



UNIVERSIDADE ESTADUAL PAULISTA

Pós-Graduação em Engenharia Elétrica

**ESTUDO E IMPLEMENTAÇÃO DE MÉTODOS DE
AVALIAÇÃO DE DISTORÇÕES EM SINAIS DE VOZ**

Giovani Cesar de Faria
Engenheiro Eletricista

Ilha Solteira - SP

1210001287



1287

Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Engenharia de Ilha Solteira
Pós-Graduação em Engenharia Elétrica

Dissertação de Mestrado

Giovani César de Faria

**ESTUDO E IMPLEMENTAÇÃO DE MÉTODOS DE
AVALIAÇÃO DE DISTORÇÕES EM SINAIS DE VOZ**

1210001287



Aluno: Giovani Cesar de Faria
Orientador: Jozué Vieira Filho

Doc. 071/2002-NRD 048/02

UNESP - "CAMPUS DE ILHA SOLTEIRA"	
SERVIÇO TÉCNICO DE BIBLIOTECA E DOCUMENTAÇÃO	
DATA DE CATEGORIA	DATA DE TOMBO
19.09.02	30.09.02
REVISÃO	TOMBO
Ailza	Te. 1287
AQUISICÃO	CLASSIFICAÇÃO
Doação Autor R\$ 10,00	F224 e

30400007
ENG. ELÉTRICA

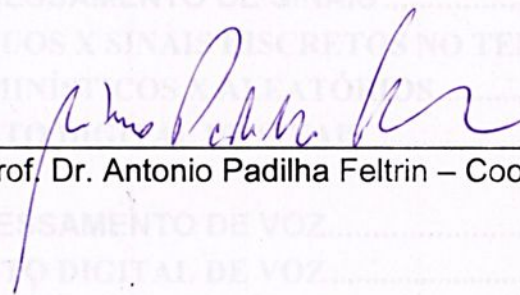
sup - 183251
sup - 54645

Ilha Solteira – SP / Dezembro de 2001

“Estudo e Implementação de Métodos de Avaliação de Distorções em Sinais de Voz”

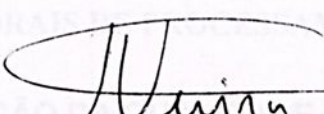
Giovani César de Faria

DISSERTAÇÃO SUBMETIDA À FACULDADE DE ENGENHARIA DE ILHA SOLTEIRA – UNESP – COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO TÍTULO DE MESTRE EM ENGENHARIA ELÉTRICA (ME).

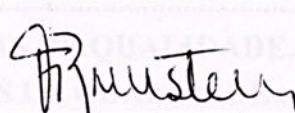


Prof. Dr. Antonio Padilha Feltrin – Coordenador

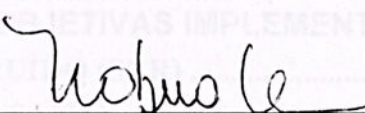
COMISSÃO EXAMINADORA:



Prof. Dr. Jozué Vieira Filho - orientador



Prof. Dr. Fernando Runstein



Prof. Dr. Nobuo Oki

Ilha Solteira – SP, dezembro de 2001



SUMÁRIO

CAPÍTULO I	INTRODUÇÃO	2
CAPÍTULO II	PROCESSO DE COMUNICAÇÃO POR VOZ	3
II.1	ANATOMIA DO SISTEMA DE PRODUÇÃO DE VOZ.....	4
II.2	FONÉTICA ACÚSTICA	5
II.3	PROPAGAÇÃO DO SOM	6
II.4	O OUVIDO E O PROCESSO AUDITIVO.....	6
II.4.1	O Mecanismo do Ouvido	7
II.4.2	O Ouvido Externo	7
II.4.3	O Ouvido Médio	8
II.4.4	O Ouvido Interno	8
CAPÍTULO III	MODELAMENTO DA PRODUÇÃO DE VOZ.....	9
III.1	MODELO DE FILTRO DISCRETO PARA PRODUÇÃO DE VOZ	10
CAPÍTULO IV	PROCESSAMENTO DE SINAIS	13
IV.1	SINAIS CONTÍNUOS X SINAIS DISCRETOS NO TEMPO.....	14
IV.2	SINAIS DETERMINÍSTICOS X ALEATÓRIOS.....	14
IV.3	PROCESSAMENTO DIGITAL DE SINAIS	14
CAPÍTULO V	PROCESSAMENTO DE VOZ.....	16
V.1	PROCESSAMENTO DIGITAL DE VOZ	16
V.2	REPRESENTAÇÃO DIGITAL DA FORMA DE ONDA DA VOZ	17
V.3	MÉTODOS TEMPORAIS DE PROCESSAMENTO	17
CAPÍTULO VI	AVALIAÇÃO DA QUALIDADE DE SINAIS DE VOZ	20
VI.1	QUALIDADE.....	20
VI.2	MEDIDAS SUBJETIVAS DE QUALIDADE.....	21
VI.3	MEDIDAS OBJETIVAS DE QUALIDADE.....	21
CAPÍTULO VII	MEDIDAS OBJETIVAS IMPLEMENTADAS	23
VII.1	RELAÇÃO SINAL/RUÍDO (SNR)	24
VII.2	MEDIDA ITAKURA-SAITO	25
VII.3	MEDIDA CEPSTRAL.....	26
VII.3.1	Processamento Homomórfico	27
VII.3.2	O Cepstro	28
VII.3.3	Deconvolução Homomórfica	28



VII.4	PSQM	29
VII.4.1	Procedimentos Objetivos de Medição	30
VII.4.2	Preparação dos Sinais de Voz	31
VII.4.3	Seleção de Parâmetros Experimentais	31
VII.4.4	Cálculo Objetivo da Qualidade	32
VII.4.5	Inicializações Globais	35
VII.4.6	Mapeamento Tempo-Frequência	35
VII.4.7	Mapeamento em Frequência e filtragem.....	36
VII.4.8	Distorção de Intensidade.....	38
VII.4.9	Modelamento Cognitivo	38
VII.4.10	Transformação de Escala de Qualidade: Objetiva → Subjetiva.....	40
CAPÍTULO VIII	MEDIDA SUBJETIVA REALIZADA	42
VIII.1	MOS	42
VIII.1.1	Escopo	42
VIII.1.2	Procedimentos de Testes Subjetivos	43
VIII.1.3	Arquivos de Voz.....	43
VIII.1.4	Seleção dos parâmetros do experimento	44
VIII.1.5	Condições de Referência	47
VIII.1.6	Escala de Qualidade de Audição.....	47
CAPÍTULO IX	SIMULAÇÕES E RESULTADOS	50
IX.1	MEDIDAS OBJETIVAS	50
IX.1.1	Relação Sinal/Ruído (SNR)	52
IX.1.2	Distância Itakura-Saito.....	54
IX.1.3	Distância Cepstral	56
IX.1.4	Comparação entre as Medidas Objetivas.....	57
IX.1.5	Diferença de Energia.....	58
IX.1.6	PSQM.....	60
IX.2	Comparação entre PSQM e as demais Medidas Objetivas	64
IX.3	MEDIDA SUBJETIVA - MOS	67
CAPÍTULO X	CONCLUSÕES	68
CAPÍTULO XI	REFERÊNCIA BIBLIOGRÁFICA	69



RESUMO

A avaliação da qualidade em sinais de voz é importante em todas as aplicações de telecomunicações que envolvem comunicação entre dois interlocutores. Essa avaliação é baseada em medidas de distorções que podem ser objetivas ou subjetivas. Neste trabalho o objetivo foi estudar, implementar e avaliar as principais medidas de distorções em sinais de voz. Inicialmente, estudou-se o processo de produção de sinais de voz e seu modelamento, com uma análise posterior do ponto de vista de processamento de sinais. Procurou-se apresentar de maneira resumida o processo auditivo humano, uma vez que uma das medidas objetivas implementadas, a PSQM, tem como princípio o modelamento da percepção humana da voz. Na seqüência foram estudadas quatro medidas objetivas e uma subjetiva, que foram, respectivamente: Relação Sinal/Ruído, Distância Itakura-Saito, Distância Cepstral, PSQM e MOS. Finalmente, todas essas medidas foram implementadas, avaliadas e comparadas entre si. Os resultados indicam que todas as medidas podem ser utilizadas, mas algumas delas não apresentam resultados satisfatórios quando a relação sinal/ruído do sinal de voz avaliado é relativamente baixa. Há também medidas que não indicam corretamente as distorções para diferentes tipos de ruído.

Com o intuito de explorar as possibilidades e técnicas de avaliação da qualidade de sinais de voz submetidos a distorções (e.g. codificação) realizou-se este trabalho apresentando várias dos métodos disponíveis (objetivos e subjetivos) e efetuando uma análise comparativa entre as técnicas. Além disso, para permitir uma visão melhor do contexto onde se insere a Análise da Qualidade de Voz, são apresentados pontos importantes do Processo de Comunicação por Voz, resumidos em dois dos principais processos envolvidos: o Processo de Produção do Sinal de Voz e o Processo Auditivo.

Capítulo I

INTRODUÇÃO

Os sistemas de comunicação vêm sofrendo uma grande evolução ao longo dos anos, evolução esta, motivada pelo crescente aumento do tráfego e pela demanda de novos serviços, cada vez mais sofisticados requerendo maior capacidade de processamento e de transmissão. Nesse sentido, surgiram inúmeras técnicas de codificação de voz que visam a diminuição das taxas de transmissão de bits sem perda da qualidade. Estas técnicas procuram representar os sinais de forma eficiente com o menor número de bits possível, garantindo sua recuperação no lado da recepção do link de comunicação.

Ao longo do texto são explorados algoritmos de avaliação dos processos (algoritmos) de codificação de voz empregados nestes sistemas de comunicação. Este interesse é motivado por uma ampla faixa de aplicações, incluindo telefonia comercial, celular, comunicação militar, correio de voz, armazenamento de sinais de voz e redes de comunicação, uma vez que ao longo do link de comunicação o sinal de voz sofre uma variedade de processamentos desde seu início (geração do sinal de voz) até o momento em que chega aos ouvidos do interlocutor no lado da recepção.

O objetivo da codificação de voz é representá-la na forma digital com o menor número possível de bits, mantendo a inteligibilidade e qualidade necessárias para cada aplicação. Em altas taxas de bits (e.g. 32 kbits/s) conseguir boa qualidade e inteligibilidade não é difícil, mas à medida que se reduz a taxa de bits (e.g. 16 kbits/s ou menos) o problema torna-se cada vez mais desafiador. Dependendo da aplicação, muitas restrições difíceis de se contornar devem ser consideradas, incluindo a da complexidade de implementação. Entretanto, o processo de codificação é responsável pela introdução de distorções que podem se tornar importantes em uma dada aplicação. Dessa forma, emergiu uma área que tem como objetivo a avaliação da performance de um sinal de voz processado, por exemplo, por codificadores / decodificadores (codecs). Todavia, avaliar a qualidade dos algoritmos de processamento não é tarefa fácil uma vez que diferentes tipos de ruído produzem efeitos diversos nos sinais de voz.

Com o intuito de explorar as possibilidades e técnicas de avaliação da qualidade de sinais de voz submetidos a distorções (e.g. codificação/decodificação) realizou-se este trabalho apresentando vários dos métodos disponíveis (objetivos e subjetivos) e efetuando uma análise comparativa entre as técnicas. Além disso, para permitir uma visão melhor do contexto onde se insere a Análise da Qualidade de Voz, são apresentados pontos importantes do Processo de Comunicação por Voz, resumidos em dois dos principais processos envolvidos: o Processo de Produção do Sinal de Voz e o Processo Auditivo.

Capítulo II**PROCESSO DE COMUNICAÇÃO POR VOZ**

Neste capítulo são discutidos temas referentes ao processo de comunicação por meio da voz. Para isso, são apresentados alguns detalhes da anatomia dos sistemas de produção da voz e auditivo. Além disso, tópicos como fonética e propagação do som são também discutidos.

Os algoritmos de codificação/decodificação trabalham com representações dos sinais de voz, ou seja, trabalham com modelos digitais da voz. Assim, para facilitar a análise e compreensão dos modelos de produção e percepção de sinais de voz existentes, um breve estudo do complexo processo de comunicação por meio da fala faz-se necessário [1].

O propósito principal da fala para os seres humanos é a comunicação entre pessoas. Este processo de comunicação por meio da voz inicia-se com a produção da mensagem no cérebro do locutor (processo neurológico). Em seguida, ocorre a conversão desta mensagem para um sinal acústico (processo bastante complexo) que é transmitido até o interlocutor (ouvinte). Neste processo, o cérebro humano produz uma sequência de comandos motores para movimentar os vários articuladores do sistema vocal de modo a produzir o sinal de voz desejado.

O processo de formação do sinal de voz pode ser esquematizado pelo diagrama da figura 1, onde são descritos alguns dos importantes articuladores do sistema vocal (língua, lábios e cordas vocais).

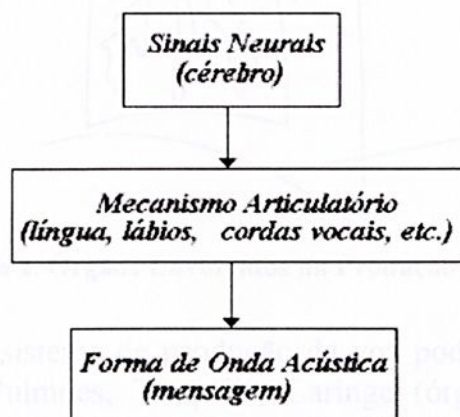


Figura 1. Formação do sinal de voz

Para comunicar-se, o ser humano efetua a transmissão da informação, que é o conteúdo da mensagem. Todavia, a fala pode ser interpretada como sendo a concatenação de um número finito de símbolos (representados pelos fonemas), que são organizados de forma coerente e compreensível de acordo com uma estrutura lingüística e regras gramaticais apropriadas. Além disso, características gerais e importantes, tais como entonação e ênfase, são adicionadas pelo locutor para conseguir o efeito desejado.

II.1 ANATOMIA DO SISTEMA DE PRODUÇÃO DE VOZ

A forma de onda da voz é resultado de uma onda de pressão acústica de som que se origina a partir de movimentos voluntários e controlados de estruturas anatômicas que formam o sistema humano de produção da fala, como indicado na figura 2 [2].

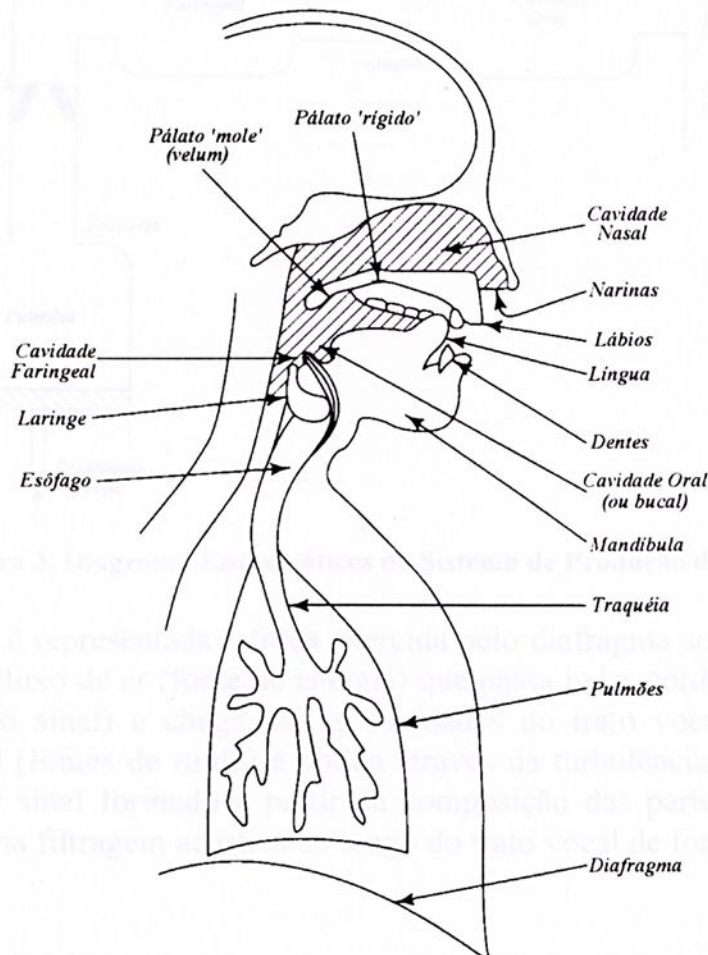


Figura 2. Órgãos Envolvidos na Produção da Voz

Os componentes do sistema de produção da voz podem, simplificadaamente, ser definidos como sendo: Pulmões, Traquéia, Laringe (órgão de produção de voz), Cavidade da Faringe (garganta), Cavidade Oral (boca) e Cavidade Nasal (nariz).

Uma análise mais detalhada das características anatômicas, cruciais para a produção da voz, inclui as cordas vocais (ou dobras vocais), o véu palatino, a língua, o maxilar, os dentes e os lábios.

Uma maneira útil de se analisar o sistema de produção da voz é trata-lo como sendo uma operação de filtragem acústica [2], onde os sinais periódicos e não-periódicos gerados são filtrados de modo a obter-se o som desejado. As principais cavidades do sistema de produção de voz (trato vocal + trato nasal) formam o principal filtro acústico. De forma esquemática, pode-se representar o sistema de produção de voz como na figura 3.

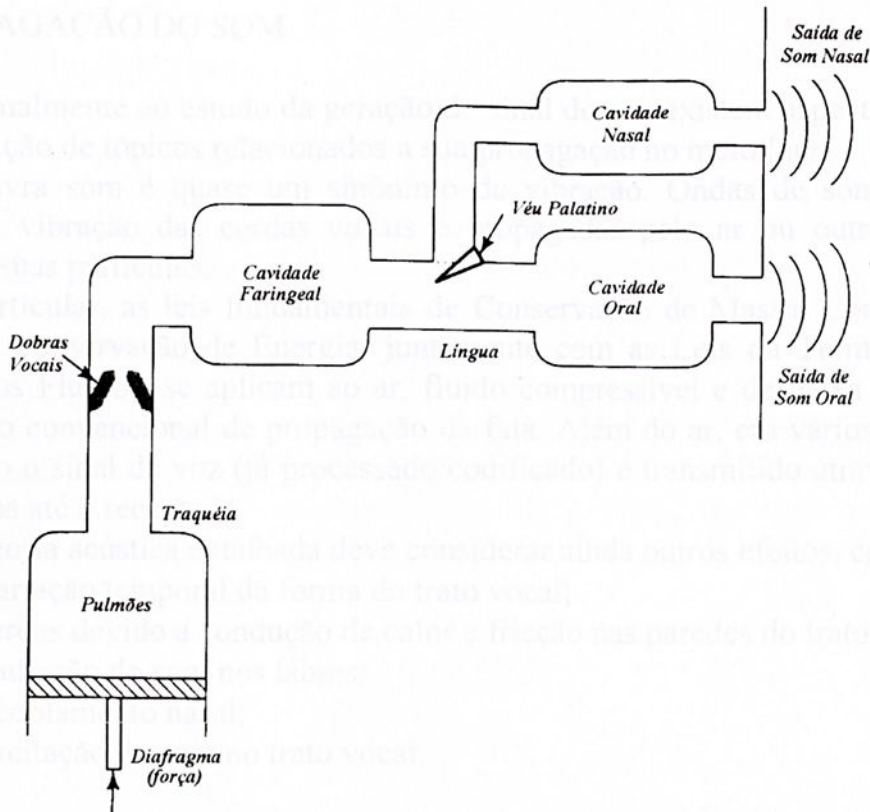


Figura 3. Diagrama Esquemáticos do Sistema de Produção de Voz

Na figura 3, é representada a força exercida pelo diafragma sobre os pulmões de modo a criar um fluxo de ar (fonte de energia) que passa pelas cordas vocais (geram a parte periódica do sinal) e chega até as cavidades do trato vocal. Já a parte não periódica do sinal (fontes de ruído) é obtida através da turbulência nos estreitamentos do trato vocal. O sinal formado a partir da composição das partes periódica e não periódica sofre uma filtragem acústica ao longo do trato vocal de forma a se produzir o sinal de voz.

II.2 FONÉTICA ACÚSTICA

Muitos dos idiomas podem ser descritos em termos de um conjunto de sons distintos (fonemas). Embora haja uma variedade de maneiras de se estudar a fonética, consideremos apenas uma caracterização acústica dos vários sons incluindo a maneira de articulação, formas de onda e sua caracterização espectrográfica [1].

De um modo geral, têm-se:

- SONOROS: produzidas pela excitação do trato vocal (fixo) por pulsos de ar quase periódicos causados pela vibração das cordas vocais;
- NÃO-SONOROS: constrição da faringe e não vibração das cordas vocais (excitação tipo ruído branco);
- PLOSIVOS: produzidos por meio da constrição total do trato vocal e liberação súbita do ar.

II.3 PROPAGAÇÃO DO SOM

Adicionalmente ao estudo da geração do sinal de voz existem aspectos referentes à caracterização de tópicos relacionados a sua propagação no meio (ar).

A palavra som é quase um sinônimo de vibração. Ondas de som podem ser criadas pela vibração das cordas vocais e propagadas pelo ar ou outro meio pela vibração de suas partículas.

Em particular, as leis fundamentais de Conservação de Massa, Conservação de Momento e Conservação de Energia, juntamente com as Leis da Termodinâmica e Mecânica dos Fluidos, se aplicam ao ar, fluido compressível e de baixa viscosidade, que é o meio convencional de propagação da fala. Além do ar, em vários sistemas de comunicação o sinal de voz (já processado/codificado) é transmitido através de cabos (óticos) e fios até o receptor.

Uma teoria acústica detalhada deve considerar ainda outros efeitos, como:

- Variação temporal da forma do trato vocal;
- Perdas devido a condução de calor e fricção nas paredes do trato vocal;
- Radiação do som nos lábios;
- Acoplamento nasal;
- Excitação do som no trato vocal.

II.4 O OUVIDO E O PROCESSO AUDITIVO

O receptor final de informação em um sistema de comunicação por voz é usualmente o ser humano. Suas habilidades de percepção ditam a precisão com que os dados (voz) devem ser processados e transmitidos. Estas habilidades prescrevem critérios de fidelidade para recepção e, de fato, acabam por determinar a capacidade necessária do canal para a transmissão das mensagens de voz.

A informação contida no sinal de voz – partindo de um locutor, atravessando um canal de transmissão e chegando ao ouvinte – deve ser considerada em um número de diferentes etapas de codificação. No lado da transmissão, as etapas devem incluir as ondas acústicas, as forças musculares manipulando o mecanismo vocal (ou sua forma física) e a excitação do trato vocal. No lado do receptor, a informação deve ser considerada em termos dos movimentos acústicos-mecânicos do transdutor auditivo (ouvido), ou em termos dos pulsos elétricos transmitidos ao cérebro através do nervo auditivo.

O diagrama da figura 4 ilustra, de maneira bastante simplificada, o processo auditivo.

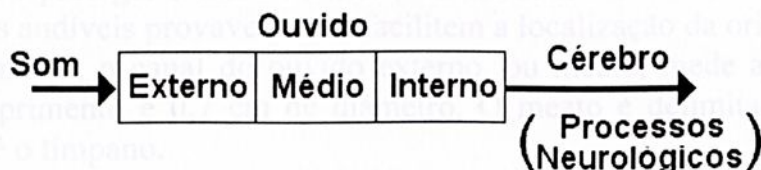


Figura 4. Processo Auditivo

II.4.1 O Mecanismo do Ouvido

O processo auditivo compreende, como resumidamente apresentado na figura 4, as operações acústicas e mecânicas do ouvido periférico, os processos para conversão dos movimentos mecânicos em impulsos elétricos no ouvido interno, a transmissão das informações neurais ao cérebro e o mecanismo final da percepção humana.

Apesar das dificuldades em se descrever precisamente cada um desses processos, é possível quantificar certos aspectos da percepção sem se saber com detalhes o que está realmente acontecendo.

O transdutor acústico básico do ser humano, ou ouvido, é mostrado esquematicamente na figura 5. Os componentes acústicos e mecânicos do órgão são convencionalmente divididos de acordo com três regiões: ouvido externo, ouvido médio e ouvido interno.

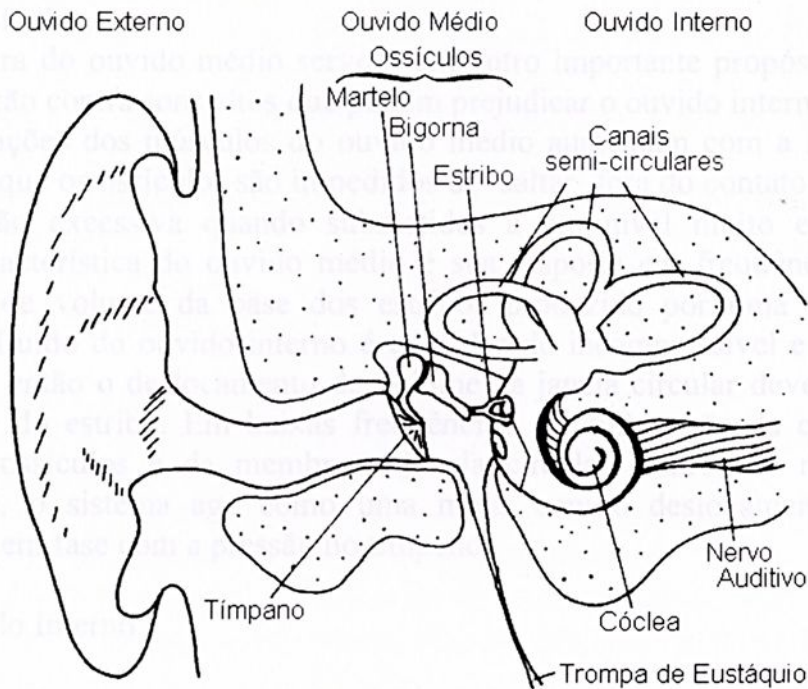


Figura 5. Diagrama Esquemático do Ouvido Humano

II.4.2 O Ouvido Externo

O termo ouvido, como é comumente utilizado, aplica-se ao que também é conhecido como orelha, que circunda a entrada do canal do ouvido externo. Sua função principal é proteger o canal externo (embora suas características direcionais em altas frequências audíveis provavelmente facilitem a localização da origem dos sons).

No ser humano, o canal do ouvido externo, ou meato, mede aproximadamente 2,7 cm de comprimento e 0,7 cm de diâmetro. O meato é delimitado por uma fina membrana que é o tímpano.

II.4.3 O Ouvido Médio

Depois de chegar ao ouvido externo o som é direcionado para dentro da cavidade do ouvido médio, que se encontra logo após o tímpano. Esta cavidade contém os chamados ossículos – martelo, bigorna e estribo. A função desses ossículos é, principalmente, a transformação de impedância do meio aéreo do ouvido externo para o meio líquido do ouvido interno. O martelo é fixado e repousa no tímpano. Ele faz contato com a bigorna, que por sua vez conecta-se por meio de uma pequena junção ao estribo. A base do estribo assenta-se em um porto, a janela oval, e é seguro por um ligamento anular. A janela oval é a entrada para o ouvido interno.

Neste processo uma onda de som, chocando-se com o ouvido externo, é conduzida ao meato externo e provoca uma vibração no tímpano. A vibração é transmitida por meio dos três ossículos ao ouvido interno. A impedância do ouvido interno é muito maior do que a do ar. Dessa forma, para uma transmissão eficiente da energia do som, se faz necessária uma transformação de impedância, que é obtida através dos ossículos.

A estrutura do ouvido médio serve a um outro importante propósito que é o de fornecer proteção contra sons altos que podem prejudicar o ouvido interno.

As contrações dos músculos do ouvido médio aumentam com a intensidade do som, de modo que os ossículos são impedidos de ‘saltar’ fora do contato um do outro e causar distorção excessiva quando submetidos a um nível muito elevado. Outra importante característica do ouvido médio é sua resposta em frequência, ou seja, o deslocamento de volume da base dos estribos produzido por uma dada onda de pressão. Se o fluido do ouvido interno é considerado incompressível e as paredes da cóclea rígidas, então o deslocamento de volume da janela circular deve ser o mesmo que o da base do estribo. Em baixas frequências, a combinação da elasticidade do tímpano, dos ossículos e da membrana janela circular controla o movimento do estribo, isto é, o sistema age como uma mola, com o deslocamento do estribo proporcional e em fase com a pressão no tímpano.

II.4.4 O Ouvido Interno

O ouvido interno, como ilustrado na figura 5, é composto pela cóclea (normalmente enrolada como uma concha em uma espiral plana com duas voltas e meia), pelo aparato vestibular e pelas terminações do nervo auditivo. É na cóclea que ocorre a transformação auditiva do sinal mecânico para os sinais neurais. Os componentes vestibulares (canais semicirculares) servem como senso de orientação espacial e aparentemente não são usados para detecção de vibrações sonoras.

A câmara da cóclea é preenchida com um líquido incolor – a perilinfa – que tem a viscosidade de aproximadamente duas vezes a da água. O comprimento do canal na concha espiral é próximo de 35 mm. A câmara da cóclea é dividida ao longo de quase todo seu comprimento por uma partição.

O ouvido interno é conectado ao ouvido médio pela base dos estribos. Com a vibração, os estribos agem como um pistão produzindo um deslocamento de volume do fluido da cóclea.



Capítulo III**MODELAMENTO DA PRODUÇÃO DE VOZ**

Neste capítulo discute-se o modelamento da produção da voz. Assim, através da análise dos tipos de excitação existentes e da elaboração de um modelo físico para o trato vocal, chega-se a uma representação do sistema de produção da voz bastante utilizado.

Em sistemas de comunicação, o sinal de voz é transmitido, armazenado e processado de diversas maneiras. Uma vez o sinal processado - por exemplo, por um codificador - espera-se recuperá-lo de forma eficiente ao final do processo, ou seja, sem prejuízo de nenhuma de suas características tidas como importantes. Para isso, a representação utilizada deve ser de tal modo que possa atender aos requisitos do sistema de comunicação utilizado.

A motivação para a compreensão do mecanismo de produção de fala baseia-se no fato de que a fala é o meio primário de comunicação do ser humano [2] e é um importante recurso para se determinar os parâmetros necessários para melhor representação da voz.

O modelo da produção de fala é baseado em três pontos importantes:

- Tipo de Excitação;
- Forma do Trato Vocal;
- Efeito da Radiação nos Lábios.

Tipos de excitação em função da excitação

1. VOCÁLICA: um movimento periódico das cordas vocais provocado por um fluxo constante de ar;
2. NÃO VOCÁLICA: uma excitação turbulenta semelhante ao ruído, causada pelo fluxo de ar através de uma constrição estreita;
3. EXPLOSIVA: causada por um acúmulo de pressão de ar atrás de uma porção totalmente fechada do trato vocal, seguida por uma repentina liberação dessa pressão;

Para muitas aplicações utilizando sinais de voz (tais como codificação, síntese e reconhecimento), uma boa performance pode ser atingida com um modelo da fala que reflita amplas características de padrões de tempo e articulação, bem como, propriedades de variação de frequência.

A configuração física mais simples que tem interpretação útil em termos do processo de produção da voz é apresentada na figura 6.



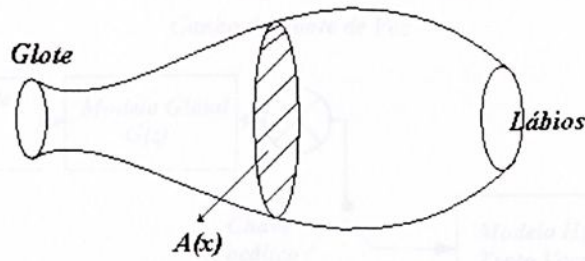


Figura 6. Tubo de Seção Não Uniforme e Variável

Na figura 6, o trato vocal é modelado como um tubo de área de seção transversal não uniforme e variante no tempo. A maneira como varia a área de seção transversal ao longo do trato vocal determina as frequências de ressonância, ou formantes, que caracterizam os sinais de voz. Dessa forma, para a obtenção de um modelo mais realístico do trato vocal deve utilizar-se um tubo de seção variável em função do tempo e do deslocamento ao longo do eixo de propagação do som. Como esta formulação pode ser extremamente complexa, um método de simplificação do modelo é baseado na representação do trato vocal por meio de uma série de tubos concatenados, como esquematizado na figura 7. Neste modelo, os A_k ($1 \leq k \leq 5$) representam as áreas de seção transversal e os L_k ($1 \leq k \leq 5$) os comprimentos de cada tubo.

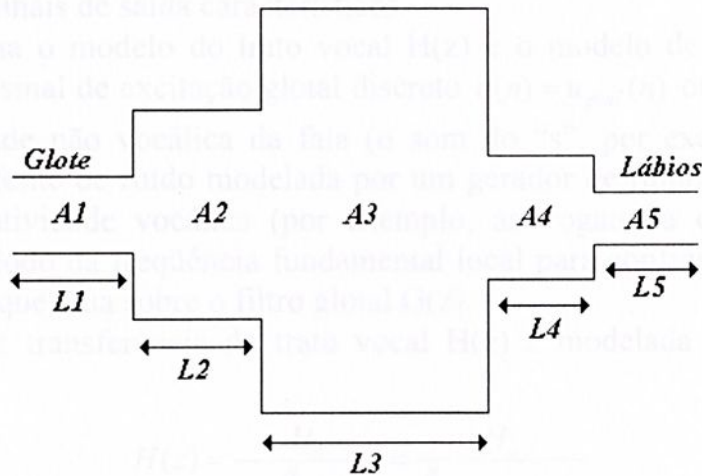


Figura 7. Modelo de Tubos do Trato Vocal

Esta configuração estabelece uma seletividade (em frequência) do tubo, que influencia o fluxo de ar que se propaga de forma a determinar suas características espectrais. Todavia, em qualquer indivíduo, as propriedades espectrais de sua fala variam com o tempo, conforme varia a forma do trato vocal.

III.1 MODELO DE FILTRO DISCRETO PARA PRODUÇÃO DE VOZ

Os componentes dedicados para processamento de sinais e os computadores cada vez mais rápidos nos permitem um grau de liberdade muito grande no desenvolvimento de sistemas de processamento digital. Assim, é interessante modelar o sinal de voz no mundo discreto. Um modelo simplificado geral linear e discreto no tempo para a produção de voz é mostrado na figura 8 [2].

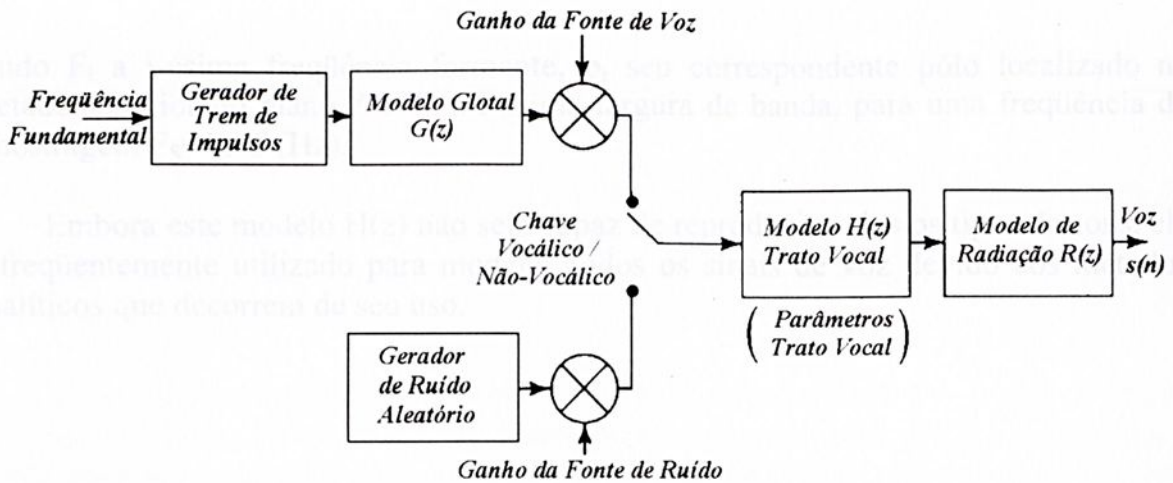


Figura 8. Modelo Discreto para Produção da Voz

Este modelo é conhecido como Modelo Geral Discreto no Tempo para geração da fala. Este nome deve-se ao fato de que os sinais e sistemas envolvidos no modelo são somente superficialmente similares ao sistema físico real, exceto no término (representado pelos lábios), onde ambos produzem formas de onda analógicas da fala. Em outras palavras, este modelo tenta representar o processo de produção de voz baseado em seus sinais de saída característicos.

Neste sistema o modelo do trato vocal $H(z)$ e o modelo de radiação $R(z)$ são excitados por um sinal de excitação glotal discreto $u(n) = u_{glote}(n)$ ou por ruído branco. Durante a atividade não vocálica da fala (o som do “s”, por exemplo), a fonte de excitação é uma fonte de ruído modelada por um gerador de ruído aleatório. Durante os períodos de atividade vocálica (por exemplo, as vogais) a excitação usa uma estimativa do período da frequência fundamental local para configurar um gerador de trem de impulsos que atua sobre o filtro glotal $G(z)$.

A função de transferência do trato vocal $H(z)$ é modelada usando a seguinte simplificação:

$$H(z) = \frac{H_o}{1 - \sum_{k=1}^N b_k z^{-k}} = \frac{H_o}{\prod_{k=1}^N (1 - p_k z^{-k})}$$

Na equação anterior, H_o representa um termo de ganho e p_k as localizações dos pólos complexos para o modelo de tubos da figura 7.

Cada par de pólos conjugados no plano Z (p_i, p_i^*) corresponde grosseiramente a uma frequência formante no espectro de $H(z)$. Como $H(z)$ deve ser um sistema estável, todos os pólos devem estar dentro do círculo unitário no plano- Z . Se os pólos no plano Z estão bem separados, boas estimativas para as frequências formantes e larguras de banda são:

$$\hat{F}_i = \left(\frac{F_s}{2\pi} \right) \cdot \tan^{-1} \left[\frac{\text{Im}(p_i)}{\text{Re}(p_i)} \right] \text{ e}$$

$$\hat{B}_i = - \left(\frac{F_s}{\pi} \right) \cdot \ln[p_i],$$

sendo F_i a i -ésima frequência formante, p_i seu correspondente pólo localizado na metade superior do plano-Z e B_i a i -ésima largura de banda, para uma frequência de amostragem $F_s = 1/T$ (Hz).

Embora este modelo $H(z)$ não seja capaz de reproduzir todos os tipos de sons, ele é frequentemente utilizado para modelar todos os sinais de voz devido aos métodos analíticos que decorrem de seu uso.

Este trabalho utiliza amplamente o conceito de processamento digital de sinais, ou seja, utilização de sinais discretos no tempo e na amplitude que são submetidos a algoritmos que executam cálculos matemáticos de modo a se obterem os respectivos sinais de saída.

Um sinal é entendido como qualquer quantidade física que varia com o tempo, espaço ou quaisquer outras variáveis independentes, sendo que neste trabalho os sinais de interesse são os sinais de voz [3].

Um segmento periódico analógico de voz é, em geral, representado como uma soma de senóides de diferentes amplitudes e frequências, ou seja:

$$s(t) = \sum_{i=1}^N A_i \cos(2\pi f_i t + \theta_i)$$

Uma maneira de analisar o conteúdo de uma informação ou mensagem contida em um certo segmento do sinal de voz é medir as amplitudes, frequências e fases contidas neste segmento.

O processamento de sinais tem função de auxiliar no processo de transformação da representação do sinal em formas alternativas que são mais apropriadas para aplicações específicas.

De maneira esquemática, pode-se representar o processamento do sinal de voz do locutor a seguir como apresentado na figura 9.

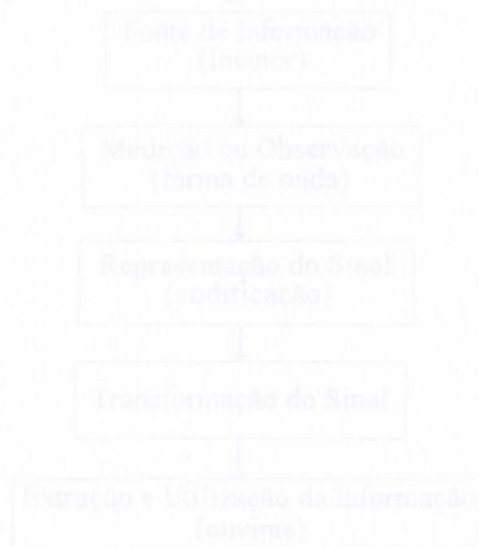


Figura 9. Processamento do sinal de voz: Locutor → Ouvinte

Capítulo IV**PROCESSAMENTO DE SINAIS**

Neste capítulo busca-se apresentar de maneira bastante ampla a importância do processamento de sinais e, em especial, do processamento de sinais de voz. Assim, são apresentados vários dos diferentes tipos de sinais, como por exemplo: sinais contínuos no tempo, discretos no tempo, determinísticos e aleatórios.

Este trabalho utiliza amplamente o conceito de processamento digital de sinais, ou seja, utiliza-se de sinais discretos no tempo e na amplitude que são submetidos a algoritmos que executam cálculos matemáticos de modo a se obterem os respectivos sinais de saída.

Um sinal é entendido como qualquer quantidade física que varia com o tempo, espaço ou quaisquer outras variáveis independentes, sendo que nesse trabalho os sinais de interesse são os sinais de voz [3].

Um segmento periódico analógico de voz é, em geral, representado como uma soma de senóides de diferentes amplitudes e frequências, ou seja:

$$u(t) = \sum_{i=1}^N A_i(t) \cdot \sin[\omega_i t + \theta_i(t)]$$

Uma maneira de analisar o conteúdo de uma informação ou mensagem contida em um curto segmento do sinal de voz é medir as amplitudes, frequências e fases contidas neste segmento.

O processamento de sinais tem função de auxiliar no processo de transformação da representação do sinal em formas alternativas que são mais apropriadas para aplicações específicas.

De maneira esquemática, pode-se representar o processamento do sinal de voz do locutor ao ouvinte como apresentado na figura 9.

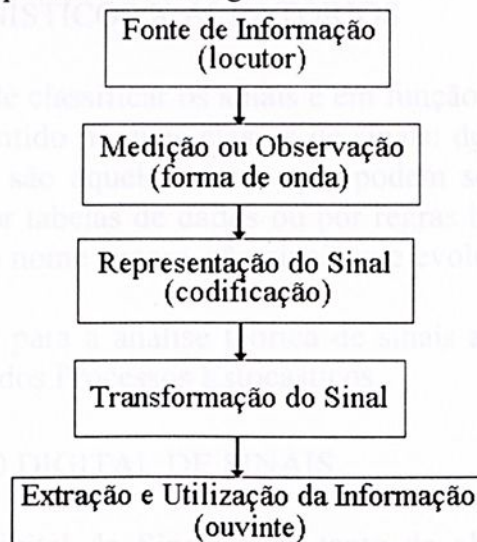


Figura 9. Processamento do sinal de voz: Locutor → Ouvinte

A figura 9 apresenta, resumidamente, os procedimentos utilizados no tratamento de um dado sinal de voz. Depois de ser gerado (fonte de informação) este sinal é medido ou observado para, em função da aplicação, passar por processos de codificação/decodificação, entre outros.

Os métodos usados no processamento de um sinal ou na análise da resposta de um sistema a um sinal dependem severamente dos atributos característicos do sinal e dos requisitos da aplicação.

IV.1 SINAIS CONTÍNUOS X SINAIS DISCRETOS NO TEMPO

Quando se trata de processamento de sinais considera-se, basicamente, dois tipos distintos de sinais, os contínuos e os discretos no tempo. Os sinais contínuos no tempo são aqueles que têm suas amplitudes definidas em um intervalo contínuo de tempo e podem ser descritos matematicamente por funções de um variável contínua. Já os sinais discretos no tempo são definidos somente em certos instantes de tempo específicos e são representados matematicamente por uma sequência de números reais ou complexos.

Quando os sinais contínuos podem assumir todos os valores possíveis de amplitude de uma faixa finita, se denominam “analógicos”. Entretanto, os sinais digitais (discretos também na amplitude) assumem valores de amplitude em um conjunto finito de valores possíveis.

Para que um dado sinal analógico possa ser processado digitalmente ele deve ser convertido para o modo digital, ou seja, discretizado no tempo e em amplitude. Para isso, ele deve passar pelo processo de amostragem e quantização. Quantização é um processo de conversão de um sinal contínuo em amplitude para um sinal discreto em amplitude, que pode ser considerado, basicamente como um processo de aproximação (arredondamento ou truncamento).

IV.2 SINAIS DETERMINÍSTICOS X ALEATÓRIOS

Uma outra maneira de classificar os sinais é em função do seu comportamento ao longo do tempo. Neste sentido há duas classes de sinais: determinísticos e aleatórios. Os sinais determinísticos são aqueles sinais que podem ser descritos por equações matemáticas explícitas, por tabelas de dados ou por regras bem definidas. Já os sinais aleatórios, como o próprio nome sugere, são sinais que evoluem ao longo do tempo de maneira imprevisível.

A base matemática para a análise teórica de sinais aleatórios é fornecida pela Teoria da Probabilidade e dos Processos Estocásticos.

IV.3 PROCESSAMENTO DIGITAL DE SINAIS

O Processamento Digital de Sinais trata tanto da obtenção de representações discretas de sinais quanto da teoria, projeto e implementação de procedimentos numéricos de processamento (algoritmos). Os processamentos digital e analógico de



sinais possuem objetivos semelhantes. Todavia, atenção especial deve ser dada ao processamento digital pelas seguintes razões:

- implementação de funções extremamente sofisticadas;
- implementações confiáveis e compactas;
- processamento em tempo real;
- flexibilidade na mudança do algoritmo (sem alterar o hardware);
- possibilidade de procedimentos que não existem no mundo analógico.

Além disso, os desenvolvimentos teóricos e os avanços na área de hardware digital levam ao fortalecimento das vantagens das técnicas de processamento digital sobre os equivalentes sistemas analógicos.

V.1. PROCESSAMENTO DIGITAL DE VOZ

Na aplicação de técnicas de processamento digital em sistemas de comunicação de voz é importante enfatizar três tópicos principais:

- Representação digital de sinais de voz;
- Implementação de técnicas de processamento sofisticadas;
- Aplicações típicas de processamento digital.

Há muitas possibilidades de representação discreta de sinais de voz. Estas representações podem ser classificadas em dois grandes grupos, como apresentado na figura 10.



Figura 10. Representações dos sinais de voz.

As representações por Forma de Onda visam exatamente a preservação da forma da onda do sinal de voz analógico através de um processo de amostragem e quantização. Já as representações Paramétricas têm como saída um modelo para produção do sinal de voz. Este último tipo de representação parte de uma representação digital da forma de onda do sinal (amostragem e quantização), que é então, processada para a obtenção dos parâmetros do modelo que melhor representa o sinal de entrada. Nas representações, após o processamento, espera-se recuperar o sinal original com o mínimo de distorção possível.



Capítulo V**PROCESSAMENTO DE VOZ**

Neste capítulo discutem-se a importância do processamento de voz e algumas das técnicas empregadas na sua análise.

A área de processamento de voz é um campo multidisciplinar que requer pesquisa nas ciências da fala e audição e em processamento digital de sinais. Ela trata do desenvolvimento e implementação de técnicas de processamento específicas para sinais de voz. Dessa forma, são exploradas as características comuns aos sinais de voz de modo a tornar mais eficiente o processamento. Neste tipo de processamento, muitas vezes, trabalha-se com representações dos sinais de voz

V.1 PROCESSAMENTO DIGITAL DE VOZ

Na aplicação de técnicas de processamento digital em sistemas de comunicação de voz é importante enfatizar três tópicos principais:

- Representação digital de sinais de voz;
- Implementação de técnicas de processamento sofisticadas;
- Aplicações típicas de processamento digital.

Há muitas possibilidades de representação discreta de sinais de voz. Estas representações podem ser classificadas em dois grandes grupos, como apresentado na figura 10.

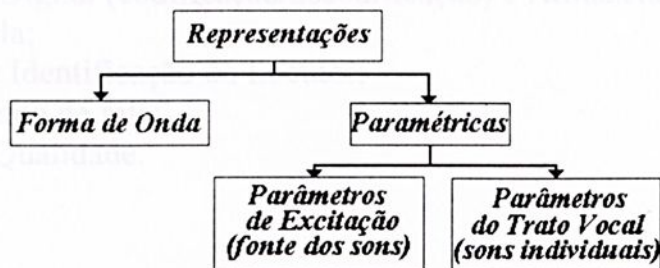


Figura 10. Representações dos sinais de voz

As representações por Forma de Onda visam exatamente a preservação da forma da onda do sinal de voz analógico através de um processo de amostragem e quantização. Já as representações Paramétricas têm como saída um modelo para produção do sinal de voz. Este último tipo de representação parte de uma representação digital da forma de onda do sinal (amostragem e quantização), que é então, processada para a obtenção dos parâmetros do modelo que melhor representa o sinal de entrada. Nas representações, após o processamento, espera-se recuperar o sinal original com o mínimo de distorção possível.

V.2 REPRESENTAÇÃO DIGITAL DA FORMA DE ONDA DA VOZ

“Watson, if I can get a mechanism which will make a current of electricity vary its intensity as the air varies in density when sound is passing through it, I can telegraph any sound, even the sound of speech” – A.G. Bell [1]

O princípio básico embutido na grande invenção de Bell é fundamental para muitos dos dispositivos, sistemas de gravação, transmissão e processamento de sinais de voz.

Na figura 11 são apresentados os quadros básicos para conversão da forma de onda analógica para sua correspondente representação digital. Assim, o sinal de voz analógico é amostrado, tendo uma representação única se é de banda limitada e se a taxa de amostragem é ao menos duas vezes a máxima frequência do sinal (Frequência de Nyquist).

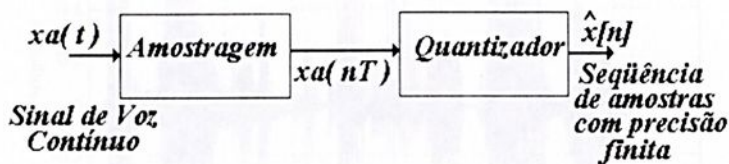


Figura 11. Representação Digital de Formas de Onda

O tipo de aplicação é, provavelmente, a consideração mais importante na escolha da representação do sinal e dos métodos de processamento a serem aplicados. Dentre as inúmeras áreas de aplicação de comunicação por voz, têm-se:

- Transmissão Digital (codificação/decodificação) e Armazenamento;
- Síntese de Fala;
- Verificação e Identificação do Locutor;
- Reconhecimento de fala;
- Melhoria da Qualidade.

V.3 MÉTODOS TEMPORAIS DE PROCESSAMENTO

O objetivo no processamento do sinal de voz é obter uma representação mais conveniente ou mais útil da informação contida no sinal. A precisão necessária no processamento é determinada pela informação a ser preservada ou intensificada no sinal [1].

Os métodos de processamento no domínio do tempo são aqueles que envolvem diretamente a forma de onda do sinal de voz. Dentre as diferentes medidas no domínio do tempo, pode-se citar:

- Taxa de Cruzamentos por Zero;
- Energia.

Estas medidas no domínio do tempo são de implementação relativamente simples utilizando o processamento digital e fornecem uma base útil para a estimação de importantes características do sinal de voz. Estas implementações têm como suposição base que as propriedades do sinal de voz variam lentamente com o tempo, ou seja, consideram o sinal de voz quasi-estacionário num intervalo curto de tempo (até 40ms). Para tanto, efetuam o processamento em intervalos de tempo do sinal (e.g. 32 ms, ou 512 amostras a 16kHz).

Na figura 12 tem-se apresentada a forma de onda de um sinal de voz típico (16000 amostras/segundo).

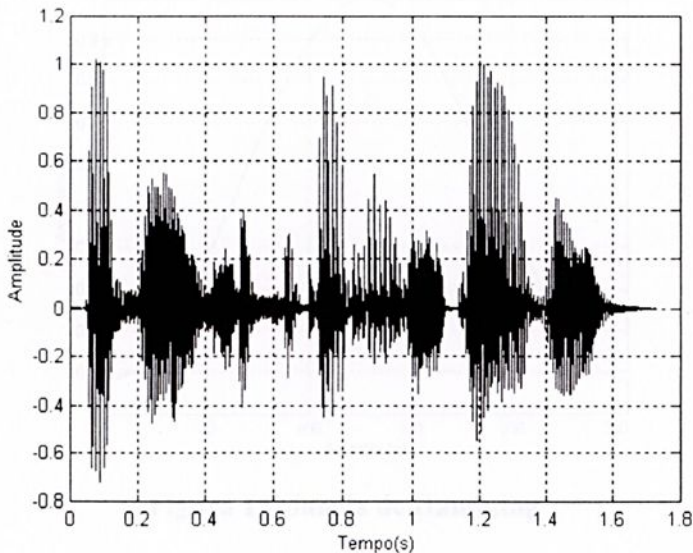


Figura 12. Sinal de Voz

Muitas das técnicas de processamento podem ser representadas matematicamente da seguinte forma:

$$Q_n = \sum_{m=-\infty}^{\infty} T\{x[m]\} \cdot w[n-m]$$

Aqui, o sinal de voz (normalmente após filtragem para isolar a banda de frequências desejada) é submetido à transformação $T[\cdot]$. A sequência resultante é então multiplicada por uma janela posicionada no instante de tempo correspondente à amostra n . O produto é então somado para todos os valores não nulos. Portanto os valores de Q_n formam uma sequência de valores médios localmente ponderados da sequência $T\{x[m]\}$. As operações necessárias para obtenção de Q_n são ilustradas na figura 13.

A equação anterior está exatamente na forma de uma convolução discreta da janela, $w[n]$, com a sequência, $T\{x[n]\}$. Assim, Q_n pode se interpretada como a saída de um Sistema Linear Invariante no Tempo com reposta ao impulso $h[n] = w[n]$.

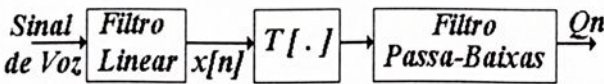


Figura 13. Representação Geral do Princípio da Análise de Curto Prazo

Como as características espectrais da voz podem mudar drasticamente ao longo do tempo, o comprimento, tipo e localização da função janela são considerações importantes. Se a janela for demasiadamente longa, a variação das características espectrais pode ter seus efeitos mascarados; se a janela for extremamente curta, tem-se como resultado imprecisões espectrais devido ao baixo número de amostras para análise. Uma janela de Hamming de 16 a 32 ms de duração pode ser usada para a análise de voz, conforme figura 14.

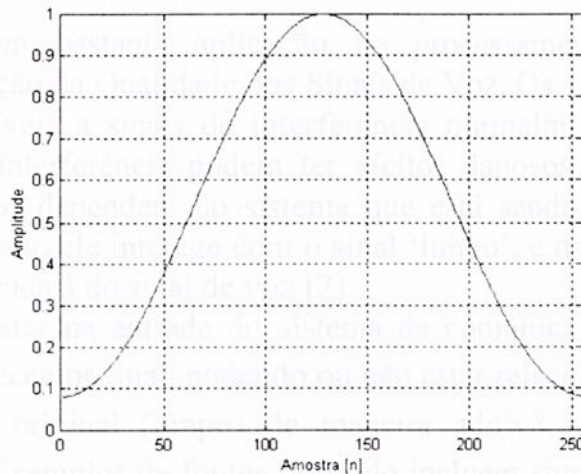


Figura 14. Janela de Hamming

Capítulo VI**AVALIAÇÃO DA QUALIDADE DE SINAIS DE VOZ**

Neste capítulo é apresentado o contexto onde o trabalho se insere, ou seja, na avaliação da qualidade de sinais de voz. Dessa forma, discute-se o conceito de qualidade e os tipos de medidas utilizadas na sua determinação, sendo que estas medidas são agrupadas em dois importantes grupos: Subjetivas e Objetivas.

Um campo com bastante aplicação no processamento de sinais de voz compreende a Avaliação da Qualidade dos Sinais de Voz. Os sistemas de comunicação por voz são susceptíveis a sinais de interferência normalmente classificados como ruído. Os sinais de interferência podem ter efeitos danosos na performance destes sistemas. Estes efeitos dependem do sistema que está sendo usado, da natureza do ruído e da maneira como ele interage com o sinal 'limpo', e da intensidade relativa do ruído, quando comparada à do sinal de voz [2].

O ruído pode estar na entrada do sistema de comunicação, no canal (meio de transmissão), ou no receptor final, podendo ou não estar relacionado ao sinal. Ele pode acompanhar o sinal original (limpo) de maneira aditiva, multiplicativa ou outra maneira mais geral. Exemplos de fontes de ruído incluem sinais de voz competitivos, sons de fundo como música, máquinas, eletrodomésticos, vento e tráfego, reverberação e ruído de canal.

A finalidade de avaliar e melhorar o sinal de voz é minimizar os efeitos do ruído nos sistemas de comunicação de voz e de intensificar as características auditivas dos sinais.

A medida de desempenho é dependente do sistema. Para sistemas que compreendem gravação de voz ruidosa, ou sistemas de comunicação analógica, o objetivo da melhoria do sinal é reforçar aspectos perceptuais do sinal ruidoso. Para sistemas digitais de comunicação, o objetivo é melhorar aspectos perceptuais do sinal codificado/decodificado.

VI.1 QUALIDADE

Pode-se definir qualidade de um sinal de voz como uma medida subjetiva que reflete o conforto auditivo que este sinal proporciona, ou seja, se o mesmo é agradável ao ouvido.

Há várias áreas em processamento de voz onde a qualidade do sinal é de grande importância: Codificação, Síntese, Reconhecimento, Melhoria do Sinal de Voz, etc. Com o aumento da complexidade e capacidade dos modernos sistemas de comunicação de sinais de voz, algoritmos mais eficientes tornam-se necessários para garantir um fluxo rápido e eficaz de informações. Para estabelecer os meios de comparação entre os algoritmos de codificação, uma variedade de técnicas de avaliação foi formulada.



De maneira geral, os testes para avaliação da qualidade dos sinais processados são divididos em duas importantes classes de medidas de qualidade:

- Medidas Subjetivas;
- Medidas Objetivas.

Melhor será a medida de avaliação de qualidade objetiva quanto mais próximo da percepção real (subjetiva) ela chegar.

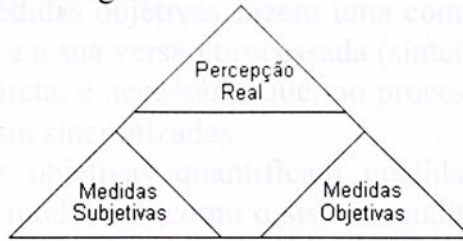


Figura 15. Percepção da voz

Na avaliação da *Qualidade* do sinal de voz a preocupação recai sobre ‘como’ um locutor exprime uma elocução e pode incluir atributos tais como: naturalidade e capacidade de identificação do locutor. Em contraste, *Inteligibilidade* preocupa-se com ‘o que’ o locutor disse, ou seja, com o conteúdo de palavras contido na mensagem.

Quando se justapõem qualidade e inteligibilidade, está se fazendo referência a todas as características que contribuem para a caracterização da voz como sendo de ‘boa’ ou ‘má’ qualidade. Interessante dizer que algumas características não são necessariamente importantes para a compreensão correta do que é dito.

De modo geral, há três abordagens em processamento de voz nas quais é desejável medir a distorção:

- PRIMEIRO: pode-se querer uma medida de qualidade para avaliar a quantidade de distorção introduzida por um dado algoritmo.
- SEGUNDO: pode-se desejar medir o nível de distorção introduzido por um canal ruidoso.
- TERCEIRO: pode-se medir a performance de um algoritmo de melhoria.

VI.2 MEDIDAS SUBJETIVAS DE QUALIDADE

As medidas subjetivas são baseadas em comparações entre o sinal original e o processado, feitas por um grupo de ouvintes. Estes ouvintes classificam subjetivamente a qualidade da voz tomando como base uma escala pré-determinada. Assim, os sinais (original e processado) são apresentados para o grupo de ouvintes que emitem notas, ou seja, classificam os sinais.

VI.3 MEDIDAS OBJETIVAS DE QUALIDADE

Em virtude dos problemas para se realizar massivamente testes subjetivos, alguns algoritmos para a estimação da qualidade do sinal de fala processado foram criados. O critério de desempenho para uma medida objetiva de qualidade do sinal de voz é a sua correlação com estimativas subjetivas de qualidade. Como o objetivo da codificação é produzir um sinal de voz que seja percebido pelo sistema auditivo de forma natural e



livre de distorções, é compreensível que medidas subjetivas de qualidade sejam meios preferíveis de classificação de qualidade.

A análise da correlação é aplicada para determinar a habilidade da medida objetiva de qualidade em prever a qualidade julgada pelos ouvintes na avaliação subjetiva.

A classificação de qualidade através de medidas objetivas fornece meios quantitativos, repetitivos de comparação da performance de codificadores de voz (VOCODERS) [2]. As medidas objetivas fazem uma comparação direta entre o sinal original (ou de referência) e a sua versão processada (sintetizada ou melhorada). Como é feita uma comparação direta, é necessário que, no processo de cálculo, as formas de onda dos sinais de voz sejam sincronizadas.

Muitas das medidas objetivas quantificam qualidade como uma medida de distância numérica ou um modelo de como o sistema auditivo avalia os sinais de fala. Como a distorção introduzida por sistemas de codificação de voz, ruído de fundo e algoritmos de melhoria varia, uma gama de medidas de qualidade e testes emergiu para diferentes aplicações.

O método normalmente utilizado na implementação de medidas objetivas é dividir o sinal de voz em quadros de curta duração, e então calcular uma medida de distância/distorção de cada um dos quadros entre o sinal original e o processado. Feito isso, pode-se formar uma medida final por meio da combinação das medidas de distorção de cada quadro ou observar os valores obtidos ao longo do processamento, formatados em um gráfico de distância/distorção.

Qualquer função que atenda às propriedades na definição acima é uma legítima métrica do vetor de espaço. Tem-se ainda que muitas das métricas utilizadas em processamento de voz são casos particulares da métrica Minkowski, ou ao menos bastante próximas.

Tomemos x_1 e y_1 os k -ésimos elementos componentes dos vetores de N elementos x e y , respectivamente. Dessa forma, a métrica de Minkowski de ordem s , ou a métrica L_s entre esses vetores é dada por:

$$d(x, y) = \left(\sum_{i=1}^N |x_i - y_i|^s \right)^{1/s}$$

Casos particulares são expressos por:

1. A métrica L_1

$$d(x, y) = \sum_{i=1}^N |x_i - y_i|$$

2. A métrica L_2 ou Euclidiana

$$d(x, y) = \sqrt{\sum_{i=1}^N |x_i - y_i|^2} = \sqrt{(x - y) \cdot (x - y)}$$

3. A métrica L_∞ ou Chebyshev (corresponde a métrica de Minkowski com $s \rightarrow \infty$)



Capítulo VII

MEDIDAS OBJETIVAS IMPLEMENTADAS

Neste capítulo, além de uma discussão sobre o significado matemático de uma medida de distância, são também apresentados alguns métodos de determinação objetivas de qualidade. Dentre os métodos disponíveis, são apresentados: Relação Sinal/Ruído, Distância Itakura-Saito, Distância Cepstral e o PSQM.

De maneira genérica, para a obtenção das chamadas medidas de distância são tomados dois vetores $\bar{x}$ e $\bar{y}$ em um espaço multidimensional. Frequentemente, deseja-se saber “quão distantes um do outro” eles estão [2]. Estes vetores corresponderão a dois pontos no tempo em uma realização de um processo aleatório. O espaço cartesiano real de dimensão N , denotado por $\mathcal{R}^N$, é a coleção de todos os vetores N -dimensionais com elementos reais. Uma métrica, $d(x,y)$, em $\mathcal{R}^N$ é uma função de valores reais com três propriedades: para todos $\bar{x}, \bar{y}, \bar{z} \in \mathcal{R}^N$,

1. $d(\bar{x}, \bar{y}) \geq 0$
2. $d(\bar{x}, \bar{y}) = 0$, se e somente se $\bar{x} = \bar{y}$
3. $d(\bar{x}, \bar{y}) \leq d(\bar{x}, \bar{z}) + d(\bar{z}, \bar{y})$

Qualquer função que atenda às propriedades na definição acima é uma legítima métrica do vetor de espaço. Tem-se ainda que muitas das métricas utilizadas em processamento de voz são casos particulares da métrica Minkowski, ou ao menos, bastante próximas.

Tomemos x_k e y_k , os k -ésimos elementos componentes dos vetores de N elementos $\underline{x}$ e $\underline{y}$, respectivamente. Dessa forma, a métrica de Minkowski de ordem s , ou a métrica l_s , entre esses vetores é dada por:

$$d_s(\bar{x}, \bar{y}) \triangleq \sqrt[s]{\sum_{k=1}^N |x_k - y_k|^s}$$

Casos particulares são expressos por:

1. A métrica l_1

$$d_1(\bar{x}, \bar{y}) = \sum_{k=1}^N |x_k - y_k|$$

2. A métrica l_2 ou Euclidiana

$$d_2(\bar{x}, \bar{y}) = \sqrt{\sum_{k=1}^N |x_k - y_k|^2} = \sqrt{[\bar{x} - \bar{y}]^T \cdot [\bar{x} - \bar{y}]}$$

3. A métrica l_∞ ou Chebyshev (corresponde a métrica de Minkowski com $s \rightarrow \infty$)



$$d_{\infty}(\bar{x}, \bar{y}) = \max_k |x_k - y_k|$$

Deve-se notar que a norma l_s de um vetor $\underline{x}$, denotada por $\|\underline{x}\|_s$, é definida como:

$$\|\underline{x}\| \triangleq \sqrt[s]{\sum_{k=1}^N |x_k|^s}$$

Dentre os vários métodos objetivos para avaliação da distorção de sinais de voz, foram selecionados os seguintes: Relação Sinal/Ruído, Distância Itakura-Saito, Distância Cepstral e PSQM (Medida Perceptual de Qualidade de Voz)

VII.1 RELAÇÃO SINAL/RUÍDO (SNR)

A Relação Sinal/Ruído (SNR) é a medida mais amplamente utilizada em sistemas analógicos e de codificação de forma de onda. Ao longo dos anos foram desenvolvidas diferentes variações, incluindo SNR clássica, SNR segmentada e SNR segmentada ponderada na frequência, entre outras [2].

É importante notar que as medidas baseadas na SNR são apropriadas somente para sistemas de codificação e melhoria que visam reproduzir a forma de onda original, pois o seu procedimento de cálculo mede o erro entre as amostras correspondentes dos sinais original e processado.

Tomemos um sinal de voz ruidoso denotado por $y[n]$ e $s[n]$ seu equivalente sem ruído. Consideremos também, $\hat{s}[n]$ o sinal correspondente ao processamento (melhoria) do sinal ruidoso. O sinal de erro no tempo pode então ser escrito como:

$$\varepsilon[n] = s[n] - \hat{s}[n]$$

$$\text{A energia do erro é: } E_{\varepsilon} = \sum_{n=-\infty}^{\infty} \varepsilon^2[n] = \sum_{n=-\infty}^{\infty} \{s[n] - \hat{s}[n]\}^2$$

$$\text{A energia contida no sinal de voz é: } E_s = \sum_{n=-\infty}^{\infty} s^2[n]$$

A resultante medida da SNR (em dB) é obtida como:

$$SNR = 10 \cdot \log_{10} \frac{E_s}{E_{\varepsilon}} = 10 \cdot \log_{10} \frac{\sum_n s^2[n]}{\sum_n \{s[n] - \hat{s}[n]\}^2}$$

Como é o caso para muitas medidas de qualidade utilizadas na avaliação de sistemas de codificação ou melhoria, o sinal de voz original $s[n]$ é necessário para se determinar o rendimento da implementação. Portanto, tais medidas são usadas em procedimentos de tempo não-real, ou seja, simulações onde ambos os sinais de voz (original e processado) estão disponíveis.

Um dos principais benefícios da SNR é a sua simplicidade matemática. Todavia, tem sido muito bem demonstrado que a SNR clássica é um estimador ruim da qualidade do sinal de voz para uma ampla faixa de distorções. Isto se deve ao fato de que a SNR não é particularmente bem co-relacionada com nenhum atributo subjetivo



da qualidade do sinal de voz, além de ponderar, de maneira uniforme, todos os erros no domínio do tempo da forma de onda da voz.

Um alto valor de SNR pode ser obtido, por exemplo, se uma elocução contém uma alta concentração de segmentos com voz, uma vez que o ruído tem um maior efeito perceptual em segmentos de baixa energia (ou segmentos sem voz).

Uma medida melhorada pode ser obtida se a SNR é calculada em quadros de curta duração (janelas) onde é tomada a média dos resultados. Esta medida é chamada SNR segmentada (SNR_{seg}), e é formulada como se segue:

$$SNR_{seg} = \frac{1}{M} \sum_{j=0}^{M-1} 10 \cdot \log_{10} \left[\sum_{n=m_j-N+1}^{m_j} \frac{s^2[n]}{\{s[n] - \hat{s}[n]\}^2} \right]$$

sendo $m_0, m_1, \dots, m_{M-1}$ os tempos finais para os M quadros, cada um dos quais com comprimento N .

A segmentação da SNR permite que a medida objetiva atribua pesos iguais para porções de alta e baixa energia do sinal de voz, tornando o processo mais eficiente.

Em alguns casos podem surgir problemas com a medida de SNR_{seg} , principalmente se quadros de silêncio absoluto forem incluídos, pois grandes valores negativos da SNR_{seg} serão gerados. Este problema é minimizado por meio da identificação de períodos de silêncio e sua exclusão dos cálculos da SNR_{seg} .

VII.2 MEDIDA ITAKURA-SAITO

Uma das características do sistema auditivo humano é que ele é relativamente insensível à distorção de fase. Isto é particularmente importante em muitos sistemas de comunicação por voz (exemplo, telefone ou rádio) [2]. Dessa forma, muitos dos sistemas de codificação focam-se somente na magnitude do espectro do sinal de voz, desprezando as características da fase. Como resultado, a forma de onda codificada pode ser muito diferente do sinal original, embora seja percebida da mesma maneira pelo ouvido humano. Medidas como as baseadas na SNR, que obtêm uma medida de distorção baseada em diferenças entre as amostras das formas de onda do sinal original e do processado, não fornecem uma medida significativa de performance quando as duas formas de onda diferem em seus espectros de fase. Todavia, embora o ouvido humano não perceba de maneira acentuada as variações da fase do sinal de voz, medidas de distância que sejam sensíveis a essa variação são importantes na medida em que tornam o processo de avaliação mais apurado.

Uma das medidas de distância de maior sucesso é a medida Itakura-Saito. Esta medida é baseada na diferença entre os modelos *somente com pólos* do sinal de referência e da forma de onda do sinal codificado/processado. A medida de distância é calculada entre conjuntos de parâmetros de Predição Linear (LP) estimados a partir de quadros sincronizados (tipicamente, quadros de 16 a 32 ms) de ambos os sinais.



A medida Itakura-Saito é muito influenciada pelas diferenças espectrais devido a não coincidência das localizações das formantes (picos no espectro de frequência), considerando que os erros de coincidência dos 'vales' espectrais não contribuem de maneira significativa para a obtenção da medida. Isto é desejável uma vez que o sistema auditivo é mais sensível a erros de localização e largura de banda das frequências formantes.

A medida Itakura-Saito, também chamada de Probabilidade Máxima de Distância Espectral, é definida como uma função de avaliação para a estimação do envelope espectral a partir de uma densidade espectral de curto prazo usando o método de probabilidade máxima [4]. A distância é representada pela equação:

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\log \frac{f(\lambda)}{g(\lambda)} + \frac{g(\lambda)}{f(\lambda)} - 1 \right) d\lambda$$

$$E = \log \left(\frac{u(f)R(f)}{u(g)R(g)} + \frac{u(g)}{u(f)R(f)} \right) \sum_{k=-p}^p A_k(f)r_k(g) - 1$$

Definindo: $d(\lambda) = \log f(\lambda) - \log g(\lambda)$

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(d(\lambda) + e^{-d(\lambda)} - 1 \right) d\lambda$$

Expandindo o integrando da equação anterior por meio da expansão de Taylor para $d(\lambda)$ na região ao redor de 0, tem-se:

$$d(\lambda) + e^{-d(\lambda)} - 1 = \sum_{i=2}^{\infty} \frac{1}{i!} \{ -d(\lambda) \}^i$$

Assim,

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\sum_{i=2}^{\infty} \frac{1}{i!} \{ -d(\lambda) \}^i \right) d\lambda$$

sendo f e g , o sinal de referência e o sinal sob observação, respectivamente, u a energia dos sinais, A_k os coeficientes dos parâmetros de probabilidade, r_k os coeficientes de autocorrelação.

VII.3 MEDIDA CEPSTRAL

Tomando como base nosso modelo de produção de voz, o sinal de saída é composto de uma sequência de excitação convoluída com a resposta impulsiva do modelo do sistema vocal. Entretanto, temos acesso somente à saída deste sistema e, freqüentemente, é desejável eliminar um de seus componentes, de modo que o outro possa ser examinado, codificado ou modelado. Porém, a eliminação de um dos dois sinais combinados é, em geral, uma tarefa muito difícil [2].

Se o objetivo for, por exemplo, avaliar algumas das propriedades do ruído, deve-se estar habilitado para extrair a informação desejada do espectro. Se o objetivo, entretanto, for remover o ruído do sinal, deve-se projetar um filtro passa-baixas para remover as componentes indesejáveis de alta frequência. Do ponto de vista do domínio da frequência, isto consiste na construção de um sistema que remova a energia espectral de alta frequência.



A análise 'Cepstral' é motivada e direcionada para problemas centrados no sinal de voz 'falado' (sem instantes de silêncio). De acordo com nosso modelo de produção de voz, o sinal de voz 'falado' é resultante da convolução da sequência de excitação $u[n]$ com a resposta impulsiva do sistema vocal $h[n]$, ou seja:

$$s[n] = u[n] * h[n]$$

Como as partes individuais não são combinadas linearmente, as técnicas lineares, aparentemente, não fornecem ajuda.

Se o propósito é determinar algumas propriedades das componentes do sinal, o Cepstro deve ser suficiente para fornecer a informação necessária. A análise Cepstral é um caso especial dentro da classe geral de métodos coletivamente conhecidos como Processamento "Homomórfico" de Sinais.

VII.3.1 Processamento Homomórfico

Um dos princípios básicos da análise de voz estabelece que se considerarmos trechos de curta duração do sinal de voz, então cada trecho pode efetivamente ser modelado como tendo sido gerado por meio da excitação de um sistema linearmente invariante no tempo. Esta excitação pode ser feita por um trem de impulsos aproximadamente periódico, por um sinal de ruído aleatório, ou por um somatório dos dois sinais.

O problema da análise de voz é estimar os parâmetros de um modelo de produção de voz e medir suas variações no tempo. Como a excitação e a resposta impulsiva de um sistema linear invariante no tempo são combinadas por meio de uma convolução, o problema da análise de voz pode também ser visto como um problema de separação das componentes desta convolução.

Sistemas homomórficos por convolução obedecem ao princípio generalizado da superposição. Este princípio, como normalmente expresso para sistemas lineares convencionais, é dado por:

$$L\{x[n]\} = L\{x_1[n] + x_2[n]\} = L\{x_1[n]\} + L\{x_2[n]\} = y_1[n] + y_2[n]$$

e

$$L\{ax[n]\} = a L\{x[n]\} = ay[n]$$

sendo que L representa o operador linear. O princípio da superposição simplesmente afirma que se um sinal de entrada é composto por uma combinação linear de sinais elementares, então a saída é uma combinação linear das saídas correspondentes, conforme apresentado na figura 16.

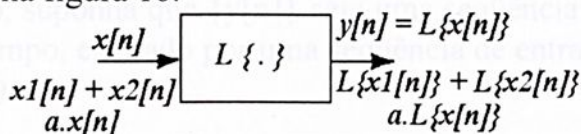


Figura 16. Representação de um sistema obedecendo ao princípio da superposição

Um resultado direto do princípio da superposição é o fato de que a saída de um sistema linear invariante no tempo pode ser expressa como uma soma de convolução, isto é:

$$y[n] = \sum_{k=-\infty}^{\infty} h[n-k].x[k] = h[n] * x[n]$$

O símbolo $*$ denota a operação de convolução discreta. Por analogia com o princípio da superposição para sistemas lineares convencionais, pode-se definir a classe de sistemas que obedece a um princípio generalizado de superposição onde a adição é substituída por uma convolução (tendo as mesmas propriedades algébricas). Assim:

$$H\{x[n]\} = H\{x_1[n] * x_2[n]\} = H\{x_1[n]\} * H\{x_2[n]\} = y_1[n] * y_2[n] = y[n]$$

Sistemas tendo a propriedade expressa pela equação anterior são denominados “Sistemas Homomórficos por Convolução”.

VII.3.2 O Cepstro

Considere uma seqüência $\{x[n]\}$ tendo uma Transformada-Z representada por $X(z)$. Se $\{x[n]\}$ é uma seqüência estável, a região de convergência de $X(z)$ inclui o círculo unitário [3]. O Cepstro complexo da seqüência $\{x[n]\}$ é definido como a seqüência $\{c_x[n]\}$, que é a transformada Z inversa de $C_x(z)$, sendo:

$$C_x(z) = \ln X(z)$$

O Cepstro complexo existe se $C_x(z)$ converge em uma região anular $r_1 < |z| < r_2$, sendo $0 < r_1 < 1$ e $r_2 > 1$. Dentro desta região de convergência, $C_x(z)$ pode ser representado pela série de Laurent:

$$C_x(z) = \ln X(z) = \sum_{n=-\infty}^{\infty} c_x[n]z^{-n}$$

Em processamento de sinais de voz o Cepstro (real) tem sido utilizado para separar e, assim, estimar o conteúdo espectral da fala a partir da frequência fundamental da voz. O Cepstro complexo é usado, na prática, para separar sinais convoluídos. O processo de separação de dois sinais convoluídos é chamado deconvolução e o uso do Cepstro complexo para realizar a separação é chamado Deconvolução Homomórfica.

VII.3.3 Deconvolução Homomórfica

É uma ferramenta útil para execução da deconvolução em várias aplicações. Para descrever este método, suponha que $\{y[n]\}$ seja uma seqüência de saída de um sistema linear invariante no tempo, excitado por uma seqüência de entrada $\{x[n]\}$. Então,

$$Y(z) = X(z).H(z)$$

sendo $H(z)$ a função de sistema. O logaritmo de $Y(z)$ é dado por:

$$C_y(z) = \ln Y(z) = \ln X(z) + \ln H(z) = C_x(z) + C_h(z)$$

Consequentemente, o Cepstro complexo da seqüência de saída $\{y[n]\}$ é expresso como a soma do Cepstro de $\{x[n]\}$ e $\{h[n]\}$, ou seja:

$$c_y[n] = c_x[n] + c_h[n]$$



Assim, observa-se que a convolução de duas seqüências no domínio do tempo corresponde à soma do Cepstro das seqüências no domínio Cepstral. O sistema para executar estas transformações é chamado Sistema Homomórfico e é ilustrado na figura 17.

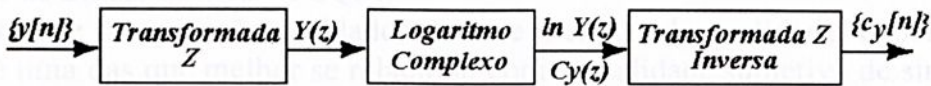


Figura 17. Sistema Homomórfico para obtenção do Cepstro

Em aplicações como processamento de voz, as características das seqüências cepstrais $\{c_x[n]\}$ e $\{c_h[n]\}$ são suficientemente diferentes, o que permite uma separação no domínio Cepstral. Especificamente, supondo que $\{c_h[n]\}$ tenha suas componentes principais (energia principal) na vizinhança de valores pequenos de n , e considerando que $\{c_x[n]\}$ tenha suas componentes concentradas em valores grandes de n , pode-se dizer que $\{c_h[n]\}$ é “Passa Baixas” e $\{c_x[n]\}$ é “Passa Altas”. Pode-se, também, separar $\{c_h[n]\}$ de $\{c_x[n]\}$ utilizando-se janelas “Passa Baixas” e “Passa Altas” apropriadas.

VII.4 PSQM

A PSQM é um método para avaliação de qualidade de voz dado pelo padrão (Recomendação) do ITU (International Telecommunication Union) – P.861 “Medição Objetiva de Qualidade de Codificadores de Voz de Banda Telefônica (300 - 3400 Hz)” [6].

A Recomendação P.861 descreve um método **objetivo** para estimação da qualidade **subjetiva** de codificadores de voz na banda telefônica (300-3400 Hz). Ela especifica a produção do sinal de voz a ser usado na medida objetiva de qualidade, o codec (codificador/decodificador) e as condições de referência para as quais o método de avaliação objetiva da qualidade tem apresentado resultados válidos. A recomendação estabelece o procedimento de cálculo da qualidade subjetiva baseada na medida de qualidade objetiva, chamado Medida de Qualidade Perceptual de Sinais de Voz (PSQM). Finalmente, permite obter uma estimativa da qualidade subjetiva a partir de resultados objetivos.

Esta Recomendação pode ser aplicada quando da avaliação dos efeitos do codec na qualidade subjetiva do sinal de voz em função do nível do sinal de entrada, dos locutores, das taxas de transmissão de bits e de transcodações.

Avaliação subjetiva da qualidade de codecs de voz é uma das tecnologias chave no desenvolvimento de redes de telecomunicação digital. A recomendação P.830 “Avaliação Subjetiva da Performance de Codecs Digitais de Largura de Banda Telefônica e Banda Larga” define metodologias subjetivas de teste para codecs de voz. Como a avaliação subjetiva de qualidade requer muito tempo e recursos financeiros, é desejável desenvolver uma metodologia objetiva para avaliação da qualidade que permita estimar a qualidade subjetiva de codecs de voz.

O desempenho destas metodologias, em termos de sua habilidade de fornecer estimativas precisas da qualidade subjetiva, tem sido investigadas no ITU-T desde os anos 80. Com isso, uma variedade de sofisticadas medidas objetivas de qualidade foi desenvolvida. Dentre as medidas desenvolvidas temos a Medida de Qualidade Perceptual de Sinais de Voz – PSQM.

Depois de comparações cuidadosas entre medidas de qualidade, concluiu-se que a PSQM é uma das que melhor se relaciona com a qualidade subjetiva de sinais de voz submetidos a um processo de codificação/decodificação.

Como mencionado anteriormente, a avaliação subjetiva de qualidade de codecs de voz pode ser feita em testes somente auditivos (via única) ou em testes de conversação (via dupla). A avaliação objetiva de qualidade descrita na recomendação P.861 “Medição Objetiva de Qualidade de Codificadores de Voz de Banda Telefônica (300 - 3400 Hz)” [6] estima a qualidade subjetiva em testes somente auditivos. Para demonstrar a performance subjetiva de um codec, os efeitos sobre uma variedade de fatores de qualidade devem ser investigados.

Quando se compara um codec com relação a outro ou com uma condição de referência baseada em resultados experimentais subjetivos, testes estatísticos que levam em consideração as distribuições dos votos subjetivos são sempre utilizados.

VII.4.1 Procedimentos Objetivos de Medição

A figura 18 ilustra o procedimento objetivo de medição.

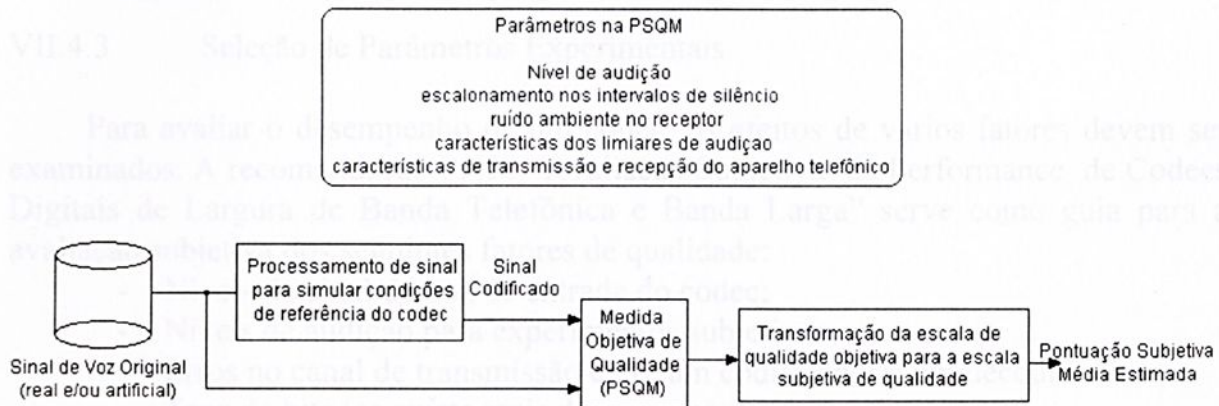


Figura 18. Procedimento de medição objetiva da qualidade

A medição objetiva da qualidade de codecs de voz requer um procedimento formado por um certo número de passos, a saber:

- Preparação dos sinais de voz, isto é, gravação da voz dos locutores e/ou geração de vozes artificiais de acordo com a Recomendação P.50 “Vozes Artificiais”;
- Seleção dos parâmetros experimentais que irão exercitar as características importantes do codec e que são passíveis de serem testadas por medições objetivas;
- Produção dos sinais de voz de referência (original) e processado (codificado);

- Cálculo da medida objetiva da qualidade do sinal de voz baseada na Medida de Qualidade Perceptual de Sinais de Voz (PSQM), utilizando os sinais de referência e codificado;
- Transformação da escala objetiva de qualidade para a escala subjetiva (quando desejável);
- Análise dos resultados.

VII.4.2 Preparação dos Sinais de Voz

Os sinais de voz para medição objetiva podem ser vozes reais ou vozes artificiais, dependendo dos objetivos do experimento. As vozes artificiais definidas na recomendação P.50 “Vozes Artificiais” reproduzem as características médias da voz humana em vários idiomas, o que é útil na estimação objetiva da qualidade subjetiva média de um codec para vários idiomas. Todavia, quando a dependência do codec em relação ao locutor ou a performance do mesmo para algum idioma em particular é a preocupação central, é recomendado que vozes reais sejam utilizadas. Em ambos os casos é importante que nenhum ruído ambiente seja adicionado.

Para que sejam obtidos resultados mais acurados, é recomendado que sejam utilizados, no mínimo, dois locutores do sexo masculino e dois do sexo feminino para cada condição de teste. Entretanto, se a dependência com o locutor deve ser testada como um fator de extrema importância, é recomendado que mais locutores sejam utilizados (8 homens, 8 mulheres e 8 crianças).

VII.4.3 Seleção de Parâmetros Experimentais

Para avaliar o desempenho de um codec, os efeitos de vários fatores devem ser examinados. A recomendação P.830 “Avaliação Subjetiva da Performance de Codecs Digitais de Largura de Banda Telefônica e Banda Larga” serve como guia para a avaliação subjetiva dos seguintes fatores de qualidade:

- Níveis do sinal de voz de entrada do codec;
- Níveis de audição para experimentos subjetivos;
- Erros no canal de transmissão entre um codificador e um decodificador;
- Taxa de bits (se existe mais do que uma disponível);
- Transcodificações;
- Ruído ambiente no transmissor;
- Música como entrada para um codec.

A medida objetiva de qualidade descrita nessa recomendação assume que:

- Os sinais de voz são "limpos" (i.e. sem ruído) no transmissor;
- Não há degradações no canal (erros de transmissão de bits, perda de quadros de sinal, etc.).

Os fatores de qualidade para os quais a Recomendação P.861 se aplica são níveis do sinal de voz de entrada, locutores (excluindo-se múltiplos locutores simultâneos) e taxa de bits e transcodificações



Adicionalmente às condições do codec, a recomendação P.830 “Avaliação Subjetiva da Performance de Codecs Digitais de Largura de Banda Telefônica e Banda Larga” recomenda o uso das condições de referência nos testes subjetivos. Estas condições são necessárias para facilitar a comparação de resultados de testes subjetivos de diferentes laboratórios ou do mesmo laboratório em épocas diferentes.

Na tabela 1 são apresentados os parâmetros utilizados no cálculo da PSQM e suas respectivas descrições.

Tabela 1. Lista de Parâmetros da PSQM

Nome	Descrição
Nb	Número de bandas no domínio de bandas críticas (Bark)
Nf	Número de amostras no quadro (tempo)
F[j]	Características de recepção em frequência do aparelho
H[j]	Características Hoth
P ₀ [j]	Limiar absoluto de audição
Δf[j]	Largura da banda j em Hertz
Δz	Largura de banda de cada sub-banda no domínio de banda crítica
γ	Expoente da função de compressão
W _{sil}	Fator de escalonamento dos intervalos de silêncio
W _{sp}	Fator de escalonamento em quadros ativos de fala

VII.4.4 Cálculo Objetivo da Qualidade

Este tópico descreve um método para medição da qualidade de sinais codificados de banda telefônica (300 - 3400 Hz) utilizando a Medida de Qualidade Perceptual de Sinais de Voz (PSQM). O objetivo da PSQM é imitar a percepção do som por parte das pessoas em situações reais. O PSQM simula experimentos nos quais as pessoas julgam a qualidade de codificadores de voz. Para tanto, realiza a comparação entre um sinal codificado e um sinal de referência. Na medida em que o PSQM é uma representação bastante fiel dos processos de percepção e julgamento humano, diferenças imperceptíveis entre os sinais de entrada e saída receberão pontuação de qualidade equivalente. Em particular, se a entrada e a saída são idênticas, o PSQM irá indicar qualidade perfeita independentemente da qualidade do sinal de entrada.

Dentro da PSQM os sinais físicos constituindo os sinais de voz original (referência) e codificado são mapeados em representações psicofísicas¹ que se comparam às nossas representações internas dos sinais de voz (as representações do cérebro).

Dentro da aproximação do PSQM, a qualidade do sinal de voz codificado é julgada com base nas diferenças na representação interna. Esta diferença é usada para o cálculo dos distúrbios do ruído como uma função do tempo e da frequência. No

¹ Psicofísica - ciência que trata da ligação entre os estímulos físicos e a maneira que eles afetam a mente.



PSQM o distúrbio de ruído médio é relacionado diretamente com a qualidade do sinal de voz codificado.

A transformação do domínio físico (externo) para o domínio psicofísico (interno) é executada por meio de três operações:

- Mapeamento Tempo-Frequência;
- Transformação ou mapeamento da escala de Frequência;
- Transformação ou mapeamento da escala de Intensidade (compressão).

Além do modelamento perceptual, o método PSQM também usa o modelamento cognitivo de modo a obter altas correlações entre medidas subjetivas e medidas objetivas.

A figura 19 mostra um modelo da filosofia básica utilizada no desenvolvimento do PSQM. Este é um modelo computacional, consistindo de um modelo perceptual e um cognitivo, que é utilizado para comparar a saída do codec de voz com o sinal de entrada.

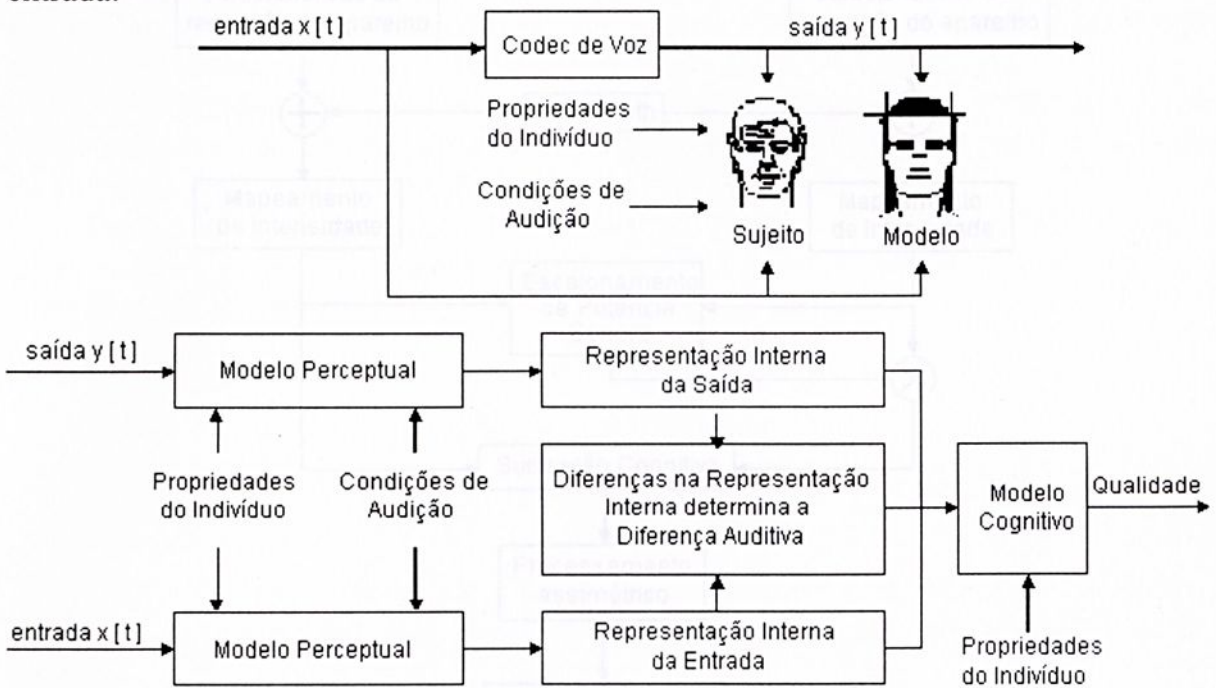


Figura 19. Modelo da Filosofia do Método PSQM

Um diagrama de blocos do algoritmo da PSQM é esquematizado na figura 20.

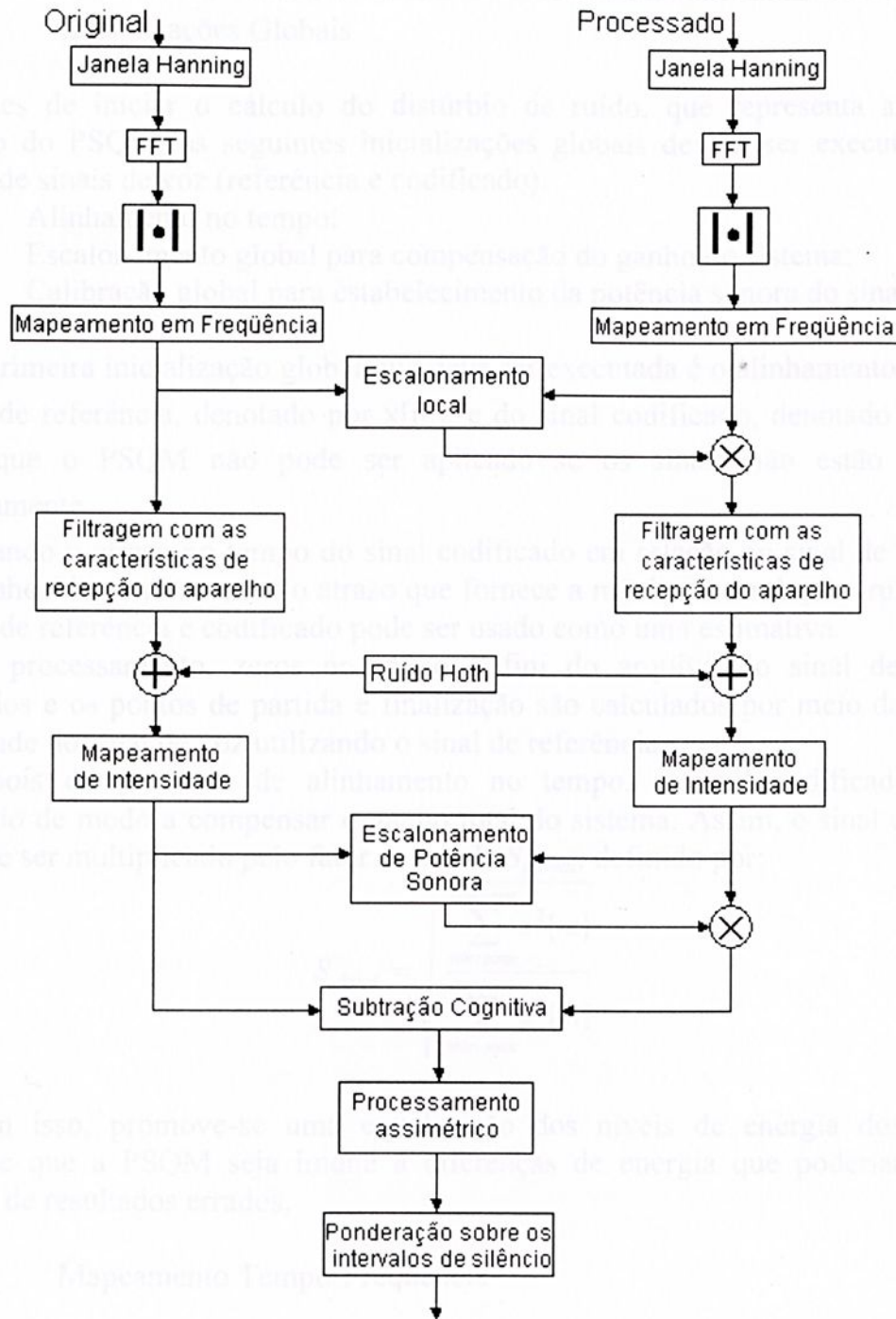


Figura 20. Algoritmo da PSQM

Na figura 20 são apresentados dois ramos de processamento que correspondem aos processos aos quais cada um dos sinais é submetido, de forma a se obter os modelos de percepção da voz.

VII.4.5 Inicializações Globais

Antes de iniciar o cálculo do distúrbio de ruído, que representa a saída do algoritmo do PSQM, as seguintes inicializações globais devem ser executadas para cada par de sinais de voz (referência e codificado):

- Alinhamento no tempo;
- Escalonamento global para compensação do ganho do sistema;
- Calibração global para estabelecimento da potência sonora do sinal.

A primeira inicialização global que deve ser executada é o alinhamento no tempo do sinal de referência, denotado por $x[m]$, e do sinal codificado, denotado por $y[m]$. Isto porque o PSQM não pode ser aplicado se os sinais não estão alinhados adequadamente.

Quando o atraso no tempo do sinal codificado em relação ao sinal de referência não é conhecido teoricamente, o atraso que fornece a máxima correlação cruzada entre os sinais de referência e codificado pode ser usado como uma estimativa.

No processamento, zeros no início e fim do arquivo do sinal de voz são descartados e os pontos de partida e finalização são calculados por meio da detecção de atividade do sinal de voz utilizando o sinal de referência.

Depois do processo de alinhamento no tempo, o sinal codificado $y[m]$ é escalonado de modo a compensar o ganho total do sistema. Assim, o sinal codificado $y[m]$ deve ser multiplicado pelo fator de escala S_{global} , definido por:

$$S_{global} = \sqrt{\frac{\sum_{m=start\ point}^{stop\ point} x^2[m]}{\sum_{m=start\ point}^{stop\ point} y^2[m]}}$$

Com isso, promove-se uma equalização dos níveis de energia dos sinais e garante-se que a PSQM seja imune a diferenças de energia que poderiam levar a obtenção de resultados errados.

VII.4.6 Mapeamento Tempo-Frequência

O mapeamento do domínio do tempo para o domínio da frequência é implementado por meio de uma transformada de Fourier de Curto Prazo com janelamento de Hanning, o que resulta em uma representação tempo-frequência com uma resolução constante em ambos os domínios.

Janelamento

O sinal de referência $x[n]$ e o sinal codificado $y[n]$ são janelados utilizando uma janela de Hanning:

$$\begin{aligned} xw[n] &= w[n] \cdot x[n] \\ yw[n] &= w[n] \cdot y[n] \end{aligned}$$



sendo $w[n]$ a função janela.

A função de janelamento é calculada como:

$$w[n] = 0.5 \left(1 - \cos \left(\frac{2\pi n}{N_f} \right) \right), \text{ para } 0 \leq n \leq N_f - 1$$

E seu formato pode ser observado na figura 21.

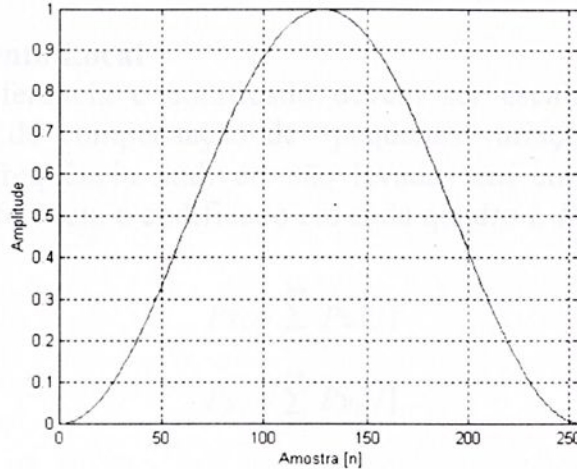


Figura 21. Janela Hanning

Os cálculos são refeitos quadro a quadro. Utilizam-se quadros de 256 amostras de comprimento para uma taxa de amostragem de 8kHz e 512 amostras para uma taxa de 16kHz. Além disso, utiliza-se uma sobreposição de 50% entre quadros adjacentes.

VII.4.7 Mapeamento em Frequência e filtragem

Este item descreve primeiro o mapeamento da escala Hertz para a escala de banda crítica, levando a uma representação SPDD (Densidade de Energia da Amostragem do Pitch) em cada quadro de amostras. A SPDD do sinal codificado é escalonada dentro de cada quadro depois que ambos os sinais original e codificado são filtrados (banda telefônica) e ruído Hoth é adicionado para simular o ambiente de comunicação. Finalmente, os sinais são filtrados com uma função de transferência característica do ouvido externo em relação ao interno.

A – Densidade de Energia da Amostragem do Pitch (SPDD)

O índice de frequência k em Hz é transformado para um índice de pitch j no domínio de banda crítica por meio de um mapeamento na escala de frequências. Primeiro, a escala de banda crítica é dividida em bandas de intervalos iguais, calculada para cada banda um valor de densidade de energia de pitch (uma amostra) a partir das amostras (normalmente mais do que uma) da densidade espectral de energia na banda correspondente na escala Hertz. As densidades amostradas de energia de pitch $Px'_i[j]$ e $Py'_i[j]$ para a banda j no quadro i são calculadas por meio de:

$$Px'_i[j] = S_p \cdot \frac{\Delta f_j}{\Delta z} \cdot \frac{1}{I_i[j] - I_j[j] + 1} \cdot \sum_{l=I_j[j]}^{I_i[j]} Px_i[k]$$

$$Py_i[j] = S_p \cdot \frac{\Delta f_j}{\Delta z} \cdot \frac{1}{I_i[j] - I_f[j] + 1} \cdot \sum_{l_i[l]}^{l_i[U]} Py_i[k]$$

sendo que $I_f[j]$ é o índice da primeira e $I_i[j]$ é o índice da última amostra na escala em Hertz para a banda j ; Δf_j é a largura de banda na banda j em Hertz, Δz é a largura de banda de cada sub-banda no domínio de banda crítica, e S_p é o fator de calibração da energia de pitch.

B – Escalonamento Local

Os sinais de referência e codificado devem ser escalonados dentro de cada quadro, como forma de compensação de pequenas variações de ganho. Somente componentes tempo-freqüência audíveis são levadas em consideração. As energias totais dos sinais de referência e codificado em cada quadro i , Px'_i e Py'_i , são calculadas como:

$$Px'_i = \sum_{j=1}^{Nb} Px_i[j]$$

$$Py'_i = \sum_{j=1}^{Nb} Py_i[j]$$

com Nb sendo o número total de bandas.

C – Filtragem de Banda Telefônica

$Px'_i[j]$ e $Py'_i[j]$ devem ser filtrados com um filtro que simula as características dos receptores telefônicos:

$$PFx_i[j] = F[j] \cdot Px'_i[j]$$

$$PFy_i[j] = F[j] \cdot Py'_i[j]$$

sendo $F[j]$ a resposta em freqüência na banda j das características de recepção do aparelho.

D – Ruído Hoth

No uso normal de telefones, o sinal de voz é perturbado por sons nas proximidades do ambiente de recepção. No PSQM, este efeito é modelado pela adição de ruído Hoth nos sinais de referência e codificado. O ruído Hoth é adicionado à densidade de energia de pitch para cada valor de j , utilizando a função Densidade Espectral de Energia:

$$PHx_i[j] = H[j] \cdot PFX_i[j]$$

$$PHY_i[j] = H[j] \cdot PFY_i[j]$$

sendo $H[j]$ o energia do ruído Hoth na banda j dado na Tabela 2.

A tabela 2 [6], embora apresentada de forma resumida, é utilizada em todo o processo de cálculo do PSQM. Dela são extraídos os valores dos parâmetros utilizados no algoritmo.



Tabela 2. Frequências de Bandas Críticas e Características de Filtragem

Banda número, j	Frequência Superior [Hz]	First FFT Bin in Band j, I _f	Last FFT Bin in Band j, I _l	Características de Recepção, F	Limiar de Audição, P ₀	Ruído Hoth, H
0	15.6	0	0	Descartadas no processamento		
1	46.9	1	1	2.45E-06	3.89E+07	1.72E+04
2	78.1	2	2	9.24E-06	1.12E+06	1.72E+04
3	109.4	3	3	3.56E-05	1.26E+05	1.72E+04
4	140.6	4	4	2.59E-04	1.86E+04	1.22E+04
5	171.9	5	5	1.18E-03	6.17E+03	8.49E+03
6	203.1	6	6	7.48E-03	2.29E+03	6.31E+03
7	234.4	7	7	3.19E-02	9.33E+02	4.91E+03
8	265.6	8	8	7.31E-02	4.37E+02	3.95E+03
9	296.9	9	9	1.37E-01	2.29E+02	3.26E+03
Continua.....						
53	3572.	109	114	1.55E-01	8.32E+00	2.69E+02
54	3765.	115	120	3.02E-02	8.32E+00	2.47E+02
55	3971.	121	127	2.03E-03	8.32E+00	2.25E+02
56	4193.	128	134	1.52E-04	8.32E+00	2.06E+02

Obs.: Baseada em uma taxa de amostragem de 16 kHz

VII.4.8 Distorção de Intensidade

Depois do cálculo das densidades de energia de pitch amostradas, que levam em consideração a filtragem de banda telefônica e o ruído Hoth, a escala de intensidade é mapeada para uma escala de potência sonora levando a uma Função Densidade de Potência Sonora Comprimida amostrada. Esta distorção de intensidade simula as transformações que ocorrem com o som ao passar do ouvido externo para o interno.

A partir das densidades de energia de pitch $PHx_i[j]$ e $PHy_i[j]$, as amostragens das densidades de potência sonora comprimida $Lx_i[j]$ e $Ly_i[j]$ são calculadas usando uma função de compressão dada por Zwicker:

$$Lx_i[j] = S_i \cdot \left(\frac{P_0[j]}{0.5}\right)^\gamma \cdot \left[\left(0.5 + 0.5 \cdot \frac{PHx_i[j]}{P_0[j]}\right)^\gamma - 1\right]$$
$$Ly_i[j] = S_i \cdot \left(\frac{P_0[j]}{0.5}\right)^\gamma \cdot \left[\left(0.5 + 0.5 \cdot \frac{PHy_i[j]}{P_0[j]}\right)^\gamma - 1\right]$$

com $P_0[j]$ como o limite interno como dado na Tabela 2, e S_i como o fator de calibração de volume de pitch. Valores negativos de $Lx_i[j]$ e $Ly_i[j]$ são alterados para zero. O valor ótimo de γ encontrado nas otimizações utilizando bases de dados que resultaram de diferentes experimentos de avaliação de sinais de voz é 0.001.

VII.4.9 Modelamento Cognitivo

No contexto do PSQM, todas as operações que não podem ser efetuadas somente no sinal referência ou somente no sinal codificado são definidas como operações cognitivas. Quatro efeitos cognitivos são discutidos aqui:

- Escalonamento de volume;
- Ruído cognitivo interno;
- Processamento assimétrico;
- Processamento de intervalos de silêncio.

A – Escalonamento de Potência Sonora

Na PSQM, a amostra da densidade de potência sonora comprimida do sinal codificado é escalonada, em cada quadro, em relação ao volume do sinal de referência:

$$Ly'_i[j] = Sl_i \cdot Ly_i[j]$$

O fator de escala Sl_i é calculado a partir da potência sonora momentânea comprimida (total) Lx_i e Ly_i :

$$Sl_i = \frac{Lx_i}{Ly_i}$$

B – Densidade de Perturbação do Ruído Amostrado

A densidade de perturbação do ruído amostrado $N_i[j]$ na banda j no quadro i é calculada como a diferença absoluta entre $Lx_i[j]$ e $Ly'_i[j]$:

$$N_i[j] = |Ly'_i[j] - Lx_i[j]| - 0.01$$

C – Processamento Assimétrico

Quando é introduzida uma nova componente tempo-frequência no sinal de voz, a qualidade subjetiva altera-se de modo a ser mais degradada do que quando uma componente de igual volume é deixada de lado pelo codec de voz. Esta assimetria é mais relevante durante os intervalos de silêncio. O ruído que está presente na fonte de sinal pode ser suprimido, levando a uma melhoria da qualidade. Se não há ruído na fonte de sinal durante os intervalos de silêncio, qualquer diferença entre os sinais de fala original e codificado leva a um decréscimo na qualidade.

Além disso, quando a componente tempo-frequência é excluída da fonte de sinal (não codificada pelo codec), o sinal remanescente ainda é um evento auditivo coerente. Se uma nova componente tempo-frequência não relacionado é introduzido no sinal codificado (ou seja, uma distorção), o sinal formado pode ser decomposto em duas partes, o sinal original e a distorção. Esta decomposição torna o ruído mais relevante.

O efeito de assimetria é quantificado por $C_i[j]$ e levado em consideração no ruído de distúrbio no quadro i , N_i :

$$N_i = \sum_{j=1}^{Nb} N_i[j] \cdot C_i[j] \cdot \Delta z$$

sendo:

$$C_i[j] = \left(\frac{PHy_i[j] + 1}{PHx_i[j] + 1} \right)^{0.2}$$

com $PHx_i[j]$ e $PHy_i[j]$ como as energias dos sinais original e codificado, respectivamente, dentro do quadro i e da banda j .



D – Distúrbio de Ruído Incluindo Processamento de Intervalos de Silêncio

No PSQM, os intervalos de silêncio são levados em consideração utilizando um fator de escala, W_{sil} , que depende do contexto dos experimentos subjetivos.

Os volumes médios de ruído, N_{spav} e N_{silav} , podem agora ser calculados em quadros ativos de fala e em quadros de silêncio, respectivamente:

$$N_{spav} = \frac{1}{M_{sp}} \sum_{i \text{ for active speech frames}} N_i$$

$$N_{silav} = \frac{1}{M_{sil}} \sum_{i \text{ for silent frames}} N_i$$

sendo M_{sp} o número de quadros de amostras com voz ativa e M_{sil} o número de quadros de silêncio.

A influência dos intervalos de silêncio depende diretamente do comprimento desses intervalos. Se o sinal original não contém qualquer intervalo de silêncio, a influência é zero. Se o sinal original contém um certo percentual de quadros de silêncio, a influência é proporcional a este percentual. Utilizando um conjunto de condições de contorno triviais, pode-se mostrar que o escalonamento correto é:

$$N_{w_{sil}} = \frac{W_{sp} \cdot p_{sp}}{W_{sp} \cdot p_{sp} + p_{sil}} \cdot N_{spav} + \frac{p_{sil}}{W_{sp} \cdot p_{sp} + p_{sil}} \cdot N_{silav}$$

com p_{sil} sendo a fração de quadros de silêncio; p_{sp} sendo a fração de quadros de fala ativos ($p_{sil} + p_{sp} = 1.0$); W_{sil} sendo o fator de escala nos intervalos de silêncio;

$W_{sp} = \frac{1 - W_{sil}}{W_{sil}}$, e $N_{w_{sil}}$ sendo distúrbio de ruído corrigido com um fator de escala W_{sil} para os intervalos de silêncio.

O distúrbio de ruído $N_{w_{sil}}$, chamado de valor PSQM nos itens seguintes, pode ser utilizado para prever a qualidade percebida subjetivamente do sinal de voz obtida no método de Índice Absoluto de Categoria (ACR) utilizando a escala de Qualidade de Audição.

NOTAS

- O valor de $N_{w_{sil}}$ tem um limite superior de 6.5;
- Para sinais de voz contendo longos intervalos de silêncio, o escalonamento é diferente daquele para sinais de voz contendo somente curtos períodos de silêncio. Além disso, o ruído na gravação dos sinais a serem utilizados também tem influência no escalonamento dos intervalos de silêncio. Para sinais de voz que não contenham intervalos de silêncio, o escalonamento não é relevante e $N_{w_{sil}}$ torna-se igual a N_{spav} .

VII.4.10 Transformação de Escala de Qualidade: Objetiva → Subjetiva

A saída do algoritmo descrito anteriormente, que é chamada de valor PSQM, indica o grau de degradação da qualidade do sinal devido à codificação do mesmo. Então, quando não é necessário estimar a qualidade subjetiva em uma escala



específica, o valor PSQM por si só já é extremamente útil. Entretanto, para estimar a qualidade subjetiva em uma escala de qualidade tal como MOS [5], o valor PSQM deve ser transformado como descrito no próximo item.

Pontuação de Opinião Média (MOS)

Na avaliação subjetiva de performance de codecs, o método ACR (Índice Absoluto de Categoria) utilizando a escala de Qualidade de Audição especificado na recomendação P.800 “Métodos para Determinação Subjetiva da Qualidade de Transmissão” [7] é freqüentemente utilizado, resultando em qualidade subjetiva em termos de MOS. Como a relação entre os valores da MOS e PSQM não são necessariamente os mesmos para diferentes idiomas ou, até mesmo, para diferentes testes subjetivos de um mesmo idioma, é difícil determinar uma função única que transforme o valor da PSQM para o valor da MOS estimada. Na prática, entretanto, é necessário derivar tal mapeamento entre ambas escalas para idiomas e testes subjetivos considerados individualmente.

VIII.1.1. Resumo

A recomendação P.800 [5] “Avaliação Subjetiva da Performance de Codecs Digitais de Larga e Banda Larga” define um método de teste específico para a avaliação de processos digitais de codificação de sinais de voz. Este método permite que os efeitos de distorção introduzidos pelo processo de codificação de voz sejam levados em consideração quando da avaliação de seu desempenho. Esta Recomendação contém diretrizes sobre como avaliar a performance de codecs digitais de maneira coerente e conclusiva.

Avaliações subjetivas de codecs digitais de sinais de voz podem, em princípio, ser conduzidas utilizando testes somente auditivos ou métodos de conversação. Em muitos casos práticos os testes somente auditivos são os métodos mais praticáveis



Capítulo VIII**MEDIDA SUBJETIVA REALIZADA**

Neste capítulo é apresentada uma explicação da avaliação subjetiva de qualidade de sinais de voz realizada.

VIII.1 MOS

A MOS (Mean Opinion Score) é uma medida para avaliação de qualidade de sinais de voz e é padronizada segundo a Recomendação da ITU (International Telecommunication Union) – P.830 “Avaliação Subjetiva da Performance de Codecs Digitais de Largura de Banda Telefônica e Banda Larga”[5]. Esta recomendação descreve métodos e procedimentos para conduzir avaliações subjetivas de performance de codecs digitais de forma a se obterem resultados comparáveis.

A Telefonia Digital em redes telefônicas chaveadas (PSTN – Public Switch Telephone Network) já se tornou lugar comum. Desde os anos 60, transmissão por meio das PSTN tem sido sinônimo de codificadores A-law (64 kbit/s) ou μ -law PCM (Modulação por Código de Pulso). A combinação de um melhor conhecimento de técnicas de processamento de sinais e avanços na tecnologia de fabricação, notadamente as técnicas LSI/VLSI (Integração em Larga Escala), levaram a um crescente interesse em métodos de codificação mais eficientes que PCM. Sendo que um grande interesse está centralizado na codificação de voz com taxas de transmissão abaixo dos 16 kbit/s.

Os testes subjetivos são os métodos mais amplamente utilizados para avaliação de performance de codecs digitais. Quando o caminho de transmissão é digital e/ou não-linear, medidas objetivas simples são insuficientes. O objetivo da metodologia de testes subjetivos é medir a degradação agregada pela parte não-linear do caminho de transmissão, e consequentemente assegurar que a performance do sistema completo seja satisfatória. Para serem apropriadas a este propósito, as medidas devem ser confiáveis e realizadas de maneira a levarem em consideração a maioria das interações entre a parte não-linear e as outras partes do sistema de transmissão.

VIII.1.1 Escopo

A recomendação P.830 [5] “Avaliação Subjetiva da Performance de Codecs Digitais de Largura de Banda Telefônica e Banda Larga” define um método de teste específico para a avaliação de processos digitais de codificação de sinais de voz. Este método permite que os efeitos de distorção introduzidos pelo processo de codificação de voz sejam levados em consideração quando da avaliação de seu desempenho. Esta Recomendação contém diretivas sobre como avaliar a performance de codecs digitais de maneira coerente e conclusiva.

Avaliações subjetivas de codecs digitais de sinais de voz podem, em princípio, ser conduzidas utilizando testes somente auditivos ou métodos de conversação. Em muitos casos práticos os testes somente auditivos são os únicos métodos praticáveis



em especial durante o desenvolvimento de codecs de voz quando uma implementação de codec em tempo real não está disponível.

VIII.1.2 Procedimentos de Testes Subjetivos

Quando da execução da *Avaliação Subjetiva* de um codec, deve-se considerar a aplicação pretendida do codec sob teste. O sistema utilizado para gravação dos sinais de voz, a seleção das condições experimentais de teste e o sistema utilizado para reproduzir os sinais de voz aos ouvintes devem ser escolhidos de modo que caracterizem uma conexão telefônica. Em particular, as características dos sistemas de gravação e reprodução devem refletir as características do transmissor e receptor que será usado na aplicação pretendida para o codec.

A execução de avaliações de codecs digitais é realizada através de um número determinado de passos que permitem avaliar o comportamento dos mesmos:

- Preparação dos sinais de voz;
- Seleção das condições de interesse a testar;
- Projeto do experimento;
- Seleção do procedimento de teste e condução do experimento;
- Análise dos resultados.

VIII.1.3 Arquivos de Voz

Para que a avaliação aproxime-se o máximo possível do sistema real, é importante utilizar uma implementação representativa do sistema de transmissão onde será usado o dispositivo sob teste.

Um outro ponto importante a se observar é o conteúdo (mensagem) dos sinais a serem reproduzidos aos ouvintes. Esses sinais de voz devem consistir de sentenças simples, curtas e significativas. Estas sentenças devem ser escolhidas de modo a serem de fácil compreensão (frases do cotidiano sem conteúdo técnico, por exemplo). Além disso, as sentenças devem ser agrupadas em conjuntos de duas ou três de tal modo que não haja conexão óbvia de sentido entre aquelas de um mesmo conjunto. É importante não utilizar sentenças muito curtas ou muito longas que dificultam a avaliação é aconselhável a utilização de sentenças que tenham duração entre 2-3 segundos.

A – Produção do Material de Teste

Também para a obtenção dos sinais de teste, deve-se atentar para vários aspectos que tornarão o material mais apropriado. Assim, por exemplo, os locutores devem estar sentados (um de cada vez) em uma sala com tempo de reverberação menor que 500 ms, e com baixo nível de ruído. Além disso, as gravações devem ser feitas em um sistema de gravação de alta qualidade para que este não seja responsável pela introdução de ruídos indesejáveis.

Outro aspecto bastante importante estabelece que os locutores devem pronunciar as sentenças de forma fluente, mas não dramática, e mantendo um nível constante que seja considerado confortável para a pronúncia.



Cuidado especial deve ser dado ao nível da fala durante a gravação. Assim, deve-se monitorar o processo de gravação para que o nível ativo da fala não se aproxime do ponto de saturação do sistema de gravação de modo a distorcer o sinal. Além disso, deve-se evitar ao máximo a introdução de ruídos espúrios (e.g. picos de chaveamento, sussurros, etc.) danosos ao sinal e ao resultado dos testes.

B – Equalização do Nível da Fala

A medição dos níveis dos sinais na parte digital da rede internacional é normalmente expressa por engenheiros de telecomunicação como y dBm0, i.e. o nível relativo a 1 mW em 600 Ω . Sinais de voz gravados digitalmente por meio de computador são tipicamente armazenados no formato inteiro. É mais conveniente especificar os níveis de tais sinais relativos à máxima energia do sinal que pode ser armazenada pelo computador, e.g. z dBov (sendo que os caracteres 'ov' significam nível do sinal digital de sobrecarga). É necessário entender a relação desses termos de modo que os níveis em dBm0 possam ser expressos em dBov. Seja a relação abaixo:

$$z \text{ dBov} = y \text{ dBm0} + C$$

Para um sistema codificado em lei A, uma senóide que excita o máximo nível tem uma energia dada por $T_{\max} = 3.14$ dBm0 (Recomendação G.711). Similarmente, para um sistema codificado por μ -law, $T_{\max} = 3.17$ dBm0.

A energia de referência para dBov é o máximo valor inteiro para o sistema em uso. Uma senóide de amplitude máxima (e.g. 32 767 para um sistema de 16-bit) terá energia -3.01 dBov. Assim, $C = -6.15$ dB para sistemas A-law e $C = -6.18$ dB para sistemas μ -law.

Um valor típico para o nível médio de voz ativa (medida de acordo com a Recomendação P.56 “Medição Objetiva do Nível de Atividade da Fala”) é -20 dBm0, correspondendo aproximadamente a -26 dBov. Assim, o nível de voz ativa para cada sentença deve ser medido, e o nível gravado ajustado para ser -27 ± 1 dBov para o sistema no formato digital. Para idiomas ou vozes onde o nível do valor médio até o valor de pico excede 23 dB, o valor médio do sinal de voz deve ser reduzido apropriadamente.

Finalmente, grupos de sentenças são combinados (com um apropriado período de silêncio entre sentenças) para produzir o sinal fonte final para ser utilizado no teste subjetivo.

VIII.1.4 Seleção dos parâmetros do experimento

Para a condução do experimento deve-se determinar condições experimentais que são convenientemente divididas naquelas que envolvem o codec sob teste e naquelas que servem como referência. As condições do codec sob teste fornecem informações essenciais sobre o desempenho do codec, enquanto que as condições de referência permitem que sejam feitas comparações consistentes quando o mesmo codec é testado em diferentes laboratórios ou no mesmo laboratório em diferentes épocas.



A – Condições do Codec

As condições de teste descritas neste item servem para indicar como a performance do codec é afetada pelas mudanças nos níveis do sinal de entrada, de audição, por diferentes locutores, pela presença de ruído de fundo, etc.

Sabe-se que codecs digitais são sensíveis ao nível de entrada do sinal de voz e que a performance do codec pode depender deste nível. Quando da avaliação da performance do codec é usual estabelecer os níveis de entrada relativos ao ponto de sobrecarga (saturação) do codec. É recomendado que sejam utilizados níveis de entrada 14, 26 e 38 dB abaixo do ponto de sobrecarga do codec (aproximadamente equivalente a -8, -20, -32 dBm0) para avaliação do impacto nas variações no nível de entrada na sua performance. Além disso, recomenda-se que ao menos três níveis de audição sejam usados quando o efeito do nível de recepção é avaliado. Estes níveis devem variar no máximo 10 dB do nível escolhido. Deve-se notar que o nível de audição escolhido é geralmente abaixo do “Nível de Audição Ótimo”, i.e. o nível de audição correspondente ao máximo valor MOS (como estabelecido através da Recomendação P.10 “Vocabulário de Termos Relativo a Qualidade de Transmissão e de Aparelhos Telefônicos”).

Um outro importante ponto para a avaliação dos codecs refere-se aos locutores. Este tipo de análise observa a influência que a alteração no padrão do sinal de voz de entrada (de masculino para feminino, por exemplo) tem sobre o resultado do processo de codificação. Para isso é recomendado que um mínimo de dois locutores de cada sexo sejam utilizados. Entretanto, se o teste da dependência do locutor é extremamente importante, é recomendado que este número seja ampliado para 8 homens, 8 mulheres e 8 crianças.

A performance do codec também deve ser observada para se saber quão robusto ele é na presença de múltiplos sinais de voz de entrada e verificar se algum efeito adverso ocorre (e.g. interrupção da transmissão, sinais espúrios e etc.).

Quando se considera somente o uso do aparelho telefônico, a mistura de dois diferentes locutores, com níveis de entrada separados de 20 dB, é normalmente suficiente. Entretanto, se o codec é usado no modo ‘conferência’, i.e. ‘*hands-free*’, onde mais do que um locutor pode falar ao mesmo tempo, então se deve assegurar que a presença de múltiplos locutores, onde a diferença nos níveis de entrada dos sinais de voz pode ser zero, não afeta adversamente o desempenho do codec.

B – Erros

A seleção das condições de erro deve ser feita no contexto das aplicações desejadas para o codec. Se o codec será usado na PSTN (Rede de Telefonia Fixa), então os erros devem ser erros de bits aleatoriamente distribuídos. Se, por outro lado, a aplicação pretendida é um ambiente de rádio, tal como rádio-móvel, então os erros ocorrerão tanto isoladamente quanto de forma distribuída e o efeito desses erros poderá ser o de corromper trechos (ou quadros) completos do sinal de voz codificado. Assim, recomenda-se que para sistemas como a PSTN (com erros distribuídos aleatoriamente), seja utilizada uma Taxa de Erro de Bits (BERs) na faixa de 0 a



1:1000. Todavia, isto será dependente do número de sentenças usadas na amostra. Em algumas circunstâncias pode ser necessário testar a BERs de até 1:100.

C – Taxas de Bits

O codec deve ser testado para todas as taxas de bits para as quais ele é capaz de operar, e.g. o codec padronizado na Recomendação G.722 estas taxas são de 48 kbit/s, 56 kbit/s e 64 kbit/s e para o codec padronizado na Recomendação G.726 são de 16 kbit/s, 24 kbit/s, 32 kbit/s e 40 kbit/s.

D – Encaminhamento Assíncrono

Quando o codec sob teste será usado em um sistema misto (analógico e digital), podem ocorrer situações onde os sinais de voz são codificados e decodificados por um codec duas ou mais vezes, i.e. canais digitais de comunicação são intercalados com canais analógicos. Em tais situações, a distorção introduzida pelo codec é acumulada em cada conversão sucessiva entre os domínios analógico e digital. Além disso, quando o codec é empregado em uma rede de comunicação, cada processo de conversão inclui distorções adicionais intrínsecas da quantização. Um exemplo de encaminhamento assíncrono é apresentado na figura 22.

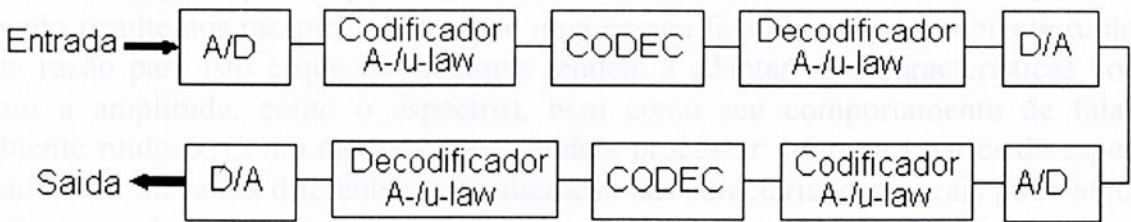


Figura 22. Encaminhamento Assíncrono

E – Encaminhamento Síncrono

Um outro tipo de situação que pode ocorrer é o da configuração do circuito de comunicação mostrado na figura 23. Neste caso, os sinais de voz são codificados e decodificados múltiplas vezes, mas sem conversões D/A e A/D ou Lei A/ μ ; ou seja, o sinal permanece no domínio digital ao longo de todo o canal de comunicação. Embora esta configuração seja menos suscetível a erros de quantização, a distorção acumulada pelos codecs continua a ocorrer e deve ser avaliada para se determinar sua influência.

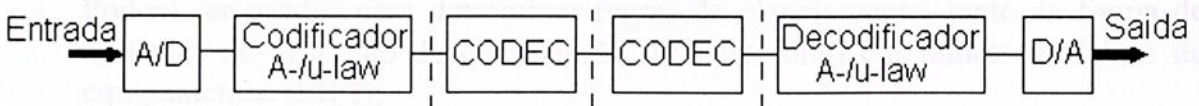


Figura 23. Encaminhamento Síncrono

F – Ruído Ambiente (emissão)

O ruído ambiente e sua interação com a voz do locutor representam importantes fatores na avaliação do codec pelas seguintes razões:

a) codecs com baixas taxas de transmissão podem usar técnicas de codificação que não são de forma de onda;

b) a aplicação do codec pode em ambiente ruidoso, por exemplo, um veículo em movimento ou um escritório ruidoso.

Devem ser executados testes suficientes com o ruído apropriado para examinar este efeito, sendo que as seguintes relações Sinal/Ruído são recomendadas:

- 30 dB para uma sala ruidosa;
- 10 dB e 20 dB ruído veicular.

Modernas técnicas de codificação de voz podem tentar suprimir ou eliminar ruído de fundo. Assim, é possível que sinais de voz que foram degradados por ruído de fundo, e subseqüentemente processados por um codec, sejam de melhor qualidade do que o sinal original de entrada.

Uma possibilidade para obtenção de sinais ruidosos de teste é que o ruído seja combinado eletricamente com as gravações originais de modo que os níveis do ruído e do sinal de voz sejam conhecidos exatamente. Todavia, não é recomendado que as gravações originais sejam feitas em um ambiente ruidoso.

Outra importante consideração é que não deve ser automaticamente suposto que a mistura de gravações separadas do sinal de voz e de ruído ambiente com nível elevado resulte nos mesmos efeitos que uma pessoa falando em um ambiente ruidoso. Uma razão para isto é que os locutores tendem a adaptar suas características vocais (tanto a amplitude, como o espectro), bem como seu comportamento de fala no ambiente ruidoso. Como alguns codecs podem processar diferentes partes do espectro de áudio de maneiras diferentes, esta alteração nas características vocais pode afetar a performance do sistema.

VIII.1.5 Condições de Referência

As condições de referência representam meios convenientes para possam ser feitas comparações significativas de resultados de testes subjetivos em diferentes laboratórios ou do mesmo laboratório em diferentes oportunidades. Estas condições incluem a condição 'melhor possível', além de condições onde degradações conhecidas e controladas tenham sido adicionadas aos sinais.

Os codecs de referência servem a dois propósitos:

- Podem ser usados para determinar regras de planejamento, tanto na forma de unidades de distorção de quantização (QDUs) como em fatores de falhas de equipamentos (EIFs);
- Podem ser usados como o padrão para julgar a performance total em termos dos parâmetros sob avaliação.

VIII.1.6 Escalas de Qualidade de Audição

Durante o processo de avaliação subjetiva da qualidade de um dado processo de codificação/decodificação deve-se escolher uma escala de avaliação adequada.



Existem muitas dessas Escalas de Qualidade ou Escalas de Opiniões, algumas são apresentadas a seguir:

A - Escala de Qualidade de Audição

Esta escala representa a qualidade da fala e é tabelada da seguinte forma:

Tabela 3. Escala de Qualidade de Audição

Pontuação	Descrição
5	Excelente
4	Boa
3	Média
2	Pobre
1	Ruim

B - Escala de Esforço de Audição

A Escala de Esforço de Audição representa o esforço necessário para entender o significado das sentenças:

Tabela 4. Escala de Esforço de Audição

Pontuação	Descrição
5	Relaxamento completo; nenhum esforço requerido.
4	Necessidade de atenção; nenhum esforço apreciável.
3	Esforço moderado requerido.
2	Considerável esforço requerido.
1	Nenhuma compreensão com qualquer esforço realizável.

C - Escala de Categoria de Degradação

Outra forma de avaliar a qualidade do sinal é determinando o Nível de Degradação introduzido:

Tabela 5. Escala de Categoria de Degradação

Pontuação	Degradação
5	Inaudível.
4	Audível, mas não é perturbadora.
3	Ligeiramente perturbadora.
2	Perturbadora.
1	Muito perturbadora.



D - Escala de Detectabilidade

Além da qualidade, pode-se medir a presença de alguma distorção ou ruído específico no sinal (eco por exemplo). Para tanto, pode-se usar a seguinte escala:

Tabela 6. Escala de Detectabilidade

Pontuação	Descrição
A	Questionável
B	Detectável
C	Não Detectável

E - Escala de Comparação

Considerando-se a apresentação de dois sinais semelhantes para os ouvintes pode-se efetuar a comparação entre os mesmos. Uma maneira de exprimir a comparação é a apresentada na tabela a seguir:

Tabela 7. Escala de Comparação

Pontuação	Descrição
3	Muito melhor
2	Melhor
1	Ligeiramente melhor
0	Quase iguais
-1	Ligeiramente pior
-2	Pior
-3	Muito pior

Figura 24. Sinal de Referência

Para permitir a avaliação dos algoritmos de cancelamento de eco, foram gerados alguns sinais de teste. Esses sinais foram gerados a partir do sinal de referência com a adição de diferentes níveis de ruído (branco e ruído de automóvel). Com isso, pode-se observar os resultados em função tanto do tipo de medida utilizada, como também do nível de distorção.

Os sinais de ruído utilizados foram um ruído branco dentro de um automóvel em movimento e um ruído branco gerado matematicamente (seqüência de números aleatórios).

Os sinais distorcidos (ou referenciados) foram criados adicionando-se ao sinal puro (referência) um sinal de ruído com três níveis diferentes. Nas figuras 25 e 26 são apresentadas as formas de onda desses sinais, que apresentam, respectivamente, boa relação sinal/ruído (SNR) e baixa SNR.



Capítulo IX**SIMULAÇÕES E RESULTADOS**

Passamos agora para a apresentação dos resultados obtidos nas simulações tanto para as medidas objetivas quanto para as subjetivas.

IX.1 MEDIDAS OBJETIVAS

Os métodos de Avaliação Objetiva apresentados nos itens anteriores foram implementados e avaliados usando o software de simulação MATLAB. O objetivo foi verificar as curvas de distorção em função do tipo e nível do ruído. Na figura 24 é apresentado um sinal de voz típico que foi tomado como sinal de referência.

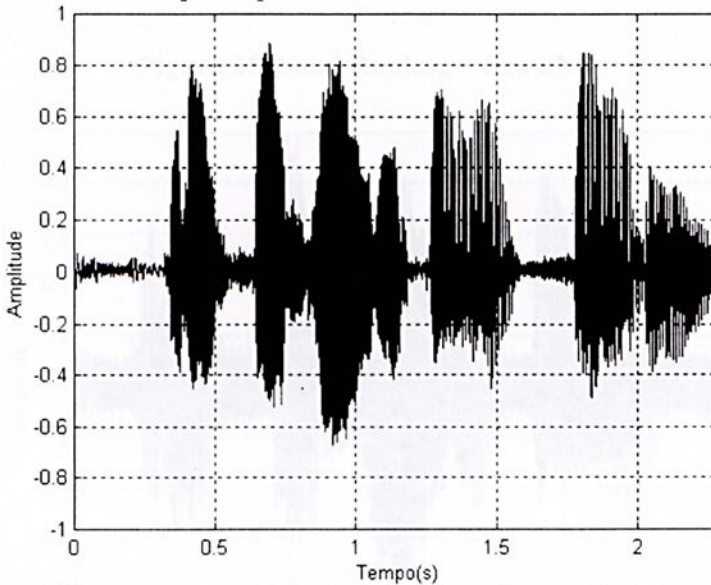


Figura 24. Sinal de Referência

Para permitir a avaliação dos algoritmos desenvolvidos foram gerados alguns sinais de teste. Esses sinais foram gerados a partir do sinal de referência com a adição de diferentes níveis de ruído (branco e ruído de automóvel). Com isso, pode-se observar os resultados em função, tanto do tipo de medida utilizada, como também do nível de distorção.

Os sinais de ruído utilizados foram um ruído gerado dentro de um automóvel em movimento e um ruído branco gerado matematicamente (seqüência de números aleatórios).

Os sinais distorcidos (ou ruidosos) foram obtidos adicionando-se ao sinal puro (referência) um sinal de ruído com três níveis diferentes. Nas figuras 25 e 26 são apresentadas as formas de onda desses sinais, que apresentam, respectivamente, boa relação sinal/ruído (SNR) e baixa SNR:



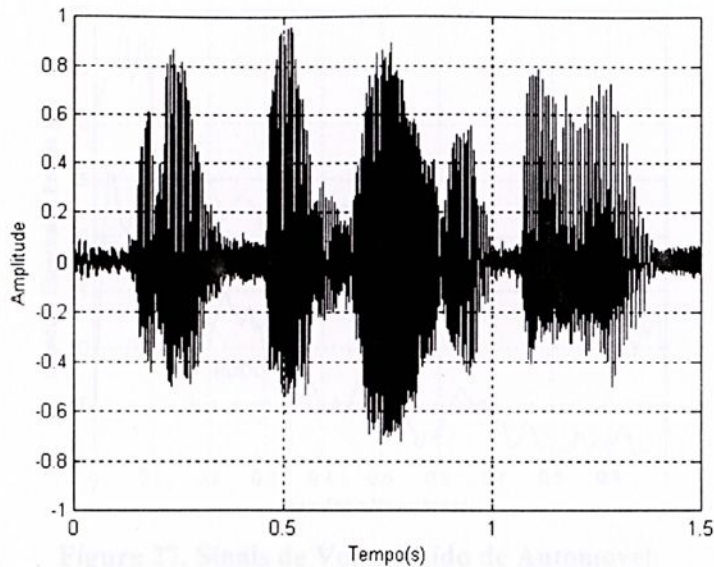


Figura 25. Sinal Ruidoso – Boa SNR

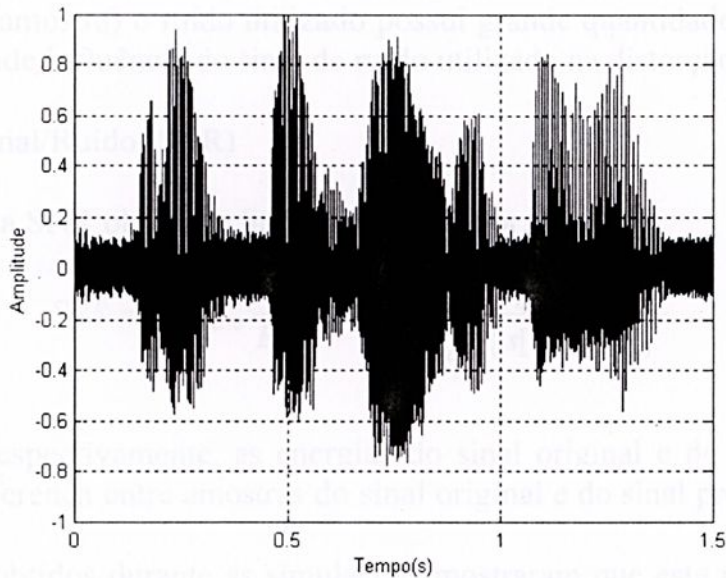


Figura 26. Sinal Ruidoso – Baixa SNR

Foram obtidos sinais com diferentes SNR (relação sinal ruído) através da adição dos sinais de ruído de acordo com a seguinte equação:

$$\text{Sinal Ruidoso} = \text{Sinal de Referência} + \alpha * \text{Ruído}$$

com ($\alpha = 0,1$, $\alpha = 0,3$ e $\alpha = 0,6$). É importante salientar que, tanto o sinal de voz como o sinal de ruído, possuíam a mesma amplitude máxima antes da aplicação da equação.

Para se ter uma idéia do nível de ruído, são apresentadas curvas das Magnitudes Espectrais de Energia dos sinais de voz original e do ruído (de automóvel) na figura 27.

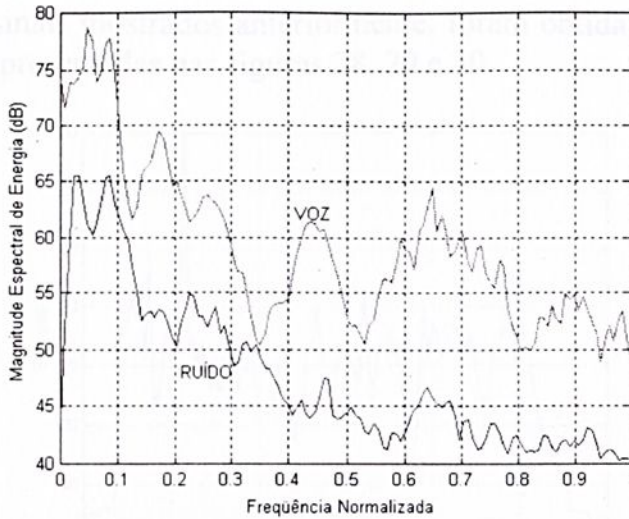


Figura 27. Sinais de Voz e Ruído de Automóvel

Na figura 27 pode-se notar que na faixa de interesse do sinal de voz ($f < 4$ kHz ou $\omega < 0.5 \pi$ rad/amostra) o ruído utilizado possui grande quantidade de energia. Este fato mostra a grande influência do sinal de ruído utilizado na distorção do sinal de voz.

IX.1.1 Relação Sinal/Ruído (SNR)

A equação da SNR obtida na literatura é dada por:

$$SNR = 10 \cdot \log_{10} \frac{E_s}{E_e} = 10 \cdot \log_{10} \frac{\sum_n s^2[n]}{\sum_n \{s[n] - \hat{s}[n]\}^2}$$

sendo E_s e E_e , respectivamente, as energias do sinal original e do erro. Este erro é calculado pela diferença entre amostras do sinal original e do sinal processado teste (ou ruidoso).

Resultados obtidos durante as simulações mostraram que esta relação, da forma como é equacionada, não representa de forma eficiente as distorções introduzidas. Um bom exemplo da deficiência desta representação seria um caso onde o sistema introduza uma distorção de modo a inverter totalmente a fase do sinal original. Esta inversão não determina diferença na percepção do conteúdo da mensagem, mas acarreta na obtenção de um valor de SNR que leva a interpretações incorretas. Para resolver esse problema, propõe-se a seguinte alteração:

$$SNR = 10 \cdot \log_{10} \frac{\sum_n s^2[n]}{\left\{ \sum_n |s[n]| - \sum_n |\hat{s}[n]| \right\}^2}$$

Nesta nova equação são tomados somente os valores absolutos das amostras do sinal original e do sinal processado, eliminando o problema de fase anteriormente relatado.

Utilizando os sinais mostrados anteriormente, foram obtidas curvas de SNR para cada sinal ruidoso, apresentadas nas figuras 28, 29 e 30.

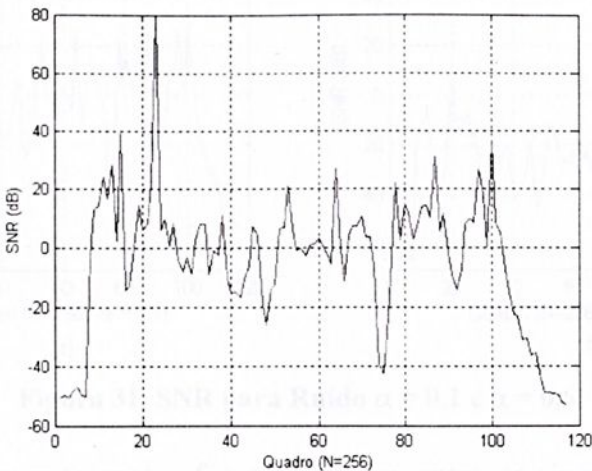


Figura 28. SNR para Ruído $\alpha = 0,1$

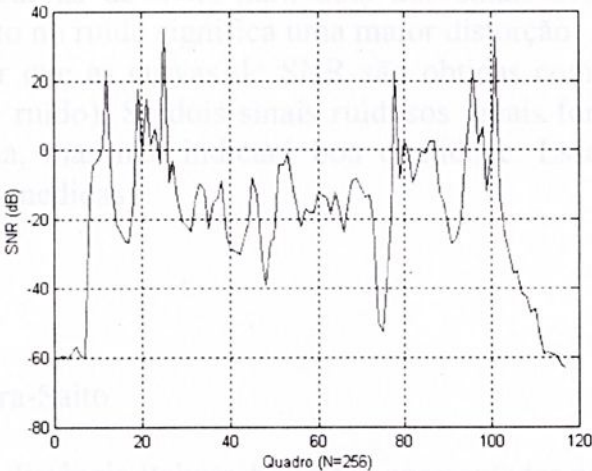


Figura 29. SNR para Ruído $\alpha = 0,3$

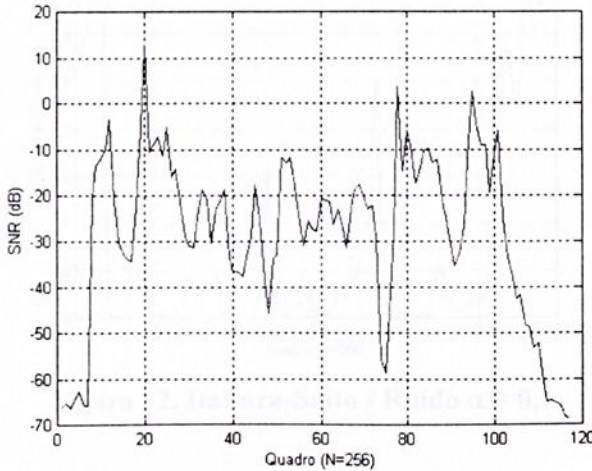


Figura 30. SNR para Ruído $\alpha = 0,6$



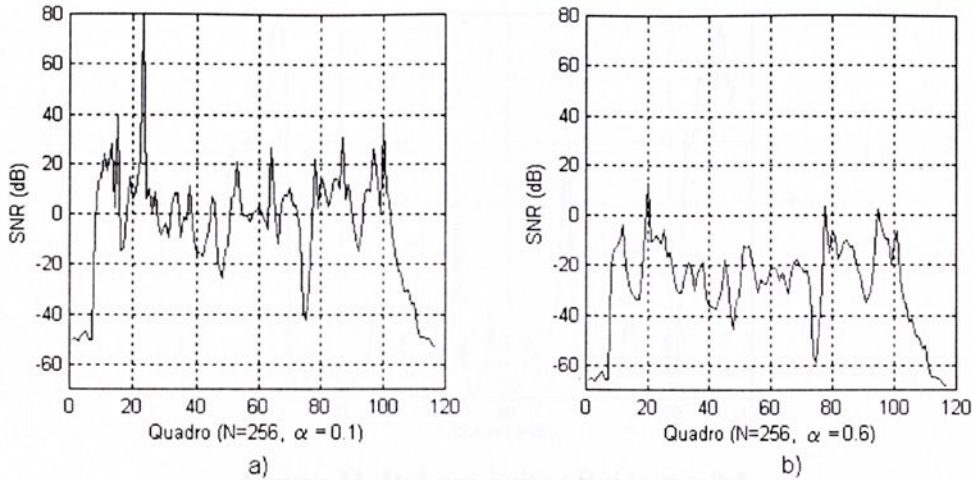


Figura 31. SNR para Ruído $\alpha = 0,1$ e $\alpha = 0,6$

Observa-se que quanto maior for a diferença entre os sinais, menor será a SNR. Então, quanto menor a SNR, maior a distorção presente no sinal ruidoso. Na figura 31 são apresentadas as curvas de SNR para dois dos sinais avaliados, onde se pode observar que o aumento no ruído significa uma maior distorção.

Deve-se ressaltar que as curvas de SNR são obtidas com base em um sinal de referência limpo (sem ruído). Se dois sinais ruidosos iguais forem utilizados, a saída da SNR será altíssima, mas não indicará boa qualidade. Essa observação é válida também para as outras medidas.

IX.1.2 Distância Itakura-Saito

As curvas para a distância Itakura-Saito são apresentadas nas figuras 32, 33 e 34.

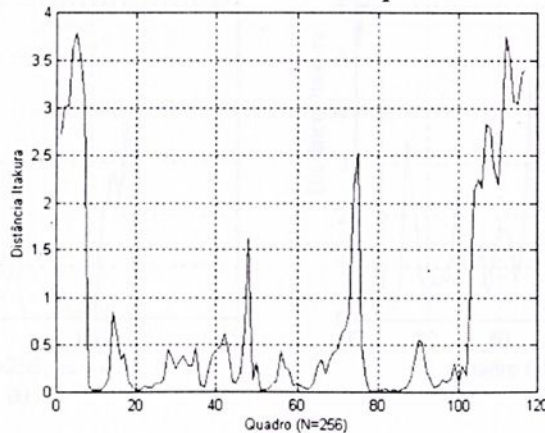


Figura 32. Itakura-Saito / Ruído $\alpha = 0,1$

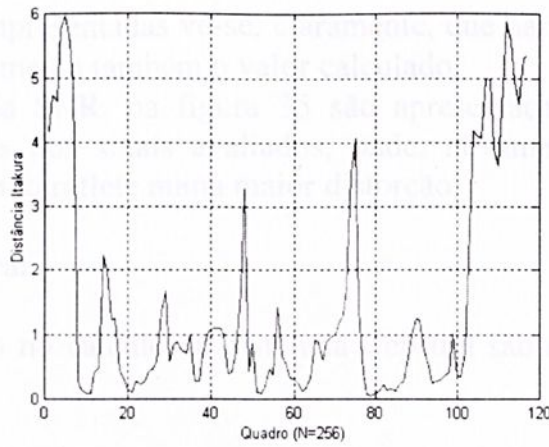


Figura 33. Itakura-Saito / Ruído $\alpha = 0,3$

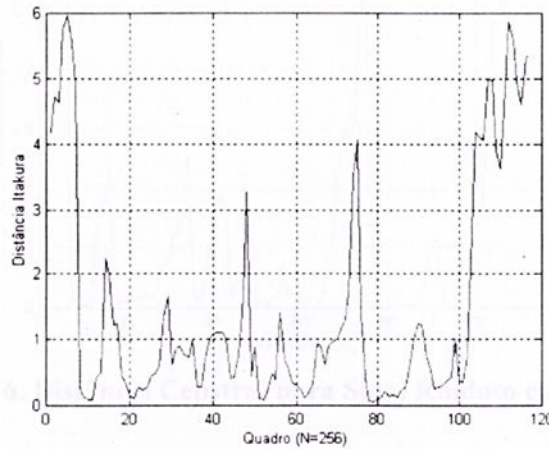


Figura 34. Itakura-Saito / Ruído $\alpha = 0,6$

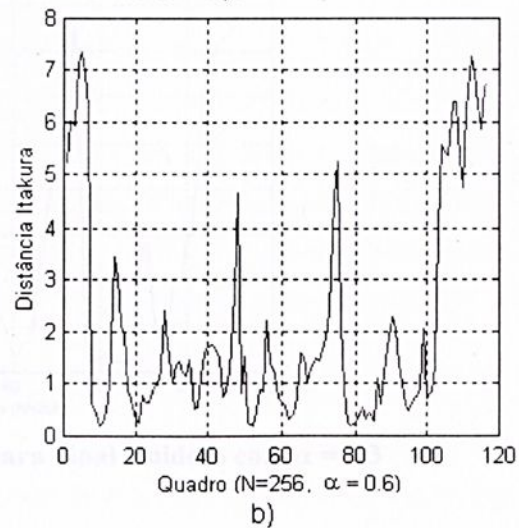
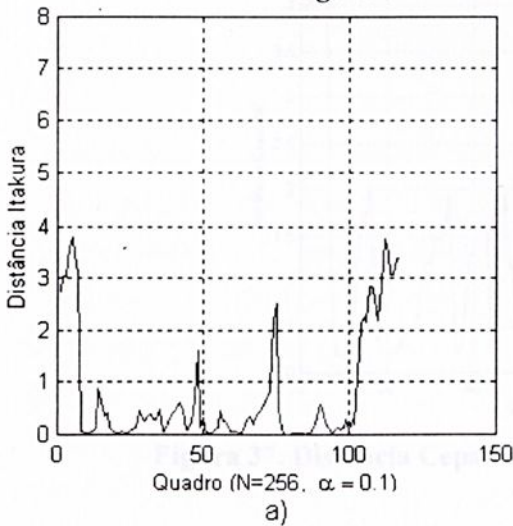


Figura 35. Itakura-Saito Ruído $\alpha = 0,1$ e $\alpha = 0,6$

O comportamento da curva para a medida Itakura-Saito é bastante diferente da SNR. Agora, quanto maior o valor apresentado, maior a medida de distorção

calculada. Nas curvas apresentadas vê-se, claramente, que na medida em que aumenta o nível de distorção, aumenta também o valor calculado.

Como no caso da SNR, na figura 35 são apresentadas as curvas da medida Itakura-Saito para dois dos sinais avaliados, onde, novamente observa-se que um aumento no nível do ruído reflete numa maior distorção.

IX.1.3 Distância Cepstral

As curvas obtidas no cálculo da distância Cepstral são apresentadas nas figuras 36, 37 e 38.

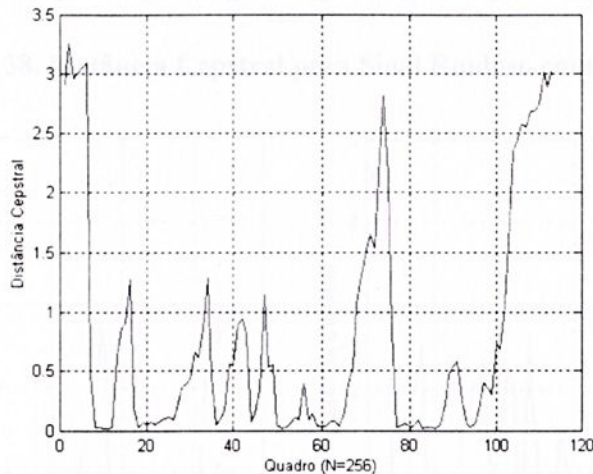


Figura 36. Distância Cepstral para Sinal Ruidoso com $\alpha = 0,1$

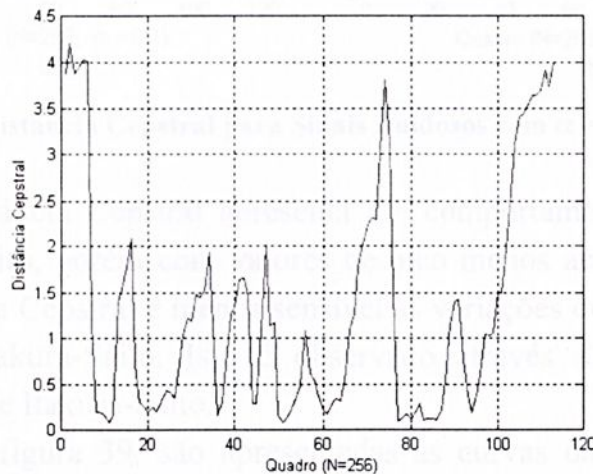


Figura 37. Distância Cepstral para Sinal Ruidoso com $\alpha = 0,3$

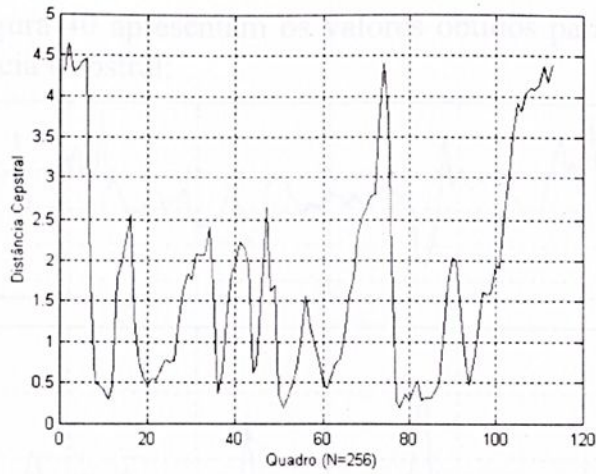


Figura 38. Distância Cepstral para Sinal Ruidoso com $\alpha = 0,6$

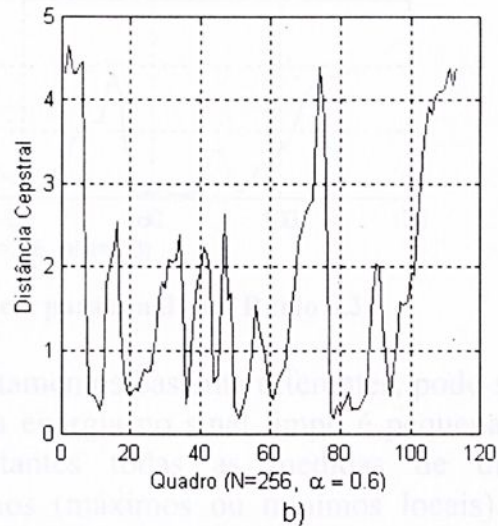
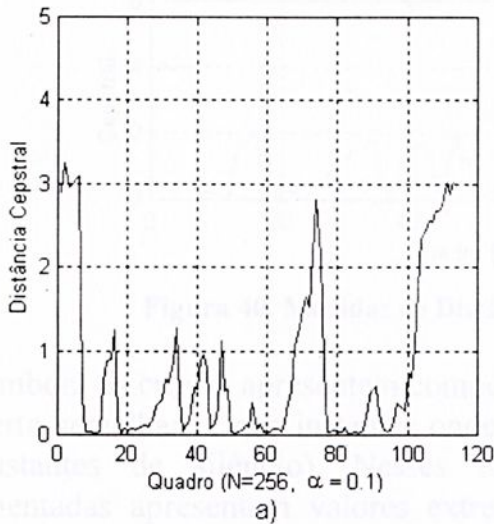


Figura 39. Distância Cepstral para Sinais Ruidosos com $\alpha = 0,1$ e $\alpha = 0,6$

A curva da distância Cepstral apresenta um comportamento semelhante ao da medida de Itakura-Saito, porém com valores de pico menos acentuados. Além disso, nota-se que a distância Cepstral é menos sensível às variações do nível de distorção do que a medida de Itakura-Saito. Isto é observado através das oscilações bruscas presentes na medida de Itakura-Saito.

Novamente, na figura 39, são apresentadas as curvas da medida de Distância Cepstral para dois dos sinais avaliados ($\alpha = 0,1$ e $\alpha = 0,6$). Observa-se que, como nos casos anteriores, um aumento no nível do ruído corresponde a uma maior distorção.

IX.1.4 Comparação entre as Medidas Objetivas

Uma análise de interesse é a comparação entre os resultados obtidos para cada uma das medidas de distorção implementadas. Para tanto, foram utilizados o sinal original (limpo) e o sinal ruidoso com $\alpha = 0,3$.

As curvas da figura 40 apresentam os valores obtidos para as medidas de SNR, Itakura-Saito e Distância Cepstral:

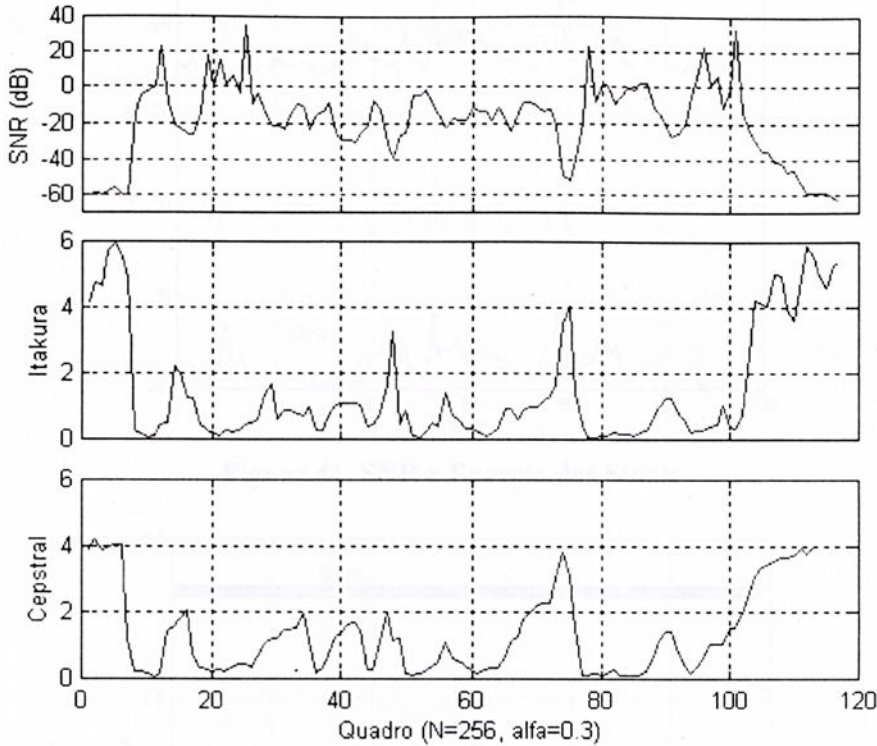


Figura 40. Medidas de Distância para Sinal com Ruído 0,3

Embora as curvas apresentem comportamentos bastante diferentes, pode-se notar uma certa semelhança nos instantes onde a energia no sinal limpo é pequena (como nos instantes de silêncio). Nesses instantes todas as medidas de distorção implementadas apresentam valores extremos (máximos ou mínimos locais). Dessa forma, pode-se concluir que no processo de cálculo de distorção os intervalos de silêncio devem ser ignorados, uma vez que podem apresentar valores muito grandes (ou pequenos) para a medida implementada.

IX.1.5 Diferença de Energia

Para verificar a influência do nível de energia nas medidas de distorção, calcularam-se as medidas de distorção entre o sinal original e o sinal original com ganhos da ordem de 200, 400 e 800% de seu máximo valor de amplitude. Dessa forma, foram obtidas as curvas apresentadas nas figuras 41 e 42.

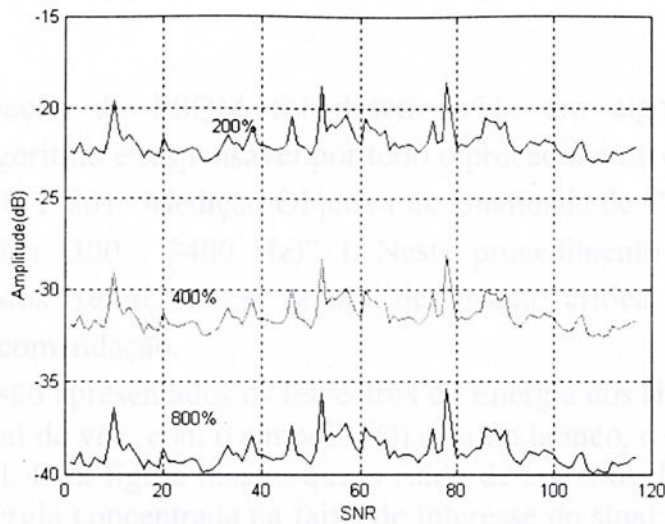


Figura 41. SNR e Energia dos Sinais

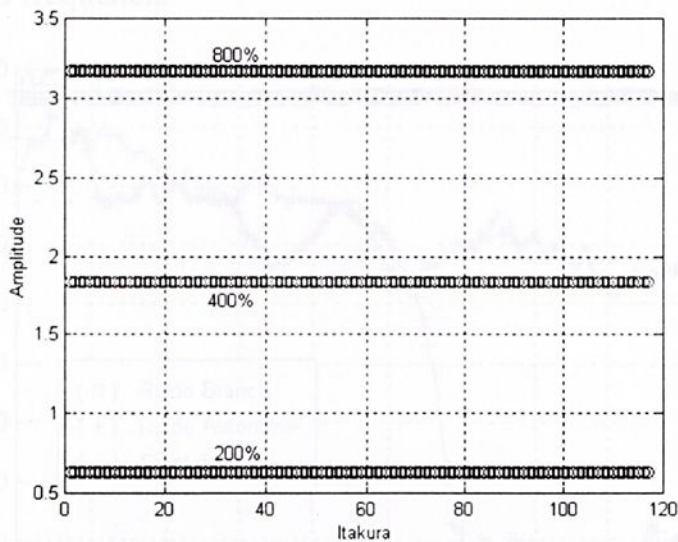


Figura 42. Medida Itakura-Saito e Energia dos Sinais

Como se pode notar da figura 42, observando as amplitudes apresentadas, as medidas de SNR e Itakura-Saito são bastante influenciadas pela diferença do nível de energia dos sinais envolvidos no cálculo. A curva para a distância Cepstral foi omitida pois os resultados obtidos não apresentaram diferença, independentemente do nível de distorção introduzido (200, 400 ou 800%).

A influência do nível de energia dos sinais envolvidos no processo de cálculo sobre a medida de distância (distorção) é uma característica indesejável, pois esta diferença de energia pode ser, por exemplo, apenas um fator de ganho que não pode ser interpretado como distorção.

IX.1.6 PSQM

Para a avaliação do PSQM foi desenvolvido um algoritmo utilizando o MATLAB. Este algoritmo é responsável por todo o procedimento de cálculo da PSQM (Recomendação ITU P.861 “Medição Objetiva de Qualidade de Codificadores de Voz de Banda Telefônica (300 - 3400 Hz)”). Neste procedimento o algoritmo utiliza parâmetros tabelados referentes a faixas de banda crítica, ruído Hoth, etc., estabelecidos na recomendação.

Na figura 43 são apresentados os Espectros de Energia dos sinais utilizados. Com o símbolo (*) o sinal de voz, com o símbolo ($\square$) o ruído branco, e com o símbolo (+) o ruído de automóvel. Esta figura mostra que o ruído de automóvel possui uma parcela significativa de energia concentrada na faixa de interesse do sinal de voz ($f < 4\text{kHz}$ ou $f_{\text{normalizada}} < 0.5\pi \text{ rad/amostra}$). Já o ruído branco, possui distribuição uniforme de energia ao longo da frequência.

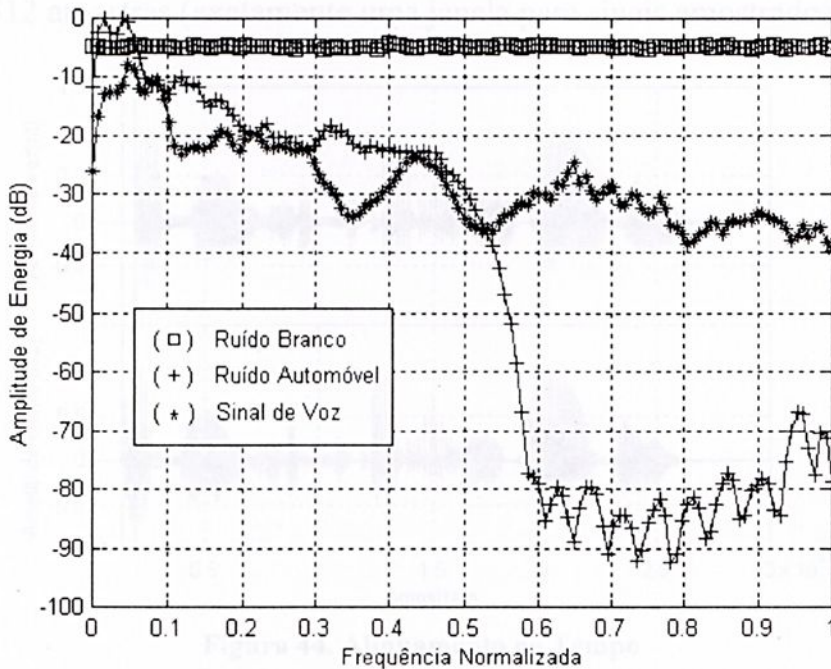


Figura 43. Espectros de Energia

Para avaliar o comportamento do algoritmo de cálculo do PSQM em função dos sinais de entrada foram utilizadas diferentes combinações de sinais.

Inicialmente, observou-se o comportamento do algoritmo quando um dado sinal de voz era comparado com ele mesmo. Ainda, utilizou-se o sinal de voz e o seu oposto para avaliação. Finalmente, comparou-se o sinal de voz com um outro com amplitude duas vezes maior que o sinal de referência.

Tabela 8. Resultados para PSQM igual a zero

Sinal	PSQM
Voz x Voz	0.0
Voz x (-Voz)	0.0
Voz x (2.Voz)	0.0

Como pode-se observar nos resultados da tabela 8 acima, a saída do algoritmo PSQM foi igual a zero para as três avaliações realizadas, como era esperado. Dessa forma, vemos que o algoritmo efetua a análise da qualidade comparando os sinais de entrada, e se estes são sinais iguais temos um processo de codificação/decodificação perfeito. Ainda, vemos que o algoritmo não é dependente da fase nem da energia dos sinais de entrada. Esses últimos resultados representam características interessantes uma vez que o ouvido humano é bastante insensível a distorções de fase, e que sinais com diferentes energias não representam, necessariamente, sinais distorcidos.

Uma outra simulação procurou demonstrar a importância do alinhamento no tempo dos sinais de entrada. O gráfico (a) da figura 44 mostra um mesmo sinal defasado de 512 amostras (exatamente uma janela para sinais amostrados a 16kHz).

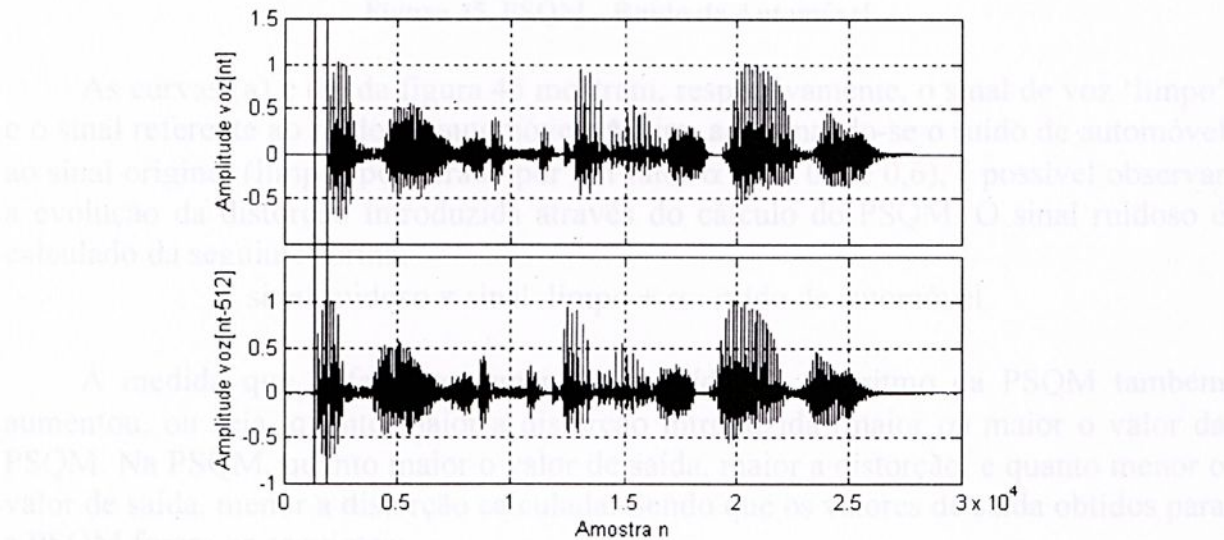


Figura 44. Alinhamento no Tempo

Submetendo esses sinais ao algoritmo da PSQM obtém-se um valor de saída igual a 4.5. Como o valor máximo de saída do algoritmo da PSQM é igual a 6.5, temos que o resultado obtido para os sinais desalinhados denota uma distorção bastante acentuada entre os sinais, ou seja, baixa qualidade do sinal sob teste. Todavia, sabendo-se que os sinais são idênticos pode-se concluir que um alinhamento mal executado/calculado pode levar a avaliações erradas sobre a qualidade do sinal.

Além disso, tomou-se como fonte de ruído um sinal gravado dentro de um automóvel em movimento de modo a verificar o comportamento do algoritmo PSQM.

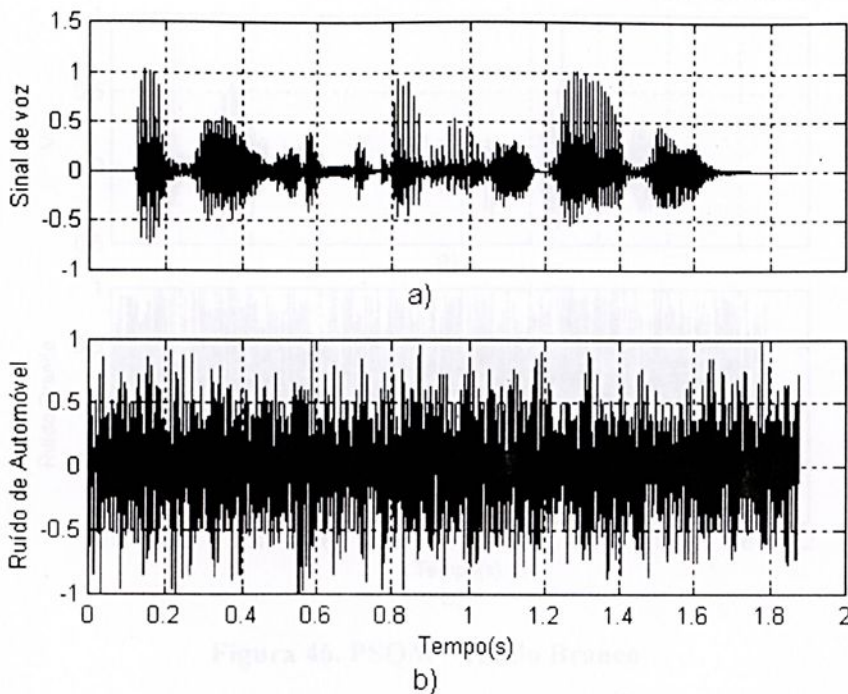


Figura 45. PSQM - Ruído de Automóvel

As curvas (a) e (b) da figura 45 mostram, respectivamente, o sinal de voz 'limpo' e o sinal referente ao ruído de automóvel. Assim, adicionando-se o ruído de automóvel ao sinal original (limpo) ponderado por um fator α (0,1, 0,3 e 0,6), é possível observar a evolução da distorção introduzida através do cálculo do PSQM. O sinal ruidoso é calculado da seguinte forma:

$$\text{sinal ruidoso} = \text{sinal limpo} + \alpha \cdot \text{ruído de automóvel}$$

À medida que o fator α aumentou, a saída do algoritmo da PSQM também aumentou, ou seja, quanto maior a distorção introduzida (maior α) maior o valor da PSQM. Na PSQM, quanto maior o valor de saída, maior a distorção; e quanto menor o valor de saída, menor a distorção calculada. Sendo que os valores de saída obtidos para a PSQM foram os seguintes:

$$\alpha = 0,1 \rightarrow \text{PSQM}_{\text{automóvel}} = 0,435$$

$$\alpha = 0,3 \rightarrow \text{PSQM}_{\text{automóvel}} = 1,108$$

$$\alpha = 0,6 \rightarrow \text{PSQM}_{\text{automóvel}} = 2,693$$

Se considerarmos que o valor máximo da PSQM é 6,5, conclui-se que os valores obtidos para $\alpha = 0,1$ e $0,3$ representam baixa distorção do sinal de referência. Já para $\alpha = 0,6$, a distorção calculada pode ser considerada média. Todavia, observando-se a evolução dos valores do algoritmo do PSQM, nota-se que para $\alpha = 0,6$ a distorção é bastante acentuada.

Para proporcionar uma avaliação mais abrangente utilizou-se também um ruído do tipo Branco. Para este sinal, foi efetuado o mesmo procedimento apresentado para o ruído de automóvel.

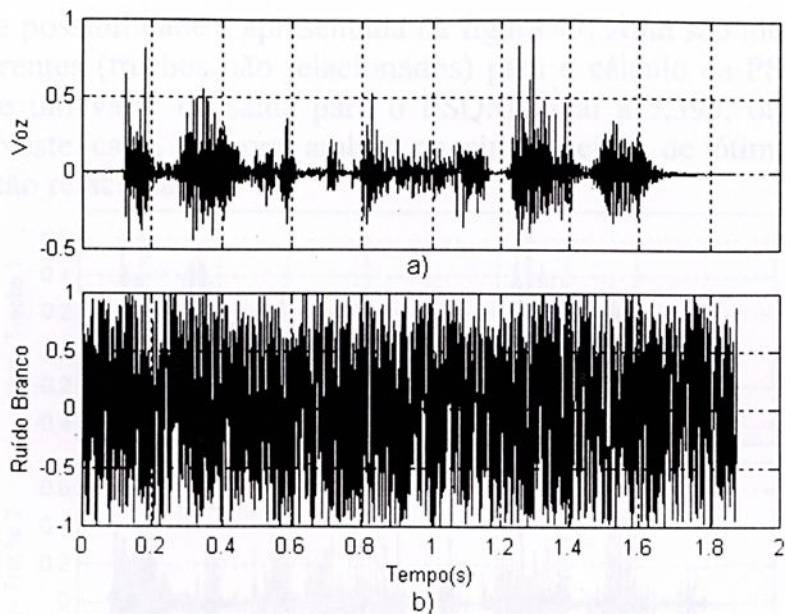


Figura 46. PSQM – Ruído Branco

Na figura 46 temos apresentados o sinal de referência (a) e o ruído branco (b). Novamente, utilizando-se a seguinte equação para a geração do sinal de teste (ruidoso):

$$\text{sinal ruidoso} = \text{sinal de referência} + \alpha \cdot \text{ruído branco}$$

com $\alpha = 0,1, 0,3 \text{ e } 0,6$.

À medida que o fator α aumentou, a saída do algoritmo da PSQM também aumentou, sendo os resultados apresentados a seguir:

- $\alpha = 0,1 \rightarrow \text{PSQM}_{\text{Branco}} = 0,127$
- $\alpha = 0,3 \rightarrow \text{PSQM}_{\text{Branco}} = 1,283$
- $\alpha = 0,6 \rightarrow \text{PSQM}_{\text{Branco}} = 4,167$

Observando-se os resultados para o ruído branco, nota-se distorção baixa para $\alpha = 0.1$ e 0.3 , e distorção alta para $\alpha = 0.6$.

Na tabela 9 estão apresentados os resultados para o ruído branco e ruído de automóvel. Como pode-se observar para $\alpha = 0.1$ o ruído de automóvel teve maior influência na distorção do sinal do que o ruído branco. Porém, para $\alpha = 0.3$ e 0.6 , o ruído branco passou a introduzir uma distorção mais acentuada.

Tabela 9. Resultados PSQM para diferentes ruídos

α	Ruído Automóvel	Ruído Branco
0.1	0.435	0.127
0.3	1.108	1.283
0.6	2.693	4.167

Uma outra possibilidade é apresentada na figura 47. Aqui são introduzidos sinais totalmente diferentes (trechos não relacionados) para o cálculo da PSQM. Para estes sinais obteve-se um valor de saída para o PSQM igual a 5,395, ou seja, distorção elevadíssima. Neste caso, embora ambos os sinais sejam de ótima qualidade, os mesmos não estão relacionados.

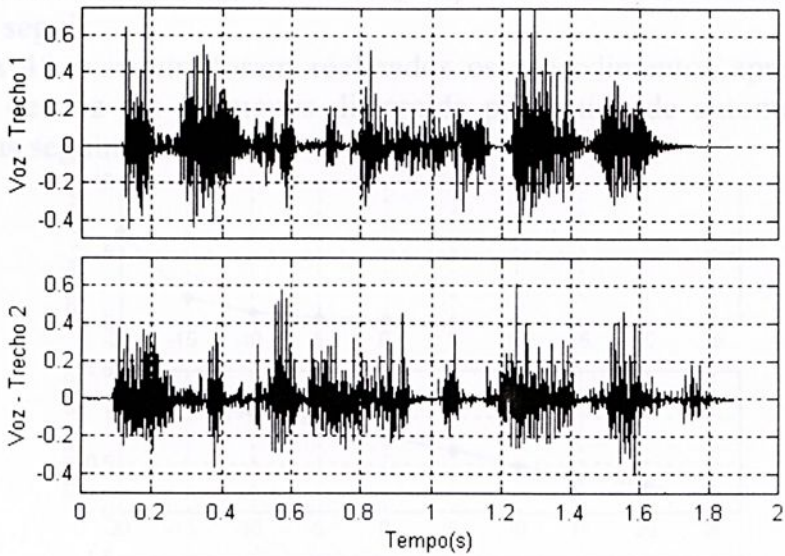


Figura 47. PSQM – Sinais Distintos

IX.2 Comparação entre PSQM e as demais Medidas Objetivas

Na tabela 10 abaixo, são agrupados todos os resultados obtidos para os diferentes tipos de medidas (SNR, Distância Cepstral, Distância Itakura-Saito e PSQM) e de sinais ruídos utilizados (Ruído Branco e de Automóvel), além da avaliação referente a defasagem dos sinais e da utilização de sinais não relacionados.

Tabela 10. Resultados para diferentes tipos de medidas ($\alpha = 0.3$)

	SNR	Cepstral	Itakura	PSQM
Ruído Automóvel	-22,022 dB	1,314	1,294	1,108
Ruído Branco	-34,002 dB	2,192	4,856	1,283
Sinais não relacionados	-21,278 dB	2,003	232,958	5,395
Sinais defasados	-16,225 dB	1,155	14,149	4,500

A distância Itakura-Saito e o PSQM são as medidas que apresentam maior variação quando os sinais não estão relacionados, ou seja, trechos de voz distintos. Já o resultado da SNR para sinais não relacionados foi praticamente equivalente ao obtido para o Ruído de Automóvel. Em todos os casos avaliados, nota-se que o ruído branco degrada o sinal de forma mais acentuada do que o ruído de automóvel. Isto deve-se pelo fato de o sinal de ruído branco possuir parcelas significativas de energia ao longo de toda a faixa de frequência. Finalmente, observando a tabela 10 acima observa-se que cada tipo de medida é mais sensível a um tipo diferente de distorção.

Uma outra análise que se mostrou bastante interessante foi a evolução do valor das medidas implementadas em função da distorção entre os sinais; para isto foram gerados vários sinais distorcidos (variando-se o parâmetro α apresentado anteriormente) que proporcionassem relações Sinal/Ruído conhecidas. Assim, estes sinais foram submetidos ao algoritmo PSQM permitindo obter os resultados que são apresentados a seguir.

Na figura 48 a seguir, foram realizados os procedimentos apresentados acima para um sinal de voz em português distorcido pelo ruído de automóvel. Com isso, foram obtidas as seguintes curvas:

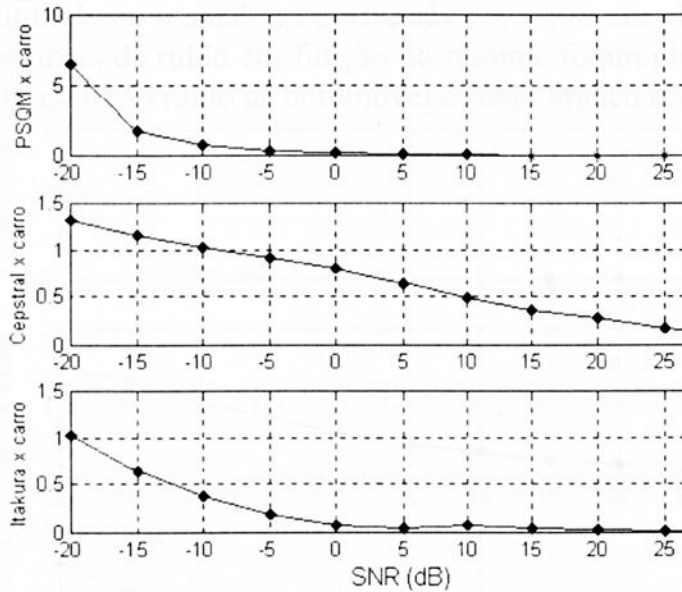


Figura 48. Sinal em Português x Ruído de Automóvel

Na figura 49, substituiu-se o ruído de automóvel por um ruído branco e se obtiveram as seguintes curvas:

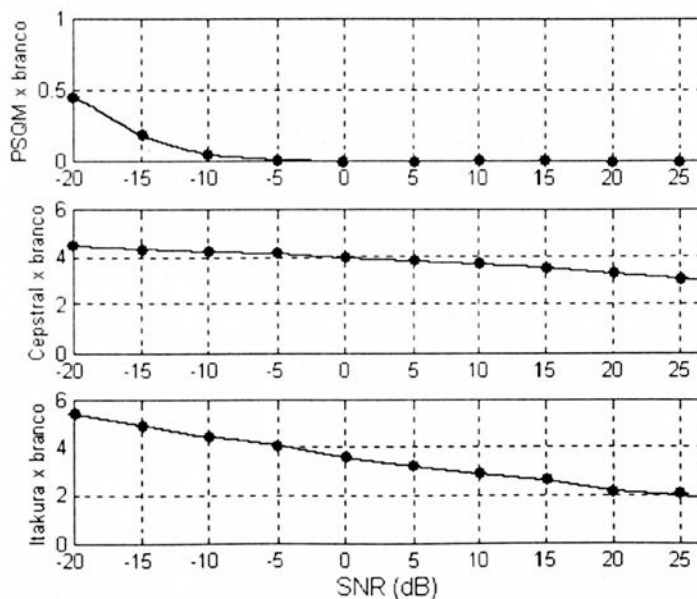


Figura 49. Sinal em Português x Ruído Branco

Observando-se as figuras 48 e 49 nota-se que somente a curva do PSQM apresentou a mesma característica de decrescimento exponencial tanto para o ruído branco como para o de automóvel. A distância Itakura-Saito foi a que apresentou maior modificação nas curvas apresentadas; passando de um decaimento aproximadamente exponencial para o ruído de automóvel para um decaimento próximo do linear para o ruído branco. Já para a distância Cepstral a característica de decaimento próxima do linear manteve-se.

Agora, substituindo-se o sinal em português por outro em alemão para observar se a influência dos sinais de ruído em função do idioma, foram plotadas as curvas das figuras 50 e 51, referentes ao ruído de automóvel e ruído branco respectivamente.

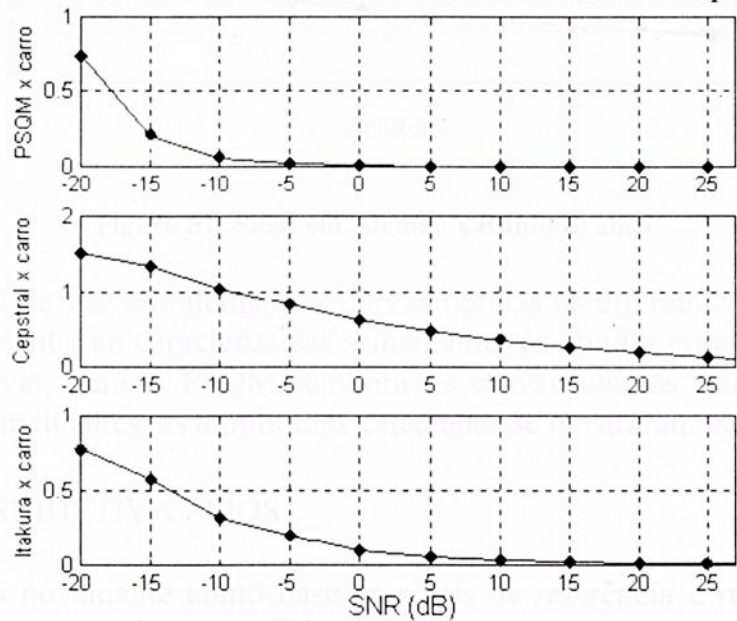


Figura 50. Sinal em Alemão x Ruído de Automóvel

	Ouvinte N°1	Ouvinte N°2	Ouvinte N°3	Ouvinte N°4	Valor Médio
Limpas x Ruído com $\alpha = 0,1$	3	2	4	3	2,25
Limpas x Ruído com $\alpha = 0,2$	2	2	3	1,5	2,1
Limpas x Ruído com $\alpha = 0,3$	1,5	1,5	2,5	1	1,6

Os resultados (pontuações) obtidos são coerentes com o resultado esperado pois, quanto maior o nível de ruído introduzido, menor a qualidade do sinal. Esse tipo de avaliação, embora eficiente, é de difícil realização e requer a participação de um grupo de ouvintes.

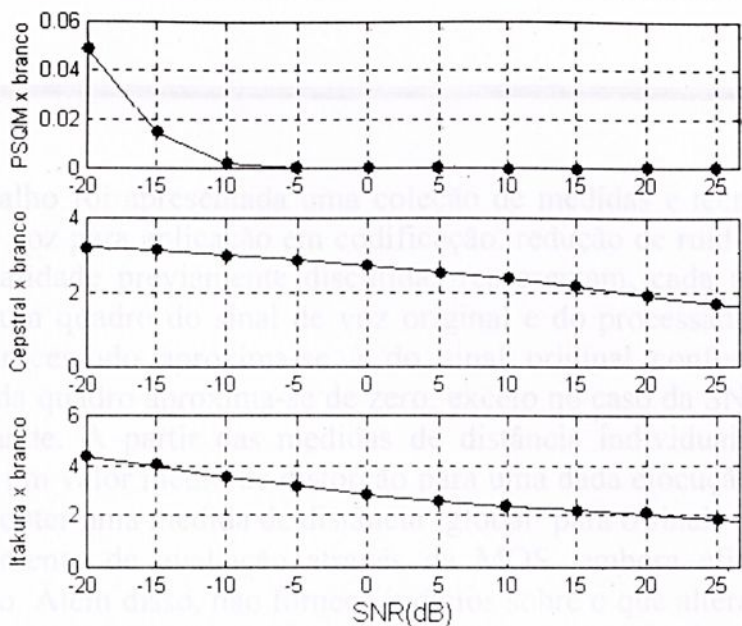


Figura 51. Sinal em Alemão x Ruído Branco

Para o sinal de voz em alemão, as curvas obtidas os diferentes tipos de medidas de distância apresentaram características semelhantes às obtidas para o sinal de voz em português. Todavia, para o PSQM, embora as curvas obtidas tenham apresentado características semelhantes, as amplitudes calculadas se mostraram bastante diferentes.

IX.3 MEDIDA SUBJETIVA - MOS

Tomaremos novamente como base os sinais de referência e ruidosos utilizados anteriormente para o cálculo das medidas objetivas.

Nesta avaliação os sinais foram apresentados para 4 diferentes ouvintes que deram notas de acordo com a diferença percebida entre os sinais. Assim, apresentando de maneira sequencial o sinal limpo e um dos sinais ruidosos, os ouvintes eram questionados sobre a distorção entre os sinais. Na escala utilizada o valor zero representa nenhuma distorção e o valor 5 representa distorção (desconforto) extremamente elevada:

	Ouvinte Nº 1	Ouvinte Nº 2	Ouvinte Nº 3	Ouvinte Nº 4	Valor Médio
Limpo x Ruidoso com $\alpha = 0,1$	3	3	4	3	3,25
Limpo x Ruidoso com $\alpha = 0,3$	2,5	2,5	3	1,5	2,40
Limpo x Ruidoso com $\alpha = 0,6$	1,5	1,5	2,5	1	1,60

Os resultados (pontuações) obtidos são coerentes com o resultado esperado, pois, quanto maior o nível de ruído introduzido, menor a qualidade do sinal. Esse tipo de medida, embora eficiente, é de difícil realização e requer a utilização de um grupo de ouvintes.

Neste trabalho foi apresentada uma coleção de medidas e técnicas de avaliação de qualidade de voz para aplicação em codificação, redução de ruído, etc. As medidas objetivas de qualidade previamente discutidas representam, cada uma, a medida de distorção entre um quadro do sinal de voz original e do processado. A qualidade do sinal de voz processado aproxima-se à do sinal original conforme a medida de qualidade de cada quadro aproxima-se de zero, exceto no caso da SNR que tende a um valor muito grande. A partir das medidas de distância individuais de cada quadro pode-se estimar um valor médio de distorção para uma dada elocução, ou um grupo de elocuições, para obter uma medida de distância 'global' para o sinal.

O procedimento de avaliação através da MOS, embora eficiente, é bastante trabalhoso e caro. Além disso, não fornece indícios sobre o que alterar para melhorar a qualidade do sinal (codificação / decodificação / transmissão). O mesmo ocorre com os métodos objetivos. Porém, por estes últimos se tratarem de procedimentos de cálculo matemático, eles permitem uma grande flexibilidade na alteração de parâmetros de forma a se verificar a sua influência nos resultados.

Já o PSQM, embora seja implementado através de um algoritmo, o resultado pode ser mapeada em uma escala de qualidade subjetiva. Pelo fato do PSQM possuir em seu processo de cálculo considerações que procuram reproduzir a percepção humana, seu valor de saída está diretamente relacionado com o julgamento de qualidade humano. Com isso, e observando-se os resultados apresentados, vê-se que o método PSQM é bastante eficiente para o cálculo da qualidade subjetiva, através de um procedimento de cálculo objetivo.

Através da análise de cada um dos padrões de medidas (objetivas e subjetivas) reconhecidamente eficientes, pode-se escolher uma medida de qualidade que seja adequada para uma aplicação em particular. Esta escolha é dependente do tipo de algoritmo usado no processamento do sinal de voz, do tipo de ruído presente neste sinal e dos requisitos do sistema, entre outros.

Medidas subjetivas de qualidade necessitam de planejamento cuidadoso, extensivos recursos de locutores e ouvintes. Medidas objetivas de qualidade são mais fáceis de aplicar, quantificam a performance do processamento através de uma medida numérica e permitem ao pesquisador/projetista alterar parâmetros de processamento e refazer todos os testes de maneira rápida.

Idealmente, seria interessante se obter um algoritmo de cálculo da qualidade do sinal de voz processado que reproduzi-se, perfeitamente, o julgamento humano.



Capítulo XI**REFERÊNCIA BIBLIOGRÁFICA**

- [1] Rabiner, L. R., Schafer, R.W., *Digital Processing of Speech Signals*, Bell Laboratories, USA, 1978.
- [2] Deller, J. R., Proakis, J. G., Hansen, John, H.L., *Discrete-Time Processing of Speech Signal*, Macmillan Publishing Company, USA, 1993.
- [3] Proakis, J.G., Manolakis, D.G., *Digital Signal Processing – Principles, Algorithms, and Applications*, Prentice Hall, Inc., USA, 1996.
- [4] Furui, S., *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, Inc., USA, 1989.
- [5] ITU-T Recomendação P.830 Avaliação Subjetiva da Performance de Codecs Digitais de Largura de Banda Telefônica e Banda Larga
- [6] ITU Recomendação P.861 Medição Objetiva de Qualidade de Codificadores de Voz de Banda Telefônica (300 - 3400 Hz)
- [7] ITU Recomendação P.800 Métodos para Determinação Subjetiva da Qualidade de Transmissão





UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Câmpus de Ilha Solteira
Programa de Pós-Graduação em Engenharia Elétrica
Av. Brasil Centro, 56
15385-000 Ilha Solteira - SP
www.dee.feis.unesp.br

