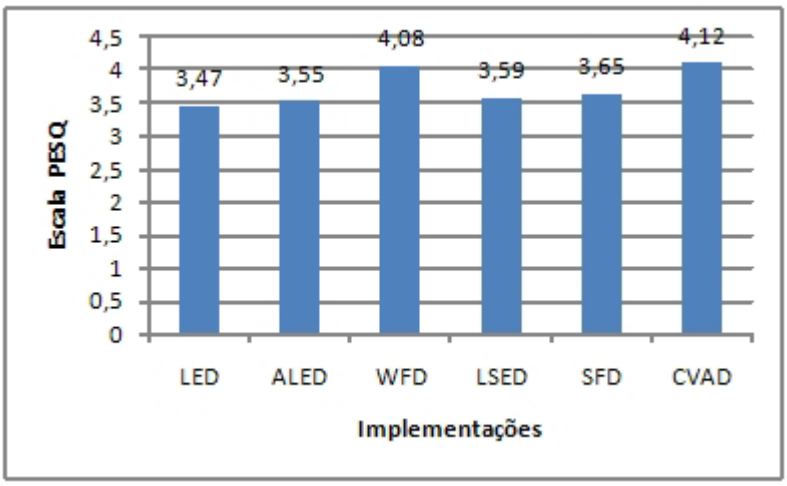


Gráfico 7 - Valor do PESQ obtido após o uso do ruído de conforto
Fonte: O autor 2009.

Ainda sobre o Gráfico 7, observa-se uma significativa melhoria da qualidade das amostras analisadas em relação ao Gráfico 4 e o Gráfico 1. A inserção de ruído não é tão impactante no resultado quanto a aplicação da técnica de recobrimento, mas sem dúvida implementa uma melhora. Outro ponto que deve ser levado em consideração, é o fato de o gerador de ruído tomar como base a energia do ruído amostrado da própria amostra original. Isso faz com que o ruído inserido pelo gerador tenha características bem próximas da amostra original.

6.3.2 Avaliação subjetiva

Dentro do cenário de análise, a Figura 33 apresenta o sinal de referência utilizado para os testes com o MOS.

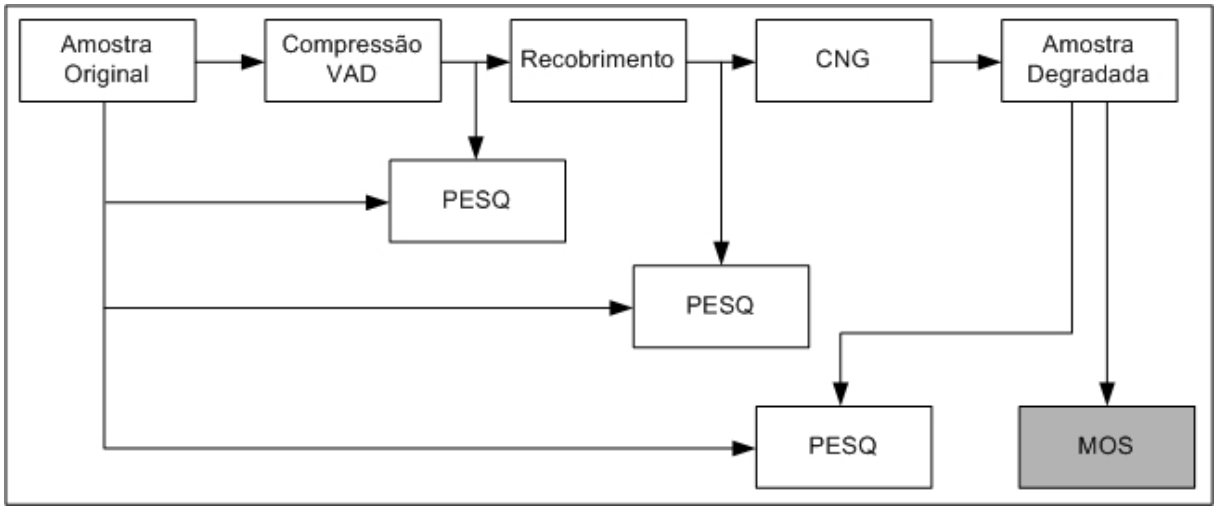


Figura 33 - Diagrama de blocos do cenário com destaque para o sinal avaliado pelo MOS
Fonte: O autor 2009.

No condizente ao nível de esforço necessário para o entendimento das quatro sentenças, o resultando é apresentado no Gráfico 8. O resultado é uma média[20] das quinze notas dadas por técnica avaliada pelo público entrevistado.

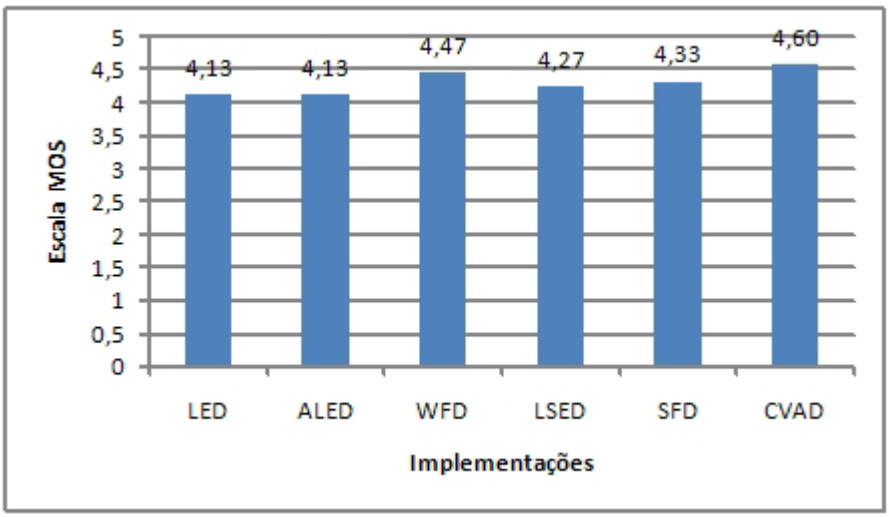


Gráfico 8 - Valor do MOS obtido após o uso do ruído de conforto
Fonte: O autor (2009).

Lembrando que para a análise dos resultados desta medição, em termos de percentual de supressão de silêncio, os valores são os mesmos apresentados no Gráfico 5, mostrado anteriormente.

6.4 CORRELAÇÃO DOS RESULTADOS

Uma análise que é bastante pertinente referente aos resultados das medições realizadas, tanto pelo PESQ quanto pelo MOS, é a correlação existente entre os dois métodos de medição da qualidade do sinal de voz. Destas relações podem-se tirar parâmetros de o quanto próximos estão os resultados. Para tanto, como descrito no capítulo quatro, foram calculadas as correlações entre os valores PESQ apresentados no Gráfico 1, Gráfico 4 e Gráfico 7 com o valor do MOS apresentado no Gráfico 8.

Neste sentido, a primeira correlação que propõe-se é entre o valor do PESQ obtido para o sinal degradado (após a aplicação das técnicas de compressão, sem o recobrimento e sem o ruído de conforto) (Gráfico 1) com o MOS (Gráfico 8). Essa correlação está identificada como "*A*" no Gráfico 9. A segunda correlação feita foi entre os valores obtidos para o PESQ, apresentados no Gráfico 4 com, novamente, o Gráfico 8 do MOS. O resultado dessa correlação foi identificado com "*B*" no Gráfico 9. Por fim, a última correlação, foi entre o resultado do PESQ apresentado no Gráfico 7 com o MOS do Gráfico 8. Essa correlação está identificada como "*C*" no Gráfico 9.

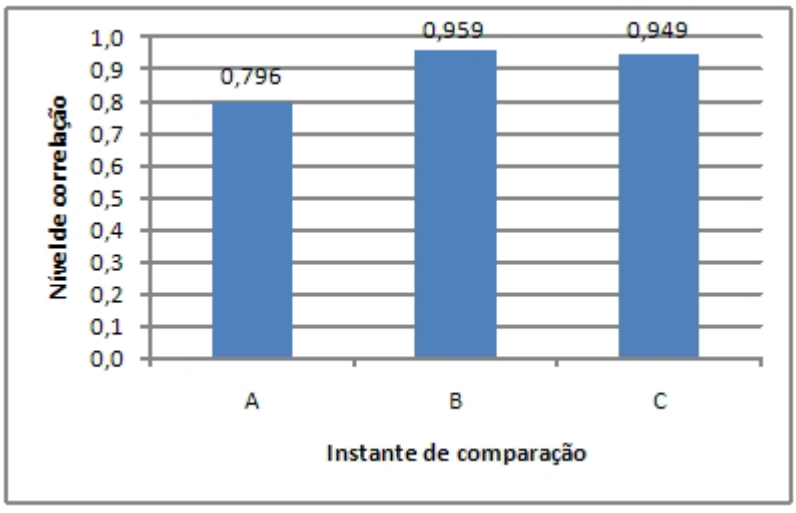


Gráfico 9 - Correlação entre respostas de nível de qualidade da voz medidos
Fonte: O autor (2009).

Considerando que, para os valores de correlação, quanto mais próximo de um estiver o resultado, mais correlacionado apresentam-se as amostras. Neste sentido o resultado se mostra

[20] A recomendação P.800 instrui para a realização do cálculo da média dos resultados obtidos, não recomendando o cálculo do desvio padrão (ITU, 1996a).

bastante satisfatório. Outro motivo que justifica o fato do resultado ser satisfatório é que, como apresentado em ITU (2001) e descrito no capítulo quatro deste, um valor de correlação médio para um comparativo entre valores MOS e PESQ, quando da análise dos mesmos parâmetros, como no caso de "*C*" (mesmo entrada e saída com os mesmos parâmetros tanto para o PESQ quanto para o MOS) é 0,935. Tudo isso, mesmo que o resultado de "*A*" tenha ficado abaixo de 0,8, mas pode-se verificar que o sinal usado para correlacionar com a saída, ainda não tinha a aplicação da técnica do recobrimento e aplicação de ruído branco. Também se verifica que esse citado sinal teve uma avaliação bastante baixa em termos de qualidade do áudio proporcionada, conforme Gráfico 1.

Outro resultado a ser observado é o fato de no Gráfico 9, o resultado de "*B*" ter sido superior ao resultado de "*C*". Isso porque o sinal do teste "*C*" era "mais completo", com recobrimento e ruído de conforto. Para esse resultado o que se especula é que como as condições de teste, especialmente as do MOS, não foram as ideais e isso pode ter alterado alguma coisa no resultado, mesmo que se tenha seguido a recomendação dentro do possível, inclusive com mais rigor que os próprios trabalhos de Sangwan (2002a) e Becvar (2007) descrevem em seus trabalhos.

7 CONCLUSÕES

Quanto aos resultados obtidos com o desenvolvimento, observa-se claramente que as duas técnicas mais elaboradas (por terem algoritmos mais complexos que as demais), a WFD e especialmente a CVAD, apresentaram menor necessidade de esforço do ouvinte para o entendimento das sentenças, em todos os testes. Com relação a isto, pode-se fazer um apontamento relevante à questão de verificação do número de cruzamentos do zero (*zero-crossing rate*) ocorridos em cada pacote de voz. As técnicas WFD e CVAD, usuárias deste algoritmo, tiveram ampla vantagem de desempenho em relação às demais. É cabível associar este resultado em particular às conclusões de Rabiner (1978). O autor destaca a complementaridade entre a verificação dos cruzamentos de zero pelo sinal de voz e a distribuição de energia do mesmo sinal para a detecção da presença de voz ativa. Isto porque, como o próprio autor referencia, altas freqüências, entre 3 kHz e 4 kHz implicam em altas taxas de cruzamentos de zeros e baixos níveis de energia e baixas freqüências, de 0 Hz a 3 kHz, implicam em baixas quantidades de cruzamentos do zero.

Quanto ao fraco desempenho das técnicas LED e ALED, e ainda a LSED em relação às demais técnicas baseadas no domínio das freqüências, também se pode apontar para as mesmas conclusões de Rabiner (1978) apontadas anteriormente. A falta de uma complementaridade para estas técnicas, já que as mesmas são somente baseadas na energia do sinal de voz, torna seu desempenho reduzido quando da busca pela detecção de pacotes de áudio com voz ativa.

Sobre a técnica SFD, apesar do bom resultado apresentado pela mesma dentro do conjunto das seis técnicas, esta técnica necessita ser testada em outros ambientes de modo a avaliar o desempenho em locais com diferentes relações entre o sinal de voz e o ruído.

Outra situação que pode se concluir é que um baixo rendimento da técnica WFD com relação ao percentual de supressão de silêncio pode estar se contra-ponto ao bom rendimento na questão da qualidade do áudio. Também deve-se considerar que, quanto menor o nível de supressão, mais próximo do áudio original será a amostra de voz sob a ação das técnicas de supressão e melhor será a qualidade do mesmo.

A técnica SFD apresenta um resultado positivamente destacável em relação às demais. Mas mais uma vez fica uma ressalva e o apontamento sobre a necessidade de mais testes para a confirmação da regularidade de bons resultados desta técnica quando da obtenção das amostras de voz em ambientes com condições de ruído adversas.

Quanto as outras quatro técnicas, as mesmas apresentaram praticamente o mesmo resultado em termos de percentual de compressão. Apenas destacando a tendência apresentada pela técnica CVAD de perda de performance em relação às demais, assim como a WFD, mas não tão acentuado como a mesma.

Em uma análise do resultado geral condizente ao percentual de supressão de silêncio obtido para amostras curtas de voz, associado ao elemento foco das técnicas de detecção e supressão de silêncio que é a economia de banda maximizada da rede IP, observa-se que se pode pensar em uma redução mínima em torno de 25% da utilização do canal ativo da rede com a aplicação destes algoritmos. Pode-se assim pensar, que em uma chamada ponto a ponto a economia estaria próxima a 50% da largura de banda demanda pela aplicação, o que confere com a afirmação de Hersent (2002).

Outro comentário a ser feito com relação aos resultados obtidos com a aplicação do MOS, é que após serem realizadas as entrevistas onde o público respondeu ao nível de esforço necessário para o entendimento das amostras, as mesmas pessoas ouviram as mesmas amostras de voz com diferentes características. Algumas destas amostras estavam sem a ação da técnica de recobrimento, outras sem a inserção de ruído branco e ainda haviam outras amostras sem pacotes de recobrimento e sem ruído ao mesmo tempo. Praticamente todos os ouvintes foram unânimes quanto à queda na qualidade do áudio e sobre o aumento do esforço necessário para o bom entendimento das sentenças.

A questão do *clipping* (corte) da voz do locutor só não foi citada para a técnica WFD, o que leva a pensar que esta mesma técnica possa ter o número de pacotes de recobrimento reduzido ou inclusive zerado, o que certamente melhoraria o seu resultado com relação ao percentual de supressão de silêncio proporcionado.

Quanto a inserção de ruído branco, foi citado pelo público ouvinte que torna-se mais agradável o áudio percebido quando da existência do ruído. Houve a solicitação apenas de tornar o ruído mais brando, ou menos intenso do que o utilizado para o experimento. Lembrando que a energia do ruído utilizada era proporcional a energia do ruído dos 200 ms iniciais de cada amostra. Isso pode significar que as amostras precisariam de um ambiente de gravação mais controlado, com menos ruído.

O fato destas últimas conclusões não terem sido relatados na forma de gráficos e tabelas, como apresentados os demais resultados, é porque não haviam sido previstos antes do início dos testes. Outro motivo foi o fato de não se ter feito um levantamento estatístico destes dados, mas que por serem considerados de alta relevância para o fechamento foram então, aqui relatados.

Por fim, como cita Davis (2002), as técnicas de VAD não afetam diretamente a qualidade do sinal de voz, quando operam de forma correta, o que do contrário pode certamente diminuir drasticamente os níveis de inteligibilidade da fala. Cita ainda Davis (2002), que demasiado uso da técnica de recobrimento pode reduzir a eficiência das técnicas em termos de economia de banda, e do contrário afetar a qualidade do áudio.

Cita ainda Davis (2002) a importância e complementariedade do uso de ruído de conforto no lado do ouvinte, o que melhora a percepção do mesmo quanto ao andamento da chamada.

Considerando essas colocações de Davis (2002), pode-se dizer que o objetivo inicial do trabalho foi alcançado. As técnicas foram implementadas e analisadas. Alguns resultados foram bastante satisfatórios e outros poderiam ser ainda melhorados.

O que sem dúvida fica claro, é a contribuição do trabalho com a possibilidade de compressão de sinais de áudio, quando em específico a aplicações de redes de pacotes, sem a necessidade do uso de codificadores mais complexos. Isso certamente tem um resultado final com campo de aplicação bastante vasto, visto que não haverá apenas economia de banda da rede IP, mas também uma diminuição do tempo total de processamento do sinal de voz.

7.1 TRABALHOS FUTUROS

Com relação à continuidade do trabalho, pode-se criar uma lista de atividades futuras por dois caminhos, mais testes apenas sobre o que já foi implementado ou novas técnicas de detecção e supressão de silêncio associadas a mais testes.

Uma abordagem válida seria trabalhar mais sobre as técnicas já implementadas e aqui apresentadas de forma a realizar mais testes de validação das mesmas. Isto incluiria mais pessoas entrevistadas quando do uso da avaliação subjetiva via P.800, ou mais amostras quando da avaliação objetiva via P.862, com diferentes condições de ambiente e submissão a diferentes ruídos correlacionados. Ainda quanto a mais amostras, mais comparativos entre as técnicas, mais amostras com diferentes locutores e diferentes características de amostras, como um diálogo entre duas pessoas, por exemplo. Utilização de amostras com maior qualidade de gravação, adquiridas em local mais compatível com as recomendações do ITU-T e fundamentalmente testes das técnicas aplicadas junto a um comunicador de voz sobre IP.

Outra abordagem seria buscar a implementação de algoritmos de VAD mais robustos de forma a dar às técnicas de detecção e supressão de silêncio uma maior independência e estabilidade no que diz respeito a sua adaptabilidade às condições adversas impostas pela inconstância do ruído ambiente. Isto poderia ser solucionado pela utilização de algoritmos adaptativos de alto desempenho. Porém existe aí um sério contraponto que seria o consumo computacional demandado por estes algoritmos que ainda são muito elevados, considerando a demanda computacional dos mesmos.

Pode-se salientar aqui que vislumbra-se a necessidade de uma futura evolução no trabalho a fim de determinar quais seriam os limites mínimo e máximo para essa adaptação do ruído de conforto, a ponto de se manter uma relação sinal ruído que permita uma qualidade de áudio aceitável para os padrões estabelecidos para telecomunicações.

Por fim, acredita-se que resultados mais satisfatórios em termos de correlação dos resultados possam ser obtidos. Para tanto, se sugere que um próximo passo seja, por exemplo, a correlação entre a quantidade de energia de cada amostra, de cada frase pré-gravada, e a energia das amostras ditas degradadas. Esse levantamento é perfeitamente factível só um pouco custoso devido a quantidade de amostras de voz trabalhadas. O que geraria um total de 72 resultados de correlações.

REFERÊNCIAS

(AGYEI-KODIE, 2003) AGYEI-KODIE, K. ***Development of Voice Activity Detection (VAD) Algorithms that is Robust Low Signal-to-Noise Ratios***. University ECE, 2003.

(BALBINOT , 2004) BALBINOT, Ricardo et al. **Voz sobre IP** - Tecnologia e tendências. Anais do XXI Simpósio Brasileiro de Telecomunicações - SBT, v. 1, Belém/PA, 2004.

(BALBINOT, 2002) BALBINOT, R. **Modelagem e Protipagem de Sistemas de Voz Sobre IP com Mecanismos de Transmissão Robusto.** 2002. Dissertação (Mestrado, Faculdade de Engenharia) - PUCRS, Porto Alegre, 2002.

(BARBEDO, 2001) BARBEDO, J. G. A. **Avaliação objetiva da qualidade de codecs de voz na faixa de telefonia**. 2001. Dissertação (Mestrado. Faculdade de Engenharia e Computação) – Unicamp, Campinas, 2001.

(BARBEDO, 2004) BARBEDO, J. G. A. **Avaliação objetiva de qualidade de sinais de áudio e voz**. 2004. Tese (..... Faculdade de Engenharia Elétrica e de Computação) - Unicamp, Campinas, 2004.

(BARCELOS, 2005) BARCELOS, A. V. **Voxcount** – Implementação de uma plataforma de contabilização aplicada à Voz sobre IP. 2005. Dissertação (..... Faculdade de Engenharia) - PUCRS, Porto Alegre, 2005.

(BECKER, 2005) BECKER, R. et al. ***A silence detection and suppression technique design for voice over IP systems. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing***. Victoria, 2005.

(BECVAR, 2007) BECVAR, Z. et al. ***Impact of Additional Noise on Subjective and Objective Quality Assessement in VoIP Multimedia Signal Processing.*** MMSP IEEE 9th Workshop on. Creta, 2007.

(BENYASSINE, 1997) BENYASSINE, A. et al. ***A Robust Low Complexity Voice Activity Detection Algorithm for Speech Communication System*.** IEEE Workshop on Speech Coding, Pocono Manor, Pennsylvania, USA, 1997.

(CAI, 2004) CAI, Libin.; ZHAO, Jiying. ***Speech quality assessment using digital watermarking***. Haptic, Audio and Visual Environments and Their Applications, 2004. HAVE 2004. Proceedings. The 3rd IEEE International Workshop on. Issue 2-3, p. 177 - 182, Oct. 2004.

(CASTELLO, 2004) CASTELLO, F. C. **Modelagem e Prototipagem de um Sistema de Voz Sobre IP baseado na implementação de Protocolos IETF aplicado a um Ambiente de Interconexão com Sistemas Telefônicos Convencionais**. 2004. Dissertação (Mestrado Faculdade de Engenharia) - PUCRS, Porto Alegre, 2004.

(CASTRO, 2006) CASTRO, F. C. C.; CASTRO, M. C. F. ***Multilayer Perceptron***. Capítulo 4. Material de aula, disciplina de Redes Neurais. Programa de Pós-Graduação em Engenharia Elétrica – PUCRS, Porto Alegre, 2006.

(CONWAY, 2000) CONWAY, A. E. **A performance monitoring system for VoIP gateways**. Workshop on Software and Performance. Ottawa, Canadá, p. 38-43, 2000.

(CORRÊA, 1996) CORRÊA, Juarez Sagebin; RODRIGUES, Sílvio Lobo. **Programas Aplicativos ao Processamento de Sinais em tempo Discreto**. Porto Alegre, EDIPUCRS, 1996.

(CORSETTI, 2004) CORSETTI, G. R. et al. **Implementação de um Filtro Adaptativo LMS Aplicado ao Cancelamento de Eco em Voz sobre IP**. II Escola Regional de Redes de Computadores, Canoas, 2004.

(DAVIS, 2002) DAVIS, Gillian. M. ***Noise reduction in speech applications***. Florida, USA: CRC Press, 2002.

(EHLERS, 2003) EHLERS, R. S. **Introdução a inferência Bayesiana**. Disponível em <http://leg.ufpr.br/~paulojus/CE227/ce227/> . Acesso em 26 de março de 2009.

(EMPIRIX, 2009) EMPIRIX. ***Assuring QoE on next generetion networks***. Whitepaper, Communications infrastructure test group. Disponível em <www.empirix.com>. Acesso em 26 de março de 2009.

(FERNANDES, 2003) FERNANDES, N. L. L. **Relação entre a Qualidade das Respostas das Recomendações G.723.1 E G.729, e o Comportamento da Rede IP de Suporte**. 2003. Tese (Mestrado em Ciências em Engenharia de Sistemas e Computação) - COPPE/UFRJ, Rio de Janeiro, 2003.

(FLEURY, 2005) FLEURY, C. A.; CARRIJO, G. A. **Quantização Vetorial Classificada Adaptativa Perceptivamente**. Congresso Nacional de Matemática Aplicada e Computacional. São Paulo, 2005.

(GONZALEZ, 1993) GONZALEZ, R. C.; WOODS, R. E. ***Digital image processing***. 3. ed. Addison-Wesley Publishing Company, 1993.

(HAYKIN, 2004) HAYKIN, S. **Sistemas de comunicação**: analógicos e digitais. 4 ed. Porto Alegre: Bookman, 2004.

(HERSENT, 2002) HERSENT, O.; GUIDE, D.; PETIT, J. P. **Telefonia IP**: Comunicação multimídia baseada em pacotes. São Paulo: Addison Wesley, 2002.

(HERSENT, 2005) HERSENT, O.; GUIDE, D.; PETIT, J. P. ***Beyond VoIP Protocols*** - *Understanding Voice Technology and Networking Techniques for IP Telephony*. John Wiley & Sons Ltd. Chichester, 2005.

(HSU, 2006) HSU, Hwei P. **Teoria e problemas de comunicação analógica e digital**. 2 ed. Porto Alegre: Bookman, 2006. 340 p.

(IETF, 1980) IETF. ***User Datagram Protocol***. Internet Engineering Task Force, 1980. (RFC768)

(IETF, 1981) IETF. ***Transmission Control Protocol***. Internet Engineering Task Force, 1980. (RFC793)

(IETF, 2002) IETF. ***Payload for Comfort Noise***. Internet Engineering Task Force, 2002. (RFC3389).

(IETF, 2003a) IETF. ***Real time protocol***. Internet Engineering Task Force, 2003. (RFC3550).

(IETF, 2003b) IETF. ***Real Time Control Protocol (RTCP)***. Internet Engineering Task Force, 2003. (RFC3605).

(IETF, 2003c) IETF. ***RTP Control Protocol Extended Reports (RTCP XR)***. Internet Engineering Task Force, 2003. (RFC3611).

(ITU, 1988a) ITU-T. ***Echo suppressors***. International Telecommunications Union, 1988. (ITU-T Recommendation G.164.0).

(ITU, 1988b) ITU-T. ***Pulse code modulation (PCM) of voice frequencies***. International Telecommunications Union, 1988. (ITU-T Recommendation G.711.0).

(ITU, 1988c) ITU-T. ***7 kHz audio-coding within 64 kbit/s***. International Telecommunications Union, 1988. (ITU-T Recommendation G.722.0).

(ITU, 1992) ITU-T. ***Coding of speech at 16 kbit/s using low-delay code excited linear prediction***. International Telecommunications Union, 1992. (ITU-T Recommendation G.728.0).

(ITU, 1993) ITU-T. ***Objective measurement of active speech level***. International Telecommunications Union, 1993. (ITU-T Recommendation P.56.0).

(ITU, 1993a) ITU-T. ***Echo cancellers***. International Telecommunications Union, 1993. (ITU-T Recommendation G.165.0).

(ITU, 1993b) ITU-T. ***Acoustic echo controllers***. International Telecommunications Union, 1993. (ITU-T Recommendation G.167.0).

(ITU, 1996a) ITU-T. ***Methods for subjective determination of transmission quality***. International Telecommunications Union, 1996. (ITU-T Recommendation P.800.0).

(ITU, 1996b) ITU-T. ***Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s***. International Telecommunications Union, 1996. (ITU-T Recommendation G.723.1).

(ITU, 1996c) ITU-T. ***Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction***. International Telecommunications Union, 1996. (ITU-T Recommendation G.729.0).

(ITU, 1996d) ITU-T. ***A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70***. International Telecommunications Union, 1996. (ITU-T Recommendation G.729.0 Annex B).

(ITU, 1996e) ITU-T. ***Reduced complexity 8 kbit/s CS-ACELP speech codec***. International Telecommunications Union, 1996. (ITU-T Recommendation G.729.0 Annex A).

(ITU, 1996f) ITU-T. ***Subjective performance assessment of telephone-band and wideband digital codecs***. International Telecommunications Union, 1996. (ITU-T Recommendation P.830).

(ITU, 1997a) ITU-T. ***Digital network echo cancellers***. International Telecommunications Union, 1997. (ITU-T Recommendation G.168.0).

(ITU, 1997b) ITU-T. ***Determination of sensitivity/frequency characteristics of local telephone systems***. International Telecommunications Union, 1997. (ITU-T Recommendation P.64.0).

(ITU, 1998) ITU-T. ***Objective quality measurement of telephone-band (300-3400 Hz) speech codecs***. International Telecommunications Union, 1998. (ITU-T Recommendation P.861).

(ITU, 2001) ITU-T. ***Perceptual Evaluation of Speech Quality***. International Telecommunications Union, 1999. (ITU-T Recommendation P.862.0).

(JIANG, 2000) JIANG, W; SCHULZRINE, H. ***Analysis of on-off Patters in VoIP and Their Effect on Voice Traffic Aggregation***. Columbia University, 2000.

(KAHRS, 1998) KAHRS, Mark; BRANDENBUG, Karlheinz. ***Applications of Digital Signal Processing to Audio and Acoustics***. New Jersey: Kluwer Academic Publishers, 1998.

(KONDOZ, 2000) KONDOZ, A.M.; EVANS, B.G. ***A High Quality Voice Coder With Integrates Echo Canceller and Voice Activity Detector for VSAT Systems***. Center for Satellite Engeneering Research, University of Surrey, 2000.

(KUROSE, 2003) KUROSE, James F.; ROSS, Keith W. **Redes de computadores e a Internet**: uma nova abordagem. 1. ed. São Paulo: Addison Wesley, 2003.

(LARMAN, 2004) LARMAN, C. **Utilizando UML e padrões**: uma introdução à análise e ao projeto orientados a objetos e ao Processo Unificado. 2 ed. Porto Alegre: Bookman, 2004.

(LYONS, 2004) LYONS, Richard. G. ***Understanding digital signal processing***. 2 ed. Nova Jersey: Pearson Education, 2004.

(MADUREIRA, 2003) MADUREIRA, L. **Delphi 6**. Porto Alegre: SENACRS, 2003.

(MELLO, 2003) MELLO, R. N. B. **Estudo comparativo da transformada karhunen-loève na compressão de imagens**. 2003. Dissertação (Programa de Pós-Graduação em Engenharia Elétrica) - UFRGS, Porto Alegre, 2003.

(MICROSOFT, 2000) MICROSOFT Corporation. **Dicionário prático de informática**. Lisboa, Portugal: McGraw-Hill, 2000.

(MONTEIRO, 2002) MONTEIRO, R. F.; ERRICO, L.; YEHIA, H. C. **Implementação de Transporte Robusto de Voz em Redes Baseadas em Protocolos IP**. XVIII SBRC. Belo Horizonte, 2002.

(NAKASHIMA, 2003) NAKASHIMA, G. Y. **Aplicação do filtro de Wiener para tratamento de sinais eletromiográficos**. 2003. Dissertação (Mestrado em Bioengenharia) - USP, São Carlos, 2003.

(NASCIMENTO, 2004) NASCIMENTO, F. A. O. **Algoritmo para Criptografia de Voz Implementado em Tempo Real em Processador de Sinais**. Itajaí, p. 30-34, 2004.

(OHRTMAN, 2004) OHRTMAN, Frank. ***Voice Over 802.11***. Norwood, MA: Artech House, 2004.

(OPPENHEIM, 1975) OPPENHEIM, A. V.; SCHAFER, R. W. ***Digital Signal Processing***. New Jersey: Prentice-Hall, 1975.

(PERCY , 2005) PERCY, Alan. ***Understanding Latency in IP Telephony***. Disponível em: <www.telephonyworld.com/training/brooktrout/iptel_latency_wp.html>. Acesso em 08/04/2009.

(PRASAD, 2002) PRASAD, R.V. et al. ***Comparation of Voice Activity Detection Algorithms for VoIP***. IEEE, Bangalore, 2002.

(RABINER, 1978) RABINER, L., R.; SCHAFER, R. W. ***Digital Processing of Speech Signals***. New Jersey: Prentice-Hall, 1978.

(RAMIREZ, 2002) RAMIREZ, J. et al. ***Efficient Voice Activity Detection Algoritms Using Long-term Speech Information*.** Dept. of Eletronics and Computer Tecnology, University of Granada, 2002.

(RANGANATHAN, 2005) RANGANATHAN, M. K.; KILMARTIN, L. ***Neural and Fuzzy Computation Techmiques for Playout Delay Adaptation in VoIP Networks***. IEEE transanctions on Neural Networks. v. 16, n. 5, 2005.

(REDDING, 2001) REDDING, C.; DEMINCO, N.; LINDNER, J. ***Voice Quality Assessment of Vocoders in Tandem Configuration***. U. S. Department of commerce. NTIA Report 01-386. Disponível em <http://www.its.bldrdoc.gov/pub/ntia-rpt/01-386/01-386.pdf> , último acesso em 10/04/2009. 2001.

(RENEVEY, 2001) RENEVEY, P; DRYGAJLO, A. ***Entropy Based Voice Activity Detection in Very Noisy Conditions***. European Conference on Speech Communication and Technology. Aalborg, Denmark, v. 3, p. 1883-1886, 2001.

(REYNOLDS, 2001) REYNOLDS, R. J. B.; RIX, A. W. ***Quality VoIP*** - *an engineering challenge*. BT Technology Journal. MA, USA, 2001.

(RIX, 2000) RIX, A. W.; HOLLIER, M. P. ***The perceptual analysis measurement system for robust end-to-endspeech quality assessment***. IEEE International Conference. Istanbul, Turkey, 2000.

(RODRIGUES, 1988) RODRIGUES, S. L**. Implementação e Avaliação do Desempenho de um Sistema Automático de Reconhecimento de Locutor pela Análise de Frases Curtas**. 1988. Tese (Mestrado em Engenharia) - IME - Instituto Militar de Engenharia, Rio de Janeiro, 1988.

(ROSE, 2007) ROSE, L. **A ética na Internet** - Anonimato e impunidade, liberdade e censura. XXX Congresso Brasileiro de Ciências da Comunicação. Santos, SP, 2007.

(ROSEMBERG, 1998) ROSEMBERG, J.; SCHULZRINNE, H. ***Internet telephony gateway location***. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. São Francisco, 1998.

(SANGWAN, 2002a) SANGWAN, A. et al. ***VAD Techniques for Real-Time Speech Transmission on the Internet***. IEEE International Conference. Jeju Island, S. Korea, 2002.

(SANGWAN, 2002b) SANGWAN, A. et al. ***Second and Third Order Adaptable Threshold for VAD in VoIP***. Beijing, China, 2002.

(SANTOS, 2006) SANTOS, M. N. **Medidas de qualidade de voz em redes IP**. 2006. Dissertação (Mestrado, Programa de Pós-Graduação em Engenharia Elétrica) - Setor de Tecnologia, UFPR, 2006.

(SCHULZRINNE, 1996) SCHULZRINNE, H. et al. **RTP**: *a transport protocol for real-time applications*. Internet Engineering Task Force, 1996. (RFC 1889)

(SHENOI, 1995) SHENOI, Kishan. ***Digital signal processing in telecommunications***. Nova Jersey: Prendice Hall , 1995.

(SMITH, 1997) SMITH, S. W. ***The Scientist and Engineer's Guide to Digital Signal Processing.*** Internet Technical Publishing, 1997. Disponível em < http://www.dspguide.com/>. Acesso em 25 de março de 2009.

(SONNINO, 2004) SONNINO, B. **Profiling na prática** - Otimize a performance de sua aplicação. Clube Delphi. 51 ed., ano IV, Neofício Editora, 2004.

(STRUM, 1988) STRUM, R.; KIRK, D. ***First Principles of Discrete System and Digital Signal Processing***. Nova Iorque: Addison-Wesley Publishing Company, 1988.

(TAGUCHI, 2003) Taguchi, A. ***Residual-Excited Linear Predictive (RELP) Vocoder system with TMS320c6711 dsk and vowel characterization.*** Dissertação (Mestrado. Departmento de Engenharia Elétrica) Universidade de Saskatchewan, Canada, 2003.

(TANYER, 1998) TANYER, S.G.; ÖZER, H. ***Voice activity Detection in Nonstationary Gaussian Noise***. Island of Rhodes, Greece, 1998.

(TANYER, 2000) TANYER, S.G.; ÖZER, H. ***Voice activity Detection in Nonstationary Gaussian Noise***. v. 8, n. 4, 2000.

(TENENBAUM, 2003) TENENBAUM, A. S. **Redes de computadores**. Rio de Janeiro: Elsevier, 2003.

(VASEGUI, 2000) VASEGUI, S. V. ***Advanced digital signal processing and noise reduction.*** 2 ed. Nova Iorque: John Wiley & Sons Ltd, 2000.

(VENDRUSCULO, 2005) VENDRUSCULO, T. **Pesquisa e desenvolvimento da Aplicação AL2G com implementação de técnicas específicas para Localização Otimizada de**

Gateways em Serviços de Telefonia IP. 2005. Dissertação (Mestrado em Engenharia Elétrica, Faculdade de Engenharia) - PUCRS, 2005.

(WAN, 1993) WAN, E. ***Time Series Prediction Using a Neural NetworkWith Embedded Tapped Delay-Lines***. MA: Addison Wesley, 1993.

(YAMADA, 2000) YAMADA, T. ***Voice Activity Detection in Noisy Environments***. University of Tsukuba, 2000.

(YOUNG, 2006) YOUNG, P. H. **Técnicas de comunicação eletrônica.** 5 ed. São Paulo: Pearson Prendice Hall, 2006.

(ZHA, 2005) ZHA, Wei; CHAN, Chan Wai-Yip. ***Objective Speech Quality Measurement Using Statistical Data Mining***. *EURASIP Journal on Applied Signal Processing*, Issue 9, p. 1410-1424, 2005.

(ZHENG, 2001) ZHENG, L.; ZHANG, L.; XU, D. ***Caracteristics of network delay and delay jitter and its effect on voice over IP (VoIP)***. Helsinki, Finland. 2001.

(ZWICKER, 1961) ZWICKER, E. ***Subdivision of the audible frequency range into critical bands***. *The Journal of the Acoustical Society of America*, Feb. 1961.

APÊNDICE A – Modelagem das técnicas de supressão de silêncio

Neste apêndice A do trabalho, são apresentados todos os diagramas desenvolvidos para a modelagem das técnicas de detecção e supressão de silêncio construídas. São aqui apresentados o diagrama de caso de uso, os diagramas de estados e o diagrama de classes, conforme segue:

a) Diagrama de casos de uso (Figura 34).

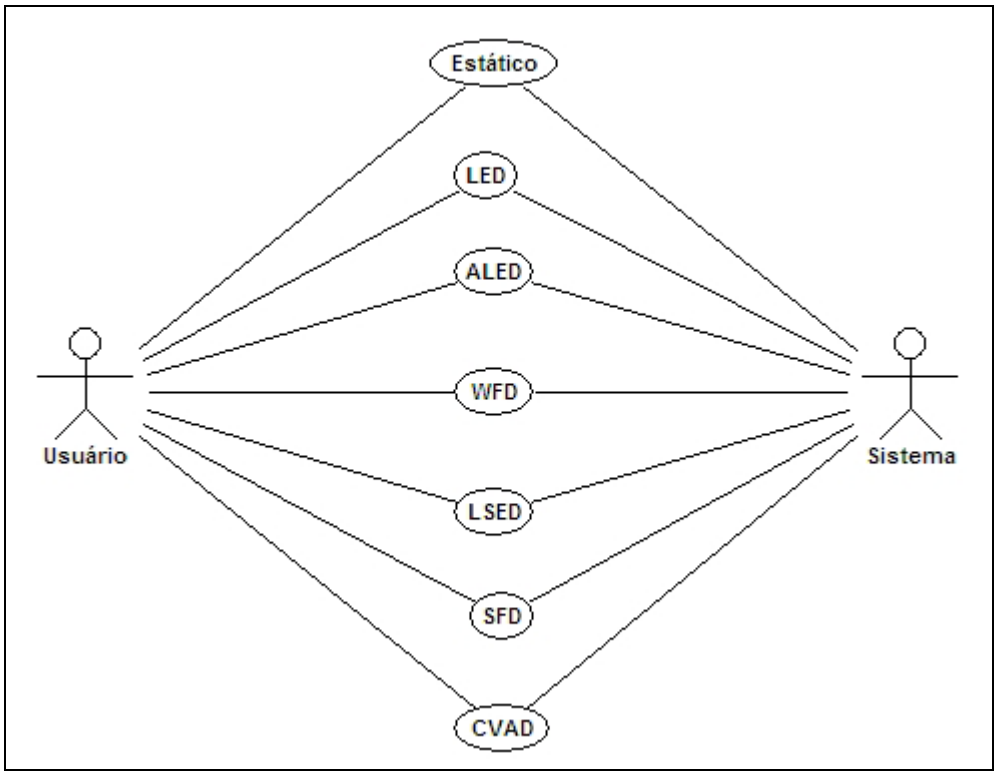


Figura 34 - Diagrama de casos de uso
Fonte: O autor (2009).

b) Diagrama de estados para o supressor estático (Figura 35)

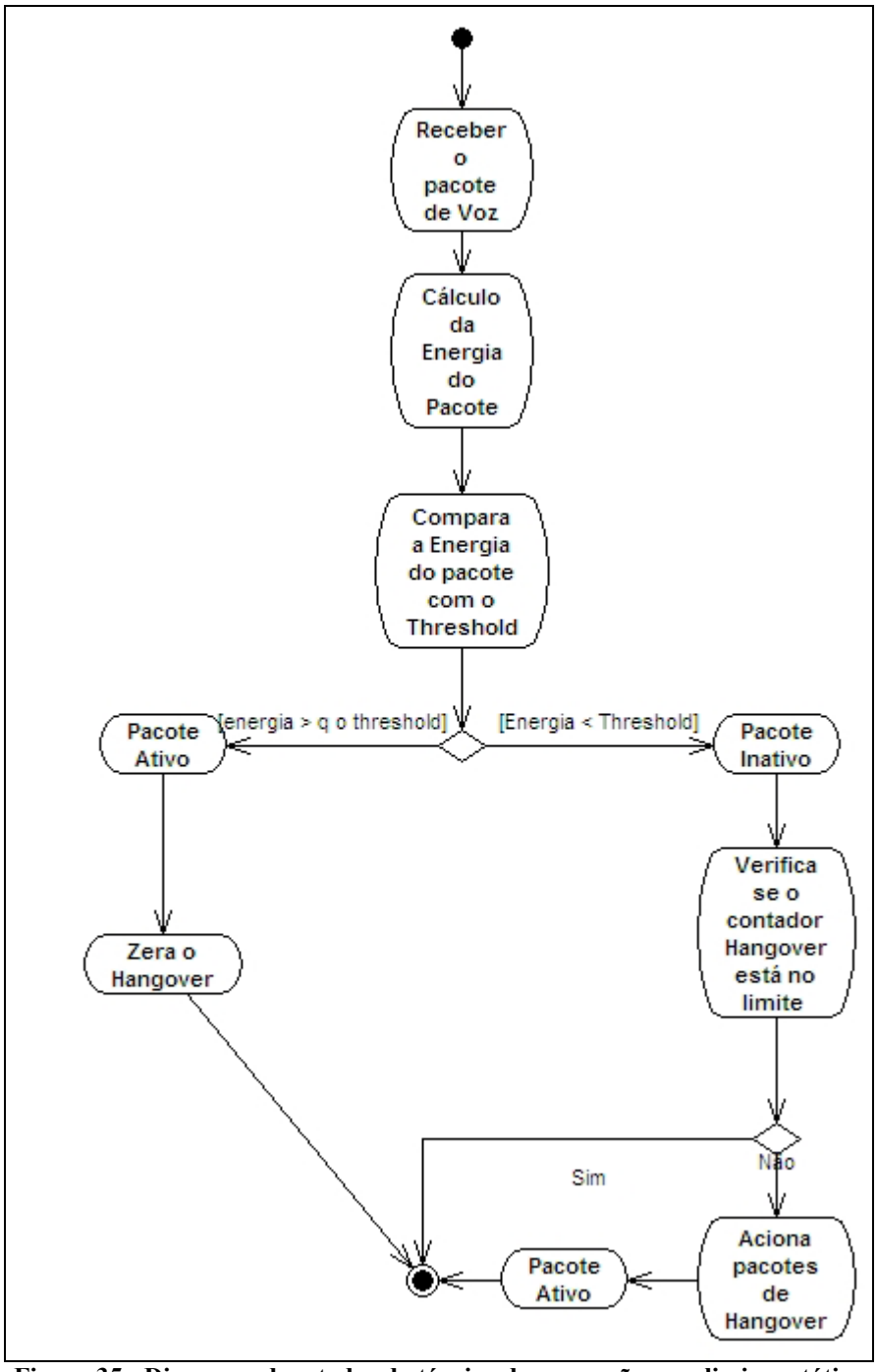


Figura 35 - Diagrama de estados da técnica de supressão com limiar estático
Fonte: O autor (2009).

c) Diagrama de estados para o LED (Figura 36)

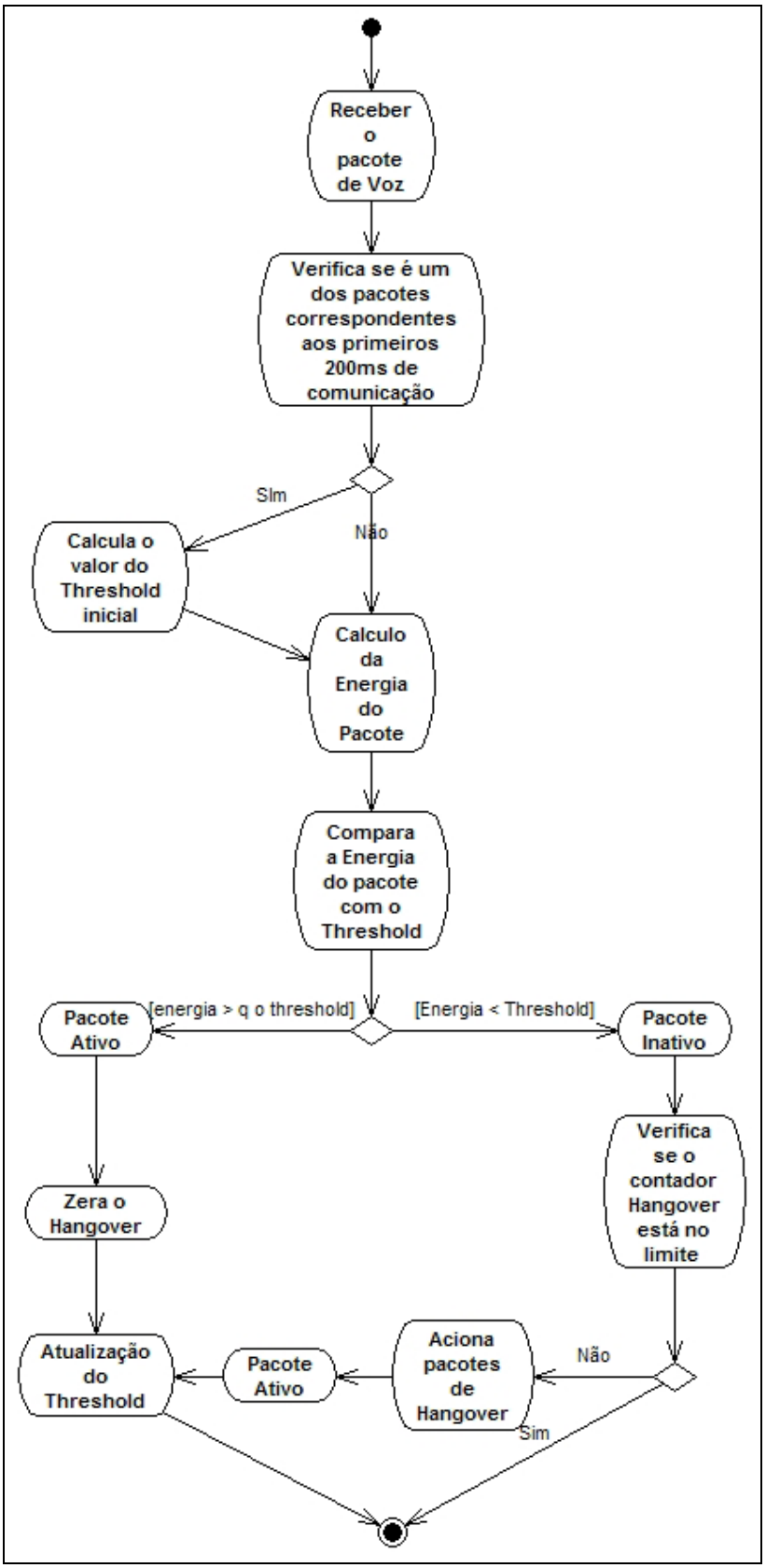


Figura 36 - Diagrama de estados da técnica LED
Fonte: O autor (2009).

d) Diagrama de estados para o ALED (Figura 37).

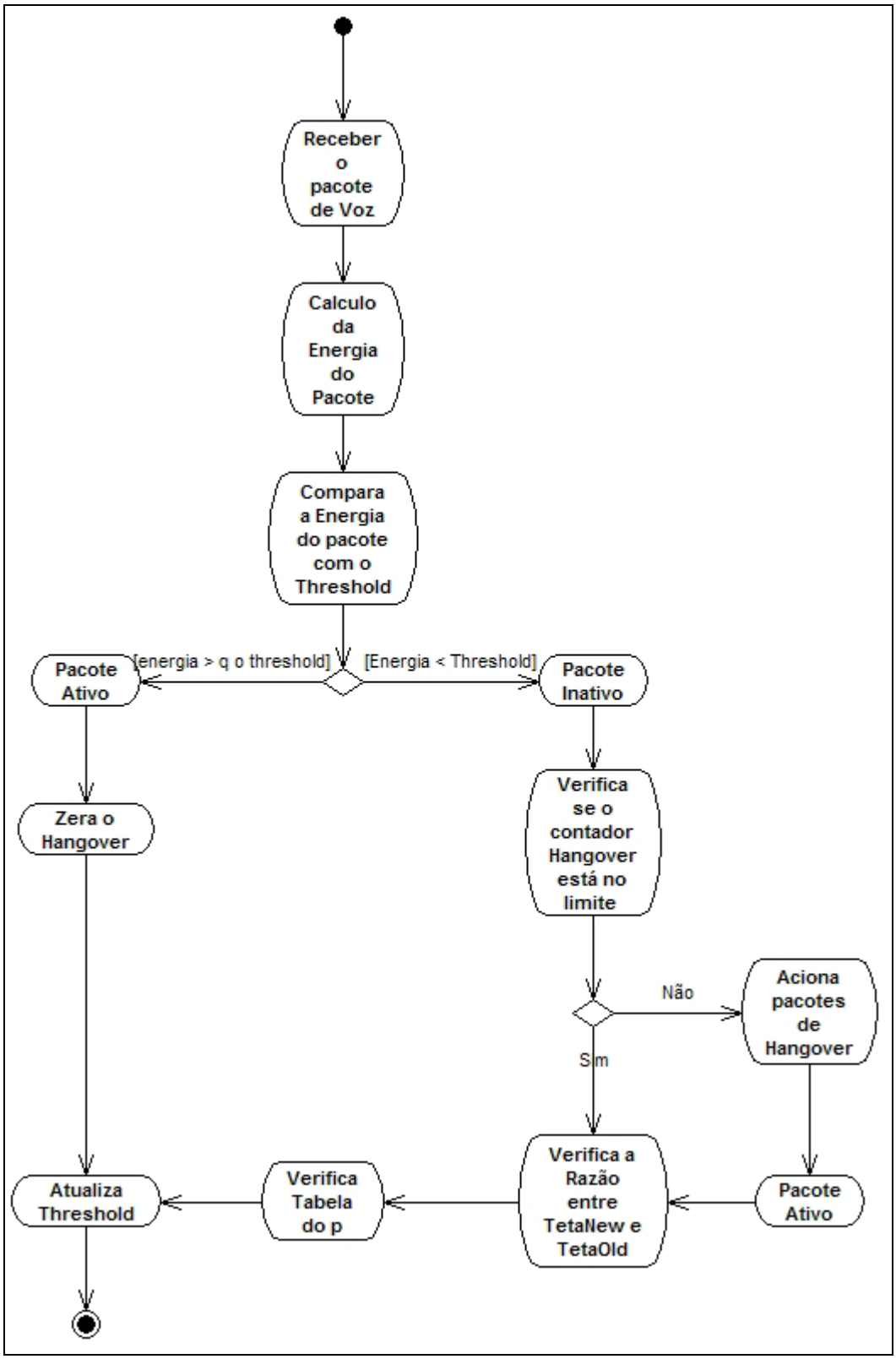


Figura 37 - Diagrama de estados da técnica ALED
Fonte: O autor (2009)

e) Diagrama de estados para o WFD (Figura 38).

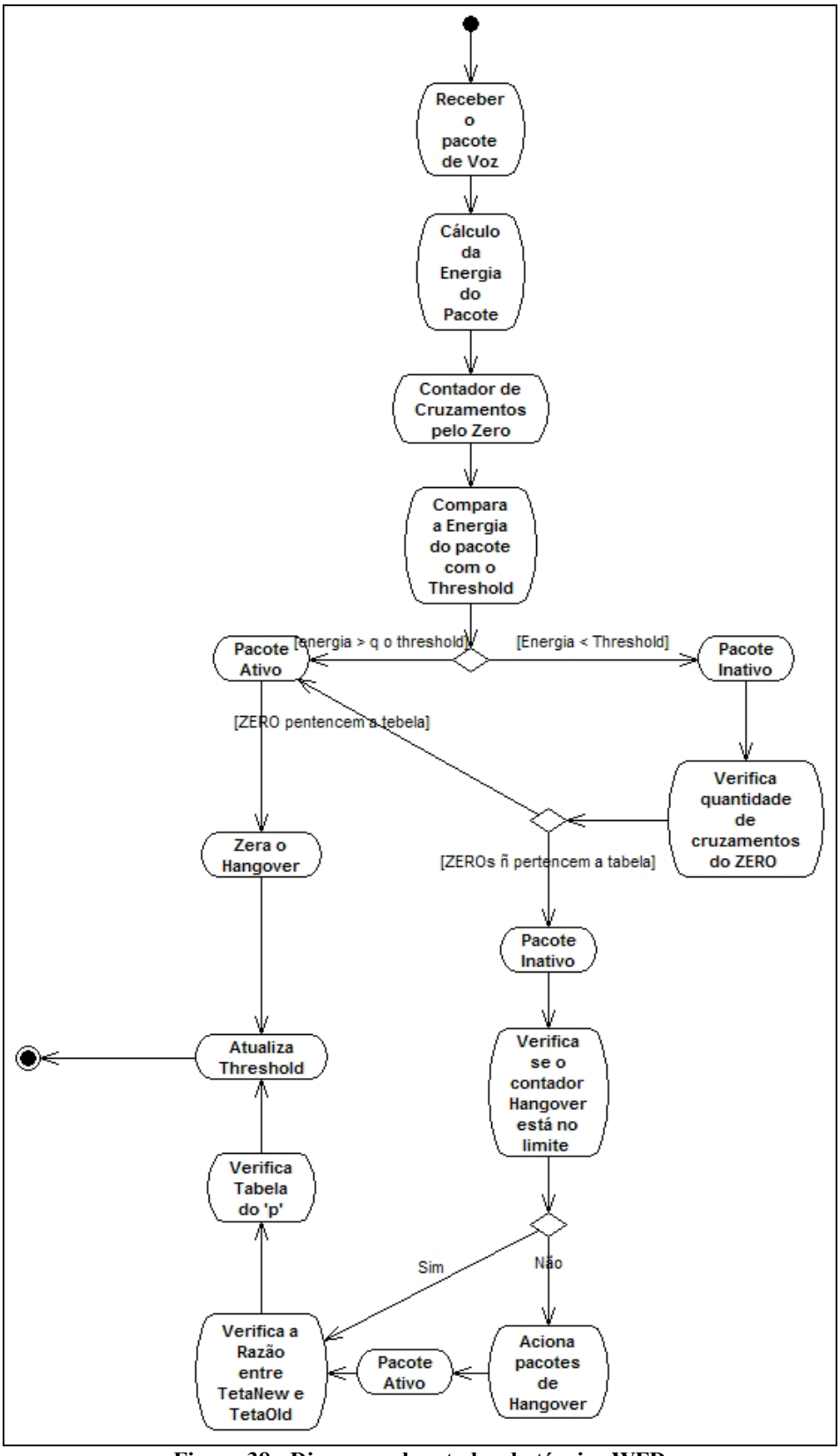


Figura 38 - Diagrama de estados da técnica WFD
Fonte: O autor (2009)

f) Diagrama de estados para o LSED (Figura 39).

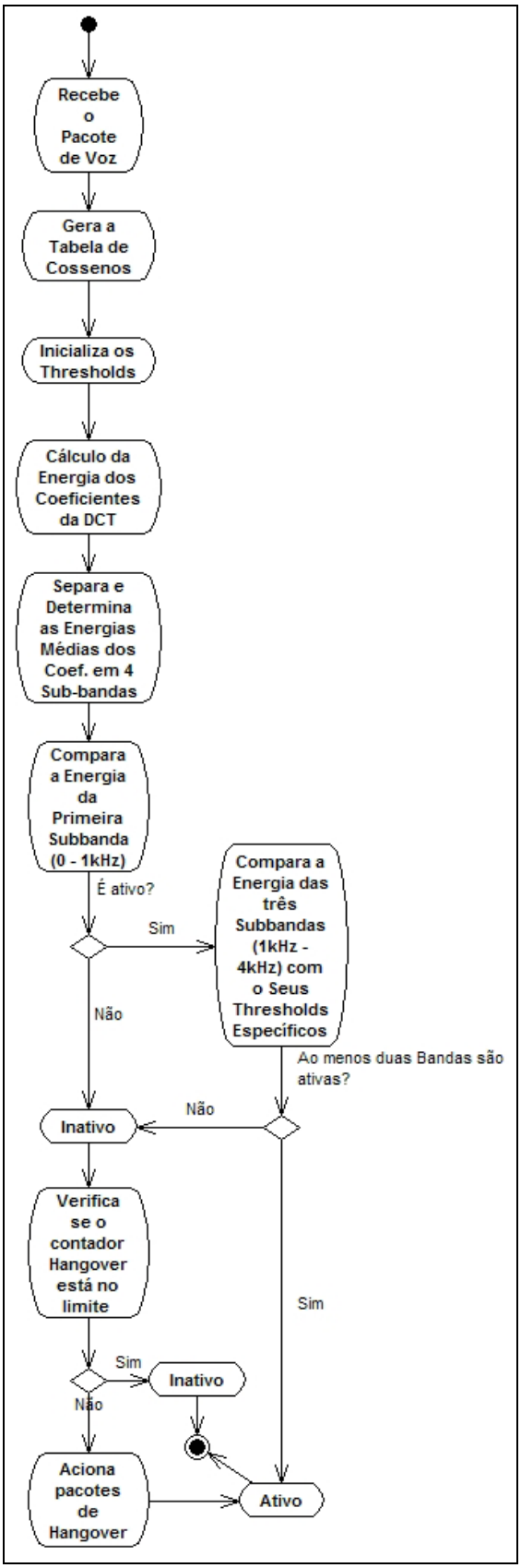


Figura 39 - Diagrama de estados da técnica LSED
Fonte: O autor (2009)

g) Diagrama de estados para o SFD (Figura 40)

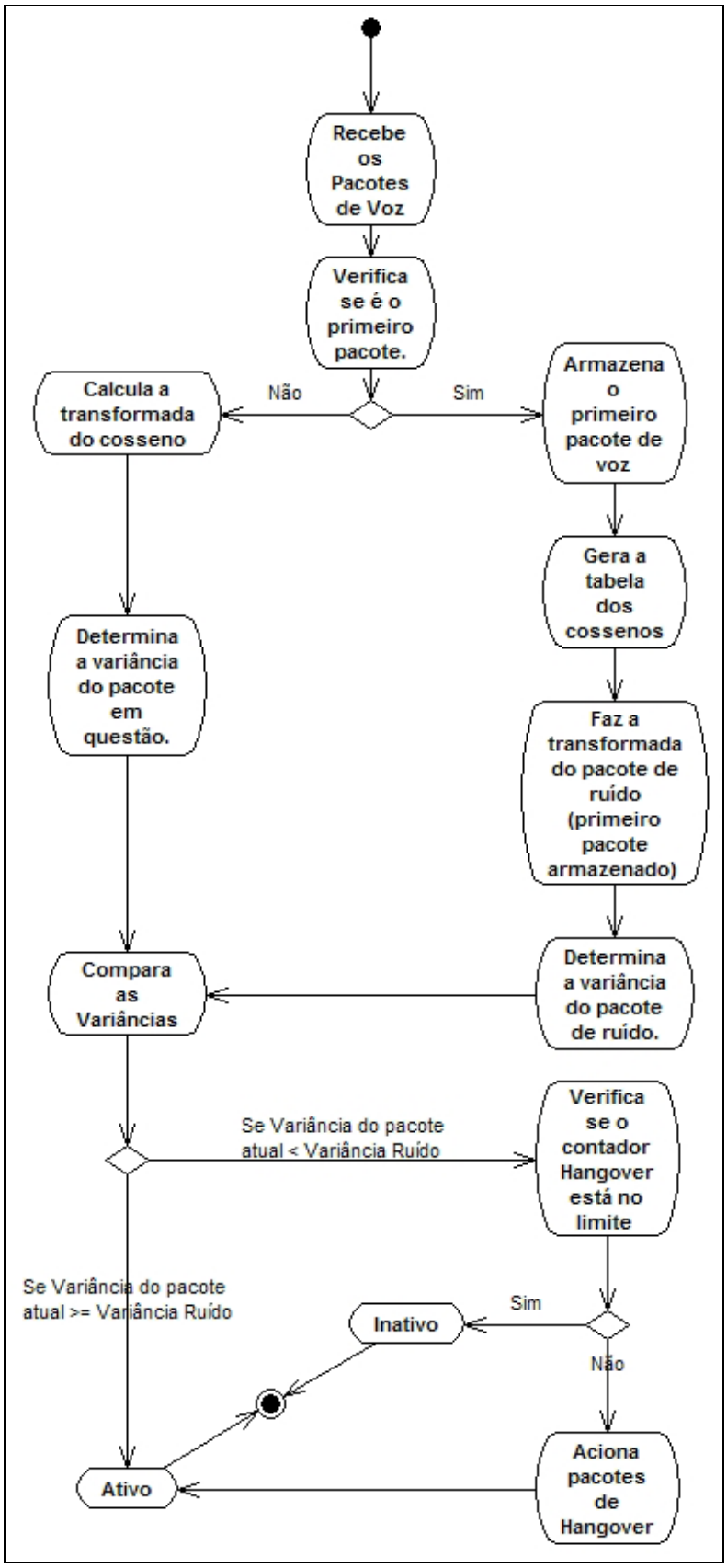


Figura 40 - Diagrama de estados da técnica SFD
Fonte: O autor (2009)

h) Diagrama de estados para o CVAD (Figura 41).

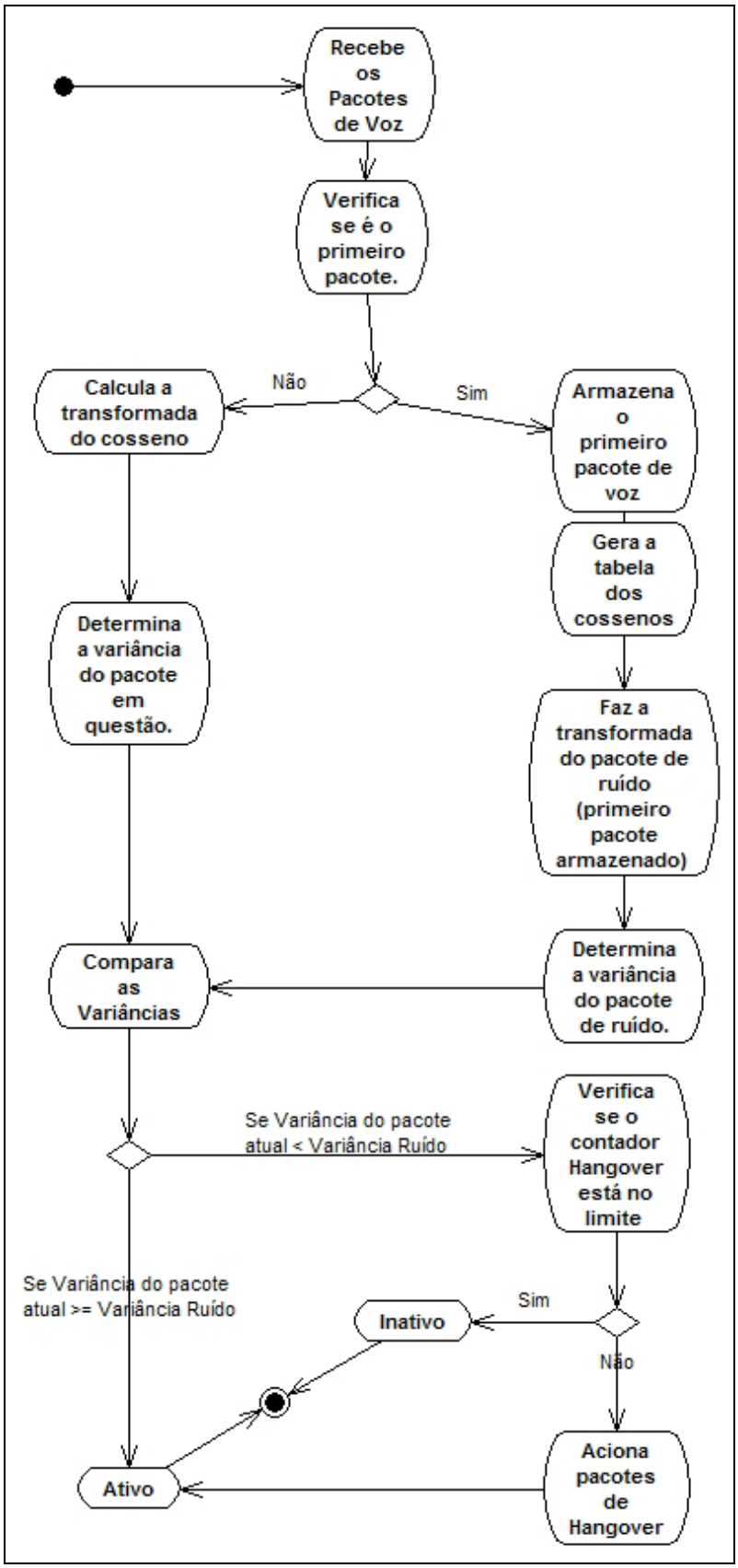


Figura 41 - Diagrama de estados da técnica CVAD
Fonte: O autor (2009)

i) Diagrama de classes para do conjunto de técnicas implementadas e usadas junto às ferramentas de desenvolvimento e teste *Wave Silence Suppression* e *Silence Suppression Tester* (Figura 42).

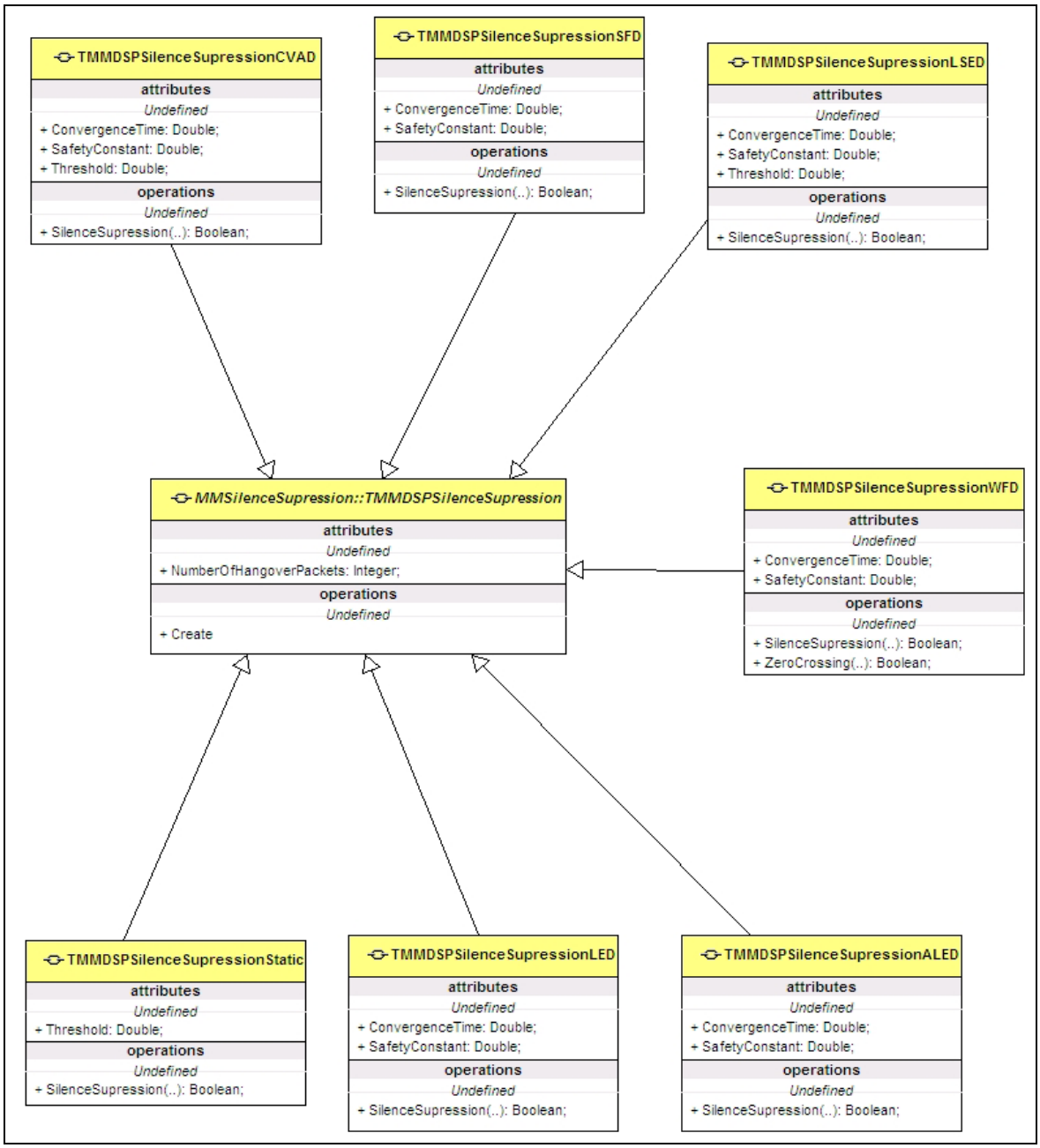


Figura 42 - Diagrama de classes das técnicas de supressão de silêncio
Fonte: O autor (2009)

APÊNDICE B – Modelagem do gerador de ruído de conforto

Neste apêndice B do trabalho, é apresentado o diagrama de classes das técnicas de geração de ruído de conforto (Figura 43) implementadas. Três classes com diferentes técnicas foram implementas, mas apenas uma foi usada na prática. Isso porque, em um primeiro momento, não foram observadas diferenças significativas entre os algoritmos implementados. O que levou a abreviar os testes quanto a geração de ruído, até por não ser o objetivo fim deste trabalho.

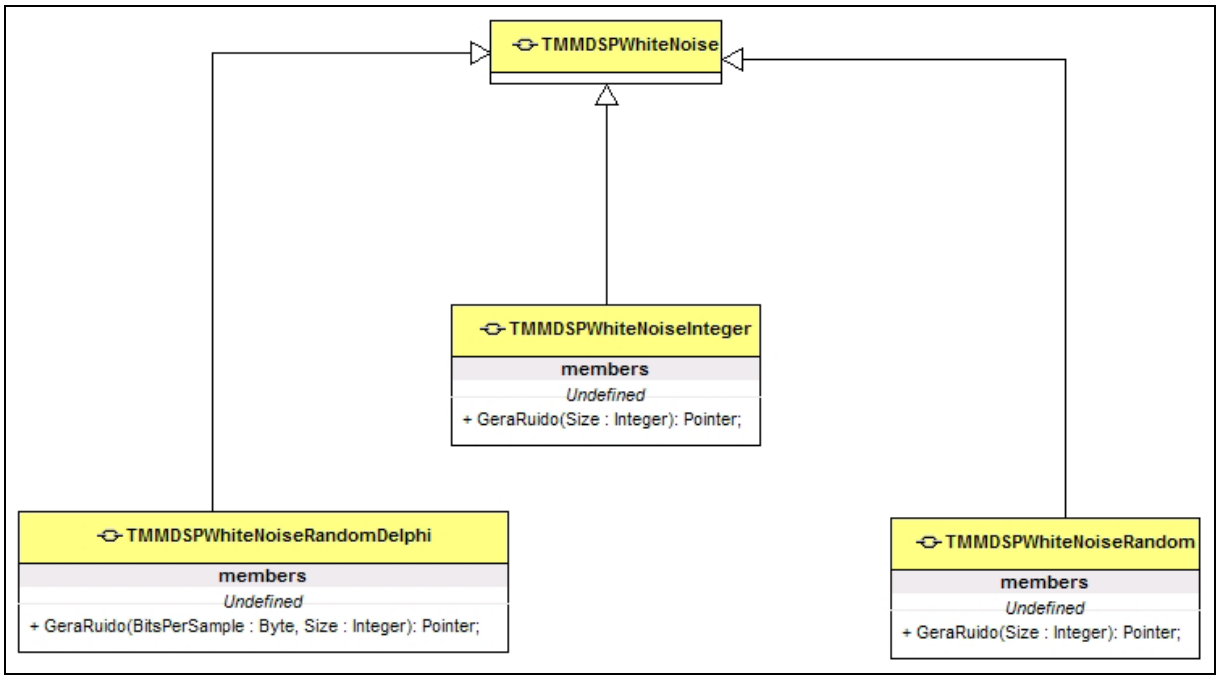


Figura 43 - Diagrama de classes das técnicas de geração de ruído de conforto
Fonte: O autor (2009)

Os demais diagramas, casos de uso e estados, não são apresentados devido a simplicidade de cada um deles, o que faz com que não agreguem informação significativa para o entendimento do trabalho.

APÊNDICE C – *Buffer* de *Dejitter*

A buferização dos frames de voz procura, primordialmente, a eliminação do efeito causado pelo *jitter*[21] na transmissão (os quais vão desde pacotes dados como perdidos até a inserção de intervalos na reprodução do áudio, com um efeito bastante comprometedor em termos de qualidade de áudio percebido pelo ouvinte) (ZHENG, 2001). Essas filas são conhecidas como *playout adaptation buffers*. Contudo, a utilização de filas para a compensação do *jitter* também resulta em sistemas que, com o uso de mecanismos de compensação e correção de erros acoplados, permitem a criação de um sistema com perdas zero, à custa da introdução de atraso (BALBINOT, 2002).

Conforme Fernandes (2003) os elementos de atraso podem ser agrupados em três grupos:

- Atraso de transmissão (TX): Formação dos quadros de voz e processamento para codificação dos quadros de voz;
- Atraso de rede: Serialização dos pacotes IP e propagação na rede (formada pela propagação através dos meios de comunicação que formam a rede e pelos tempos de enfileiramentos nos nós de rede);
- Atraso de recepção (RX): Atraso gerado pelo *buffer* de *dejitter* e tempo de processamento para decodificação dos quadros de voz.

O dimensionamento do *buffer* para a compensação do *jitter* e para a eliminação da percepção de perda de pacotes devido a esse *jitter* está diretamente relacionado aos atrasos observados comumente em redes IP (FERNANDES, 2003).

Entre os diversos mecanismos existentes para o correto dimensionamento do buffer, são de particular interesse todos aqueles que possibilitem um redimensionamento dinâmico do mesmo em razão das condições da rede. A utilização conjunta desses mecanismos com a realimentação provida pelo RTCP (RTCP - *Real-Time Transport Control Protocol*) (EITF, 2003b), particularmente aquela observada em seus relatórios (os quais, pelo processamento dos dados, permitem inclusive a determinação do *throughput*[22] médio disponível no receptor) (SCHULZRINNE, 1996).

---

[21] *Jitter* é uma variação estatística do retardo na entrega de dados em uma rede, ou seja, pode ser definida como a medida de variação do atraso entre os pacotes sucessivos de dados (BARCELOS, 2005).

[22] *Throughput* é a medida de velocidade de transferência de dados empregue num sistema de comunicação complexo; ou medida da velocidade de processamento de dados num sistema de computador (MICROSOFT, 2000).

Assim, objetivando contribuir para minimizar estes problemas junto a aplicação, implementou-se um *buffer* de *dejitter* com a função de sincronização do tempo de reprodução dos pacotes de áudio recebidos do locutor, a reordenação dos pacotes quando da chegada fora de ordem, e também permitir ao sistema receptor gerenciar determinadas perdas de pacotes da rede, a fim de minimizar a degradação da qualidade do áudio para o ouvinte.

A Figura 44, de forma ilustrativa, apresenta o gerenciamento de reordenamento e sequencimaneto no tempo dos pacotes pelo *buffer* de *dejitter*. Primeiro os pacotes ordenados são enviados pelo *host* para a rede, sendo na sequência os mesmos desordenados e sujeitos a variação de atraso (diferente dos 20 ms iniciais) na sua ordenação. Por fim, os pacotes são reordenados e sequenciados dentro do *buffer* de *degitter* via identificação do *sequence number* e do *timestamp* do protocolo RTP (IETF, 2003a) (TENENBAUM, 2003).

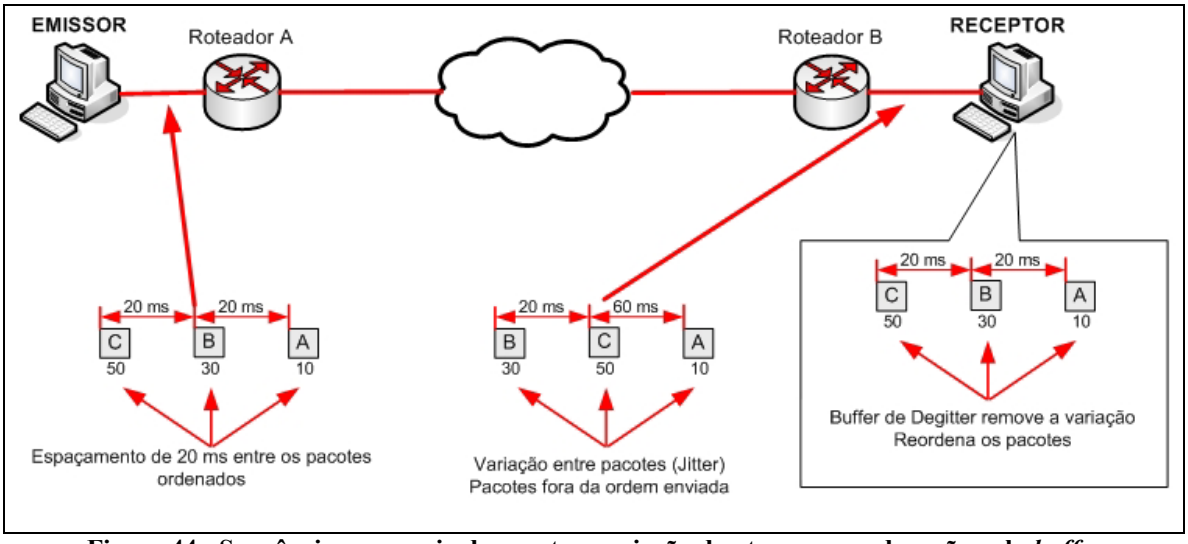


Figura 44 - Sequência com envio de pacotes, variação do atraso e reordenação pelo *buffer*
Fonte: O autor (2005).

Basicamente, o *buffer* de *dejitter* implementado apresenta via interface, o reordenamento de valores que representam experimentalmente o *sequence number* dos pacotes RTP (IETF, 2003a) (TENENBAUM, 2003) da aplicação de VoIP.

Quanto a variações de atraso, o *buffer* analisa o *Timestamp* do pacote RTP (IETF, 2003a) (TENENBAUM, 2003), a fim de remover essas variações introduzidas pela rede e realizar uma reprodução sincronizada.

O que pode-se vislumbrar aqui é também que esse reordenamento e remoção da variação do atraso só são possíveis enquanto essas variáveis estiverem dentro do limite de tempo do próprio *buffer*. Em geral para aplicações de VoIP, os *buffer* de recepção mais robustos costumam ter seu tamanho (quantidade de pacotes gerenciáveis) adaptativo, mas tendo o mesmo uma limitação proporcional a 80 ms de áudio em média. Isso significa que, caso o *buffer* tenha 80 ms de tamanho, e como no exemplo da Figura 44, o pacote *A* tenha

chegado e tenha sido reproduzido, o pacote *C* também tenha chegado, mas o pacote *B* ainda não, o mesmo terá no máximo 80 ms para chegar antes que o pacote *C* seja jogado para frente e tenha os seus dados reproduzidos. Neste caso, mesmo que o pacote *B* venha a chegar depois desses 80 ms, ele será descartado por estar demasiadamente atrasado. Caso não ocorresse o descarte, o mesmo seria reproduzido fora da ordem, acabando por degradar de forma mais significativa a qualidade do áudio reproduzido.

ANEXO A – RECOMENDAÇÃO P.800 (MOS)

Tratando-se de sistemas de voz, seja no uso de telefonia convencional ou de voz digitalizada trafegando por uma rede IP, a identificação da qualidade da fala é importante para mensurar o quão satisfatório é o sistema. Quanto mais próximos os sinais de entrada e saída do sistema, melhor a qualidade do serviço observado (FERNANDES, 2003).

Para as técnicas de detecção e supressão de silêncio não existe nenhuma técnica que possa fazer uma avaliação específica e isolada das mesmas. Para este trabalho foi pensado que o emprego de determinadas técnicas de avaliação sobre os algoritmos de VAD seria uma forma de certificação, ou não, da correta implementação dos mesmos, e ou da verificação, ou não, de características desejadas para o sistema. A grande questão pensada foi em como fazer isto.

Considerando a existência de diversos métodos para avaliação da qualidade da voz disponíveis pelo mercado, sendo alguns deles recomendações do ITU-T, optou-se por buscar algo que pudesse ser adaptado para a presente situação. Dentre uma série de técnicas de avaliação, o escolhido foi o de avaliação subjetiva da voz pelo teste de avaliação por escuta da recomendação P.800 (ITU 1996). Como citado por Fernandes (2003) neste tipo de teste não é esperado obter o mesmo padrão de realismo como o alcançado no teste por conversação desta mesma recomendação P.800.

Na seqüência deste Anexo A é feita a descrição de partes da recomendação P.800, especificamente no que é de interesse deste trabalho.

P.800

Sumário

Esta recomendação descreve métodos e procedimentos que permitem uma avaliação subjetiva da qualidade da transmissão telefônica. De forma corriqueira é conhecida como *Mean Opinion Score* (MOS), mas é importante salientar que esta é apenas uma das formas de pontuação citadas nesta recomendação.

Origem

A Recomendação P.800 foi revisada pelo Grupo de Estudos 12 do ITU-T, entre 1993 e 1996, sendo aprovada em 30 de agosto de 1996 pela resolução n° 1 do *World Telecommunication Standardization Conference* (WTSC).

Escopo

A P.800 contém sugestões para condução de testes subjetivos de qualidade de transmissão em laboratório. Provê métodos considerados convenientes para determinarem o quão satisfatório é o desempenho de dada conexão telefônica.

Entende-se como avaliação subjetiva, aquela que usa procedimentos de conversação ou apenas escuta como métodos para testes, com o fim de aferir a qualidade da transmissão por equipamentos ou serviços de telecomunicações.

Os métodos indicados têm aplicações genéricas, qualquer que sejam os fatores presentes de degradação. Exemplos desses fatores são: perda, ruído de circuito, erros de transmissão, ruído de ambiente, eco, distorção não linear, tempo de propagação, etc. Combinações de dois ou mais desses fatores também são considerados.

Método Recomendado - Teste de Opinião por Escuta

Neste tipo de teste não é esperado obter o mesmo padrão de realismo como o alcançado em testes de conversação. O método de teste mais recomendado para opinião em escuta é a classificação por categoria absoluta (ACR - *Absolute Category Rating*), sendo bastante estável e com aplicação em conexões telefônicas analógicas e digitais.

O teste de escuta tem uso direto na qualificação de sistemas de transmissão que sejam essencialmente unidirecionais. Os resultados obtidos por esse tipo de teste podem ser usados, com alguma reserva, na qualificação de conversações sobre sistemas bidirecionais, como a rede pública telefônica.

Método de teste ACR

Serão descritos os diversos procedimentos para gravação e escuta, que devem ser seguidos neste método.

A fim de eliminar variações indesejáveis na fonte da fala, as amostras devem ser obtidas respeitando-se alguns critérios:

i. Ambiente de gravação

A pessoa que irá falar deve estar sentada dentro de uma sala silenciosa, de volume entre 30 e 120 m^3 e com tempo de reverberação menor que 500 ms (preferencialmente entre 200 e 300 ms). O nível de ruído na sala deve estar abaixo de 30 dBA, sem picos dominantes no espectro.

ii. Sistema de gravação

O sistema de gravação deve ser de alta qualidade (semelhante aos de estúdios de gravação) e possuir um dos itens a seguir:

– Um gravador de fita convencional com duas trilhas e equalização fixa;

– Um processador digital de áudio de dois canais, com um gravador de alta qualidade ou *Digital Audio Tape* (DTA);

– Um sistema de armazenamento digital controlado por computador.

iii. Fala

O tipo de fala a ser usada consiste de sentenças curtas e simples, escolhidas ao acaso e que sejam de fácil entendimento. Podem ser retiradas de jornais ou literatura não técnica, por exemplo. Deve ser formada uma lista contendo sentenças, sem conexão óbvia entre elas. Cada sentença não pode ser muito curta ou longa em demasia, com tempo de pronúncia ideal entre dois e três segundos cada.

O responsável pelo experimento decide quantas sentenças são necessárias para formar cada grupo para amostra de fala. O mínimo de dois e máximo de cinco é o recomendado.

Grupos são combinados em listas com cinco ou dez grupos cada, de modo que a lista completa é usada como uma série de amostras sujeitas ao mesmo tratamento, mas com nível de escuta ou outros parâmetros alterados, enquanto a lista é pronunciada.

Exemplo de material de fala:

Não existe nada para ser visto;
Eu quero um minuto com o inspetor;
Ele precisa de dinheiro?

iv. Procedimentos de gravação

A fala deve ser gravada utilizando-se um microfone linear e um amplificador de baixo ruído, conforme especificado na publicação IEC 581-5. O microfone deve estar posicionado a uma distância entre 140 e 200 mm dos lábios. Em alguns casos faz-se necessário o uso de um anteparo “corta-sopro”, para que a respiração do orador não seja notada.

Dois sistemas de gravação são usados simultaneamente: um grava toda a faixa de freqüência da fala em um canal e o outro a fala em resposta do telefone, no canal correspondente. Este procedimento é necessário para o caso de ser preciso comparar as duas versões.

O nível da fala é definido pela recomendação P.56 (ITU 1993) e observado durante toda a gravação.

Para reduzir o risco de resultados dependentes de peculiaridades das vozes escolhidas para as falas, é essencial que mais de uma voz masculina e feminina sejam usadas de forma balanceada.

Da mesma forma que foram listados critérios relativos à fala, também deve-se observar procedimentos rígidos para a escuta, conforme listados a seguir:

i) Ambiente de escuta

Deve obedecer às mesmas condições da sala de gravação, atendendo também aos critérios de ruído ambiente.

ii) Sistema de escuta

O sistema de telefone local, sistema de alto-falante, etc., deve ser calibrado de acordo com a recomendação P.64 (ITU, 1997b). É recomendado que as características de sensibilidade de freqüência de recepção sejam medidas pelo menos duas vezes, no início e no fim de cada experimento. Qualquer variação significativa entre as medidas pode invalidar o experimento.

iii) Ouvintes

As pessoas escolhidas para os testes devem ser usuários de telefone, escolhidas ao acaso. Não podem estar envolvidas com atividades de medida de desempenho de sistemas telefônicos, ou trabalharem com assuntos relacionados à codificação de voz. Além disso, essas pessoas não podem ter participado de testes subjetivos há pelo menos 6 meses, não podem ter participado de testes de opinião por escuta há pelo menos um ano e por fim, nunca devam ter escutado a mesma lista de sentenças antes. Caso não seja possível atender ao descrito, isto deve ser registrado na conclusão dos resultados.

iv) Escalas de opinião recomendadas pelo ITU-T

Várias escalas de julgamento com cinco níveis podem ser empregadas, dependendo do seu propósito. A forma de apresentação e as palavras usadas nos experimentos subjetivos têm grande importância. As escalas de opinião mostradas a seguir são freqüentemente adotadas pelo ITU-T:

- Escala de qualidade de escuta (Tabela 7)

Tabela 7 - MOS

Qualidade da fala	Escala
Excelente	5
Boa	4
Fraca	3
Pobre	2
Ruim	1

Fonte: ITU (1996a).

A avaliação qualitativa dessa escala (Tabela 7) é representada pelo símbolo MOS (pontuação de opinião média da qualidade de escuta, ou simplesmente pontuação de opinião média).

– Escala de esforço de escuta (Tabela 8)

Tabela 8 - MOSle

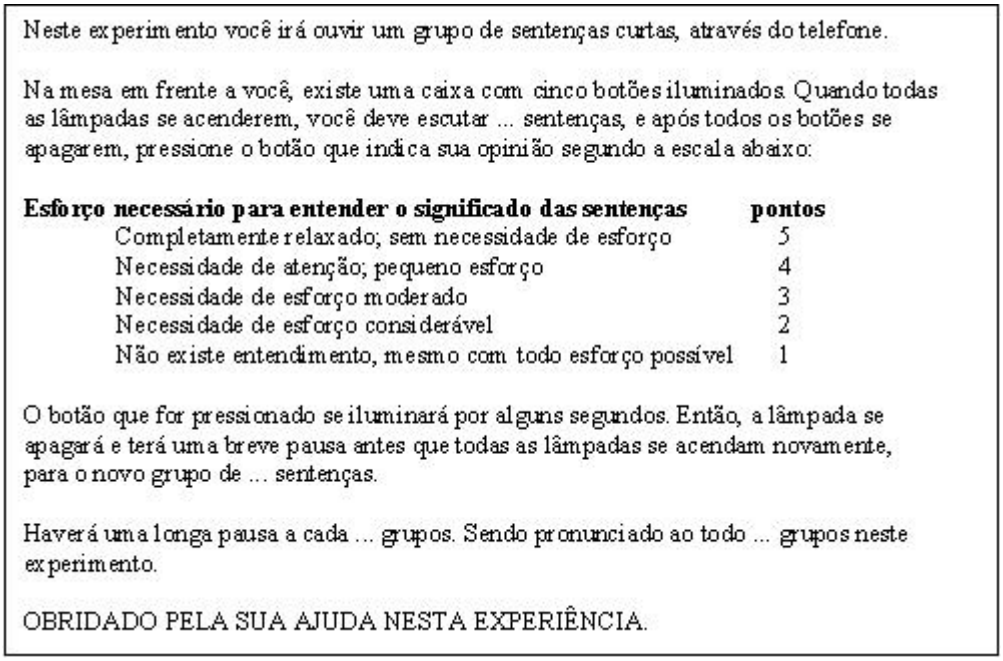
Esforço necessário para entender o significado das sentenças.	Pontos
Completamente relaxado; sem necessidade de esforço.	5
Necessidade de atenção; pequeno esforço.	4
Necessidade de esforço moderado.	3
Necessidade de esforço considerável.	2
Não existe entendimento, mesmo com todo o esforço possível.	1

Fonte: ITU (1996a).

A avaliação qualitativa dessa escala (Tabela 8) é representada pelo símbolo MOS$_{LE}$. Quando não é possível o uso da notação com texto subscrito, pode ser adotado o símbolo MOSle.

v) Instruções aos ouvintes

Um exemplo típico de instruções é apresentado na Figura 45. Elas devem ser apresentadas antes do início do experimento, podendo ser verbais, caso necessário. Após serem completamente entendidas, o ouvinte deve escutar algumas sentenças para praticar o emprego da pontuação. Nenhuma opinião que direcione serem exemplos com boa ou má qualidade deve ser sugerida a ele. Nem deve ser submetido a uma grande quantidade de exemplos, que cubra toda a faixa de pontuação. Depois das sentenças preliminares exemplificadoras, o ouvinte deve ter tempo suficiente para tirar qualquer tipo de dúvida, desde que não seja de cunho técnico. Perguntas técnicas só poderão ser respondidas após o término do experimento.



Neste experimento você irá ouvir um grupo de sentenças curtas, através do telefone.

Na mesa em frente a você, existe uma caixa com cinco botões iluminados. Quando todas as lâmpadas se acenderem, você deve escutar ... sentenças, e após todos os botões se apagarem, pressione o botão que indica sua opinião segundo a escala abaixo:

Esforço necessário para entender o significado das sentenças	**pontos**
Completamente relaxado; sem necessidade de esforço	5
Necessidade de atenção; pequeno esforço	4
Necessidade de esforço moderado	3
Necessidade de esforço considerável	2
Não existe entendimento, mesmo com todo esforço possível	1

O botão que for pressionado se iluminará por alguns segundos. Então, a lâmpada se apagará e terá uma breve pausa antes que todas as lâmpadas se acendam novamente, para o novo grupo de ... sentenças.

Haverá uma longa pausa a cada ... grupos. Sendo pronunciado ao todo ... grupos neste experimento.

OBRIDADO PELA SUA AJUDA NESTA EXPERIÊNCIA.

Figura 45 - Exemplo de instruções quando da aplicação do MOS
Autor: ITU (1996a).

vi) Análise estatística e resultados

A média numérica deve ser calculada para cada condição de nível de escuta, e devem ser listadas para uma inspeção inicial, de modo que se possa observar os efeitos das vozes masculinas e femininas.

O cálculo do desvio padrão para cada condição em separado não é recomendado. O limite de confiança deve ser avaliado com cuidado.

Para auxílio na análise dos dados, pode-se desenhar gráficos mostrando a pontuação em função de parâmetros do teste, por exemplo: MOS x Atenuação do circuito. Sendo que o MOS sempre deve ser mostrado no eixo vertical.

ANEXO B – *REAL-TIME TRANSPORTE PROTOCOL* (RTP)

A medida que o rádio da Internet, a telefonia da Internet, a música por demanda, a videoconferência, o vídeo sob demanda, e outras aplicações de multimídia se tornavam mais comuns, as ficou clara a necessidade da existência de um protocolo de tempo real que atendesse as necessidades das mesmas. Desse modo foi criado o RTP. Ele é descrito na RFC (RFC – *Request for Comments*) 3550 (IETF, 2003a).

A posição do RTP na pilha de protocolos é questionável. Decidiu-se que ele deveria ser inserido no espaço do usuário e, desse modo, ser (normalmente) executado sobre o UDP (UDP - *User Datagram Protocol*) (IETF, 1980). Pode-se entender melhor isso quando visto o funcionamento do RTP.

Aplicações multimídia consistem em vários fluxos de áudio, vídeo, texto e ainda outros fluxos. Esses fluxos são armazenados na biblioteca RTP como cita Tenenbaum (2003), que se encontra juntamente da aplicação do usuário. O RTP efetua a multiplexação dos fluxos e os codifica em pacotes RTP, que são então colocados em um soquete. Na outra extremidade do soquete, junto ao sistema operacional, os pacotes UDP são gerados e incorporados a pacotes IP. Se o computador estiver em uma rede Ethernet, os pacotes IP serão inseridos em quadros Ethernet para a transmissão. A pilha de protocolos para essa situação é mostrada na Figura 46.

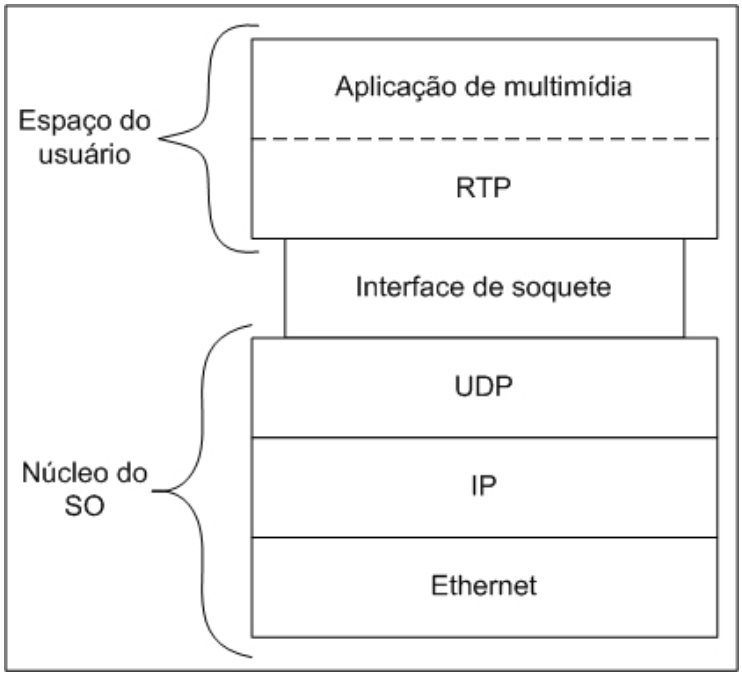


Figura 46 - Pilha de protocolos com o posicionamento do RTP
Autor: Tenenbaum (2003).

A formação dos quadros Ethernet, com o encapsulamento do RTP é mostrado na Figura 47.

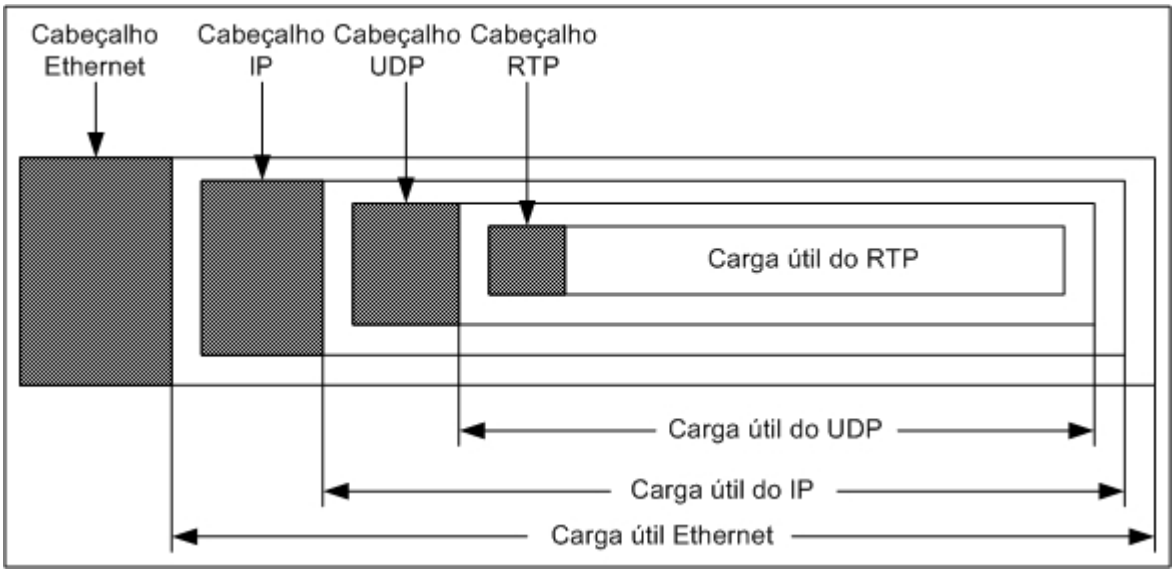


Figura 47 - Encapsulamento Ethernet com o RTP como carga
Autor: Tenenbaum (2003).

Como consequência dessa estrutura, pode ficar um pouco complexo definir em que camada o RTP está, se junto aplicação ou junto a camada de transporte. Como ele funciona junto ao programa de aplicação, fica parecendo mais um protocolo de aplicação. Por outro lado, o RTP é um protocolo genérico e independente das aplicações que apenas fornecem recursos de transporte, e assim também é semelhante a um protocolo de transporte. Segundo Tenenbaum (2003), a melhor definição é que o RTP possa ser protocolo de transporte implementado na camada de aplicação.

A função básica do RTP é multiplexa diversos fluxos de dados de tempo real, sobre um único fluxo de pacotes UDP. O fluxo UDP pode ser enviado a um único destino ou a vários destinos. Como o RTP utiliza o UDP padrão, seus pacotes não são tratados de maneira especial por roteadores.

Cada pacote enviado em um fluxo RTP recebe um número uma unidade maior que seu antecessor. Essa numeração permite ao destino identificar se algum pacote está faltando. Se um pacote for omitido por algum motivo, o mesmo não será reenviado, como por exemplo, no TCP (TCP - *Transmission Control Protocol*) (IETF, 1981). A retransmissão não é uma opção prática, pois o pacote retransmitido provavelmente chegaria com um atraso que inviabilizaria sua utilização pela aplicação de tempo real, as quais o protocolo se destina. Como consequência, o RTP não oferece nenhum controle de fluxo, nenhum controle de erros, nenhuma confirmação e nenhum mecanismo para solicitar retransmissões.

O cabeçalho do RTP é apresentado na Figura 48. Ele consiste em três palavras de 32 bits cada, e potencialmente extensões. Assim, até o campo do *Synchronization source identifier* o cabeçalho tem 12 bytes.

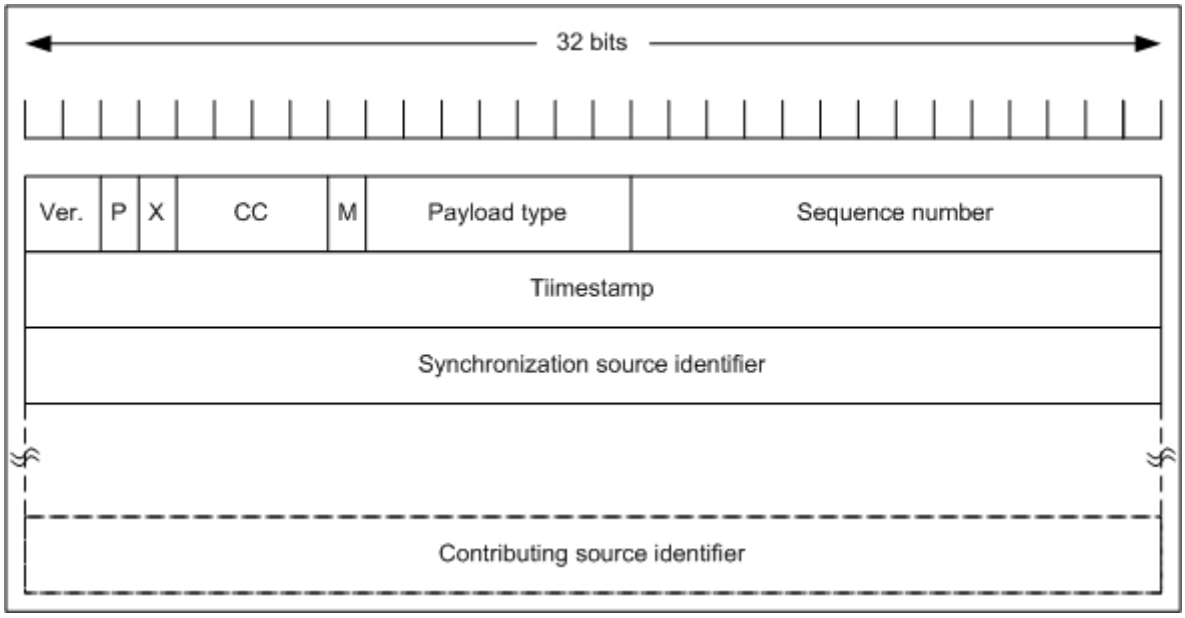


Figura 48 - Cabeçalho do RTP
Autor: Tenenbaum (2003) e IETF (2003a).

- Versão: A primeira palavra contém o campo de versão, que atualmente é a 2;
- O bit *P* indica que o pacote foi completado até chegar a um múltiplo de 4 bytes. O último byte de preenchimento informa quantos bytes foram acrescentados;
- O bit *X* indica que um cabeçalho de extensão está presente;
- O campo *CC* informa quantas origens de contribuição estão presentes, de 0 a 15;
- O bit *M* é um bit marcador específico da aplicação. Ele pode ser usado para marcar o começo de um quadro de vídeo, o começo de uma palavra em um canal de áudio;
- O campo *Payload type* informa que algoritmo de codificação foi usado;
- O *Sequence number* é um contador incrementado em cada pacote RTP enviado;
- O *Timestamp* é produzido pela origem do fluxo para anotar quando a primeira amostra no pacote foi realizada;
- *Synchronization source identifier* informa a que fluxo o pacote. Esse identificador é usado para multiplexar e demultiplexar vários fluxos de dados em um único fluxo de pacotes UDP;
- *Contributing source identifiers*, se estiverem presentes, serão aplicados quando houver misturadores (*mixer*) de áudio em estúdio.

O protocolo RTP tem uma extensão dita RTCP. O RTCP está especificado na RFC 3605 (IETF, 2003b) funciona realizando o monitoramento da aplicação junto a rede. A primeira função pode ser a de *feedback* sobre o atraso, *jitter*, largura de banda, congestionamento e outras características da rede. Essas informações podem ser usadas pelo processo de codificação para aumentar a taxa de dados, e oferecer melhor qualidade, quando a rede estiver respondendo bem, e para reduzir a taxa de dados quando houver problemas. Fornecendo esses relatórios de comportamento da rede de forma contínua, os algoritmos de codificação e inclusive de supressão de silêncio, quando no caso de voz, podem ser adaptados continuamente a ponto de sempre oferecer a melhor qualidade o possível para a aplicação.