Dissertação de Mestrado

Inatel
Instituto Nacional de Telecomunicações

AVALIAÇÃO DA RELAÇÃO ENTRE QUALIDADE PERCEPTUAL DA FALA E TAXA DE ACERTO DE SISTEMAS DE RECONHECIMENTO DE FALA EM AMBIENTES RUIDOSOS

ANDRÉ GODOI CHIOVATO

DEZEMBRO / 2005

Avaliação da Relação entre Qualidade Perceptual da Fala e Taxa de Acerto de Sistemas de Reconhecimento de Fala em Ambientes Ruidosos

André Godoi Chiovato

Dissertação apresentada ao Instituto Nacional de Telecomunições, como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Carlos Alberto Ynoguti

Co-orientador: Prof. Dr. Francisco José Fraga da Silva

Santa Rita do Sapucaí
2005

Dissertação defendida e aprovada em 16/12/2005, pela comissão julgadora:

Prof. Dr. Carlos Alberto Ynoguti - DTE / Instituto Nacional de Telecomunicações - INATEL

Prof. Dr. Francisco José Fraga da Silva - PSI / Escola Politécnica da Universidade de São Paulo - POLI/USP

Prof. Dr. Miguel Arjona Ramirez - PSI / Escola Politécnica da Universidade de São Paulo - POLI/USP

Prof. Dr. Adonias Costa da Silveira
Coordenador do Curso de Mestrado

A meus pais Maurício e Adriana, pelos princípios ensinados que tanto honro e pela motivação em enfrentar os problemas deste objetivo alcançado. A meus irmãos Anelice e Marcell, que tanto amo.

Agradecimentos

Agradeço a Deus, por ter me dado saúde, paciência, serenidade e discernimento perante os desafios e as tomadas de decisões.

Ao Professor Doutor Francisco José Fraga da Silva pelos 5 anos de assistência ao meu trabalho, incluindo a pesquisa na Iniciação Científica e o curso de mestrado, pilares da minha formação científica.

Ao Professor Doutor Carlos Alberto Ynoguti, que me acolheu muito bem na difícil substituição do cargo de orientador. Obrigado pelos seus conselhos enquanto trabalhamos juntos no Programa de Estágio Docente do Inatel. Foram importantes para a minha visão de mundo e formação como educador.

À Dóris Vono, pela paciência e compreensão dos meus inúmeros dias de estudo. Por ter me escutado nestes momentos finais tão difíceis.

À Dair, por me apoiar e me oferecer cuidados e carinho. Sua casa foi o meu segundo lar.

Ao meu amigo Ramon, que inúmeras vezes me ajudou nos problemas técnicos.

Aos meus amigos Antônio André, Mário, Nilson e Rogério pelo apóio e alegres encontros naquele quiosque de Ubatuba.

À secretária Marilena Tobias pelos pedidos de ajuda atendidos com total prontidão durante o meu Programa de Estágio Docente.

Aos colegas Weberth e Antônio pela ajuda prestada no sistema operacional Linux.

Aos professores e funcionários do Inatel, que de alguma forma, sempre me ajudaram.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ACR	Absolute Category Rating
ATH	Absolute Threshold Hearing
DSR	Distributed Speech Recognition
EMSR	Ephraim and Malah noise Suppression Rule
EQM	Erro Quadrático Médio
ETSI	European Telecommunication Standard Institute
FFT	Fast Fourier Transform
HMM	Hidden Markov Models
HTK	Hidden Markov Models Toolkit
IDCT	Inverse Discrete Cosine Transform
IRS	Intermediate Reference System
IVR	Interactive Voice Response
MMSE	Minimum Mean-Square Error
MOS	Mean Opinion Score
NMT-PSS	Noise Masking Threshold - Power Spectral Subtraction
NR-WI008	Bloco Noise Reduction do algoritmo WI008
NR-WI008 + SWP	Bloco SNR-dependent Waveform Processing com o algoritmo NR-WI008
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure
PSD	Power Spectral Density
PSS	Power Spectral Subtraction
RAF	Reconhecimento Automático de Fala
RAFAR	Reconhecimento Automático de Fala para Ambiente Ruidoso

RMS	Root Mean Square
SPL	Sound Pressure Level
STQ Aurora	Speech processing, Transmission and Quality aspects: Working Group Aurora
STSA	Short-Time Spectral Amplitude
STSS	Short-Time Spectral Subtraction
SWP	SNR-dependent Waveform Processing
VAD	Voice Activity Detector
VADNest	Voice Activity Detector for Noise estimation
WI007	Front-end Feature Extraction Algorithm (Aurora 1)
WI008	Advanced Front-end Feature Extraction Algorithm (Aurora 1)

Lista de Símbolos

Símbolos utilizados no Capítulo 3

α	Parâmetro fixo usado na determinação do parâmetro atenuação espectral $G(w)$.
$\alpha(q, w)$	Parâmetro usado na determinação da função atenuação espectral $G(q, w)$, varia a cada quadro q e para a cada freqüência w.
α_{min}	Parâmetro fixo usado na determinação de $\alpha(q, w)$. Representa o menor valor que o parâmetro $\alpha(q, w)$ pode assumir.
α_{max}	Parâmetro fixo usado na determinação de $\alpha(q, w)$. Representa o maior valor que o parâmetro $\alpha(q, w)$ pode assumir.
β	Parâmetro fixo usado na determinação do parâmetro atenuação espectral.
$\beta(q, w)$	Parâmetro usado na determinação da função de atenuação espectral $G(q, w)$, varia a cada quadro q e para cada freqüência w.
β_{min}	Parâmetro fixo usado na determinação de $\beta(q, w)$. Representa o menor valor que o parâmetro $\beta(q, w)$ pode assumir.
β_{max}	Parâmetro fixo usado na determinação de $\beta(q, w)$. Representa o maior valor que o parâmetro $\beta(q, w)$ pode assumir.
γ	Parâmetro fixo usado na determinação do parâmetro atenuação espectral.
μ	Parâmetro utilizado no cálculo de R_{prio}.
ν	Parâmetro utilizado no cálculo de R_{prio} utilizado para ponderação entre os quadros $q-1$ e $q-2$.
$ATH(f)$	Absolute Threshold Hearing - Limiar Absoluto de Audição em cada freqüência [Hz].
B_k	Energia em cada banda crítica k somada.
C_k	Matriz resultante da convolução de B_k com a função espalhamento S_k que é o espectro de banda crítica espalhado.

$d(n)$	Ruído aditivo no domínio do tempo.
$\hat{D}(w)$	Estimativa do espectro médio do ruído a cada freqüência w.
f	Freqüência em Hz.
F_α	Função que leva a máxima redução do ruído residual para um limiar de mascaramento mínimo e a mínima redução do ruído residual para um limiar de mascaramento máximo.
F_β	Função que leva a máxima redução do ruído residual para um limiar de mascaramento mínimo e a mínima redução do ruído residual para um limiar de mascaramento máximo.
$G(w)$	Função ganho.
k	O número de banda crítica.
O_k	Limiar relativo para cada banda crítica k.
$P(w)$	Espectro de potência do sinal.
$R_{post}(q,w)$	Relação sinal ruído *a posteriori*, determinado a cada quadro q e para cada freqüência w.
$R_{prio}(q,w)$	Relação sinal ruído *a priori*, determinado a cada quadro q e para cada freqüência w.
$s(n)$	Sinal de fala limpo no domínio do tempo.
$\hat{S}(w)$	Estimativa do espectro do sinal de fala limpa a cada freqüência w.
S_k	Função de espalhamento.
$\hat{S}(q,w)$	Estimativa do espectro do sinal de fala limpa a cada quadro q e para cada freqüência w.
T_k	Limiar de mascaramento do ruído para cada banda crítica k.
$T(q,w)$	Limiar de mascaramento de ruído, varia a cada quadro e para cada freqüência w.
$y(n)$	Sinal de fala ruidosa no domínio do tempo.
$Y(w)$	Espectro do sinal de fala ruidosa a cada freqüência w.

Símbolos utilizados no Capítulo 4

$\eta(f,t)$	Relação sinal ruído para cada quadro t e freqüência f.
$(c(1),\ldots,c(12)+En)$	Coeficientes Mel-cepstrais.
En	Coeficiente de energia calculado a cada quadro.
$H(f,t)$	Resposta impulsiva do filtro de Wiener da primeira etapa do projeto.
$H_2(f,t)$	Resposta impulsiva do filtro de Wiener da segunda etapa do projeto.
p	Coeficiente de pré-ênfase.
$S_{in}(n)$	Sinal de fala de entrada no tempo.
S_{in_PSD}	Densidade Espectral de Potência média do sinal de entrada.
$S_{in}(f,t)$	Sinal de fala de entrada na freqüência f para cada quadro t.
S_{den}	Espectro do sinal de fala com menos ruído da primeira etapa do filtro de Wiener.
S_{den2}	Espectro do sinal de fala com menos ruído da segunda etapa do filtro de Wiener.
S_N	Estimativa do espectro do ruído.

Símbolos utilizados no Capítulo 5

L_p	Função ponderação aplicada a cada quadro N.
p	Constante utilizada na função de ponderação. Seu valor deve ser > 1.

Resumo

Este trabalho tem como objetivo avaliar distorção produzida no sinal de fala ruidoso ao ser realçado pelos algoritmos de redução de ruído. Esta avaliação é feita através da comparação entre taxa de acerto (%) de um sistema padronizado de Reconhecimento Automático de Fala (RAF) e medidas objetivas do índice (PESQ-MOS) da qualidade perceptual do sinal de fala, obtidas após aplicação de métodos de redução de ruído.

O cenário de testes, realizado sobre a base de dados de fala ETSI STQ-Aurora DSR Working Group e um sistema de reconhecimento padronizado, avaliou os seguintes algoritmos: WI008 (padrão ETSI STQ-Aurora), EMSR (algoritmo de supressão de ruído tradicional de Ephraim e Malah), NMT-PSS (algoritmo do tipo subtração espectral com características psico-acústicas) e EMSR + NMT-PSS (algoritmo baseado na regra de supressão de Ephraim e Malah, mas com o conceito de limiar de mascaramento do ruído).

Além disso, uma curva que modela a relação matemática entre o índice PESQ-MOS e a Taxa de Reconhecimento (%) é proposta. A intenção é predizer, em determinadas situações, o desempenho do sistema de RAF através da ferramenta PESQ. A aproximação é baseada na Curva Logística, cujos parâmetros de configuração possuem significados físicos validados pelos resultados experimentais.

E por fim, são apresentadas algumas análises que apontam vantagens e desvantagens dos tipos de ruído da base Aurora com relação ao desempenho do sistema de RAF padronizado.

Abstract

The goal of this work is to evaluate the distortion of the noisy speech signal being after enhanced by noise-reduction algorithms. This is performed by comparison of word accuracy (%) of a standardized Automatic Speech Recognition (ASR) system and objective measures of perceptual speech quality (PESQ-MOS score), obtained after applying noise-reduction methods.

The test scenario, composed of ETSI STQ-Aurora DSR Working Group database and a standardized ASR system, evaluated the following algorithms: WI008 (ETSI STQ-Aurora standard), EMSR (Ephraim and Malah noise Suppressor Rule algorithm), NMT-PSS (Noise Masking Threshold - Power Spectral Subtraction) and EMSR + NMT-PSS (EMSR algorithm with the concept of noise masking threshold).

Moreover, a curve that models the relationship between PESQ-MOS score and Recognition Rate (%) is proposed. The purpose is to predict, under certain conditions, the system performance by means of the PESQ evaluation. This approximation is based in the Logistic Curve, which configuration parameters have physical meanings, validated by experimental results.

Finally, some analysis are presented to indicate the advantages and disadvantages of several noise types present at Aurora 1 database over recognition system performance.

Capítulo 1

Introdução

A fala pode ser considerada como o meio mais amigável de comunicação entre os seres humanos, pois este sinal também carrega informações sobre o orador, sua emoção, e a língua por meio da qual ele se expressa. Portanto, não surpreende que as tecnologias de processamento da fala abram um imenso leque de aplicações no mundo moderno. Cada aplicação em vista tem que ser adequadamente analisada para saber se o processamento da fala pode ajudar, dadas as atuais potencialidades da tecnologia de fala. Além disto, um estudo completo dos fatores humanos envolvidos, como o aparelho falante e a audição, é imprescindível, pois cada vez mais, as técnicas se preocupam em modelá-los matematicamente.

Reconhecimento Automático de Fala (RAF) é a capacidade de uma máquina converter linguagem falada em palavras reconhecidas. Estas palavras podem ser a saída final do sistema ou a entrada de um sistema de processamento de linguagem natural. A ação em questão, que é uma função da aplicação, pode ser por exemplo uma agenda de telefone celular acionada por voz, um pedido de informação ou a conversão de uma entrada falada em texto. Sistemas de RAF podem ser valiosos em situações onde os olhos e/ou as mãos do operador estão ocupados em outras tarefas, como no caso de um piloto em uma aeronave ou automóvel, ou para pessoas deficientes.

Ainda não existe, e não se sabe se algum dia existirá, um sistema capaz de reconhecer a fala de qualquer pessoa, em ambiente indeterminado, pronunciada de qualquer maneira e abrangendo um vocabulário ilimitado em qualquer idioma.

Hoje em dia os reconhecedores de alto desempenho conseguem-no restringindo muitos dos parâmetros mencionados, isto é, limitam-se a um objetivo concreto, formando os assim chamados sistemas orientados a tarefas específicas, tais como reserva de passagens aéreas, acesso a dados bancários (seqüências de dígitos), comandos de fala para automóveis, etc. As principais especificações ou restrições de um sistema de reconhecimento automático de fala podem ser resumidas na

seguinte tabela:

Parâmetros	**Faixa de Variação**
Modo de Pronúncia	De palavras isoladas a fala contínua
Estilo de Pronúncia	De leitura a fala espontânea
Treinamento	De dependente de locutor a independente de locutor
Vocabulário	De pequeno ($<$ 20 palavras) a grande ($>$ 20.000 palavras)
SNR	De alta ($>$ 30 dB) a baixa ($<$ 10 dB)
Transdutor	De microfone com cancelamento de ruído a microfone de carvão

1.1 Reconhecimento Robusto para Condições Adversas

O sinal de fala é deteriorado em muitas situações por fatores do tipo:

1. Fisiológicos
    - distorções introduzidas pelo próprio locutor, como o efeito Lombard [18] (o aumento em esforço vocal quando locutor está num ambiente ruidoso), alta aceleração gravitacional, resfriados, sussurro e rouquidão;
    - uso de máscara de oxigênio em aeronaves;
2. Acústicos
    - saturação;
    - reverberação;
    - ruído sonoro do ambiente;
3. Canal de comunicação
    - eco;
    - atraso;
    - jitter;
    - latência;
    - perda de pacotes;
    - erros de bit;
    - atraso de grupo;

- desvanecimento por multi-percurso;
- ruído aditivo;
- interferência entre canais de comunicação adjacentes;
- ruído convolucional;

4. Codificação

- vários esquemas de compactação em cascata;
- erros e desacordo entre codecs (codificadores/decodificadores) de voz;

Dentre todos os problemas citados, para este trabalho apenas os *ruídos sonoros do ambiente* do fator *acústico* serão abordados, especificamente aqueles que representam os ambientes reais para a comunicação móvel (mais detalhes no Capítulo 5):

- Metrô
- Multidão de gente
- Carro
- Salão de exposições
- Restaurante
- Rua
- Aeroporto
- Estação de trem

O ser humano é capaz de reconhecer a fala sob condições de baixa relação sinal-ruído (inferior a 5 dB, dependendo do espectro de freqüências do ruído). Mas um reconhecedor automático de fala normalmente não consegue trabalhar em condições tão adversas. Então, o desenvolvimento de sistemas ótimos requer um tratamento sofisticado do sinal antes do reconhecimento, isto é, um aprimoramento do pré-processamento como o uso de microfones que possuem cancelamento de ruído, métodos que realçam os segmentos "falados" contaminados por segmentos ruidosos e técnicas que compensam o maior número de efeitos listados acima. Na seção seguinte, algumas das principais técnicas de aquisição robusta do sinal de fala implementadas no pré-processamento dos sistemas de RAF serão apresentadas.

1.2 Aquisição Robusta do Sinal de Fala no Pré-processamento

Nesta seção, serão apresentados os principais mecanismos dedicados ao aprimoramento da taxa de acerto dos Sistemas de Reconhecimento Automático de Fala para Ambientes Ruidosos - RAFAR, isto é, dos sistemas que trabalham com o sinal de fala degradado, ou quando as características do sinal de fala no ambiente de treinamento são diferentes daquelas nas quais será realizado o reconhecimento.

Em outras palavras, os desafios do reconhecimento robusto incluem solucionar as degradações acústicas produzidas pelo ruído aditivo, os efeitos lineares e não-lineares das filtragens, transduções ou transmissões, bem como as fontes de interferência e a redução da taxa de acerto devido às alterações na pronúncia da fala causadas pela intensidade do ruído. Alguns destes problemas estão ilustrados na Figura 1.1.

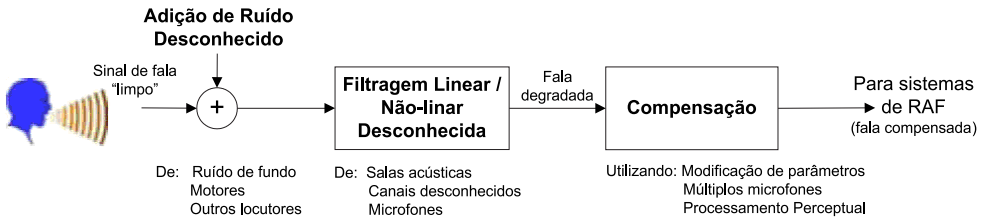

Figura 1.1: *Representação de algumas fontes de ruído (que podem degradar a taxa de reconhecimento de fala) e alguns procedimentos de compensação que propiciam uma aquisição robusta de informações do sinal de fala.*

Vale ressaltar que o foco deste trabalho está no bloco Compensação da figura anterior, com apresentações de diversos algoritmos que realçam o sinal de fala, descritos nos Capítulos 2 e 3.

1.2.1 Microfones com Cancelamento de Ruído

Os principais problemas existentes ao adquirir o sinal de fala através de microfones são: presença do ruído aditivo e ruído convolucional, que certamente afetarão o processo de reconhecimento de fala.

Para minimizar o ruído convolucional, do tipo reverberação, a primeira tentativa é trabalhar com microfones bem próximos da boca. Muitos experimentos com sistemas de RAF provaram a sua eficiência, principalmente quando utilizado microfones do tipo *headset*. Com relação à influência do canal de comunicação, no caso a função de transferência do próprio microfone, a solução seria adotar

um equipamento com resposta em freqüência adequada para a faixa de voz, que proporcione uma transdução sem distorções.

Contudo, para o caso dos reconhecedores no interior de automóveis, há mais um problema: o microfone de cabeça não é um atrativo, devido ao seu incômodo e conseqüentemente, as posições dos microfones devem ser otimizadas para a obter o mesmo desempenho [1]. Observe a Figura 1.2 indicando as possíveis posições do microfone.

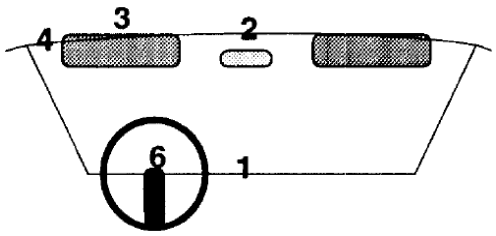

Figura 1.2: *Posições do Microfone. Pos. 1: no painel entre o passageiro e o motorista. Pos. 2: no teto do carro perto do espelho refletor. Pos. 3: na parte superior do pára-brisa de frente e à direita do motorista. Pos. 4: na parte superior do pára-brisa de frente e à esquerda do motorista. Pos. 5: acima da cabeça do motorista. Pos. 6: no volante. Pos. 7: no teto, entre o passageiro e o motorista. Segundo os estudos de [1], cada posição apresenta uma acústica diferente, sendo a Pos. 3 a que apresentou os melhores resultados com base na SNR e na taxa de acerto (utilizando o algoritmo de extração J-RASTA-PLP (ver 1.2.2) e o sistema HMM).*

Quanto ao ruído aditivo (ruído de fundo), uma tentativa de minimizá-lo seria utilizar um microfone direcional (útil quando a fonte de ruído vem de uma direção específica). Porém, este tipo de microfone perde a sua eficiência se as condições do ruído forem bastante severas, como por exemplo, no interior de um automóvel em alta velocidade e com os vidros abertos, ou no interior de um *cockpit* de uma aeronave.

E ainda, em altas freqüências (acima de 3 kHz), muitos dos microfones tornam-se direcionais [18]. Por exemplo, é comum o microfone de lapela possuir uma resposta em freqüência que enfatiza as componentes de alta freqüência (por serem direcionais) para compensar sua localização (abaixo e fora do fluxo sonoro emitido da boca do locutor).

Para melhorar a SNR, pode-se usar vários microfones com algoritmos que processam várias fontes de entrada. Como por exemplo, os algoritmos do tipo Cancelamento Adaptativo de Ruído, que utilizam um microfone (canal de referência) para estimar as características do sinal do ruído e outro para a captura do sinal de fala (canal primário), e em seguida subtrair um do outro. Observe a Figura 1.3 que apresenta o processo de Cancelamento Adaptativo utilizando dois

microfones. Ambos os sinais de entrada são processados e as duas interferências minimizadas. Maiores detalhes sobre esta técnica podem ser encontradas em [2], [3], [4] e [5].

Um ponto importante do Cancelamento Adaptativo de Ruído é a necessidade de capturar amostras similares do ruído tanto no microfone com sinal de fala corrompido quanto no microfone de referência, sem perdas ou incoerências, para garantir a sua eficiência. No entanto, em algumas aplicações, nem sempre isso é possível, pois ambos os microfones podem estar muito distantes, e então o ruído do microfone de referência não será similar ao do sinal de fala, dificultando a sua estimação, ou podem estar muito próximos, e então o sinal de fala poderá ser capturado também pelo microfone de referência, resultando numa degradação da informação útil.

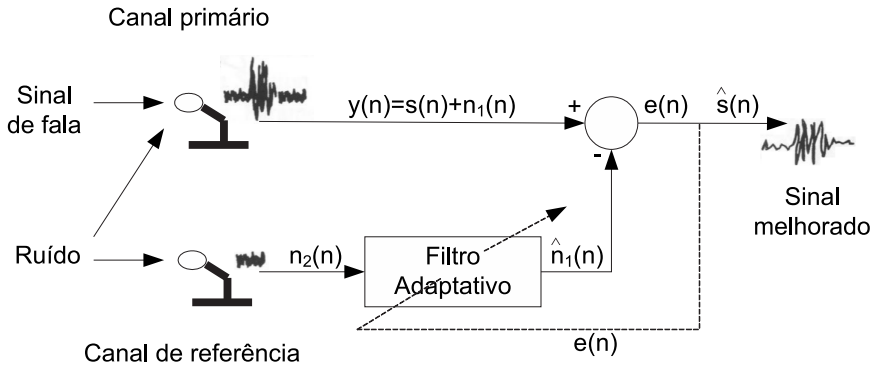

Figura 1.3: *Diagrama de blocos do Cancelamento Adaptativo de Ruído. O sinal $n_2(n)$ é filtrado para produzir uma saída estimada de $n_1(n)$, o qual é subtraído da entrada primária $y(n)$ para gerar um sinal realçado $\hat{s}(n)$.*

Uma outra técnica semelhante, é o cancelamento por Multi-microfones, o qual utiliza vários microfones posicionados em locais diferentes do ambiente e que, ao serem alinhados em fase e somados em amplitude, haverá um cancelamento natural do ruído [10]. Conseqüentemente, o sinal de fala é realçado. Esta técnica é bastante empregada no interior de aeronaves com hélice [11].

Outras aproximações mais conhecidas são os algoritmos [5], [6], [7], [8] e [9], que utilizam métodos adaptativos baseados na minimização da energia média quadrática. Estes algoritmos admitem que os sinais desejados são independentes de todas as fontes de interferência, fornecendo bons resultados quando o ruído é apenas aditivo. Entretanto, o mesmo não acontece quando o ruído é também convolucional.

Uma última abordagem do processamento de múltiplos microfones são os algoritmos baseados na correlação cruzada, os quais reforçam o som proveniente

de uma posição particular. Embora estes algoritmos sejam atrativos porque se assemelham ao sistema binaural humano, apresentam uma modesta melhoria com relação ao método Multi-microfones [10].

1.2.2 Algoritmos de Extração de Parâmetros Robustos do Sinal de Fala

O objetivo básico dos algoritmos que serão apresentados nesta seção, é extrair, em ambientes ruidosos, os parâmetros que representam o sinal de fala, combatendo principalmente o ruído aditivo e o convolucional.

Espectro Relativo com Predição Linear Perceptual - RASTA-PLP

Ao longo do amadurecimento natural das idéias propostas na área de processamento de voz, especialmente dos métodos robustos que extraem os parâmetros do sinal de fala, surgiu um aprimoramento da análise PLP (*Perceptual Linear Predictive*), conhecido como RASTA-PLP (*RelAtive SpecTrAl PLP*). Para um melhor entendimento, a Figura 1.4 apresenta as duas análise separadamente.

O princípio básico da análise PLP é aproximar o espectro auditivo * da fala por um modelo matemático "de-pólos" com características psico-acústicas, como a relativa insensibilidade do ouvido humano às variações muito lentas na freqüência e o fato de que o ruído de fundo, dependendo da sua intensidade, não prejudica a comunicação oral entre as pessoas.

A ordem desse filtro, que modela o espectro auditivo, é que especifica a quantidade de detalhes do sinal de fala [12]. Como é capaz de predizer a envoltória das formantes do sinal, esta técnica é bastante voltada aos sistemas com dependência de locutor, além de realizar uma redução da quantidade de informação, tornando-se computacionalmente eficiente. Estas características são bastante atraentes por serem úteis aos sistemas de RAF.

A análise RASTA-PLP, tem os seguintes objetivos: em primeiro lugar, eliminar algumas componentes do espectro auditivo (ruído aditivo) antes de estimar o modelo "de-pólos" e em segundo, combater o ruído convolucional [13].

Observe os dois algoritmos no diagrama de blocos apresentado na Figura 1.4. Primeiramente, ambos têm a ponderação da magnitude do espectro por meio da função Banda Crítica (banco de filtros na escala Bark). Somente para o RASTA, aplica-se um logaritmo na saída do banco possibilitando a subtração média da influência do ruído convolucional causado pelo canal de comunicação

*Por espectro auditivo, entende-se o espectro realmente "ouvido" pelo ser humano, incluindo o processamento realizado pelo ouvido interno

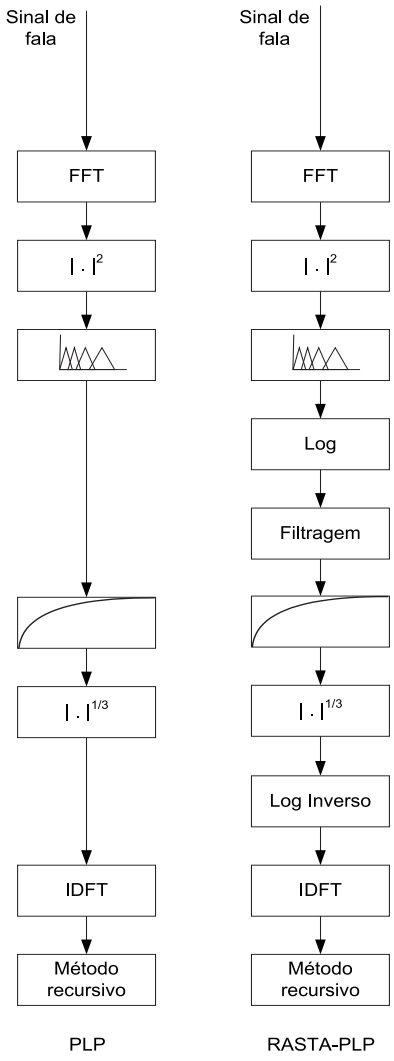

Figura 1.4: *Diagrama de blocos comparativo entre a análise PLP e RASTA-PLP sobre o sinal de fala.*

(microfones). Em seguida, uma filtragem temporal da trajetória, do tipo passa-altas ou passa-faixa, é realizada sobre os coeficientes das bandas críticas, que serve para compensar os efeitos do ruído convolucional.

Em outras palavras, a porção passa-altas do filtro (equivalente a um passa-faixa) alivia os efeitos do ruído aditivo e a porção passa-baixas alisa algumas das variações rápidas entre *frames*, inerentes à segmentação do sinal de fala. As variações lentas em freqüência do sinal de fala, correspondentes à influência do canal, também são suprimidas e assim, a representação paramétrica se torna menos sensível aos efeitos do ambiente (robustez) [13].

Em seguida, em ambos os algoritmos, multiplica-se o sinal por uma curva fixa *loudness*, atuando como uma pré-ênfase psico-física correspondente ao comportamento do ouvido humano. Depois, comprime-se as suas amplitudes através da Lei

da Potência de Audição, que consta de uma função raíz cúbica que simula a não-linearidade entre a intensidade do som e seu *loudness* percebido (também devido às características inerentes do ouvido humano). Por fim, para o PLP, aplica-se a IDFT (*Inverse Discrete Fourier Transform*) e uma solução qualquer do tipo auto-regressiva para os coeficientes, como por exemplo, o algoritmo de *Durbin*. O mesmo vale para o RASTA-PLP, mas antes da IDFT, aplica-se a transformação anti-logarítmica (exponencial). Como resultado, tem-se um modelo relativamente simples para ambos os casos.

J-RASTA-PLP

Várias melhorias do RASTA-PLP foram propostas, que dependem basicamente do tipo de filtragem e do banco de filtros não-linear (bandas críticas) utilizados. Por exemplo, na forma original, o RASTA-PLP minimiza certas componentes espectrais que se tornam aditivas no domínio espectral logarítmico (visto anteriormente). Entretanto, algumas componentes ruidosas aditivas e descorrelacionadas se tornam dependentes do sinal depois da operação logarítmica sobre o espectro de potência, impossibilitando a remoção pelo algoritmo RASTA-PLP [15]. Para resolver este problema, uma melhoria foi proposta em [16], que substitui a transformada logarítmica do RASTA-PLP por $y = ln(1 + J.X_k)$, onde J é uma constante positiva dependente da média da energia do ruído e k é o número da banda crítica. A vantagem desta modificação é um aumento na robustez contra o ruído aditivo e convolucional, se comparado com o RASTA-PLP [17].

RASTA-MFCC

Por último, tem-se a melhoria do RASTA envolvendo o domínio mais popular dentre os sistemas de RAF, o domínio *cepstral*, que depende tanto do tipo de normalização CMN (*Cepstral Mean Normalization*) quanto das características do ruído que deseja-se atacar. Este algoritmo fornece a melhor representação paramétrica e a compactação de informação mais eficiente (mais detalhes na seção 8.2.4.2 de [18]). A diferença entre o RASTA-MFCC (*Mel-Frequency Cepstral Coefficient*) e o RASTA-PLP é o tipo de filtragem temporal da trajetória sobre os parâmetros da fala, onde o primeiro possui um banco com bandas-críticas cuidadosamente ajustadas de acordo com as formantes do espectro, que são enfatizadas, resultando numa melhora na taxa de reconhecimento [19]. As técnicas de normalização CMN foram desenvolvidas pelo grupo de estudos da fala da Universidade Carnegie Mellon [20], [21], [22], [23], [24] e [25].

1.2.3 Processamento Fisiológico

Atualmente, há um grande número de esquemas de pré-processamento que envolvem aspectos fisiológicos e perceptuais do ouvido humano, tipicamente constituindo-se de um banco de filtros passa-faixa (representando a seletividade em freqüência do ouvido humano) seguido por mecanismos de extração de informações do sinal de fala tanto no domínio do tempo quanto na freqüência.

Todos os esforços estão voltados aos sistemas de RAFAR (*Reconhecimento Automático de Fala para Ambientes Ruidosos*), para melhorar a SNR e realçar a fala presente no sinal ruidoso. Como exemplos: Cohen, 1989 [26]; Ghitza, 1988 [27]; Lyon, 1982 [28]; Seneff, 1988 [29]; Hermansky, 1990 [12]; Patterson, Robinson, et al., 1991 [30]. Além disso, têm-se Ephraim e Malah, 1984 [43]; Virag, 1999 [40] e algoritmos padronizados pelo ETSI, WI007 (2000) [58] e WI008 (2002) [57], os quais compõem o foco principal deste trabalho (mais detalhes nos Capítulos 2 e 3).

1.3 Estrutura da Dissertação

A dissertação está organizada da seguinte maneira:

No Capítulo 2 são apresentadas as principais características dos algoritmos de subtração espectral de tempo curto. São apresentados também, três algoritmos de realce da fala: em primeiro, o algoritmo EMSR com uma explanação da regra de supressão de ruído de Ephraim e Malah. Em segundo, o algoritmo NMT-PSS proposto por Virag, com uma descrição teórica da estimação do parâmetro limiar de mascaramento do ruído, e por último, o algoritmo EMSR + NMT-PSS, o qual está fundamentado no EMSR, porém com conceitos do limiar de mascaramento do ruído, proposto em [37].

No Capítulo 3 é apresentado o algoritmo de pré-processamento avançado WI008, definido pelo ETSI como o algoritmo padrão de extração de parâmetros da fala para as comunicações móveis européias. Além disso, o algoritmo que serviu de base para esta padronização, o WI007, também será apresentado.

No Capítulo 4, é descrita a ferramenta de avaliação perceptual da qualidade da fala, o PESQ, padronizada pelo ITU-T.

No Capítulo 5, é especificada a base de dados e o sistema de reconhecimento utilizados.

No Capítulo 6, são descritos os resultados obtidos e as condições dos ensaios. As conclusões são apresentadas no Capítulo 7.

O Anexo A contém a descrição detalhada do algoritmo PESQ utilizado no

Capítulo 6.

O Anexo B apresenta tabelas com os valores da Taxa de Reconhecimento (%) e da pontuação PESQ-MOS para cada tipo de ruído de cada cenário de teste.

O Anexo C apresenta tabelas com valores da Taxa de Reconhecimento (%) do algoritmo WI008 utlizados no Capítulo 6, determinado pelo ETSI como sendo o algoritmo com o melhor desempenho.

Capítulo 2

Supressão de Ruídos em Sinais de Fala

O foco deste capítulo é apresentar três algoritmos de redução de ruído do tipo *single channel* aplicados no pré-processamento de sistemas de reconhecimento: EMSR (*Ephraim and Malah noise Suppression Rule*), NMT-PSS (*Noise Masking Threshold - Power Spectral Subtraction*) e EMSR + NMT-PSS (modificação da regra Ephraim e Malah proposta em [37], com a introdução do conceito de limiar de mascaramento do ruído para trabalhar a baixa SNR (< 10dB)).

2.1 Algoritmos de Atenuação Espectral de Tempo Curto

No desenvolvimento de diversos algoritmos na área de realce de sinais de voz com apenas um canal, a subtração espectral de tempo-curto (STSS - *Short-Time Spectral Subtraction*) da fala tem sido amplamente utilizada. A idéia básica é usar a magnitude espectral de tempo-curto da fala ruidosa e recuperar uma estimativa da amplitude espectral de tempo-curto da fala limpa removendo o ruído aditivo. A Figura 2.1 mostra uma representação básica das técnicas de processamento que utilizam STSS.

Um dos problemas dos algoritmos baseados na subtração espectral, é o surgimento de um resíduo chamado "ruído musical" na fala melhorada [38], que consiste de picos isolados e/ou cristas curtas no espectrograma, que correspondem aos pontos do quadro espectral atual onde a magnitude local excedeu a estimativa de ruído médio.

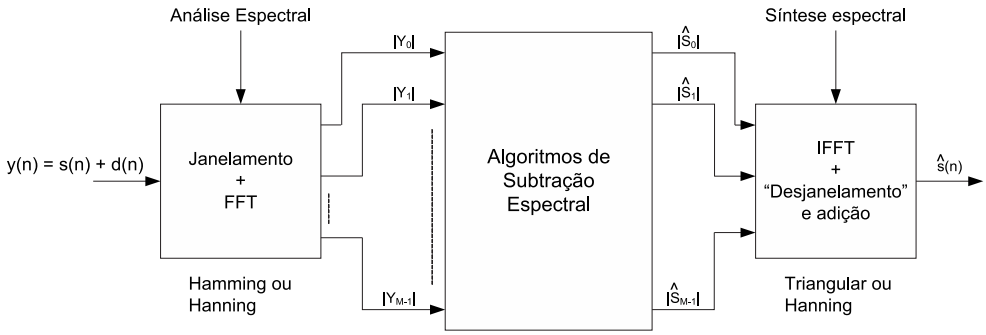

Figura 2.1: *Diagrama simplificado da subtração espectral de tempo-curto. Na entrada tem-se o sinal de fala corrompido $y(n)$. A análise e síntese é feita quadro a quadro num tempo curto (10 a 40 ms) de acordo com o conceito de estacionariedade do sinal de fala.*

2.1.1 Algoritmos Subtrativos Convencionais

Subtração Espectral

A subtração espectral é um método de subtração de ruído baseado na técnica de estimação STSS. É muito conhecida por possuir um conceito simples e ser de fácil implementação. Foi apresentada pela primeira vez por Boll em 1979 [39].

Para a análise desta técnica, seja um sinal de fala $s(n)$ corrompido por ruído aditivo estacionário $d(n)$. Desta forma, tem-se:

$$y(n) = s(n) + d(n) \tag{2.1}$$

A fala e o ruído são assumidos como sendo descorrelacionados e o processamento é feito a cada quadro no domínio da freqüência. O propósito da subtração espectral é obter uma estimativa do sinal limpo, que denomina-se de $\hat{s}(n)$, a partir do sinal ruidoso $y(n)$ e de uma estimativa do ruído $d(n)$.

No algoritmo básico de subtração espectral de potência PSS (*Power Spectral Subtraction*) a magnitude da transformada de Fourier de tempo-curto (FFT) é estimada como:

$$|\hat{S}(w)|^2 = \begin{cases} |Y(w)|^2 - |\hat{D}(w)|^2, & se \quad |Y(w)|^2 > |\hat{D}(w)|^2 \\ 0 & c.c. \end{cases} \tag{2.2}$$

onde $|\hat{D}(w)|^2$ representa uma estimativa do espectro médio de potência do ruído. O espectro do ruído é estimado nas pausas que ocorrem durante a fala.

A fase da fala ruidosa não é modificada, baseado no fato da distorção da fase ter pouca influência na percepção humana. Portanto, o melhor resultado possível, em qualquer algoritmo do tipo subtrativo, é obtido ao sintetizar a fala utilizando

a magnitude espectral do sinal limpo e a fase espectral do sinal ruidoso. Esta situação é chamada de *limite teórico* [40].

Subtração Espectral como Filtragem

São algoritmos do tipo subtrativos que podem ser vistos também como algoritmos de atenuação espectral de tempo curto, onde é feita uma filtragem da fala ruidosa com um filtro variável no tempo, que depende do espectro do sinal ruidoso e do espectro estimado do ruído. O processo é equivalente à multiplicação da magnitude do espectro de tempo curto da fala ruidosa pela função ganho:

$$|\hat{S}(w)| = G(w).|Y(w)|, \quad 0 \leq G(w) \leq 1 \tag{2.3}$$

A função ganho mais conhecida para filtros do tipo PSS, foi proposta por Berouti *et al.* [41].

$$G(w) = \begin{cases} \left[1 - \alpha . \left(\frac{|\hat{D}(w)|}{|Y(w)|}\right)^{\gamma}\right]^{1/\gamma}, & se \quad \left(\frac{|\hat{D}(w)|}{|Y(w)|}\right)^{\gamma} < \frac{1}{\alpha + \beta} \\ \left[\beta . \left(\frac{|\hat{D}(w)|}{|Y(w)|}\right)^{\gamma}\right]^{1/\gamma}, & c.c. \end{cases} \tag{2.4}$$

Esta função, $G(w)$, caracteriza um dos algoritmos do tipo subtrativo mais flexíveis. Os parâmetros responsáveis por esta flexibilidade são α, β e γ. Ao reduzir o α aumenta-se o ruído musical e diminui a distorção da fala. A redução de β também aumenta o ruído musical, mas reduz o ruído de fundo que permanece na fala melhorada. O parâmetro γ apenas diz respeito ao formato da curva de transição de $G(w) = 1$ (onde a componente espectral não é modificada) para $G(w) = 0$ (onde a componente espectral é suprimida).

A escolha adequada destes três parâmetros é fundamental, mas quando se trabalha com sinais que apresentam baixa relação sinal-ruído (menor que 10 dB), torna-se impossível minimizar a distorção da fala e o ruído musical, simultaneamente. Mas, ao considerar a SNR de cada componente espectral dos quadros anteriores, é possível minimizar o efeito do ruído musical enquanto mantém-se a distorção da fala no mínimo possível. Este é o foco da próxima técnica, EMSR, que será discutida.

2.1.2 Algoritmo EMSR

A técnica EMSR (*Ephraim e Malah noise Suppression Rule*) é um sistema de melhoria de sinais de fala, capturados com ruído pelo mesmo microfone (*single channel*), que é baseado na derivação de um estimador ótimo de amplitude (magnitude) espectral de tempo curto (STSA - *Short-Time Spectral Amplitude*), o qual busca minimizar o erro quadrático médio (MMSE - *Minimum Mean-Square Error*) entre a amplitude espectral de tempo curto (STSA) original e sua estimação.

De uma maneira em geral, o algoritmo EMSR tenta encontrar um compromisso entre a quantidade de ruído a ser removida e as distorções introduzidas no sinal de voz devido ao processamento realizado. Em outras palavras, ele realiza uma redução moderada do ruído de fundo (que não apresenta eficiência para SNRs menores que 10 dB) enquanto evita completamente o aparecimento do ruído musical.

O algoritmo original foi proposto por Ephraim e Malah pela primeira vez em 1983 [42]. No ano seguinte, o seu desenvolvimento matemático foi detalhado em outra publicação [43]. Seguindo os mesmos princípios, outras regras de supressão foram propostas [43] e [44]. Neste trabalho, será focado somente o princípio original de EMSR desenvolvido em [43], pois o mecanismo responsável por eliminar o ruído é basicamente o mesmo em todas as regras de supressão propostas. A Figura 2.2 mostra o algoritmo em diagrama de blocos.

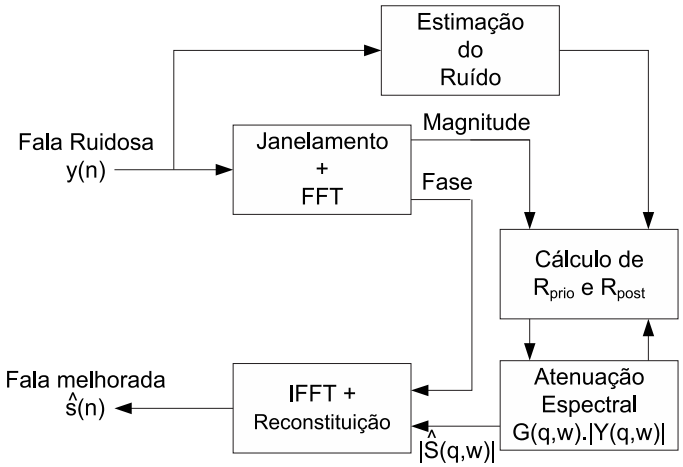

Figura 2.2: *Diagrama de blocos do algoritmo original Ephraim e Malah.*

Como já foi dito, o EMSR é um tipo de algoritmo de atenuação espectral de tempo curto. Seu ganho espectral é dado como $G(q, w)$, isto é, varia com o espectro e quadro, que é aplicado sobre a magnitude de cada componente espectral ruidosa também de tempo curto $|Y(q, w)|$, equivalente à Equação (2.3). A função

ganho $G(q, w)$ é dada por:

$$G(q,w) = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+R_{post}}\right)\left(\frac{R_{prio}}{1+R_{post}}\right)}.M\left[(1+R_{post}\left(\frac{R_{prio}}{1+R_{post}}\right)\right] \quad (2.5)$$

$$M[\Theta] = \exp\left(-\frac{\Theta}{2}\right)\left[(1+\Theta)I_0\left(\frac{\Theta}{2}\right) + \Theta I_1\left(\frac{\Theta}{2}\right)\right] \quad (2.6)$$

onde I_0 e I_1 são as equações de Bessel modificadas.

Isso é feito afim de obter-se uma estimativa da magnitude $|\hat{S}(q,w)|$ da componente espectral w da fala limpa no *q-ésimo* quadro. Na Equação (2.5), omitiu-se a indexação quadro q e a freqüência w.

O ganho $G(q, w)$, depende de dois parâmetros, apresentados nas Equações (2.7) e (2.8).

$$R_{post}(q,w) = \begin{cases} \dfrac{|Y(q,w)|^2}{|\hat{D}(w)|^2} - 1, & se \quad \dfrac{|Y(q,w)|^2}{|\hat{D}(w)|^2} > 1 \\ 0, & c.c. \end{cases} \quad (2.7)$$

onde $|\hat{D}(w)|^2$ é a potência estimada do ruído na freqüência w. O parâmetro R_{post} é a relação sinal-ruído *a posteriori*, computada do quadro de tempo curto atual q para cada componente espectral w.

$$R_{prio}(q,w) = (1-\mu).R_{post}(q,w) + \mu.\frac{|G(q-1,w).Y(q-1,w)|^2}{\hat{D}(w)^2} \quad (2.8)$$

O parâmetro R_{prio} é a relação sinal-ruído *a priori*. Pode-se verificar que na primeira parcela de (2.8) efetua-se uma ponderação de $(1-\mu)$ na relação sinal ruído do quadro atual, R_{post}. Ainda em (2.8), tem-se o fator $G(q-1,w).Y(q-1,w)$, que é uma estimativa do espectro do sinal sem ruído do quadro anterior. Desta forma, tem-se na segunda parcela de (2.8) a relação sinal-ruído do quadro anterior, que é ponderada pelo parâmetro μ. Pode-se então compreender a razão da denominação *a priori*, pois o cálculo de R_{prio} depende de uma estimativa do sinal limpo, a qual é feita baseada em um quadro prévio, enquanto para a determinação de R_{post} não necessita-se de nenhum conhecimento prévio do sinal ruidoso no quadro atual.

A influência dos parâmetros R_{post} e R_{prio}

O parâmetro R_{prio}, relação sinal-ruído *a priori*, é obtido pela Equação (2.8) e é o parâmetro dominante em (2.5). Como pode ser visto na Figura 2.3, fortes atenuações são obtidas somente se R_{prio} é baixa e baixas atenuações são obtidas somente se R_{prio} é alta.

Quando R_{prio} é baixa, R_{post} atua como parâmetro de correção (como mostra o lado esquerdo da Figura 2.3). Quando R_{prio} é baixa e R_{post} é alta, o efeito é oposto ao intuitivamente esperado e existe uma atenuação muito forte. Este comportamento é conseqüência do desacordo entre as relações sinal-ruído *a priori* e a *posteriori*, sendo muito útil na eliminação do ruído residual.

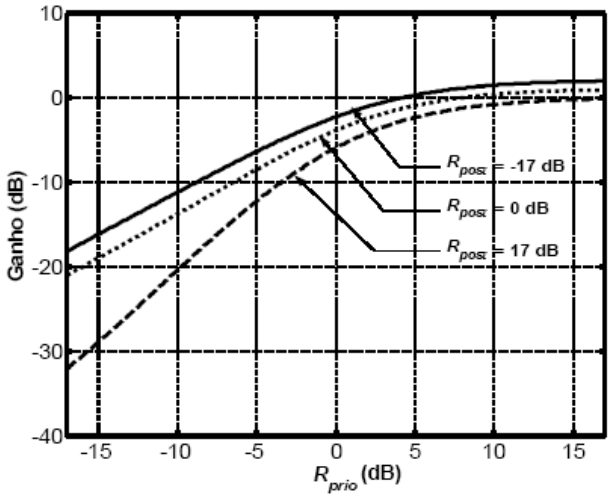

Figura 2.3: *O ganho EMSR versus R_{prio}, para diferentes valores de R_{post}.*

Um estudo experimental de (2.8) [42] mostra dois comportamentos diferentes para a relação sinal-ruído *a priori*, os quais estão expressos na Figura 2.4.

1. Quando a relação sinal-ruído *a posteriori*, R_{post}, fica abaixo ou próxima a 0 dB, R_{prio} corresponde a uma versão altamente suavizada de R_{post} ao longo de sucessivos quadros. Conseqüentemente, a variância de R_{prio} para cada freqüência, é muito menor que a de R_{post} ao longo de sucessivos quadros. Pode-se notar, no lado esquerdo da Figura 2.4, como a curva de R_{prio} é muito mais suave que a curva de R_{post}.

2. Quando R_{post} é muito maior que 0 dB, R_{prio} segue R_{post} ao longo de sucessivos quadros. Como pode ser visto no lado direito da Figura 2.4, nos últimos 20 quadros, R_{prio} segue R_{post} com apenas um quadro de atraso. Este comportamento explica-se ao serem feitas as seguintes considerações com relação a (2.8):

- Quando R_{prio} é alta, $G(q,w) \cong 1$ (lado direito da Figura 2.3), portanto tem-se:

$$R_{prio}(q,w) \cong (1-\mu).R_{post}(q,w) + \mu.\frac{|Y(q-1,w)|^2}{\hat{D}(w)^2} \tag{2.9}$$

- Como $R_{post} \gg 1$:

$$R_{prio}(q,w) \cong (1-\mu).R_{post}(q,w) + \mu.R_{post}(q-1,w) \tag{2.10}$$

- E considerando que μ é geralmente escolhido próximo a 1 (para estes experimentos, 0.98), a Equação (2.8) pode ser aproximada por:

$$R_{prio} \cong (\mu).R_{post}(q-1,w) \tag{2.11}$$

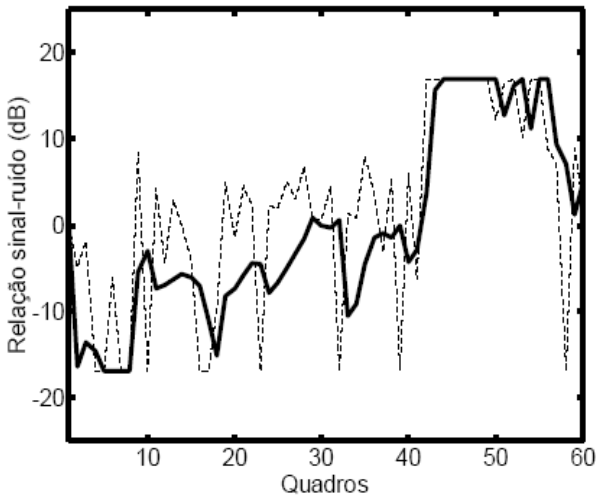

Figura 2.4: *As relações sinal-ruído R_{post} e R_{prio} ao longo de sucessivos quadros. Curva de linha contínua: R_{prio}; Curva de linha pontilhada: R_{post}. Nos 40 primeiros quadros, o sinal contém somente ruídos na freqüência escolhida e para os 20 quadros seguintes, surge uma componente com mais de 15 dB de relação sinal-ruído na freqüência mostrada.*

Quando o nível do sinal é bem superior ao do ruído, como ocorre nos últimos 20 quadros da Figura 2.4, $R_{prio}(q,w)$ não é mais uma versão suavizada da relação sinal-ruído e sim uma versão atrasada de R_{post}, o que é importante no caso de sinais não-estacionários.

2.1.3 Algoritmo NMT-PSS

A preocupação com o efeito psico-acústico do ruído surgiu na codificação [45] e compressão [46] de sinais de audio que seriam transmitidos em canais de baixa taxa de transmissão ou que seriam armazenados (como, por exemplo, o padrão MPEG [47]). Em pesquisas recentes, o modelo auditivo humano também tem sido bastante explorado [40], [48], [49], com o objetivo de realizar uma melhoria dos sinais de fala utilizando critérios baseados na percepção humana. Nestes algoritmos, como é o caso do NMT-PSS (*Noise Masking Threshold - Power Spectral Subtraction*), a preocupação não é remover completamente o ruído do sinal, especialmente o ruído musical, e sim atenuá-lo abaixo do limiar auditivo. No contexto de algoritmos subtrativos STSS, isso reduz a quantidade de modificação na amplitude espectral, reduzindo artefatos audíveis e contribuindo para obter-se um sinal de alta qualidade.

Antes de descrever o algoritmo NMT-PSS, deve-se levantar algumas propriedades do sistema auditivo humano e apresentar o cálculo do limiar de mascaramento do ruído. As propriedades do sistema auditivo humano são:

- **Limiar Auditivo Absoluto:** define a pressão mínima para que os sons sejam audíveis. Devido à sensibilidade seletiva do ouvido humano, esta pressão varia consideravelmente com a freqüência. A Figura 2.5 mostra o nível de pressão do som SPL (*Sound Pressure Level*) mínimo necessário para que 10%, 50% e 90% de pessoas, de 20 a 25 anos de idade, possam ouvir um tom de teste em ambiente silencioso.

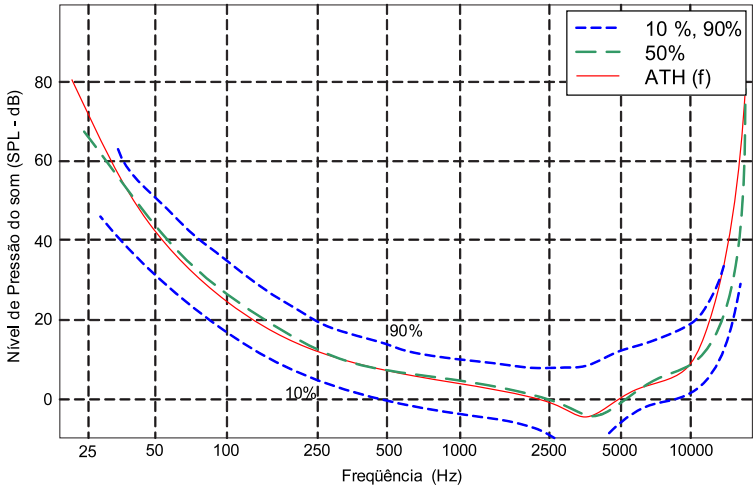

Figura 2.5: *Limiar de audição no silêncio. A linha contínua corresponde ao limiar absoluto de audição (ATH - Absolute Threshold of Hearing).*

- **Banda Crítica:** de acordo com [50], banda crítica é uma faixa de freqüência dentro da qual um som complexo não pode ter todas as suas componentes individualmente identificadas. Dependendo da diferença de intensidade entre as componentes de freqüência de um determinado som, elas só poderão ser individualmente distinguidas se ocorrerem em bandas críticas diferentes. Mais detalhes quanto às faixas de freqüência correspondentes a cada banda crítica do ouvido humano podem ser encontrados em [40].

- **Mascaramento simultâneo:** ocorre no domínio da freqüência, onde um sinal fraco se torna inaudível pela presença de um sinal mais forte e de freqüência próxima, ocorrendo simultaneamente *. Quando um tom de menor intensidade não puder mais ser percebido, diz-se então que ele foi "mascarado". A Figura 2.7 mostra a quantidade de mascaramento provido por um tom de 1kHz para vários níveis L_M de pressão absoluta SPL do som.

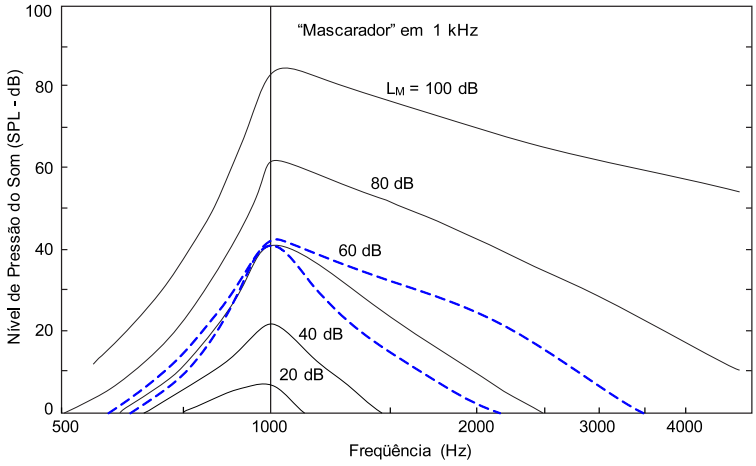

Figura 2.6: *Curvas de mascaramento para tom "mascarador" de 1 kHz, adaptado de [51]. É importante destacar que estas curvas são apenas médias, visto que elas variam de pessoa para pessoa, o que está ilustrado nas curvas pontilhadas que mostram o mascaramento causado por um tom puro de 60 dB para duas pessoas diferentes.*

O cálculo do limiar de mascaramento do ruído é dado na Figura 2.7. No primeiro bloco, o espectro do sinal $y(n)$ é dividido de acordo com as bandas críticas e a energia de cada banda, B_k, é calculada. Para aplicações em processamento de voz, na faixa de telefonia, tem-se o número de bandas críticas k

*Este modelo apresenta um bom desempenho mesmo não considerando o mascaramento temporal, que se refere ao efeito do mascaramento de sinais ocorrendo com uma pequena diferença de tempo.

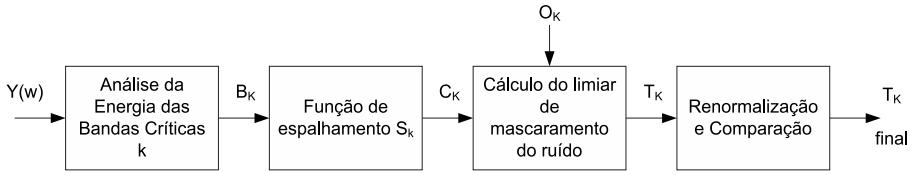

Figura 2.7: *Cálculo do limiar de mascaramento.*

= 18 que cobrem a faixa de 0 a 4 kHz. O segundo bloco, Função de Espalhamento S_K [51], trata da estimação dos efeitos do mascaramento entre diferentes bandas críticas, pois embora a percepção do ouvido aconteça em bandas críticas diferentes, porções do sinal de cada banda interferem nas bandas próximas, ocasionando espalhamento de energia. O terceiro bloco, Cálculo do Limiar de Mascaramento do Ruído, encontrado em [50], [52] e [53], pode ser de dois tipos: *limiar para tom mascarando ruído* (estimado em $14.5 + B_k$ dB abaixo do espectro de banda crítica espalhada C_k) e *limiar para ruído mascarando tom* (estimado como 5.5 dB abaixo de C_k). Após determinar se o sinal tem natureza tonal ou de ruído, calcula-se o Limiar de Mascaramento do ruído T_k subtraindo o limiar relativo O_k (de Sinha e Tewfik [†] [54]) do espectro de banda crítica espalhado C_k. Por último, é realizado uma renormalização da energia estimada (multiplicando T_k pelo inverso do ganho da energia provocado pelo espalhamento) e uma comparação com as medidas de limiar absoluto da audição [55] (qualquer limiar T_k abaixo do limiar absoluto da audição é substituído pelo limiar absoluto para aquela banda crítica).

No método para o cálculo do limiar de ruído, como descrito acima, o limiar de mascaramento de ruído T_k deve ser calculado a partir de um espectro de potência de fala limpa. No entanto, na prática, apenas o sinal de fala ruidosa está disponível. Então, é feita uma estimativa do sinal de fala limpa usando-se um esquema simples de subtração espectral de potência.

Virag [40], utiliza a estimativa do limiar de mascaramento do ruído para cada banda crítica T_k para ajustar os parâmetros α e β da Equação (2.4), a cada quadro q e para cada freqüência w. Na Figura 2.8 tem-se o diagrama de blocos do algoritmo NMT-PSS.

A adaptação dos parâmetros é realizada da seguinte forma:

$$\alpha(q,w) = F_\alpha\left[\alpha_{min}, \alpha_{max}, T(q,w)\right] \tag{2.12}$$

$$\beta(q,w) = F_\beta\left[\beta_{min}, \beta_{max}, T(q,w)\right] \tag{2.13}$$

[†]Esse método considera que o sinal de fala, em média, possui uma natureza tonal em bandas críticas mais baixas e uma natureza de ruído em bandas críticas mais altas.

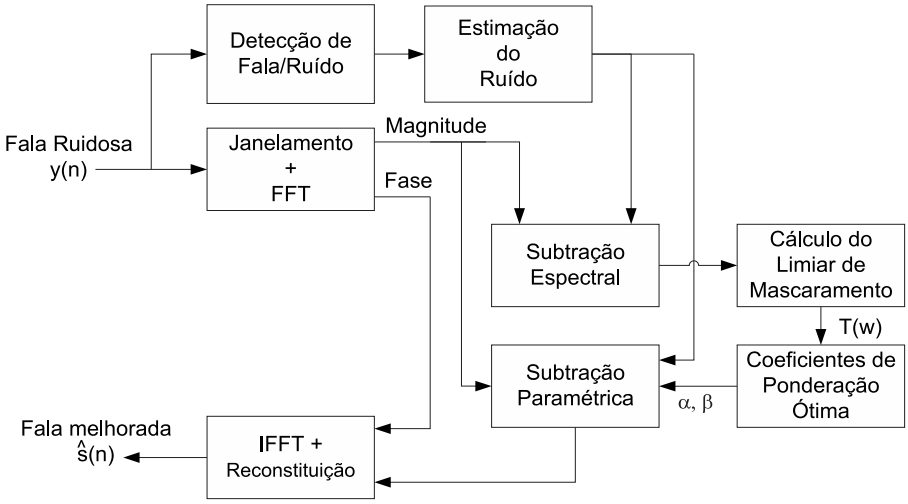

Figura 2.8: *Esquema do melhoramento perceptual da fala NMT-PSS.*

onde α_{min}, α_{max}, β_{min} e β_{max} limitam $\alpha(q,w)$ e $\beta(q,w)$ e $T(q,w)$ é a estimativa do limiar de mascaramento como visto anteriormente. As funções F_α e F_β levam a uma máxima redução de ruído residual para mínimos limiares de mascaramento e vice-versa. Em outras palavras, $F_\alpha = \alpha_{max}$ se $T(q,w) = T(q,w)_{min}$ e $F_\alpha = \alpha_{min}$ se $T(q,w) = T(q,w)_{max}$.

Os parâmetros $T(q,w)_{min}$ e $T(q,w)_{max}$ são os valores mínimo e máximo do limiar de mascaramento atualizados a cada quadro. Os valores de F_α entre esses dois extremos são interpolados linearmente com base no valor de $T(q,w)$. As mesmas considerações podem ser feitas com respeito a F_β.

Após terem sido ajustados os parâmetros α e β, baseado em $T(q,w)$, estes valores são substituídos na Equação 2.4. O parâmetro $\alpha(q,w)$ será denominado de *parâmetro perceptual de atenuação* quando utilizado no algoritmo da seção seguinte.

Os melhores valores escolhidos em [40], com melhor compromisso entre ruído residual e distorção da fala, foram:

$$\alpha_{min} = 1 \qquad \alpha_{max} = 6$$
$$\beta_{min} = 0 \qquad \beta_{max} = 0.02$$

Parâmetros fixos:

$$\gamma = \gamma_1 = 2 \qquad \gamma_2 = 0.5$$

2.1.4 Algoritmo EMSR + NMT-PSS

Como já mencionado, o algoritmo de Ephraim e Malah permite apenas uma moderada redução de ruído, o que é insuficiente quando a relação sinal-ruído é muito baixa (SNR < 10 dB), isto é, uma quantidade considerável do ruído original permanece no sinal melhorado. Por esta razão, este novo algoritmo é baseado no EMSR, mas com modificações que permitiram obter melhor desempenho no caso de sinais ruidosos com relações sinal-ruído baixas, usando o conceito de limiar de mascaramento auditivo explicado na seção anterior. Este algoritmo foi proposto em [37].

O algoritmo também é baseado na atenuação espectral de tempo curto por filtragem. A função ganho $G(q,w)$ utilizada é a mesma que foi apresentada na Equação 2.5 da seção 2.1.2. No entanto, as relações sinal-ruído *a posteriori* e a *priori*, que são básicas para o cálculo de $G(q,w)$, foram modificadas, sendo agora derivadas por meio das seguintes relações:

$$R_{post}(q,w) = \frac{|Y(q,w)|^2}{\alpha(q,w)|\hat{D}(w)|^2} - 1 \tag{2.14}$$

$$\begin{aligned} R_{prio}(q,w) = &[(1-\mu)R_{post}(q,w)] + \mu.\nu.G^2(q-1,w).\frac{|Y(q-1,w)|^2}{\alpha(q,w)|\hat{D}(w)|^2} + \\ &+ \mu.(1-\nu).G^2(q-2,w)\frac{|Y(q-2,w)|^2}{\alpha(q,w).|\hat{D}(w)|^2} \end{aligned} \tag{2.15}$$

onde μ e ν foram, experimentalmente, escolhidos como 0.96 e 0.75, respectivamente. O parâmetro $\alpha(q,w)$ representa o fator perceptual de atenuação, cujo cálculo foi detalhado na seção anterior. A Figura 2.9 apresenta o diagrama de blocos funcionais do algoritmo.

Justificativas para as alterações introduzidas no EMSR

1. O parâmetro perceptual de atenuação $\alpha(q,w)$, é utilizado seguindo a mesma idéia de Berouti [41]. Como apresentado na seção 2.1.1, Berouti propôs em seu algoritmo a sobre-estimação do ruído, fixando um valor de α que, multiplicado pela potência espectral média do ruído, eliminaria maior quantidade de ruído de fundo. Já na seção anterior, Virag aplica o parâmetro $\alpha(q,w)$ na Equação (2.4) de Berouti, mas agora em função do limiar de mascaramento do ruído, obtendo assim, melhores resultados que Berouti. No entanto, ambos algoritmos ainda mantiveram ruído musical no sinal de fala melhorado. Por esta razão, este algoritmo propõem utilizar o parâmetro de atenuação

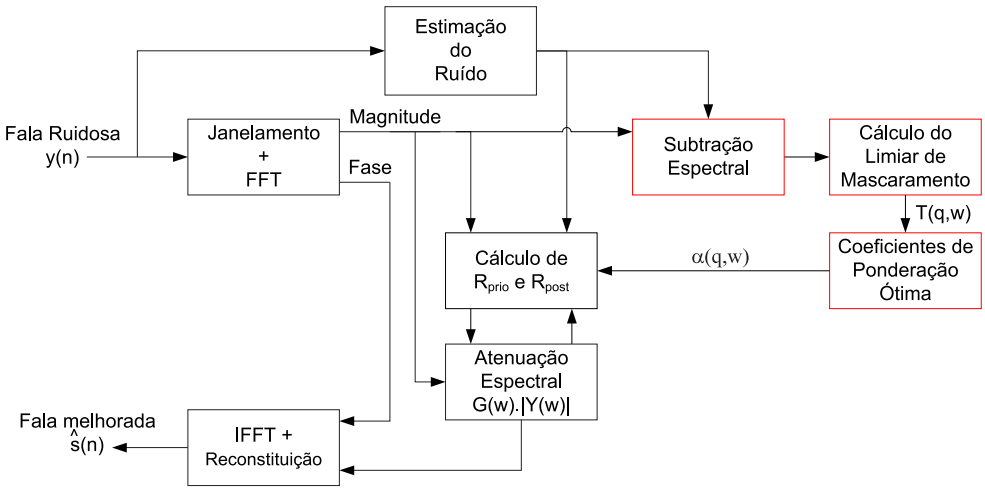

Figura 2.9: *Diagrama em blocos do EMSR + NMT-PSS. Em destaque, as modificações propostas.*

$\alpha(q, w)$ nas Equações (2.14) e (2.15), que na verdade são adaptações do método EMSR, devido à eficiência deste na eliminação do ruído musical. Nestas equações, quando o parâmetro $\alpha(q, w)$ é multiplicado pela estimativa de potência espectral média do ruído, o espectro médio do ruído é sobre-estimado, o que resulta em uma atenuação adicional do ruído de fundo. Os parâmetros básicos, escolhidos empiricamente para o cálculo de $\alpha(q, w)$ em (2.12), com melhor compromisso entre ruído residual e distorção da fala, foram:

$$\alpha_{min} = 0.75$$
$$\beta_{min} = 2.5$$

Ao introduzir o parâmetro $\alpha(q, w)$, constatou-se experimentalmente que o parâmetro $\mu = 0.96$ proporcionou melhores resultados perceptuais se comparado ao valor $\mu = 0.98$ no método original de Ephraim e Malah.

2. Outra modificação é a presença do terceiro termo na equação do parâmetro R_{prio} ao longo de sucessivos quadros, a fim de promover e eliminação do ruído residual. Com o intuito de aumentar ainda mais este efeito de suavização, este algoritmo introduz o terceiro termo na Equação (2.15) levando em conta mais de um quadro anterior na derivação da expressão de R_{prio}. Este aumento na suavização foi possível através do parâmetro $\nu = 0.75$, que, além de garantir maior influência do quadro $(q-1)$ na determinação de R_{prio}, também permite considerar para esta determinação o quadro $(q-2)$.

Capítulo 3

Algoritmo de Extração de Parâmetros do Pré-processamento Avançado WI008

3.1 Motivação

De posse de todos os problemas fisiológicos, acústicos, influências do canal de comunicação e codificação citados no Capítulo 1, entende-se que robustez é um assunto essencial no preparo de sistemas práticos com tecnologia de RAF. Como por exemplo, os dispositivos portáteis (telefones celulares) onde os diferentes ambientes ou canais presentes, interferem na qualidade da fala e reduzem o desempenho do sistema de reconhecimento.

Visando as aplicações de RAF em telefonia móvel celular, foi criado um novo conceito, denominado Reconhecimento de Fala Distribuído (DSR - *Distributed Speech Recognition*). Neste novo paradigma, o pré-processamento (extração de parâmetros) do sistema RAF fica por conta do terminal móvel. Os parâmetros são então transmitidos para uma central de processamento, onde fica o *back-end* (reconhecimento de padrões) do sistema de RAF.

Desta forma, consegue-se aliviar a carga de processamento do sistema central (por exemplo, um portal de voz ou *Call Center*) e ao mesmo tempo, minimiza-se o problema da degradação do sinal de fala devido ao canal de transmissão.

Como o intuito de padronizar um sistema europeu de DSR, foi criado o grupo de pesquisa e desenvolvimento STQ Aurora (*Speech processing, Transmission and Quality aspects: Working Group Aurora*), cuja missão era a de promover

uma concorrência pública com finalidade de escolher o melhor algoritmo de pré-processamento para o sistema DSR europeu [59].

O algoritmo vencedor * foi selecionado em fevereiro de 2002 e formou-se um padrão do Instituto Europeu de Normas de Telecomunicações (ETSI) para o sistema DSR europeu. Este algoritmo ficou conhecido como *Advanced DSR Front-end*, sob a sigla WI008.

Será descrito agora, a parte de redução do ruído deste algoritmo.

3.2 Visão Geral

No modelo DSR, os parâmetros da fala são calculados e compactados no terminal (móvel ou fixo) do usuário e então transmitidos sobre a rede até o servidor. No servidor, os parâmetros são descompactados e o processo de reconhecimento é realizado. Os principais componentes do pré-processamento estão apresentados na Figura 3.1. O processo é dividido em duas partes: o lado do terminal (usuário) e o lado do servidor. Como já foi falado, o cálculo mais pesado do pré-processamento é feito no terminal. Neste ponto, os parâmetros cepstrais melhorados (ou "de-noised") são calculados no bloco Extração de Parâmetros (em destaque cinza da Figura 3.1). Os parâmetros cepstrais então são compactados no bloco Compressão de Parâmetros e processados no último bloco, denominado de *Framing*, *Bit-stream*, *Formatting and Error Protection*.

Do lado do servidor, os parâmetros recebidos são decodificados nos blocos *Bit-stream Decoding* e *Error Mitigation*, em seguida, descompactados no Descompactador de Parâmetros. O bloco *Server Feature Processing* desempenha um estágio de processamento de parâmetros num nível computacional relativamente baixo, consistindo principalmente de cálculos de derivadas primeira e segunda dos coeficientes cepstrais. Finalmente, os parâmetros entram no bloco *Back-end*, onde é realizado o final do processo de reconhecimento de fala.

Há duas considerações importantes a serem feitas neste capítulo. Nas seções seguintes, será descrito com mais detalhes a parte dos componentes do pré-processamento (em destaque na Figura 3.1) que aumentam a robustez do sistema de RAF, apresentado nos parágrafos anteriores. Os processos relativos à compressão e codificação e aqueles realizados no servidor (decodificação, descompactação, processamento de parâmetros e *back-end*), não serão abordados nesta dissertação, pois está fora do escopo proposto.

*O algoritmo vencedor da concorrência foi obtido de uma colaboração entre as empresas Motorola, France Telecom e Alcatel, fruto do conhecimento acumulado das três empresas nas áreas de redução de ruído e realce de fala [57].

A segunda consideração é relativo ao uso do WI008 nos testes presentes no Capítulo 6. Pelo fato deste algoritmo passar todo o sinal de fala para o domínio cepstral no bloco Cálculo Cepstral (ainda no terminal do usuário), o ponto de captura de amostras de fala para as análises propostas por este trabalho, é a saída do bloco Processamento da Forma de Onda. O motivo é a necessidade de se coletar amostras do sinal de fala ainda no domínio do tempo para o processo de síntese e comparação da qualidade do realce do sinal de fala ruidoso utilizando a ferramenta de avaliação perceptual da qualidade da fala, conhecida como PESQ. O Capítulo 4 oferece mais detalhes sobre essa ferramenta.

3.3 Pré-processamento Robusto ao Ruído

Nesta dissertação, considera-se apenas a versão de amostragem de 8 kHz do algoritmo *Advanced Front-end WI008*. As extensões para 11 e 16 kHz estão descritas em detalhes em [60]. No algoritmo WI008, os parâmetros cepstrais com redução de ruído são calculados a partir da entrada de um sinal digital proveniente do ADC (*Analog to Digital Converter*) interno do celular.

No bloco Redução de Ruído, foi utilizado um esquema de redução de ruído composto por um filtro de Wiener "deformado" [†] com dois estágios, e cada estágio é uma combinação de um filtro de Wiener de dois passos [56] e de uma redução de ruído no domínio do tempo, conforme descrito em [63]. Na saída deste bloco, as amostras são coletadas para uma análise da qualidade, através da comparação entre o sinal de fala original e o sinal de fala com ruído reduzido. Mais detalhes serão apresentados nos Capítulos 4 e 6.

Depois da redução do ruído (ver a Figura 3.1), o Processamento da Forma de Onda do tipo "SNR-dependente" [‡] (*SNR-dependent Waveform Processing* [61]) é aplicado ao sinal "de-noised" (termo dado ao sinal de fala que passou pelo primeiro processo de redução de ruído, i. e., o bloco Redução de Ruído). Neste ponto, as amostras de audio são capturadas para análise de qualidade do sinal de fala realçado.

O sinal de saída do Processamento da Forma de Onda é utilizado para o Cálculo Cepstral. Finalmente, uma Equalização Cega (*Blind Equalization*) [62] é aplicada sob os parâmetros cepstrais.

Do lado do servidor, o sinal proveniente do canal entra no bloco Decodificação,

[†]A expressão "deformado" é uma tentativa de tradução do termo *warped*, que indica que o filtro de Wiener é projetado para trabalhar com um eixo de freqüências na escala Mel.

[‡]A expressão "SNR-dependente" será utilizada a partir deste ponto sem as aspas, em uma tentativa de abreviar o seu significado completo: "dependente da relação sinal-ruído".

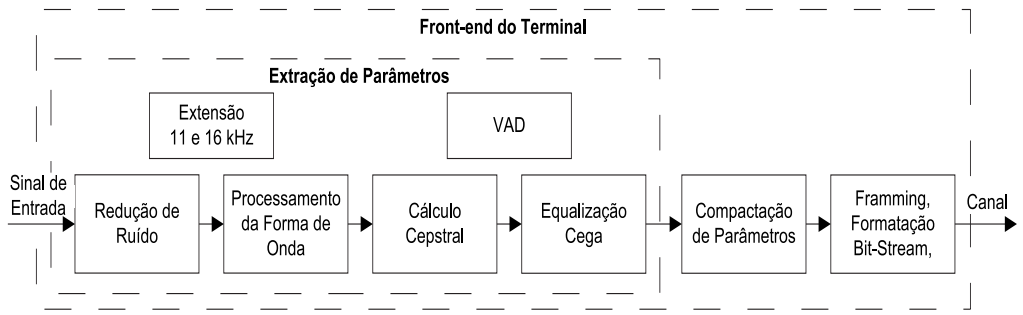

Figura 3.1: *Diagrama de blocos do algoritmo WI008, no lado do terminal.*

Bit-Stream e Mitigação de Erro. Em seguida, é feita a descompactação dos parâmetros e depois, no bloco Processamento dos Parâmetros, os parâmetros dinâmicos (primeira e segunda derivada) são calculados a partir dos parâmetros cepstrais. Também neste lado, o coeficiente energia é calculado e os vetores-parâmetros entram no *Back-end*, conforme ilustrado na Figura 3.2.

As seções seguintes descrevem os blocos denominados "Redução de Ruído" e "Processamento da Forma de Onda do tipo SNR-dependente". Os demais blocos não serão explicados nesta dissertação, pois conforme mencionado, não foram utilizados nos testes descritos no Capítulo 6.

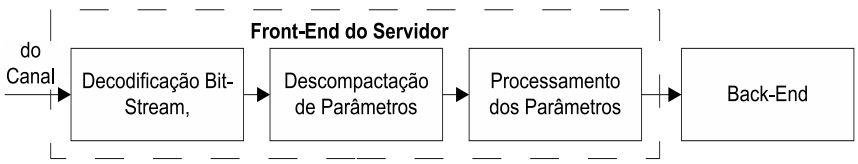

Figura 3.2: *Diagrama de blocos do algoritmo WI008, no lado do servidor.*

3.3.1 Redução do Ruído

A redução do ruído (linha tracejada do diagrama de blocos da Figura 3.3) é realizada por um filtro de Wiener, de dois passos (Figura 3.4). O primeiro e o segundo passo são similares, mas não idênticos.

O processamento é feito quadro-a-quadro, onde o comprimento do quadro é de 25 ms, com 10 ms de deslocamento (sobreposição). O espectro do sinal é calculado usando uma FFT de 256 pontos.

Depois, o espectro de cada quadro é suavizado da seguinte maneira: faz-se a média aritmética da PSD de duas freqüências consecutivas dos 129 primeiros pontos da FFT, obtendo assim, um espectro suavizado com apenas 65 pontos.

Além do espectro suavizado, é calculado em paralelo, um espectro médio, que

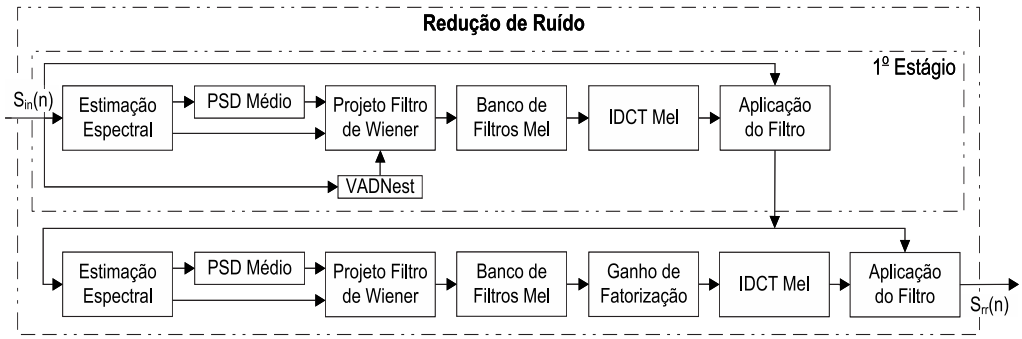

Figura 3.3: *Bloco Redução de Ruído. Cada bloco Projeto F.W. possui duas etapas.*

pe obtido por meio da média aritmética do espectro suavizado de dois quadros consecutivos.

O espectro do quadro atual e a correspondente classificação de quadros em fala ou silêncio é feita no bloco que detecta a atividade da voz, chamado de VADNest (*Voice Activity Detector for Noise estimation*). Estas informações são utilizadas no bloco Projeto Filtro de Wiener, para estimar as características em freqüência do filtro de Wiener.

O *VADNest* é um detector de atividade da voz baseado na energia. O quadro atual é nomeado como *fala* quando a diferença entre o *log* energia do quadro atual e o *log* energia de todo o sinal estimado, excede um limiar definido. Os quadros nomeados de *silêncio* são usados para atualizar a estimação do ruído.

A característica do filtro de Wiener no domínio da freqüência é estimada em dois passos, como mostra a Figura 3.4.

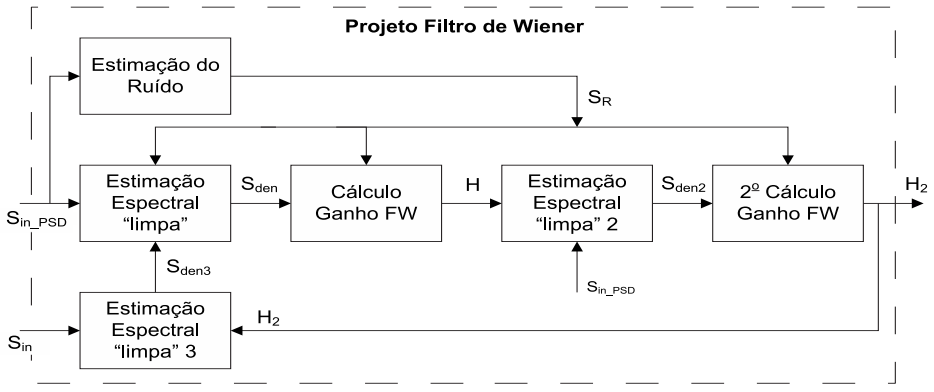

Figura 3.4: *Diagrama de blocos do Projeto do Filtro de Wiener.*

A primeira estimativa do filtro de Wiener é obtida a partir do espectro do sinal com menos ruído (que será tratado como sinal *de-noised*) S_{den} e a estimativa do ruído S_N como:

$$H(f,t) = \frac{\eta(f,t)}{1+\eta(f,t)} \text{ com } \eta(f,t) = \frac{S_{den}(f,t)}{S_N(f,t)} \tag{3.1}$$

onde S_{den} é computado como:

$$S_{den}(f,t) = \beta S_{den3}(f,t-1) + (1-\beta)\max\{S_{in\_PSD}(f,t) - S_N(f,t), 0\} \tag{3.2}$$

onde β=0.98 e o espectro "de-noised" S_{den3} é calculado a partir do quadro anterior como:

$$S_{den3}(f,t-1) = H_2(f,t-1)S_{in}(f,t-1) \tag{3.3}$$

A segunda característica em freqüência do filtro de Wiener é obtida a partir do segundo espectro estimado "de-noised" S_{den2} e do ruído estimado S_N como:

$$H_2(f,t) = \frac{\eta(f,t)}{1+\eta(f,t)} \text{ com } \eta(f,t) = \max\{\frac{S_{den2}(f,t)}{S_N(f,t)}, \eta_{th}\} \tag{3.4}$$

onde a variável η_{th}=0,079432823 corresponde à máxima atenuação de -11,33 dB e S_{den2} é calculado aplicando o primeiro filtro de Wiener no espectro do sinal de entrada, isto é

$$S_{den2}(f,t) = H(f,t)S_{in\_PSD}(f,t) \tag{3.5}$$

No bloco *Mel Filter-Bank*, a característica do filtro de Wiener é suavizada e transformada para a escala "deformada" de freqüências Mel através de 23 janelas triangulares. Tais janelas de freqüência coincidem com aquelas tradicionalmente utilizadas no bloco Cálculo Cepstral. A resposta impulsiva do filtro de Wiener é obtida através da transformada da inversa do co-seno do espectro na escala Mel no bloco MEL IDCT. Esta resposta impulsiva é truncada (comprimento de 17 amostras) e então janelada pela janela Hanning. O sinal "de-noised" é obtido pela convolução do sinal ruidoso de entrada com a resposta impulsiva do filtro de Wiener assim obtida.

Como apresentado na Figura 3.3, o sinal com ruído reduzido do primeiro estágio entra no segundo estágio, onde o segundo filtro de Wiener é projetado e aplicado mais uma vez para reduzir o ruído. A principal diferença entre os dois estágios é o bloco Ganho de Fatorização utilizado no segundo estágio. Neste bloco, uma redução dinâmica de ruído do tipo SNR-dependente é realizada da seguinte forma: uma atenuação mais forte é aplicada naqueles quadros totalmente ruidosos, com o intuito de limpá-los, enquanto que uma atenuação menos

agressiva é aplicada naqueles quadros que contém fala contaminada pelo ruído. Tal coeficiente de Fatorização pode assumir valores entre 0.1 e 0.8, o que pode significar uma atenuação de apenas 10% dos quadros *fala + ruído* e de até 80% dos quadros *ruído puro*. Foi observado que o ganho de fatorização tem um desempenho de maior exatidão no segundo estágio do que no primeiro, devido ao fato de que o sinal como um todo possui uma SNR maior, proveniente da redução do ruído efetuada no primeiro estágio. Outra diferença é que no segundo estágio não se usa o VADNest, e assim, o espectro do ruído é atualizado a cada quadro e não somente quando há quadros sem sinal de fala.

Utilizando estes dois estágios, foi possível ganhar mais flexibilidade no projeto do filtro de Wiener. Note que o sinal de entrada de cada estágio tem uma SNR diferente - no primeiro estágio, a SNR do sinal de entrada pode ser muito baixo, enquanto que no segundo estágio, a SNR do sinal de entrada é maior. Assim, em cada estágio, diferentes decisões são tomadas dependendo da atual SNR (este comportamento não-linear seria difícil de realizar num filtro de Wiener de um único estágio).

3.3.2 Processamento da Forma de Onda SNR-dependente

Em segmentos sonoros do sinal de voz, as formas de onda exibem uma quase-periodicidade, com máximos e mínimos devido à excitação glotal do trato vocal (máximos para a glote fechada e mínimos para a glote aberta). Ao contrário, a energia do ruído interferente de fontes externas pode ser considerada relativamente constante dentro do período de *pitch*. Conseqüentemente, dentro do período de *pitch* da fala ruidosa, a SNR é variável. No Processamento da Forma de Onda do tipo SNR-dependente (SWP) (método que só possui influência positiva se aplicado depois de um processo de redução do ruído), as porções de SNR alta da forma de onda são enfatizadas e as porções onde a SNR é baixa são "desenfatizadas" pela função de janelamento e ponderação. Pode-se observar o funcionamento do algoritmo no diagrama de blocos da Figura 3.5.

As porções de SNR alta são detectadas como máximos do contorno da energia suavizada computada a partir da forma de onda. O contorno da energia é suavizado através de um filtro FIR de média movente e o janelamento é retangular. A Figura 3.6 a seguir ilustra as diversas etapas deste processo.

A SNR tem uma inclinação reduzida a partir de um máximo para outro (ou de um pico para outro), ou seja, os primeiros 80% do intervalo entre os dois máximos são enfatizados e os últimos 20% são "de-enfatizados" pela função de

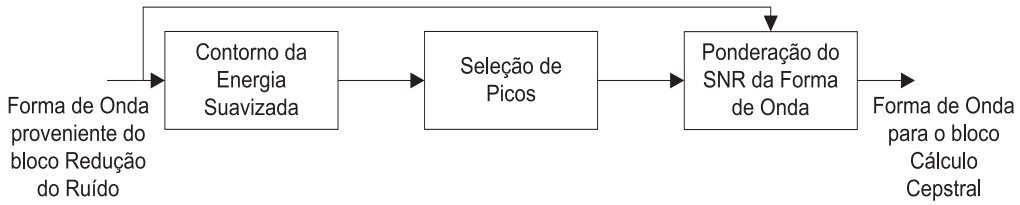

Figura 3.5: *Principais componentes do Processamento da Forma de Onda do tipo SNR-dependente.*

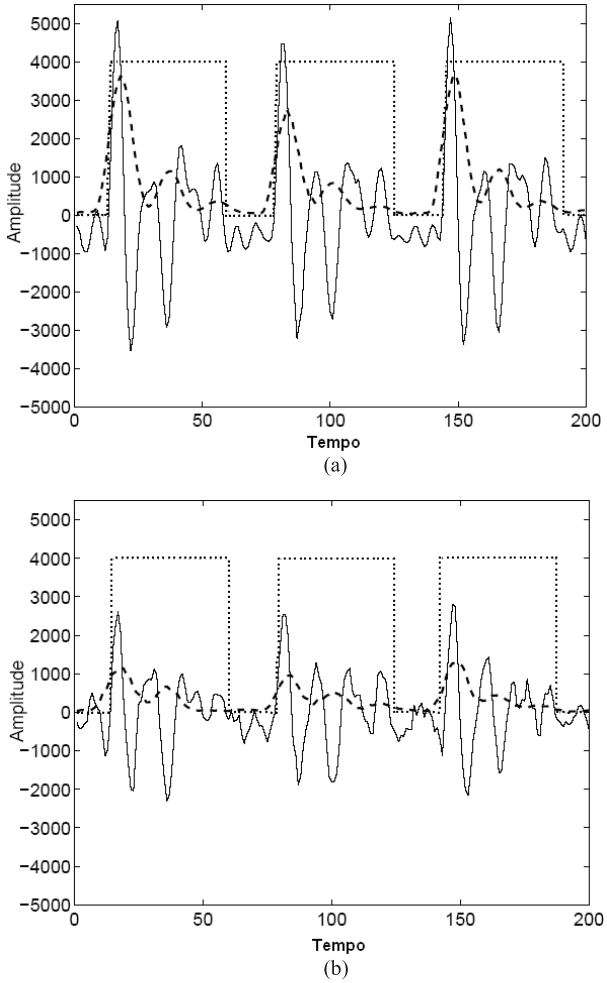

Figura 3.6: *Seleção de Picos: a linha contínua em (a) mostra uma típica forma de onda "limpa" dentro de um quadro sonoro. Também pode-se observar o contorno da energia suavizada (linha tracejada) e a correspondente função de janelamento retangular (linha pontilhada). Em (b), tem-se uma versão de baixa SNR (SNR = 0 dB) do mesmo quadro de fala da parte (a). Em ambos os casos, o processamento de redução de ruído (Filtro Wiener de dois estágios) foi aplicado.*

ponderação. Esta operação, de fato, aumenta a SNR dos quadros sonoros e realça a periodicidade do sinal. Mais detalhes do SWP pode ser encontrado no artigo [60].

3.3.3 Cálculo Cepstral - Algoritmo WI007

O bloco Cálculo Cepstral presente no algoritmo WI008 é denominado de Algoritmo de Extração de Parâmetros do Pré-processamento, WI007, que também foi desenvolvido pelo grupo STQ Aurora e padronizado pelo ETSI em abril de 2000 [58].

Este algoritmo foi adotado neste trabalho como o ponto de partida das análises apresentadas no Capítulo 6, pois ele não aplica nenhum método de realce da fala ruidosa. Apenas realiza a extração dos parâmetros *mel-cepstrais*. O seu diagrama de blocos está mostrado na Figura 3.7.

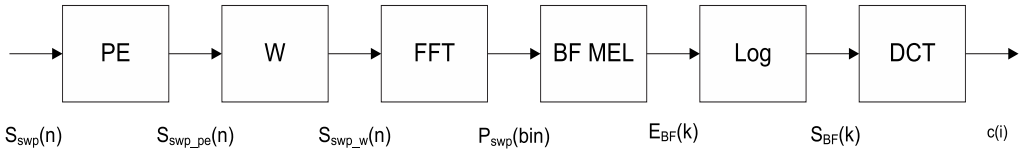

Figura 3.7: *Principais componentes do bloco Cálculo Cepstral.*

As amostras do sinal de entrada passam por uma pré-ênfase, são multiplicadas pela janela Hamming e depois a Transformada Rápida de Fourier é calculada. A filtragem Mel tem como idéia básica combinar a informação contida na freqüência (FFT) com as bandas de freqüências seletivas do ouvido humano. Seus principais parâmetros são: 23 bandas de freqüência, $f_{inicial} = 64$Hz, $f_{final} = 4$ kHz e $f_s = 8$kHz (taxa de amostragem). A função *log* natural é aplicada na saída do bloco Banco de Filtros Mel e por último 13 coeficientes cepstrais são calculados através da Transformada Discreta do Cosseno, $(c(1), c(2), \ldots, (12) + En)$.

Para o algoritmo WI008, a pré-ênfase é aplicada na saída do bloco Processamento da Forma de Onda do tipo SNR-dependente. Porém, foi constatado que há uma melhora nos resultados ao modificar o coeficiente de pré-ênfase original, isto é, passando $p = 0{,}97$ para $p = 0{,}9$. O motivo é a tentativa de aproveitar a pré-ênfase natural do banco de filtros Mel, pois a sua saída não é energeticamente normalizada. Também, uma maior robustez ao ruído é observada quando estima-se o espectro de potência ao invés do espectro de magnitude antes de se aplicar o banco de filtros [61].

Capítulo 4

Avaliação Perceptual da Qualidade da Voz - PESQ

4.1 Introdução

Neste capítulo será apresentada uma ferramenta de análise objetiva criada para avaliar a qualidade perceptual da fala, o PESQ (*Perceptual Evaluation of Speech Quality*), bem como as motivações para o seu desenvolvimento e as razões pelas quais esta ferramenta é usada ao invés do seu antecessor PSQM (*Perceptual Speech Quality Measure*).

Baseado na pontuação MOS (*Mean Opinion Score*), o PESQ foi selecionado pelo ITU (*Internacional Telecommunication Union*) numa competição que procurou determinar o melhor modelo de avaliação perceptual da qualidade da fala e abranger as exigências das operadoras de telefonia mundiais, isto é, ser capaz de suportar diversas condições e aplicações. Como por exemplo, avaliar os codificadores de voz, testar as redes de telecomunicações fim-a-fim e avaliar a qualidade da fala melhorada pelos algoritmos que realçam os segmentos de fala quando contaminados pelo ruído.

Para uma noção básica do seu funcionamento, um diagrama de blocos é apresentado na Figura 4.1 e descrito sucintamente na Tabela 4.1.

Este modelo foi aprovado pelo ITU-T (Setor de Padronização das Telecomunicações do ITU) na Recomendação P.862 de fevereiro de 2001 [66]. A pontuação PESQ, resultante da comparação entre o sinal original e o degradado (este cenário de comparação é denominado de *sistema intrusivo*), apresenta uma correlação de mais de 93% [66] com a medida subjetiva MOS, com a vantagem de ser bem mais prática de se obter, pois os métodos subjetivos demandam tempo e pessoas.

Devido a esta praticidade, os trabalhos mais recentes têm utilizado esta me-

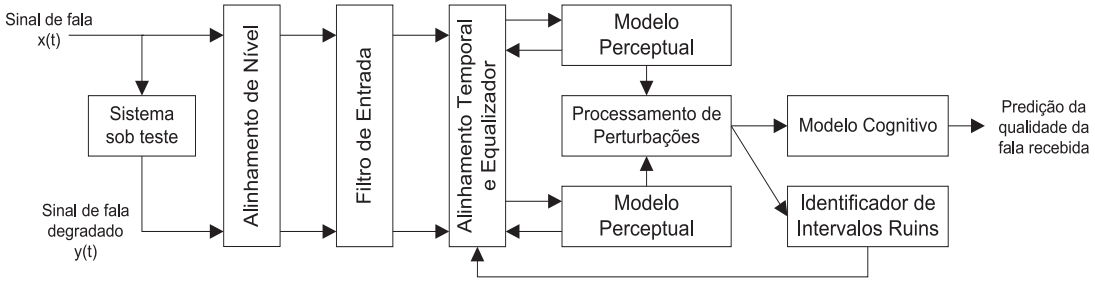

Figura 4.1: *Diagrama de blocos simplificado do algoritmo PESQ.*

Tabela 4.1: *Apresentação dos blocos funcionais do algoritmo PESQ.*

Processo	**Descrição sucinta**
Alinhamento de Nível	Para permitir a comparação do sinal, a fala de referência e o sinal degradado são ajustados a fim de apresentar um mesmo nível de potência.
Filtro de Entrada	Faz uma compensação no sinal devido às filtragens inerentes aos sistemas de telefonia.
Alinhamento Temporal	O sistema sob teste pode conter atrasos variáveis. O PESQ utiliza uma poderosa técnica para identificar e compensar mudanças de atraso perceptualmente irrelevantes.
Modelamento Perceptual	Os sinais de referência e o degradado passam através por um processamento temporal e espectral que simula a percepção subjetiva do sistema de audição humano.
Processamento de Perturbações Modelamento Cognitivo	Trata-se de uma análise das perturbações do sinal distorcido que são perceptualmente audíveis para o ouvinte. Estas são processadas através de um modelo cognitivo que atribui uma pontuação para a qualidade da fala baseada no MOS (-0.5 = pior caso e 4.5 = melhor caso).

dida por ser bastante confiável nas avaliações quando o quesito é qualidade da fala produzidas pelos sistemas; especialmente daqueles que implementam técnicas que reduzem o ruído de sinais de fala com baixa SNR (< 10 dB) [70] e [67]. Assim, através do PESQ é possível avaliar o desempenho destes sistemas através da medida da qualidade perceptual da fala processada. Esta utilização da ferramenta PESQ é um dos principais assuntos propostos por este trabalho, com resultados

apresentados no Capítulo 6.

No entanto, para classificar um sistema é importante definir primeiro o que é qualidade. Por analogia, considera-se um ramo de transporte. Não o de telecomunicações, mas o de viagens aéreas. O mesmo "transporte" oferece economia, negócios, e serviços de primeira classe. Anos de análise da satisfação dos clientes, via pesquisas de opinião pública, deram às empresas de linhas aéreas informações suficientes (quanto à percepção dos passageiros) para determinar o que realmente influencia a qualidade do serviço prestado. Então qualidade não é somente técnica, qualidade é a percepção dos consumidores quanto a um serviço ou produto.

4.2 Qualidade da Fala

Já a medida subjetiva da qualidade da fala, conhecida como VQM (*Voice Quality Measurement*), avalia a opinião média dos usuários de telecomunicações. O método mais exato de obter uma medida da qualidade da voz seria realmente perguntar diretamente aos usuários de telefonia. Assim, num caso ideal os usuários seriam continuamente interrompidos durante uma ligação para responder o que acham da qualidade da conexão. Contudo, por razões óbvias, esta não é uma solução prática, então métodos de testes funcionais e objetivos para avaliar a qualidade da fala, como o PSQM e o PESQ, foram criados.

Estes métodos avaliam a qualidade da fala através da escala de pontos de 1 a 5 (ver Anexo A trata da pontuação específica do PESQ), freqüentemente utilizada pelas aplicações do ITU-T e dependendo do idioma, deve-se usar uma pronúncia equivalente ao inglês, pois este fato pode resultar em pequenas variações se comparado ao texto original utilizado nos testes subjetivos [71]. Na Tabela 6.13, tem-se a pontuação para a medida da qualidade da fala.

Tabela 4.2: *Escala da opinião da medida de qualidade da fala utilizada no desenvolvimento do PESQ.*

Qualidade da Fala	**Pontuação**
Excelente	5
Bom	4
Satisfatório	3
Pobre	2
Ruim	1

A qualidade avaliada a partir desta pontuação (pontuação média da qualidade da fala ou simplesmente, pontuação média da opinião) é representada pela sigla

MOS.

Na seção seguinte, serão apresentados os fatores que afetam a qualidade da fala.

4.3 Fatores que Afetam a Qualidade da Fala

Há muitos fatores, comuns para todas as tecnologias de telecomunicações, que tornam o diálogo de uma ligação telefônica "pobre" em qualidade. Por exemplo, tem-se ruído sonoro do ambiente, eco, atraso, saturação, erros e desacordos entre *codecs* (codificadores e decodificadores) de voz. Na tecnologia voz sobre o protocolo IP (*VoIP*), há fatores prejudiciais específicos causados pelo roteamento de pacotes na rede, como jitter, latência e perda de pacotes, enquanto que para as comunicações móveis, tem-se erros de bit, atraso de grupo, desvanecimento por multi-percurso e vários esquemas de compressão em cascata. Já para os sistemas de reconhecimento de fala, tem-se o ruído convolucional e o aditivo, distorções e limitações do microfone bem como as distorções introduzidas pelo próprio locutor, como o efeito de Lombard, alta aceleração gravitacional, resfriados, sussurro e rouquidão.

Dependendo da aplicação, pode haver várias combinações em cascata dos problemas citados. Por exemplo, considere uma ligação internacional de longa distância para uma central de auto-atendimento ao cliente com recursos de reconhecimento de fala, originada de um telefone móvel dentro do carro, cuja rede de transporte entre origem e destino é VoIP, e o cidadão está rouco.

4.4 Ferramenta de Medição da Qualidade da Fala

Neste capítulo, além da explicação detalhada do algoritmo PESQ no Anexo A, será apresentado resumidamente na seção 4.4.1, o seu antecessor: o PSQM (*Perceptual Speech Quality Measurement*). O objetivo dessa abordagem é apontar as falhas do primeiro algoritmo de medição da qualidade da fala (PSQM) que motivaram o desenvolvimento de um novo algoritmo chamado PESQ.

4.4.1 Perceptual Speech Quality Measurement - PSQM

O primeiro padrão internacional para a medição da qualidade perceptual da fala foi validado apenas para a avaliação de codificadores de voz, na largura de banda do telefone (300 - 3400 Hz) utilizada nas redes de transporte do sistema telefônico. Em fevereiro de 1996, o ITU-T lançou a Recomendação P.861 [74]. Este método

mostrou a maior correlação entre as medidas objetivas e subjetivas quando comparado com outros quatro métodos [73].

Contudo, o propósito desta recomendação foi bastante limitado, tornando-se obsoleto devido aos seguintes motivos: além de avaliar somente o desempenho dos codificadores da fala na faixa de telefonia, do ponto de vista teórico, o PSQM possui baixa correlação entre as qualidades objetiva e subjetiva ao avaliar as modernas distorções das redes que transportam a voz. Como por exemplo, a perda de pacotes e o atraso variável devido ao roteamento de pacotes da rede VoIP. O método PSQM não havia sido projetado para tais fins (não possui alinhamento temporal entre o sinal de fala original e o degradado) e conseqüentemente ocorre uma errônea avaliação da qualidade perceptual da fala.

E por fim, há distorções cujo distúrbio perceptual é subestimado pelo PSQM, como por exemplo, distorções de curta duração e alto volume. Já as distorções lineares, causadas pelas filtragens da rede sob teste, são super-estimadas [74].

4.4.2 Inovações do PESQ

Em busca da solução dos problemas citados na seção anterior, pesquisadores da Psytechnics Limited, Royal PTT Netherland e Philips Research [66] criaram um algoritmo com desempenho superior, mas ainda fundamentado no PSQM, onde as duas principais inovações são: o método que identifica e compensa o atraso temporal variável entre o sinal de referência e o degradado, e o modelo cognitivo aprimorado.

O algoritmo de compensação do atraso baseado na percepção é descrito com detalhes em [68] e será resumido na seção A.13 do Anexo A.

Quanto ao modelo cognitivo, embora a idéia básica do PESQ seja a mesma do PSQM (ambos os sinais, original e degradado, são mapeados e representados através de um modelo perceptual), há uma diferença entre cada representação: o modelo cognitivo do PESQ é superior, permitindo uma predição da qualidade da fala percebida sob o sinal degradado através de um melhor tratamento dos efeitos cognitivos avaliados (os parâmetros psico-acústicos utilizados no mapeamento não serão ressaltados nesta dissertação pois este assunto foge do escopo proposto, mas podem ser encontrados em [76]).

Os efeitos cognitivos considerados tanto pelo PSQM quanto pelo PESQ são: as distorções assimétricas e as diferentes ponderações das distorções durante os segmentos de fala e de silêncio. Embora estes efeitos são modelados para permitir a maior correlação entre as pontuações objetiva e subjetiva, para o PSQM não é o suficiente quando utilizado em redes de transporte de voz modernas.

O efeito dos distúrbios assimétricos é causado quando um codificador/decodificador distorce o sinal de entrada. Ao tentar reconstituir este sinal original na outra ponta da comunicação através da introdução de novas componentes espectrais que o integram, tem-se uma tarefa ineficiente pois o ouvido humano é percetualmente mais sensível quanto às componentes extras ou espúrios. Então o sinal resultante ainda será interpretado por duas diferentes percepções: o sinal de entrada e o sinal distorcido, deixando claramente uma distorção audível [77].

Como solução, o codec (ou codificador/decodificador) omiti componentes espectrais através de um fator de multiplicação corretivo assimétrico. Assim o resultado pode não ser decomposto da mesma forma e a distorção é menos censurada pois o ouvido humano é menos sensível com a ausência de algumas componentes presentes no sinal original.

Este fator de multiplicação corretivo é muito simples. Ele é modelado no PSQM utilizando a relação de potência entre o sinal de saída com o de entrada em função do tempo e da freqüência.

No PESQ, este efeito possui uma melhor representação através de uma ponderação complexa dos distúrbios causados pela introdução de novas componentes espectrais. Isto é, embora tais componentes sejam ponderadas por uma assimetria similar ao do PSQM, o PESQ não utiliza um fator simples. Trata-se de uma combinação entre o distúrbio total (proveniente das pequenas diferenças inaudíveis devido ao mascaramento em função do tempo e da freqüência) e o assimétrico ponderado (distúrbio total, multiplicado pelo fator corretivo assimétrico em função do tempo e da freqüência, como aplicado no PSQM e citado no parágrafo anterior) por locução envolvida. Mais detalhes podem ser encontrados no Anexo A.

O segundo efeito cognitivo, descrito em [65], relaciona com os distúrbios que ocorrem durante os períodos de voz ativa, onde há maior capacidade de desfigurar a qualidade do que nos intervalos de silêncio. No PSQM eles são modelados por um fator de ponderação que pode ser ajustado de acordo com o contexto dos testes. Entretanto, no caso do PESQ, o ITU-T não permitiu nenhum ajuste, mas um procedimento de diferentes ponderações no tempo, visando o desempenho ótimo nos diversos ensaios, foi permitido. Observe a seguinte ponderação L_p na equação 4.1 aplicada ao longo da duração da locução:

$$L_p = \left(\frac{1}{N} \sum_{n=1}^{N} \text{distúrbio}[n]^p \right)^{\frac{1}{p}} \tag{4.1}$$

onde N é o comprimento total de quadros e $p > 1{,}0$. Esta ponderação L_p

enfatiza os distúrbios sonoros quando comparada com a média (no tempo) normal L_1, conduzindo para uma correlação melhor entre pontuação objetiva e subjetiva [78], [79].

E por fim, uma outra grande diferença entre o PSQM e o PESQ é a compensação parcial das distorções lineares (filtragem) comumente encontrada nos sistemas sob testes. Sabe-se que, as distorções lineares são menos perceptíveis do que as não-lineares. Conseqüentemente, as mínimas diferenças lineares entre o sinal original e o degradado são compensadas. Já os efeitos mais ásperos ou variações rápidas são apenas parcialmente compensados, mantendo assim um mínimo de resíduo que contribui para a percepção geral do distúrbio. Maiores detalhes podem ser encontrados no Anexo A.

A compensação parcial da resposta em freqüência citada no parágrafo anterior também cria um impacto sobre as diferentes compensações parciais do ganho quando aplicado em sucessivos quadros. Esta compensação do ganho é uma parte essencial para qualquer medida objetiva da qualidade da fala de sistemas, pois quanto mais se reduz ou se minimiza as variações do ganho, menor será o choque quanto à qualidade da fala percebida (i. e., variações rápidas ou intensas podem causar o efeito contrário). Este é um dos principais problemas no desenvolvimento de medições da qualidade da fala percebida dos sistemas: a forma de como a variação do ganho é tratada e como ela se une com o efeito assimétrico [74]. O detalhamento de todos os passos que compõem o algoritmo PESQ estão descritos no Anexo A.

Capítulo 5

Material Utilizado nos Experimentos

Este capítulo descreve a base de dados de fala projetada para avaliar o desempenho de algoritmos de reconhecimento de fala em condições ruidosas e o software que compõe o próprio sistema de reconhecimento.

5.1 Base de Dados

5.1.1 Origem

Conhecida como Aurora 1, a base de dados de fala utilizada é derivada da base TIDigits, especificamente da parte que contém gravações de vozes adultas masculinas e femininas, pronunciando dígitos conectados em seqüencia de até 7 números [32].

Cada locução, originalmente gravada a uma freqüência de amostragem de 20 kHz, é subamostrada para 8 kHz através de um filtro passa-baixas, com banda passante entre 0 e 4 kHz. A quantização é linear com 16 bits e canal mono.

5.1.2 Aplicando Filtros

Para considerar as condições reais dos terminais e equipamentos de telecomunicações, dois tipos de filtros são aplicados em todas as locuções. Conhecidos como filtro G.712 e filtro MIRS, as suas respostas em freqüência correspondem ao padrão definido pelo ITU [33]. Observe a Figura 5.1 que apresenta a resposta em freqüência de cada filtro.

A maior diferença entre ambos é a resposta plana do G.712 entre 300 e 3400 Hz, contra a atenuação em baixas freqüências presente no MIRS. Este último

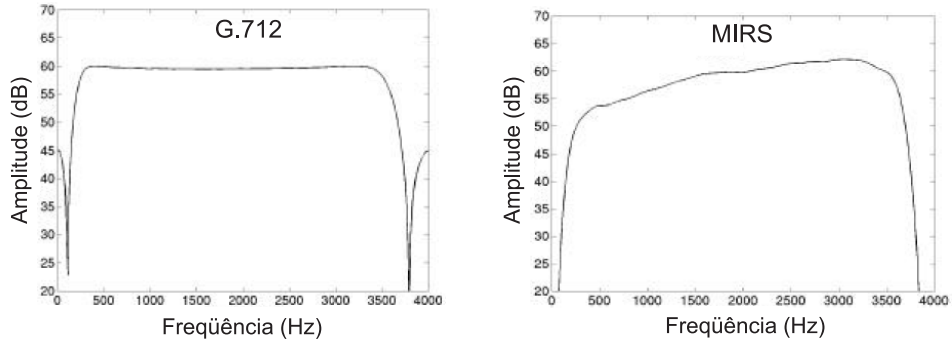

Figura 5.1: *Respostas em freqüência dos filtros padronizados pelo ITU.*

simula o comportamento de alguns equipamentos presentes na especificação GSM 03.50 [34].

5.1.3 Adicionando os Ruídos

Os ruídos são artificialmente adicionados à base TIDigits filtrada. Para determinar a relação sinal-ruído desejada (SNR), deve-se primeiramente definir o que é "relação sinal-ruído", SNR, pois ela depende da escala de freqüência selecionada.

Desta forma, a SNR é definido como a razão entre a energia do ruído e do sinal de fala após a filtragem G.712, pois este filtro garante a manutenção do espectro original de ambos sinais (ver o espectro plano na Figura 5.1). Para determinar a energia da fala, utiliza-se um software baseado na recomendação P.56 do ITU [35]. A energia do ruído é calculada a partir do valor RMS (*Root Mean Square*) através do mesmo software, onde um segmento de ruído de mesmo comprimento do sinal de fala é aleatoriamente selecionado da sua gravação total.

Os sinais ruidosos são selecionados e gravados em diversos locais reais que abrangem a maioria das situações onde os terminais (fixos ou móveis) de telecomunicações se encontram. São eles:

- Metrô (*Suburban train*)
- Multidão de gente (*Babble*)
- Carro (*Car*)
- Salão de exposições (*Exhibition hall*)
- Restaurante (*Restaurant*)
- Rua (*Street*)

- Aeroporto (*Airport*)
- Estação de trem (*Train station*)

Cada ruído é adicionado artificialmente à base TIDigits a SNRs de 20dB, 15dB, 10dB, 5dB, 0dB e -5dB.

5.1.4 Cenários de Treinamento e Teste

Esta base possui dois modos de treinamento para os sistemas de RAF, definidos como:

1. Treinamento com dados limpos (condições limpas)
2. Treinamento com dados limpos e ruidosos (múltiplas condições)

A vantagem do treinamento utilizando apenas os dados limpos é o modelamento da fala sem distorções, pois tal modelo é o mais apropriado para representar as informações disponíveis da fala. Porém, sistemas treinados com estes dados só obtêm um bom desempenho se testados apenas com dados limpos, pois este modelo não possui nenhuma informação quanto às possíveis distorções causadas pelo ruído.

Ao contrário, os sistemas treinados com múltiplas condições, podem apresentar ótimos resultados quando testados com dados ruidosos, pois os seus modelos detêm informações dos tipos de ruído. O maior desempenho é obtido quando os modelos são treinados e testados nas mesmas condições ruidosas.

No modo *treinamento com dados limpos*, 8.440 sentenças são selecionadas da parte que corresponde ao treinamento da base TIDigits, contendo gravações de 55 adultos masculinos e 55 femininos. Estes sinais são filtrados com o filtro G.712 sem adicionar nenhum ruído.

As mesmas 8.440 sentenças são selecionados para o segundo modo de treinamento, *treinamento com dados limpos e ruidosos*. Elas são igualmente divididas dentro de 20 diretórios, isto é, 422 locuções para cada diretório. Os 20 diretórios representam 4 diferentes cenários de ruído a 5 diferentes SNRs. Em outras palavras, adicionando 4 ruídos (*Suburban*, *Babble*, *Car* e *Exhibition hall*) a SNRs de 20, 15, 10 e 5 dB mais a condição do sinal de fala sem ruído (condição *clean*), com 422 locuções para cada, totaliza-se 8.440 sentenças para o modo *treinamento múltiplas condições*. Tanto o sinal de fala quanto o de ruído são filtrados pelo G.712 antes de serem somados.

Para avaliar os sistemas, 4.004 locuções de 52 locutores masculinos e 52 femininos que compõem a parte de testes da base TIDigits, são selecionadas para

montar três diferentes conjuntos de testes, denominados TESTE-A, TESTE-B e TESTE-C. As locuções de teste são diferentes das de treinamento, pois o sistema é independente de locutor.

Para facilitar o entendimento da montagem de cada conjunto, primeiramente é preciso definir o que é um *subconjunto*: são 1.001 locuções (4.004 dividido em quatro partes) com um sinal ruidoso adicionado a SNRs de 20, 15, 10, 5, 0 e -5dB, além da sétima condição, *clean*. Cada conjunto de teste é composto por 4 *subconjuntos*, onde cada um possui 1.001 locuções vezes 7 SNRs. Novamente, tanto o sinal de fala quanto o de ruído são filtrados pelo G.712.

Os conjuntos de teste são:

- TESTE-A: testa o sistema considerando os mesmos tipos de ruído utilizados no treinamento. Os ruídos são:
  - *Suburban train*
  - *Babble*
  - *Car*
  - *Exhibition hall*

  Cada ruído é adicionado em cada um dos *subconjuntos*, isto é, 4 (ruídos) vezes 7 (SNRs) vezes 1.001 (locuções) = 28.028 sentenças. Note que esse teste possui as mesmas condições ruidosas do treinamento, e conseqüentemente resultará num alto desempenho da taxa de acerto.

- TESTE-B: testa o sistema com diferentes tipos de ruído daqueles presentes no treinamento. Os ruídos são:
  - *Restaurant*
  - *Street*
  - *Airport*
  - *Train station*

  Mais uma vez, cada ruído é adicionado em cada um dos *subconjuntos*: 4 (ruídos) vezes 7 (SNRs) vezes 1.001 (locuções) = 28.028 sentenças.

- TESTE-C: testa o sistema com 2 tipos de ruído, porém com um outro padrão de filtragem (MIRS) presente nos equipamentos e terminais de telecomunicações da especificação GSM 03.50. Os ruídos são:
  - *Suburban train* (proveniente do TESTE-A)

– *Street* (proveniente do TESTE-B)

Note que para o treinamento, ambos ruídos são filtrados com G.712. Cada ruído é adicionado com SNRs de 20, 15, 10, 5, 0, -5 e a condição *clean*. Desta forma, têm-se 2 (ruídos) vezes 7 (SNRs) vezes 1.001 (locuções) = 14.014 sentenças. Para este cenário, ambos os sinais (fala e ruído) passam pelo filtro MIRS antes de serem somados.

5.2 Sistema de Reconhecimento HTK

O sistema de reconhecimento, do tipo independente de locutor e palavras conectadas (dígitos), é construído utilizando o software HTK (*Hidden Markov Models Toolkit*) [36], versão 3.1. As locuções são processados por modelos de palavra (HMM) com os seguintes parâmetros:

- 16 estados por palavra
- Modelo do tipo *left-to-right* sem saltos entre estados (com exceção dos modelos de pausa, que serão apresentados a seguir)
- Mistura de 3 Gaussianas por estado
- Matriz de variância diagonal para todos os coeficientes acústicos (sem matriz de covariância completa)

O vetor de parâmetros *mel-cepstrais* possui dimensão 39, isto é, 12 coeficientes *mel cepstrais* mais o logaritmo da energia do quadro atual, todos com primeira e segunda derivada. O coeficiente zero não é utilizado.

Contudo, para modelar as pausas presentes tanto no começo e no fim quanto entre os dígitos de cada locução, dois modelos de pausa são utilizados. O primeiro, nomeado de "sil" (silêncios antes e depois da locução) consiste de 3 estados, onde cada estado possui uma mistura de 6 Gaussianas. A estrutura de transição é apresentada na Figura 5.2. O segundo modelo, chamado de "sp" (silêncios entre palavras) consiste de um único estado que apresenta as mesmas características do estado 2 do primeiro modelo (sil).

Quanto ao treinamento, ele é feito em várias etapas aplicando a re-estimação Baum-Welch,

1. Inicializa-se todas os modelos de palavra e o modelo de pausa "sil" com uma média e variância global, e com 1 Gaussiana por estado;

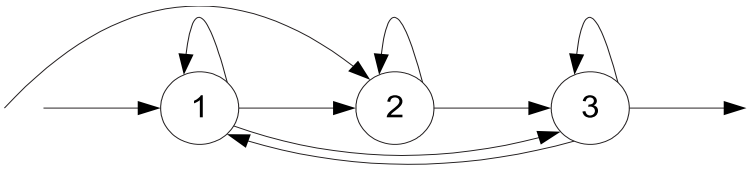

Figura 5.2: *Possível transições do modelo de pausa "sil".*

2. São feitas três iterações de re-estimação Baum-Welch (no modo *pruning* para reduzir a carga computacional);

3. Introduz-se o modelo de pausa entre dígitos, aumenta-se o número de Gaussianas do modelo "sil" para 2 e aplica-se mais três iterações de re-estimação Baum-Welch;

4. Aumenta-se o número de Gaussianas dos modelos de palavra para 3 e dos modelos de pausa para 6, e aplica-se mais sete iterações de re-estimação Baum-Welch.

Na etapa de reconhecimento, utiliza-se o algoritmo de Viterbi [36].

Capítulo 6

Resultados e Discussões

Neste capítulo, os resultados serão apresentados através de 4 vertentes:

1. Avaliações Iniciais, que serão como um ponto de partida dos resultados experimentais. O objetivo é testar o sistema de reconhecimento sem algoritmo de realce da fala no pré-processamento e utilizar os resultados para comparações;

2. Avaliação da Eficiência dos Algoritmos, que servirá para levantar:

   (a) A taxa de reconhecimento de cada algoritmo;

   (b) A qualidade perceptual da fala que os algoritmos alcançaram.

3. Modelamento matemático da Curva PESQ-MOS versus Taxa de Reconhecimento (%), com o objetivo de predizer a taxa de acerto a partir do índice PESQ-MOS, particularizando o cenário de teste;

4. Interpretações finais sobre o comportamento de alguns ruídos da base.

6.1 Avaliações Iniciais

O ponto de partida das análises será o resultado da taxa de acerto do sistema HTK com pré-processamento sem nenhuma técnica capaz de realçar os trechos de fala do sinal ruidoso, como é o caso do algoritmo WI007 da seção 3.3.3. Através dele, apenas os parâmetros *mel-cepstrais* são extraídos do sinal de fala ruidoso ou limpo.

O objetivo desta avaliação é testar o sistema aplicando o TESTE-A, TESTE-B e TESTE-C, tanto no sistema com treinamento em *múltiplas condições* quanto em *condições limpas*.

- Treinamento em Múltiplas Condições:

  1. Aplicando o TESTE-A:

     A Tabela 6.1 apresenta a taxa de reconhecimento quando se aplica o teste com as mesmas condições ruidosas do treinamento (ruídos *Subway*, *Babble*, *Car* e *Exhibition*). Sabe-se que a degradação do desempenho aumenta quando a SNR diminui. Porém, ao analisar a média vertical marginal da tabela, note que as degradações não são significativamente diferentes entre os diversos ruídos na faixa de 0 a 20 dB. Isso significa que o modelo de palavras e pausas (seção 5.2) do sistema contribui bastante para o bom desempenho. Para o TESTE-A, obteve-se uma média global de 87,81%.

  2. Aplicando o TESTE-B:

     Na Tabela 6.2, têm-se os resultados da taxa de reconhecimento ao treinar um sistema com os ruídos *Subway*, *Babble*, *Car* e *Exhibition* e testá-lo com outros, como *Restaurant*, *Street*, *Airport* e *Train-station*. O desempenho não é tão menor se comparado com o TESTE-A. Na média (de 0 a 20 dB), obteve-se 86,27%. Observa-se apenas uma leve piora para os ruídos que não estão no treinamento. Desta forma, parece que os ruídos deste teste apresentam características espectrais similares aos ruídos do treinamento [59].

  3. Aplicando o TESTE-C:

     Para este teste, que corresponde parcialmente a uma avaliação com as mesmas condições ruidosas do treinamento, porém aplicando o filtro MIRS (ao invés do G.712), os resultados estão detalhados na Tabela 6.3. De acordo com a média global da taxa de acerto (de 0 a 20 dB), houve uma leve degradação(83,77%), devido às diferenças entre as respostas em freqüência deste filtro com relação ao filtro G.712, utilizado no treinamento [59].

Tabela 6.1: *TESTE-A - Taxa de acerto em (%) para treinamento com múltiplas condições.*

SNR/dB	Subway	Babble	Car	Exhibition	Média
clean	98.68	98.52	98.39	98.49	98.52
20	97.61	97.73	98.03	97.41	97.69
15	96.47	97.04	97.61	96.67	96.94
10	94.44	95.28	95.74	94.11	94.89
5	88.36	87.55	87.80	87.60	87.82
0	66.90	62.15	53.44	64.36	61.71
-5	26.13	27.18	20.58	24.34	24.55
Média 0 a 20dB	88.75	87.95	86.52	88.03	**87.81**

Tabela 6.2: *TESTE-B - Taxa de acerto em (%) para treinamento com múltiplas condições.*

SNR/dB	Restaurant	Street	Airport	Train-station	Média
clean	98.68	98.52	98.39	98.49	98.52
20	96.87	97.58	97.44	97.01	97.22
15	95.30	96.31	96.12	95.53	95.81
10	91.96	94.35	93.29	92.87	93.11
5	83.54	85.61	86.25	83.52	84.73
0	59.29	61.34	65.11	56.12	60.46
-5	25.51	27.60	29.41	21.07	25.89
Média 0 a 20dB	85.39	87.03	87.64	85.01	**86.27**

Tabela 6.3: *TESTE-C - Taxa de acerto em (%) para treinamento com múltiplas condições.*

SNR/dB	Subway (MIRS)	Street(MIRS)	Média
clean	98.50	98.58	98.54
20	97.30	96.55	96.92
15	96.35	95.53	95.94
10	93.34	92.50	92.92
5	82.41	82.53	82.47
0	46.82	54.44	50.63
-5	18.91	24.24	21.57
Média 0 a 20dB	83.24	84.31	**83.77**

- Treinamento em Condições Limpas:

1. Aplicando o TESTE-A:

   Para o sistema treinado com dados limpos e testado com dados ruidosos deste teste, os resultados estão na Tabela 6.4. Note que o desempenho é bem pior se comparado com o treinamento com múltiplas condições, devido à falta de informações relativas às características do ruído, tanto nos modelos de palavras quanto nos modelos de pausa (*sil* e *sp*). É de se esperar que ruídos que apresentam segmentos fortemente não-estacionários degradem mais o desempenho do sistema. Dentre os ruídos que apresentam segmentos fortemente não-estacionários (*Subway*, *Restaurant*, *Babble*, *Street*, *Airport* e *Train-station*), a pior taxa de reconhecimento é do ruído *Babble*, com 49,88%.

2. Aplicando o TESTE-B:

   A Tabela 6.5 apresenta os resultados. A menor taxa de acerto é do ruído *Restaurant*, com 52,59%, embora o desempenho do *Airport* e *Train-station* não apresentem tantas diferenças, 53,25% e 55,63%, respectivamente. Os motivos são os mesmos apresentados no item anterior (presença de segmentos não estacionários).

3. Aplicando o TESTE-C:

   Este teste apresenta um resultado inesperado. Note na Tabela 6.6 a melhora da taxa de reconhecimento para o ruído *Street* quando utiliza-se o filtro MIRS ao invés do G.712 no treinamento limpo, que parece ser devido à atenuação das componentes de baixa freqüência, onde se encontra a maior parte da energia do ruído *Street* [59].

Tabela 6.4: *TESTE-A - Taxa de acerto em (%) para treinamento em condições limpas.*

SNR/dB	Subway	Babble	Car	Exhibition	Média
clean	98.93	99.00	98.96	99.20	99.02
20	97.05	90.15	97.41	96.39	95.25
15	93.49	73.76	90.04	92.04	87.33
10	78.72	49.43	67.01	75.66	67.70
5	52.16	26.81	34.09	44.83	39.47
0	26.01	9.28	14.46	18.05	16.95
-5	11.18	1.57	9.39	9.60	7.93
Média 0 a 20dB	69.48	49.88	60.60	65.39	**61.34**

Tabela 6.5: *TESTE-B - Taxa de acerto em (%) para treinamento em condições limpas.*

SNR/dB	Restaurant	Street	Airport	Train-station	Média
clean	98.93	99.00	98.96	99.20	99.02
20	89.99	95.74	90.64	94.72	92.77
15	76.24	88.45	77.01	83.65	81.33
10	54.77	67.11	53.86	60.29	59.00
5	31.01	38.45	30.33	27.92	31.92
0	10.96	17.84	14.41	11.57	13.69
-5	3.47	10.46	8.23	8.45	7.65
Média 0 a 20dB	52.59	61.51	53.25	55.63	**55.74**

Tabela 6.6: *TESTE-C - Taxa de acerto em (%) para treinamento em condições limpas.*

SNR/dB	Subway (MIRS)	Street(MIRS)	Média
clean	99.14	98.97	99.05
20	93.46	95.13	94.29
15	86.77	88.91	87.84
10	73.90	74.43	74.16
5	51.27	49.21	50.24
0	25.42	22.91	24.16
-5	11.82	11.15	11.48
Média 0 a 20dB	66.16	66.11	**66.14**

Vale ressaltar que todos os resultados apresentados nas tabelas desta seção são iguais aos resultados apresentados em [59], confirmando assim, que o procedimento foi corretamente realizado.

Para as próximas seções e futuras comparações, este cenário com os resultados do algoritmo WI007 que acabou de ser apresentado, será chamado de "Nenhum processo", relativo à ausência de técnicas no pré-processamento que realçam a fala contaminada por ruído.

Os resultados das quatro técnicas de realce da fala (EMSR, NMT-PSS, EMSR + NMT-PSS e WI008), já apresentados em capítulos anteriores, serão apresentados na próxima seção.

6.2 Avaliação da Eficiência dos Algoritmos

O objetivo desta seção é fazer uma comparação da eficiência de dois fatores entre os algoritmos supressores de ruído do sinal de fala:

1. A taxa de reconhecimento;
2. A qualidade perceptual da fala.

6.2.1 Taxa de Reconhecimento

Sabe-se que a principal filosofia dos algoritmos supressores de ruído abordados neste trabalho é realçar o sinal de voz a partir do sinal ruidoso. Porém, em condições críticas de SNR (entre -5 e 10 dB), a taxa de reconhecimento de alguns algoritmos tende a cair por causa da introdução de resíduos no sinal de fala "melhorado", que desfavorece ainda mais o desempenho.

Assim, através da Taxa de Reconhecimento (%) alcançada pelo sistema de RAF usando os algoritmos listados abaixo, é possível determinar dentro de um dado cenário de teste, qual aquele que apresenta o melhor desempenho:

- WI008, com dois ensaios:
  1. **NR-WI008 (*Noise Reduction WI008*):** aproveita apenas o bloco inicial Redução de Ruído, utilizando filtragens para a redução do ruído aditivo (seção 3.3.1);
  2. **NR-WI008 + SWP (*NR-WI008 with SNR-dependent Waveform Processing*):** aproveita o bloco Redução de Ruído + Processamento da Forma de Onda do tipo SNR-dependente, utilizando filtragens de redução do ruído, ponderações e estimações espectrais adaptativas da SNR a cada quadro (seção 3.3.2);
- **EMSR + NMT-PSS:** conceitos psico-acústicos em conjunto com uma técnica original para estimar a SNR a cada quadro e suprimir o ruído através de ponderações;
- **NMT-PSS:** subtração espectral tradicional do espectro ruidoso com conceitos psico-acústicos;
- **EMSR:** uso de uma técnica original para estimar a SNR e suprimir o ruído através de ponderações, com eficiência apenas para SNR $>$ 10 dB.

Primeiramente, cada algoritmo de realce foi utilizado para processar as locuções de ambos os modos de treinamento: múltiplas condições (ruido + sinal limpo) e condição limpa (sinal livre de ruído). Em seguida, para cada modo de treinamento, foi aplicado os TESTE-A, B e C (igualmente com prévia aplicação dos algoritmos de realce sobre as locuções) e os resultados apresentados em tabelas a seguir.

- Treinamento em múltiplas condições:

  1. Aplicando o TESTE-A:

     Na Tabela 6.7 há um comparativo entre os algoritmos utilizando a média da Taxa de Reconhecimento (%) para cada SNR dos ruídos desse teste (*Subway*, *Babble*, *Car* e *Exhibition*).

     Para avaliar cada algoritmo na faixa de SNR crítica, uma outra média é tirada dos resultados médios de -5 a 10 dB. O algoritmo NR-WI008 + SWP obteve o melhor desempenho, com 75,04% de acerto. O motivo parece estar na sua complexa seqüência de métodos, como os filtros de Wiener em cascata, estimação espectral adaptativa da SNR, detector de atividade de voz, que ao somarem os seus esforços, reduzem o ruído aditivo e convolucional. Vale lembrar que o algoritmo NR-WI008 + SWP não deixa ruído musical no sinal de fala melhorado. Ao contrário, o algoritmo NMT-PSS (que deixa o ruído musical no sinal), resultou na pior Taxa de Reconhecimento dentre os algoritmos de realce da fala com média de 72,03% de acerto.

  2. Aplicando o TESTE-B:

     Da mesma forma do teste anterior, a Tabela 6.8 apresenta a média da Taxa de Reconhecimento (%) que cada algoritmo juntamente com o sistema HTK alcançou para os ruídos *Restaurant*, *Street*, *Airport* e *Train-station*. Novamente, o NR-WI008 + SWP obteve o melhor desempenho com média de 72,50% (entre -5 e 10 dB).

     Note que os algoritmos do tipo STSS ficaram praticamente empatados. Parece que neste teste, o ganho na redução do ruído aditivo foi maior que os prejuízos causados pelo ruído musical. Em outras palavras, o ruído musical é obscurecido porque os modelos de palavras e pausas foram testados por um cenário de ruídos de fundo totalmente desconhecidos.

  3. Aplicando o TESTE-C:

     A análise é feita da mesma forma e pode ser vista na Tabela 6.9. No entanto, os ruídos de fundo utilizados são *Subway* e *Street* com filtragem MIRS, ao invés do G.712 utilizado no treinamento. Uma melhora inesperada foi a do algoritmo EMSR, com 69,38% de acerto. Esta melhoria parece vir da resposta em freqüência do filtro MIRS, que atenua as componentes de baixa freqüência onde se concentra boa parte da energia dos ruídos *Subway* e *Street*.

Porém, o mesmo não se observa no algoritmo EMSR + NMT-PSS que segue praticamente o mesmo método de supressão, apresentando uma média um pouco menor, 68,23%. O motivo parece estar na compensação das baixas freqüências realizada pela curva do limiar auditivo absoluto do ouvido humano (Figura 2.5) deste algoritmo.

A pior Taxa (%) foi do algoritmo NR-WI008 + SWP, com 64,77%. Parece que o comportamento da resposta em freqüência do MIRS não contribuiu com os blocos *Redução do Ruído* e *Processamento da Forma de Onda do tipo SNR-dependente* em SNR crítica.

Outro ponto importante é a Taxa (%) do algoritmo NMT-PSS, com 65,15%, que parece confirmar a hipótese que quando os modelos de palavras e pausas do sistema possuem informações sobre os ruídos de fundo, o ruído musical é perceptível e o desempenho sofre degradações. No entanto, como foi visto no TESTE-B anterior, embora o ruído *Street* não faz parte do treinamento e conseqüentemente, causa uma mistura de características espectrais, ainda assim não ofusca a detecção do ruído musical.

Tabela 6.7: *Média da Taxa de Reconhecimento (%) dos ruídos do TESTE-A (Subway, Babble, Car e Exhibition) para cada SNR e para todos os algoritmos de realce com sistema treinado em múltiplas condições. Em destaque, o algoritmo com o melhor desempenho.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	98.52	97.80	98.79	98.70	98.57	98.76
20	97.70	98.16	98.12	98.09	97.96	98.12
15	96.95	97.58	97.62	97.40	97.28	97.34
10	94.89	95.97	95.71	95.38	95.41	95.52
5	87.83	91.29	90.40	89.24	88.70	89.87
0	61.71	73.35	73.95	71.19	69.86	71.79
-5	24.55	37.61	40.09	36.76	34.15	37.81
Média -5 a 10	67.25	74.55	**75.04**	73.14	72.03	73.75

Tabela 6.8: *Média da Taxa de Reconhecimento (%) para todos os algoritmos e ruídos do TESTE-B (Restaurant, Street, Airport e Train-station) e para todos os algoritmos de realce com sistema treinado em múltiplas condições. Em destaque, o algoritmo com o melhor desempenho.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	98.52	97.80	98.79	98.70	98.57	98.76
20	97.23	97.93	98.22	97.87	98.11	98.01
15	95.82	96.96	97.31	96.98	97.20	96.87
10	93.12	94.32	94.76	94.21	94.33	94.24
5	84.73	87.59	88.05	85.89	85.83	85.39
0	60.47	69.00	70.66	64.74	64.05	64.05
-5	25.90	35.15	36.56	29.77	28.34	28.28
Média -5 a 10	66.05	71.52	**72.50**	68.65	68.14	67.99

Tabela 6.9: *Média da Taxa de Reconhecimento (%) para todos os algoritmos e ruídos do TESTE-C (Subway e Street) com filtragem MIRS e para todos os algoritmos de realce com sistema treinado em múltiplas condições. Em destaque, o algoritmo com o melhor desempenho.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	98.54	97.69	98.68	98.50	98.36	98.59
20	96.93	97.92	97.78	97.64	97.34	97.64
15	95.94	96.85	96.68	96.59	96.01	96.58
10	92.92	94.01	93.39	93.91	92.74	93.76
5	82.47	84.67	83.50	86.15	83.70	86.52
0	50.63	57.77	56.31	64.81	59.77	65.80
-5	21.55	25.89	25.90	28.06	24.38	31.46
Média -5 a 10	61.90	65.58	64.77	68.23	65.15	**69.38**

- Treinamento em condições limpas:

  1. Aplicando o TESTE-A:

     Na Tabela 6.10, os valores médios da Taxa de Reconhecimento (%) são apresentados. Note que os próximos 3 testes foram aplicados sob o sistema com treinamento limpo. Portanto, é de se esperar que o desempenho seja pior que os 3 testes anteriores, pois os modelos de palavras e pausas do sistema não possuem informações quanto ao comportamento dos tipos de ruído. Contudo, ainda assim o algoritmo NR-WI008 + SWP obteve o melhor desempenho entre os demais, com 65,59 % de

acerto e o motivo é o mesmo: eficiência dos diversos métodos trabalhando "em equipe", como apresentadas na seção 3.3.

Outro resultado esperado é desempenho ruim do algoritmo NMT-PSS com 33,68% de acerto, devido ao ruído musical remanescente no sinal, que desta vez foi bastante prejudicial e perceptível aos modelos HMM, pois a referência estava limpa.

2. Aplicando o TESTE-B:

   A Tabela 6.11 apresenta os valores médios da Taxa de Reconhecimento (%). Com 64,05%, o algoritmo NR-WI008 + SWP possui o melhor desempenho. O pior é o algoritmo NMT-PSS, com 33,52% de acerto. Note que desta vez, o prejuízo do ruído musical foi maior que o ganho na redução do ruído aditivo. Esta análise é óbvia, pois este teste foi aplicado sob um sistema treinado com locuções livres de ruído.

3. Aplicando o TESTE-C:

   O algoritmo NR-WI008 + SWP ganhou com 57,64%, como mostra a Tabela 6.12. Note que neste TESTE-C houve uma melhora geral para os demais algoritmos, incluindo o Nenhum Processo, se comparados com os 2 testes A e B anteriores para condição limpa. Este ensaio aponta uma forte evidência de que a filtragem MIRS contribui com o desempenho. Todavia, o algoritmo NMT-PSS mais uma vez apresentou a taxa mais baixa devido ao ruído musical, com 35,04% de acerto.

Tabela 6.10: *Média da Taxa de Reconhecimento (%) dos ruídos do TESTE-A (Subway, Babble, Car e Exhibition) para cada SNR e para todos os algoritmos de realce com sistema treinado em condição limpa. Em destaque, o algoritmo com o melhor desempenho.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	99.02	98.48	98.93	98.96	98.98	99.01
20	95.25	96.77	97.76	95.90	92.57	96.71
15	87.33	94.40	96.11	91.14	83.56	93.52
10	67.71	88.74	92.05	80.87	64.92	84.41
5	39.47	75.88	82.10	60.59	39.89	66.54
0	16.95	50.86	60.68	34.05	19.24	39.56
-5	7.94	21.40	27.54	10.85	10.70	14.95
Média -5 a 10	33.02	59.22	**65.59**	46.59	33.68	51.36

No Anexo B, têm-se as tabelas com valores da Taxa (%) de cada algoritmo separadas por ruído e TESTE.

Tabela 6.11: *Média da Taxa de Reconhecimento (%) para todos os algoritmos e ruídos do TESTE-B (Restaurant, Street, Airport e Train-station) e para todos os algoritmos de realce com sistema treinado em condição limpa. Em destaque, o algoritmo com o melhor desempenho.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	99.02	98.48	98.93	98.96	98.98	99.01
20	92.77	96.15	97.68	94.90	93.12	96.01
15	81.34	93.44	95.63	88.65	83.67	90.58
10	59.01	87.23	91.31	76.25	65.79	78.95
5	31.93	73.16	80.53	55.30	38.88	58.12
0	13.70	48.81	57.61	29.54	18.88	32.00
-5	7.65	19.96	26.78	8.72	10.53	9.66
Média -5 a 10	28.07	57.29	**64.05**	42.45	33.52	44.68

Tabela 6.12: *Média da Taxa de Reconhecimento (%) para todos os algoritmos e ruídos do TESTE-C (Subway e Street) com filtragem MIRS e para todos os algoritmos de realce com sistema treinado em condição limpa. Em destaque, o algoritmo com o melhor desempenho.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	99.06	98.40	98.73	99.06	98.84	99.07
20	94.30	95.01	97.08	95.67	92.07	96.62
15	87.84	91.08	94.94	91.32	85.75	93.78
10	74.17	83.17	88.97	82.63	68.86	86.45
5	50.24	67.89	74.89	65.62	42.18	72.30
0	24.17	40.42	45.68	40.56	19.30	48.33
-5	11.49	19.02	21.02	19.60	9.83	23.30
Média -5 a 10	40.01	52.62	**57.64**	52.10	35.04	57.59

6.2.2 Avaliação Perceptual da Qualidade da Fala

Dadas as atuais potencialidades da ferramenta de análise perceptual da qualidade da fala, o PESQ, (discutidas no Capítulo 4), o objetivo desta seção é avaliar a quantidade de distorção que os algoritmos de supressão do ruído introduzem no sinal de fala, com o intuito de melhorar a SNR. Para isto, o cenário de testes constou de:

- Locuções de dígitos conectados da base de dados Aurora.
- SNRs de 20, 15, 10 5, 0 e -5 dB;

- Os tipos de ruído que representam os ambientes reais para a comunicação móvel:
  - Com filtragem G.712:
    - Subway
    - Babble
    - Car
    - Exhibition
    - Restaurant
    - Street
    - Airport
    - Train-station
  - Com filtragem MIRS:
    - Subway
    - Street
- Técnicas de realce do sinal de fala ruidosa:
  - Nenhum Processo;
  - NR-WI008;
  - NR-WI008 + SWP;
  - EMSR;
  - NMT-PSS;
  - EMSR + NMT-PSS;
- Os cenários de testes:
  1. TESTE-A;
  2. TESTE-B;
  3. TESTE-C.

Aplicando o PESQ

O PESQ é uma medida objetiva da similaridade entre 2 sinais de fala (o de referência e o degradado) contendo a mesma sentença . Observe a Figura 6.1.

Para este trabalho, o sinal de referência equivale às locuções limpas (subconjunto *clean*) de cada TESTE (A, B e C). O sinal degradado corresponde às

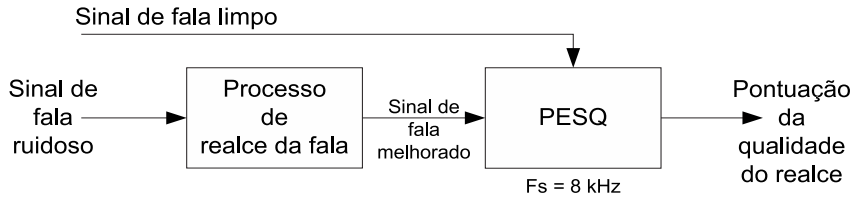

Figura 6.1: *Cálculo da pontuação PESQ entre dois sinais de fala.*

respectivas locuções ruidosas do subconjunto *clean* de cada TESTE, isto é, *clean* somado artificialmente com ruídos à SNRs de 20, 15, 10, 5, 0 e -5 dB, porém processadas pelos algoritmos de supressão (que neste caso, passa a ser o *sinal de fala melhorado* da Figura 6.1).

O exemplo a seguir mostra a linha de comando para realizar uma avaliação PESQ (com freqüência de amostragem de 8000 Hz) entre um sinal limpo e o seu respectivo sinal ruidoso (SNR de 10 dB) selecionado do TESTE-A não processado por nenhum algoritmo:

pesq +8000 MAH_462_CLEAN.wav MAH_462_SNR10.wav

Sinal de fala limpo	Sinal de fala melhorado	Algoritmo	PESQ
MAH_462_CLEAN.wav	MAH_462_SNR10.wav	Nenhum	2.1103

Desta forma, comparando o sinal limpo com o sinal ruidoso (realçado ou não), é possível traduzir através da pontuação PESQ alcançada (entre -0,5 = pior caso e 4,5 = melhor caso ou nenhuma distorção), a qualidade da supressão do ruído que o algoritmo realizou sobre o sinal de fala ruidoso. Como já visto no Capítulo 4, na maioria dos casos, a pontuação PESQ se assemelha com a pontuação MOS da Tabela 6.13, isto é, ficando entre 1,0 e 4,5.

Tabela 6.13: *Escala de pontuação MOS da medida de qualidade da fala.*

Qualidade da Fala	**Pontuação**
Excelente	5
Bom	4
Satisfatório	3
Pobre	2
Ruim	1

Resultados da Avaliação Perceptual

Primeiramente, antes de se apresentar os resultados, deve-se levantar as seguintes considerações para uma boa interpretação:

1. Note que os resultados PESQ apresentados nas Tabelas 6.14, 6.15 e 6.16 não foram cruzados com as Taxas de Reconhecimento (%) do sistema. Assim, por enquanto não faz sentido comparar o PESQ entre sistemas ora treinados em condição limpa ora ou em múltiplas condições. Porém, esta relação será levantada na seção 6.3.

2. As linhas *clean* (SNR) das Tabelas não foram processadas pelos algoritmos. Conseqüentemente, é intuitivo esperar a maior pontuação PESQ ou distorção nula, 4,5, ao comparar a condição *clean* (sinal limpo) com ela mesma para todos os algoritmos. Por esta razão, todas as pontuações PESQ alcançadas nestes testes são ditas "valores absolutos", pois representam qual a similaridade real entre *sinal de fala melhorado* e o *sinal de fala limpo* da Figura 6.1. Caso os valores da linha *clean* tivessem sido realçados antes da análise PESQ, teria-se os "valores relativos" da qualidade percebida, pois o sinal de referência poderia ser modificado pelo realce. Esta comparação não traduziria verdadeiramente a similaridade real entre *sinal de fala melhorado* e o *sinal de fala limpo*, ou seja, não avaliaria a qualidade real do processo de supressão de ruído de cada algoritmo.

3. Como já foi dito na listagem da seção 6.2.2, o algoritmo WI008 passou por dois ensaios, onde no primeiro, as amostras do sinal de fala foram extraídas na saída do bloco Redução do Ruído. Este ensaio foi chamado de "NR-WI008" (*Noise Reduction* WI008). No segundo ensaio, as amostras foram extraídas na saída do bloco subseqüente, Processamento da Forma de Onda do tipo SNR-dependente (SWP). Este ensaio foi denominado de "NR-WI008+SWP" (*Noise Reduction* WI008 + *SNR-dependent Waveform Processing*). Os motivos destes ensaios são:

   (a) Avaliar a eficiência tanto do bloco Redução de Ruído (NR) quanto do NR + SWP, analisando o compromisso entre supressão de ruído e distorção causada no sinal de fala, ou seja, medir o PESQ;

   (b) Após o bloco SWP, há uma passagem do sinal de fala para o domínio cepstral, que se considerada, tornaria impossível a reconstituição do sinal no domínio do tempo e armazenamento em arquivo tipo WAVE, necessário para a avaliação PESQ.

Nas Tabelas 6.14, 6.15 e 6.16 a seguir, têm-se os valores médio da pontuação PESQ para o TESTE-A, B e C.

1. Aplicando o TESTE-A:

   Na Tabela 6.14, note que os melhores algoritmos para a faixa de SNRs analisada (de -5 a 20 dB), foram o NR-WI008, NR-WI008 + SWP e EMSR + NMT-PSS por não introduzirem ruído musical no sinal de fala resultante. O mesmo não acontece com o algoritmo NMT-PSS, que apresenta o resíduo no sinal de fala, degradando a qualidade do seu sinal realçado. Uma ressalva: o algoritmo NR-WI008 apresentou as menores pontuações PESQ nas SNRs de -5 e 0 dB, devido às distorções causadas sob o sinal de fala ao tentar suprimir o ruído.

Tabela 6.14: *Média da pontuação PESQ para todos os algoritmos que processaram os ruídos do TESTE-A (Subway, Babble, Car e Exhibition). Em destaque, a maior pontuação.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7962	**3.1340**	3.0235	3.1027	3.0311	3.0537
15	2.4897	2.8028	2.7946	**2.8312**	2.7475	2.7712
10	2.1803	2.4557	2.5411	**2.5493**	2.4475	2.4755
5	1.8781	2.0784	**2.2335**	2.2292	2.1006	2.1479
0	1.6069	1.6642	1.8637	**1.8646**	1.7066	1.7975
-5	1.3497	1.2689	**1.4773**	1.4659	1.2946	1.4370

2. Aplicando o TESTE-B:

   Na Tabela 6.15, note que o algoritmo que menos causou degradação no sinal de fala (para SNR < 5 dB) ao tentar suprimir o ruído aditivo, foi o WI008-NR + SWP. Ao contrário, o algoritmo NMT-PSS causou forte degradações no sinal realçado por manter o ruído musical em toda a faixa de SNRs. Observe que, o algoritmo NR-WI008 causou mais distorções no sinal nas SNRs críticas -5 e 0 dB do que o ruído musical no algoritmo NMT-PSS para a mesma faixa.

3. Aplicando o TESTE-C:

   Neste teste, os resultados estão na Tabela 6.16, onde o algoritmo EMSR + NMT-PSS apresentou praticamente, o melhor desempenho em todas as SNRs. O mesmo raciocínio da influência do ruído musical e das distorções sob o sinal, vale para este teste: o algoritmo NMT-PSS perde eficiência na

Tabela 6.15: *Média da pontuação PESQ para todos os algoritmos que processaram os ruídos do TESTE-B (Restaurant, Street, Airport e Train-station). Em destaque, a maior pontuação.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.8881	**3.1822**	3.0526	3.1310	3.0557	3.0998
15	2.5888	2.8512	2.8244	**2.8581**	2.7716	2.8189
10	2.2778	2.5021	2.5655	**2.5662**	2.4664	2.5171
5	1.9576	2.1187	**2.2542**	2.2400	2.1189	2.1830
0	1.6424	1.7063	**1.8850**	1.8683	1.7266	1.8241
-5	1.3465	1.2996	**1.4843**	1.4554	1.3065	1.4385

faixa de SNR crítica (agora de -5 a 5 dB) devido ao resíduo remanescente e o algoritmo NR-WI008 devido às suas distorções sob o sinal de fala para a mesma faixa.

Tabela 6.16: *Média da pontuação PESQ para todos os algoritmos que processaram os ruídos do TESTE-C (Subway e Street), porém com filtragem MIRS. Em destaque, a maior pontuação.*

SNR /dB	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT-PSS	NMT-PSS	EMSR
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7203	**3.0206**	2.9008	2.9923	2.8992	2.9539
15	2.4143	2.6759	2.6390	**2.7106**	2.6069	2.6628
10	2.1187	2.3310	2.3690	**2.4150**	2.2916	2.3647
5	1.8575	1.9498	2.0503	**2.1041**	1.9619	2.0579
0	1.6059	1.5582	1.6829	**1.7767**	1.6222	1.7432
-5	1.3651	1.2361	1.3508	**1.4689**	1.3146	1.4346

No Anexo B, têm-se as tabelas com valores PESQ de cada algoritmo separadas por ruído e TESTE.

Curva SNR vs PESQ

Uma outra análise realizada para cada TESTE, foi sobre a relação SNR vs PESQ, que parece ser aproximadamente linear. Observe os gráficos das Figuras 6.2, 6.3 e 6.4. A pontuação PESQ acompanha o aumento da relação sinal-ruído linearmente para todos os ruídos correspondentes aos ambientes reais da comunicação móvel analisados.

1. TESTE-A:

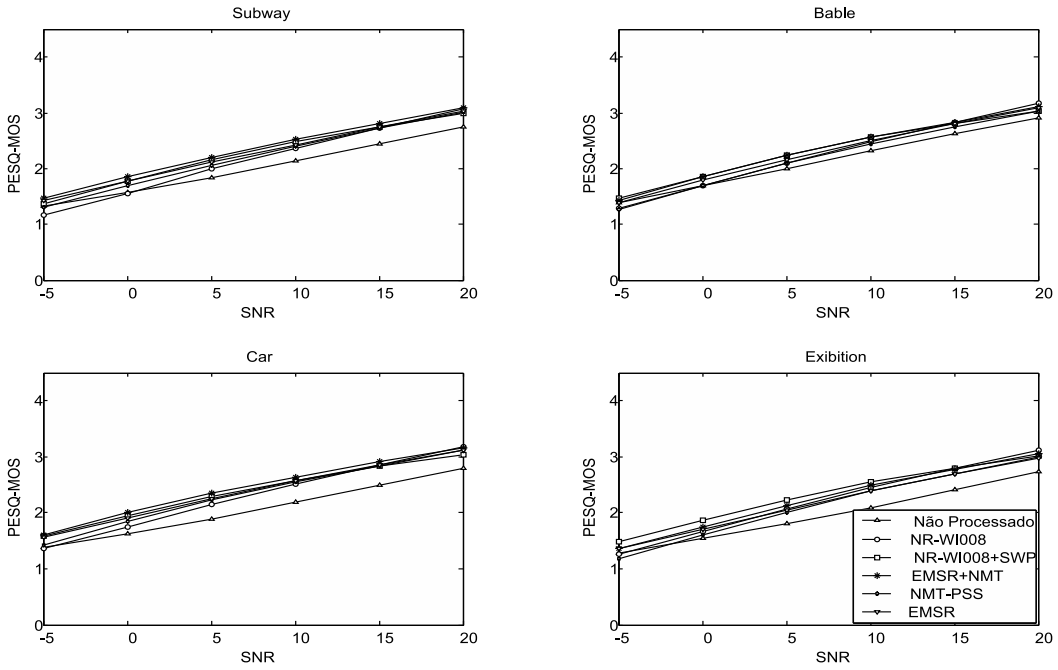

Figura 6.2: *Relação entre SNR e PESQ para os ruídos do TESTE-A. As SNRs são: 20, 15, 10, 5, 0 e -5 dB.*

2. TESTE-B:

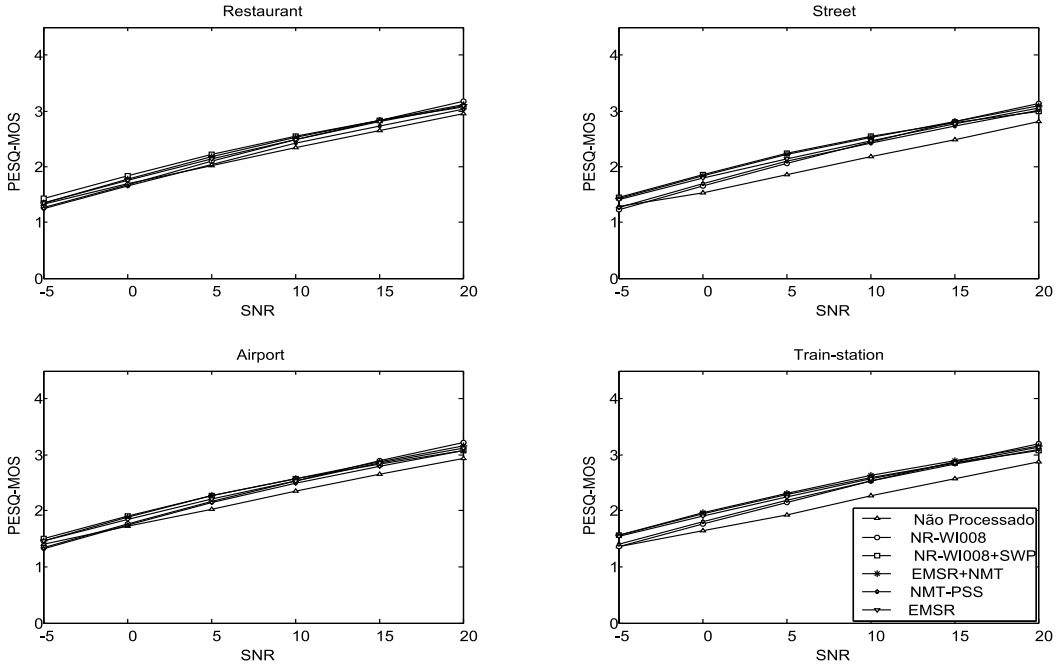

Figura 6.3: *Relação entre SNR e PESQ para os ruídos do TESTE-B. As SNRs são: 20, 15, 10, 5, 0 e -5 dB.*

3. TESTE-C:

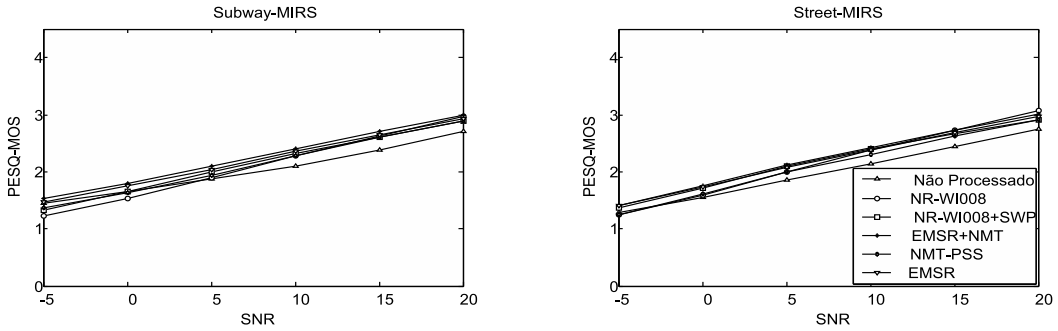

Figura 6.4: *Relação entre SNR e PESQ para os ruídos do TESTE-C. As SNRs são: 20, 15, 10, 5, 0 e -5 dB.*

6.3 Modelamento da Curva PESQ-MOS vs Taxa de Reconhecimento (%)

Os estudos apresentados em [67], sugerem uma aproximação polinomial empírica (de 4ª ordem) que modela a relação entre a pontuação PESQ-MOS e a Taxa de Reconhecimento de fala (%), para vários SNRs e ruídos aditivos. A sua finalidade é tentar predizer através da ferramenta PESQ, a taxa de reconhecimento do sistema de RAF dentro de um vasto cenário de condições adversas.

No entanto, esta predição parece ser válida apenas quando se particulariza um cenário de teste com várias SNRs (do tipo TESTE-A, B ou C) e não quando analisa-se um cenário com vários tipos de testes. É sobre este assunto que esta seção está dedicada.

Este trabalho sugere uma nova aproximação, também empírica, que modela a mesma relação PESQ-MOS vs Taxa de Reconhecimento (%) para cada tipo de ruído, com o mesmo objetivo de tentar predizer a taxa de acerto do sistema. Esta aproximação é baseada na Curva Logística. As motivações que nortearam esta nova abordagem são:

- Aproveitar a alta correlação entre as características gráficas da relação PESQ-MOS vs Taxa de Reconhecimento (%) com a Curva Logística, que faz uma excelente aproximação devido a grande quantidade de dados disponível em cada experimento, como listada na seção 6.2.2, isto é, para cada algoritmo analisado, a curva possui 70 pontos (7 SNRs vezes 10 ruídos) que deverão ser interpolados por uma única função com apenas 3 parâmetros de configuração e com baixo erro;

- Ao contrário da função polinomial que interpola apenas os pontos dados sem oferecer um ponto de convergência da relação PESQ-MOS vs Taxa (%), a Curva Logística aproxima-se, assintoticamente, da Taxa de Reconhecimento (%) para o sinal de fala limpo do sistema (SNR *clean*), ao mesmo tempo que a pontuação PESQ-MOS aumenta em direção à qualidade perceptual "boa";

- Cada parâmetro de configuração da Curva Logística adquire um significado físico que valida essa aproximação.

6.3.1 Parâmetros da Curva Logística

Inicialmente, esta curva possui apenas dois parâmetros de configuração, a e b. Mas um terceiro parâmetro, c, é adicionado para permitir um *offset*, flexibilizando o ajuste vertical. Na Equação (6.1) tem-se a Curva Logística, onde os parâmetros de configuração foram determinados empiricamente para modelar a curva PESQ-MOS vs Taxa de Reconhecimento (%) para cada algoritmo considerado nos ensaios, através de uma varredura de faixa de valores em busca do menor Erro Quadrático Médio (EQM).

$$f(x) = \left(\frac{1}{1 + e^{b - a.x}} - c \right) .100 \tag{6.1}$$

onde x é será o índice PESQ-MOS e $f(x)$ a taxa de acerto em %.

A faixa de valores varrida para cada parâmetro a, b e c da Curva Logística foi de 0 a 12, porém os valores mínimos, médios e máximos obtidos estão listados abaixo para cada tipo de treinamento. O Erro Quadrático Médio de cada algoritmo, pode ser encontrado na Tabela 6.17.

- Treinamento e teste em condições ruidosas:

$$\begin{array}{ll} 3{,}5 \leq \text{a} \leq 5{,}13 & a_{medio} = 4{,}02 \\ 5{,}2 \leq \text{b} \leq 7{,}9 & b_{medio} = 6{,}24 \\ 0{,}013 \leq \text{c} \leq 0{,}027 & c_{medio} = 0.019 \end{array}$$

- Treinamento em condição limpa e teste em condições ruidosas:

$$\begin{array}{ll} 2{,}9 \leq \text{a} \leq 3{,}70 & a_{medio} = 3{,}23 \\ 4{,}8 \leq \text{b} \leq 6{,}50 & b_{medio} = 6{,}20 \\ 0{,}003 \leq \text{c} \leq 0{,}023 & c_{medio} = 0.012 \end{array}$$

Os valores de a, b e c do primeiro treinamento apresentado acima, serão adotados para plotar as curvas a seguir que exemplificarão o comportamento de cada parâmetro. O Erro Quadrático Médio da curva PESQ-MOS vs Taxa (%) de cada algoritmo para cada tipo de treinamento, pode ser encontrado na Tabela 6.3.1.

Tabela 6.17: *Valores dos parâmetros de configuração da curva logística para cada algoritmo submetido ao TESTE-A, B e C e com o respectivo Erro Quadrático Médio - EQM.*

Cenário	Algoritmo	a	b	c	EQM
HMM Ruidoso	Nenhum Processo	5,13	7,90	0,0230	0,024
	NR-WI008	3,70	5,23	0,0230	0,000817
	NR-WI008 + SWP	3,70	5,90	0,0133	0,000441
	EMSR + NMT-PSS	3,87	6,33	0.0167	0.0013
	NMT-PSS	3,77	5,70	0,0200	0,00874
	EMSR	4,00	6,37	0,0200	0,0016
Valor Médio		**4,02**	**6,24**	**0,019**	**0,00615**
HMM Limpo	Nenhum Processo	3,70	7,40	0,010	0,0055
	NR-WI008	2,87	4,80	0,023	0,0018
	NR-WI008 + SWP	3,37	5,87	0,0100	0,000764
	EMSR + NMT-PSS	3,17	6,47	0,0100	0,0028
	NMT-PSS	2,97	6,40	0,0033	0,0011
	EMSR	3,33	6,33	0,0167	0,0039
Valor Médio		**3,23**	**6,21**	**0,012**	**0,00264**

As Figuras 6.5, 6.6 e 6.7 apresentam as curvas para os valores de a, b e c médios, máximos e mínimos. Quanto ao significado físico de cada parâmetro, o mesmo foi obtido através de resultados experimentais que serão discutidos a seguir.

O parâmetro a determina o grau de inclinação da curva. O seu significado físico pode ser interpretado como "sensibilidade de um sistema de reconhecimento de fala", pois quanto maior o seu valor, mais íngreme será a curva, conseqüentemente, a Taxa de Reconhecimento (%) sobe rapidamente para uma pequena variação positiva da pontuação PESQ-MOS, e vice-versa. Por exemplo, observe a Figura 6.5 com diferentes valores de a. Para $a = 5,13$ tem-se o sistema mais sensível, pois com uma pequena variação negativa de 0,5 em torno do valor 2,0 do PESQ-MOS, a Taxa (%) se reduz em quase 20%.

Ao contrário, para $a = 3,5$, tem-se um sistema de RAF menos sensível à variação da qualidade perceptual. No entanto, isto não significa que o sistema seja mais robusto, apenas significa que a taxa de acerto varia menos para um variação pequena do índice MOS. Um sistema robusto seria aquele que possuísse uma curva PESQ vs Taxa de acerto o mais convexa possível, ou seja, "mais logarítmica" (ou então, "menos linear"), mantendo alta a taxa de acerto para

uma ampla faixa de valores PESQ.

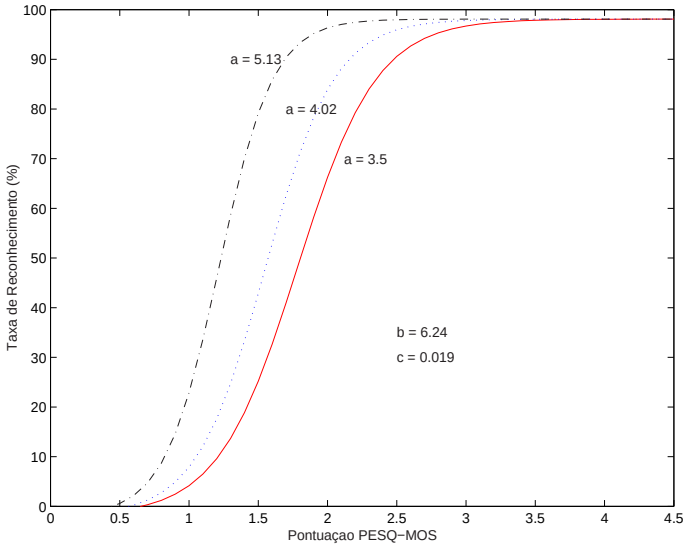

Figura 6.5: *Exemplo de diferentes valores do parâmetro de configuração a, onde* a_{medio} *= 4,02 ,* a_{max} *= 5,13 e* a_{min} *= 3,5 foram adquiridos levantando a curva para todos os algoritmos de um sistema, treinado em condições ruidosas e testado com TESTE-A, B e C.*

O parâmetro b desloca a curva horizontalmente, especificamente, para a esquerda quando b diminui e para a direita quando ele aumenta.

Este parâmetro b somente adquire significado físico ao ser associado ao parâmetro a. Assim, pode-se afirmar com certeza que um sistema de RAF com a alto e b baixo, é um sistema robusto, pois um valor pequeno de b provoca o início da subida da curva para baixos índices PESQ-MOS, enquanto que um valor alto de a faz com que esta "subida" seja íngreme, atingindo um patamar elevado de taxa de acerto mesmo para valores obviamente pequenos de PESQ-MOS.

A Figura 6.6 mostra b com diferentes valores. Para $b = 5,2$, tem-se o sistema com o melhor desempenho.

Por último, o parâmetro de configuração adicional c, que permite deslocar a curva na vertical, fisicamente possui uma relação com a Taxa de Reconhecimento (%) para um sistema em condições favoráveis (livre de ruído), ou seja, quanto menor o valor de c, mais alta será a taxa de acerto para PESQ-MOS = 4,5 (sinal completamente limpo). Observe a Figura 6.7 com diferentes valores de c.

6.3.2 Análise Experimental da Aproximação

Nesta seção, dois gráficos obtidos experimentalmente serão apresentados para validar a aproximação proposta.

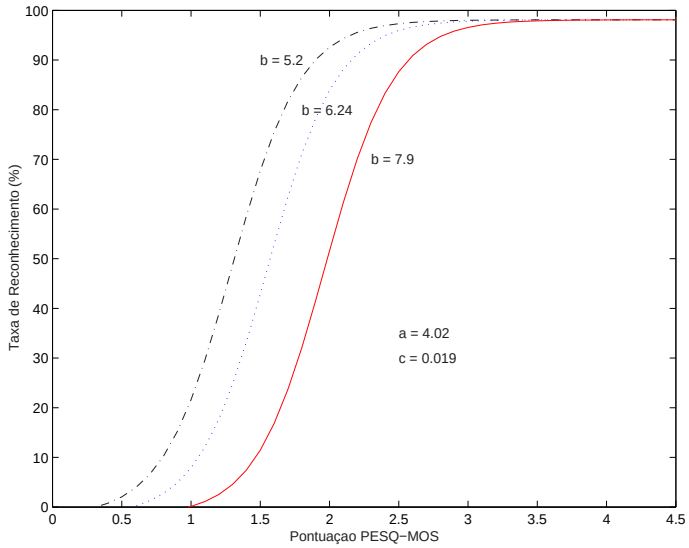

Figura 6.6: *Exemplo de diferentes valores do parâmetro de configuração b, onde* b_{medio} *= 6,24 ,* b_{max} *= 7,9 e* b_{min} *= 5,2 foram adquiridos levantando a curva para todos os algoritmos de um sistema treinado e testado com locuções ruidosas.*

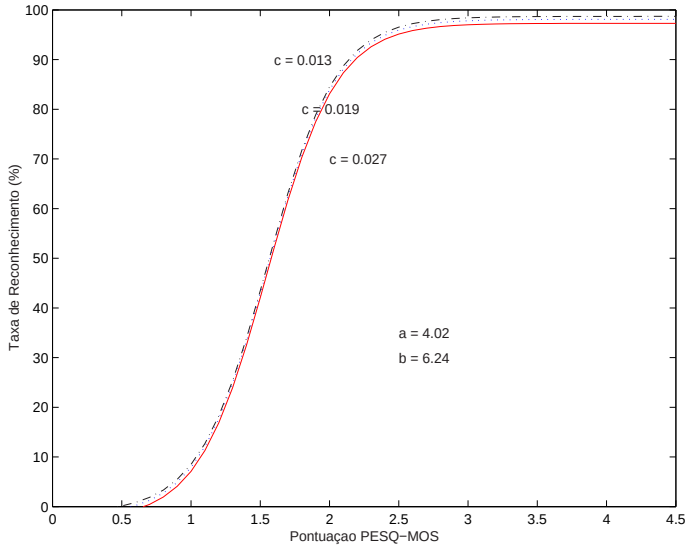

Figura 6.7: *Exemplo de diferentes valores do parâmetro de configuração c, isto é,* c_{medio} *= 0.019,* c_{max} *= 0.027 e* c_{min} *= 0.013 que foram adquiridos levantando a curva para todos os algoritmos de um sistema treinado e testado com locuções ruidosas.*

A Figura 6.8, apresenta um panorama geral resultante dos TESTE-A, B e C de todos os algoritmos com um sistema treinado em condição ruidosa.

Note que a curva do sistema com Nenhum Processo possui um grau de inclinação maior que os demais algoritmos. Então, a hipótese que explica o significado físico do parâmetro *a* da Curva Logística parece ser verdadeira, pois não foi realizado nenhum esforço para eliminar o ruído. Além disso, a Tabela 6.17

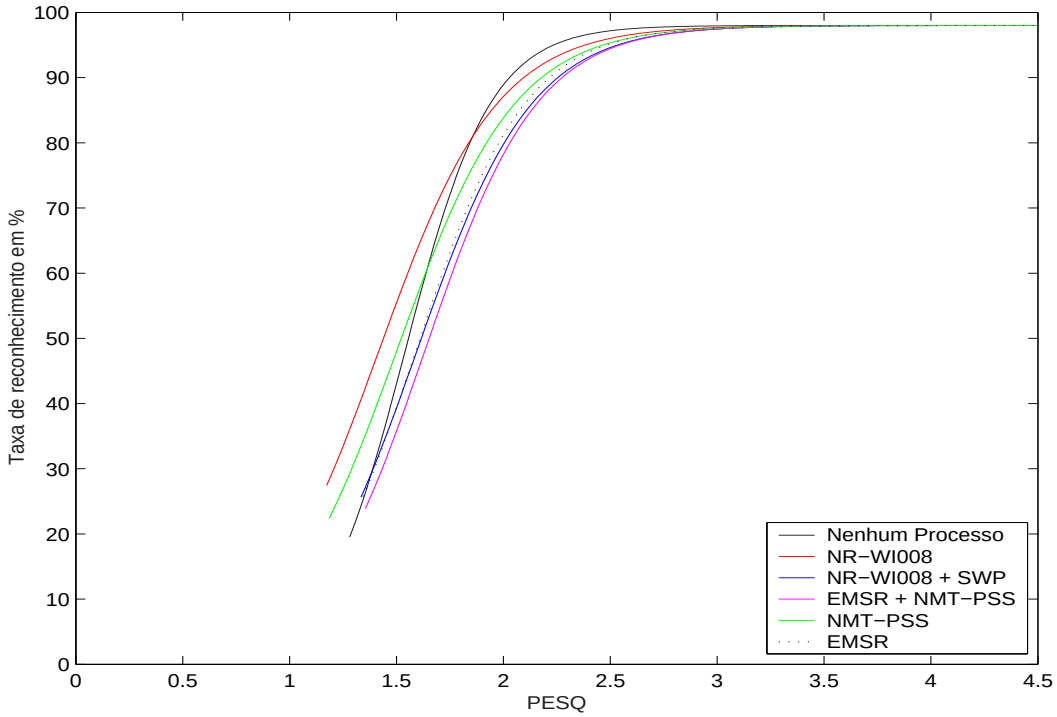

Figura 6.8: *Comparação da aproximação Logística para cada algoritmo de pré-processamento incluindo todos os ruídos dos ambientes reais da comunicação móvel. Cada curva possui 7 SNR (clean, 20, 15, 10, 5, 0 e -5) vezes 10 tipos de ruído (Subway, Babble, Car, Exhibition, Restaurant, Street, Airport, Train-station, Subway-MIRS e Street-MIRS), totalizando 70 pontos. O sistema de reconhecimento foi treinado e testado em múltiplas condições.*

apresenta o maior valor de a para o ensaio Nenhum Processo.

Na Figura 6.9, tem-se a curva média de um sistema com treinamento limpo submetido aos TESTES A, B e C.

Neste gráfico, é possível observar o significado físico do parâmetro b. Como já era esperado, a curva do algoritmo NMT-PSS apresenta o pior desempenho, pois a Taxa (%) começa a cair já para valores relativamente altos da pontuação PESQ-MOS (em torno de 3,4). Como já se sabe, o motivo é o ruído musical remanescente no sinal de fala "melhorado".

Observe também, que as curvas dos algoritmos NR-WI008 e NR-WI008 + SWP parecem confirmar os resultados esperados dos métodos utilizados.

O algoritmo NR-WI008 possui como principal método uma filtragem de Wiener relativamente complexa, que serve para reduzir ao máximo o ruído aditivo, porém causando uma pequena deformação do sinal (veja as Tabelas 6.14, 6.15 e 6.16) mesmo quando a SNR está alta. Em outras palavras, em alta SNR, a distorção do sinal de fala causada por este algoritmo torna-se sensível a ponto de reduzir levemente a taxa de acerto, enquanto que em baixa SNR, o ruído é tão prejudi-

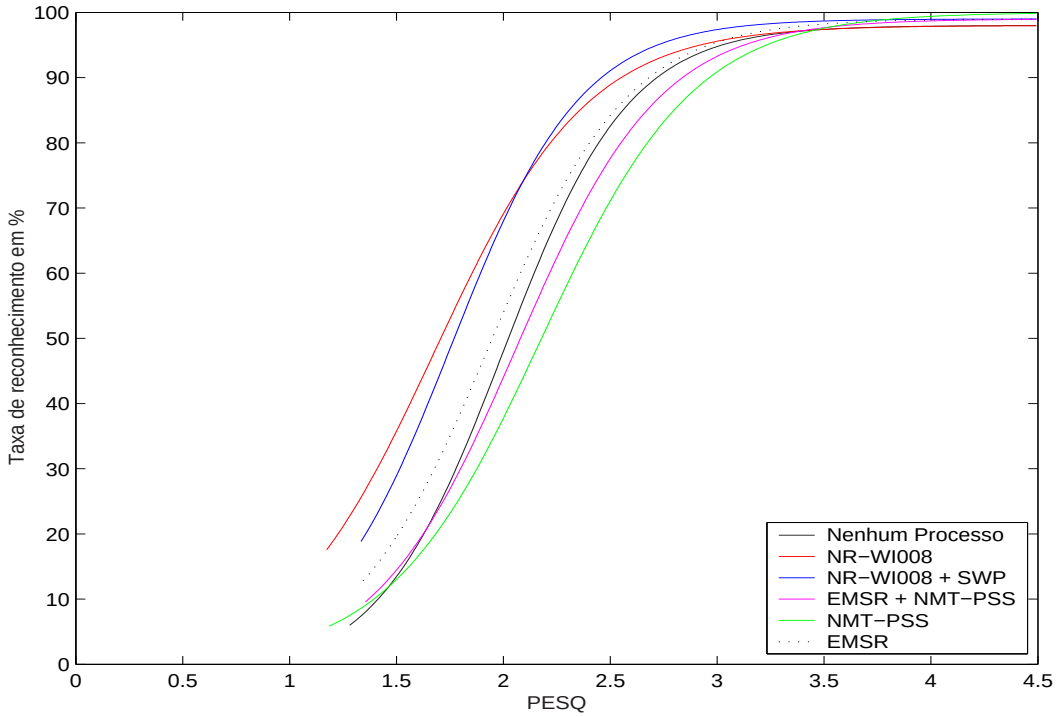

Figura 6.9: *Comparação da aproximação Logística para cada algoritmo de pré-processamento incluindo todos os ruídos dos ambientes reais da comunicação móvel. Cada curva possui 7 SNR (clean, 20, 15, 10, 5, 0 e -5) vezes 10 tipos de ruído (Subway, Babble, Car, Exhibition, Restaurant, Street, Airport, Train-station, Subway-MIRS e Street-MIRS), totalizando 70 pontos para interpolação. O sistema de reconhecimento foi treinado e testado em condição limpa.*

cial, que mesmo o algoritmo distorcendo o sinal durante o processo de redução de ruído, há uma melhora na taxa de acerto.

Assim, parece que é uma vantagem não aplicar nenhum algoritmo quando for possível garantir numa determinada aplicação, SNR alta, pois como foi visto, o ensaio Nenhum Processo apresentou o melhor desempenho para esta condição. Ao passo que, quando não for possível garantir SNR alta, utiliza-se o algoritmo NR-WI008, isto é, WI008 do ETSI.

O algoritmo NR-WI008 + SWP tenta reduzir o ruído do sinal trabalhando com ponderações que dependem da estimação espectral adaptativa da SNR. Como a SNR já foi melhorada no bloco anterior (NR - *Noise Reduction*), a redução passa a ser moderada, ou seja, sem causar deformação da forma de onda do sinal. Por isso a curva apresenta uma Taxa (%) maior para PESQ em torno de 2,5, devido à maior facilidade em realçar um sinal com ruído reduzido. Porém, a inclinação é um pouco maior do que o algoritmo NR-WI008 para a faixa de pontuação 1,5 a 2,0, isto é, menos resistência à variação do PESQ. O motivo são as deformações causadas pelo NR no bloco anterior.

Contudo, isso não parece ser uma desvantagem porque o ganho já foi obtido: o PESQ do algoritmo NR-WI008 + SWP é maior do que o algoritmo NR-WI008. Conseqüentemente, a SNR melhora e o desempenho do sistema aumenta.

Outro resultado esperado é do algoritmo EMSR, que provou melhor eficiência sobre os algoritmos NMT-PSS e EMSR + NMT-PSS quando o sistema é treinado em condição limpa. O motivo parece estar na contribuição negativa do limiar de mascaramento do ruído. Talvez seja preferível fazer uma redução adaptativa baseada na SNR a obscurecer o ruído através do limiar de mascaramento auditivo do ouvido humano.

6.4 Interpretação do Comportamento dos Ruídos

Nesta seção, serão apresentadas algumas interpretações que descrevem o influência de alguns ruídos sobre o desempenho dos sistemas, apontando as suas vantagens e desvantagens.

A Figura 6.10 mostra a aproximação da relação PESQ-MOS vs Taxa (%) para cada tipo de ruído. Cada curva foi obtida interpolando os resultados das Taxas (%) e PESQ-MOS, obtidas pelos algoritmos de realce de um sistema treinado e testado em condições ruidosas.

Os pontos observados foram: estacionariedade, influência da filtragem MIRS e espalhamento dos pontos da curva de sistemas com treinamento limpo.

1. Estacionariedade

   - Analisando os ruídos *Subway* e *Exhibition* da Figura 6.10, observa-se que o sistema apresenta um melhor desempenho devido ao menor grau de inclinação da curva e a manutenção da taxa de acerto em alta para uma diminuição da pontuação PESQ, como foi visto na seção 6.3.1. No entanto, o ruído *Exhibition* é considerado estacionário e o *Subway*, não-estacionário [59].

     Assim, a fato da estacionariedade do ruído parecer não influenciar o desempenho destes algoritmos de realce, caracteriza bons sistemas de pré-processamento. No entanto, o fator que parece tornar um ruído mais prejudicial do que outro (como por exemplo, o *Restaurant*) são as suas características espectrais;

   - Uma outra análise para os mesmos ruídos (*Subway* e *Exhibition*) da Figura 6.10, é a semelhança do comportamento espectral entre ambos

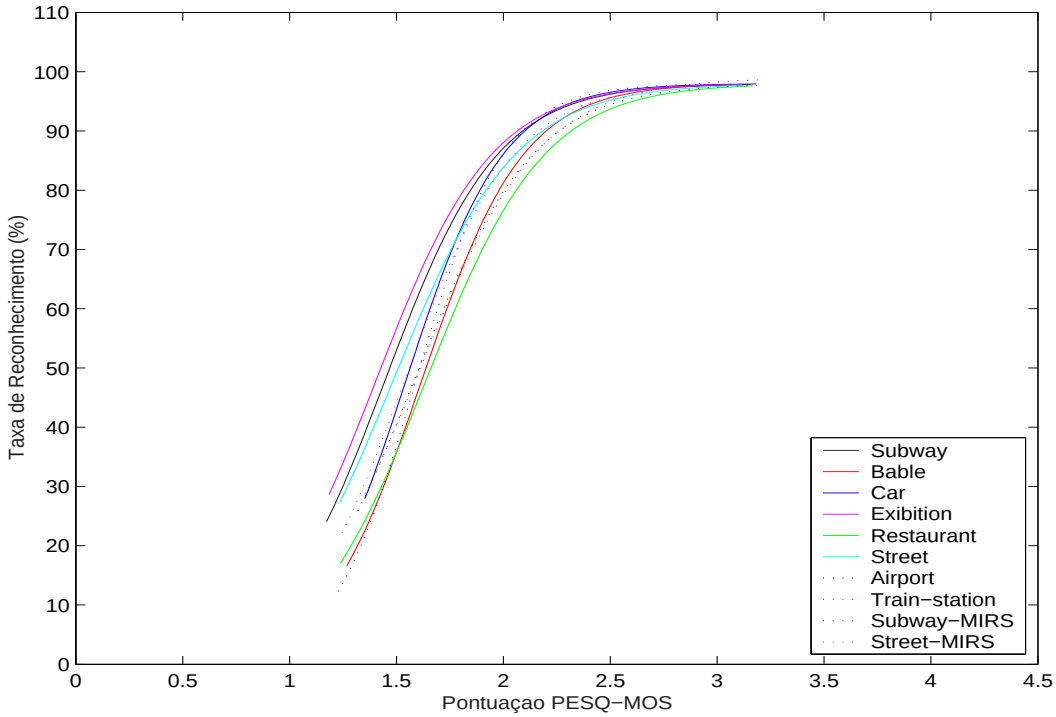

Figura 6.10: *Aproximação média da relação PESQ-MOS vs Taxa (%) para cada tipo de ruído processado pelos algoritmos de realce. O sistema foi treinado e testado em múltiplas condições. Cada curva possui 42 pontos interporlados: 7 (SNRs) vezes 6 (algoritmos).*

(as curvas se mantém paralelas). Isso parece apontar a seguinte possibilidade: predizer o desempenho destes dois ruídos com apenas um dos ensaios (*Subway* ou *Exhibition*), economizando tempo e dinheiro;

2. Influência da filtragem MIRS

   - Na Figura 6.10, o ruído *Subway-MIRS* (filtragem MIRS de equipamentos GSM definido pelo ITU), possui o maior grau de inclinação, diminuindo a Taxa de Reconhecimento (%) rapidamente para uma pequena variação da pontuação PESQ-MOS. O motivo parece estar na resposta em freqüência do filtro MIRS, que atenua as componentes de menor freqüência que coincidem com boa parte da faixa da inteligibilidade da voz humana. Com isso há uma predominância das componentes de maior freqüência no sinal de fala ruidoso, prejudicando o sistema. É um comportamento semelhante ao do ouvido humano, que se irrita facilmente com a presença de espúrios de alta freqüência se comparado com os de baixa freqüência. Observe a resposta em freqüência do filtro MIRS na Figura 6.12, retirada do artigo [59]. Na Figura 6.13,

tem-se o espectro médio do ruído *Subway*, com energia razoável nas componentes de alta freqüência.

3. Espalhamento dos pontos da curva de sistemas com treinamento limpo:

   - Como já era esperado, a nuvem de pontos da Figura 6.11 é mais dispersa para um sistema treinado na condição limpa. Se não há informações sobre os tipos de ruído nos modelos de palavras e pausas, é normal o sistema apresentar uma queda na sua robustez, principalmente para os ruídos *Restaurant* e *Babble*.

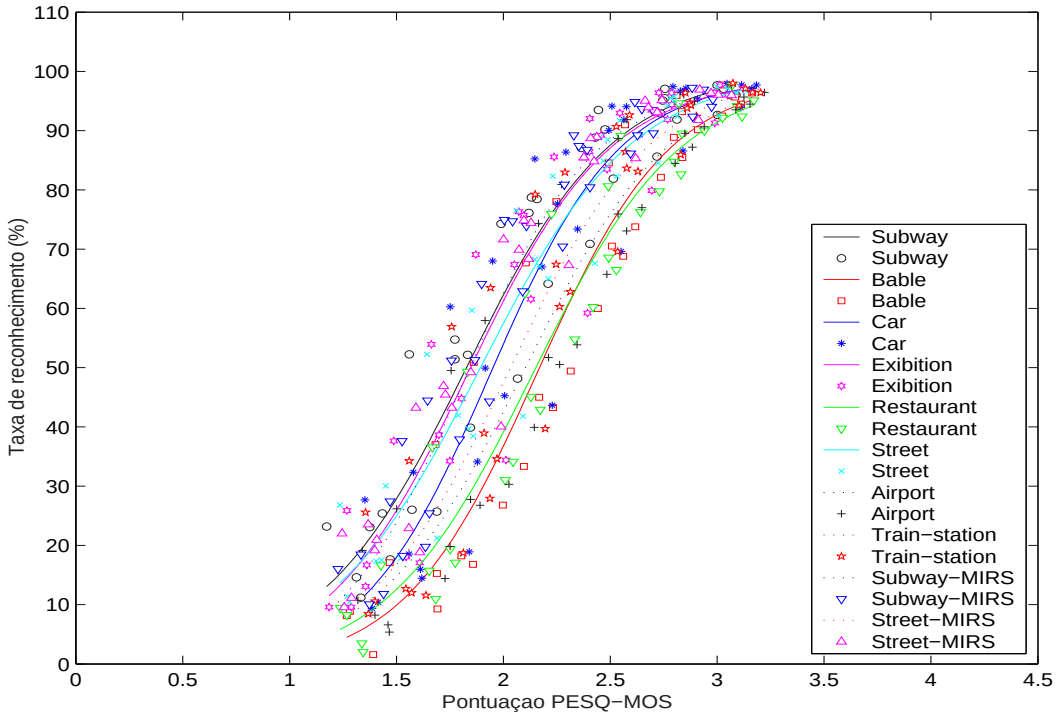

Figura 6.11: *Aproximação média da relação PESQ-MOS vs Taxa (%) para cada tipo de ruído processado pelos algoritmos de realce. O sistema foi treinado em condição limpa e testado em condições ruidosas. Cada curva possui 42 pontos interporlados: 7 SNRs vezes 6 algoritmos.*

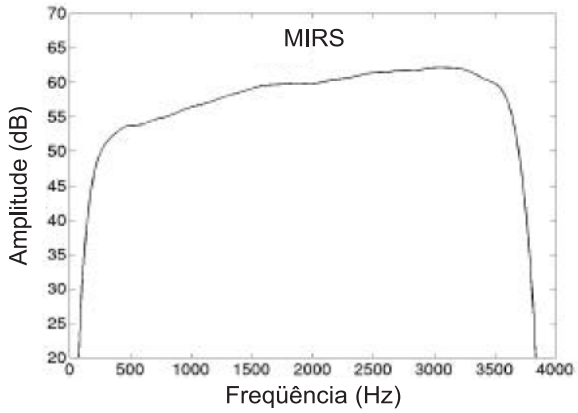

Figura 6.12: *Resposta em freqüência do filtro MIRS.*

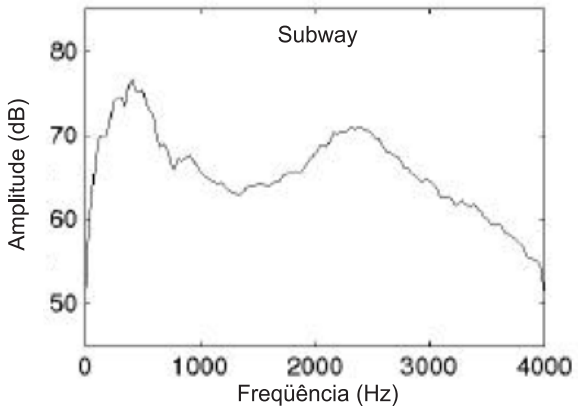

Figura 6.13: *Espectro médio do ruído real Subway da base de dados Aurora retirado do artigo [59].*

Capítulo 7

Conclusões

O trabalho desenvolvido nesta dissertação teve como objetivos principais, fazer uma avaliação perceptual da qualidade da fala realçada que os algoritmos de pré-processamento do sistema de reconhecimento alcançaram em ambiente ruidoso e comparar o desempenho deste mesmo sistema através da taxa de acerto obtida.

Observou-se que cada teste apresenta alguma particularidade, pois as condições de treinamento e testes mudam. Por exemplo, considerando o treinamento com múltiplas condições para os testes iniciais, o TESTE-A apresentou o melhor desempenho, pois os ruídos são os mesmos do treinamento. O bom desempenho do TESTE-B do mesmo treinamento mostrou que boa parte das características espectrais dos seus ruídos são adequadamente aproximadas pelas dos ruídos do treinamento. O TESTE-C apresentou uma piora no desempenho quando se utiliza o filtro da especificação GSM 03.50 (filtro MIRS), mostrando assim a importância do ruído convolucional (distorção no processo de RAF).

Para o treinamento em condição limpa, o TESTE-A e B resultaram num menor desempenho se comparado com o treinamento com múltiplas condições. O motivo é claro: os modelos de palavras e pausas do sistema não possuem nenhuma informação quanto ao comportamento de cada tipo de ruído. E ainda, observou-se que existem ruídos que degradam mais que outros, como é caso do *Babble* e *Restaurant*, pois apresentam segmentos fortemente não-estacionários. No TESTE-C, houve uma melhora inesperada para o ruído *Street*. Parece que a contribuição é da resposta em freqüência do filtro MIRS, que atenua as componentes de baixa freqüência, onde também está a maior parte da energia do ruído em questão.

Na avaliação e comparação da Taxa de Reconhecimento entre todos os algoritmos, foi observado que o ruído musical é menos perceptível quando o cenário de testes do sistema possui diferentes ruídos daqueles utilizados na fase de treinamento. Em outras palavras, o ganho na redução do ruído aditivo desconhecido

(TESTE-B em múltiplas condições) é maior que a degradação do ruído musical.

No entanto, esta situação se inverte quando os HMMs são testados com os mesmos ruídos do treinamento: o ruído musical é o maior causador da degradação do sistema, como mostra o TESTE-A sob múltiplas condições.

Para os demais TESTES (A, B e C) aplicados em sistemas com treinamento limpo, na comparação da Taxa de Reconhecimento, as conclusões são diferentes, pois as condições são outras. Ocorre degradações devido ao ruído musical em todos os TESTES, pois além dos HMMs não possuírem dados sobre o seu comportamento, ele é o maior agravante do desempenho do sistema.

Na avaliação perceptual da qualidade da fala realçada, foi possível observar que praticamente não há um algoritmo com a melhor pontuação PESQ para todas as SNRs e cenários de teste. Embora alguns obtiveram as melhores pontuações em determinados TESTES, os demais não apresentaram resultados tão diferentes.

Desta forma, os critérios de escolha dentre os algoritmos analisados, poderão ser a faixa de SNR predominante da aplicação e a carga computacional requerida. Por exemplo, um sistema em que predomina uma SNR relativamente alta (15 dB), poderá ser utilizado o algoritmo EMSR + NMT-PSS, considerado computacionalmente leve. Ou ainda, não utilizar nenhum algoritmo. Para uma SNR crítica, (em torno de 0 dB), pode-se utilizar um algoritmo mais pesado, como o WI008.

Além disso, parece que a melhor pontuação PESQ nem sempre é daquele algoritmo que fornece a maior Taxa de Reconhecimento (%), como foi o caso do algoritmo NR-WI008 + SWP. No entanto, vale lembrar que neste trabalho, não foi possível avaliar a qualidade perceptual deste algoritmo por inteiro, pois num dado instante do seu processo, as amostras do sinal de fala passam para o domínio cepstral, dificultando a sua reconstituição no formato WAVE, necessário na medição PESQ.

Também, um efeito claramente observável é a deformação que o algoritmo NR-WI008 causou no sinal de fala realçado em condições críticas da SNR (em torno de 0 dB), diminuindo a sua qualidade perceptual. Neste caso, o bloco *Redução de Ruído* do algoritmo WI008 parece ter uma menor relação "redução de ruído aditivo" versus "deformação do sinal de fala" do que o algoritmo NR-WI008 + SWP. Porém, numa visão de conjunto, o WI008 está sendo considerado pelo ETSI, como o melhor algoritmo, devido às suas altas Taxas de Reconhecimento (%) (Anexo C).

Quanto ao ruído musical, foi possível mensurar a sua influência negativa sob a qualidade do sinal de fala realçado, ao se analisar o algoritmo NMT-PSS, que obteve as menores pontuações PESQ ao longo da faixa da SNR de 5 a 20 dB dos TESTE-A e B, e de 10 a 20 dB do TESTE-C. Para este último teste, parece que

não foi uma vantagem, do ponto de vista da qualidade perceptual, atenuar as componentes de baixa freqüência através do filtro MIRS, pois juntamente com a redução da energia do ruído, perde-se também um pouco da energia da faixa do sinal de voz.

Quanto à aproximação que representa a relação PESQ-MOS vs Taxa de Reconhecimento (%), observou-se uma excelente aproximação devido à grande quantidade de dados disponível em cada experimento, que puderam ser modelados por uma única função com apenas 3 parâmetros de configuração (a, b e c) e com baixo EMQ. Além disso, ao contrário da função polinomial que interpola apenas os pontos dados sem oferecer um ponto de convergência da relação PESQ-MOS vs Taxa (%), a Curva Logística aproxima-se, assintoticamente, da Taxa de Reconhecimento (%) para o sinal de fala limpo do sistema (SNR *clean*), ao mesmo tempo que a pontuação PESQ-MOS aumenta em direção à qualidade perceptual "boa", como se esperava.

Um dos pontos mais importantes desta aproximação, é o significado físico de cada parâmetro, que já foi devidamente apresentado no capítulo anterior.

E por fim, ao analisar alguns fatos inerentes aos ensaios, como a estacionariedade, a influência da filtragem MIRS e o espalhamento dos pontos da curva PESQ-MOS vs Taxa (%) de sistemas com treinamento limpo, foi possível levantar algumas hipóteses.

Por exemplo, a curva PESQ-MOS vs Taxa (%) de alguns ruídos, apresenta um melhor desempenho, especificamente dos ruídos *Subway* e *Exhibition*, devido ao menor grau de inclinação da curva. Conseqüentemente, há uma manutenção da taxa de acerto em alta para uma diminuição da pontuação PESQ. Por isso, a estacionariedade do ruído parece não influenciar o desempenho destes algoritmos de realce, caracterizando bons sistemas de pré-processamento.

Quanto à influência da filtragem MIRS, o ruído *Subway-MIRS* possui o maior grau de inclinação, diminuindo a Taxa de Reconhecimento (%) rapidamente para uma pequena variação da pontuação PESQ-MOS, pois a resposta em freqüência do filtro MIRS atenua as componentes de menor freqüência, que coincidem com uma parte da faixa da inteligibilidade da voz humana.

Sobre o espalhamento dos pontos da curva de sistemas com treinamento limpo, como já era esperado, a nuvem de pontos é mais dispersa. Se não há informações sobre os tipos de ruído nos modelos de palavras e pausas, é normal o sistema apresentar valores bem diferentes para cada ruído e condição. E ainda, uma queda na sua robustez, principalmente para os ruídos *Restaurant* e *Babble*.

Anexo A

Algoritmo PESQ

Como dito antes, o algoritmo PESQ segue os mesmos passos do PSQM [65], [74], mas com as modificações citadas na seção 4.4.2. Cada um dos passos consecutivos será descrito a seguir.

A.1 Calibração

O primeiro passo do algoritmo PESQ é compensar de forma geral, o ganho do sistema que está sob teste. Esta etapa consiste de um escalonamento global dos sinais envolvidos, o original $X(t)$ e o degradado $Y(t)$, para o mesmo nível de potência.

Desta forma, o PESQ assume duas considerações: o nível de audição subjetivo é uma constante, entorno de 79 dB SPL (*Sound Pressure Level*) para o ouvido humano [72]. A segunda consideração diz respeito às variações entre os níveis dos sinais gravados dentro de um ensaio subjetivo simples: são pequenas variações.

O alinhamento de nível do PESQ é executado com base na potência da banda passante do filtro (300 - 3000 Hz) de ambos os sinais, original e degradado. Ao lado do alinhamento de nível no domínio do tempo, há o alinhamento no domínio da freqüência, logo que houver a análise tempo e freqüência (cálculo da FFT). Tal alinhamento é feito através da geração de uma forma de onda senoidal, com freqüência de 1000 Hz e uma amplitude de 40 dB SPL. Esta senoide também é transformada para o domínio da freqüência através da Transformada Rápida de Fourier (FFT) janelada com 32 ms de comprimento de quadro. Depois de converter a escala do eixo das freqüências para a escala Bark modificada, a amplitude de pico da densidade de potência resultante é então normalizada pela multiplicação de um fator de escalonamento S_p de potência de 10^4 [69]. O mesmo tom de referência de 40 dB SPL é utilizado para calibrar a escala *loudness* do

modelo psico-acústico. Depois de deformar o eixo da intensidade (ou eixo vertical) para uma escala *loudness* utilizando as leis de Zwicker [76], a densidade *loudness* sobre a escala de freqüência Bark é normalizada para 1 utilizando o fator de escalonamento *loudness* S_1 [69].

A.2 Filtragem do Receptor

Assume-se que o processo de audição é executado através de um dispositivo manual com resposta em freqüência correspondente ou a um receptor IRS (*Intermediate Reference System*) [81] ou às características de um receptor IRS modificado [72]. Um modelo de avaliação perceptual da qualidade da fala humana foi determinado para computar os sinais que são ouvidos subjetivamente. Para isso, ambos os sinais, original e degradado, são filtrados pelo receptor IRS. No PESQ, isso é implementado através da FFT do sinal completo, em seguida, passando por um filtro com uma resposta linear às características do receptor IRS [81], seguida pela inversa da FFT do sinal completo. Como resultado, tem-se o $X_{IRSS}(t)$ e o $Y_{IRSS}(t)$ do sinal de entrada $X_S(t)$ e saída $Y_S(t)$, respectivamente.

A.3 Cálculo da Fala Ativa

Em ambos os arquivos de fala, original e degradado, se iniciados ou finalizados com grandes intervalos de silêncio, poderiam influenciar nos seus respectivos cômputos dos valores médios de distorção. Desta forma, uma estimativa das porções silenciosas é feita do começo ao fim de cada arquivo. Se a soma da amplitude absoluta de cinco amostras consecutivas do sinal original ultrapassar um limiar, aquela posição classifica-se como "início" ou "fim" de um intervalo ativo. O intervalo entre este início e fim é definido como um intervalo de tempo de voz ativa.

A.4 Decomposição Tempo-Freqüência e Modificação do Eixo Tempo

O ouvido humano desempenha uma transformação tempo-freqüência. No PESQ, este fenômeno é caracterizado por uma FFT de tempo curto com uma janela Hanning de 32 ms. A sobreposição de quadros é de 50%. O espectro de potência (valor absoluto do sinal complexo) é armazenado em vetores separados para os sinais: original e degradado. A informação da fase dentro de cada quadro é

descartada e todos os cálculos são baseados apenas na representação da potência $PX_{WIRSS}(f)_n$ e $PY_{WIRSS}(f)_n$.

Os pontos de início de cada quadro do sinal degradado são deslocados sobre o atraso observado no estimador de atraso variável [68], enquanto que o eixo do tempo do sinal original é mantido sem alterações. Se o atraso aumentar, porções do sinal degradado são excluídas durante o processamento e para os casos onde o atraso diminui, porções do sinal degradado são repetidas. Esta modificação do eixo do tempo, forneceu de forma geral, os melhores resultados da correlação com a qualidade da fala percebida subjetivamente. Uma pequena extensão desta estratégia é dada na seção A.12.

A.5 Cálculo da Densidade de Potência do Sinal

A escala Bark indica que as baixas freqüências do sistema auditivo humano têm uma melhor resolução em freqüência ao se comparar com as altas freqüências. Esta técnica é implementada multiplicando os correspondentes índices da FFT pelas correspondentes bandas da escala Bark e em seguida, fazendo uma soma normalizada da potência por banda. A função que transforma Hertz em Bark é a mesma citada na literatura de processamento de voz. Os sinais resultantes são conhecidos como densidade de potência do sinal $PPX_{WIRSS}(f)_n$ e $PPY_{WIRSS}(f)_n$.

A.6 Compensação da Resposta em Freqüência Linear

Para tratar da filtragem de sistemas sob teste, uma média do espectro de potência do sinal original e do degrado é calculada (esta média só é feita sobre os quadros que foram considerados como quadros de voz ativa). Então, para cada índice Bark, um fator de compensação é calculado a partir da relação entre o espectro degradado e o espectro original. A maior compensação não deve ultrapassa 20 dB. A densidade de potência do sinal original de cada quadro n é então multiplicada por este fator de compensação parcial para haver uma equalização a partir do sinal original para o degradado. Isto resulta numa versão filtrada da densidade de potência do sinal original $PPX'_{WIRSS}(f)_n$.

Esta compensação parcial é utilizada porque várias filtragens atrapalham o ouvinte enquanto que uma filtragem moderada raramente afeta a qualidade geral percebida, principalmente se nenhuma referência é disponibilizada para o indivíduo para comparação. E por fim, a compensação é feita somente no sinal

original porque o sinal degradado é o único que será julgado pelas pessoas num ensaio ACR (*Absolute Category Rating*).

A.7 Compensação do Ganho Variante no Tempo

Nesta etapa, temos um cálculo da relação da densidade de potência entre os arquivos original e o degradado. Em seguida, o resultado é limitado na faixa de $\{3.10^{-4}, 5\}$ e ainda um filtro passa-baixa de primeira ordem é aplicado nesta relação ao longo do eixo do tempo (a constante de tempo deste filtro é de aproximadamente 16 ms). O motivo de se usar um passa-baixas é que dependendo da idade e do sexo do locutor, entre 60 e 90% da energia da fala natural está abaixo de 500 Hz.

Por fim, a densidade de potência do sinal distorcido em cada quadro n é então multiplicada por esta relação, resultando numa densidade de potência do sinal distorcido com ganho compensado parcialmente, o $PPY'_{WIRSS}(f)_n$.

A.8 Cálculo da Densidade Loudness

Depois da compensação parcial tanto por filtragem quanto por variações do ganho em tempo curto (discutidas na seção anterior), as densidades de potência de ambos os sinais são transformadas para novo um volume de sonoridade que possa ser mais bem ouvido (*loudness*), utilizando as leis de Zwicker [76], como mostra a Equação A.1:

$$LX(f)_n = S_1 \left[\frac{P_O(f)}{0,5}\right] \times \left\{\left[0,5 + 0,5\frac{PPX'_{WIRSS}(f)_n}{P_O(f)}\right] - 1\right\} \qquad \text{(A.1)}$$

onde $P_O(f)$ é o limiar absoluto de audição e S_1, o fator de escalonamento *loudness*. O mesmo é feito para a densidade de potência do sinal degradado, substituindo o $PPX'_{WIRSS}(f)_n$ por $PPY'_{WIRSS}(f)_n$. O resultado é um vetor de duas dimensões, e ambos, $LX(f)_n$ e $LY(f)_n$, são conhecidos como *densidade loudness*.

A.9 Cálculo da Densidade de Distúrbio

Neste ponto, é calculada uma subtração entre a *densidade loudness* distorcida e a original. Quando esta diferença é positiva, significa que componentes como o

ruído foram adicionados. E quando a diferença é negativa, componentes do sinal original foram excluídos. Este vetor de diferenças é conhecido como *densidade de distúrbio bruto*.

Para que este cálculo seja possível, um vetor (bi-dimensional) com valores de mascaranento é determinado em cada quadro, como mostra o artigo [69], e em seguida há uma seqüência de regras aplicada em tais quadros:

- Se a densidade de distúrbio bruta é positiva e maior que o valor de mascaramento, este é subtraído do distúrbio bruto.
- Se a densidade de distúrbio bruta oscila entre mais e menos do valor da magnitude do mascaramento, a densidade de distúrbio é jogada para zero.
- Se a densidade de distúrbio é mais negativa do que o valor negativo do mascaramento, o valor do mascaramento é somado com a densidade de distúrbio.

Este método modela o processo de pequenas diferenças inaudíveis na presença de sinais sonoros (onde há o mascaramento) em cada quadro. O resultado é uma densidade de distúrbio $D(f)_n$ como uma função no tempo (do quadro n) e na freqüência.

A.10 Modelamento dos Efeitos Assimétricos

O efeito assimétrico, como já foi citado anteriormente, é causado pela distorção do codificador/decodificador sobre o sinal de entrada. E em geral, este fato complica a reconstituição deste sinal ao tentar introduzir novas componentes de freqüência e tempo, e o resultado final acaba sendo decomposto claramente em dois sinais com distorções audíveis: o sinal de entrada e o distorcido [73]. Quando o codificador/decodificador omiti componentes de freqüência e tempo, o resultado final pode não ser decomposto da mesma forma, e ainda a distorção ser menos censurada. Este efeito é modelado pelo cálculo da densidade de distúrbio assimétrico $DA(f)_n$ por quadro, através da multiplicação entre a densidade de distúrbio $D(f)_n$ e um fator assimétrico (citados na seção A.9).

Este fator assimétrico é igual à razão da densidade de potência do original e do distorcido levado à potência de 1,2. Se o fator assimétrico é menor que 3, então é jogado para zero. Caso ele exceda 12, seu valor é ceifado [69].

A.11 Agregando a Densidades de Distúrbios à Freqüência e ao Processamento dos Intervalos de Silêncio

A densidade de distorção e a densidade de distúrbio assimétrico são somadas ao longo do eixo da freqüência utilizando duas normas L_p diferentes (que serão mostradas na seção A.13) e uma ponderação dos quadros suaves (que possuem baixa sonoridade) como mostra as Equações A.2 e A.3:

$$D_n = M_n \sqrt[3]{\sum_{f=1}^{k} [|D(f)_n| W_f]^3} \tag{A.2}$$

$$DA_n = M_n \sum_{f=1}^{k} [|DA(f)_n| W(f)] \tag{A.3}$$

onde k é o número de bandas Bark e M_n é um fator de multiplicação igual a $[(potencia\ \ do\ \ quadro\ \ original + 10^5)/10^7]^{-0,04}$, resultando numa ênfase dos distúrbios que ocorreram durante os segmentos de silêncio do sinal de fala original, e W_f é uma série de constantes proporcionais aos índices Bark [69]. Estes valores provenientes de cada somatório, D_n e DA_n, são chamados de "quadros distorcidos". Observe a Figura A.1.

Se o sinal distorcido possui um decréscimo no atraso maior que 16 ms (metade de um quadro da FFT), a estratégia mencionada na seção A.4 é aplicada. Excepcionalmente neste caso, no cálculo da qualidade objetiva da fala é melhor ignorar os quadros distorcidos durante um decréscimo em atraso tão pequeno ($< 16\ ms$), ao invés de repeti-los como manda a regra. Como conseqüência, os quadros distorcidos serão jogados para zero todas as vezes que isso ocorrer. Por fim, os quadros distorcidos são notificados como D'_n e DA'_n. Observe o início da Figura A.2.

A.12 Realinhamento dos Intervalos Ruins

Quando uma seqüência de quadros de fala seguida de um quadro distorcido ocorre acima de um determinado limite, classifica-se como um intervalo de quadros ruim. Na minoria dos casos, há uma injustiça: a medida objetiva prediz grandes quantidades de distorções sobre um número mínimo de quadros ruins devido à má observação do atraso temporal no pré-processamento dos algoritmos avaliadores da qualidade da fala. Como solução, para os denominados "intervalos ruins",

um novo valor de atraso é estimado através da localização da máxima correlação entre o sinal original e o sinal degradado pré-compensado do atraso observado no pré-processamento (veja os blocos "Identificação e Compensação do Atraso" da Figura A.1). Quando uma máxima da correlação cruzada está abaixo de um limiar, conclui-se que o intervalo possui um equilíbrio de ruído (i.e., ruído do original contra ruído do degradado) e o intervalo não é taxado como ruim. Conseqüentemente, o processamento para aquele intervalo é imediatamente interrompido.

Por fim, têm-se os quadros finais distorcidos D_n^n e DA_n^n, cujo são utilizados para computar a qualidade geral percebida. Observe a Figura A.2.

A.13 Agregando os Distúrbios no Tempo

Nesta última etapa, são realizados dois somatórios utilizando a ponderação que enfatiza os distúrbios sonoros L_p citada na Equação 4.1. Somando os quadros considerados distorcidos no tempo, tem-se a pontuação PESQ final (veja a Figura A.2).

A primeira soma é feita a cada 32 ms, envolvendo 20 quadros ($N = 20$) da respectiva distorção (D_n ou DA_n) separadamente e cada uma emprega o somatório ponderado L_6 ($p = 6$). Este somatório também é sobreposto de 50% e nenhuma função janela é utilizada. Denomina-se estes somatórios de "parciais".

Depois que estes dois somatórios de 20 em 20 quadros (parciais) foram realizados, um outro somatório é aplicado, mas agora com $p = 2$ (isto é, L_2) sobre todo o intervalo de voz ativa (não entra os segmentos de silêncio) do arquivo de fala. Denomina-se este de somatório "completo".

O motivo de se utilizar um maior p nos somatórios "parciais" e um menor no somatório "completo" é devido ao fato de quando houver distúrbios locais nos somatórios parciais, este não afete o restante do sinal, ou seja, se um trecho dentro do arquivo de fala está distorcido, a qualidade dos demais permanece intacta.

A.14 Computando a Pontuação PESQ

A pontuação final do PESQ é uma combinação linear do valor ponderado de distorções com o valor ponderado de distúrbios assimétricos. Esta combinação linear foi otimizada num vasto cenário de testes subjetivos, e depois de mapeados, a escala de pontuação do PESQ ficou de -0.5 até 4.5 (-0.5 = pior caso e 4.5 = melhor caso ou nenhuma distorção), embora na maioria dos casos o resultado se assemelha com a pontuação MOS, isto é, ficando entre 1.0 e 4.5. Lembrando que

a escala MOS está fundamentada nos ensaios subjetivos ACR [71], [72]. Observe a Tabela 6.13.

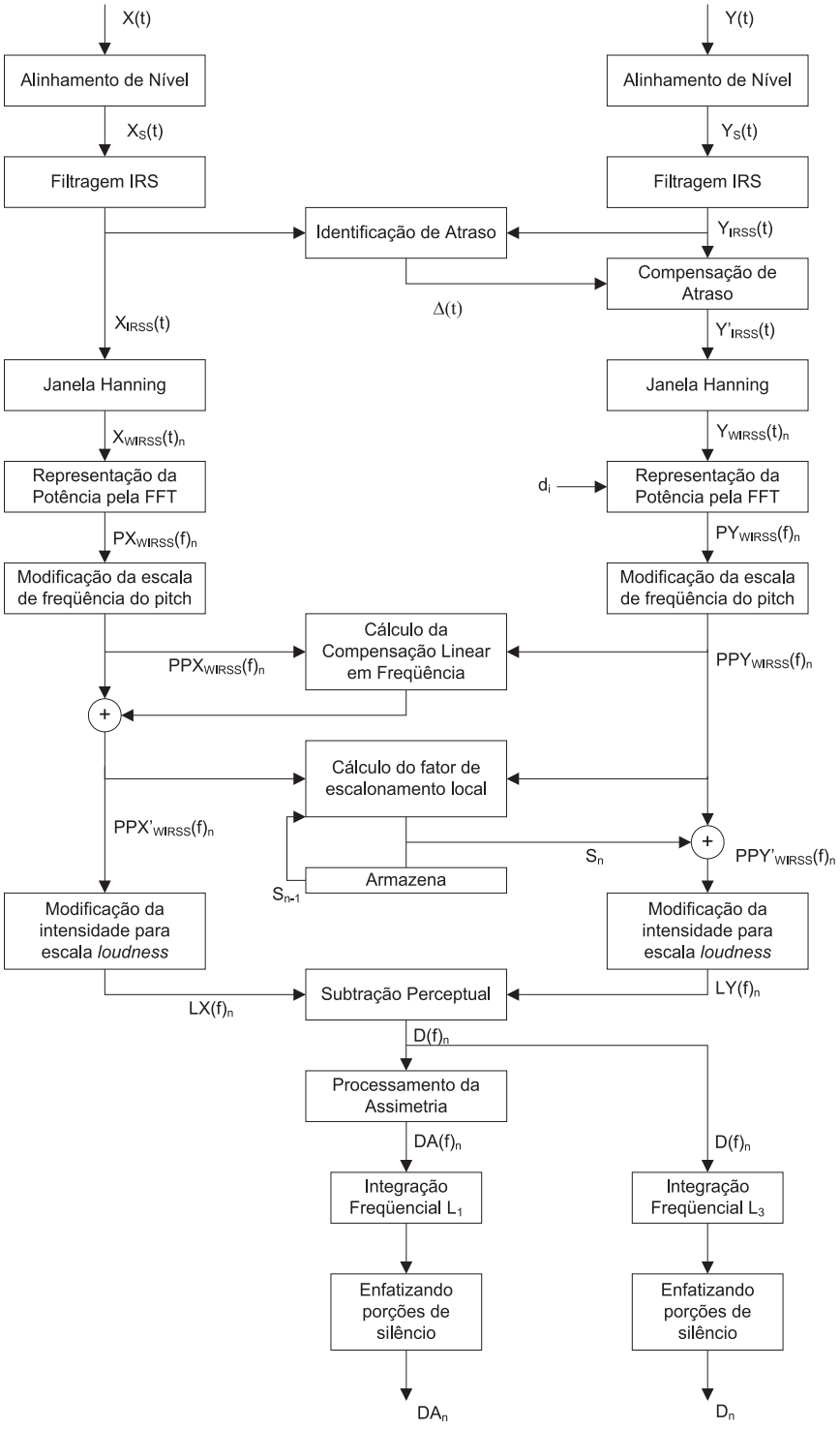

Figura A.1: *Principais componentes do Processamento da Forma de Onda do tipo SNR-dependente.*

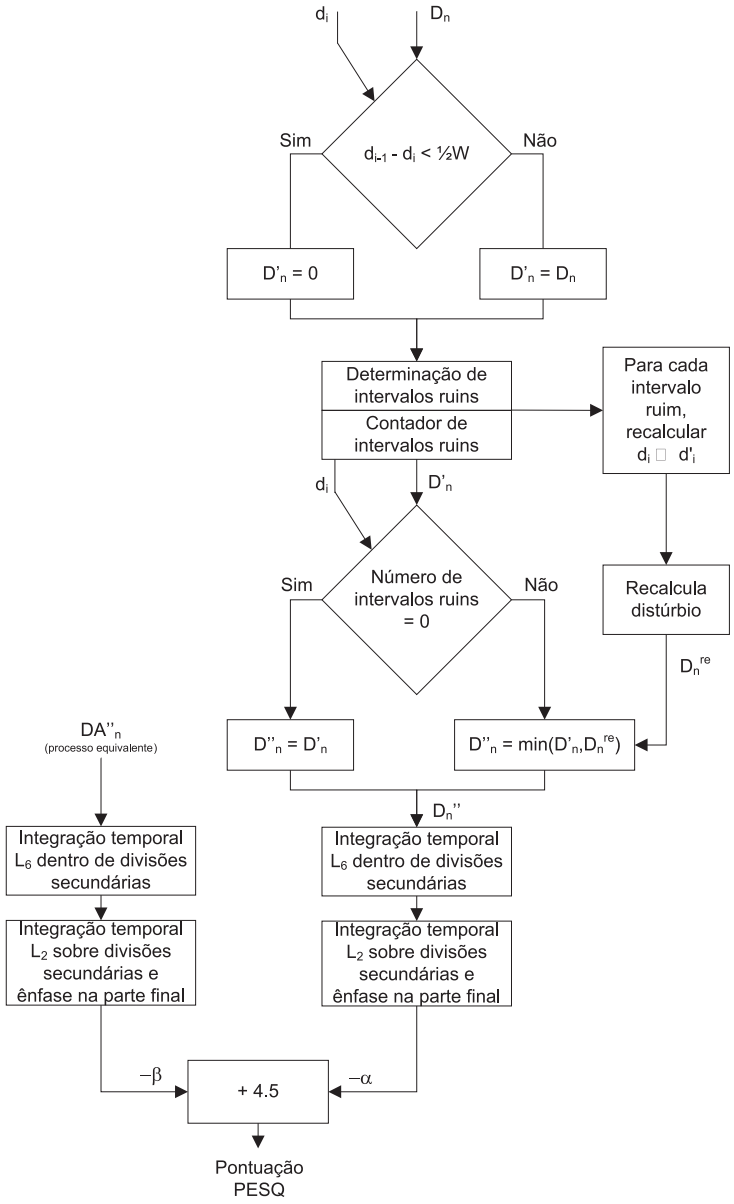

Figura A.2: *Principais componentes do Processamento da Forma de Onda do tipo SNR-dependente. O cálculo do DA''_n é equivalente aos blocos do D''_n apresentado.*

A.15 Desempenho do PESQ

Nesta seção, o desempenho do PESQ é comparado com o do PSQM [65], [74] nas Figuras de A.3 até A.6, utilizando o cálculo da correlação. As figuras mostram a correlação de coeficientes entre cada um dos dois métodos de avaliação e o subjetivo MOS para alguns testes de qualidade audível ACR. A Figura A.3 apresenta 19 testes contendo principalmente codificadores/decodificadores de redes móveis. Na Figura A.4, os resultados de 9 testes predominantemente para redes fixas, são fornecidos. Já a Figura A.5, apresenta 10 testes com VoIP com diferentes tipos de

codificadores/decodificadores. E finalmente, na Figura A.6, são apresentados os resultados de 8 testes conduzidos pelo PESQ (com idiomas diferentes e condições incluindo ruído de fundo) em laboratórios independentes, utilizando dados desconhecidos. Tanto os 22 testes de desempenho do ITU-T publicamente conhecidos em [71] e [72], quanto os 8 testes do PESQ, obtiveram uma média de correlação de 0.935. Este fato reafirma a estabilidade do modelo para os cenários de teste desconhecidos.

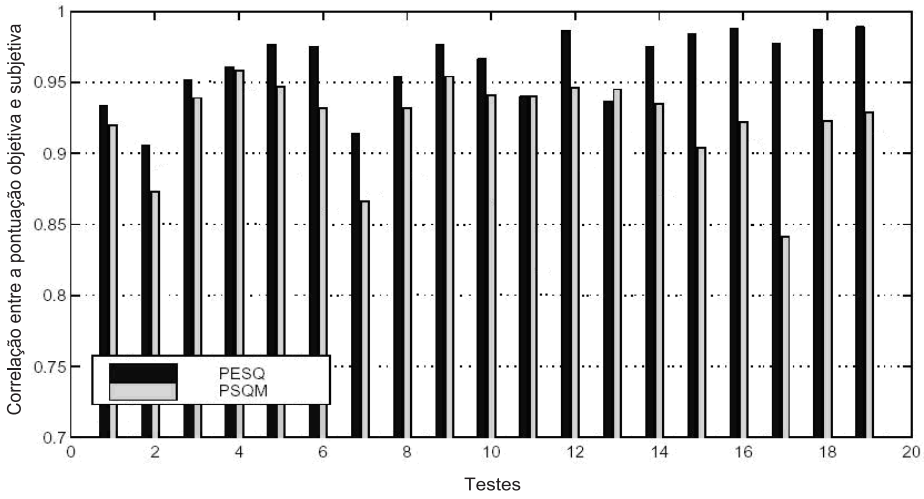

Figura A.3: *Resultado do PESQ e PSQM [65], [74] para o desempenho da rede móvel. Coeficientes de correlação por ensaio, depois de um mapeamento polinomial de terceira ordem.*

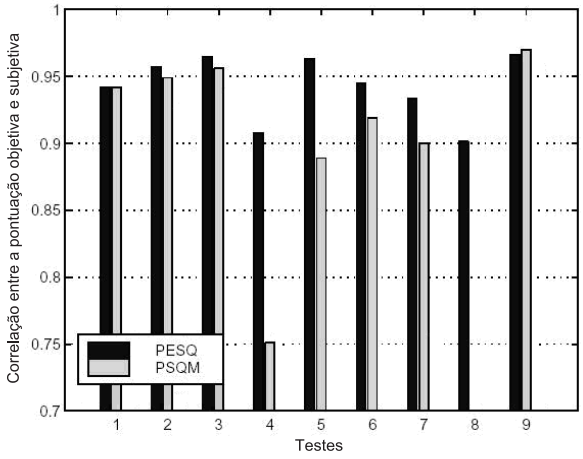

Figura A.4: *Resultado do PESQ e PSQM [65], [74] para o desempenho da rede fixa. Coeficientes de correlação por experimento, depois de um mapeamento polinomial de terceira ordem. A pontuação do teste 8 do PSQM está abaixo do fundo da escala.*

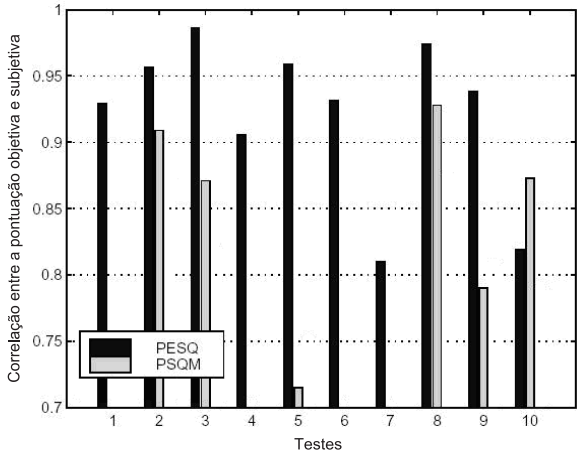

Figura A.5: *Resultado do PESQ e PSQM [65], [74] para o desempenho do VoIP. Coeficientes de correlação por experimento, depois de um mapeamento polinomial de terceira ordem. A pontuação dos testes 1, 4, 6 e 7 para o PSQM, está abaixo do fundo da escala.*

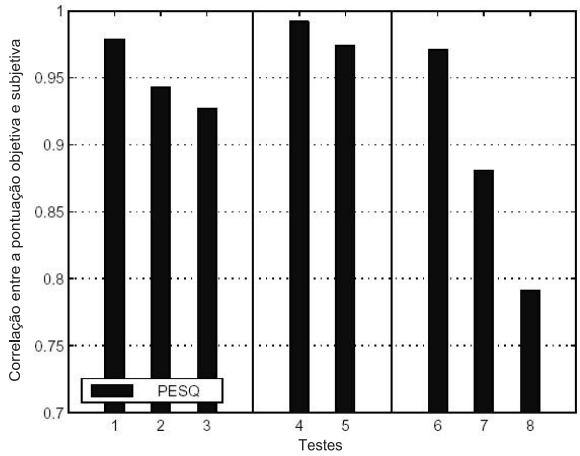

Figura A.6: *Resultados independentes para testes subjetivos desconhecidos (apenas para o PESQ). Coeficientes de correlação por ensaio, depois de um mapeamento polinomial de terceira ordem.*

A.16 Aplicações Atuais do PESQ

O PESQ foi desenvolvido para uma vasta escala de distorções, mas ainda não é o ponto final da avaliação da qualidade da fala. Por exemplo, o modelo psicoacústico não está completamente preparado para avaliar sinais de música, pois há características próprias que o mascaramento desenvolvido não suporta. Nesta seção, será dado um panorama geral das aplicações que o PESQ suporta e onde ele falha.

Logo a seguir, é apresentado um resumo das condições que o PESQ foi testado e aprovado com desempenho aceitável (detalhes mais aprofundados do propósito

do método podem ser encontrados na Recomendação P.862 [66]), e ainda, algumas aplicações dedicadas à prestação de serviços mais modernos:

- Distorções de codificadores: avalia as formas de onda dos codecs em redes já instaladas ou durante a instalação (e.g., G.711, G726, G.727);
- Equipamentos de transmissões e avaliação de erros por perda de pacotes com múltiplos codecs: CELP/codecs híbridos a 4 kbits/s ou acima (e.g., GSM, AMR, CDMA, TDMA, ACELP, VCELP, etc.);
- Deformação do tempo: variação do atraso;
- Ambiente ruidoso: neste caso, deverão estar disponíveis para o PESQ os sinais: limpo e ruidoso;
- Avaliação de sistemas IVR (*Interactive Voice Response*): sistema que responde automaticamente às perguntas feitas pelos consumidores e processadas através do reconhecimento de fala. Podem ser inúteis quando tem-se uma baixo SNR ou conectividade pobre (por exemplo, de ligações móveis). Nestes casos onde a qualidade é pobre, faz sentido para os sistemas IVR reconhecer a deficiência e re-encaminhar a ligação para um atendente real (um indivíduo) ao invés de pedir para o usuário repetir o que foi solicitado, várias e várias vezes;
- Avaliação de sistemas *Voice Recording*: há intenções legais em ter uma cópia de segurança do diálogo de uma ligação telefônica, como nos serviços de transações bancárias ou de negócios. Neste caso, é interessante fazer uma medição da qualidade da voz para alertar o sistema quando a qualidade da voz atingir um patamar de ligação não inteligível e conseqüentemente, tornando a gravação inútil. A melhoria da qualidade, assegura gravações de voz sempre inteligível.

O PESQ não atende os seguintes pontos:

- Eco do locutor, onde o indivíduo ouve a sua própria voz atrasada;
- Uma modulação do lado do locutor, onde o indivíduo ouve sua própria voz distorcida;
- Medições intrusivas, onde apenas o sinal de saída do sistema está disponível para avaliação;
- Música.

A.17 Conclusões

Para avaliações da qualidade dos sinais de fala da banda do telefone (30 - 3400 Hz), o algoritmo PESQ possui um desempenho muito maior do que os primeiros modelos de avaliação de codificadores de voz, principalmente o PSQM P.861. Em fevereiro de 2001, o PESQ substituiu estes modelos tornando-se a nova Recomendação P.862 do ITU-T. As maiores vantagens do PESQ sobre o PSQM são:

- Inclusão de uma compensação perceptual e dinâmica de atraso, a qual permite várias avaliações de distorções no eixo do tempo (veja [68]);
- Inclusão de uma ponderação L_p no tempo, a qual modela corretamente o maior peso que o subjetivo fornece num curto distúrbio sonoro;
- Um melhor modelamento para os efeitos assimétricos, onde há uma diferença de distorção entre as componentes de freqüência que são introduzidas, contra as componentes que são excluídas;
- A habilidade de tratar de forma correta, as distorções da resposta em freqüência;
- Um aprimoramento do escalonamento da potência local, o qual cuida da influência perceptual de acordo com as variações do ganho.

O PESQ foi avaliado sobre uma larga escala de codificadores/decodicadores de fala e testes de redes de telefonia. E conseqüentemente, foram encontrados valores da predição da qualidade da fala, com correlação entre a pontuação objetiva e subjetiva de 0,935, tornando-o um representante significativo do histórico evolutivo dos métodos que avaliam a qualidade da fala.

Anexo B

Valores Taxa de Acerto (%) e PESQ-MOS

Neste Anexo, para cada algoritmo estão expostos todos os valores da taxa de reconhecimento e do índice PESQ-MOS da fala, separados por ruído e TESTE.

B.1 Cenário: Treinamento em múltiplas condições submetido ao TESTE-A

Tabela B.1: *Tabela do ruído **Subway** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-A.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.68	97.79	98.89	98.86	98.65	98.99
20	97.61	98.22	98.13	98.13	97.70	98.07
15	96.47	97.08	97.21	97.11	96.56	97.24
10	94.44	95.58	94.81	94.90	94.81	95.33
5	88.36	90.42	88.73	88.85	87.93	90.08
0	66.90	72.34	70.74	73.26	70.10	74.85
-5	26.13	37.37	36.63	41.11	35.83	43.54
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7555	3.0668	2.9999	3.0874	3.0015	3.0301
15	2.4444	2.7357	2.7542	2.8118	2.7182	2.7452
10	2.1307	2.3647	2.4752	2.5140	2.4049	2.4290
5	1.8326	1.9889	2.1573	2.2090	2.0666	2.1194
0	1.5728	1.5600	1.7730	1.8451	1.6887	1.7732
-5	1.3333	1.1736	1.3747	1.4704	1.3126	1.4339

Tabela B.2: *Tabela do ruído **Babble** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-A.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.52	97.88	98.31	98.49	98.28	98.58
20	97.73	98.19	98.04	97.88	97.85	97.88
15	97.04	97.28	97.85	97.25	97.28	96.98
10	95.28	95.56	95.92	94.83	95.41	94.56
5	87.55	88.54	89.54	87.00	87.30	87.30
0	62.15	64.93	67.02	59.67	61.52	57.71
-5	27.15	28.39	30.53	17.78	22.64	15.66
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.9097	3.1638	3.0352	3.1114	3.0232	3.084
15	2.6167	2.8350	2.8103	2.8372	2.7373	2.7977
10	2.3147	2.4931	2.5691	2.5602	2.4416	2.5077
5	1.9973	2.1053	2.2476	2.2322	2.0957	2.1674
0	1.6921	1.6826	1.8612	1.8584	1.6894	1.8037
-5	1.3911	1.2813	1.4680	1.4265	1.2670	1.3946

Tabela B.3: *Tabela do ruído **Car** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-A.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.39	97.41	98.93	98.54	98.39	98.54
20	98.03	98.36	98.48	98.27	98.09	98.18
15	97.61	98.24	98.15	98.09	98.06	98.15
10	95.74	97.32	97.05	96.66	96.45	96.69
5	87.80	95.53	93.29	92.81	91.56	93.05
0	53.44	79.84	81.06	80.50	77.45	81.39
-5	20.58	41.63	46.02	47.06	39.37	48.37
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7931	3.1842	3.0463	3.1624	3.1108	3.1143
15	2.4932	2.8536	2.8273	2.9081	2.8399	2.8489
10	2.1807	2.5077	2.5753	2.6384	2.5504	2.5638
5	1.8781	2.1474	2.2921	2.3473	2.2285	2.2535
0	1.6184	1.7514	1.9500	2.0056	1.8401	1.9148
-5	1.3856	1.3522	1.5787	1.6115	1.4139	1.5587

Tabela B.4: *Tabela do ruído **Exhibition** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-A.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.49	98.12	99.01	98.92	98.95	98.92
20	97.41	97.87	97.81	98.09	98.18	98.33
15	96.67	97.72	97.25	97.13	97.22	96.98
10	94.11	95.40	95.06	95.12	94.97	95.50
5	87.60	90.65	90.03	88.28	88.00	89.05
0	64.36	76.30	76.98	71.34	70.38	73.22
-5	24.34	43.04	47.18	41.07	38.75	43.66
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7265	3.1210	3.0127	3.0497	2.989	2.9863
15	2.4043	2.787	2.7867	2.7677	2.6944	2.6928
10	2.0950	2.4574	2.5447	2.4846	2.3932	2.4013
5	1.8045	2.0721	2.2369	2.1282	2.0114	2.0513
0	1.5443	1.6628	1.8704	1.7495	1.6082	1.6981
-5	1.2886	1.2686	1.4876	1.3554	1.1847	1.3607

B.2 Cenário: Treinamento em múltiplas condições submetido ao TESTE-B

Tabela B.5: *Tabela do ruído **Restaurant** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-B.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.68	97.79	98.89	98.86	98.65	98.99
20	96.87	98.00	98.22	97.85	98.19	97.88
15	95.30	96.35	96.96	96.81	97.33	96.50
10	91.96	92.69	93.86	92.97	93.46	92.51
5	83.54	84.46	85.02	81.73	82.68	79.98
0	59.29	60.64	63.95	51.27	55.48	50.14
-5	25.51	24.38	25.88	16.12	21.49	14.31
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.9402	3.1715	3.0698	3.1170	3.0251	3.0932
15	2.6413	2.8363	2.8203	2.8309	2.7304	2.8002
10	2.3342	2.4909	2.5496	2.5281	2.4194	2.4925
5	2.0083	2.1009	2.2256	2.1723	2.0452	2.1284
0	1.6834	1.6684	1.8318	1.7747	1.6531	1.7514
-5	1.3378	1.2681	1.4267	1.3544	1.2370	1.3442

Tabela B.6: *Tabela do ruído **Street** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-B.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.52	97.88	98.31	98.49	98.28	98.58
20	97.58	97.58	97.94	97.49	97.19	97.70
15	96.31	96.86	96.95	96.83	96.31	96.80
10	94.35	94.56	94.86	94.53	93.86	94.65
5	85.61	88.60	88.54	87.88	85.76	88.06
0	61.34	69.71	72.04	69.92	64.30	70.22
-5	27.60	37.15	37.97	36.79	30.02	36.97
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7997	3.1327	2.9985	3.0879	3.0076	3.0488
15	2.4885	2.7978	2.7825	2.8142	2.7234	2.7673
10	2.1696	2.4523	2.5441	2.5303	2.4273	2.4680
5	1.8592	2.0615	2.2309	2.2113	2.0909	2.1460
0	1.5215	1.6429	1.8534	1.8374	1.6902	1.7888
-5	1.2811	1.2345	1.4494	1.4305	1.2708	1.4051

Tabela B.7: *Tabela do ruído **Airport** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-B.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.39	97.41	98.93	98.54	98.39	98.54
20	97.44	97.97	98.09	97.79	98.30	97.94
15	96.12	97.20	97.35	96.57	97.46	96.66
10	93.29	94.72	94.54	94.09	94.33	94.51
5	86.25	88.85	89.41	86.64	86.79	85.89
0	65.11	73.43	72.05	67.94	67.07	66.03
-5	29.41	37.61	37.64	28.60	29.08	26.75
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.9393	3.2212	3.0683	3.1533	3.0869	3.1247
15	2.6482	2.8938	2.8433	2.8838	2.8035	2.8485
10	2.3448	2.5366	2.5783	2.5768	2.4841	2.5362
5	2.0261	2.1637	2.2725	2.2633	2.1442	2.2106
0	1.7269	1.7558	1.9140	1.8912	1.7506	1.8464
-5	1.3991	1.3405	1.5012	1.4666	1.3191	1.4609

Tabela B.8: *Tabela do ruído **Train-station** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-B.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.49	98.12	99.01	98.92	98.95	98.92
20	97.01	98.15	98.64	98.36	98.77	98.55
15	95.53	97.41	97.99	97.69	97.69	97.50
10	92.87	95.31	95.77	95.25	95.68	95.28
5	83.52	88.46	89.23	87.29	88.09	87.63
0	56.12	72.23	74.58	69.82	69.36	69.82
-5	21.07	41.47	44.74	37.55	32.77	35.08
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.8731	3.2035	3.0740	3.1657	3.1033	3.1323
15	2.5772	2.8769	2.8514	2.9034	2.8293	2.8597
10	2.2627	2.5285	2.5899	2.6297	2.5346	2.5717
5	1.9369	2.1486	2.2879	2.3131	2.1954	2.2469
0	1.6377	1.7579	1.9407	1.9700	1.8125	1.9099
-5	1.3679	1.3553	1.5597	1.5703	1.3990	1.5439

B.3 Cenário: Treinamento em múltiplas condições submetido ao TESTE-C

Tabela B.9: *Tabela do ruído **Subway-MIRS** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-C.*

SNR [dB]	**Nenhum Processo**	**NR-WI008**	**NR-WI008 + SWP**	**EMSR + NMT**	**NMT-PSS**	**EMSR**
clean	98.50	97.61	98.77	98.59	98.59	98.71
20	97.30	98.28	97.88	97.85	97.39	97.82
15	96.35	96.99	96.50	97.02	96.28	96.78
10	93.34	93.95	93.28	94.17	93.15	93.92
5	82.41	85.57	83.14	87.17	84.89	87.69
0	46.82	56.65	53.61	65.49	59.59	67.18
-5	18.91	23.76	22.54	28.19	22.94	33.53
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.6989	2.9740	2.8833	2.9801	2.8855	2.9365
15	2.3914	2.6270	2.6147	2.7016	2.5950	2.6478
10	2.1079	2.2851	2.3296	2.4054	2.2772	2.3533
5	1.8668	1.8982	2.0046	2.0897	1.9343	2.0427
0	1.6536	1.5259	1.6456	1.7949	1.6353	1.7574
-5	1.4399	1.2265	1.3341	1.5299	1.3733	1.4703

Tabela B.10: *Tabela do ruído **Street-MIRS** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em múltiplas condições e submetido ao TESTE-C.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	98.58	97.76	98.58	98.40	98.13	98.46
20	96.55	97.55	97.67	97.43	97.28	97.46
15	95.53	96.70	96.86	96.16	95.74	96.37
10	92.50	94.07	93.50	93.65	92.32	93.59
5	82.53	83.77	83.86	85.13	82.50	85.34
0	54.44	58.89	59.01	64.12	59.95	64.42
-5	24.24	28.02	29.26	27.93	25.82	29.38
	PESQ	**PESQ**	**PESQ**	**PESQ**	**PESQ**	**PESQ**
clean	4.5000	4.5000	4.5000	4.5000	4.5000	4.5000
20	2.7418	3.0672	2.9183	3.0045	2.9129	2.9712
15	2.4372	2.7248	2.6633	2.7196	2.6187	2.6777
10	2.1295	2.3769	2.4084	2.4247	2.3059	2.3762
5	1.8481	2.0013	2.0959	2.1186	1.9894	2.0731
0	1.5582	1.5905	1.7203	1.7586	1.6091	1.7289
-5	1.2902	1.2456	1.3674	1.4078	1.2558	1.3988

B.4 Cenário: Treinamento em condição limpa submetido ao TESTE-A

Tabela B.11: *Tabela do ruído **Subway** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-A.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	98.93	98.37	99.02	98.96	98.86	98.93
20	97.05	95.92	97.64	96.04	92.66	97.05
15	93.49	92.88	95.46	91.86	85.63	95.06
10	78.72	87.07	90.21	81.89	70.89	88.82
5	52.16	74.27	78.45	64.17	48.14	76.11
0	26.01	52.23	54.74	39.88	25.70	51.40
-5	11.18	23.18	23.09	17.65	14.61	25.39

Tabela B.12: *Tabela do ruído **Babble** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-A.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	99.00	98.64	98.55	99.09	98.97	99.09
20	90.15	96.34	97.70	94.47	92.47	95.68
15	73.76	93.20	95.92	85.46	82.13	88.88
10	49.43	84.55	90.99	68.80	59.98	70.53
5	26.81	67.68	78.02	43.26	33.34	45.01
0	9.28	37.00	50.88	16.81	15.27	18.23
-5	1.57	8.86	17.08	-3.30	8.16	-0.82

Tabela B.13: *Tabela do ruído **Car** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-A.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	98.96	98.45	98.93	98.66	98.93	98.93
20	97.41	97.70	97.97	97.11	93.86	97.73
15	90.04	97.14	96.69	95.35	86.58	96.69
10	67.01	94.12	94.04	89.29	69.61	91.83
5	34.09	85.24	86.37	73.37	43.66	77.66
0	14.46	60.27	68.00	45.27	18.91	49.93
-5	9.39	27.68	32.36	15.96	10.41	18.55

Tabela B.14: *Tabela do ruído **Exhibition** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-A.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	99.20	98.46	99.20	99.11	99.17	99.07
20	96.39	97.10	97.72	95.99	91.30	96.39
15	92.04	94.38	96.36	91.89	79.88	93.46
10	75.66	89.20	92.97	83.49	59.21	86.45
5	44.83	76.33	85.56	61.56	34.40	67.39
0	18.05	53.93	69.08	34.25	17.06	38.66
-5	9.60	25.89	37.64	13.08	9.60	16.69

B.5 Cenário: Treinamento em condição limpa submetido ao TESTE-B

Tabela B.15: *Tabela do ruído **Restaurant** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-B.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	98.93	98.37	99.02	98.96	98.86	98.93
20	89.99	95.09	97.61	92.39	92.17	94.14
15	76.24	89.50	94.60	82.62	79.77	84.77
10	54.77	80.63	89.10	66.47	60.24	68.56
5	31.01	62.54	75.90	42.89	34.11	45.07
0	10.96	36.57	49.31	17.10	15.75	19.37
-5	3.47	8.26	16.64	-0.03	9.43	2.00

Tabela B.16: *Tabela do ruído **Street** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-B.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	99.00	98.64	98.55	99.09	98.97	99.09
20	95.74	96.67	97.43	96.28	92.59	97.10
15	88.45	95.04	95.50	92.38	84.52	94.23
10	67.11	88.91	91.69	82.32	67.59	84.89
5	38.45	76.51	82.32	64.99	41.81	68.26
0	17.84	52.24	59.67	39.69	21.22	41.93
-5	10.46	26.84	30.02	17.53	11.37	17.35

Tabela B.17: *Tabela do ruído **Airport** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-B.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	98.96	98.45	98.93	98.66	98.93	98.93
20	90.64	96.45	97.70	94.48	93.50	95.68
15	77.01	94.96	96.06	87.21	84.46	89.56
10	53.86	88.70	91.83	73.07	65.76	75.93
5	30.33	74.35	80.91	50.52	39.93	51.71
0	14.41	49.51	57.98	26.78	19.83	27.77
-5	8.23	19.15	26.19	5.37	10.74	6.59

Tabela B.18: *Tabela do ruído **Train-station** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-B.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	99.20	98.46	99.20	99.11	99.17	99.07
20	94.72	96.39	97.99	96.45	94.23	97.10
15	83.65	94.26	96.36	92.38	85.93	93.77
10	60.29	90.68	92.63	83.12	69.58	86.42
5	27.92	79.23	82.97	62.78	39.68	67.42
0	11.57	56.90	63.47	34.59	18.73	38.94
-5	8.45	25.58	34.25	12.00	10.58	12.68

B.6 Cenário: Treinamento em condição limpa submetido ao TESTE-C

Tabela B.19: *Tabela do ruído **Subway-MIRS** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-C.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	99.14	98.31	98.93	99.08	98.83	99.08
20	93.46	94.04	97.24	95.27	92.35	96.93
15	86.77	89.22	94.84	89.56	86.15	93.64
10	73.90	80.87	89.22	80.44	70.43	87.41
5	51.27	64.14	74.92	62.88	44.30	74.70
0	25.42	37.61	44.46	37.89	19.77	51.21
-5	11.82	16.03	18.48	18.24	10.07	27.42

Tabela B.20: *Tabela do ruído **Street-MIRS** com valores da Taxa de Acerto (%) e PESQ-MOS por SNR, para todos os algoritmos de realce com sistema treinado em condição limpa e submetido ao TESTE-C.*

SNR [dB]	Nenhum Processo	NR-WI008	NR-WI008 + SWP	EMSR + NMT	NMT-PSS	EMSR
clean	98.97	98.49	98.52	99.03	98.85	99.06
20	95.13	95.98	96.92	96.07	91.78	96.31
15	88.91	92.93	95.04	93.08	85.34	93.92
10	74.43	85.46	88.72	84.82	67.29	85.49
5	49.21	71.64	74.85	68.35	40.05	69.89
0	22.91	43.23	46.89	43.23	18.83	45.44
-5	11.15	22.01	23.55	20.95	9.58	19.17

Anexo C

Taxas de Reconhecimento (%) do Algoritmo WI008

Este Anexo apresenta os valores da taxa de reconhecimento do algoritmo WI008 por inteiro, isto é, considerando todos os seus blocos funcionais originais padronizados pelo ETSI. Lembrando que os valores da taxa de acerto deste mesmo algoritmo apresentados no Capítulo 6 consideram ora apenas o bloco **Redução de Ruído**, ora os blocos **Redução de Ruído + Processamento da Forma de Onda com SNR-dependente**.

O motivo, como é mencionado no mesmo capítulo, é a passagem do sinal de fala para o domínio *cepstral* realizada pelo algoritmo WI008, impossibilitando analisar paralelamente a qualidade perceptual alcançada por cada bloco através da ferramenta PESQ (que trabalha com locuções no formato WAVE).

Tabela C.1: *TESTE-A - Taxa de acerto em (%) para treinamento com múltiplas condições.*

SNR/dB	Subway	Babble	Car	Exhibition	Média
clean	99.02	98.82	98.99	99.14	98.99
20	98.62	98.58	98.54	98.24	98.50
15	97.54	97.91	98.42	97.56	97.86
10	95.36	96.07	97.38	95.34	96.04
5	91.43	90.21	93.83	90.10	91.39
0	75.28	68.71	80.67	76.00	75.17
-5	39.85	30.05	40.41	44.99	38.83
Média 0 a 20dB	91.65	90.30	93.77	91.45	91.79

Tabela C.2: *TESTE-B - Taxa de acerto em (%) para treinamento com múltiplas condições.*

SNR/dB	Restaurant	Street	Airport	Train-station	Média
clean	99.02	98.82	98.99	99.14	98.99
20	98.10	98.13	98.63	98.80	98.42
15	96.93	97.85	98.03	97.69	97.63
10	94.81	95.59	95.94	96.05	95.60
5	87.14	90.39	91.44	90.16	89.78
0	65.58	73.85	75.78	74.05	72.32
-5	28.52	38.88	40.95	41.75	37.53
Média 0 a 20dB	88.51	91.16	91.96	91.35	90.75

Tabela C.3: *TESTE-C - Taxa de acerto em (%) para treinamento com múltiplas condições.*

SNR/dB	Subway (MIRS)	Street(MIRS)	Média
clean	98.99	98.85	98.92
20	98.07	97.94	98.00
15	97.54	97.73	97.64
10	95.58	95.31	95.45
5	88.95	87.52	88.24
0	66.99	65.63	66.31
-5	30.43	30.62	30.53
Média 0 a 20dB	89.43	88.83	89.13

Tabela C.4: *TESTE-A - Taxa de acerto em (%) para treinamento em condições limpas.*

SNR/dB	Subway	Babble	Car	Exhibition	Média
clean	99.08	99.03	99.05	99.23	99.10
20	97.85	98.25	98.36	97.81	98.07
15	96.38	96.74	97.52	96.70	96.84
10	92.26	91.99	95.29	92.59	93.03
5	83.88	80.68	88.73	84.05	84.34
0	61.93	51.12	66.06	63.50	60.65
-5	31.07	18.95	29.82	33.20	28.26
Média 0 a 20dB	86.46	83.76	89.19	86.93	86.59

Tabela C.5: *TESTE-B - Taxa de acerto em (%) para treinamento em condições limpas.*

SNR/dB	Restaurant	Street	Airport	Train-station	Média
clean	99.08	99.03	99.05	99.23	99.10
20	98.07	97.64	98.42	98.43	98.14
15	95.33	96.58	97.05	96.76	96.43
10	89.87	92.74	93.26	96.86	93.18
5	76.05	83.28	83.54	84.20	81.77
0	50.26	59.70	60.24	62.23	58.11
-5	18.39	29.23	27.32	29.56	26.13
Média 0 a 20dB	81.92	85.99	86.50	87.70	85.53

Tabela C.6: *TESTE-C - Taxa de acerto em (%) para treinamento em condições limpas.*

SNR/dB	Subway (MIRS)	Street(MIRS)	Média
clean	99.02	99.03	99.03
20	97.36	97.67	97.52
15	95.30	95.74	95.52
10	90.33	90.75	90.54
5	78.88	78.48	78.68
0	52.59	52.12	52.36
-5	25.15	26.12	25.64
Média 0 a 20dB	82.89	82.95	82.92

Referências Bibliográficas

[1] SMOLDERS, J., CLAES, T., SABLON, G., VAN COMPERNOLLE, D., "On the Importance of the Microphone Position for Speech Recognition in the Car". *IEEE Trans. Acoust., Speech and Signal Processing*, pp.I/429-I/432, 1994.

[2] WIDROW, B., GLOVER, J., MCCOOL, J., KAUNITZ, J., WILLIAMS, C., HEARN, R., ZEIDLER, J., DONG, E. and GOODLIN, R., "Adaptative Noise Canceling: Principles and Applications". *Proc. IEEE*, 63(12):1151-1162, 1975.

[3] POWELL, G., DARLINGTON, P., and WHEELER, P., "Practical Adaptive Noise Reduction in the Aircraft Cockpit Environment". *Proceedings of the ICASSP*, pp.173-176, 1987.

[4] NAKADAI, Y. and SUGAMURA, N., "A Speech Recognition Method for Noise Enviroments Using Dual Inputs". *In ICSLP*, pp. 1141-1144, 1990.

[5] VAN COMPERNOLLE, D., MA, W., XIE, F., and VAN DIEST, M., "Speech Recognition in Noisy Enviroments with the Aid of Microphone Arrays". *Speech Communication*, 9(5-6):433-442.

[6] FROST III, O., "An Algorithm for Linearly Constrained Adaptive Array Processing". *Proc. IEEE*, 60(8):926-935, 1972.

[7] FARRELL, K., MAMMONE, R., and FLANAGAN, J., "Beamforming Microphone Arrays for Speech Enhancement". *Proceedings of the ICASSP*, pp. 285-288, 1992.

[8] GRIFFITHS, L. and JIM, C., "An Alternative Aprroach to Linearly Constrained Adaptive Beamforming". *IEEE Trans. Antennas Propag.*, AP-30:27-34, 1982.

[9] SLYH, R. and MOSES, R., "Microphone Array Speech Enhancement in Overdetermined Signal Scenarios". *Proceedings of the ICASSP*, pp. II.347-II.350, 1993.

[10] SULLIVAN, T. and STERN, R., "Multi-microphone Correlation-based Processing for Robust Speech Recognition". *Proceedings of the ICASSP*, pp. II.91-II.94, 1993.

[11] ZANGI, K., "A New Two-sensor Active Noise Cancellation Algorithm". *Proceedings of the ICASSP*, pp. II.351-II.354, 1993.

[12] HERMANSKY, H., "Perceptual Linear Predictive (PLP) Analysis of Speech". *Journal of the Acoustical Society of America* 87, pp. 1738-1752, Apr. 1990.

[13] HERMANSKY, H., MORGAN, N., BAYYA, A. and KOHN, P., "RASTA-PLP Speech Analysis Technique". *Proceedings of the ICASSP*, pp. I.121-I-124, 1992.

[14] HERMANSKY, H., MORGAN, N. and HIRSCH, H.,"Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing". *Proceedings of the ICASSP*, pp. II.83-II-86, 1993.

[15] HERMANSKY, H. and MORGAN, N., "RASTA Processing of Speech". *IEEE Trans. on Speech and Audio Processing*, 2(4):578-589, 1994.

[16] MORGAN, N., HERMANSKY, H., "RASTA Extensions: Robustness to Additive and Convolutional Noise". *ESCA Workshop "Speech Processing in Adverse Conditions"*, 1992.

[17] KOEHLER, J., MORGAN, N., HERMANSKY, H., HIRSCH, H. G., and TONG, G., "Integrating RASTA-PLP Into Speech Recognition". *IEEE Trans. Acoust., Speech and Signal Processing*, vol. 1, pp. 421-424, Adelaide, Australia, 1994.

[18] JUNQUA, J. C., HATON, J. P., "Robustness in Automatic Speech Recognition - Fundamentals and Applications". *Kluwer Academic Publishers*, Norwell, Massachusetts, 1996.

[19] SHEN, J-L., HWANG, W-L. and LEE, L-S., "Robust Speech Recognition Features Based on Temporal Trajectory Filtering of Frequency Band Spectrum". *Proceedings of the ICASSP*, pp. II.50-II-54, 1995.

[20] ACERO, A., "Acoustical and Environment Robustness in Automatic Speech Recognition". *Ph.D. thesis*, Carnegie Mellon University, 1990.

[21] ACERO, A. and STERN, R., "Envoronment Robustness in Automatic Speech Recognition". *Proceedings of the ICASSP*, pp. 849-852, 1990.

[22] LIU, F.-H., ACERO, A. and STERN, R., "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering". *Proceedings of the ICASSP*, pp. I.257-I.260, 1992.

[23] LIU, F.-H., STERN, R., ACERO, A. and MORENO, P., "Environment Normalization for Robust Speech Recognition Using Cepstral Normalization". *Proceedings of the ICASSP*, pp. II.61-II.64.

[24] LIU, F.-H., STERN, R., HUANG, X. and ACERO, A., "Efficient Cepstral Normalization for Robust Speech Recognition". *In ARPA Human Language Technology Workshop*, pp. 69-74, 1993.

[25] STERN, R., LIU, F.-H., MORENO, P. and ACERO, A., "Signal Processing for Robust Speech Recognition". *Proceedings of the ICASSP*, pp. 1027-1030, 1994.

[26] COHEN, J. R., "Appication of an Auditory Model to Speech Recognition". *Journal of the Acoustical Society of America*, 85(6):2623-2629, 1989.

[27] GHITZA, O., "Temporal Non-place Information in the Auditory-nerve Firing Patterns as a Front End for Speech Recognition in a Noisy Environment". *Journal of Phonetics*, 16(1):109-124, 1988.

[28] LYON, R. F., "A Computational Model of Filtering, Detection and Compression in the Cochlea". *IEEE Trans. Acoust., Speech and Signal Processing*, pp. 1282-1285, Institute of Electrical and Electronic Enginners, 1982.

[29] SENEFF, S., "A Joint Synchrony/Mean-rate Model of Auditory Speech Processing". *Journal of Phonetics*, 16(1):55-76, 1988.

[30] PATTERSON, R. D., ROBINSON, K., HOLDSWORTH, D., ZHANG, C. and ALLERHAND, M., "Complex Sounds and Auditory Images". *In Auditory Physiology and Perpection*, pp. 429-446, Pergamon Press, 1991.

[31] COLE, A., R., MARIANI, J., USZKOREIT, H., ZAENEN, A., ZUE, V., "Survey of the State of the Art in Human Language Technology". *Directorate XIII-E of the Commission of the European Communities, Center for Spoken Language Understanding*, Oregon Graduate Institute, Nov. 1995.

[32] LEONARD, G., G., "A Database for Speaker Independent Digit Recognition". *Proceedings of the ICASSP*, Vol.3, p.42.11, 1948.

[33] ITU Recommendation G.712, "Transmission Performance Characteristic of Pulse Code Modulation Channels". Nov. 1996.

[34] ETSI-SMG Technical Specification, "European Digital Cellular Telecommuncation System (Phase 1), Tranmission Planning Aspects for the Speech Service in GSM PLMN System". GSM03.50, versaion 3.4.0, Jul 1994.

[35] ITU Recommendation P.56, "Objective Measurement of Active Speech Level". Mar. 1993.

[36] YOUNG, S., EVERMANN, G., KERSHAW, D., MOORE, G., ODELL J., OLLASON, D., VALTCHEV, V. and WOODLAND, P., "HTK Book (for HTK Version 3.1)". Cambrigde University Engineering Departament, Dec. 2001.

[37] ABRANCHES, L. K. S. and SILVA, F. J. F., "Speech Enhancement System Based on Nonlinear Spectral Attenuation Using a Noise Masking Threshold". 6^{th} *IASTED International Conference on Signal and Image Processing - SIP2004*, Honolulu, Hawaii, Aug. 2004.

[38] BRILINGER, D. R., "Time Series Data Analysis and Theory". San Francisco: Holden-Day, 1981.

[39] BOLL, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27*, pp. 113-120, Apr. 1979.

[40] VIRAG, N., "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System". *IEEE Trans. on Speech and Audio Processing*, vol. 7, n. 2, pp. 126-137, March 1999.

[41] BEROUTI, M., SCHWARTZ, R. and MAKHOUL, J., "Enhancement of Speech Corrupted by Acoustic Noise". *Proceedings of the ICASSP*, Washington, DC, vol. 4, pp. 208-211, Apr.1979.

[42] EPHRAIM, Y. and MALH, D., "Speech Enhancement Using Optimal Nonlinear Spectral Amplitude Estimation". *IEEE Trans. Conf. Acoust. Speech, Signal Processing*, pp. 1118-1121, Apr. 1983.

[43] EPHRAIM, Y. and MALH, D., "Speech Enhancement Using Minimum Mean-Square Error Short-Time Spectral Amplitude Estimation". *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-32, n. 6, pp. 1109-1121, Dec. 1984.

[44] EPHRAIM, Y. and MALH, D., "Speech Enhancement Using Minimum Mean-Square Error log-Spectral Estimation". *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-33, n. 2, pp. 443-445, Apr. 1985.

[45] JONHSTON, J. D., "Transform Coding of Audio Signal Using Perceptual Noise Criteria". *IEEE Journal. Select. Areas Commum*, vol. 6, pp. 314-323, Feb. 1988.

[46] International Organization for Standardization, ISO/IEC 11172-3: Information Technology - Coding of Moving Picture and Associated Audio for Digital Storage Media at Up to About 1.5 Mbits/s - Part 3: Audio, [S.L.], 1993.

[47] AMBIKAIRAJAH, E., DAVIS, A. G. and WONG, T. K., "Auditory Masking and MPEG-1 Audio Compression". *Journal of the Acoustical Society of America*, pp. 165-175, 1997.

[48] TSOUKALAS, D., PARASKEVASA, M. and MORJOPOULOS, J., "Speech Enhancement Using Psicho-acoustic Criteria". *Proceedings of the ICASSP*, Minneapolis, MN, pp. 359-361, Apr. 1993.

[49] USAGAWA, T., IWATA, M. and EBATA, M., "Speech Parameter Extraction in Noise Environment Using a Masking Model". *Proceedings of the ICASSP*, Adelaide, Australia, vol. II, pp. 81-84, Apr. 1994.

[50] FLETCHER, H., "Auditory Patterns, Review of Modern Physics". vol. 12, pp. 47-65, 1940.

[51] ZWICKER, E. and FASTL, H., "Psychoacoustic". Springer, 2nd ed., 1999.

[52] SCHROEDER, M. R., ATAL, B. S. and HALL, J. L., "Optimizing Digital Speech Coders by Exploting Masking Properties of the Human Ear". *Journal of Acoustic Soc. of America*, vol. 66, pp. 1647-1652, Dec. 1979.

[53] SCHARF, B., "Foundations of Modern Auditory Treory". Cap. 5, New York Academic, 1970.

[54] SINHA, D. and TEWFIK, A. H., "Low Bit Rate Transparent Audio Compression Using Adapted Walelets". *IEEE Trans. Signal Processing*, vol. 41, pp. 3463-3479, Dec. 1993.

[55] TERHARDT, E., STOLL, G. and SEEWANN, M., "Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals". *Journal of the Acoustical Society of America*, vol, 71, pp. 679-688, Mar. 1982.

[56] AGARWAL, A., CHENG, Y.M., "Two-Stage Mel-Warped Wiener Filter for Robust Speech Recognition". *Proc. ASRU'99*, 1999.

[57] ETSI standard doc. "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm". ETSI ES 202 050 v0.1.0 (2002-04), Apr. 2002.

[58] ETSI standard doc. "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms". ETSI ES 201 108 v1.1.2(2000-04)

[59] HIRSCH, H.-G., PEARCE, D., "The Aurora Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions". *ISCA ITRW ASR 2000*, Sep. 2000.

[60] MACHO, D., CHENG, Y.M., "Robust Wideband ASR Front-end Based on a Extention of Narrowband Robust Front-end". *Proc. ICSLP'00*, 2000

[61] MACHO, D., CHENG, Y.M., "Use of Voicing Information to Improve the Robustness of the Spectral Parameter Set". *Proc. ICSLP'00*, 2000.

[62] MAUUARY, L., "Blind Equalization in the Cepstral Domain for Robust Telephone based Speech Recognition". *Proc. EUSPICO'98*, Vol. 1, pp. 359-363, 1998.

[63] NOÉ et al., "Noise Reduction for Noise Robust Feature Extraction for Distributed Speech Recognition". *Proc. Eurospeech'01*, 2001.

[64] JUANG, B. H., RABINER, L. R., "Fundamentals of Speech Recognition". Ed. Prentice-Hall, Signal Processing Series, New Jersey, 1993.

[65] BEEREENDS, J. G. and STEMERDINK, J. A., "A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation". *Journal of Audio Eng. Soc.*, vol. 42, pp. 115 - 123, Mar. 1994.

[66] BEERENDS J., G., RIX A.W., HOLLIER M. P. and HEKSTRA A. P., "Perceptual Evaluation of Speech Quality (PESQ), The New ITU Standard for End-to-End Speech Quality Assessment". *Journal of Audio Eng. Soc.*, vol. 50, no. 10 pp., 755-778, Oct. 2002.

[67] SUN, H., SHUE, L., CHEN, J., "Investigations into the Relationship Between Measurable Speech Quality and Speech Recognition Rate for Telephony Speech". *Proceeding of ICASSP*, pp. 865-868, 2004.

[68] RIX, A. W., M. P. HOLLIER, A. P. HEKSTRA, and J. G. BEERENDS, "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment Part I - Time-Delay Compensation". *Journal of Audio Eng. Soc.*, vol. 50, pp., 755-764 (this issue).

[69] BEERENDS, J. G., HEKSTRA, A. P., RIX, A. W. and HOLLIER, M. P., "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment Part II - Psychoacoustic Model". *Journal of Audio Eng. Soc.*, vol. 50, pp., 765-778 (this issue).

[70] YOU, C. H., KOH S. N. and RAHARDJA, S., "Kalman Filtering Speech Enhancement incorporating masking properties for mobile communication in a car environment". *Proceedings of the IEEE International Conference on Multimedia and Expo - ICME2004*, Taipei, Taiwan, Jun. 2004.

[71] ITU-T Rec.P.800, "Methods for Subjective Determination of Transmission Quality". *International Telecommunication Union*, Geneva, Switzerland, Aug. 1996.

[72] ITU-T Rec.P.830, "Subjective Performance Assessment of Telephone-Band and Wideband Digital Codec". *International Telecommunication Union*, Geneva, Switzerland, Feb. 1996.

[73] ITU-T Study Group 12, "Review of Validation Tests for Objective Speech Quality Measures". Doc. COM 12-74, Mar. 1996.

[74] ITU-T Rec. P.861, "Objective Quality Measurement of Telephone-Band (300 - 3400 Hz) Speech Codecs". *International Telecommunications Union*, Geneva, Switzerland, Aug. 1996.

[75] VORAN, S., "Objective Estimation of Perceived Speech Quality - Part I and II: Development of the Measuring Normalizing Block Technique". *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 371-390, Jul. 1999.

[76] ZWICKER, E. and FELDTKELLER, R., "Das Ohr als Nachrichtenempfänger". Hirzel Verlang, Stuttgart, Germany, 1967.

[77] BEERENDS, J. G., "Modeling Cognitive Effects that Play a Role in the Perception of Speech Quality". *Speech Quality Assessment, DEGA, ITG, and EURASIP, Eds.*, pp. 1-9, Bochum, Germany, Nov. 1994.

[78] ETSI/TM/TM5/TCH-HS, "Correlation of a Perceptual Speech Quality Measure with the Subjective Quality of the GSM Candidate Half Rate Speech Codecs". Tech. Doc. 92/44, Dec. 1992.

[79] HOLLIER, M. P., HAWKSFORD, M. O., and GUARD, D. R., "Error Activity and Error Entropy as a Measure of Psychoacoustic Significance in the Perceptual Domian". *IEEE Proc. - Vision, Image and Signal Process.*, vol. 141, pp. 203-208, Jun. 1994.

[80] ITU-T Study Group 12, "Improvement of the P.861 Perceptual Speech Quality Measure". Doc. COM 12-20, Dec. 1997.

[81] ITU-T Rec. 48, "Specification for an Intermediate Reference System". *International Telecommunication Union*, Geneva, Switzerland, 1989.