Melhoria de sinais de fala baseada em atenuação espectral não-linear utilizando propriedades da percepção auditiva

Lidiane Krebsky da Silveira Abranches

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

Orientador: Prof. Dr. Francisco José Fraga da Silva

Santa Rita do Sapucaí
2004

Dissertação defendida e aprovada em 28/09/2004, pela comissão julgadora:

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

Prof. Dr. Francisco José Fraga da Silva - DTE / Instituto Nacional de Telecomunicações - INATEL

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

Coordenador do Curso de Mestrado

Prof. Dr. Adonias Costa da Silveira

Ao meu esposo Rogério e aos meus pais
Adonias e Miriam , que sempre me
motivaram a realizar este sonho e alcançar
este meu objetivo com êxito.

Agradecimentos

Ao Professor Doutor Francisco José Fraga da Silva pela dedicação e participação constante no desenvolvimento deste trabalho. Obrigada por me ajudar a progredir na arte da pesquisa e em me oferecer sempre tão pronta atenção.

A todos os colegas e amigos mais próximos pelo apoio e amizade nestes últimos anos. A todos os professores e funcionários do Inatel que, direta ou indiretamente, contribuíram na preparação deste trabalho.

Ao INATEL pela concessão da bolsa e também pela oportunidade de participação no PED - Programa de Estágio Docente, que muito me enriqueceu. À FAPEMIG e à CAPES pelo apoio financeiro dado ao projeto de pesquisa.

Agradeço aos meus pais, Miriam e Adonias, que acompanharam a minha luta no desenvolvimento deste trabalho me oferecendo todo apoio e encorajamento necessário.

E, em especial, agradeço ao meu esposo Rogério, que foi meu maior incentivador e que não mediu esforços para que eu pudesse com tranquilidade desenvolver meus estudos. Sua presença ao meu lado, todo amor e carinho dispensados foi o que tornou este trabalho possível.

E, principalmente, agradeço a Deus que me capacitou desenvolver este trabalho com saúde e paz dirigindo meus passos em todos os instantes, pelas lutas e vitórias que certamente me tornaram uma pessoa melhor.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ATH	*Absolute Threshold Hearing* - Limiar absoluto de audição
EMSR	*Ephraim and Malah noise Suppression Rule* - Regra de supressão de ruído de Ephraim e Malah
FDP	Função densidade de probabilidade
FFT	*Fast Fourier Transform* - Transformada rápida de Fourier
IFFT	*Inverse Fast Fourier Transform* - Transformada rápida de Fourier inversa
MMSE	*Minimum Mean Square Error* - Mínimo erro médio quadrático
MOS	*Mean Opinion Score* - Pontuação média de opinião
NMT-PSS	*Power Spectral Subtraction based on Noise Making Threshold* - Subtração espectral de potência baseada no limiar de mascaramento do ruído
PDF	*Probability Density Function* - Função Densidade de Probabilidade
PESQ	*Perceptual Evaluation of Speech Quality* - Avaliação perceptual de qualidade de fala
PSS	*Power Spectral Subtraction* - Subtração espectral de potência
SFM	*Spectral Flatness Measure* - Medida de Planura Espectral
SNR	*Signal Noise Rate* - Relação sinal-ruído
SpEAR	*Speech Enhancement Assessment Resource* - Recurso de avaliação de melhoria de fala

SPL *Sound Pressure Level* - Nível de pressão do som

STSS *Short-Time Spectral Subtraction* - Subtração espectral de tempo curto

Lista de Símbolos

Símbolos utilizados no Capítulos 2

α	Parâmetro fixo usado na determinação do parâmetro atenuação espectral $G(\omega)$
β	Parâmetro fixo usado na determinação do parâmetro atenuação espectral
γ	Parâmetro fixo usado na determinação do parâmetro atenuação espectral
$d(n)$	Ruído aditivo no domínio do tempo
$\hat{D}(\omega)$	Estimativa do espectro médio do ruído a cada frequência ω
$G(\omega)$	Função ganho
$s(n)$	Sinal de fala limpa no domínio do tempo
$\hat{S}(\omega)$	Estimativa do espectro do sinal de fala limpa a cada frequência ω
$y(n)$	Sinal de fala ruidosa no domínio do tempo
$Y(\omega)$	Espectro do sinal de fala ruidosa a cada frequência ω

Símbolos utilizados no Capítulo 3

$\alpha(q,\omega)$	Parâmetro usado na determinação da função de atenuação espectral $G(q,\omega)$, varia a cada quadro q e para cada freqüência ω.
α_{min}	Parâmetro fixo usado na determinação de $\alpha(q,\omega)$. Representa o menor valor que o parâmetro $\alpha(q,\omega)$ pode assumir.
α_{max}	Parâmetro fixo usado na determinação de $\alpha(q,\omega)$. Representa o maior valor que o parâmetro $\alpha(q,\omega)$ pode assumir.
α	Coeficiente de tonalidade
$\beta(q,\omega)$	Parâmetro usado na determinação da função de atenuação espectral $G(q,\omega)$, varia a cada quadro q e para cada freqüência ω.
β_{min}	Parâmetro fixo usado na determinação de $\beta(q,\omega)$. Representa o menor valor que o parâmetro $\beta(q,\omega)$ pode assumir.
β_{max}	Parâmetro fixo usado na determinação de $\beta(q,\omega)$. Representa o máximo valor que o parâmetro $\beta(q,\omega)$ pode assumir.
γ	Parâmetro fixo usado na determinação do parâmetro atenuação espectral
A_m	Média aritmética do espectro de potência
$ATH(f)$	Absolute Threshold Hearing - Limiar absoluto de audição em cada freqüência [Hz]
B_k	Energia em cada banda crítica k somada
b_{i_k}	Limite de freqüência inferior da banda crítica k
b_{s_k}	Limite de freqüência superior da banda crítica k
C_k	Matriz resultante da convolução de B_k com a função espalhamento S_k que é o espectro de banda crítica espalhado'
f	Freqüência em Hertz [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_m	Média geométrica do espectro de potência
k	O número da banda crítica
O_k	Limiar relativo para cada banda crítica k
$P(\omega)$	Espectro de potência do sinal
S_k	Função de espalhamento
SFM_{dB}	Spectral Flatness Measure (em dB)
T_k	Limiar de mascaramento do ruído para cada banda crítica k
$T(q, \omega)$	Limiar de mascaramento de ruído, varia a cada quadro e para cada frequência ω

Símbolos utilizados nos capítulos 4 e 5

α_k	A fase da k-ésima componente espectral S_k
$\alpha(q,\omega)$	Parâmetro usado na determinação da função de atenuação espectral $G(q,\omega)$, varia a cada quadro q e para cada freqüência ω.
γ_k	Relação sinal-ruído *a posteriori*
ϑ_k	A fase da k-ésima componente espectral Y_k
$\lambda_d(k)$	Variância da k-ésima componente do ruído
$\lambda_s(k)$	Variância da k-ésima componente da fala
μ	Parâmetro utilizado no cálculo de $R_{prio}(q,\omega)$
ν	Parâmetro utilizado no cálculo de $R_{prio}(q,\omega)$ utilizado para ponderação entre os quadros $q-1$ e $q-2$
ξ_k	Relação sinal-ruído *a priori*
A_k	A amplitude da k-ésima componente espectral S_k
$\hat{A}_k$	A estimativa de A_k
$d(n)$	Ruído aditivo no domínio do tempo
D_k	A k-ésima componente espectral do sinal de ruído $d(n)$
$\hat{D}(\omega)$	Potência estimada do ruído para cada componente espectral ω
$E\{.\}$	Operador esperança matemática
$G(q,\omega)$	Ganho espectral que é aplicado a cada valor do espectro de tempo curto
R_k	A amplitude da k-ésima componente espectral Y_k
R_{min}	Parâmetro que limita o mínimo valor de $R_{prio}(q,\omega)$ e $R_{post}(q,\omega)$
R_{max}	Parâmetro que limita o máximo valor de $R_{prio}(q,\omega)$ e $R_{post}(q,\omega)$
$R_{post}(q,\omega)$	Relação sinal ruído *a posteriori*, determinado a cada quadro q e para cada freqüência ω
$R_{prio}(q,\omega)$	Relação sinal ruído *a priori*, determinado a cada quadro q e para cada freqüência ω
$s(n)$	Sinal de fala limpa no domínio do tempo
S_k	A k-ésima componente espectral do sinal de fala limpa $s(n)$
$\hat{S}(q,\omega)$	Estimativa do espectro do sinal de fala limpa a cada quadro q e para cada freqüência ω
$T(q,\omega)$	Limiar de mascaramento de ruído, varia a cada quadro e para cada frequência ω

v_k	Parâmetro dependente de ξ_k e γ_k
$y(n)$	Sinal de fala ruidosa no domínio do tempo
Y_k	A k-ésima componente espectral do sinal de fala ruidosa $y(n)$
$Y(q,\omega)$	Valor do espectro ruidoso de tempo curto

Resumo

Este trabalho tem como objetivo apresentar um sistema de melhoria de sinais de fala, de canal único, que seja capaz de melhorar a qualidade e inteligibilidade de sinais de fala degradados por ruído aditivo. O algoritmo proposto é baseado na regra de supressão de ruído de Ephraim e Malah, mas com modificações introduzidas para tratar sinais ruidosos de baixa relação sinal-ruído (SNR < 10 dB). A principal modificação foi feita introduzindo-se o conceito de limiar de mascaramento do ruído, que é uma propriedade bem conhecida do sistema auditivo humano. Desta forma, a solução tornou-se correlacionada com a percepção humana.

Estas modificações foram feitas no sistema original de Ephraim e Malah com o objetivo de aumentar a redução de ruído e simultaneamente permitir uma maior redução do fenômeno de ruído musical. Resultados experimentais mostraram que uma melhoria geral na qualidade da fala foi obtida para todos os tipos de ruído aproximadamente estacionários considerados no experimento, em uma ampla faixa de relação sinal-ruído.

Abstract

In this work, the goal is to present a new single channel speech enhancement system to improve the quality and intelligibility of speech signals degraded by additive noise. The proposed algorithm is based on the Ephraim and Malah noise suppressor rule, but with some modifications added in order to deal with noisy speech presenting very low signal-to-noise ratios (SNR $<$ 10 dB). The principal modification was done by introducing the concept of noise masking threshold, which is a well-known property of the human auditory system. This consideration make the solution correlated with human perception.

These modifications were done with the aim of increasing the amount of noise reduction and simultaneously providing a more efficient elimination of the musical noise phenomenon. Evaluation results have shown that an improvement on the overall speech quality was achieved for all types of nearly stationary noise considered in the experiment, in a very wide signal-to-noise ratio range.

Capítulo 1

Introdução

1.1 A Melhoria da qualidade de sinais de fala - *Speech Enhancement*

Grande parte das relações humanas acontecem através da comunicação oral. Esta realidade impulsiona o mercado de telecomunicações na direção de um contínuo crescimento de aplicações de sistemas de processamento de voz. Entretanto, paralelamente, surgem os problemas decorrentes da presença de diversos tipos e níveis de ruído ambiental, comprometendo drasticamente o desempenho de sistemas que posteriormente se utilizarão da voz (por exemplo: codificadores e reconhecedores de fala). Dentre tantas aplicações, o uso progressivo das comunicações móveis destacou a importância de se ter algoritmos eficientes de melhoria da qualidade de sinais de fala.

Nos últimos anos, uma grande variedade de técnicas têm sido desenvolvidas com o objetivo de reduzir o ruído presente na fala. No entanto, permanece o desafio de encontrar técnicas cada vez mais eficientes para realizar esta tarefa. De maneira geral, as técnicas de redução de ruído têm como objetivo encontrar um compromisso entre a quantidade de ruído a ser removida e as distorções introduzidas no sinal de voz devidas ao processamento realizado.

Sistemas de melhoria de fala podem ser desenvolvidos segundo um critério matemático, utilizando uma modelagem específica da fala. Algumas técnicas baseiam-se em modelos de processos estocásticos, já outras estão fundamentadas em aspectos importantes da percepção humana. Como por exemplo, concentrar os esforços na melhoria da qualidade das consoantes sabendo que sua influência na inteligibilidade é muito grande, inteiramente desproporcional à quantidade de energia no sinal, em comparação com as vogais. Outra classificação importante dos métodos de melhoria é se estes utilizam apenas um canal ou múltiplos canais.

Para aplicações de canal único, a fala e o ruído ocupam o mesmo canal e a estimativa de ruído deve ser feita nos períodos onde não há atividade de voz. Já nos métodos que se utilizam de mais de um canal, um canal contém fala com ruído aditivo e um segundo canal é utilizado apenas para captar as amostras de ruído.

Neste trabalho, serão abordados os problemas e soluções relativos a sistemas de melhoria de apenas um canal, baseados na atenuação espectral de tempo curto. Como a maioria dos métodos de atenuação espectral de tempo curto utilizam-se dos princípios da subtração espectral, os conceitos básicos desta técnica serão explicados em primeiro lugar.

Os algoritmos baseados na subtração espectral possuem como vantagem a facilidade de implementação, bem como variadas opções de adaptação dos parâmetros de subtração de acordo com a aplicação desejada. No entanto, uma desvantagem significativa destas técnicas de atenuação de tempo curto é a introdução de ruído musical na fala melhorada. Este tipo de ruído, muitas vezes, é mais incômodo do que o próprio ruído original.

Diversas técnicas já foram criadas na tentativa de reduzir este ruído musical residual, dentre elas [1] [2] [3] [4] [5]. No entanto, tais técnicas conseguiram somente reduzir este efeito, sem eliminá-lo completamente. Em geral, quando é feita a tentativa de eliminar o fenômeno do ruído musical, realiza-se uma sobre-estimação do espectro médio do ruído. Porém, esta técnica traz uma consequência indesejável: a fala ruidosa é atenuada mais que o desejado, causando distorções audíveis no sinal de fala [6].

A regra de supressão de ruído proposta por EPHRAIM & MALAH [7] [8], EMSR - *Ephraim and Malah noise Suppression Rule* - realiza uma redução moderada de ruído, enquanto evita completamente o aparecimento do ruído musical. Em relações sinal-ruído médias (entre 10 e 15 dB), o ruído remanescente é suficientemente pequeno para prover uma boa qualidade perceptual da fala melhorada. No entanto, o mesmo não ocorre para altos níveis de ruído. Em baixas relações sinal-ruído (SNR < 10 dB), o desempenho não é tão eficaz, pois não aplica uma forte atenuação sobre o ruído indesejado, ou seja, uma considerável quantidade do ruído original permanece no sinal melhorado.

1.2 Objetivos deste trabalho

Com base nas considerações acima, o presente trabalho propõe um novo esquema de melhoria de fala com seu ponto central baseado na regra de Ephraim e Malah. Modificações foram realizadas com o objetivo de obter-se um resultado satisfatório para falas ruidosas que apresentam baixa relação sinal-ruído (SNR < 10 dB).

Para compor o sistema e proporcionar uma ferramenta capaz de trabalhar com sinais de baixa relação sinal-ruído, foi introduzido o conceito de limiar de mascaramento de ruído, que é uma propriedade bem conhecida do sistema auditivo humano, já amplamente usado em codificação de áudio de banda larga [9]. A idéia de explorar as propriedades de mascaramento do sistema auditivo humano surgiu do sistema proposto por VIRAG [10], onde foram adaptados os parâmetros básicos de subtração espectral com base no limiar de mascaramento do ruído. Desta forma, obteve-se uma melhoria da qualidade de fala em comparação com métodos clássicos, como o algoritmo de subtração espectral de potência [1], subtração espectral de potência modificado [2] e subtração espectral não linear [3]. A propriedade auditiva enuncia que um ouvinte humano não percebe ruído aditivo desde que este permaneça abaixo deste limiar de mascaramento. Mas a utilização isolada deste método, embora atinja altos níveis de redução de ruído e possua um desempenho muito superior a outros métodos, como pode ser visto em [10], apresenta a desvantagem de não eliminar completamente o ruído musical.

Portanto, no sistema proposto, o limiar de mascaramento do ruído é utilizado para adaptar a regra de supressão de ruído de Ephraim e Malah (EMSR), que por si mesma é capaz de eliminar completamente o ruído musical. A introdução do limiar permitirá melhorar o desempenho do método EMSR com respeito à quantidade de redução de ruído quando o sinal de fala ruidosa apresenta uma relação sinal-ruído baixa (SNR < 10 dB). Pode-se dizer então que o sistema proposto é baseado em um critério matemático-estatístico , devido ao modelamento utilizado no método EMSR, sendo ao mesmo tempo, também, um sistema que se utiliza de um critério perceptual, devido à utilização do limiar de mascaramento do ruído.

As contribuições deste trabalho foram publicadas em dois artigos científicos, apresentados nos congressos SIP-2004 (*Signal and Image Processing)* [14] e IWT-2004 (*International Workshop on Telecommunications*) [15].

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 apresentadas, também, as características do processamento de fala baseado em análise e síntese de quadros de tempo curto.

O capítulo 3 apresenta uma descrição teórica da estimação do parâmetro limiar de mascaramento do ruído e a forma de obtenção do parâmetro perceptual de atenuação, que introduz na solução proposta uma correlação com a percepção auditiva.

No capítulo 4 é feita uma explanação da regra de Ephraim e Malah, que é a base da solução proposta.

No capítulo 5 é apresentado o sistema proposto, aspectos da implementação e os resultados obtidos nas simulações. É especificada a base de dados utilizada e o ambiente de simulação utilizado para desenvolvimento do sistema.

No Capítulo 6 são descritas as conclusões deste trabalho e apresentadas as sugestões para outras pesquisas nesta área.

O anexo A contém a derivação da distribuição Rayleigh para as amplitudes espectrais, utilizada no Capítulo 4 na derivação do estimador de amplitude de Ephraim e Malah.

Os anexos B, C, D e E são utilizados no Capítulo 4 no desenvolvimento estatístico e matemático para obtenção do estimador de amplitude de Ephraim e Malah.

O anexo F apresenta o desempenho detalhado do método proposto, comparando-o com outros métodos existentes. Contém os resultados obtidos em todas as locuções utilizadas da base de dados.

Capítulo 2

Algoritmos de subtração espectral de tempo-curto

2.1 Subtração espectral de tempo-curto

No desenvolvimento dos mais 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.

Na entrada, temos um sinal de fala corrompido $y(n)$. A transformação realizada no sinal é feita por meio de um sistema padrão de análise e síntese operando quadro a quadro. Existem diversos métodos para o processamento de análise e síntese, mas a transformada de Fourier do sinal de tempo curto é o mais utilizado.

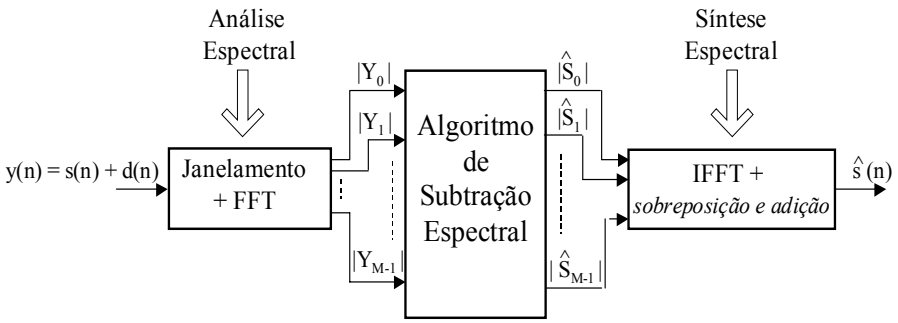

Figura 2.1: *Diagrama básico da subtração espectral de tempo curto.*

O processamento baseado em quadros não é motivado apenas pela disponibilidade da FFT (Fast Fourier Transform), mas principalmente pelo fato da fala ser estacionária em tempo-curto (10 a 40 ms). Além disso, será possível um processamento em tempo real (com um pequeno atraso), o que é necessário para a maioria das aplicações. Desta forma, a subtração é atualizada a cada quadro.

Para a análise, o sinal será dividido em quadros de tamanho N. Este tamanho N deve ser escolhido de acordo com as propriedades da fala, ou seja, o sinal da fala pode ser considerado estacionário para quadros de duração entre 10 a 40ms. A janela de análise é deslocada de M amostras, sendo $M < N$, (usualmente $M = \frac{N}{2}$, ou menor ainda) gerando uma sobreposição de quadros. Em seguida, o quadro é multiplicado pela janela de Hamming (ou Hanning) e posteriormente a FFT é aplicada ao sinal.

O espectro do sinal de saída é obtido após a subtração do ruído ser feita sobre cada quadro do sinal ruidoso. O sinal resultante é transformado de volta ao domínio do tempo por meio da FFT inversa e o sinal de saída é sintetizado. Esta síntese é, usualmente, realizada multiplicando cada quadro pela função inversa de janela de análise, seguida de um janelamento triangular (ou Hanning) e depois da adição dos quadros, sobrepostos em 50%.

2.2 Algoritmos Subtrativos

2.2.1 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 [1].

Para a análise desta técnica, consideremos um sinal de fala $s(n)$ corrompido por ruído aditivo estacionário $d(n)$. Desta forma, temos

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

A fala e o ruído são assumidos como sendo descorrelacionados e o processamento é feito a cada quadro no domínio da frequência. O propósito da subtração espectral é obter uma estimativa do sinal limpo, que denominamos aqui 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

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{D}(\omega)|^2, & se \ |Y(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0, & \text{caso contrário} \end{cases} \tag{2.2}$$

onde $|\hat{D}(\omega)|^2$ representa uma estimativa do espectro médio de potência do ruído. O espectro médio 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 usando a magnitude espectral do sinal limpo e a fase espectral do sinal ruidoso. Esta situação é chamada de limite teórico [10] e será considerada quando forem apresentados os resultados obtidos neste trabalho no capítulo 5.

2.2.2 Subtração Espectral como Filtragem

Os algoritmos do tipo subtrativo 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 de subtração espectral é equivalente à multiplicação da magnitude do espectro de tempo curto da fala ruidosa pela função ganho, a saber

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

Temos então, o filtro para PSS (Power Spectral Subtraction) correspondente a equação (2.2):

$$G(\omega) = \begin{cases} \sqrt{1 - \dfrac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}}, & se \ |Y(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0, & \text{caso contrário} \end{cases} \tag{2.4}$$

Desta forma, compreende-se melhor o processo de atenuação espectral de tempo-curto. As componentes espectrais resultantes estão em função do quanto elas excedem a estimativa de ruído. A consequência é que as regiões do espectro onde a energia da fala é alta, comparada com o ruído, sofrem pouca alteração, enquanto regiões compostas somente por ruído são, praticamente, suprimidas. Nestes dois casos, a função ganho assume um valor dependente do inverso da relação sinal ruído de cada componente espectral, a cada quadro.

Um dos algoritmos mais conhecidos de atenuação espectral foi proposto por

BEROUTI et al.[2]. A função ganho deste algoritmo é dada por

$$G(\omega) = \begin{cases} \left\{ 1 - \alpha . \left[\frac{|\hat{D}(\omega)|}{|Y(\omega)|} \right]^{\gamma} \right\}^{\frac{1}{\gamma}}, & se \quad \left[\frac{|\hat{D}(\omega)|}{|Y(\omega)|} \right]^{\gamma} < \frac{1}{\alpha + \beta} \\ \left\{ \beta . \left[\frac{|\hat{D}(\omega)|}{|Y(\omega)|} \right]^{\gamma} \right\}^{\frac{1}{\gamma}}, & \text{caso contrário} \end{cases} \tag{2.5}$$

Dentre os algoritmos do tipo subtrativo, este é um dos 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 $G(\omega) = 1$ (onde a componente espectral não é modificada) para $G(\omega) = 0$ (onde a componente espectral é suprimida).

A adequada escolha 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.

2.3 Ruído musical

Como já foi mencionado no Capítulo 1, o ruído residual musical é um fenômeno que surge ao utilizar-se técnicas de subtração espectral de tempo curto. Muitas vezes a qualidade da fala se torna pouco natural e este ruído residual é mais incômodo que o ruído original. Ele ocorre porque a magnitude do espectro de tempo curto possui grandes variações na fala ruidosa [11]. Após realizada a atenuação espectral, o espectro da magnitude de tempo curto nas bandas de frequência que continham apenas ruído apresentam agora uma sucessão de picos espectrais espaçados aleatoriamente, correspondendo aos pontos do quadro espectral atual onde a magnitude local excede a estimativa de ruído médio. Diversas modificações têm sido aplicadas com o objetivo de reduzir este efeito, como o cálculo da magnitude espectral média [1], sobre-estimação da potência do ruído [2], subtração espectral não-linear [3]. Mas estas técnicas não são capazes de eliminar o fenômeno do ruído musical, apenas reduzi-lo.

O termo *musical* é uma referência à presença de tons puros de curta duração no ruído residual.

Este trabalho tem como principal objetivo uma melhoria real da fala ruidosa, aplicando um algoritmo de atenuação espectral de tempo curto capaz de tornar imperceptível o ruído residual musical, em sinais com relação sinal ruído entre 0

e 15 dB.

Capítulo 3

Melhoria de sinais de fala utilizando propriedades de mascaramento do sistema auditivo humano

3.1 Introdução

A preocupação com o efeito psicoacústico do ruído surgiu na codificação [9] e compressão [12] de sinais de áudio que seriam transmitidos em canais de baixa taxa de transmissão ou seriam armazenados (como, por exemplo, o padrão MPEG [16]). Em pesquisas recentes, o modelo auditivo humano também tem sido bastante explorado [10] [17] [18], com o objetivo de realizar uma melhoria dos sinais de fala utilizando critérios baseados na percepção humana. Nestes algoritmos a preocupação não é remover completamente o ruído do sinal e sim atenuá-lo abaixo do limiar auditivo. No contexto de algoritmos subtrativos STSS, isto reduz a quantidade de modificação na amplitude espectral, reduzindo artefatos audíveis e contribuindo para obter-se um sinal de alta qualidade.

3.2 Propriedades do sistema auditivo humano

3.2.1 Limiar Auditivo Absoluto

Os sons necessitam ter uma pressão mínima para serem audíveis. Devido à sensibilidade seletiva do ouvido humano, esta pressão varia consideravelmente com a frequência. O limiar absoluto de audição também varia de pessoa para pessoa, de acordo com as características individuais e com a idade. A Figura 3.1 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.

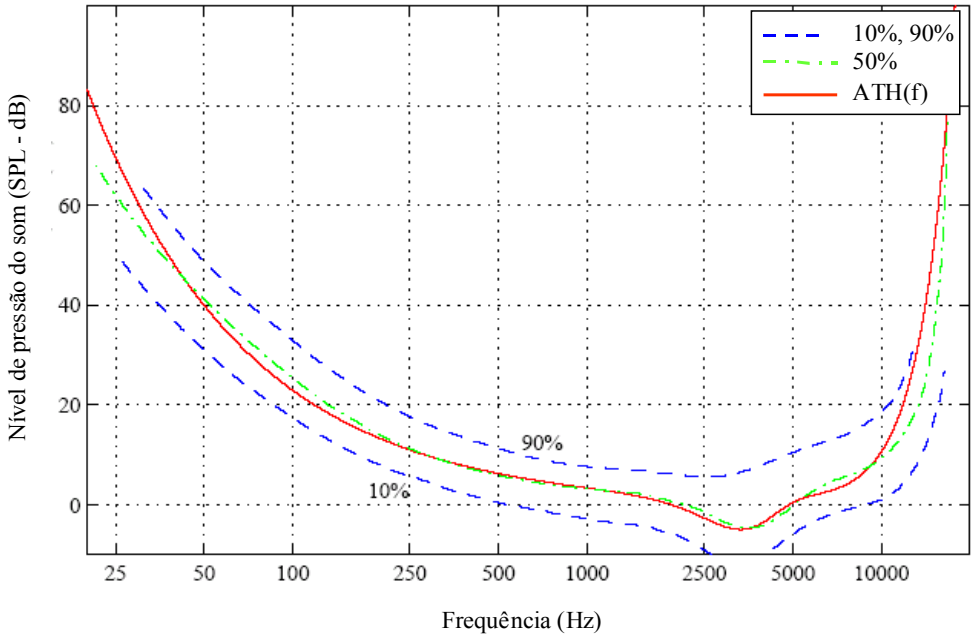

Figura 3.1: *Limiar de audição no silêncio: Nível de pressão do som (*SPL - Sound Pressure Level*)- requerido para que 10%, 50% e 90% das pessoas possam detectar um tom de teste. A linha contínua corresponde ao limiar absoluto de audição (*ATH - Absolute Threshold Hearing*) e apresenta a aproximação da equação (3.9). Adaptado de [22].*

3.2.2 Banda crítica

Experimentos voltados à percepção humana, como por exemplo aquele realizado por Fletcher [19], mostram que a banda crítica é uma faixa de frequência dentro da qual um som complexo não pode ter todas as suas componentes individualmente identificadas. Dependendo da diferença de intensidade entre as componentes de

frequência de um determinado som, elas só poderão ser individualmente distinguidas se ocorrerem em bandas críticas diferentes. Cada faixa de frequência corresponde a uma banda crítica e a largura da banda crítica aumenta com a frequência, conforme apresentado na Tabela 3.1 [10].

3.2.3 Mascaramento

O modelo utilizado considera apenas o mascaramento simultâneo (mascaramento no domínio da frequência): um sinal fraco se torna inaudível pela presença de um sinal mais forte ocorrendo simultaneamente. 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.

Como visto na seção anterior, só é possível distinguir dois tons quando a diferença de frequência entre os dois estiver acima do valor da banda crítica na faixa analisada. Essa percepção será mais difícil, quanto maior for a diferença de intensidade entre esses tons. Quando um tom de menor intensidade não puder mais ser percebido, diz-se então que ele foi "mascarado".

A Figura 3.2 mostra a quantidade de mascaramento provido por um tom de 1KHz para vários níveis L_M de pressão absoluta SPL do som. Pode-se notar que as inclinações da curva variam com o nível L_M. 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.

3.3 Cálculo do limiar de mascaramento do ruído

As propriedades de mascaramento do ouvido humano, assim como sua capacidade de seletividade de frequência, permite o cálculo do limiar de mascaramento do ruído. Este cálculo define um limiar de amplitude espectral, abaixo do qual todas as componentes espectrais são mascaradas na presença do sinal mascarador. Este cálculo é realizado conforme [10] e [9] e descrito nas etapas a seguir.

3.3.1 Etapa 1 - Análise da energia do sinal em cada banda crítica

O sinal ruidoso $y(n)$ é analisado, como já foi comentado no capítulo anterior, por meio de quadros de N amostras sucessivas. Usando a Transformada Rápida de

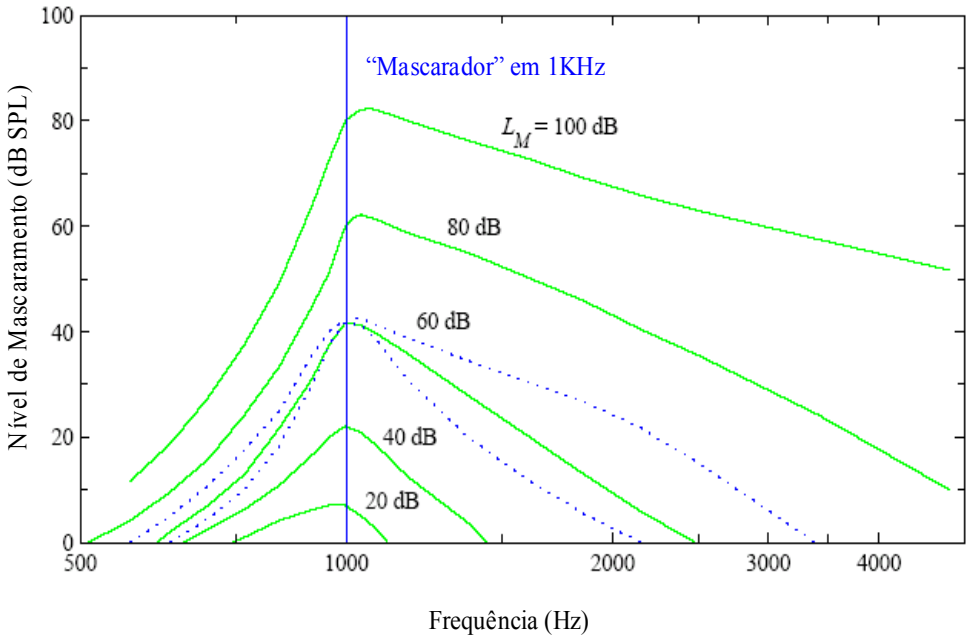

Figura 3.2: *Curvas de mascaramento para tom "mascarador" de 1kHz. Adaptado de [22].*

Fourier (*FFT - Fast Fourier Transform*) obtém-se o espectro complexo ($Re(\omega)$ e $Im(\omega)$), o qual é convertido em espectro de potência:

$$P(\omega) = Re^2(\omega) + Im^2(\omega) \tag{3.1}$$

De acordo com a Tabela 3.1 , o espectro é divido em bandas críticas e a energia de cada banda crítica é somada, ou seja

$$B_k = \sum_{\omega=bi_k}^{bs_k} P(\omega) \tag{3.2}$$

onde bi_k e bs_k são os limites de frequência inferior e superior, respectivamente, da banda crítica k e B_k é a energia na banda crítica k. O índice k depende da taxa de amostragem. Para aplicações em processamento de voz, na faixa de telefonia, $k = 18$ bandas críticas cobrem uma faixa de frequência até 4.000 Hz, conforme pode ser observado na Tabela 3.1.

Esta soma de energia, com base nas bandas críticas, traduz a transformação de frequência em espaço que acontece na membrana basilar[1], que se comporta de

[1] A membrana basilar situa-se no interior da cóclea, que é um canal espiral, cheio de líquido, que se comunica através de três pequenos ossos (martelo, bigorna e estribo) com a membrana timpânica. Ela está localizada no ouvido interno.

forma variada nas diversas frequências.

Tabela 3.1: *Escala de bandas críticas com as respectivas frequências em [Hz] e intervalos de* bins *da FFT de 256 pontos para frequência de amostragem de 8 kHz.*

N^oda banda crítica k	Freq.[Hz]	Intervalos da FFT
1	0 - 94	1 - 3
2	94 - 187	4 - 6
3	187 - 312	7 - 10
4	312 - 406	11 - 13
5	406 - 500	14 - 16
6	500 - 625	17 - 20
7	625 - 781	21 - 25
8	781 - 906	26 - 29
9	906 - 1094	30 - 35
10	1094 - 1281	36 - 41
11	1281 - 1469	42 - 47
12	1469 - 1719	48 - 55
13	1719 - 2000	56 - 64
14	2000 - 2312	65 - 74
15	2312 - 2687	75 - 86
16	2687 - 3125	87 - 100
17	3125 - 3687	101 - 118
18	3687 - 4000	119 - 128

3.3.2 Etapa 2 - Aplicação da função de espalhamento no espectro de banda crítica

A função de espalhamento é usada para estimar os efeitos do mascaramento entre diferentes bandas críticas. Pois, embora a percepção do ouvido aconteça em bandas críticas, porções do sinal de cada banda interferem nas bandas próximas ocasionando espalhamento da energia. A função espalhamento, apresentada na figura 3.3, tem inclinações aproximadas de +25 e -10 dB/banda crítica. É uma aproximação razoável para os dados experimentais obtidos por Zwicker [22] e apresentados anteriormente.

Para se obter o espectro de banda crítica espalhado C_k é feita uma convolução do espectro de energia na escala perceptual B_k com a função de espalhamento S_k

$$C_k = S_k * B_k \tag{3.3}$$

De acordo com [20], um modelo matemático de função de espalhamento satisfatório é dado, em decibéis, por

$$S_k = 15,81 + 7,5(k + 0,4) - 17,5\sqrt{1 + (k + 0,474)^2} \tag{3.4}$$

O deslocamento em número de bandas críticas varia de $+17$ a -17 para aplicações em processamento de voz na faixa de telefonia. A função é apresentada na Figura 3.3.

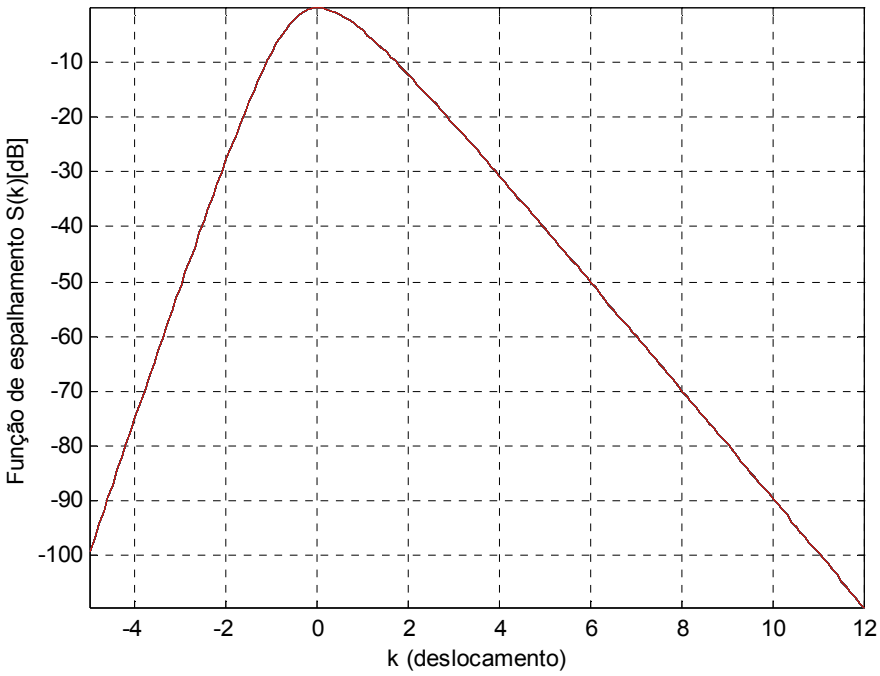

Figura 3.3: *Função de espalhamento: usada para o cálculo do limiar de mascaramento do ruído.*

3.3.3 Etapa 3 - Cálculo do limiar de mascaramento do ruído

Em diversas referências encontradas na literatura, como [19], [20] e [23], são detalhados dois tipos de limiares de mascaramento. O primeiro é o limiar para tom mascarando ruído, ele é estimado como $(14.5 + k)$ dB abaixo de C_k (sendo k

o número da banda crítica). E o segundo é o limiar para ruído mascarando tom, estimado como 5.5 dB abaixo de C_k, e uniforme dentro de uma mesma banda crítica.

Para determinar se o sinal tem natureza tonal ou de ruído é determinada a Medida de Planura Espectral (*SFM - Spectral Flatness Measure*). A SFM (em dB) é definida como

$$SFM_{dB} = 10log_{10}\frac{G_m}{A_m} \tag{3.5}$$

onde G_m e A_m representam a média geométrica e aritmética, respectivamente, do espectro de potência. Deste valor um coeficiente de tonalidade é gerado, a saber

$$\alpha = min\left(\frac{SFM_{dB}}{SFM_{dBmax}}, 1\right) \tag{3.6}$$

onde $SFM_{dBmax} = -60$ dB representa o SFM de um sinal completamente tonal, resultando em um coeficiente $\alpha = 1$. Um SFM de $0dB$ indicaria um sinal de natureza inteiramente ruidosa e um coeficiente $\alpha = 0$. Exemplificando, um SFM de -30 dB resulta em $\alpha = 0.5$ e um $SFM = -70$ dB resultaria em $\alpha = 1$.

Ao finalizar a determinação da natureza do sinal, através do cálculo de α, pode ser efetuado o cálculo do limiar relativo (ou limiar de *offset*) O_k, em decibels, para a energia de mascaramento em cada banda crítica k

$$O_k = \alpha(14 + i) + (1 - \alpha)5.5 \tag{3.7}$$

Para reduzir a complexidade e carga computacional do algoritmo, na implementação, foi utilizado um método mais simples para o cálculo de O_k proposto por Sinha & Tewfik em [21]. Esse método considera que o sinal de fala, em média, tem uma natureza tonal em bandas críticas mais baixas e uma natureza de ruído em bandas críticas mais altas, conforme apresentado na Figura 3.4.

O limiar relativo O_k é então subtraído do espectro de banda crítica espalhado C_k e assim é obtido o limiar de mascaramento do ruído T_k:

$$T_k = 10^{log_{10}(C_k)-(O_k/10)} \tag{3.8}$$

3.3.4 Etapa 4 - Renormalização e comparação com o limiar absoluto de audição

A função de espalhamento, ao ser convoluída com B_k, aumenta a energia estimada em cada banda. Para existir uma relação com o limiar de energia do espectro

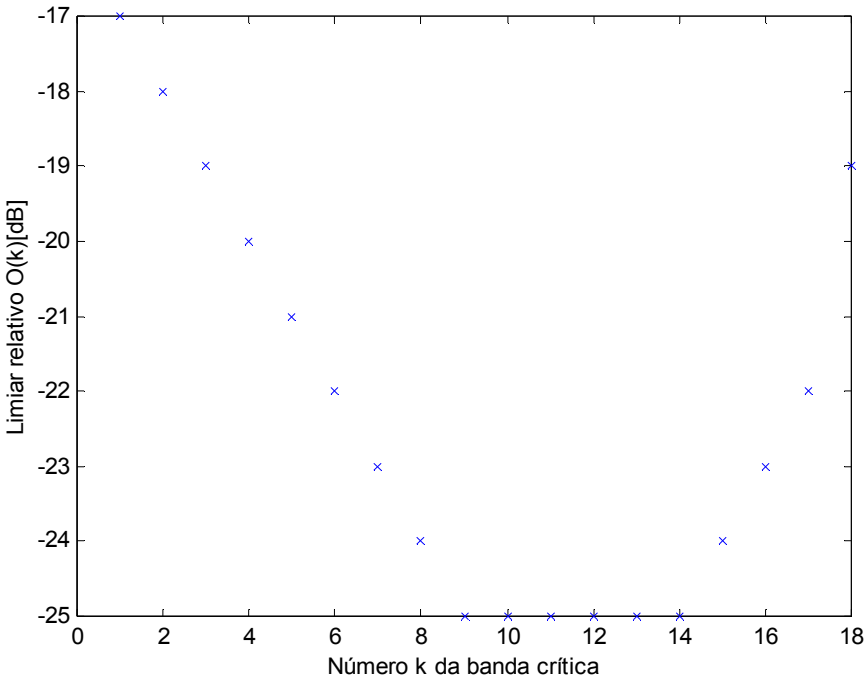

Figura 3.4: *Limiar relativo: usado para o cálculo do limiar de mascaramento do ruído.*

original seria necessário realizar uma deconvolução, pois a convolução feita com B_k deveria ser desfeita. Mas, devido ao formato da função de espalhamento, esse processo de deconvolução é muito instável. Por isso, em seu lugar é realizada uma renormalização. Isso é feito multiplicando cada limiar de mascaramento do ruído T_k pelo inverso do ganho de energia provocado pelo espalhamento.

Após efetuada a renormalização, o limiar deve ser comparado com as medidas de limiar absoluto da audição. Isto é realizado da seguinte forma: qualquer limiar de mascaramento do ruído T_k que estiver abaixo do limiar absoluto da audição é substituído pelo limiar absoluto para aquela banda crítica. Como existe uma variação do limiar absoluto dentro da banda crítica, uma média é calculada entre os valores absolutos das fronteiras da banda crítica.

A determinação do limiar absoluto (em dB) é feita através da equação (3.9), que é uma aproximação do limiar absoluto proposta para utilização em processamento de sinais [24]. Esta aproximação está representada na curva contínua da Figura 3.1.

$$ATH(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5\exp^{-0.6(f/1000-3.3)^2} + 10^{-3}\left(\frac{f}{1000}\right)^{4} \qquad (3.9)$$

onde f é a frequência em $[Hz]$.

Para ilustrar, a Figura 3.5 e a Figura 3.6 mostram o limiar de mascaramento do ruído para um mesmo quadro da fala limpa e ruidosa, respectivamente. A locução está amostrada em 8 KHz, portanto consideram-se 18 bandas críticas.

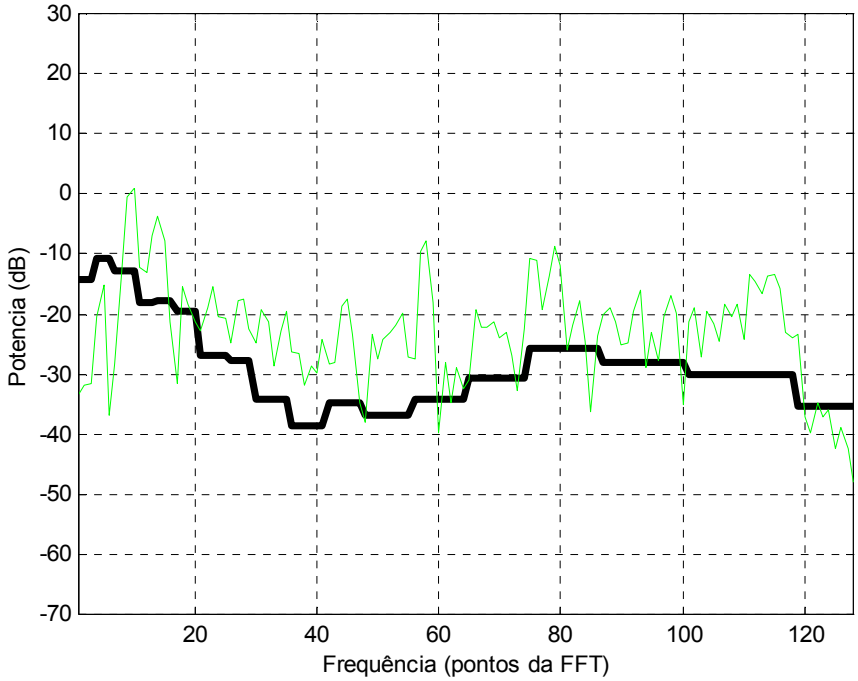

Figura 3.5: *Exemplo do limiar de mascaramento T_k. A linha mais espessa representa o limiar de mascaramento T_k nas 18 bandas críticas. A linha menos espessa representa o espectro de potência de 32ms de fala limpa.*

No método para 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 ruidoso está disponível. Então, é feita uma estimativa do sinal de fala limpa usando-se um esquema simples de subtração espectral de potência.

3.4 Cálculo dos parâmetros de subtração

VIRAG [10], em seus experimentos, usa a estimativa do limiar de mascaramento do ruído T_k para ajustar os parâmetros α e β da Equação (2.5) , a cada quadro q e para cada frequência ω.

A adaptação dos parâmetros é realizada da seguinte forma:

$$\alpha(q,\omega) = F_\alpha[\alpha_{min}, \alpha_{max}, T(q,\omega)] \tag{3.10}$$

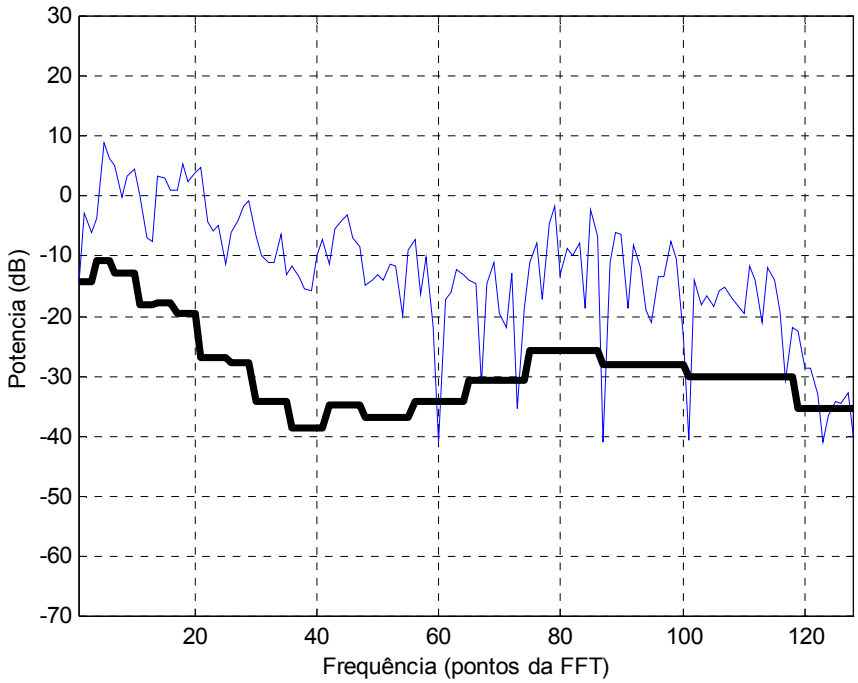

Figura 3.6: *Exemplo do limiar de mascaramento T_k. A linha mais espessa representa o limiar de mascaramento T_k nas 18 bandas críticas. A linha menos espessa representa 32ms de fala corrompida com ruído no interior da cabine de uma aeronave F16 ($SNR = 2.74dB$).*

$$\beta(q,\omega) = F_\beta[\beta_{min}, \beta_{max}, T(q,\omega)] \tag{3.11}$$

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

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

Após terem sido ajustados os parâmetros α e β, baseado em $T(q,\omega)$, estes valores são substituídos na Equação (2.5).

O parâmetro $\alpha(q,\omega)$ será denominado de *parâmetro perceptual de atenuação* quando utilizado no algoritmo proposto no Capítulo 5.

3.5 Implementação do algoritmo NMT-PSS

Denominamos de NMT-PSS (*Noise Masking Threshold - Power Spectral Subtraction*) o algoritmo que realiza o melhoramento do sinal de fala de acordo com o método explicado neste capítulo.

3.5.1 Características do sistema

- A entrada do sistema proposto é um sinal de fala amostrado em 16 kHz, oriundo da base de dados SpEAR [37], que foi degradado com ruído aditivo.
- Para o sistema proposto, o sinal foi submetido a um filtro passa-baixas com frequência de corte em 3400 kHz e a seguir subamostrado em 8 kHz.
- Cada quadro de análise consiste de 256 amostras da fala degradada. A sobreposição dos quadros é de 50%. Desta forma, cada quadro tem uma duração de 32 ms, com 16 ms de sobreposição entre quadros consecutivos.
- A análise espectral dos quadros é realizada por meio de uma FFT usando a janela de Hamming.
- É feita a estimativa da magnitude espectral utilizando o método NMT-PSS, que é então multiplicada pela exponencial complexa da fase ruidosa antes de proceder-se à volta para o domínio do tempo por meio da IFFT. A janela Hanning é utilizada na síntese.

3.5.2 Escolha de parâmetros

O algoritmo aqui denominado como NMT-PSS foi implementado de forma similar ao apresentado por VIRAG [10]. Após feito o cálculo do limiar de mascaramento do ruído $T(\omega)$, foi determinado $\alpha(q,\omega)$ e $\beta(q,\omega)$ diretamente na Equação (2.5), como detalhado nas seções anteriores deste capítulo.

Os parâmetros básicos escolhidos em [10], com melhor compromisso entre ruído residual e distorção da fala, foram

$\alpha_{min} = 1$ e $\alpha_{max} = 6$

$\beta_{min} = 0$ e $\beta_{max} = 0.02$

Parâmetros fixos: $\gamma = \gamma_1 = 2$ e $\gamma_2 = 0.5$

3.5.3 Resultados Obtidos

Os resultados de desempenho deste algoritmo serão posteriormente utilizados quando comparados com o algoritmo proposto neste trabalho, no Capítulo 5. A Figura 3.7 ilustra o desempenho deste algoritmo NMT-PSS. Pode-se verificar nos espectrogramas que o sinal resultante do algoritmo NMT-PSS elimina grande parte do ruído, bem mais que a clássica subtração espectral PSS, mas permanece ainda uma quantidade considerável de ruído musical (que pode ser identificado como pequenas "ilhas"isoladas no espectrograma), o qual deseja-se eliminar.

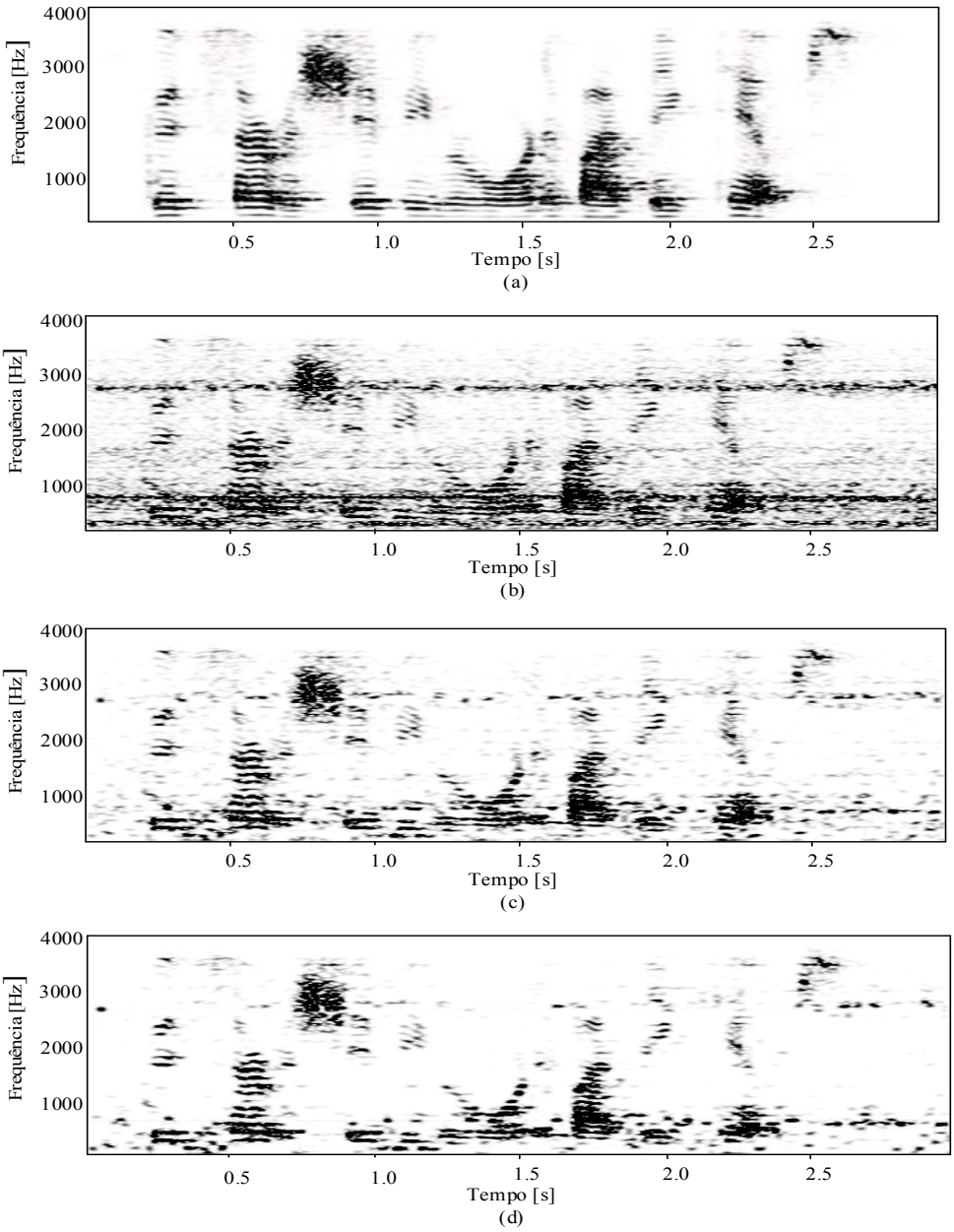

Figura 3.7: *Espectrogramas: (a)Sinal de fala limpa da frase em inglês "Good service should be rewarded by big tips", (b)Sinal de fala ruidoso (corrompido por ruído aditivo no interior da cabine de uma aeronave F16), (c)Sinal resultante da Subtração espectral - PSS e (d)Sinal resultante do algoritmo NMT-PSS.*

Capítulo 4

Método proposto por Ephraim e Malah para supressão de ruído

4.1 Introdução

O método EMSR (*Ephraim and Malah noise Supression Rule)* é um sistema de melhoria de sinais de fala baseado na derivação de um estimador ótimo de amplitude (magnitude) espectral de tempo curto (*STSA - Short-time spectral amplitude*). O estimador de amplitude derivado busca minimizar o erro quadrático médio (*MMSE - Minimum Mean-Square Error*) entre a amplitude espectral de tempo curto (STSA) original e sua estimação.

A técnica EMSR foi apresentada por Ephraim e Malah pela primeira vez em 1983 [7]. No ano seguinte, o seu desenvolvimento matemático foi detalhado em outra publicação [8]. Seguindo os mesmo princípios, outras regras de supressão foram propostas [8] [27]. Neste trabalho, será focado somente o princípio original de EMSR desenvolvido em [8], pois o mecanismo responsável por eliminar o ruído musical é basicamente o mesmo em todas as regras de supressão propostas. Nos estudos iniciais de Ephraim e Malah [7] [8], a redução de ruído sem introduzir ruído musical foi mencionada simplesmente como uma descoberta experimental. Mais tarde, os mecanismos que permitiam a eliminação do ruído musical foram investigados e detalhados por CAPPÈ [25]. Neste capítulo, todos estes mecanismos serão devidamente apresentados.

4.2 Modelo estatístico utilizado

Na derivação do estimador ótimo MMSE STSA, o foco foi colocado na necessidade de estimar o módulo de cada coeficiente de expansão em série de Fourier do sinal de fala em um dado quadro de análise a partir das observações ruidosas naquele mesmo quadro. A razão desta formulação é devida ao fato de os coeficientes de expansão de Fourier serem as amostras de sua transformada de Fourier e, também, pela relação próxima entre a expansão da série de Fourier e a transformada discreta de Fourier que permitirá uma implementação eficiente através da FFT. Para derivar este estimador, seria necessário um prévio conhecimento da distribuição de probabilidade dos coeficientes de expansão de Fourier do sinal da fala e do ruído. No entanto, como na prática isso não acontece optou-se por assumir um modelo estatístico razoável. Este modelo utiliza-se das propriedades estatísticas assintóticas dos coeficientes de expansão de Fourier quando o tempo de análise tende a infinito ($T \longrightarrow \infty$), assumindo que as partes reais e imaginárias dos coeficientes de cada processo (sinal de fala e ruído) podem ser modeladas como variáveis aleatórias gaussianas estatisticamente independentes. A média de cada coeficiente é assumida como zero e a variância varia no tempo, devido à não estacionariedade da fala (em tempo longo).

O Teorema do Limite Central motiva a utilização do modelo estatístico Gaussiano, visto que os coeficientes de expansão de Fourier são, na verdade, uma soma ponderada (ou integral) das variáveis aleatórias resultantes das amostras do processo no domínio do tempo[1].

Destaca-se, ainda, que esse modelo estatístico Gaussiano resulta em uma distribuição Rayleigh (ver Anexo A) para a amplitude de cada componente espectral da fala que assume probabilidade insignificante para realizações de baixa amplitude.

4.2.1 Independência estatística no modelo gaussiano

Assumir independência estatística no modelo Gaussiano é, na verdade, equivalente à hipótese de que os coeficientes de expansão de Fourier são descorrelacionados. É possível demonstrar que a correlação entre os coeficientes aproxima-se de zero à medida que o comprimento do quadro de análise tende para o infinito [28]. Em nosso sistema, o comprimento do quadro de análise, como já mencionado no Capítulo 2, não pode ser muito grande devido à hipótese de quase-

[1]Embora não seja mencionado por Ephraim e Malah, pode-se reforçar este argumento se considerarmos que a função de autocorrelação de tempo longo média de um sinal de fala tende a zero para amostras espaçadas de mais de 3ms [26]. Ou seja, dentro de um quadro de fala de 30 ms, a maior parte das amostras possuem baixa correlação entre si reforçando o argumento de os coeficientes de Fourier apresentarem partes real e imaginária com distribuição gaussiana (Teorema do Limite Central).

estacionariedade do sinal de fala ser válida somente para quadros de duração entre 10 e 40ms. Com isso, ao limitar-se o tamanho do quadro de análise, pode-se estar provocando um certo grau de correlação entre os coeficientes espectrais. No entanto, com o objetivo de simplificar o algoritmo resultante, mantém-se essa consideração de independência estatística. O que ocorre na prática é aplicar-se sobre o sinal ruidoso uma janela de Hamming, o que torna as componentes espectrais mais distantes descorrelacionadas. Mas a correlação entre as componentes espectrais adjacentes é aumentada, em consequência do uso da janela de Hamming que resulta em um lóbulo principal mais largo e lóbulos laterais menores se comparado com a aplicação de uma janela retangular.

4.2.2 A validade do modelo gaussiano

É importante destacar que os pesquisadores não chegaram a um acordo quanto à melhor distribuição de probabilidade das componentes espectrais da fala. Os estudos apontaram para distribuições distintas. Por exemplo, algumas pesquisas [29] [30] apresentam o modelo gaussiano como sendo a distribuição mais aproximada, enquanto outra [31] já apresentou a distribuição Gama. Portanto, a validade do modelo estatístico gaussiano proposto baseia-se nos resultados obtidos no trabalho realizado por Ephraim e Malah [8].

4.3 Derivação do estimador de Amplitude

É apresentada nesta seção a derivação do estimador de amplitude de forma bastante detalhada. O desenvolvimento matemático completo apresentado nesta seção não foi encontrado em nenhuma referência bibliográfica, sendo necessário desenvolver as equações básicas disponíveis em [8], para, através de diversas etapas de cálculo, chegar à equação final do estimador de amplitude. Será constatado que nenhuma das etapas no desenvolvimento do estimador foi omitida. Desta forma, pretende-se que esta seção sirva de material consulta para as pesquisas baseadas no Supressor de Ruído de Ephraim e Malah, nas quais seja conveniente uma compreensão aprofundada do desenvolvimento matemático que permitiu a obtenção do estimador de Amplitude.

Assim como em (2.1), representa-se o sinal de fala ruidosa $y(n)$, no intervalo de observação $[0, N-1]$, como a fala limpa $s(n)$ corrompida pelo ruído aditivo $d(n)$

$$y(n) = s(n) + d(n), \quad 0 \leq n \leq N-1 \tag{4.1}$$

Na análise a ser feita, temos

- $S_k \triangleq A_k exp(j\alpha_k)$: a k-ésima componente espectral do sinal $s(n)$ no intervalo [0,$N-1$];
- D_k : a k-ésima componente espectral do ruído $d(n)$ no intervalo [0,$N-1$];
- $Y_k \triangleq R_k exp(j\vartheta_k)$: a k-ésima componente espectral do sinal ruidoso $y(n)$ no intervalo [0,$N-1$];
- $\sigma_s^2(k) = \dfrac{\lambda_s(k)}{2}$: variância da k-ésima componente da fala, e
- $\sigma_d^2(k) = \dfrac{\lambda_d(k)}{2}$: variância da k-ésima componente do ruído

De acordo com o modelo anteriormente escolhido, assumem-se, portanto, as seguintes hipóteses: A_k é uma variável aleatória com distribuição Rayleigh, com a fase α_k uniformemente distribuída no intervalo $[0, 2\pi]$; D_n possui partes real e e imaginária com distribuição gaussiana de média zero e variâncias iguais (Ver Anexo A). Também assume-se que A_k, α_k e D_k são estatisticamente independentes.

As componentes espectrais de Y_k são obtidas a partir de $y(n)$ por meio de Transformada Discreta de Fourier: (4.2), a saber

$$Y_k = \sum_{n=0}^{N-1} y(n)\, exp\left(-j\frac{2\pi}{N}kn\right), \quad k = 0, 1, 2, ..., N-1 \tag{4.2}$$

As componentes espectrais S_k e D_k, assim como Y_k, são obtidas em (4.2).

O objetivo é estimar o módulo A_k, a partir de um sinal corrompido $y(n)$ no intervalo [0,T].

É necessário minimizar a seguinte medida de distorção

$$E\{(A_k - \hat{A}_k)^2\} \tag{4.3}$$

onde $E\{.\}$ representa o operador esperança estatística.

Dadas as observações ruidosas $y(n), 0 \leq n \leq N-1$, este estimador ótimo em um sentido de mínimo erro médio quadrático (MMSE) é (ver Anexo B) dado por:

$$\hat{A}_k = E\{A_k | y(n), \quad 0 \leq n \leq N-1\} \tag{4.4}$$

O sinal corrompido $y(n)$ será tratado em termos de suas componentes espectrais Y_k, as quais são modeladas como variáveis aleatórias gaussianas[2]. Ou seja, a partir de componentes $\{Y_0, Y_1, ...\}$, deseja-se obter uma estimativa de A_k:

[2]Como as partes real e imaginária de Y_k são variáveis aleatórias gaussianas de acordo com o modelo assumido, Y_k é modelada por uma distribuição gaussiana conjunta.

$$\hat{A}_k = E\left\{A_k | Y_0, Y_1, ...\right\} \tag{4.5}$$

Como as componentes espectrais foram assumidas como estatisticamente independentes, o estimador de amplitude MMSE pode ser derivado apenas com base em Y_K:

$$\hat{A}_k = E\left\{A_k | Y_k\right\} \tag{4.6}$$

Utilizando a regra de Bayes, conforme apresentado no Anexo C, obtemos:

$$E\left\{A_k | Y_k\right\} = \frac{\int_0^\infty \int_0^{2\pi} a_k p(Y_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(Y_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k} \tag{4.7}$$

onde $p\left(.\right)$ é a função densidade de probabilidade da variável aleatória associada à variável do seu argumento.

Assumindo o modelo estatístico apresentado anteriormente, em que assume-se distribuição Rayleigh para a amplitude de cada componente espectral da fala a_k e uma distribuição uniforme entre 0 e 2π para sua respectiva fase α_k (Anexo A), temos:

$$p(a_k) = \begin{cases} \dfrac{2a_k}{\lambda_s(k)} exp\left(-\dfrac{a_k^2}{\lambda_s(k)}\right), & se \;\; a_k \in [0, \infty) \\ 0, & c.c. \end{cases} \tag{4.8}$$

e

$$p(\alpha_k) = \begin{cases} \dfrac{1}{2\pi}, & se \;\; \alpha_k \in [0, 2\pi) \\ 0, & c.c. \end{cases} \tag{4.9}$$

Portanto, sabendo da independência estatística entre as variáveis, temos:

$$p\left(a_k, \alpha_k\right) = \frac{a_k}{\pi.\lambda_s(k)} exp\left\{-\frac{a_k^2}{\lambda_s(k)}\right\} \tag{4.10}$$

Também, visto que as partes real e imaginária de D_k são variáveis aleatórias gaussianas e estatisticamente independentes, de média zero, e que possuem a mesma variância, temos (ver Anexo A):

$$p\left(Y_k|a_k,\alpha_k\right) = \frac{1}{\pi.\lambda_d(k)}\, exp\left\{-\frac{1}{\lambda_d(k)}|Y_k - a_k e^{j\alpha_k}|^2\right\} \tag{4.11}$$

Podemos analisar graficamente o significado de (4.11), observando a Figura 4.1.

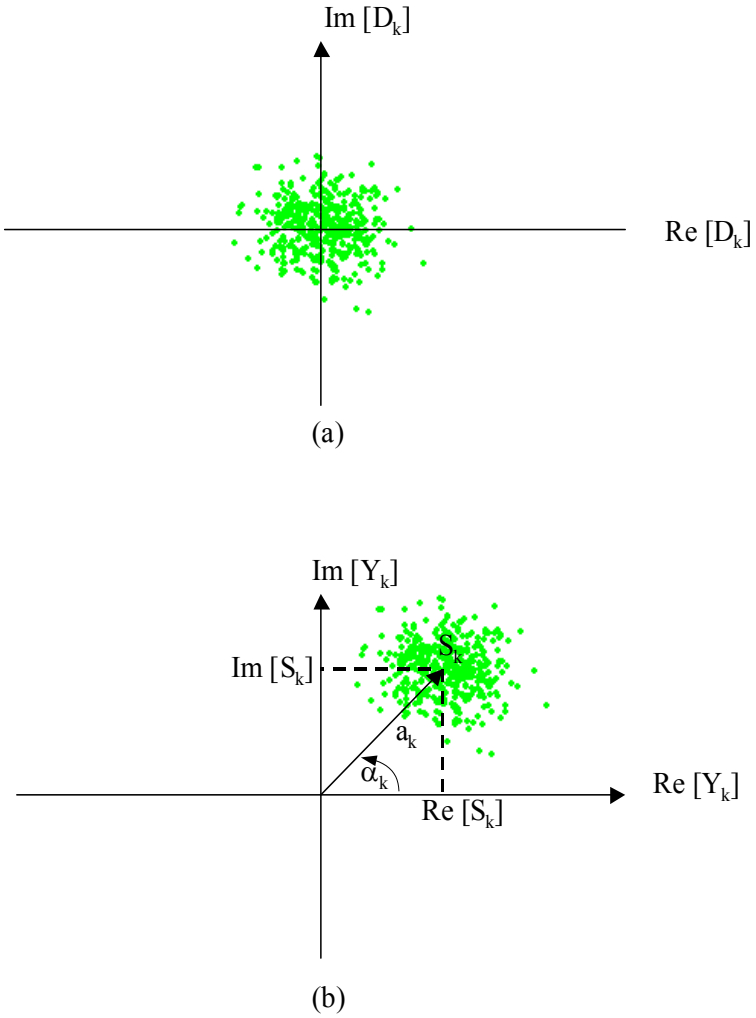

Figura 4.1: *Representação gráfica do comportamento de $Y_k|a_k,\alpha_k$ $(Y_k|S_k)$.*

Na Figura 4.1(a), vemos a representação gráfica simplificada das amostras de uma componente espectral do ruído (D_k) no plano complexo. Na Figura 4.1(b), vemos a variável aleatória $Y_k|S_k$ representada em um plano complexo como uma "nuvem" de pontos com distribuição gaussiana em torno do ponto médio $S_k \triangleq a_k e^{j\alpha_k}$, mostrando a variação causada no sinal limpo após a introdução do ruído.

Substituindo (4.10) e (4.11) em (4.7), temos

$$\hat{A}_k = \frac{N(\hat{A}_k)}{D(\hat{A}_k)}$$

$$\hat{A}_k = \frac{\int_0^{+\infty}\int_0^{2\pi} a_k \left[\frac{1}{\pi\lambda_d(k)}\, exp\left(-\frac{1}{\lambda_d(k)}|Y_k - a_k e^{j\alpha_k}|^2\right)\right]\left[\frac{a_k}{\pi\lambda_s(k)}\, exp\left(-\frac{a_k^2}{\lambda_s(k)}\right)\right] d\alpha_k da_k}{\int_0^{+\infty}\int_0^{2\pi} \left[\frac{1}{\pi\lambda_d(k)}\, exp\left(-\frac{1}{\lambda_d(k)}|Y_k - a_k e^{j\alpha_k}|^2\right)\right]\left[\frac{a_k}{\pi\lambda_s(k)}\, exp\left(-\frac{a_k^2}{\lambda_s(k)}\right)\right] d\alpha_k da_k} \tag{4.12}$$

Consideremos por enquanto apenas o numerador $N(\hat{A}_k)$ de (4.12) para simplificar a apresentação dos cálculos:

$$N(\hat{A}_k) = \int_0^{+\infty} \frac{a_k^2}{\pi\lambda_s(k)}\, exp\left(-\frac{a_k^2}{\lambda_s(k)}\right) \underbrace{\int_0^{2\pi} \frac{1}{\pi\lambda_d(k)}\, exp\left(-\frac{1}{\lambda_d(k)}\underbrace{|Y_k - a_k e^{j\alpha_k}|^2}_{\text{(II)}}\right) d\alpha_k\, da_k}_{\text{(I)}} \tag{4.13}$$

Da equação (4.13), sabendo que $Y_k = R_k exp(j\vartheta_k)$, desenvolvemos a parcela (II):

$$\begin{aligned} |Y_k - a_k e^{j\alpha_k}|^2 &= |R_k[cos(\vartheta_k) + jsen(\vartheta_k)] - a_k[cos(\alpha_k) + jsen(\alpha_k)]|^2 \\ &= [R_k cos(\vartheta_k) - a_k cos(\alpha_k)]^2 + [R_k sen(\vartheta_k) - a_k sen(\alpha_k)]^2 \\ &= R_k^2 cos^2(\vartheta_k) - 2R_k a_k cos(\vartheta_k)cos(\alpha_k) + a_k^2 cos^2(\alpha_k) + \\ &\quad R_k^2 sen^2(\vartheta_k) - 2R_k a_k sen(\vartheta_k)sen(\alpha_k) + a_k^2 sen^2(\alpha_k) \end{aligned} \tag{4.14}$$

Utilizando as relações trigonométricas apresentadas em (4.15) e (4.16):

$$sen^2(u) + cos^2(u) = 1 \tag{4.15}$$

$$cos(u).cos(v) + sen(u).sen(v) = cos(u - v) \tag{4.16}$$

Obtém-se resultado da parcela (II) desenvolvida em (4.14) :

$$|Y_k - a_k e^{j\alpha_k}|^2 = R_k^2 - 2R_k a_k cos(\alpha_k - \vartheta_k) + a_k^2 \tag{4.17}$$

Aplicando o resultado de (4.17) na parcela (I) de (4.13):

$$(\mathrm{I}) = \frac{1}{\pi\lambda_d(k)} \int_0^{2\pi} exp\left[-\left(\frac{R_k^2 - 2R_k a_k cos(\alpha_k - \vartheta_k) + a_k^2}{\lambda_d(k)}\right)\right] d\alpha_k$$

$$(\mathrm{I}) = \frac{1}{\pi\lambda_d(k)}\, exp\left[-\frac{(R_k^2 + a_k^2)}{\lambda_d(k)}\right] \int_0^{2\pi} exp\, \frac{[2R_k a_k cos(\alpha_k - \vartheta_k)]}{\lambda_d(k)}\, d\alpha_k \qquad (4.18)$$

Sabendo que a função de Bessel modificada (do primeiro tipo) de ordem zero é definida como:

$$I_0(z) = \frac{1}{2\pi} \int_0^{2\pi} exp\,[zcos\beta] d\beta \qquad (4.19)$$

façamos as seguintes substituições:

$$\beta = \alpha_k - \vartheta_k \begin{cases} d\beta = d\alpha_k \\ \alpha_k = 0 \Rightarrow \beta = -\vartheta_k \\ \alpha_k = 2\pi \Rightarrow \beta = 2\pi - \vartheta_k \end{cases}$$

$$z = \frac{2R_k a_k}{\lambda_d(k)}$$

Desta forma, obtemos:

$$2\pi I_0(z) = \int_{-\vartheta_k}^{2\pi - \vartheta_k} exp\,[zcos\beta] d\beta = \int_0^{2\pi} exp\,[zcos\beta] d\beta \qquad (4.20)$$

A igualdade acima é válida, pois $exp\,[zcos\beta]d\beta$ é uma função periódica com período 2π na variável β.

De (4.18), usando (4.20), obtemos:

$$(I) = \frac{1}{\pi\lambda_d(k)}\, exp\left[-\frac{(R_k^2 + a_k^2)}{\lambda_d(k)}\right] 2\pi I_0(z)\, d\alpha_k \qquad (4.21)$$

Substituindo a equação (4.21) em (4.13), temos:

$$N(\hat{A}_k) = \int_0^{+\infty} \frac{a_k^2}{\pi\lambda_s(k)}\, exp\left(-\frac{a_k^2}{\lambda_s(k)}\right) \frac{2\pi}{\pi\lambda_d(k)}\, exp\left[-\frac{(R_k^2 + a_k^2)}{\lambda_d(k)}\right] I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k \qquad (4.22)$$

$$N(\hat{A}_k) = \frac{2\pi}{\pi^2 \lambda_s(k)\lambda_d(k)} \int_0^{+\infty} a_k^2 \, exp\left(-\frac{a_k^2}{\lambda_s(k)}\right) \, exp\left[-\frac{(R_k^2 + a_k^2)}{\lambda_d(k)}\right] I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k$$

$$N(\hat{A}_k) = \frac{2\pi}{\pi^2 \lambda_s(k)\lambda_d(k)} \int_0^{+\infty} a_k^2 \, exp\left[-a_k^2\left(\frac{1}{\lambda_s(k)} + \frac{1}{\lambda_d(k)}\right)\right] \, exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k$$

Definindo:

$$\frac{1}{\lambda(k)} = \frac{1}{\lambda_s(k)} + \frac{1}{\lambda_d(k)} \tag{4.23}$$

Temos:

$$N(\hat{A}_k) = \frac{2\pi}{\pi^2 \lambda_s(k)\lambda_d(k)} \, exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) \int_0^{+\infty} a_k^2 \, exp\left(\frac{-a_k^2}{\lambda(k)}\right) \, I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k \tag{4.24}$$

Tendo simplificado o numerador $N(\hat{A}_k)$ em (4.24), devido à semelhança com o denominador $D(\hat{A}_k)$, obtém-se:

$$D(\hat{A}_k) = \frac{2\pi}{\pi^2 \lambda_s(k)\lambda_d(k)} \, exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) \int_0^{+\infty} a_k \, exp\left(\frac{-a_k^2}{\lambda(k)}\right) \, I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k \tag{4.25}$$

Substituindo (4.25) e (4.21) em (4.12), obtem-se

$$\hat{A}_k = \frac{N(\hat{A}_k)}{D(\hat{A}_k)} = \frac{\frac{2\pi}{\pi^2 \lambda_s(k)\lambda_d(k)} \, exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) \int_0^{+\infty} a_k^2 \, exp\left(\frac{-a_k^2}{\lambda(k)}\right) \, I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k}{\frac{2\pi}{\pi^2 \lambda_s(k)\lambda_d(k)} \, exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) \int_0^{+\infty} a_k \, exp\left(\frac{-a_k^2}{\lambda(k)}\right) \, I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k}$$

$$\hat{A}_k = \frac{\int_0^{+\infty} a_k^2 \, exp\left(\frac{-a_k^2}{\lambda(k)}\right) \, I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k}{\int_0^{+\infty} a_k \, exp\left(\frac{-a_k^2}{\lambda(k)}\right) \, I_o\left(\frac{2R_k a_k}{\lambda_d(k)}\right) da_k} \tag{4.26}$$

Definem-se, então, os seguintes parâmetros:

$$v_k \triangleq \frac{\xi_k}{1 + \xi_k} . \gamma_k \tag{4.27}$$

onde temos:

$$\xi_k \triangleq \frac{\lambda_s(k)}{\lambda_d(k)} \tag{4.28}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{4.29}$$

Seguindo MCAULAY & MALPASS [33] γ e ξ são interpretadas como sendo as relações sinal-ruído *a posteriori* e *a priori*, respectivamente.

Baseado nos parâmetros definidos em (4.27), (4.28), (4.29) e (4.23) , a partir de (4.26) obtém-se (ver Anexo D):

$$\hat{A}_k = \frac{\int_0^{+\infty} a_k^2\, exp\left[-a_k^2\left(\frac{1}{\lambda(k)}\right)\right]\, I_o\left(2a_k\sqrt{\frac{v_k}{\lambda(k)}}\right) da_k}{\int_0^{+\infty} a_k\, exp\left[-a_k^2\left(\frac{1}{\lambda(k)}\right)\right]\, I_o\left(2a_k\sqrt{\frac{v_k}{\lambda(k)}}\right) da_k} \tag{4.30}$$

Desenvolvendo (4.30) (Anexo E), chega-se ao seguinte resultado

$$\hat{A}_k = \Gamma(1.5).\frac{\sqrt{v_k}}{\gamma_k}.M(-0.5;1;-v_k)R_k \tag{4.31}$$

onde $M(a;b;x)$ é a função confluente hipergeométrica e $\Gamma(.)$ é a função gama, sendo que $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$.

Usando a seguinte relação [44]

$$M(-0.5;1;-x) = exp\,(-x/2)[(1+x)I_0(x/2) + xI_1(x/2)] \tag{4.32}$$

temos finalmente que

$$\hat{A}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k}\, exp\,(-\frac{v_k}{2})\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k \tag{4.33}$$

onde $I_0(.)$ e $I_1(.)$ são as funções de Bessel modificadas de ordem zero e ordem um, respectivamente.

4.4 A Regra de supressão de Ephraim e Malah

O algoritmo EMSR é um tipo de algoritmo de atenuação espectral de tempo curto. O estimador desenvolvido na seção anterior permite apresentar o ganho espectral $G(q,\omega)$, que é aplicado sobre a magnitude de cada componente espectral ruidosa de tempo curto $|Y(q,\omega)|$. Isso é feito afim de obter-se uma estimativa da

magnitude $|\hat{S}(q,\omega)|$ da componente espectral ω da fala limpa no q-ésimo quadro:

$$|\hat{S}(q,\omega)| = G(q,\omega)|Y(q,\omega)| \tag{4.34}$$

A Equação (4.34) possui o mesmo significado da equação (4.33), porém utilizando uma notação semelhante àquela utilizada nos capítulos 2 e 3, enquanto que (4.33) é apresentada segundo a notação original utilizada por EPHRAIM & MALAH.

A Equação (4.35), proposta por CAPPÈ [25], apresenta uma forma equivalente à equação (4.33), que por sua nomenclatura mais didática, será utilizada a partir desta seção. Assim,

$$G = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+R_{post}}\right)\left(\frac{R_{prio}}{1+R_{prio}}\right)}.M\left[(1+R_{post})\left(\frac{R_{prio}}{1+R_{prio}}\right)\right] \tag{4.35}$$

onde

$$M[\Theta] = \exp(-\frac{\Theta}{2})[(1+\Theta)I_0(\frac{\Theta}{2}) + \Theta I_1(\frac{\Theta}{2})] \tag{4.36}$$

A passagem da notação utilizada por EPHRAIM & MALAH para a notação de CAPPÈ é feita por meio das seguintes relações:

$$\begin{aligned} R_{post} &= \gamma_k - 1 \\ R_{prio} &= \xi_k \end{aligned} \tag{4.37}$$

A função $M[\Theta]$ apresentada na equação (4.35) está relacionada a equação (4.32) que substituída em (4.31) , temos $-x = -v_k$. E utilizando as equações (4.27), (4.28) e (4.29) e a notação proposta por CAPPÈ, temos:

$$v_k = (1+R_{post})\left(\frac{R_{prio}}{1+R_{prio}}\right)$$

Na Equação (4.35), omitiu-se a indexação quadro q e freqüência ω. O ganho $G(q,\omega)$, como visto em (4.35), depende de dois parâmetros, apresentados em (4.38) e (4.39):

$$R_{post}(q,\omega) = \begin{cases} \dfrac{|Y(q,\omega)|^2}{|\hat{D}(\omega)|^2} - 1, & se \quad \dfrac{|Y(q,\omega)|^2|}{\hat{D}(\omega)|^2} > 1 \\ 0, & c.c. \end{cases} \tag{4.38}$$

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

$$R_{prio}(q,\omega) = (1-\mu).R_{post}(q,\omega) + \mu.\frac{|G(q-1,\omega).Y(q-1,\omega)|^2}{\hat{D}(\omega)^2} \tag{4.39}$$

O parâmetro R_{prio} é a relação sinal-ruído *a priori*. Pode-se verificar que na primeira parcela de (4.39) é efetuada uma ponderação de $(1-\mu)$ na relação sinal ruído do quadro atual, R_{post}. Ainda em (4.39), temos o fator $G(q-1,\omega).Y(q-1,\omega)$, que é uma estimativa do espectro do sinal sem ruído do quadro anterior. Desta forma, temos na segunda parcela de (4.39) 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 apenas do sinal ruidoso no quadro atual.

O parâmetro μ foi escolhido, experimentalmente, como 0.98 em [25], com o objetivo de eliminar o efeito de ruído musical, característica que será explicada na seção 4.6.3.

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

O parâmetro R_{prio}, relação sinal-ruído *a priori*, é avaliada pela relação recursiva em (4.39) e é o parâmetro dominante em (4.35), como pode-se ver na Figura 4.2. 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 4.2). 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 musical, como será detalhado mais adiante na seção 4.6.2.

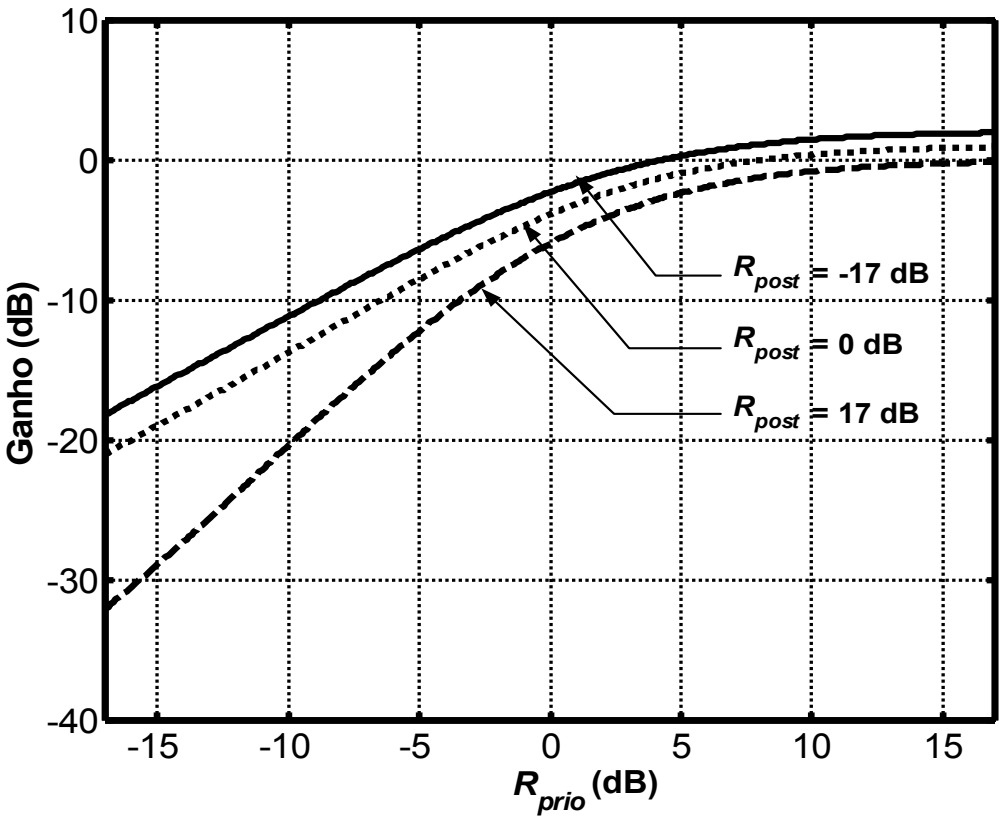

Figura 4.2: *O ganho EMSR versus R_{prio}, para diferentes valores de R_{post}.*

Um estudo experimental de (4.39) mostra dois comportamentos diferentes para a relação sinal-ruído *a priori*, os quais estão expressos na Figura 4.3.

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 uma dada frequência, é muito menor que a de R_{post} ao longo de sucessivos quadros. Pode-se notar, no lado esquerdo da Figura 4.3, 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 4.3, 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 à (4.39):

- Quando $R_{prio}(q.\omega)$ é alta, $G(q,\omega) \cong 1$ (lado direito da Figura 4.2), portanto temos:.

$$R_{prio}(q,\omega) \cong (1-\mu).Rpost(q,\omega) + \mu.\frac{|Y(q-1,\omega)|^2}{\hat{D}(\omega)^2} \tag{4.40}$$

- Como $R_{post}(q,\omega) >> 1$:.

$$R_{prio}(q,\omega) \cong (1-\mu).R_{post}(q,\omega) + \mu.R_{post}(q-1,\omega) \tag{4.41}$$

- E considerando que μ é geralmente escolhido próximo a 1, a equação (4.39) pode ser aproximada por:.

$$R_{prio}(q,\omega) \cong (\mu).Rpost(q-1,\omega) \tag{4.42}$$

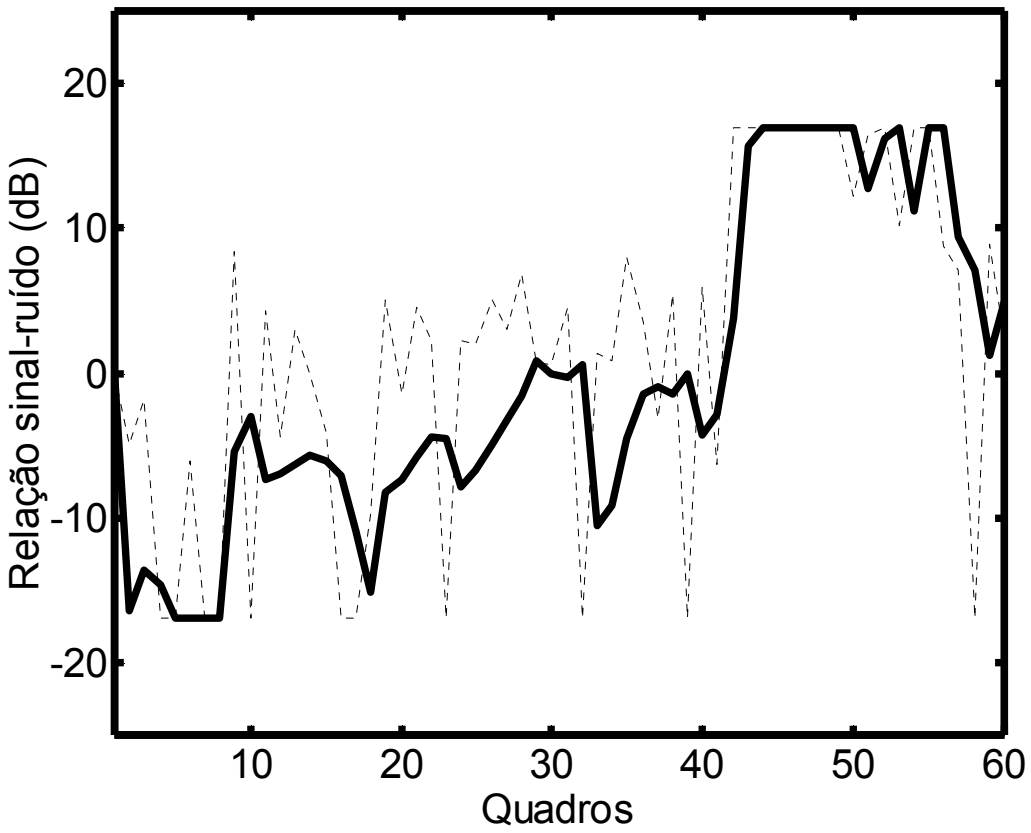

Figura 4.3: *As relações sinal-ruído R_{post} e R_{prio} ao longo de sucessivos quadros. Curva de linha contínua: R_{prio} ; Curva de linha pontilhada: R_{post}. Nos 40 primeiros quadros, o sinal contém somente ruído na frequência escolhida e para os 20 quadros seguintes, surge uma componente com mais de 15 dB de relação sinal-ruído na frequência mostrada.*

Quando o nível do sinal é bem superior ao do ruído, como ocorre nos últimos 20 quadros da Figura 4.3, $R_{prio}(q,\omega)$ 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.

4.6 A eliminação do ruído musical

4.6.1 A suavização de R_{prio}

Nas regiões do espectro que correspondem somente a ruído, a suavização de R_{prio} permite reduzir os efeitos do ruído musical, pois R_{post} tende a $-\infty$ em média e R_{prio} tem uma variância reduzida. Como a função ganho $G(q,\omega)$ de EMSR depende principalmente de R_{prio}, o ganho $G(q,\omega)$ não vai apresentar grandes variações ao longo de sucessivos quadros. Desta forma, o ruído musical, que é composto de componentes senoidais que aparecem e desaparecem rapidamente ao longo dos quadros, é significativamente reduzido.

A idéia de calcular a atenuação do espectro de tempo curto através de uma média ao longo de sucessivos quadros já foi explorada anteriormente por BOLL [1], mas a não-linearidade do procedimento adotado por Ephraim e Malah garantiu um desempenho muito superior.

4.6.2 Desacordo entre R_{post} e R_{prio}

Outro efeito será útil na eliminação do ruído musical: nas regiões de apenas ruído, constatamos que o valor médio de R_{prio} está em média bem abaixo de 0 dB (como vemos na Figura 4.3), portanto, é improvável que qualquer medida de relação sinal-ruído atinja altos valores. Por esta razão, nas situações em que R_{post} é alta e R_{prio} é baixa, a atenuação é aumentada fortemente, conforme é observado no lado esquerdo da Figura 4.2. Desta forma, componentes do espectro atual maiores que o nível médio do ruído são fortemente atenuadas. Esta característica é muito útil nos casos em que o ruído de fundo é não-estacionário, evitando que surja ruído musical quando o ruído exceder suas características médias.

4.6.3 A influência do parâmetro μ

É necessário que haja um compromisso entre o grau de suavização de R_{prio} nas regiões de puro ruído e o nível de distorção causado ao sinal quando ocorre uma transição para uma região de alta relação sinal-ruído. Em simulações, nota-se que em trechos de uma locução onde há apenas ruído, R_{prio} possui:

- valor médio proporcional a $(1-\mu)$, se μ for próximo a 1 (maior que 0.9)

- desvio padrão proporcional a $(1-\mu)$, se μ for próximo a 1 (maior que 0.9)

Desta forma, μ deverá ser o mais próximo possível de 1 para evitar o aparecimento de ruído musical. Mas, por outro lado, quando surge uma componente do sinal de fala a EMSR reage rapidamente fazendo o parâmetro ganho $G(q,\omega)$ sair de um valor baixo e se aproximar de 1, desde que a relação sinal-ruído da componente do sinal seja maior que $1/(1-\mu)$. As simulações têm mostrado que, para componentes de sinal com menor relação sinal-ruído, R_{prio} leva mais tempo para atingir seu valor final, o que resulta em uma atenuação indesejada das componentes de baixa amplitude do sinal.

Podemos compreender a atuação do parâmetro μ ao comparar os gráficos apresentados na Figura 4.4, onde o valor médio de R_{prio} é: (a) -9.5 dB, (b) -16 dB e (c) -6.8 dB. Pode-se vizualizar, também, a diferença de variâncias: em (b) a variância é claramente bastante reduzida. Estes resultados, a princípio, nos levam a classificar a escolha em (b), $\mu = 0.998$, como aquela que provê melhores resultados, pois nos quadros de ruído

- R_{prio} médio é menor, o que leva a uma atenuação maior do ruído, como visto na Figura 4.2.
- Menor variância de R_{prio} implica em menor variância do parâmetro ganho $G(q,\omega)$, o que ajuda a evitar ruído musical.

No entanto, pode-se observar na Figura 4.5 (ampliação do trecho transitório das curvas da Figura 4.4) o atraso significativo ocorrido em (b), entre o surgimento da componente do sinal de fala e o momento em que $R_{prio}(q,\omega)$ assume um valor significantivo, acima de 0 dB. Isto é resultado da relação $1/(1-\mu)$ citada acima e faz com que a componente do sinal seja atenuada de forma incorreta nos primeiros quadros de alta relação sinal ruído. Na prática, o uso de valores tão próximos a 1 como visto em (b), resultarão em distorções audíveis nos trechos de transição de silêncio para fala.

4.6.4 A limitação de R_{prio}

Um último procedimento é utilizado para diminuir ainda mais qualquer ruído musical remanescente:

$$R_{min} \leq R_{prio} \leq R_{max} \tag{4.43}$$

$$R_{min} \leq R_{post} \leq R_{max} \tag{4.44}$$

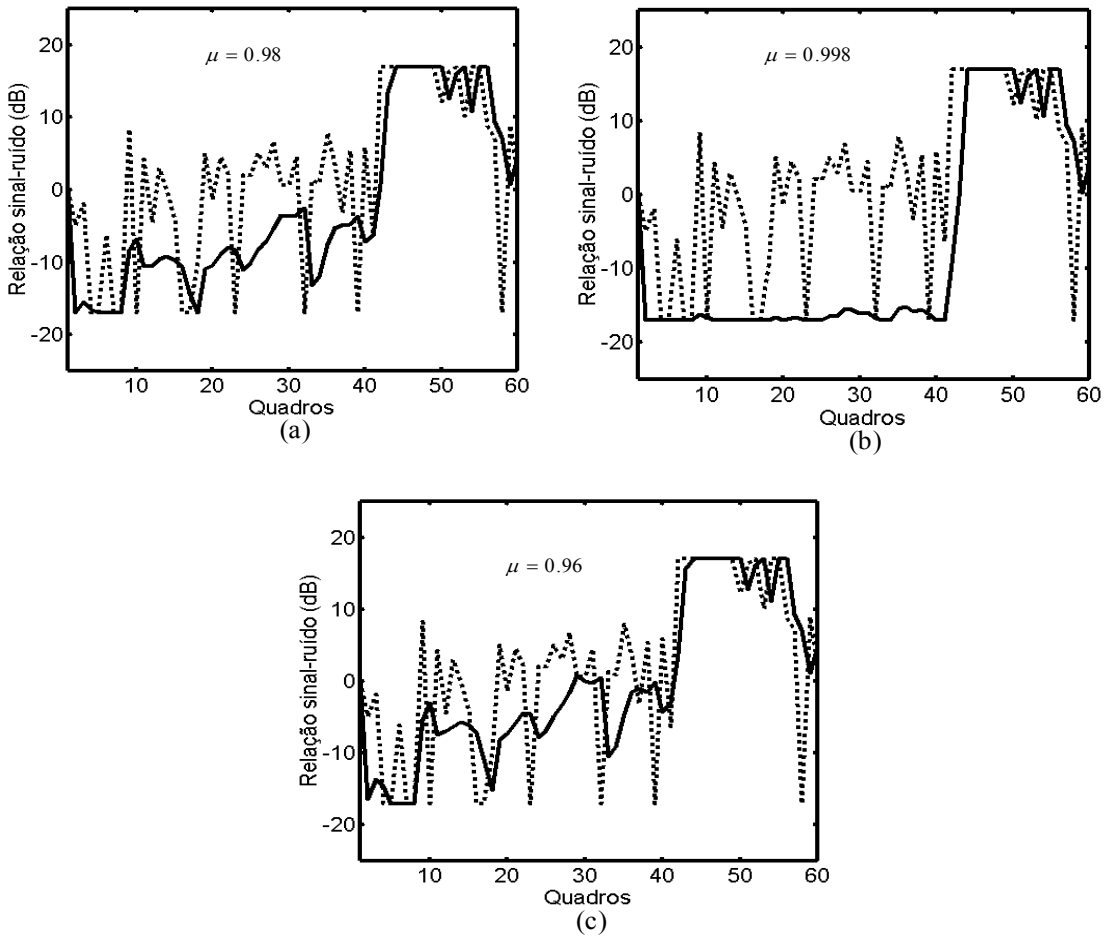

Figura 4.4: *As relações sinal-ruído R_{post} e R_{prio} ao longo de sucessivos quadros. Curva de linha contínua: R_{prio} ; Curva de linha pontilhada: R_{post}. (a) $\mu = 0.98$, (b) $\mu = 0.998$ e (c) $\mu = 0.96$.*

Qualquer irregularidade nos valores de R_{prio} e R_{post}, que poderia gerar ruído musical, é contornada ao limitar-se as fronteiras máximas e mínimas de R_{prio} e R_{post}. Na prática, o que foi constatado foram valores isolados de R_{prio} e R_{post} que tendiam a $+\infty$ ou a $-\infty$, resultando em um valor discrepante de ganho e, consequentemente, o aparecimento de ruído musical. Para implementação eficiente em hardware, é possível tabular os valores do ganho G para R_{prio} e R_{post} entre R_{min} e R_{max}, sem necessidade de realizar o cálculo de G a cada quadro e para cada freqüência.

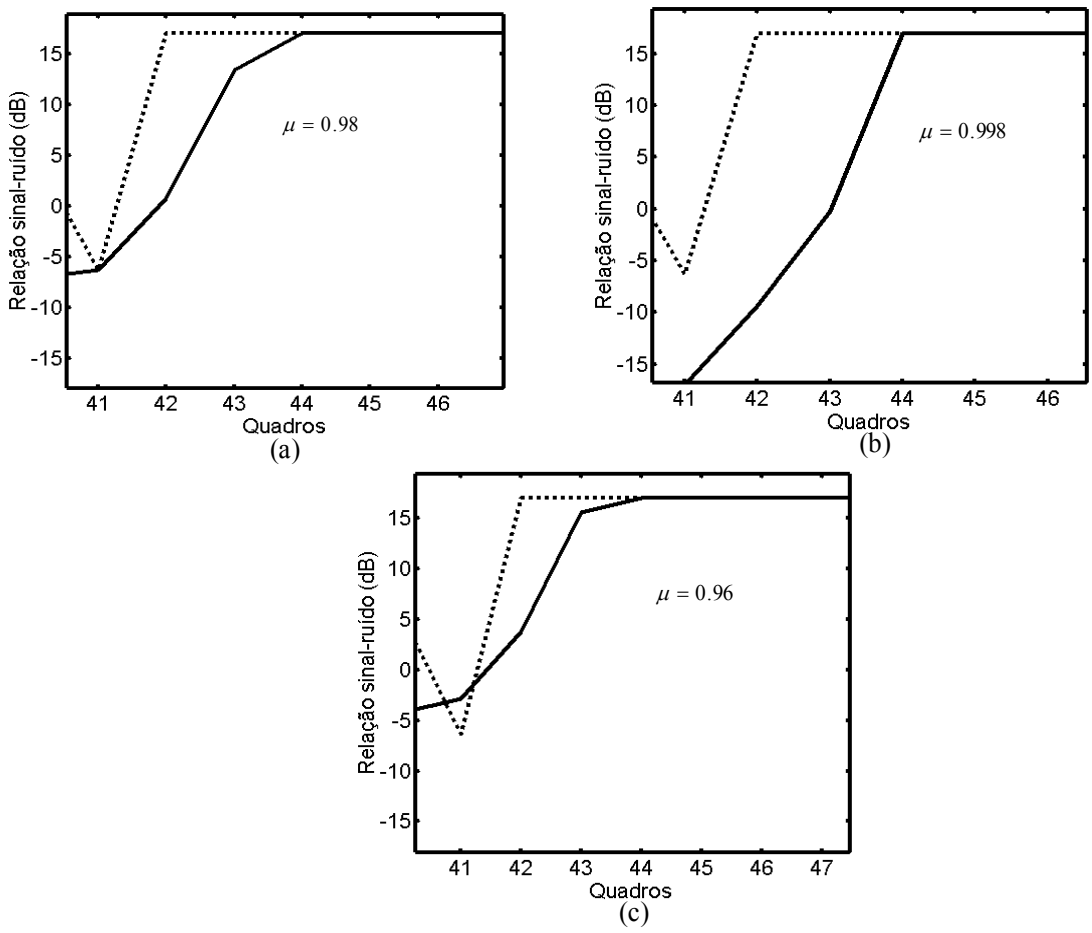

Figura 4.5: *As relações sinal-ruído R_{post} e R_{prio} ao longo dos quadros de transição. Curva de linha contínua: R_{prio} ; Curva de linha pontilhada: R_{post}. (a) $\mu = 0.98$, (b) $\mu = 0.998$ e (c) $\mu = 0.96$.*

4.7 A implementação do algoritmo de Ephraim e Malah

O algoritmo foi implementado segundo apresentado no diagrama em blocos da Figura 4.6. Esta implementação foi realizada em MatLab, sendo esta parte importante deste trabalho por possibilitar uma completa compreensão deste algoritmo em todos seus detalhes. As equações (4.38), (4.39) e (4.35) correspondem ao núcleo deste algoritmo.

A base de dados utilizada será descrita no capítulo 5, seção 5.6.1, assim como os tipos de ruído sob análise.

4.7.1 Características do sistema

As características básicas deste sistema são as mesmas do algoritmo apresentado no Capítulo 3 (seção 3.5.1), visto que os resultados de desempenho dos algoritmos do Capítulo 3 e 4 serão utilizados para verificar o desempenho do algoritmo proposto neste trabalho, o qual terá estas mesmas características básicas.

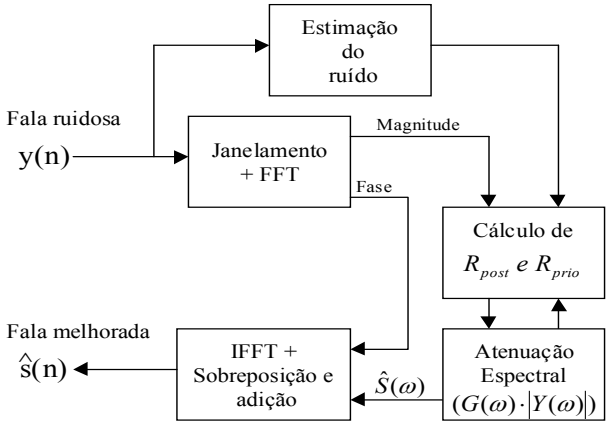

Figura 4.6: *Diagrama em blocos do algoritmo original proposto por Ephraim e Malah.*

4.7.2 Escolha de parâmetros

Os parâmetros básicos utilizados foram:

- $\mu = 0.98$ (em 4.39); pois segundo CAPPÈ [25] este valor apresentou o melhor compromisso entre a eliminação de ruído de fundo, a diminuição de ruído musical e a distorção introduzida no sinal. Foram realizados testes também para $\mu = 0.96$ e $\mu = 0.998$ e foi confirmado a obtenção de melhores resultados para $\mu = 0.98$ na solução de EPHRAIM & MALAH. Os resultados obtidos para $\mu = 0.98$ e $\mu = 0.96$ nesta solução são apresentados nas tabelas 5.4, 5.5 e 5.6 no Capítulo 5 para efeito de comparação de desempenho com o algoritmo proposto.

- Para limitar as fronteiras de R_{prio} e R_{post}, $-17dB \leq R_{post} \leq 17dB$ e $-17dB \leq R_{prio} \leq 17dB$. Valores escolhidos empiricamente, proporcionando melhores resultados, diferentemente dos limites de -15 e 15 dB propostos por CAPPÈ [25].

4.7.3 Resultados

Os resultados de desempenho deste algoritmo serão posteriormente utilizados quando comparados com o algoritmo proposto neste trabalho, no Capítulo 5.

Os espectros apresentados na Figura 4.7 ilustram o desempenho do algoritmo EMSR. Pode-se verificar no espectrograma (d) que o sinal resultante do algoritmo EMSR elimina grande parte do ruído do sinal original sem introduzir ruído musical, diferentemente do que é visto no algoritmo PSS em (c), onde o ruído musical pode ser identificado como "ilhas" isoladas no espectrograma. No entanto, ainda permanece parte do ruído original no sinal, o qual deverá ser reduzido pelo algoritmo proposto neste trabalho.

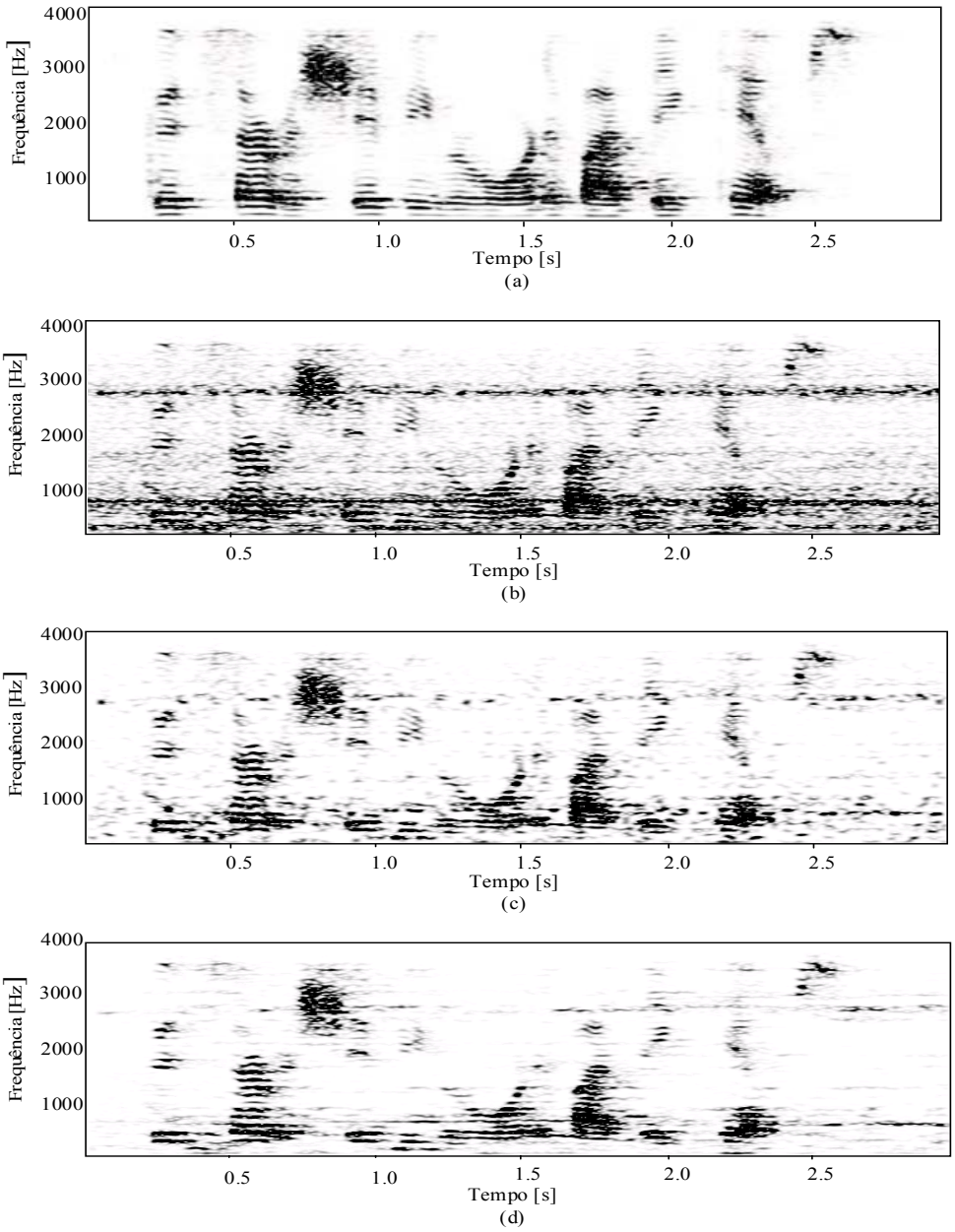

Figura 4.7: *Espectrogramas: (a)Sinal de fala limpa da frase "Good service should be rewarded by big tips", (b)Sinal de fala ruidoso (corrompido por ruído aditivo no interior da cabine de uma aeronave F16), (c)Sinal resultante da Subtração espectral - PSS e (d)Sinal resultante do algoritmo EMSR.*

Capítulo 5

Algoritmo de supressão de ruído proposto

5.1 Motivação

O método proposto por Ephraim e Malah e detalhado no capítulo anterior mostrou muita eficiência em reduzir ruído de fundo sem introdução de ruído musical. No entanto, como já mencionado anteriormente, ele permite apenas uma moderada redução de ruído, o que é insuficiente quando a relação sinal-ruído é muito baixa (SNR < 10 dB), ou seja, uma quantidade considerável do ruído original permanece no sinal melhorado. Por esta razão, neste trabalho é proposto um algoritmo cujo núcleo está baseado nas regras de Ephraim e Malah, mas com modificações que permitirão obter melhor desempenho no caso de sinais ruidosos com relações sinal-ruído baixas, usando o conceito de limiar de mascaramento auditivo explicado no Capítulo 3.

5.2 A função Ganho

O algoritmo proposto também é um tipo de algoritmo de atenuação espectral de tempo curto. A função ganho $G(q,\omega)$ utilizada é a mesma que foi apresentada no capítulo anterior em (4.35), que, para efeitos didáticos, é apresentada novamente:

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

$$M[\Theta] = \exp(-\frac{\Theta}{2})[(1+\Theta)I_0(\frac{\Theta}{2}) + \Theta I_1(\frac{\Theta}{2})] \tag{5.1}$$

O ganho espectral $G(q,\omega)$, assim como no algoritmo EMSR do capítulo 4, é aplicado sobre cada componente espectral ruidosa de tempo curto $Y(q,\omega)$.

No entanto, no algoritmo proposto neste capítulo, as relações sinal-ruído *a posteriori* e *a priori*, que são básicas para o cálculo de $G(q,\omega)$, foram modificadas, sendo agora derivadas por meio das seguintes relações:

$$R_{post}(q,\omega) = \frac{|Y(q,\omega)|^2}{\alpha(q,\omega)|\hat{D}(\omega)|^2} - 1 \tag{5.2}$$

$$\begin{aligned} R_{prio}(q,\omega) &= (1-\mu)R_{post}(q,\omega)] + \mu.\nu.G^2(q-1,\omega).\frac{|Y(q-1,\omega)|^2}{\alpha(q,\omega)|\hat{D}(\omega)|^2} + \\ &\quad \mu.(1-\nu).G^2(q-2,\omega)\frac{|Y(q-2,\omega)|^2}{\alpha(q,\omega).|\hat{D}(\omega)|^2} \end{aligned} \tag{5.3}$$

onde μ e ν foram, experimentalmente, escolhidos como 0.96 e 0.75, respectivamente. O parâmetro $\alpha(q,\omega)$ representa o parâmetro perceptual de atenuação, cujo cálculo detalhado foi apresentado no Capítulo 3.

5.3 Alterações introduzidas pelo novo algoritmo

São apresentadas a seguir as justificativas para as alterações introduzidas nos parâmetros R_{post} e R_{prio}, que garantiram maior eficiência ao algoritmo proposto.

5.3.1 Introdução do *parâmetro perceptual de atenuação* $\alpha(q,\omega)$

O parâmetro perceptual de atenuação $\alpha(q,\omega)$, que é empregado no algoritmo proposto, é utilizado seguindo a mesma idéia de BEROUTI [2]. Como apresentado no capítulo 2, 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. Em seguida VIRAG [10], como visto no Capítulo 3, aplica o parâmetro $\alpha(q,\omega)$ na equação de BEROUTI

(2.5), também com o objetivo de realizar uma atenuação adicional através de sobre-estimação do ruído. Porém, ela propõe que $\alpha(q,\omega)$ varie de acordo com o limiar de mascaramento do ruído, tornando a regra de atenuação espectral mais flexível e mais relacionada com as características do ouvido humano, obtendo assim melhores resultados que BEROUTI. No entanto, ambos algoritmos ainda mantiveram ruído musical no sinal de fala melhorado.

Por esta razão, no desenvolvimento do algoritmo proposto neste capítulo, optou-se por utilizar o parâmetro perceptual de atenuação $\alpha(q,\omega)$, que continua sendo determinado com base no limiar de mascaramento do ruído e calculado como apresentado no capítulo 3. No entanto, este parâmetro é utilizado agora nas equações (5.2) e (5.3), que na verdade são adaptações do método EMSR, devido à eficiência deste na eliminação de ruído musical. Nestas equações, quando o parâmetro $\alpha(q,\omega)$ 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. A determinação do parâmetro $\alpha(q,\omega)$, a cada quadro, depende diretamente do limiar de mascaramento do ruído $T(q,\omega)$, exatamente como apresentado no Capítulo 3. Quando $T(q,\omega)$ é alto, significa que o nível de ruído naquela frequência pode ser também alto porque ele será mascarado pela fala. Assim, $\alpha(q,\omega)$ deve ser colocado em baixo nível para evitar distorções audíveis na fala. Ao contrário, quando $T(q,\omega)$ é baixo, mesmo baixos níveis de ruído, na frequência ω, serão perceptualmente incômodos. Desta forma, $\alpha(q,\omega)$ deve ter um nível mais alto com o objetivo de reduzir ao máximo o ruído residual.

Os parâmetros básicos, escolhidos empiricamente para o cálculo de $\alpha(q,\omega)$ em (3.10) , com melhor compromisso entre ruído residual e distorção da fala, foram:

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

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

5.3.2 Consideração do quadro $(q-2)$ no cálculo de R_{prio}

Outra importante diferença entre o algoritmo proposto e o padrão EMSR é a presença do terceiro termo na equação do parâmetro R_{prio} em (5.3). No capítulo anterior, foi explicada a importância da suavização da variação de R_{prio} ao longo de sucessivos quadros, a fim de promover a eliminação do ruído musical. Com o

intuito de aumentar ainda mais este efeito de suavização, a nossa proposta foi a de introduzir o terceiro termo da equação (5.3), 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 $(q-2)$.

Com o objetivo de melhorar ainda mais o desempenho do algoritmo, foi feita a tentativa de incluir o quadro $(q-3)$ na determinação de R_{prio}. Mas, como não houve melhoria de desempenho, esta última alteração não foi implementada na versão final do algoritmo.

5.4 Resultados obtidos pelo algoritmo proposto

O efeito de ambas as alterações, apresentadas na seção anterior, pode ser visualizado comparando as curvas apresentadas na Figura 5.1. É apresentado o comportamento de R_{post} e R_{prio} para o algoritmo EMSR na Figura 5.1(a) e para o algoritmo proposto na Figura 5.1(b).

Pode-se constatar que, nos primeiros 40 quadros do sinal, na Figura 5.1(b) R_{prio} tem um valor médio muito menor que o valor médio observado nos mesmos primeiros 40 quadros do sinal na Figura 5.1(a). E, como foi mostrado na Figura 4.2, apresentada novamente como Figura 5.2, o parâmetro R_{prio} é o parâmetro dominante na determinação do Ganho $G(q,\omega)$. Portanto, quando se verifica que no algoritmo proposto o valor R_{prio} médio é menor que o obtido pelo algoritmo EMSR nos trechos de ruído, isto significa que o algoritmo proposto executa uma atenuação muito mais forte da fala ruidosa nas frequências em que há somente ruído, se comparado com a técnica EMSR.

Observa-se que o efeito de suavização da variação de R_{prio} no algoritmo proposto também aumentou, ou seja, nos primeiros 40 quadros, a curva de R_{prio} apresentada em 5.1(b) é ainda mais suave que a curva apresentada em 5.1(a). Conseqüentemente, o algoritmo proposto provoca uma menor variação do Ganho $G(q,\omega)$ e, com isso, uma maior redução do fenômeno do ruído musical, como já foi explicado anteriormente.

O desempenho comparativo desta solução, com relação ao método EMSR padrão, será apresentado mais adiante, mensurando a melhoria obtida no sistema proposto.

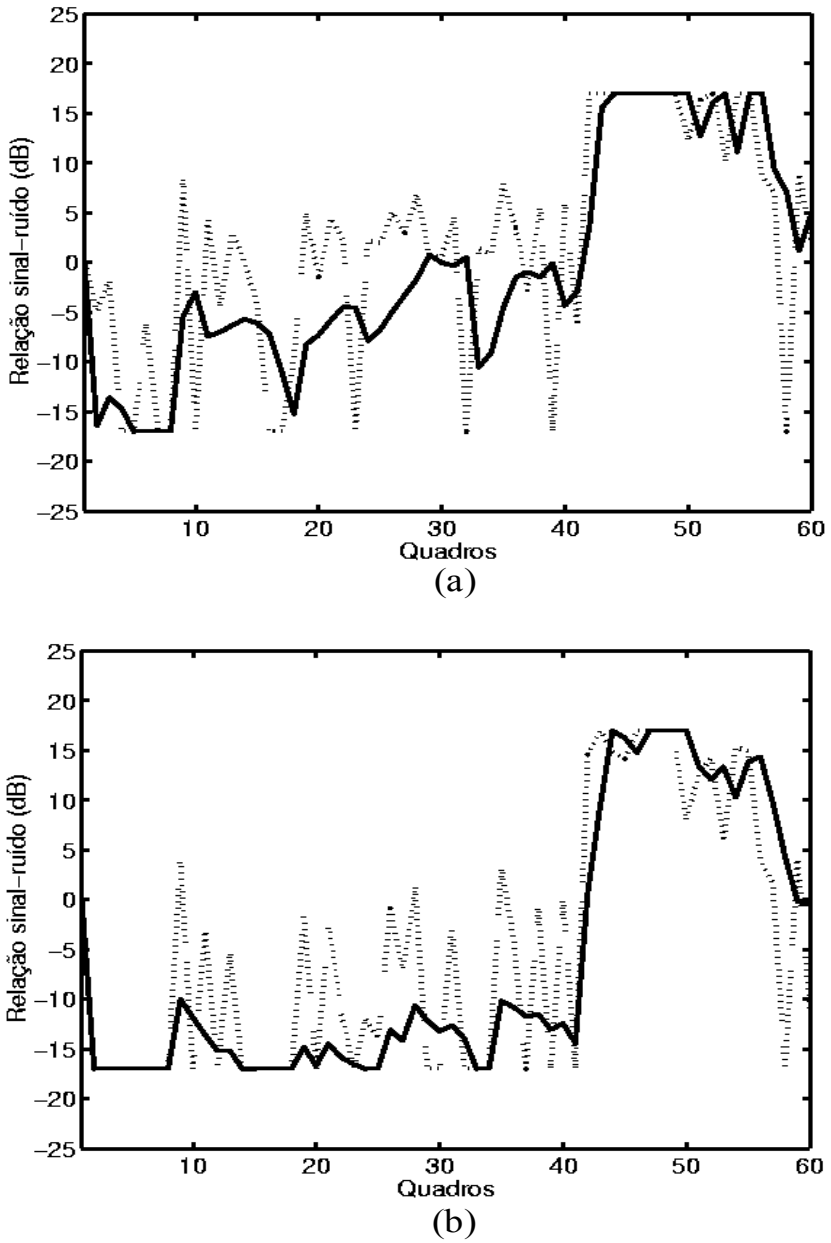

Figura 5.1: *(a)Gráfico extraído do Capítulo 4. As relações sinal-ruído R_{post} e R_{prio} ao longo de sucessivos quadros do algoritmo original EMSR. Curva de linha contínua: R_{prio} e Curva de linha pontilhada: R_{post}. (b)As relações sinal-ruído R_{post} e R_{prio} ao longo de sucessivos quadros do algoritmo proposto, calculados por (5.2) e (5.3). Curva de linha contínua: R_{prio} e Curva de linha pontilhada: R_{post}. As curvas (a) e (b) foram obtidas na mesma locução e na mesma frequência ω.*

5.5 Diagrama em blocos do algoritmo proposto

Uma visão completa do algoritmo proposto está apresentada no diagrama em blocos da Figura 5.3.

O sistema de melhoria proposto é composto pelos seguintes passos principais:

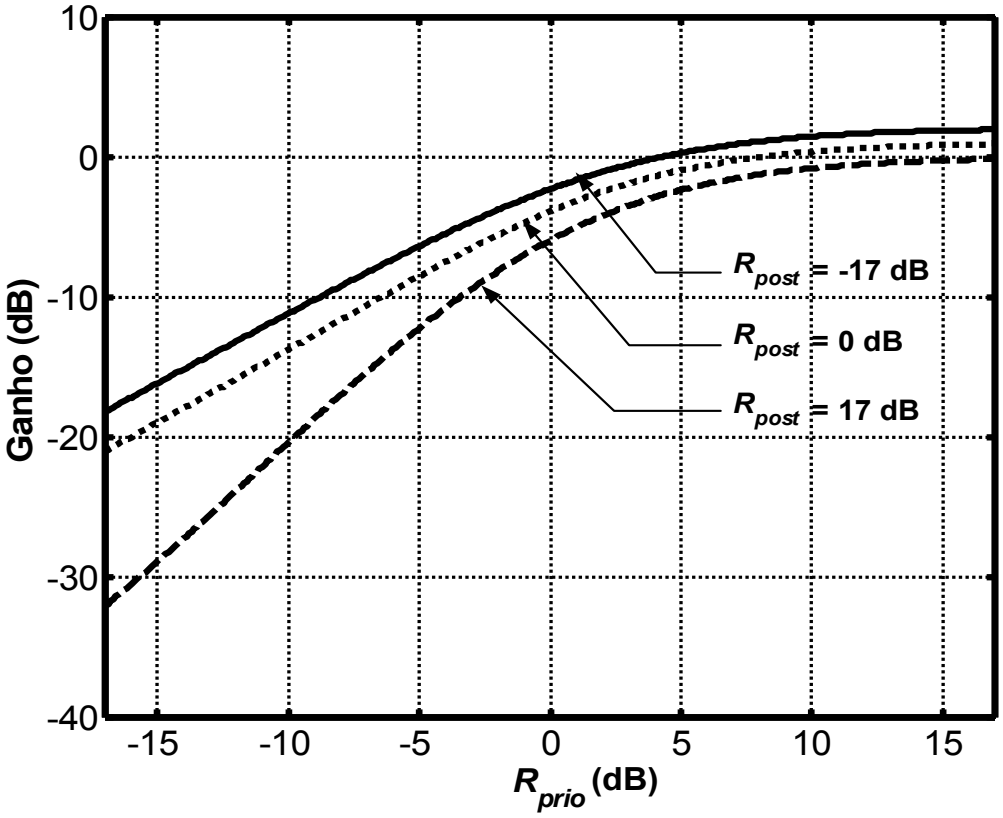

Figura 5.2: *O ganho EMSR versus a R_{prio}, para diferentes valores de R_{post}.*

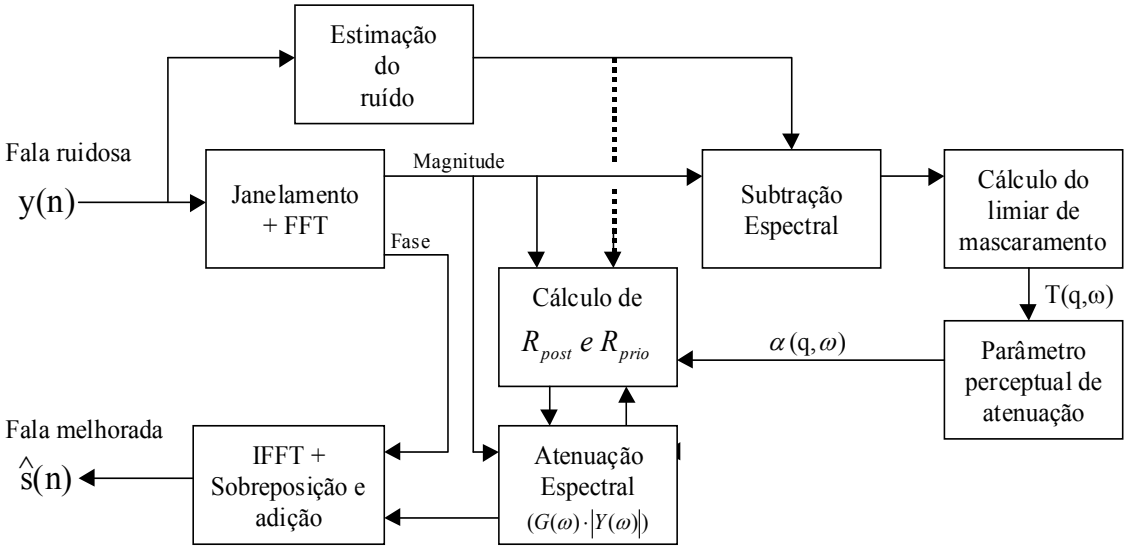

Figura 5.3: *Diagrama em blocos do algoritmo proposto.*

1. Análise espectral.
2. Estimação do ruído médio.
3. Cálculo do limiar de mascaramento do ruído $T(q, \omega)$.
4. Adaptação no tempo e na frequência do fator perceptual de atenuação $\alpha(q, \omega)$ baseado no limiar $T(q, \omega)$.
5. Cálculo de R_{post} e R_{prio}, de acordo com (5.2) e (5.3), usando o fator perceptual de atenuação $\alpha(q, \omega)$.
6. Cálculo da magnitude espectral da fala usando a função ganho $G(q, \omega)$ baseado em R_{post} e R_{prio}.
7. Transformada inversa de Fourier.

A determinação do limiar $T(q, \omega)$, como já foi mencionado no Capítulo 3, deve ser feita a partir de um sinal de fala limpa. Contudo, o sistema proposto dispõe apenas do sinal de fala ruidoso inicial. Então, é feita uma estimativa aproximada do sinal de fala limpa através de um simples esquema de subtração espectral de potência (PSS) para o cálculo de $T(q, \omega)$, conforme ilustrado no diagrama em blocos da Figura 5.3.

5.6 Descrição do sistema

5.6.1 Base de dados utilizada

A maioria das bases de dados de sinais de fala disponíveis hoje em domínio público foram projetadas para experimentos de reconhecimento de fala. Elas contêm locuções de um grande número de pessoas em ambientes livres de ruído ou em ambientes ruidosos, mas sua estrutura está direcionada para a aplicação de reconhecimento de fala. Neste trabalho, foi utilizada a base de dados SpEAR[1] [37], já mencionada nos capítulos 3 e 4, mas não detalhada. Esta é uma ferramenta especificamente desenvolvida para avaliar o desempenho de algoritmos de melhoria de sinais de fala, visto que tanto os sinais ruidosos quanto os sinais limpos de referência estão disponíveis. Nesta base de dados, os sinais ruidosos foram obtidos somando acusticamente o sinal limpo e o ruído em um ambiente controlado. Os sinais limpos foram retirados da base de dados TIMIT, mas eles foram reproduzidos e depois regravados juntamente com os sinais ruidosos. Vários

[1] Speech Enhancement Assessment Resource (SpEAR) Database. Beta Release v1.0. CSLU, Oregon Graduate Institute of Science and Technology. E. Wan, A. Nelson, and Rick Peterson.

tipos de ruído em vários níveis foram combinados com a fala limpa, totalizando 33 arquivos .wav ruidosos. A Tabela 5.1 e a Tabela 5.2 apresentam as locuções utilizadas e os diferentes tipos de ruído que foram considerados para teste do algoritmo.

Tabela 5.1: *Tabela das locuções originais utilizadas para teste*

Nome do arquivo(.wav)	Frase	Locutor
bigtips / tips	*Good service should be rewarded by big tips*	Masculino
peaches	*The fifth jar contains big, juicy peaches.*	Masculino
draw	*Draw every outer line first, then fill in the interior.*	Feminino
butter	*Butterscotch fudge goes well with vanilla ice cream.*	Masculino
vega	Vocal por Susan Vega	Feminino

Tabela 5.2: *Tipos de ruído combinados às locuções apresentadas na Tabela 5.1.*

Tipo de Ruído	Especificação do Ruído
F16	Corresponde ao ruído no interior da cabine de uma aeronave F16[1].
Factory	Corresponde ao ruído de fundo presente em uma fábrica de carros.
Pink	Corresponde ao ruído rosa.
White	Corresponde ao ruído branco.
Volvo	Corresponde ao ruído no interior de um automóvel Volvo[2].

[1] F16 a uma velocidade de 500 nós e altitude de 300-600 pés.

[2] No asfalto, a 120 Km/h, quarta marcha e em condições chuvosas.

5.6.2 Ferramenta para análise do desempenho do algoritmo proposto

Com o objetivo de comparar o desempenho do algoritmo proposto com outros algoritmos já existentes, foi realizada uma avaliação objetiva da qualidade da fala melhorada usando a pontuação PESQ (*Perceptual Evaluation of Speech Quality*). Esta ferramenta foi padronizada pelo ITU (*International Telecommunications Union*) como um modelo de avaliação perceptual de qualidade de fala que apresenta bom desempenho para diversas aplicações, como por exemplo, para avaliação de codificadores de voz ou em testes de redes fim a fim de diversos tipos. Este modelo foi aprovado pelo ITU-T na Recomendação P.862, em fevereiro de

2001 [38]. Os trabalhos mais recentes têm utilizado esta medida por ser bastante confiável para avaliação de desempenho de sistemas de melhoria de sinais de fala [40].

As pontuações PESQ apresentam uma correlação de mais de 95% com a medida subjetiva MOS[2] (Mean Opinion Score), com a vantagem de ser bem mais fácil de obter, pois os métodos subjetivos, como o MOS, dependem de muito tempo e de muitas pessoas [38].

O software PESQ utiliza a escala de opinião ACR (*absolute category rating*) [42] [43] apresentada na Tabela 5.3[3].

Tabela 5.3: *Escala de pontuação de qualidade ACR. Adaptado de [38]*

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

Para ilustrar como é utilizada a ferramenta PESQ, apresentamos a sua linha de comando na Figura 5.4.

Para utilização do algoritmo PESQ deve ser informada a taxa de amostragem (no nosso caso 8 kHz), o nome do arquivo que contém a locução de referência (que é a locução limpa) e por último a locução a ser avaliada.

A ferramenta é encontrada na página do ITU [39].

5.7 Avaliação do desempenho

A Figura 5.5 mostra que no algoritmo proposto, assim como visto no método EMSR do capítulo 4, não existem "ilhas" ao longo do espectrograma, o que caracteriza a ausência de ruído musical no sinal de fala melhorado.

No entanto, não é possível, apenas vizualizando o espectrograma, mensurar a melhoria obtida quando se compara o método proposto com outros métodos. Por esta razão, a comparação de desempenho do algoritmo proposto com outros

[2]O MOS é a medida subjetiva de qualidade de voz mais utilizada, principalmente para avaliação de algoritmos de codificação de fala. Este método consiste de uma pontuação de qualidade média a partir da avaliação de um grupo de ouvintes quanto à impressão subjetiva da qualidade da fala.

[3]Há uma pequena diferença na escala do PESQ com relação à escala ACR: a máxima pontuação do PESQ é 4,5 ao invés de 5 (obtém-se 4,5 ao comparar um sinal com ele mesmo)

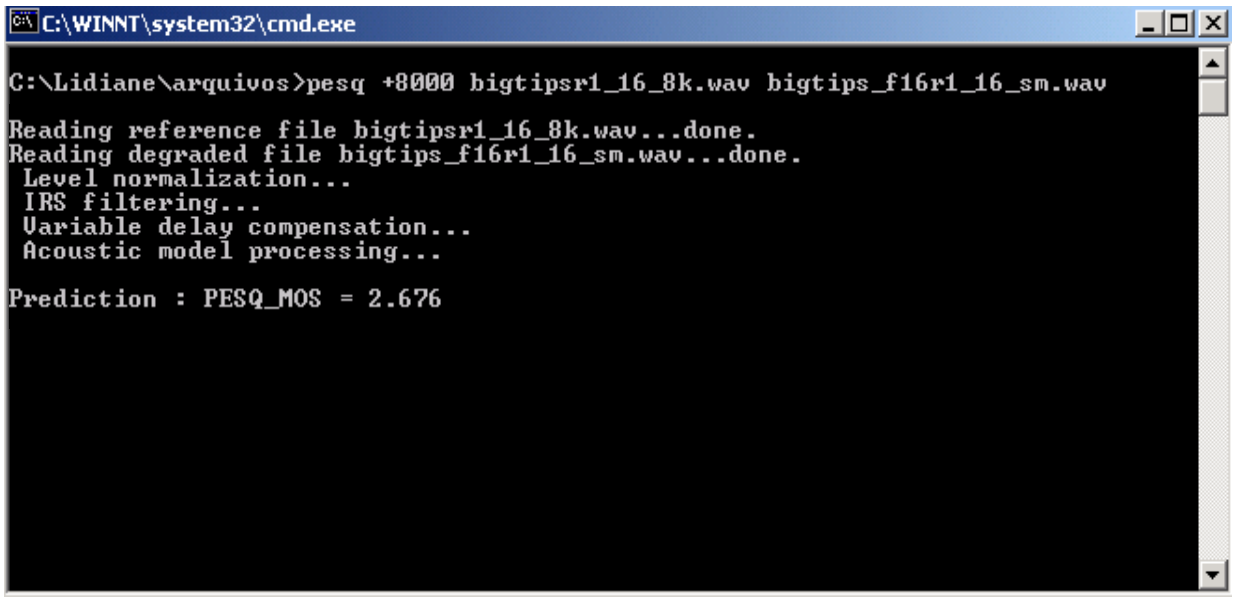

Figura 5.4: *Utilização da ferramenta PESQ.*

métodos será feito com base na pontuação PESQ que já foi apresentada em seção anterior. Os métodos utilizados nesta comparação são:

1. Método PSS (Power Spectral Subtraction)
2. Método NMT - PSS (Power Spectral Subtraction based on Noise Masking Threshold)
3. Método EMSR original ($\mu = 0.98$)[25]
4. Método EMSR ($\mu = 0.96$)
5. Método proposto neste trabalho

É importante ressaltar que além do algoritmo proposto, parte deste trabalho foi a implementação dos métodos acima citados, em MatLab, seguindo passo-a-passo as referências bibliográficas apresentadas ao longo dos capítulos. O método PSS apresentado no Capítulo 2, o NMT-PSS segundo o Capítulo 3, o EMSR seguindo os passos apresentados no Capítulo 4 e o método proposto de acordo com o apresentado no início deste capítulo. Nestas tabelas é apresentado também o limite teórico que , como mencionado no Capítulo 2, é o melhor resultado possível em qualquer algoritmo subtrativo, é obtido ao sintetizar a fala usando a magnitude espectral do sinal original limpo e a fase espectral do sinal ruidoso.

As tabelas 5.4, 5.5 e 5.6 apresentam resultados médios obtidos a partir dos 33 arquivos .wav ruidosos (escolhidos da base de dados SpEAR) com relações sinal-ruído nas faixas de 0 a 5 dB, 5 a 10 dB e 10 a 15 dB, respectivamente. A

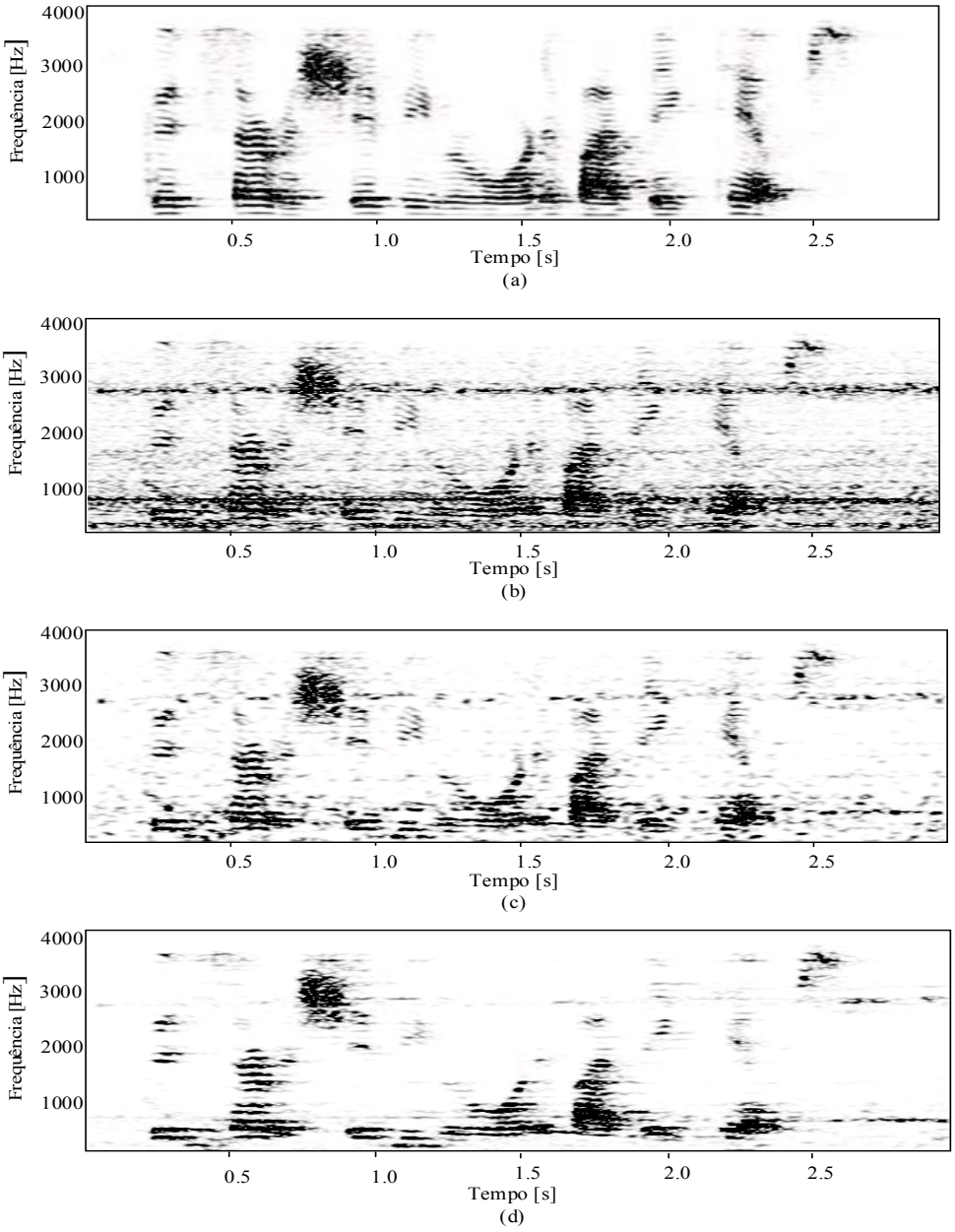

Figura 5.5: *Espectrogramas: (a)Sinal de fala limpa da frase "Good service should be rewarded by big tips", (b)Sinal de fala ruidoso (corrompido por ruído aditivo no inteior da cabine de uma aeronave F16), (c)Sinal resultante da Subtração espectral - PSS e (d)Sinal resultante do algoritmo proposto.*

tabela original contendo todos os resultados para cada uma das locuções escolhidas encontra-se no Anexo F. É importante ressaltar que foi utilizada a maior parte da base de dados, para uma ampla faixa de relações sinal-ruído. Evitou-se utilizar locuções em que a fala limpa tivesse um tamanho diferente da fala com ruído, o que impediria uma eficiente comparação de desempenho após o processamento.

Tabela 5.4: *Média de medidas PESQ-MOS para sinais de fala melhorados usando diferentes métodos. Sinais ruidosos originais com relação sinal-ruído entre 0 e 5 dB.*

Tipo de ruído	White	Pink	F16	Factory
Relação sinal-ruído média (dB)	(3,22dB)	(2,78dB)	(2,65dB)	(3,49dB)
Sem processamento (ruidoso)	1,980	1,917	2,049	2,414
PSS ($a = 2$)	2,250	2,126	2,229	2,550
NMT-PSS	2,416	2,286	2,356	2,628
EMSR ($\mu = 0.96$)	2,494	2,399	2,495	2,762
EMSR ($\mu = 0.98$)	2,505	2,404	2,496	2,765
Algoritmo proposto	**2,610**	**2,515**	**2,603**	**2,858**
Limite teórico	3,998	3,878	3,917	3,993

Tabela 5.5: *Média de medidas PESQ-MOS para sinais de fala melhorados usando diferentes métodos. Sinais ruidosos originais com relação sinal-ruído entre 5 e 10 dB.*

Tipo de ruído	Pink	F16	Volvo	Factory
Relação sinal-ruído média (dB)	(6,97dB)	(6,21dB)	(7,89dB)	(5,17dB)
Sem processamento (ruidoso)	1,878	2,194	3,182	2,213
PSS ($\alpha = 2$)	2,231	2,432	3,498	2,466
NMT-PSS	2,446	2,609	3,487	2,575
EMSR ($\mu = 0.96$)	2,502	2,775	3,679	2,635
EMSR ($\mu = 0.98$)	2,522	2,783	3,678	2,635
Algoritmo proposto	**2,668**	**2,894**	**3,707**	**2,768**
Limite teórico	3,791	3,985	4,212	3,922

Tabela 5.6: *Média de medidas PESQ-MOS para sinais de fala melhorados usando diferentes métodos. Sinais ruidosos originais com relação sinal-ruído entre 10 e 15 dB.*

Tipo de ruído	F16	Pink
Relação sinal-ruído média (dB)	(14,85dB)	(12,13dB))
Sem processamento (ruidoso)	2,647	2,500
PSS ($\alpha = 2$)	2,937	2,973
NMT-PSS	3,065	3,167
EMSR ($\mu = 0.96$)	3,288	3,268
EMSR ($\mu = 0.98$)	3,283	3,291
Algoritmo proposto	**3,300**	**3,420**
Limite teórico	4,117	4,062

Pode-se observar nas tabelas 5.4, 5.5 e 5.6, que, para todas as faixas de relação sinal-ruído, o resultado médio do algoritmo proposto foi superior aos resultados dos outros algoritmos. Foram introduzidos nas tabelas, para comparação, os resultados obtidos utilizando-se o método EMSR com o parâmetro $\mu = 0.96$, visto que este é o valor do parâmetro μ utilizado no algoritmo proposto.

5.7.1 Considerações da implementação computacional

A implementação foi efetuada em MatLab®, versão 5.3, utilizando um computador com processador Pentium 4 com 256Mbytes de memória. É importante ressaltar que o tempo de processamento das locuções ruidosas, pelo algoritmo proposto, sempre foi inferior ao tempo de duração das locuções. Desta forma, o algoritmo proposto pode ser usado em qualquer aplicação de comunicação em tempo real, pois mesmo em uma linguagem computacional interpretada, como o MatLab®, o tempo de processamento é baixo.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Conclusões

O trabalho desenvolvido nesta dissertação teve como objetivo principal apresentar um algoritmo de melhoria de fala (de canal único) degradada por ruído aditivo. Esse algoritmo foi desenvolvido para que fosse capaz de tratar com eficiência sinais ruidosos, incluindo aqueles de baixa relação sinal-ruído (SNR<10 dB). O sistema proposto foi baseado no método de supressão de ruído desenvolvido por Ephraim e Malah (EMSR), que consiste em estimar de forma otimizada a amplitude espectral de tempo curto (STSA), sob um critério MMSE e assumindo um determinado modelo estatístico. Neste trabalho, foi realizado um estudo estatístico e matemático detalhado de todas as expressões básicas apresentadas por Ephraim e Malah em seu método original, as quais foram apresentadas no Capítulo 4, bem como em diversos Anexos. Também foi feita a implementação prática do mesmo, para posterior comparação com o método proposto.

O algoritmo EMSR é muito eficiente na eliminação do fenômeno de ruído musical. No entanto, não realiza uma suficiente redução do ruído original para sinais de baixa relação sinal-ruído (<10dB). Por isso, o algoritmo proposto utilizou-se do conceito de mascaramento auditivo, introduzindo um parâmetro que permitisse uma atenuação adicional. Esse parâmetro foi denominado parâmetro de atenuação perceptual, que está diretamente ligado com o limiar de mascaramento do ruído, o qual foi detalhado no Capítulo 3. Isso resultou em uma solução com uma maior redução de ruído e tornou a solução mais correlacionada com a percepção auditiva humana.

A idéia de utilizar um parâmetro de atenuação dependente do limiar de mascaramento e aplicar sobre a equação de ganho de BEROUTI (2.5), como mostrado no Capítulo 3, foi apresentada por VIRAG [10]. Neste trabalho, foi realizada uma implementação similar à de VIRAG (NMT-PSS) para comparar o seu desempenho

com o desempenho do método proposto.

A contribuição principal deste trabalho consistiu em usar o parâmetro de atenuação espectral não mais na equação do ganho de Berouti (2.5), mas sim na equação do ganho do método EMSR, por meio da modificação das expressões de cálculo das relações sinal-ruído *a priori* e *a posteriori*.

Além de introduzir o parâmetro de atenuação perceptual no algoritmo proposto, a consideração do quadro $(q-2)$ no cálculo da relação sinal-ruído *a priori* (R_{prio}) permitiu dar a este parâmetro um maior efeito de suavização ao longo dos quadros, garantindo assim maior eficiência na eliminação do fenômeno do ruído musical.

O resultado deste algoritmo mostrou-se bastante eficiente na redução de ruído de fundo, quando comparado com alguns outros métodos de melhoria de fala existentes e na eliminação do ruído musical. Ao analisar o espectrograma resultante do algoritmo proposto, na Figura 5.5 é possível notar a inexistência do ruído musical, caracterizado nos espectrogramas como pequenas "ilhas ".Por meio dos resultados apresentados nas tabelas da última seção do Capítulo 5, através da pontuação PESQ, foi comprovada a eficiência do método proposto. Verificou-se que o método proposto é superior a todos os outros métodos analisados, considerando diversos tipos e níveis de ruído.

A base de dados utilizada, SpEAR database, garantiu confiabilidade aos resultados obtidos, visto fornecer tanto o sinal limpo quanto o sinal ruidoso para avaliação de desempenho através da ferramenta PESQ, detalhada na seção 5.6.2 do Capítulo 5. Sua maior desvantagem é não possuir uma grande quantidade de locuções, totalizando apenas 5 frases diferentes, como mostra a Tabela 5.1. No entanto, a base utilizada foi considerada suficiente pela variedade em tipos e níveis de ruído.

É importante destacar que o algoritmo proposto também se mostrou computacionalmente eficiente, podendo ser implementado para a utilização em sistemas de comunicação em tempo-real.

6.2 Trabalhos Futuros

Durante o desenvolvimento do sistema foi possível identificar algumas possibilidades de extensão deste trabalho, a fim de tentar melhorar ainda mais a qualidade do sinal. Estas sugestões são relatadas a seguir.

6.2.1 Introduzir um método de detecção de voz

Para trabalhar com sinais de fala de maior duração e ruídos menos estacionários seria necessário introduzir um método de detecção de sinal de voz (VAD - *Voice Activity Detection*). Deste modo seria possível detetar novos trechos de silêncio para estimar novamente o espectro do ruído, atualizando assim a estimativa de ruído que, devido à sua não estacionariedade, poderá ser bastante diferente ao longo do tempo. No estágio atual, o sistema usa apenas o início de cada frase para estimar o ruído.

6.2.2 Testar o aumento da eficiência de sistemas de reconhecimento de fala ruidosa

Pode-se testar a eficiência do algoritmo proposto quando utilizado antes de um sistema de reconhecimento de fala ruidosa. Esta seria uma outra forma de medir o desempenho do sistema desenvolvido. Neste caso, a etapa de síntese do sinal de fala poderia ser omitida, uma vez que os parâmetros utilizados nos sistemas de reconhecimento de fala poderiam ser obtidos diretamente no domínio da freqüência.

6.2.3 Modificar o modelo estatístico para amplitude do sinal de fala adotado por Ephraim e Malah.

No desenvolvimento do estimador de amplitude proposto por Ephraim e Malah, o modelo estatístico adotado para a amplitude espectral do sinal de fala (A_k) foi a Distribuição Rayleigh. A proposta seria utilizar, por exemplo, a distribuição Gama ou Laplaciana e comparar os resultados obtidos.

6.2.4 Propor um algoritmo baseado em outra solução de Ephraim e Malah.

Além do estimador de amplitude original de Ephraim e Malah, apresentado no Capítulo 4, estes mesmos autores desenvolveram outro estimador [27], que busca minimizar a seguinte distorção

$$E[logA_k - log\hat{A}_k]$$

resultando em

$$\hat{A}_K = exp\{E[lnA_k|Y_k]\}$$

Poder-se-ia desenvolver um algoritmo similar ao proposto neste trabalho, mas baseado neste estimador de amplitude, o qual minimiza o erro médio quadrático do espectro logarítmico.

6.2.5 Considerar a incerteza de presença de fala no sinal ruidoso.

No modelo original EMSR utilizado, considera-se que a fala está presente durante todo o sinal ruidoso. Ephraim e Malah propõem um outro método [8] que considera um estimador que é derivado assumindo que o sinal está presente nas observações ruidosas com probabilidade $p < 1$.

A proposta seria desenvolver um algoritmo que utilize este modelamento de estimador de amplitude e acrescente a ele o conceito de mascaramento auditivo de forma similar ao realizado neste trabalho.

Anexo A

Derivação da distribuição Rayleigh para as amplitudes espectrais

A distribuição Rayleigh é caracterizada pela *fdp*

$$f_R(r) = \begin{cases} \frac{r}{\sigma^2} exp \frac{-r^2}{2\sigma^2}, & r \geq 0 \\ 0, & r < 0 \end{cases} \tag{A.1}$$

Segundo mostra LATHI [36], uma variável aleatória Rayleigh pode ser derivada de duas variáveis aleatórias gaussianas independentes. Sendo estas variáveis partes real e imaginária, a e b, respectivamente, temos a representação cartesiana abaixo:

$$z = a + jb \rightarrow |z|^2 = a^2 + b^2$$

Além de a e b serem consideradas como gaussianas independentes, ambas possuem média zero e mesma variância. Portanto temos,

$$f_A(a) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\, \left(\frac{-a^2}{2\sigma^2}\right) \tag{A.2}$$

$$f_B(b) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\, \left(\frac{-b^2}{2\sigma^2}\right) \tag{A.3}$$

Então,

$$f_{AB}(a,b) = f_A(a)f_B(b) = \frac{1}{2\pi\sigma^2}\, exp\, \frac{-(a^2+b^2)}{2\sigma^2} \tag{A.4}$$

$$f_{AB}(a,b) = f_A(a)f_B(b) = \frac{1}{2\pi\sigma^2}\, exp\, \frac{-(|z|^2)}{2\sigma^2} \tag{A.5}$$

Como mostra a Figura A.1, podemos vizualizar a representação de $p(a,b)$:

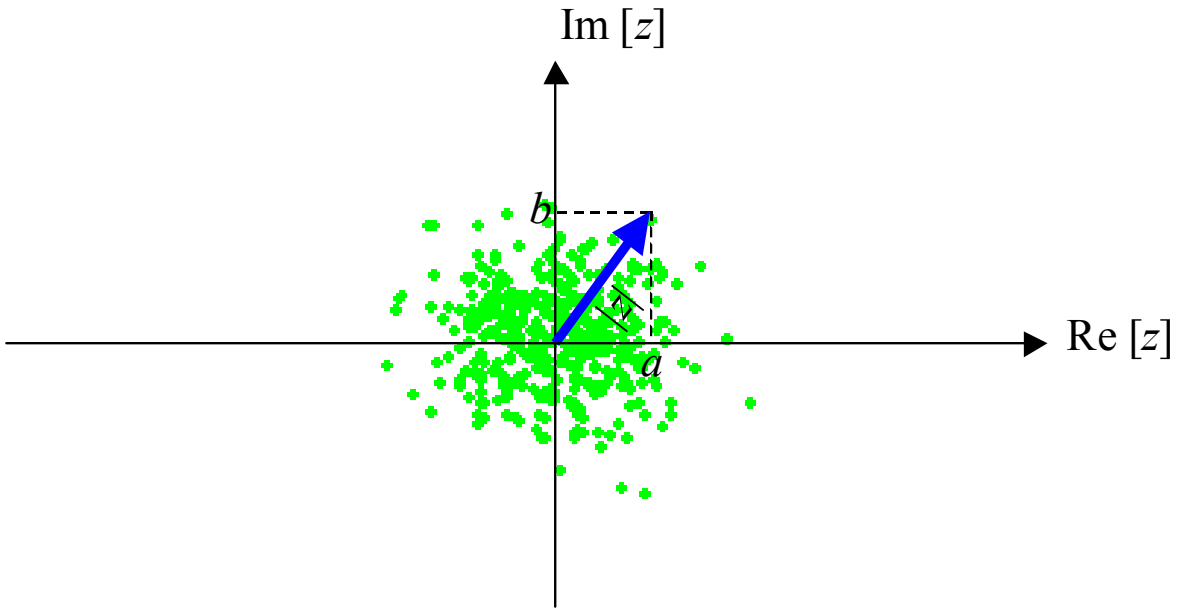

Figura A.1: *Representação da pdf conjunta* $p(a,b)$.

Os pontos no plano a e b podem ser descritos em coordenadas polares (r,θ) conforme Figura A.2, onde:

$$r = \sqrt{a^2+b^2} \tag{A.6}$$

$$\theta = tan^{-1}\left(\frac{b}{a}\right) \tag{A.7}$$

$$z = re^{j\theta} \tag{A.8}$$

Na Figura A.2(a) a região hachurada representa $r < \mathrm{r} \leq r+dr$ e $\theta < \Theta \leq \theta+d\theta$ (onde dr e $d\theta$ ambos $\rightarrow 0$). Por isso, se $f_{r\Theta}(r,\theta)$ é a FDP conjunta de r e Θ,

então a probabilidade da observação de r e Θ nesta região é $f_{r,\Theta}(r,\theta)drd\theta$. Mas sabe-se, também, que esta probabilidade é $f_{AB}(a,b)$ vezes a área $rdrd\theta$ da região hachurada.

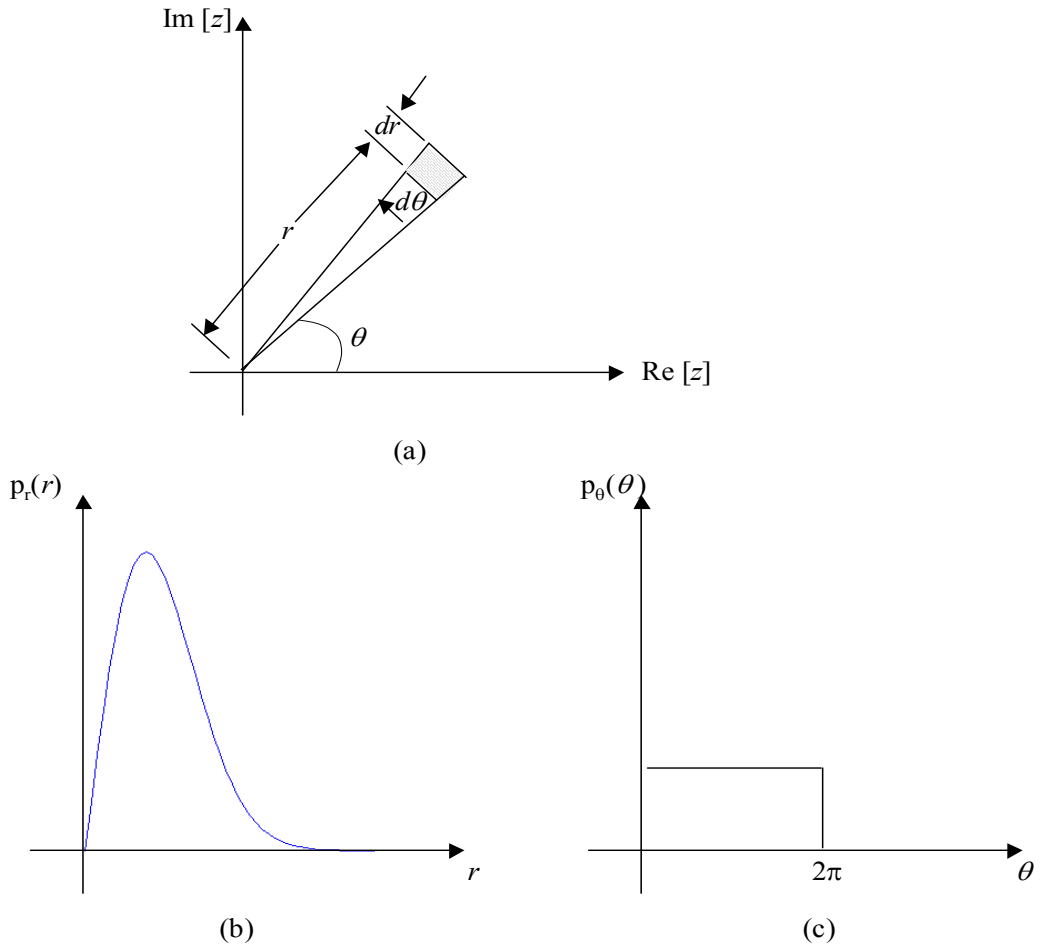

Figura A.2: *Derivação da densidade Rayleigh.*

Então:

$$\frac{1}{2\pi\sigma^2}\, exp\, \frac{-(a^2+b^2)}{2\sigma^2} rdrd\theta = f_{r\Theta}(r,\theta)drd\theta \tag{A.9}$$

E:

$$\begin{aligned} f_{r\Theta}(r,\theta) &= \frac{r}{2\pi\sigma^2}\, exp\, \frac{-(a^2+b^2)}{2\sigma^2} \\ &= \frac{r}{2\pi\sigma^2}\, exp\, \frac{-r^2}{2\sigma^2} \end{aligned} \tag{A.10}$$

e também:

$$f_r(r) = \int_{-\infty}^{+\infty} p_{r\Theta}(r,\theta)d\theta \tag{A.11}$$

Como Θ existe apenas na região $(0, 2\pi)$,

$$\begin{aligned} f_r(r) &= \int_0^{2\pi} \frac{r}{2\pi\sigma^2} exp\left(\frac{-r^2}{2\sigma^2}\right) d\theta \\ &= \frac{r}{\sigma^2} exp\left(\frac{-r^2}{2\sigma^2}\right), r \geq 0 \end{aligned} \tag{A.12}$$

Note que r é sempre maior que 0. Da mesma forma, encontra-se

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 \leq \Theta < 2\pi \\ 0, & \text{caso contrário} \end{cases} \tag{A.13}$$

A *fdp* $p_r(r)$ é uma função densidade Rayleigh. Ambos, $f_r(r)$ e $f_\Theta(\theta)$, são mostrados na Figura A.2 (b) e (c), respectivamente.

Anexo B

Estimador da amplitude MMSE (I)

ESPERANÇA CONDICIONAL:

A Esperança Condicional pode ser vista como uma função de x:$g(x) = E[Y|x]$ [41]. Portanto faz sentido considerar uma variável aleatória $g(X) = E[Y|X]$. Desta forma, pode-se imaginar que um experimento aleatório é executado e um valor para X é obtido, dizemos que $X = x_0$, e finalmente o valor $g(x_0) = E[Y|x_0]$ é produzido. Estamos interessados em $E[g(X)] = E[E[Y|X]]$. Em particular vamos provar que:

$$E[Y] = E[E[Y|X]] \tag{B.1}$$

O lado direito é representado por:

$$E[E[Y|X]] = \int_{-\infty}^{+\infty} E[Y|X] f_X(x) dx \quad ; \text{X contínuo} \tag{B.2}$$

A equação (B.1) é provada considerando que X e Y são variáveis aleatórias conjuntamente contínuas:

$$\begin{aligned} E[g(X)] = E[E[Y|X]] &= \int_{-\infty}^{+\infty} E[Y|X] f_X(x) dx \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f_Y(y|x) dy f_X(x) dx \\ &= \int_{-\infty}^{+\infty} y \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{+\infty} y f_Y(y) dy = E[Y] \end{aligned} \tag{B.3}$$

MÍNIMO ERRO MÉDIO QUADRÁTICO:
O objetivo é minimizar:

$$E[(Y - g(X))^2] \tag{B.4}$$

Primeiro consideremos o caso onde $g(X)$ é limitado a uma função linear de X [41]:

$$minE[(Y - a)^2] = E[Y^2] - 2aE[Y] + a^2 \tag{B.5}$$

O melhor valor de a é obtido tomando a derivada com relação a a. Temos então,

$$a = E[Y] \tag{B.6}$$

Consideremos agora o caso onde g(X) é uma função não-linear. Utilizando (B.1), temos

$$\begin{aligned} E[(Y - g(x))^2] &= E[E[(Y - g(x))^2|X] \\ &= \int_{-\infty}^{+\infty} E[(Y - g(x))^2|X = x] f_X(x) dx \end{aligned} \tag{B.7}$$

O integrando acima é positivo para todo x. Portanto, a integral é minimizada minimizando $E[(Y - g(x))^2|X = x]$ para cada x. Mas $g(x)$ é uma constante à medida que a esperança condicional está envolvida. Então o problema é equivalente a (B.5) e a "constante" que minimiza $E[(Y - g(x))^2|X = x]$ é:

$$g(x) = E[Y|X = x] \tag{B.8}$$

Anexo C

Estimador da amplitude MMSE (II)

$$E\{A_k|Y_k\} = \int_{-\infty}^{+\infty} a_k.p(a_k|Y_k)da_k \tag{C.1}$$

A regra de Bayes mostra que:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} = \frac{P(A,B)}{P(B)} \tag{C.2}$$

Baseado na regra de Bayes, mostrada em (C.2), temos então:

$$p(a_k|Y_k) = \frac{p(Y_k|a_k).p(a_k)}{p(Y_k)} = \frac{p(a_k,Y_k)}{p(Y_k)} \tag{C.3}$$

Substituindo (C.3) em (C.1):

$$E\{A_k|Y_k\} = \frac{1}{p(Y_k)} \int_{-\infty}^{+\infty} a_k.p(a_k,Y_k)da_k \tag{C.4}$$

Através do teorema de função densidade de probabilidade marginal, sabemos que:

$$p(a_k,Y_k) = \int_{-\infty}^{+\infty} p(a_k,\alpha_k,Y_k)d\alpha_k \tag{C.5}$$

Substituindo (C.5) em (C.4), temos:

$$E\{A_k|Y_k\} = \frac{1}{p(Y_k)} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a_k.p(a_k,\alpha_k,Y_k)d\alpha_k da_k \tag{C.6}$$

Utilizando Bayes novamente, obtemos:

$$E\{A_k|Y_k\} = \frac{1}{p(Y_k)} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a_k.p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k \tag{C.7}$$

Vamos desenvolver agora $p(Y_k)$, que é o denominador em (C.7):
Utilizando, novamente, o teorema das fdp's marginais:

$$p(Y_k) = \int_{-\infty}^{+\infty} p(Y_k, a_k)da_k \tag{C.8}$$

E, mais uma vez, aplicando a regra de Bayes:

$$p(Y_k, a_k) = \int_{-\infty}^{+\infty} p(Y_k, a_k, \alpha_k)d\alpha_k = \int_{-\infty}^{+\infty} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k \tag{C.9}$$

Substituindo (C.9) em (C.8), obtemos

$$p(Y_k) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k \tag{C.10}$$

E, finalmente, substituindo (C.10) em (C.7), temos

$$E\{A_k|Y_k\} = \frac{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a_k.p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k} \tag{C.11}$$

Como a magnitude espectral a_k é sempre positiva e a fase α_k varia entre 0 e 2π, alteramos os limites das integrais, resultando em:

$$E\{A_k|Y_k\} = \frac{\int_{0}^{+\infty} \int_{0}^{2\pi} a_k.p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k}{\int_{0}^{+\infty} \int_{0}^{2\pi} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k} \tag{C.12}$$

Anexo D

Estimador da amplitude MMSE (III)

Deseja-se mostrar a igualdade:

$$\frac{R_k}{\lambda_d(k)} = \sqrt{\frac{v_k}{\lambda(k)}} \tag{D.1}$$

Substituindo (4.28) e (4.29) em (4.27):

$$v_k = \frac{\frac{\lambda_s(k)}{\lambda_d(k)}}{1+\frac{\lambda_s(k)}{\lambda_d(k)}}.\frac{R_k^2}{\lambda_d(k)} = \frac{\frac{\lambda_s(k)}{\lambda_d(k)}}{\lambda_d(k)+\lambda_s(k)}.R_k^2 = \frac{\lambda_s(k)}{\lambda_s(k)+\lambda_d(k)}.\frac{R_k^2}{\lambda_d(k)}$$

$$R_k = \sqrt{\lambda_d(k)[\lambda_s(k)+\lambda_d(k)]\frac{v_k}{\lambda_s(k)}} = \sqrt{\left(\lambda_d(k)+\frac{\lambda_d^2(k)}{\lambda_s(k)}\right)v_k}$$

Considerando a equação (4.23):

$$R_k = \sqrt{\left[\lambda_d(k)+\lambda_d^2(k)\left(\frac{1}{\lambda(k)}-\frac{1}{\lambda_d(k)}\right)\right]v_k} = \sqrt{\left[\lambda_d(k)+\frac{\lambda_d^2(k)}{\lambda(k)}-\lambda_d(k)\right]v_k} = \lambda_d(k)\sqrt{\frac{v_k}{\lambda(k)}}$$

$$\frac{R_k}{\lambda_d(k)} = \sqrt{\frac{v_k}{\lambda(k)}} \tag{D.2}$$

A forma apresentada em (4.30) é resultado da igualdade apresentada em (D.2).

Anexo E

Estimador da amplitude MMSE (IV)

Para desenvolvimento de (4.30), utilizaremos algumas fórmulas. Primeiramente, (6.631.1) de [32]:

$$\int_0^\infty x^\mu e^{-\alpha x^2} J_\nu(\beta x) dx = \frac{\beta^\nu \Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\mu + \frac{1}{2}\right)}{2^{\nu+1}\alpha^{\frac{1}{2}(\mu+\nu+1)}\Gamma(\nu+1)} \, {}_1F_1\left(\frac{\nu+\mu+1}{2}; \nu+1; -\frac{\beta^2}{4\alpha}\right) \tag{E.1}$$

onde ${}_1F_1(a;b;x)$ é a função confluente hipergeométrica.

Existem outras notações usadas para a função ${}_1F_1(a;b;x)$, incluindo $F(\alpha;\beta;x)$, $M(a;b;z)$ e $\Phi(\alpha;\beta;z)$.

A função confluente hipergeométrica $\Phi(\alpha;\beta;z)$ é definida por (9.210.1) em [32]:

$$\Phi(\alpha;\beta;z) = 1 + \frac{\alpha}{\beta}\frac{z}{1!} + \frac{\alpha(\alpha+1)}{\beta(\beta+1)}\frac{z^2}{2!} + \frac{\alpha(\alpha+1)(\alpha+2)}{\beta(\beta+1)(\beta+2)}\frac{z^3}{3!} + ... \tag{E.2}$$

Também será utilizada a fórmula (8.406.3) de [32], que é a relação entre a função de Bessel modificada de enésima ordem e a correspondente função de Bessel:

$$I_n(z) = j^{-n} J_n(jz) \tag{E.3}$$

onde $j = \sqrt{-1}$

E, finalmente, (9.212.1) de [32]:

$$\Phi(\alpha;\gamma;z) = e^z\,\Phi(\gamma-\alpha;\gamma;-z) \tag{E.4}$$

Usando (E.3) em (4.30), temos que:

$$I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}a_k\right) = J_0\left(j.2\sqrt{\frac{v_k}{\lambda_k}}a_k\right) \tag{E.5}$$

Para utilizar (E.1) no numerador de (4.30), façamos:
$x = a_k$; $\mu = 2$; $\alpha = 1/\lambda(k)$; $\beta = 2j\sqrt{\frac{v_k}{\lambda_k}}$ e consequentemente, $\beta^2 = -4\frac{v_k}{\lambda_k}$
Para utilizar-se (E.1) no denominador de (4.30), façamos:
$x = a_k$; $\mu = 1$; $\alpha = 1/\lambda(k)$; $\beta = 2j\sqrt{\frac{v_k}{\lambda_k}}$ e consequentemente, $\beta^2 = -4\frac{v_k}{\lambda_k}$
Desenvolve-se então (4.30) utilizando (E.1):

$$\hat{A}_k = \frac{\dfrac{\Gamma(1.5)}{2.(\lambda(k))^{-3/2}.\Gamma(1)}.{}_1F_1(1.5\ 1\ v_k)}{\dfrac{\Gamma(1)}{2.(\lambda(k))^{-1}\Gamma(1)}.{}_1F_1(1\ 1\ v_k)} \tag{E.6}$$

Através da equação (E.4), temos:

$$\hat{A}_k = \Gamma(1.5).\lambda(k)^{1/2}.\frac{exp(v_k).{}_1F_1(-0.5;1;-v_k)}{exp(v_k).{}_1F_1(0;1;-v_k)} \tag{E.7}$$

Mas ${}_1F_1(0;1;-v_k) = 1$ (basta substituir $\alpha = 0$ em (E.2)).
Então temos que:

$$\hat{A}_k = \Gamma(1.5).\lambda(k)^{1/2}.{}_1F_1(-0.5;1;-v_k) \tag{E.8}$$

Podemos apresentar (E.8) com o formato final apresentado por Ephraim e Malah [8]:

$$\hat{A}_k = \Gamma(1.5).\frac{\sqrt{v_k}}{\gamma_k}.{}_1F_1(-0.5;1;-v_k)R_k \tag{E.9}$$

DEMONSTRAÇÃO DA EQUIVALÊNCIA ENTRE (E.8) e (E.9):
Desejamos mostrar que:

$$\lambda(k)^{1/2} = \frac{\sqrt{v_k}}{\gamma_k}.R_k \tag{E.10}$$

Para isso serão utilizadas (4.27), (4.28), (4.29) e (4.23), resultando em:

$$\begin{aligned}\frac{\sqrt{v_k}}{\gamma_k}.R_k &= \frac{\sqrt{\frac{\xi_k}{1+\xi_k}.\gamma_k}}{\gamma_k}.R_k = \frac{\sqrt{\frac{\xi_k}{1+\xi_k}}}{\sqrt{\gamma_k}}.R_k = \frac{\sqrt{\frac{\frac{\lambda_s(k)}{\lambda_d(k)}}{1+\frac{\lambda_s(k)}{\lambda_d(k)}}}}{\sqrt{\frac{R_k^2}{\lambda_d(k)}}}.R_k = \frac{\sqrt{\frac{\frac{\lambda_s(k)}{\lambda_d(k)}}{\frac{\lambda_s(k)+\lambda_d(k)}{\lambda_d(k)}}}}{\frac{R_k}{\sqrt{\lambda_d(k)}}}.R_k \\ &= \frac{\sqrt{\frac{\lambda_s(k)}{\lambda_d(k)}}.\sqrt{\frac{\lambda_d(k)}{\lambda_s(k)+\lambda_d(k)}}}{\frac{1}{\sqrt{\lambda_d(k)}}} = \sqrt{\frac{\lambda_s(k).\lambda_d(k)}{\lambda_s(k)+\lambda_d(k)}} = \sqrt{\lambda(k)} \qquad \text{(E.11)}\end{aligned}$$

Anexo F

Resultados obtidos para cada locução

Tabela F.1: *Tabela que apresenta os resultados de pontuação PESQ-MOS obtidos para cada uma das 33 locuções ruidosas (em arquivo.wav) utilizando diferentes métodos de melhoria de sinais de fala.*

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
1. bigtips + ruído F16 (SNR = 2,74)	
bigtips + ruído F16 (Amostrado em 8kHz)	2,004
Método Proposto	**2,687**
Método EMSR ($\mu = 0.98$)	2,548
Método EMSR ($\mu = 0.96$)	2,540
Método NMT-PSS	2,394
Método PSS	2,252
Limite Teórico	3,921
2. bigtips + ruído Factory (SNR = 3,33)	
bigtips + ruído Factory (Amostrado em 8kHz)	2,337
Método Proposto	**2,842**
Método EMSR ($\mu = 0.98$)	2,759
Método EMSR ($\mu = 0.96$)	2,748
Método NMT-PSS	2,592
Método PSS	2,515
Limite Teórico	3,997

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
3. bigtips + ruído rosa (SNR = 2,16)	
bigtips + ruído rosa (Amostrado em 8kHz)	1,909
Método Proposto	**2,661**
Método EMSR ($\mu = 0.98$)	2,520
Método EMSR ($\mu = 0.96$)	2,519
Método NMT-PSS	2,423
Método PSS	2,257
Limite Teórico	3,946
4. bigtips + ruído branco (SNR = 2,37)	
bigtips + ruído branco (Amostrado em 8kHz)	1,905
Método Proposto	**2,676**
Método EMSR ($\mu = 0.98$)	2,536
Método EMSR ($\mu = 0.96$)	2,517
Método NMT-PSS	2,437
Método PSS	2,212
Limite Teórico	4,017
5. bigtips + ruído Volvo (SNR = 7,22)	
bigtips + ruído Volvo (Amostrado em 8kHz)	3,267
Método Proposto	**3,632**
Método EMSR ($\mu = 0.98$)	3,678
Método EMSR ($\mu = 0.96$)	3,690
Método NMT-PSS	3,432
Método PSS	3,526
Limite Teórico	4,174
6. butter + ruído F16 (SNR = 2,09)	
butter + ruído F16 (Amostrado em 8kHz)	2,255
Método Proposto	**2,724**
Método EMSR ($\mu = 0.98$)	2,610
Método EMSR ($\mu = 0.96$)	2,611
Método NMT-PSS	2,488
Método PSS	2,353
Limite Teórico	4,043

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
7. butter + ruído Factory (SNR = 3,07)	
butter + ruído Factory (Amostrado em 8kHz)	2,460
Método Proposto	**2,885**
Método EMSR ($\mu = 0.98$)	2,784
Método EMSR ($\mu = 0.96$)	2,787
Método NMT-PSS	2,697
Método PSS	2,579
Limite Teórico	4,067
8. butter + ruído Pink (SNR = 2,09)	
butter + ruído Pink (Amostrado em 8kHz)	2,003
Método Proposto	**2,534**
Método EMSR ($\mu = 0.98$)	2,371
Método EMSR ($\mu = 0.96$)	2,386
Método NMT-PSS	2,261
Método PSS	1,997
Limite Teórico	4,015
9. draw + ruído F16 (SNR = 4,35)	
draw + ruído Pink (Amostrado em 8kHz)	2,023
Método Proposto	**2,435**
Método EMSR ($\mu = 0.98$)	2,340
Método EMSR ($\mu = 0.96$)	2,353
Método NMT-PSS	2,289
Método PSS	2,221
Limite Teórico	3,839
10. draw + ruído Factory (SNR = 5,17)	
draw + ruído Factory (Amostrado em 8kHz)	2,213
Método Proposto	**2,768**
Método EMSR ($\mu = 0.98$)	2,635
Método EMSR ($\mu = 0.96$)	2,635
Método NMT-PSS	2,575
Método PSS	2,466
Limite Teórico	3,922

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
11. draw + ruído Pink (SNR = 3,87)	
draw + ruído Pink (Amostrado em 8kHz)	1,923
Método Proposto	**2,510**
Método EMSR ($\mu = 0.98$)	2,388
Método EMSR ($\mu = 0.96$)	2,388
Método NMT-PSS	2,296
Método PSS	2,190
Limite Teórico	3,881
12. draw + ruído White (SNR = 4,08)	
draw + ruído White (Amostrado em 8kHz)	2,054
Método Proposto	**2,543**
Método EMSR ($\mu = 0.98$)	2,473
Método EMSR ($\mu = 0.96$)	2,470
Método NMT-PSS	2,394
Método PSS	2,288
Limite Teórico	3,978
13. draw + ruído Volvo (SNR = 8,79)	
draw + ruído Volvo (Amostrado em 8kHz)	3,149
Método Proposto	**3,791**
Método EMSR ($\mu = 0.98$)	3,762
Método EMSR ($\mu = 0.96$)	3,760
Método NMT-PSS	3,591
Método PSS	3,612
Limite Teórico	4,238
14. peaches + ruído F16 (SNR = 3,26)	
draw + ruído F16 (Amostrado em 8kHz)	2,294
Método Proposto	**2,611**
Método EMSR ($\mu = 0.98$)	2,608
Método EMSR ($\mu = 0.96$)	2,603
Método NMT-PSS	2,368
Método PSS	2,354
Limite Teórico	3,960

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
15. peaches + ruído Factory (SNR = 4,07)	
peaches + ruído Factory (Amostrado em 8kHz)	2,446
Método Proposto	**2,846**
Método EMSR ($\mu = 0.98$)	2,752
Método EMSR ($\mu = 0.96$)	2,751
Método NMT-PSS	2,596
Método PSS	2,556
Limite Teórico	3,914
16. peaches + ruído Pink (SNR = 2,77)	
peaches + ruído Pink (Amostrado em 8kHz)	2,169
Método Proposto	**2,623**
Método EMSR ($\mu = 0.98$)	2,553
Método EMSR ($\mu = 0.96$)	2,545
Método NMT-PSS	2,367
Método PSS	2,295
Limite Teórico	3,914
17. peaches + ruído Volvo (SNR = 7,68)	
peaches + ruído Volvo (Amostrado em 8kHz)	3,131
Método Proposto	**3,699**
Método EMSR ($\mu = 0.98$)	3,595
Método EMSR ($\mu = 0.96$)	3,586
Método NMT-PSS	3,439
Método PSS	3,357
Limite Teórico	4,224
18. tips + ruído F16 (SNR = 0,37)	
tips + ruído F16 (Amostrado em 8kHz)	1,754
Método Proposto	**2,471**
Método EMSR ($\mu = 0.98$)	2,325
Método EMSR ($\mu = 0.96$)	2,322
Método NMT-PSS	2,183
Método PSS	1,986
Limite Teórico	3,841

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
19. tips + ruído F16 (SNR = 3,26)	
tips + ruído F16 (Amostrado em 8kHz)	1,965
Método Proposto	**2,688**
Método EMSR ($\mu = 0.98$)	2,546
Método EMSR ($\mu = 0.96$)	2,538
Método NMT-PSS	2,413
Método PSS	2,207
Limite Teórico	3,889
20. tips + ruído F16 (SNR = 5,23)	
tips + ruído F16 (Amostrado em 8kHz)	2,117
Método Proposto	**2,822**
Método EMSR ($\mu = 0.98$)	2,692
Método EMSR ($\mu = 0.96$)	2,688
Método NMT-PSS	2,555
Método PSS	2,355
Limite Teórico	3,956
21. tips + ruído F16 (SNR = 7,20)	
tips + ruído F16 (Amostrado em 8kHz)	2,271
Método Proposto	**2,966**
Método EMSR ($\mu = 0.98$)	2,874
Método EMSR ($\mu = 0.96$)	2,862
Método NMT-PSS	2,662
Método PSS	2,508
Limite Teórico	4,013
22. tips + ruído F16 (SNR = 9,17)	
tips + ruído F16 (Amostrado em 8kHz)	2,423
Método Proposto	**3,112**
Método EMSR ($\mu = 0.98$)	3,057
Método EMSR ($\mu = 0.96$)	3,055
Método NMT-PSS	2,817
Método PSS	2,681
Limite Teórico	4,066

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
23. tips + ruído F16 (SNR = 10,16)	
tips + ruído F16 (Amostrado em 8kHz)	2,496
Método Proposto	**3,176**
Método EMSR ($\mu = 0.98$)	3,136
Método EMSR ($\mu = 0.96$)	3,135
Método NMT-PSS	2,905
Método PSS	2,763
Limite Teórico	4,082
24. tips + ruído F16 (SNR = 12,14)	
tips + ruído F16 (Amostrado em 8kHz)	2,644
Método Proposto	**3,308**
Método EMSR ($\mu = 0.98$)	3,290
Método EMSR ($\mu = 0.96$)	3,295
Método NMT-PSS	3,060
Método PSS	2,944
Limite Teórico	4,124
25. tips + ruído F16 (SNR = 14,11)	
tips + ruído F16 (Amostrado em 8kHz)	2,802
Método Proposto	**3,416**
Método EMSR ($\mu = 0.98$)	3,423
Método EMSR ($\mu = 0.96$)	3,435
Método NMT-PSS	3,229
Método PSS	3,104
Limite Teórico	4,144
26. tips + ruído F16 (SNR = 16,08)	
tips + ruído F16 (Amostrado em 8kHz)	2,985
Método Proposto	**3,495**
Método EMSR ($\mu = 0.98$)	3,524
Método EMSR ($\mu = 0.96$)	3,550
Método NMT-PSS	3,350
Método PSS	3,267
Limite Teórico	4,178

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
27. tips + ruído F16 (SNR = 17,95)	
tips + ruído F16 (Amostrado em 8kHz)	3,084
Método Proposto	**3,528**
Método EMSR ($\mu = 0.98$)	3,570
Método EMSR ($\mu = 0.96$)	3,601
Método NMT-PSS	3,374
Método PSS	3,344
Limite Teórico	4,187
28. vega + ruído Pink (SNR = 3,03)	
vega + ruído Pink (Amostrado em 8kHz)	1,583
Método Proposto	**2,288**
Método EMSR ($\mu = 0.98$)	2,171
Método EMSR ($\mu = 0.96$)	2,155
Método NMT-PSS	2,084
Método PSS	1,893
Limite Teórico	3,633
29. vega + ruído Pink (SNR = 5,00)	
vega + ruído Pink (Amostrado em 8kHz)	1,729
Método Proposto	**2,471**
Método EMSR ($\mu = 0.98$)	2,339
Método EMSR ($\mu = 0.96$)	2,323
Método NMT-PSS	2,266
Método PSS	2,062
Limite Teórico	3,710
30. vega + ruído Pink (SNR = 6,97)	
vega + ruído Pink (Amostrado em 8kHz)	1,877
Método Proposto	**2,667**
Método EMSR ($\mu = 0.98$)	2,518
Método EMSR ($\mu = 0.96$)	2,499
Método NMT-PSS	2,447
Método PSS	2,230
Limite Teórico	3,787

SINAL DE VOZ CORROMPIDO	PONTUAÇÃO PESQ-MOS
31. vega + ruído Pink (SNR = 8,94)	
vega + ruído Pink (Amostrado em 8kHz)	2,027
Método Proposto	**2,867**
Método EMSR ($\mu = 0.98$)	2,709
Método EMSR ($\mu = 0.96$)	2,685
Método NMT-PSS	2,624
Método PSS	2,400
Limite Teórico	3,877
32. vega + ruído Pink (SNR = 14,85)	
vega + ruído Pink (Amostrado em 8kHz)	2,500
Método Proposto	**3,420**
Método EMSR ($\mu = 0.98$)	3,291
Método EMSR ($\mu = 0.96$)	3,268
Método NMT-PSS	3,167
Método PSS	2,973
Limite Teórico	4,062
33. vega + ruído Pink (SNR = 18,76)	
vega + ruído Pink (Amostrado em 8kHz)	2,867
Método Proposto	**3,702**
Método EMSR ($\mu = 0.98$)	3,642
Método EMSR ($\mu = 0.96$)	3,633
Método NMT-PSS	3,506
Método PSS	3,376
Limite Teórico	4,175

Referências Bibliográficas

[1] 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.

[2] BEROUTI, M., SCHWARTZ, R. and MAKHOUL, J., Enhancement of speech corrupted by acoustic noise, in *Proc.IEEE ICASSP*, Washington, DC,vol.4, pp. 208- 211, Apr. 1979

[3] LOCKWOOD, P. and BOUDY,J., Experiments with a nonlinear spectral subtractor (NSS),hidden Markov models and projection, for robust recognition in cars, *Speech Communication*, v.11, pp. 215-228, Jun. 1992.

[4] MOORER, J.A. and BERGER, M., Linear-phase bandsplitting: Theory and applications, *J.Audio Eng.Soc.*, vol.34,no.3, pp.143-152, 1996

[5] VASEGHI, S. and FRAYLING-CORK, R., Restoration of old gramophone recordings, *J.Audio Eng.Soc.*, vol.40,no.10, pp.791-801, 1992

[6] CAPPE O. and LAROCHE, J., Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings, *IEEE Trans.Speech Audio Processing*, vol.3, pp.84-93, 1994.

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

[8] EPHRAIM, Y. and MALAH, D., Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator , *IEEE Trans.Acoust.,Speech and Signal Processing*, vol.ASSP-32,no.6, pp.1109-1121, Dec.1984.

[9] JOHNSTON, J.D., Transform Coding of Audio Signals using Perceptual Noise Criteria, *IEEE J.Select.Areas Commun.*,vol.6, pp.314-323, Feb.1988.

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

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

[12] 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.

[13] DELLER Jr.,J., HANSEN, J. and PROAKIS J., *Discrete-Time Processing of Speech Signals*, NY: IEEE Press, 2000.

[14] SILVA,F.J.F. and ABRANCHES,L.K.S., 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.

[15] ABRANCHES,L.K.S. and SILVA,F.J.F., Speech Enhancement based on masking properties using a nonlinear short-time spectral attenuation, *International Workshop on Telecommunications (IWT)*, Brazil, Aug.2004.

[16] 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.

[17] TSOUKALAS, D., PARASKEVASa M. and MORJOPOULOS J., Speech Enhancement using psicho-acoustic criteria, *IEEE ICASSP*, Minneapolis, MN, pp.359-361, Apr. 1993

[18] USAGAWA, T.,IWATA, M. and EBATA M., Speech parameter extraction in noise environment using a masking model, *IEEE ICASSP*, Adelaide, Australia, vol.II, pp. 81-84, Apr. 1994.

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

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

[21] SINHA, D. and TEWFIK, A.H., Low bit rate transparent audio compression using adapted wavelets, *IEEE Trans.Signal Processing*, vol.41, pp.3463-3479, Dec.1993

[22] ZWICKER, E. and FASTL, H., *Psychoacoustics*, Springer Verlag, 2nd ed., 1999.

[23] SCHARF, B.., *Foundations of Modern Auditory Theory*, cap.5. New York Academic, 1970.

[24] TERHARDT E., STOLL G. e SEEWANN M., Algorithm for extraction of pitch and pitch salience from complex tonal signals, *J. Acoust. Soc. Am.*, vol. 71, pp. 679-688, Mar. 1982.

[25] CAPPE, O., Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor, *IEEE Transactions on Speech and Audio Processing*, vol.2, n.2, pp. 345-349, Apr.1994.

[26] RABINER L.R. and SCHAFER, *Digital Processing of Speech Signals*, Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1978.

[27] EPHRAIM, Y. and MALAH, D., Speech Enhancement Using a Minimum Mean-Square Error log-Spectral Amplitude Estimator , *IEEE Trans.Acoust.,Speech and Signal Processing*, vol.ASSP-33,n.2,pp.443-445, Apr.1985.

[28] DAVENPORT, W.B. and ROOT, W.L., *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, ch.7,appendix1, 1960.

[29] ZELINSKI, R. and NOLL, P., Adaptive Tranform coding of speech signals, *IEEE Trans.Acoust.,Speech,Signal Processing*, vol.ASSP-25,pp.306, Aug.1977.

[30] TRIBOLET J.M. and Crochiere R.E., Frequency domain coding of speech, *IEEE Trans.Acoust.,Speech, Signal Processing*, vol.ASSP-27, pp.522, Oct.1979.

[31] PORTER J.E. and BOLL S.F., Optimal estimators for spectral restoration of noisy speech, *Proc.IEEE Int.Conf.Acoust.,Speech,Signal Processing*, pp.18A.2.1 - 18A2.4, Mar.1984.

[32] GRADSHTEYN, I.S. and RYZHIK, I.M., *Table of Integrals, Series and Products* , New York: Academic, 1980.

[33] MCAULAY, R.J. and MALPASS, M.L., Speech enhancement using a soft-decision noise suppression filter , *IEEE Trans. Acoust., Speech, Signal Processing*, vol ASSP-28, pp. 137-145, Apr.1980.

[34] LIM, J.S. and OPPENHEIM, A.V., Enhancement and bandwidth compression of noisy speech, *Proc.IEEE ICASSP*, vol.67,pp. 1586-1604, Dec. 1979.

[35] ATAL, B.S. and HANAUER, S.L., Speech analysis and syntesis by linear prediction of the speech wave, *The Journal of the Acoustical Society of America*, vol.50, pp.637 - 655, 1971.

[36] LATHI,B.P., *Modern Digital and Analog Communication Systems*, Oxford University Press,pp.461-463, 1998.

[37] WAN E., NELSON A. and PETERSON R. Oregon Graduate Institute Assessment Resource. Disponível em: WWW. URL: <http://cslu.ece.ogi.edu/nsel/data/SpEAR_database.html>. Acessado em: 02 de Junho de 2004.

[38] 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.

[39] Download PESQ. Disponível em: http://www.itu.int/rec/recommendation.asp ?type=folders&lang=e&parent=T-REC-P.862. Acessado em: 20 de Agosto de 2004.

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

[41] LEON-GARCIA, A., *Probability and Random Processes for Electrical Engineering second edition*,pp.215-248, Addison-Wesley, 1994.

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

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

[44] RICE S.O., *Statistical Properties of a Sinewave Plus Random Noise*, Bell System Tech J., pp. 109-157, Jan.1948.