UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE ENGENHARIA DE ILHA SOLTEIRA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Estudo e implementação de uma técnica de redução de ruído em sinais de voz baseada na subtração espectral e em critérios psicoacústicos

Allan Zukeran Kanda

Orientador: Prof. Dr. Jozué Vieira Filho

Ilha Solteira – SP, Fevereiro de 2010

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Ilha Solteira

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"Estudo e implementação de uma técnica de redução de ruído em sinais de voz baseada na subtração espectral e em critérios psicoacústicos"

ALLAN ZUKERAN KANDA

Orientador: Prof. Dr. Jozué Vieira Filho

Dissertação apresentada à Faculdade de Engenharia - UNESP – Campus de Ilha Solteira, para obtenção do título de Mestre em Engenharia Elétrica.
Área de Conhecimento: Telecomunicações.

Ilha Solteira – SP
Fevereiro/2010

FICHA CATALOGRÁFICA

Elaborada pela Seção Técnica de Aquisição e Tratamento da Informação
Serviço Técnico de Biblioteca e Documentação da UNESP - Ilha Solteira.

Kanda, Allan Zukeran.
K16e Estudo e implementação de uma técnica de redução de ruído em sinais de voz baseada na subtração espectral e em critérios psicoacústicos / Allan Zukeran Kanda. -- Ilha Solteira : [s.n.], 2010
84 f. : il.

Dissertação (mestrado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Telecomunicações, 2010.

Orientador: Josué Vieira Filho

l. ANIQUE. 2. Subtração espectral. 3. Minimização do erro quadrático médio. 4. Ruído – Redução. 5. Psicoacústicas – Técnicas.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

CAMPUS DE ILHA SOLTEIRA

FACULDADE DE ENGENHARIA DE ILHA SOLTEIRA

CERTIFICADO DE APROVAÇÃO

TÍTULO: Estudo e implementação de uma técnica de redução de ruído em sinais de voz baseada na subtração espectral e em critérios psicoacústicos

AUTOR: ALLAN ZUKERAN KANDA
ORIENTADOR: Prof. Dr. JOZUE VIEIRA FILHO

Aprovado como parte das exigências para obtenção do Título de MESTRE em ENGENHARIA ELÉTRICA , Área: AUTOMAÇÃO, pela Comissão Examinadora:

Prof. Dr. JOZUE VIEIRA FILHO
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira

Profa. Dra. SUELY CUNHA AMARO MANTOVANI
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira

Prof. Dr. MARCO APARECIDO QUEIROZ DUARTE
Departamento de Matemática / Universidade Estadual de Mato Grosso do Sul

Data da realização: 25 de fevereiro de 2010.

Dedico este trabalho a Deus e toda minha família e amigos
que fizeram e fazem parte de minha vida

Agradecimentos

À Deus.

Aos meus pais, Claudio Sueki Kanda e Akemy Zukeran Kanda.

À Minhas irmãs Claudia Zukeran Kanda e Cristina Yayoi Zukeran Kanda.

Aos amigos Carlos, Evandro, Leonardo, Leandro e em especial à Adriana.

Ao meu orientador professor Jozué Vieira Filho.

À Faculdade UNESP de Ilha Solteira.

RESUMO

A proposta deste trabalho é aprimorar a performance da técnica de redução de ruído, *subtração espectral* baseado na relação *SNR a Priori*, através da implementação de dois novos parâmetros *Potência de Articulação* e *Não-Articulação* obtidas a partir de algumas técnicas psicoacústicas. Faz-se um estudo da anatomia do sistema de audição humana e algumas limitações físicas, com o objetivo de entender o princípio básico da técnica *ANIQUE*, que é um sistema de avaliação objetiva de voz e têm como princípio o modelamento da percepção humana da voz. Através do modelo *ANIQUE* são extraídas as principais técnicas psicoacústicas para obtenção dos novos parâmetros, *Potência de Articulação* e *Não-Articulação*. Procurou-se apresentar de maneira resumida o processo de equacionamento das técnicas de redução de ruído em sinais de voz e das técnicas psicoacústicas. Posteriormente são descritos todos os processos das técnicas utilizadas que foram simuladas utilizando a linguagem de programação do *MatLab*®, seguido das avaliações objetivas dos sinais processados pelo software *PESQ*, que é um programa de avaliação objetiva de voz. Os resultados mostram que a implementação das técnicas psicoacústicas foram eficazes para melhorar a performance da técnica *subtração espectral* baseada na relação *SNR a Priori*.

Palavras-chaves: ANIQUE. Subtração espectral. Minimização do erro quadrático médio. Redução de ruído. Técnicas psicoacústicas.

ABSTRACT

The purpose of this work is to enhance the performance of noise reduction techniques based on spectral subtraction, which take in account the a priori signal-to-noise (*SNR a Priori)* to be estimated considering psychoacoustic criteria. in order to understand the basic principle of the *ANIQUE,* which is a psychoacoustic based technique used to evaluate the quality of speech signals, it was necessary to develop a study of the anatomy of the human hearing and some physical limitations, From the *ANIQUE* are extracted new parameters namely *Articulation* and *Non-Articulation Powers*, used to estimate the *SNR_prio*. As a result, it was obtained a new spectral based technique which was implemented in the *MatLab*® environment and evaluated using the objective quality measure for speech signal simulations namely PESQ. The results show that the implementation of psychoacoustic techniques were effective in enhance the performance of the spectral subtraction technique based on *SNR a Priori*.

Keywords: ANIQUE. Spectral subtraction. Noise reduction. Articulation power. Non-articulation power. Psychoacoustic techniques.

LISTA DE FIGURAS

Figura 1 – Anatomia simplificada do ouvido 16
Figura 2 – Ouvido médio. 17
Figura 3 – Estrutura da cóclea. 19
Figura 4 – Anatomia simplificada de um corte transversal da cóclea. 20
Figura 5 – Propriedades da membrana basilar 21
Figura 6 – Frequência característica ao longo da cóclea. 22
Figura 7 – Anatomia do órgão de Corti 23
Figura 8 – Limiar absoluto de audibilidade 26
Figura 9 – Ilustração dos experimentos para identificação das bandas críticas 28
Figura 10 – Principais tipos de mascaramento. 30
Figura 11 – Ilustração do efeito do mascaramento simultâneo 32
Figura 12 – Exemplo de ruído de banda estreita mascarando tom 33
Figura 13 – Exemplo de sinal tonal mascarando ruído de banda estreita 34
Figura 14 – Diagrama de blocos de sistemas intrusivos e não intrusivos 37
Figura 15 – Diagrama de blocos da técnica *ANIQUE* 39
Figura 16 – Diagrama de blocos do bloco *"Estimação de Qualidade"*. 40
Figura 17 – Esquema básico da subtração espectral 45
Figura 18 – Diagrama de blocos do sistema da técnica de redução de ruído com *ANIQUE* 53
Figura 19 – Janela Hamming com largura de banda de 512 amostras. 54
Figura 20 – Diagrama do bloco *"Estimação de Qualidade"*. 56
Figura 21 – Resposta impulsiva do filtro gammatone 57
Figura 22 – Resposta em frequência do filtro gammatone 57
Figura 23 – Resposta em freqüência do banco de filtros de banda crítica 58
Figura 24 – Exemplo da envoltória do sinal e seu espectro de modulação 60
Figura 25 – Resposta em frequência do banco de filtros de modulação 61
Figura 26 – Pequeno trecho das respostas em frequência dos filtros separadamente. . . 68
Figura 27 – Forma de onda do sinal 1 sem adição de ruído 71
Figura 28 – Forma de onda do sinal 1 com a relação *SNR*=0dB 71
Figura 29 – Forma de onda do sinal 1 com a relação *SNR*=5dB 72

Figura 30 – Forma de onda do sinal 1 com a relação *SNR*=10dB. 72
Figura 31 – Forma de onda do sinal 1 com a relação *SNR*=15dB 72
Figura 32 – Forma de onda do sinal 2 com a relação *SNR*=10dB 73
Figura 33 – Forma de onda do sinal 2 com a relação *SNR*=10dB processado pela filtragem. 73
Figura 34 – Avaliações objetivas dos sinais processados com SNR=0dB 74
Figura 35 – Avaliações objetivas dos sinais processados com SNR=5dB 75
Figura 36 – Avaliações objetivas dos sinais processados com SNR=10dB 75
Figura 37 – Avaliações objetivas dos sinais processados com SNR=15dB 76

LISTA DE QUADROS

Quadro 1 – Nível de pressão sonora para exemplos do cotidiano. 25
Quadro 2 – Frequências características das bandas crítica. 29
Quadro 3 – Frequência características e largura de banda dos filtros de modulação. . 61
Quadro 4 – Avaliações dos sinais processados com *SNR*=0dB. 74
Quadro 5 – Avaliações dos sinais processados com *SNR*=5dB. 74
Quadro 6 – Avaliações dos sinais processados com *SNR*=10dB. 75
Quadro 7 – Avaliações dos sinais processados com *SNR*=15dB. 76

LISTA DE SÍMBOLOS E ABREVIATURAS

ACR	Teste de Qualidade Absoluta
AM	Modulação em Amplitude
CDMA	Sistema de Múltiplo Acesso por Divisão de Código
DVS	Detecção de Voz-Silêncio
GSM	Sistema de Comunicação Móvel Global
ISTFT	Transformada Inversa de Fourier de Curto Prazo
ITU-T	Seção de Padronização da área de Telecomunicações do ITU - União Internacional de Telecomunicações
MMSE	Minimização do Erro Quadrático Médio
MMSE+SNR Prio	Minimização do Erro Quadrático Médio baseado na relação SNR a Priori
MSV	Melhoramento de Sinais de Voz
MOS	Medida de Qualidade de Voz
PESQ	Avaliação Perceptual de Qualidade de Voz
PSQM	Sistema de Medida de Análise Perceptual
QoS	Qualidade de Serviço
SE	Subtração Espectral
SE+SNR Prio	Subtração Espectral baseado na relação SNR a Priori
SE+SNR Prio+P.Art	Subtração Espectral baseado na relação SNR a Priori com os parâmetros de Articulação
SNR	Relação sinal/ruído
SNR Prio	Relação Sinal/Ruído a Priori
SNR Post	Relação Sinal/Ruído a Poteriori
SPL	Nível de Pressão Sonora
STFT	Transformada de Fourier de Curto Prazo
UMTS	Sistema de Telecomunicação Móvel Universal
VoIP	Sistema de Comunicação Via Protocolo de Internet
$\hat{\bullet}$	Indica Estimação
$\Im\{\bullet\}$	Transformada de Fourier

$h_k(t)$	Resposta Impulsiva dos Filtros de Banda Crítica do Canal Cóclea *k*
$H_k(s)$	Resposta em Frequência dos Filtros de Banda Crítica do Canal Cóclea *k*
ERB_k	Largura de Banda Retangular Equivalente do Canal Cóclea *k*
$\Gamma(N)$	Função Gamma
$\tilde{s_k}(n)$	Transformada de Hilbert de $s_k(n)$
$\gamma_k(n)$	Envoltória do Sinal Filtrado pelo Canal Cóclea *k*
$\Psi_{k,A}(m)$	Espectro da Envoltória de Articulação do Canal Cóclea *k* e quadro *m*
$\Psi_{k,N}(m)$	Espectro da Envoltória de Não-Articulação do Canal Cóclea *k* e quadro *m*
$\Lambda_A(m,n)$	Potência de Articulação do quadro *m*
$\Lambda_N(m,n)$	Potência de Não-Articulação do quadro *m*
$\sigma_r^{\,2}(\bullet)$	Estimação da Potência do Ruído
$\lvert Y(\bullet)\rvert^2$	Potência do sinal Ruidoso
$\lvert V(\bullet)\rvert^2$	Potência do sinal de Voz
$H_S(\omega)$	Resposta em frequência do filtro da subtração espectral com os parâmetros SNR_Prio e as Potências de Articulação
$H_{So}(\omega)$	Resposta em frequência do filtro da subtração espectral com os parâmetros SNR_Prio
$H_{CS}(\omega)$	Resposta em frequência do filtro em cascata

SUMÁRIO

1 Introdução 13
1.1 Motivação 13
1.2 Objetivo 14
1.3 Organização do Trabalho 14

2 Sistema Auditivo Humano 15
2.1 Sistema Auditivo Humano 15
2.3 Ouvido Externo 16
2.4 Ouvido Médio 17
2.5 Ouvido Interno 18
2.5.1 Cóclea 19
2.5.2 Membrana Basilar 20
2.5.3 Órgão de Corti 22

3 Fenômenos Auditivos. 24
3.1 Percepção de Volume de Som 24
3.2 Resposta em Frequência do Sistema Auditivo 25
3.3 Limiar Absoluto de Audibilidade em Silêncio 26
3.4 Bandas Críticas 27
3.5 Mascaramento 29
3.5.1 Mascaramento Não Simultâneo 31
3.5.2 Mascaramento Simultâneo 31

4 Sistema ANIQUE de Avaliação Objetiva de Voz. 35
4.1 Sistema de Avaliação Objetiva de Voz 35
4.2 Modelos Intrusivos e Não-Intrusivos 37
4.3 ANIQUE 38
4.4 Banco de Filtros Cocleares e Envoltória Temporal 39
4.5 Banco de Filtros de Modulação 41
4.6 Análise de Articulação 41

5 Técnicas de Redução de Ruído Baseada na Relação SNR a Priori 42
5.1 Subtração Espectral 42
5.2 Minimização do Erro Quadrático Médio 46
5.3 Relação SNR na Redução de Ruído em Sinais de Voz 48
5.4 Técnicas de MSV baseadas na Relação SNR Posteriori 49
5.5 Técnicas de MSV baseadas na Relação SNR Priori 50
5.6 Técnicas Clássicas de Redução de Ruído Usando a SNR Prio 50

6 Obtenção de um Filtrobaseado na SE e em Critérios Psicoacústicos 52
6.1 Incorporação de Técnicas Psicoacústicas na Redução de Ruído 52
6.2 Potência de Articulação e Não-Articulação 55
6.3 Banco de Filtro de Banda Crítica e Envoltória Temporal 55
6.4 Banco de Filtro de Modulação 60
6.5 Análise de Articulação 62
6.6 Estimação da Potência do Ruído 63
6.7 Estimação da SNR Posteriori 64
6.8 Estimação da SNR Priori 65
6.9 Procedimento de Filtragem 66

7 Simulações e Resultados. 69
7.1 Sinais Utilizados nas Simulações 70
7.2 Sinais Processados 73

8 Conclusões 77

Referências. 79

Apêndice A – Definições Complementares 83

CAPÍTULO 1

Introdução

1.1 Motivação

A melhoria da qualidade dos sinais de fala presentes nos sistemas de telecomunicações tem sido foco de intensos estudos nas últimas décadas (PELLOM; HASSEN, 1998). Em praticamente todas as aplicações de transmissão de voz a qualidade da comunicação pode ser comprometida pela presença de elementos que degradam o sinal, como o ruído ambiente, reverberação, perdas devidas à codificação em enlaces digitais e concorrência de outras conversações ou de outras fontes de sinal. Tais elementos podem afetar o sinal de diversas formas, reduzindo sua inteligibilidade, aumentando o cansaço do ouvinte, tornando a conversação pouco natural, ou ainda, afetando a eficiência de outros sistemas que se utilizarão desses sinais posteriormente, como reconhecedores ou codificadores de voz. Os métodos de melhoria da qualidade dos sinais de fala buscam, portanto, identificar e extrair os elementos que degradam a qualidade do sinal, realçando a informação de fala, possibilitando assim uma melhor comunicação entre as partes envolvidas.

Na maioria das aplicações práticas, um algoritmo eficiente de redução de ruído deve ser capaz de melhorar o sinal ruidoso nos aspectos auditivo (inteligibilidade) e físico (recuperação de onda original), além de garantir uma baixa carga de processamento computacional, necessária para uma implementação em tempo real. Esta redução de ruído é denominada de *Melhoramento de Sinais de Voz* (*MSV*).

1.2 Objetivo

O objetivo deste trabalho é aplicar os princípios de piscoacústicas em uma técnica de redução de ruído baseada na subtração espectral e na na relação *SNR a Priori*. A aplicação da *SNR_prio* em técnicas baseadas na subtração espectral e em outras técnicas de princípio equivalente, denominadas de técnicas clássicas de *MSV*, reduz o ruído residual do sinal de voz processado, melhorando a inteligibilidade e mantendo um bom nível de redução de ruído. Através de avaliações objetivas de qualidade de voz, mostra-se que a incorporação de técnicas psicoacústicas possibilita melhorar técnicas de redução de ruído baseadas na subtração espectral e na *SNR_prio*. As características psicoacústicas exploradas no trabalho são baseadas numa técnica de avaliação de qualidade de voz denominada ANIQUE (An Auditory Model for Single-Ended Speech Quality Estimation) O software utilizado para a implementação dos algoritmos de redução de ruído foi o *MatLab*® e para avaliação dos sinais processados foi utilizado a *PESQ*, que é uma medida objetiva de qualidade de voz.

1.3 Organização do Trabalho

Para melhor entendimento das técnicas psicoacústicas que foram extraídas da *ANIQUE*, no capítulo 2 apresenta-se uma análise da anatomia de todo o sistema de audição humana. No capítulo 3 apresentam-se algumas limitações do sistema de audição humana, que são importantes para compreensão dos motivos das utilizações das técnicas psicoacústicas.

Medidas de avaliação objetivas, como a *ANIQUE* e outras, são descritas de maneira resumida no capítulo 4. Posteriormente, no capítulo 5 são descritos o modelamento matemático das técnicas de redução de ruído, a *subtração espectral* e *minimização do erro quadrático médio* e os parâmetros de redução de ruído *SNR_Post* e a *SNR_Prio*.

No capítulo 6 é apresentado todo o sistema de redução de ruído baseado nos novos parâmetros extraídos da técnica *ANIQUE* e na *subtração espectral* baseada na *SNR a Priori*. No capítulo 7 são apresentados os resultados para várias simulações no capítulo 8, finalizando, apresenta-se uma conclusão final do trabalho.

CAPÍTULO 2

Sistema Auditivo Humano

Pode-se definir a psicoacústica como o estudo fisiológico da audição, sendo que o objetivo das pesquisas em psicoacústica é entender o funcionamento do processo auditivo, ou seja, como os sons chegam aos ouvidos e são processados, pelos mesmos e pelo cérebro, de modo a dar ao ouvinte informações úteis sobre o mundo à sua volta.

Para entender as técnicas psicoacústicas utilizadas na *ANIQUE*, de onde são extraídos os parâmetros para melhoramento do desempenho da técnica de redução de ruído, é necessário o conhecimento básico do funcionamento de audição humana. Portanto, neste capítulo é apresentada uma análise da anatomia de todo o sistema de audição humana.

2.1 Sistema Auditivo Humano

A maioria dos sistemas de percepção do ser humano, entre eles o sistema auditivo, não é precisa e possuem limitações físicas. Para entender um pouco mais de suas limitações, uma análise mais detalhada de sua anatomia proporcionará uma base para o estudo de algumas técnicas psicoacústicas.

Ondas de som são propagações de vibrações de um meio físico geradas através da vibração de um corpo. Normalmente, o meio físico é o ar e a onda sonora corresponde à variação da pressão atmosférica de suas partículas.

A Figura 1 apresenta a anatomia do sistema auditivo, que é dividido em três partes conhecidas como: ouvido externo, ouvido médio e ouvido interno.

O sistema auditivo humano funciona com base em operações acústicas e mecânicas do

ouvido externo, no processamento feito no ouvido médio para conversão dos movimentos mecânicos em impulsos elétricos e na transmissão das informações neurais do ouvido interno para o cérebro.

2.2 Ouvido Externo

O ouvido externo compreende desde a orelha até o canal externo, terminando no tímpano. A orelha tem a função de proteger o canal externo e de acentuar certas frequências, ajudando a localizar as fontes sonoras e direcioná-las para o canal externo do ouvido. Sua forma ajuda o ouvido a perceber se o som está à frente ou atrás do ouvinte com boa acuidade, e também acima ou abaixo (com menor precisão).

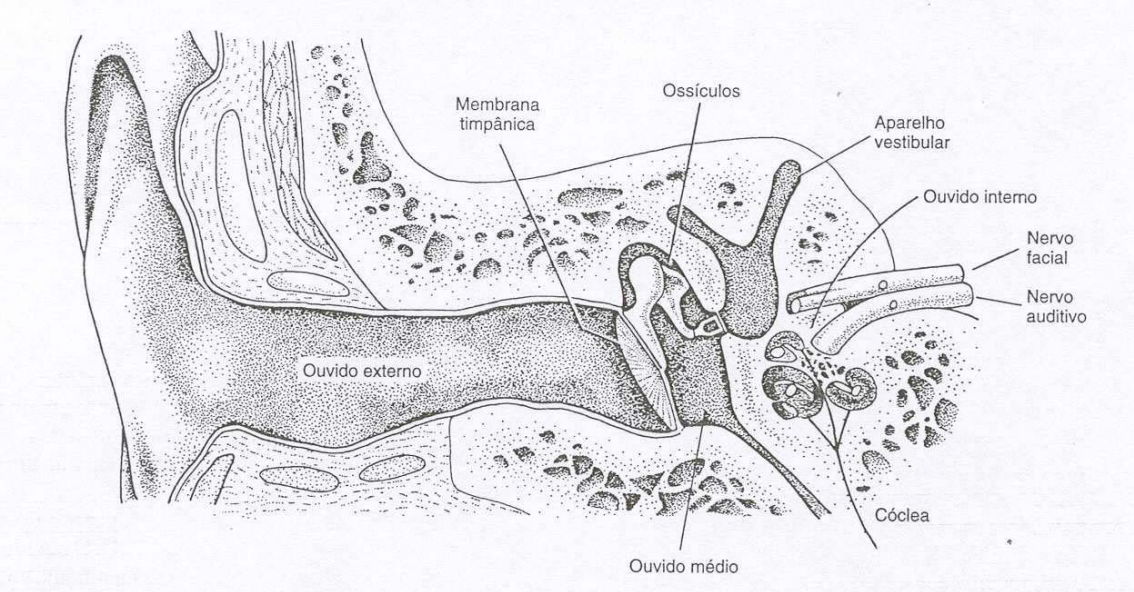


Figura 1 – Anatomia simplificada do ouvido (BERNE; LEVY, 2000).

O ouvido externo condiciona o sinal acústico que chega ao conduto auditivo, podendo aumentar a pressão no tímpano em até 15 dB para as frequências de 3 kHz a 5 kHz (GIGUERE; WOODLAND, 1982), o que melhora a sensibilidade para a audição dos sinais de fala. Uma vez que a variação de pressão sonora chega ao tímpano, ela faz com que este vibre, realizando a conversão da energia sonora em energia mecânica.

2.3 Ouvido Médio

As principais estruturas do ouvido médio são a membrana timpânica, a cadeia ossicular, com os respectivos ligamentos e músculos, e a cavidade preenchida com ar na qual estão localizados os ossículos. Sua principal função é melhorar a transmissão sonora entre o ouvido externo e o ouvido interno. A Figura 2 ilustra o ouvido médio.

Os ossículos têm a função de transformar a impedância acústica do sinal que entra no ouvido. Isso é necessário porque o meio externo (o ar) e o ouvido interno possuem diferentes resistências à propagação da onda. A resistência do fluído do ouvido interno é mais alta que aquela do ar, fazendo com que os ossículos atuem como conversores de impedância. Esta transformação de impedância ocorre devido ao efeito de alavanca que existe entre o martelo e a bigorna e a diferença entre as áreas do tímpano e da parte do estribo que está em contato com a janela oval, que concentra a energia imposta no sistema. Esse efeito poderá resultar, para o ouvido interno, em um aumento de até 30 dB entre os níveis de pressão sonora no tímpano e na janela oval (Backus, 1969).

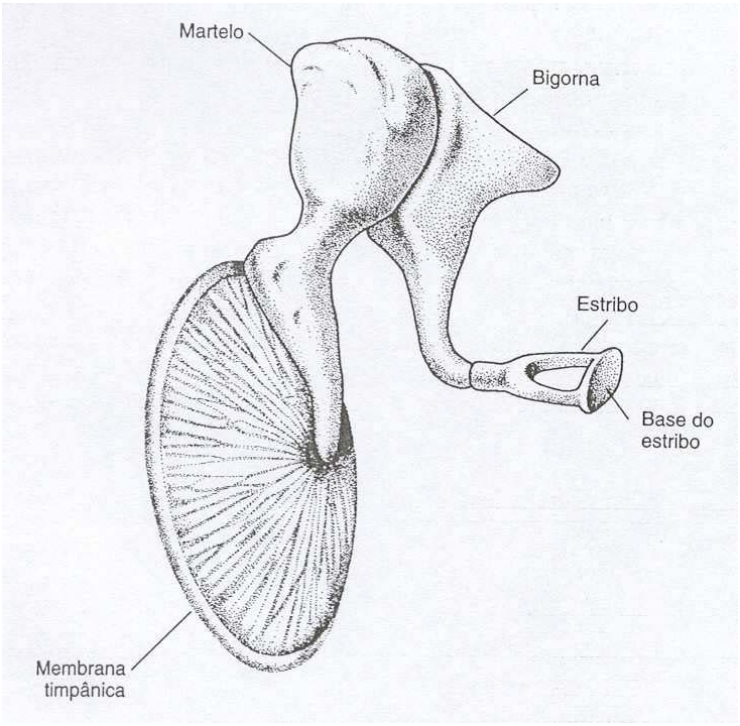


Figura 2 – Ouvido médio (BERNE; LEVY, 2000).

Outra função do ouvido médio é a proteção do ouvido interno contra danos causados por vibrações provenientes de ondas sonoras. Essa proteção é realizada por dois grupos de músculos que entram em ação e se contraem automaticamente em resposta aos sons com níveis de pressão sonora acima de 75 dB SPL (Nível de Pressão Sonora), enrijecendo o sistema e fazendo com que a transmissão de energia não seja muito eficiente (BACKUS, 1969).

O primeiro grupo muscular tem o objetivo de atenuar o movimento do martelo ao se contrair e, consequentemente, atenuar a vibração que passa pelo ouvido médio. O segundo grupo tem o objetivo de manter o osso estribo longe da janela oval, visando enfraquecer a vibração que passará para o ouvido interno. Este efeito é conhecido como reflexo acústico e aproximadamente de 12 a 14 dB de atenuação são conseguidos nesse processo, mas esses valores são para sons abaixo de 1 kHz somente. Essa reação de contração não é instantânea e leva de 60 a 120 ms em média para entrar em funcionamento, de modo que o ouvido não é protegido para sons muito impulsivos (como por exemplo, o som de uma arma de fogo).

O processo de transformação do sinal acústico é chamado de função de transferência do ouvido médio e é equivalente a uma filtragem passa baixas com corte em 5 kHz, com uma sobre elevação na faixa entre 2 kHz e 5 kHz e um pico em torno de 3,5 kHz. Como essa filtragem não altera o espectro de forma significativa, ela é em geral, desconsiderada para sinais com faixa até 5 kHz.

O ouvido médio ainda tem as funções de realizar o casamento de impedância acústica, filtrar sons de baixa frequência em ambientes barulhentos e diminuir a sensibilidade para a própria fala.

2.4 Ouvido Interno

O ouvido interno é formado pela cóclea, labirinto e canal interno. Da cóclea sai o nervo auditivo via canal interno, que é ósseo, por onde também passam os nervos faciais (responsável pela movimentação de músculos da face) e o aparelho vestibular (responsável pelo equilíbrio) (GLASBERG; MOORE, 1990).

2.4.1 Cóclea

A cóclea é uma estrutura rígida na forma de caracol preenchida por fluídos incompressíveis, sendo um dos órgãos principais da audição. Ela é responsável pela conversão das vibrações mecânicas, que chegam do ouvido médio, em impulsos elétricos. A cóclea é dividida por duas membranas, ao longo de seu comprimento de aproximadamente 30mm, que são: a membrana vestibular e membrana basilar.

A cóclea ainda contém muitas outras partes, como o órgão de Corti, de fundamental importância para a audição. A Figura 3 ilustra a estrutura da cóclea e a Figura 4 ilustra seu corte transversal.

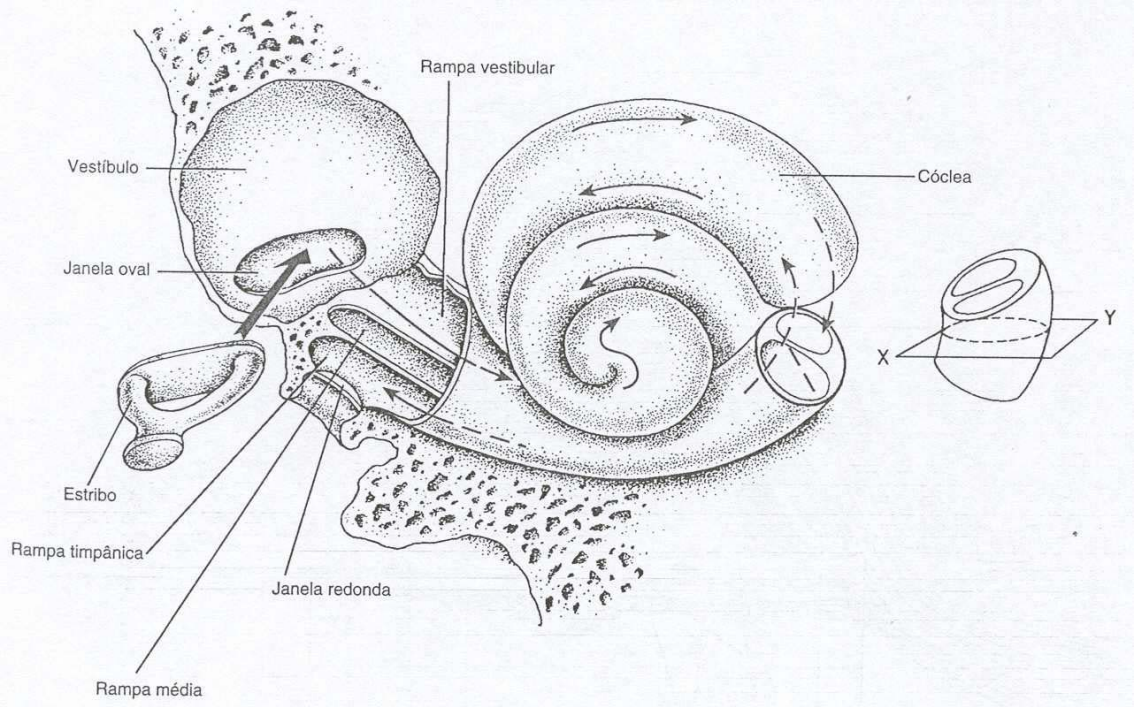


Figura 3 – Estrutura da cóclea (BERNE, 2000).

Pelo osso estribo são passadas as vibrações do ouvido médio para o ouvido interno, o qual se movimenta para dentro e para fora do ouvido interno através da janela oval, e deslocam-se através do fluído. O diâmetro do tímpano é de 15 a 30 vezes maior do que o da janela oval, amplificando a pressão transmitida para o ouvido interno, e essa variação de pressão na cóclea faz com que a membrana basilar movimente-se transversalmente. Este movimento é detectado pelo órgão de Corti, que realiza a conversão de energia mecânica em impulsos elétricos.

As células ciliadas do órgão de Corti são sensíveis à variações de cerca de 60 dB, enquanto o intervalo de sensibilidade da audição é da ordem de 100 dB (ZWICKER; FASTL, 1999).

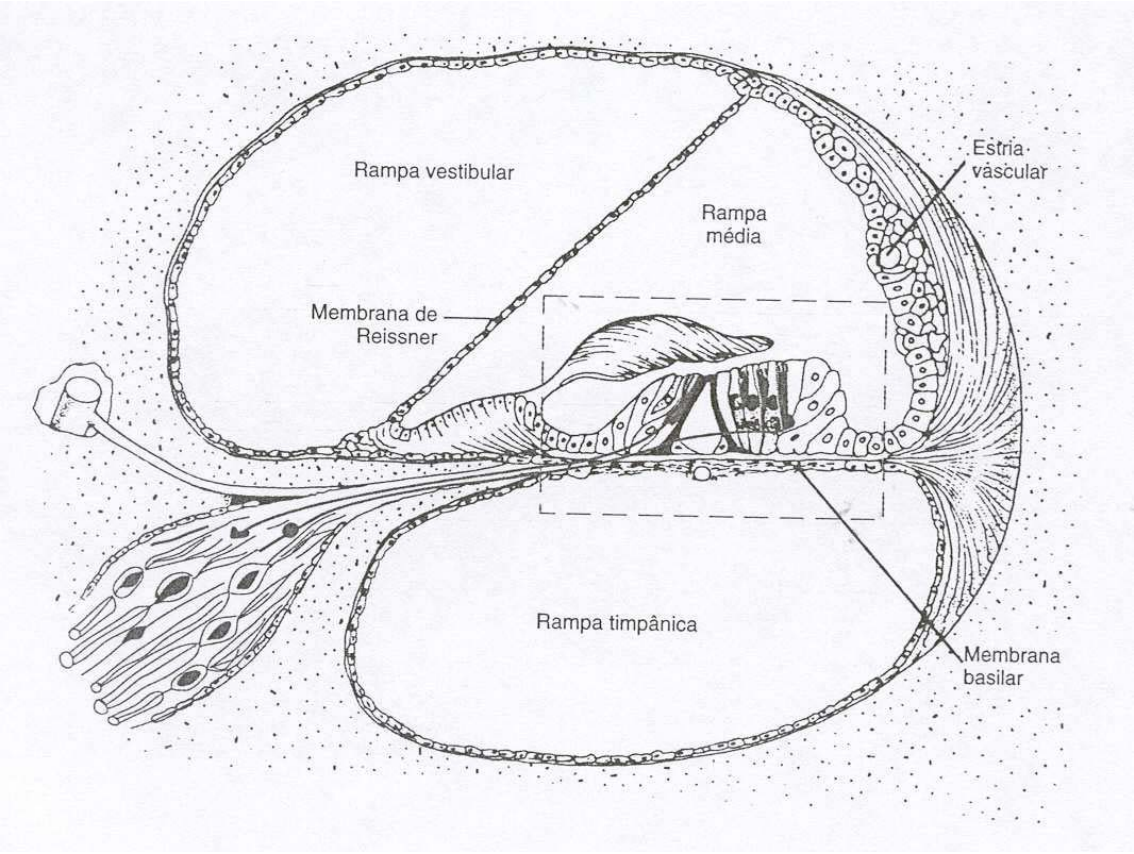


Figura 4 – Anatomia simplificada de um corte transversal da cóclea (BERNE; LEVY, 2000).

2.4.2 Membrana Basilar

A membrana basilar é responsável pelo processo de percepção do som, fazendo uma análise das frequências. Essa membrana se estende por todo o comprimento da cóclea, sendo mais fina e rígida perto da base (extremidade mais próxima do ouvido médio), e mais grossa e menos rígida na outra extremidade, respondendo às variações de pressão que ocorrem no fluído dentro da cóclea.

O estímulo a uma resposta que possui apenas um componente em frequência (tom puro) ocorre na forma de uma onda que se propaga ao longo de toda a membrana, com maior amplitude em uma região específica que dependente da frequência especifica do estímulo. Para as altas frequências, a amplitude máxima da onda ocorre próximo à base da membrana e, para as baixas frequências, a amplitude máxima ocorre próxima à outra extremidade. Portanto, a membrana basilar comporta-se como um analisador de espectro, na qual ocorre uma associação posição-frequência.

Na Figura 5(a), apresenta-se a amplitude da vibração da membrana basilar em função da distância de sua base para um sinal com duas componentes de frequência – uma alta e outra baixa. Nota-se que a amplitude de vibração não é simétrica em relação ao seu máximo. Na Figura 5(b), são mostradas as componentes de sua estrutura e na Figura 5(c) apresenta-se a relação entre a frequência do sinal e a posição da oscilação ao longo da membrana. Por fim, na Figura 5(d), pode-se observar a relação a rigidez da membrana em função da distância da base.

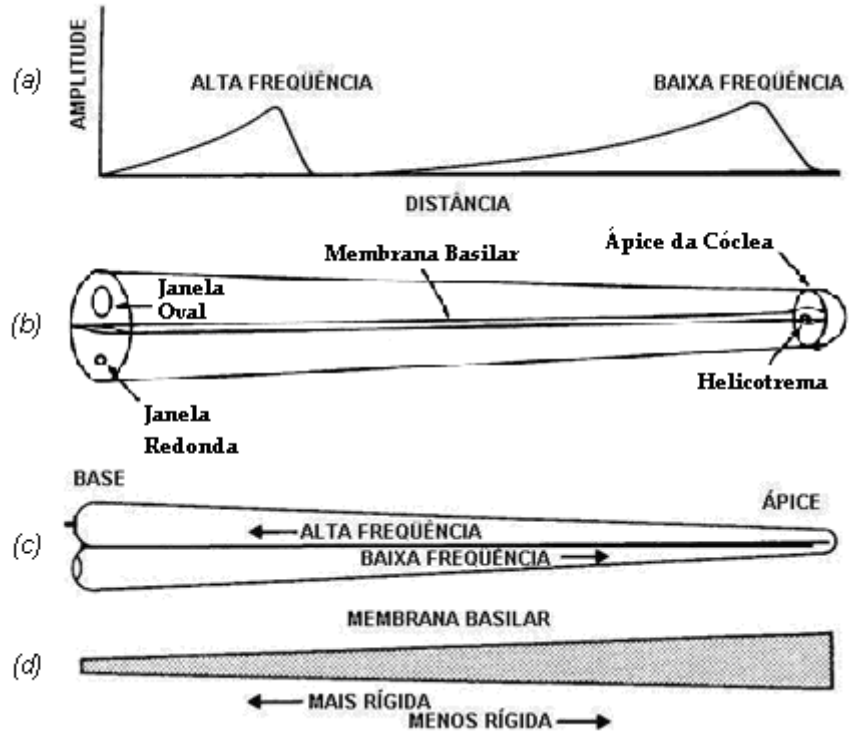


Figura 5 – Propriedades da membrana basilar, (a) amplitude da vibração da membrana basilar em função da distância de sua base, (b) componentes de sua estrutura , (c) relação entre a frequência do sinal e a posição da oscilação ao longo da membrana, (d) relação entre a rigidez da membrana em função da distância da base (LEITE, 2003).

Em uma determinada frequência, cada região da membrana basilar possui seu pico de oscilação, que é denominada frequência característica. Na Figura 6, observa-se a distribuição das frequências características ao longo da cóclea.

Na membrana basilar ainda existe duas estruturas: as fibras basilares e o órgão de Corti.

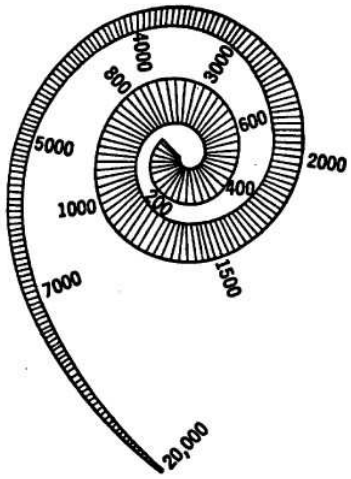


Figura 6 – Frequência característica ao longo da cóclea (LEITE, 2003).

2.4.3 Órgão de Corti

O processo de transformação dos movimentos da membrana basilar em impulsos nervosos para envio do cérebro é feito pelas células do órgão de Corti. Ele está localizado sobre a membrana basilar e contém cerca de 20.000 fibras basiliares, que ficam em contato com o nervo auditivo.

As fibras basilares são pequenas estruturas delgadas com comprimentos que variam ao longo da membrana, sendo mais curtas junto à janela oval e mais longas no ápice da cóclea. Com o movimento da membrana basilar, as células ciliadas transformam o movimento das fibras basilares em impulsos nervosos, que são transmitidos pelo nervo coclear para a região específica do córtex cerebral. A Figura 7 ilustra a anatomia do órgão de Corti.

Como cada ponto da membrana basilar possui uma frequência característica específica, a curva de resposta em frequência das vibrações presentes na janela oval é equivalente à de um filtro passa-faixa com fator de qualidade aproximadamente constante, resultando numa melhor resolução nas baixas frequências.

Assim, as fibras basilares localizadas na região de altas frequências características respondem em uma maior faixa de frequências do que as fibras na região de baixas frequências características.

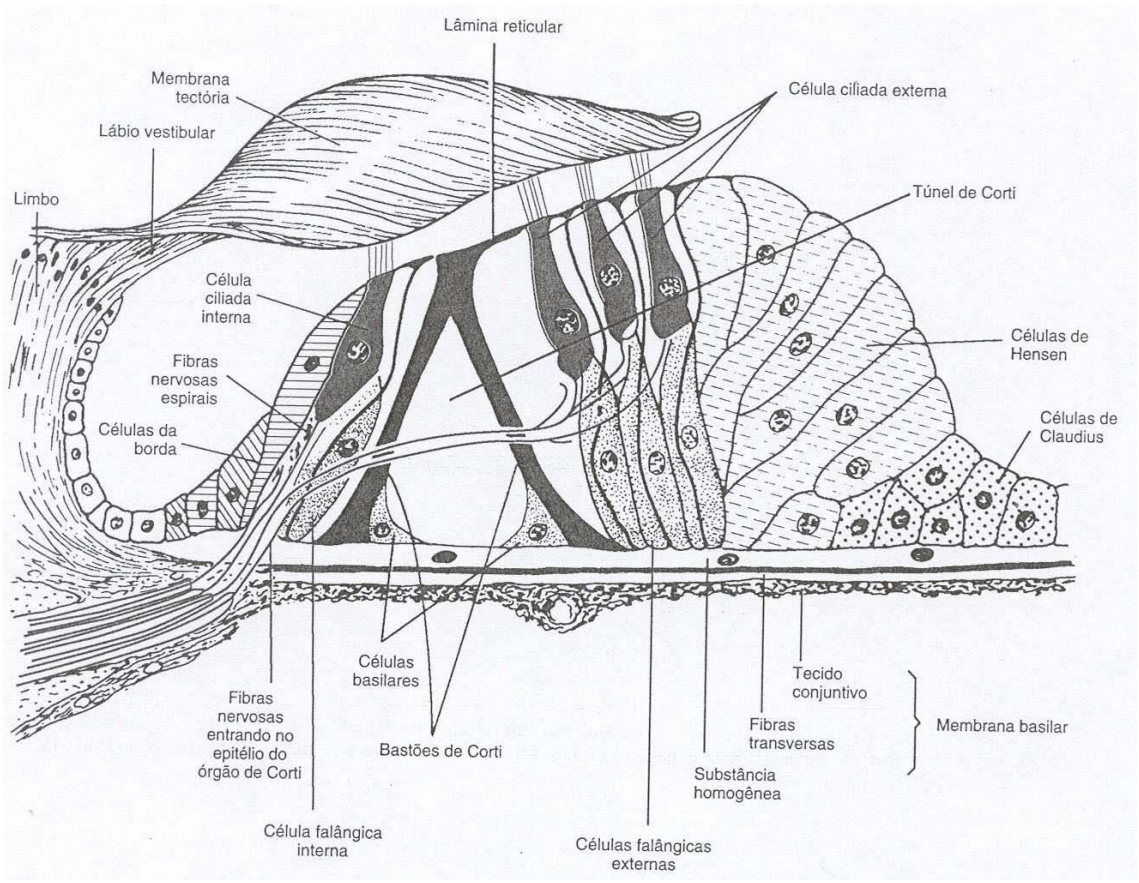


Figura 7 – Anatomia do Órgão de Corti (ZWICKER; FASTL, 1999).

Um comportamento similar é obtido ao se traçar a curva de resposta ao longo da membrana basilar para um tom em uma frequência específica. Para cada frequência, há um ponto da membrana basilar em que a vibração é máxima. A posição desse ponto, medida a partir do helicotrema é, aproximadamente, proporcional ao logaritmo da frequência do som. Ao redor desse ponto haverá uma faixa, de cerca de 1,5 mm, onde a vibração estará presente, atenuando-se conforme se afasta do ponto. Tal faixa determina o conceito de banda crítica, como será visto no capítulo 3.

CAPÍTULO 3

Fenômenos Auditivos

Para entendimento das técnicas psicoacústicas extraídas da técnica de avaliação de qualidade de voz *ANIQUE*, no capítulo 2 foi apresentada toda anatomia do sistema de audição humano. Neste capítulo apresentam-se algumas limitações do sistema de audição humano, importantes para compreensão dos motivos das utilizações das técnicas psicoacústicas.

3.1 Percepção de Volume de Som

No ser humano a percepção de volume não reflete o que ocorre com a pressão do ar. O sistema auditivo humano suporta variações de mais de 1.000.000 vezes a pressão da onda sonora no ar, mas não há sensação de um aumento tão grande de volume nessa situação. O sistema é mais sensível à variações em baixas pressões do que em as altas. Portanto, não existe uma linearidade entre a pressão do ar e a percepção de volume. Devido a essa não linearidade, as ondas sonoras são normalmente caracterizadas pelo seu nível logarítmico, que tem uma melhor relação com a percepção de volume.

A unidade mais usada para o nível de pressão sonora é a SPL (Nível de Pressão Sonora), que expressa o nível de pressão sonora em relação a um nível de referência (pressão sonora do limiar de audibilidade em 1 kHz) (Painter; Spanias, 2000), que é dada por:

$$L = 20\log_{10}\left(p/p_0\right) \ \ (\mathrm{dB_{SPL}}) \tag{3.1}$$

onde, $p_0 = 22\mu Pa$ e p em Pa (Pascal).

Na Quadro (1), são apresentados alguns exemplos de níveis de pressão sonora, em dB_{SPL}, onde o limiar da dor apresenta-se próximo a $130dB_{SPL}$.

Quadro 1 – Nível de pressão sonora para exemplos do cotidiano.

Situação	Pressão Sonora (dB_{SPL})
Limiar de Audibilidade	0
Murmúrio	30
Conversação Normal	60
Trânsito Pesado	80
Indústria Mecânica	100
Britadeira	120
Limiar da Dor	130
Motor de Jato	150

3.2 Resposta em Frequência do Sistema Auditivo

Assim como a percepção de volume, a percepção de frequência do sistema auditivo também não é linear. O ser humano consegue distinguir com mais precisão variações em baixas frequências do que em altas. Essa não linearidade acontece devido à estrutura física da membrana basilar, sendo a variação da largura e da rigidez em função da distância da base os principais fatores que explicam essa não linearidade. Portanto, a maior parte da membrana responde a sons com frequência inferior a 3 kHz, onde se encontra a maior quantidade de informação necessária para o entendimento da fala.

3.3 Limiar Absoluto de Audibilidade em Silêncio

O limiar absoluto de audibilidade em silêncio é caracterizado pela quantidade de energia necessária para que o ouvinte possa detectar um som com apenas um componente em frequência (um tom) em um ambiente em silêncio absoluto. Este limiar pode ser aproximado pela seguinte expressão analítica (Leite, 2003, Painter; Spanias, 2000)

$$T(f) = 3{,}64(f/1000)^{-0{,}8} - 6{,}5e^{-0{,}6(f/1000-3{,}3)^2} + 10^{-3}(f/1000)^4 \quad \text{(dB SPL)} \qquad (3.2)$$

O primeiro termo de $T(f)$ descreve o corte nas baixas frequências; o segundo descreve o aumento de sensibilidade do ouvido para a faixa de frequências em torno de 3 kHz; e o último descreve o corte nas altas frequências. O gráfico da Figura 8 foi obtido através dessa expressão e representa o limiar absoluto de audibilidade.

O primeiro termo pode ser interpretado como um resultado do ruído interno (causado por atividade muscular, fluxo de sangue etc.), ao passo que os dois últimos termos são interpretados como a característica de transferência do ouvido médio para o interno.

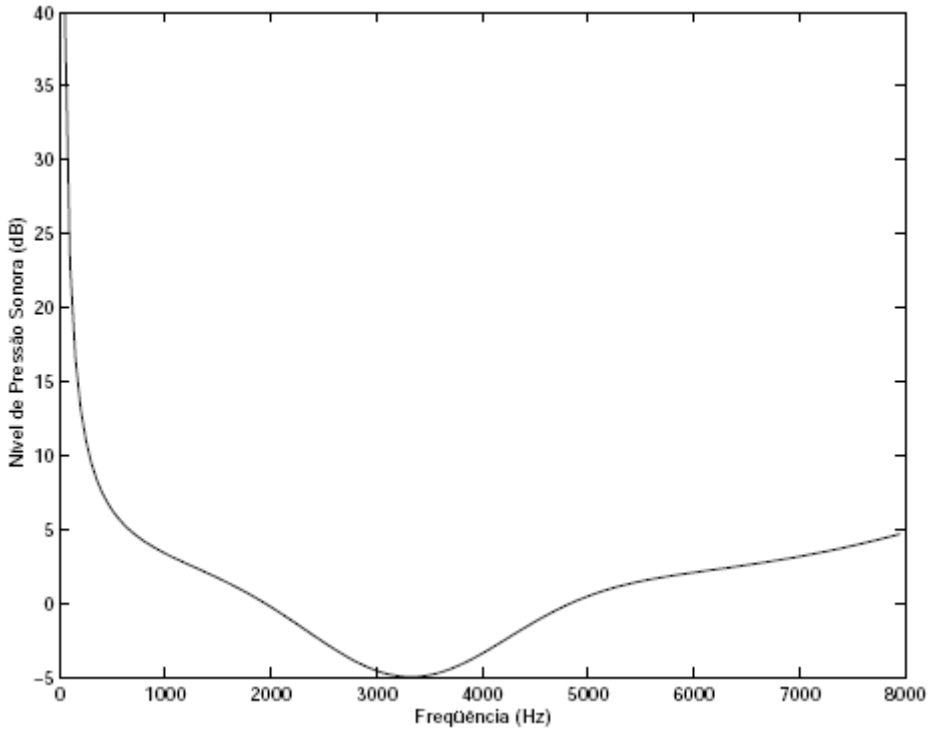


Figura 8 – Limiar absoluto de audibilidade.

3.4 Bandas Críticas

Uma banda crítica define uma faixa de frequências em torno de cada frequência característica associada a cada ponto da membrana basilar. Ela é uma faixa de frequência tomada ao redor de uma frequência central, na qual as respostas subjetivas do sistema auditivo mudam abruptamente (MOORE, 1997). Portanto, o sistema auditivo possui um comportamento diferente para sinais dentro e fora de uma banda crítica. A seguir, são apresentados dois experimentos típicos que demonstram a existência das bandas críticas (PAINTER; SPANIAS, 2000), onde ilustra-se na Figura 9 esses experimentos.

O primeiro experimento emprega um ruído de banda estreita com um determinado nível SPL. Ao aumentar a largura de banda deste ruído com o nível SPL constante, a intensidade de ruído percebida por um determinado ouvinte irá se manter constante. Isso será mantido até que se atinja um valor limite para a largura de banda do ruído. A partir desse limite, o ouvinte em questão perceberá um aumento na intensidade do ruído. Neste exemplo, a banda crítica é a máxima largura de banda em que o ouvinte não perceberá aumento da intensidade.

No segundo experimento, emprega-se um ruído de banda estreita e dois tons puros, com mesmo nível SPL, separados por uma distância Δf. Para uma determinada relação sinal/ruído, o ruído de banda estreita não será percebido na presença dos tons. Esse fenômeno chama-se de mascaramento auditivo e será apresentado mais adiante. Ao se aumentar a distância em freqüência (Δf) entre os tons, o ruído de banda estreita irá se manter imperceptível até o limite da banda crítica; neste instante, o ouvinte começará a perceber a existência do ruído. Esse mesmo experimento pode ocorrer invertendo-se os papéis, ou seja, um tom sendo mascarado por dois ruídos de banda estreita enquanto estes estão dentro da banda crítica.

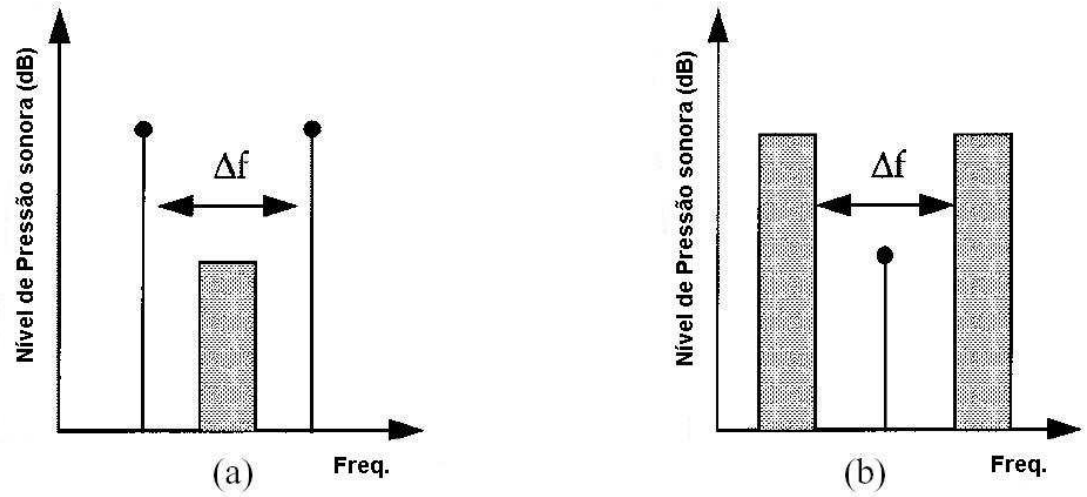


Figura 9 – Ilustração dos experimentos para identificação das bandas críticas (PAINTER; SPANIAS, 2000).

É importante notar que nos dois exemplos anteriores a banda crítica depende do ouvinte em questão e da frequência central do ruído de banda estreita. A partir de medidas realizadas para um grande número de ouvintes, uma aproximação usada para a banda crítica é dada por (PAINTER; SPANIAS, 2000)

$$BW_c(f) = 25 + 75\left[1 + 1{,}4(f/1000)^2\right]^{0.69} \quad \text{(Hz)} \tag{3.3}$$

Apesar das bandas críticas serem contínuas na frequência, para aplicações práticas é comum ser utilizado um conjunto discreto. O conjunto discreto mais utilizado, e que será utilizado no modelo perceptual estudado, está apresentado na Quadro (2) (CAVE, 2002), denominada escala Bark.

Segundo Pohlmann (1995), as características das bandas críticas estão intimamente ligadas às da membrana basilar, onde cada banda crítica corresponde a cerca de 1,3 mm de espaçamento na membrana basilar, o que corresponde a aproximadamente 100 Hz para frequências abaixo de 500 Hz e equivale, aproximadamente, a 20% da frequência central da banda para frequências acima de 1000 Hz (em direção à janela oval). Portanto, a resposta de amplitude em frequência, para cada banda crítica, pode ser modelada como a de um filtro passa-faixas com largura de faixa crescente com a frequência.

A não linearidade na largura das bandas críticas e sua dependência da frequência podem ser explicadas pelo fato de que a associação entre posição e frequência que ocorre na membrana basilar não é linear, como visto anteriormente.

Embora exista uma banda crítica ao redor de cada frequência, na maioria das aplicações adotam-se dos valores mostrados na Quadro (2). A distância de uma banda crítica é conhecida como um Bark. A função a seguir permite converter frequências em Hertz para a escala Bark (PAINTER; SPANIAS, 2000):

$$z(f) = 13\arctan(0{,}00076\, f) + 3{,}5\arctan\left[(f/7500)^2\right] \text{ (Bark)} \tag{3.4}$$

Quadro 2 – Frequências características das bandas críticas.

Bark	Frequências Características (Hz)	Bark	Frequências Características (Hz)
1	50	13	1850
2	150	14	2150
3	250	15	2500
4	350	16	2900
5	450	17	3400
6	570	18	4000
7	700	19	4800
8	840	20	5800
9	1000	21	7000
10	1170	22	8500
11	1370	23	10500
12	1600		

3.5 Mascaramento

O efeito de mascaramento ocorre quando um som, denominado de mascarado, se torna imperceptível para um ouvinte devido à presença de outro som, denominado de mascarador.

Esse fenômeno ocorre com muita frequência no cotidiano de todas as pessoas. Um exemplo é o som de um despertador de um relógio de pulso que é perceptível em locais tranquilos, mas pode ser imperceptível em locais barulhentos como em um *shopping center* ou um show de rock. Isto indica que o limiar de audibilidade depende do ambiente.

Para a área de codificação de sinais de áudio o estudo dos princípios de mascaramento é muito importante. Atualmente, os mais importantes algoritmos de codificação de áudio de alta fidelidade utilizam informações relativas ao mascaramento para diminuir sua taxa de compressão ou melhorar sua qualidade para uma mesma taxa. Essa diminuição ocorre quando os codificadores adicionam o ruído de codificação, de maneira que eles sejam mascarados, ou seja, que sua potência esteja abaixo do limiar de mascaramento.

Com o objetivo de dificultar a pirataria na área de áudio, pesquisas estão sendo feitas buscando-se adicionar informações aos sinais de áudio (copyright, permissões etc.), de maneira que as informações adicionadas façam parte do sinal de áudio, não sendo perceptível ao ouvinte, e cuja remoção não seja possível sem a destruição, ao menos parcial, do sinal de áudio.

O mascaramento é normalmente classificado em duas categorias principais: simultâneo e não simultâneo. A Figura 10 ilustra esses tipos de mascaramento, onde pode-se observar o limiar de mascaramento (linha pontilhada) em função do tempo, na qual o sinal mascarador (linha sólida) está presente por 200 ms, enquanto que seus efeitos estão presentes por cerca de 450 ms.

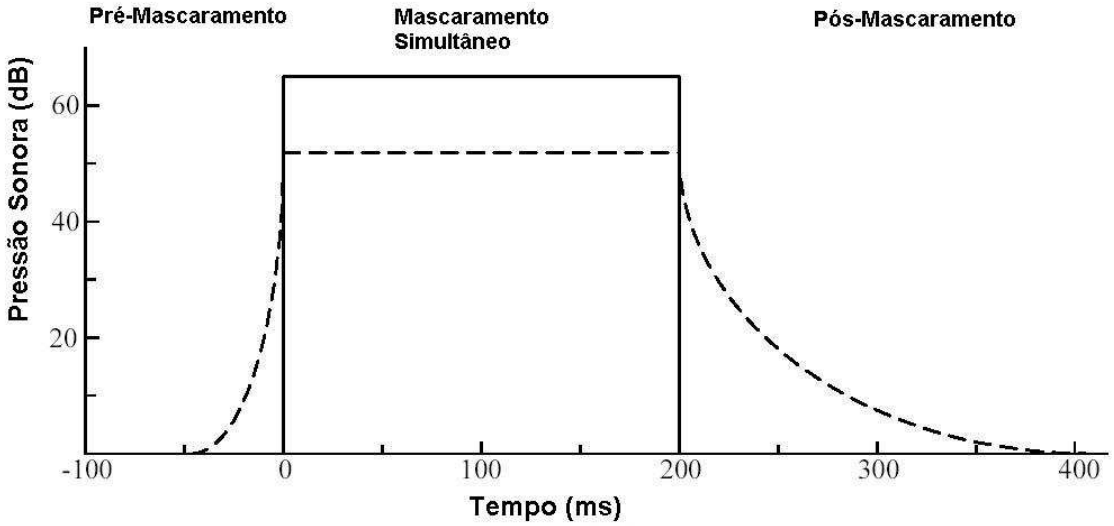


Figura 10 – Principais tipos de mascaramento (CAVE, 2002).

3.5.1 Mascaramento Não Simultâneo

O mascaramento não simultâneo é aquele que ocorre na ausência de um sinal mascarador. Ele pode ocorrer anteriormente à presença do sinal mascarador (pré-mascaramento), ou posteriormente (pós-mascaramento), como se mostra na Figura 10.

O pré-mascaramento ocorre, principalmente, devido à limitação da resolução temporal do sistema auditivo humano. Seu efeito significativo tem a duração de cerca de 2 ms. Devido a essa característica, o pré-mascaramento tem recebido bem menos atenção do que os demais tipos de mascaramento. Estudos mostram que 2 ms antes da presença do sinal mascarador o limiar de mascaramento já é 25 dB inferior ao limiar do mascaramento simultâneo (Painter; Spanias, 2000). O pós-mascaramento tem efeito bem mais significativo do que o pré-mascaramento. Seus efeitos são observados em até 200 ms após a presença do sinal mascarador. De acordo com Moore (1995) há três fatores que contribuem para o pós-mascaramento: a continuação da vibração da membrana basilar após o término do sinal mascarador, a fadiga do nervo auditivo (ou tempo para sua adaptação à ausência do sinal mascarador) e a continuidade neural produzida pelo sinal mascarador em um nível mais alto.

3.5.2 Mascaramento Simultâneo

Mascaramento simultâneo ocorre quando dois tons de frequências próximas se encontram em uma mesma banda crítica, de modo que o tom de maior amplitude se sobrepõe ao de menor amplitude, num processo denominado de percepção sonora.

Observa-se na Figura 10 que o mascaramento simultâneo é o mais importante dos mascaramentos, pois ele atinge os maiores níveis de pressão sonora. A presença de um som de sinal mascarador cria tamanha excitação na membrana basilar e nas células ciliadas do órgão de Corti que as oscilações provocadas pelo sinal mascarado não são percebidas pelo ouvinte.

Um exemplo apresentado em Cave (2002) mostra que se um ruído com largura de banda de 1 Bark e nível de 40 dB for adicionado a um tom puro de 20 dB dentro da mesma banda crítica, será observado um aumento de apenas 0,04 dB no nível de pressão sonora.

O mascaramento simultâneo pode ser facilmente observado com um exame de audiometria na presença do sinal mascarador. A Figura 11, ilustra a alteração do limiar de audibilidade devido à presença de um tom com nível de pressão sonora de 70 dB_{SLP} e com frequência de 1 kHz. Qualquer sinal com intensidade inferior à do limiar de mascaramento será mascarado.

A seguir, apresenta-se o mascaramento simultâneo para diferentes combinações de sinais mascaradores e mascarados.

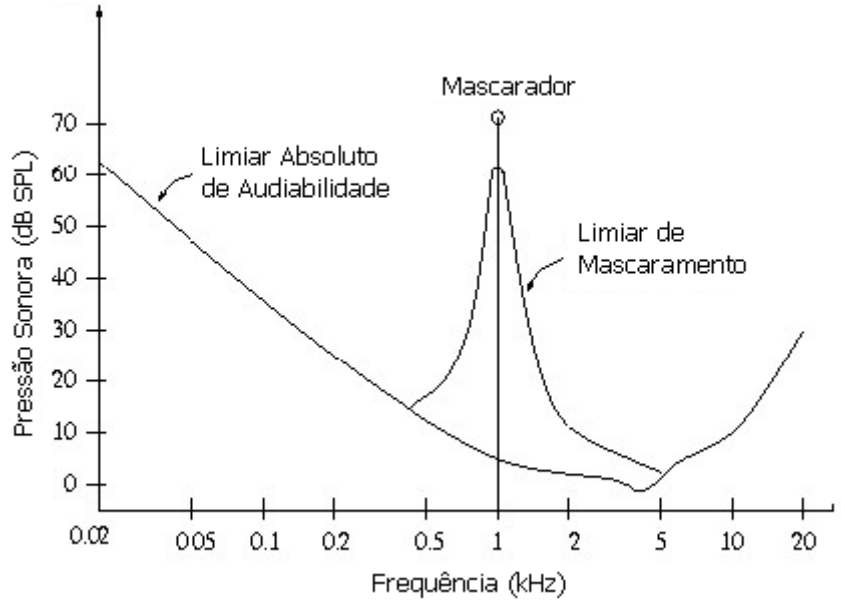


Figura 11 – Ilustração do efeito do mascaramento simultâneo (LEITE, 2003).

Ruído Mascarando Tom

Nessa configuração, um ruído de banda estreita mascara um tom puro. O mascaramento acontece somente quando a intensidade do tom mascarado é menor do que um determinado limiar, que está diretamente relacionado à intensidade do ruído mascarador e à frequência do sinal mascarado. O limiar possui seu valor máximo quando o tom mascarado está presente na frequência central do ruído mascarador (PAINTER; SPANIAS, 2000).

Na maioria dos estudos, o limiar de mascaramento para esse cenário varia aproximadamente em 5 dB. Com isso, pode acontecer de um ruído de menor intensidade mascarar um tom de maior intensidade.

Na Figura 12 há um ruído com largura de banda de 1 Bark, frequência central de 410 Hz e intensidade de 80 dB_{SPL}, mascarando um tom de 76 dB_{SPL} de mesma frequência central.

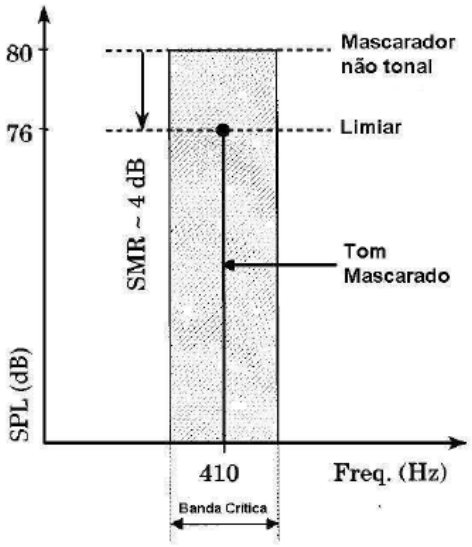


Figura 12 – Exemplo de ruído de banda estreita mascarando tom (PAINTER; SPANIAS, 2000).

Tom Mascarando Ruído

Contrário ao que ocorre com o ruído mascarando tom, nessa configuração um tom mascara um ruído de banda estreita, dado que o espectro do ruído esteja abaixo de um limiar diretamente relacionado à intensidade do tom mascarador. Para esse cenário, o limiar de mascaramento varia entre 21 e 28 dB (SCHROEDER; ATAL; HALL,1979). Com isso, observar-se uma assimetria no poder de mascaramento do ruído e do tom, na qual o ruído possui um poder de mascaramento muito maior.

Como acontece na configuração do ruído mascarando o tom, o limiar de mascaramento possui seu valor máximo quando o tom mascarador está no centro do espectro do ruído mascarado. A Figura 13 ilustra essa configuração de mascaramento.

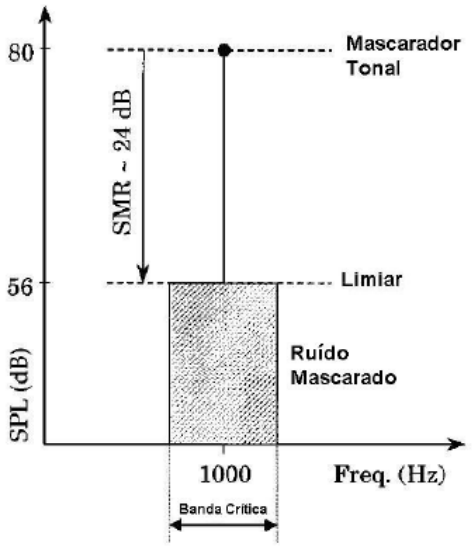


Figura 13 – Exemplo de sinal tonal mascarando ruído de banda estreita (PAINTER; SPANIAS, 2000).

Ruído Mascarando Ruído

A configuração de ruído de banda estreita mascarando ruído de banda estreita é mais complexa de ser analisada que a de ruído mascarando tom e vice-versa. Limiares da ordem de 26 dB já foram observados para esse tipo de mascaramento (PAINTER; SPANIAS, 2000).

Tom Mascarando Tom

A configuração de tom mascarando tom tem pouca utilidade para a área de codificação de áudio ou fala. Isso se deve ao fato de que os cenários de mascaramento para sinais de áudio e fala são mais complexos do que sinais puramente tonais.

CAPÍTULO 4

Sistema ANIQUE de Avaliação Objetiva de Qualidade de Voz

Neste capítulo é feita uma apresentação resumida sobre a importância dos modelos objetivos de avaliação de qualidade de voz e apresenta-se um estudo mais detalhado da técnica *ANIQUE,* de onde obtém-se os parâmetros necessários para aprimoramento da técnica de redução de ruído em sinais de voz proposta neste trabalho.

4.1 Sistema de Avaliação Objetiva de Voz

As modernas redes de telecomunicações estão cada vez mais complexas. Além disso, a rede tradicional de telefones públicos existente está conectada com as mais modernas redes de comunicação, como as redes baseadas no Sistema de Comunicação Móvel Global (GSM), no Sistema de Múltiplo Acesso por Divisão de Código (CDMA), no Sistema de Telecomunicações Móvel Universal (UMTS) e no Sistemas de Comunicação Via Internet (VoIP). Considerando que essas redes de comunicações são altamente distribuídas e são algumas vezes conectadas entre si em chamadas telefônicas, o número de fatores que degradam a qualidade auditiva do sinal de voz transmitido é elevado. Além disso, essas redes de comunicação têm que lidar com a relação entre qualidade de serviço e custo de operação. Boa qualidade com baixo custo operacional é o objetivo que se busca. Dessa forma, a avaliação da qualidade de sinais de voz sobre a moderna rede de telecomunicações é muito importante não somente para o projeto do sistema de rede de comunicação e desenvolvimento, mas também para o sustento da qualidade de serviço (QoS).

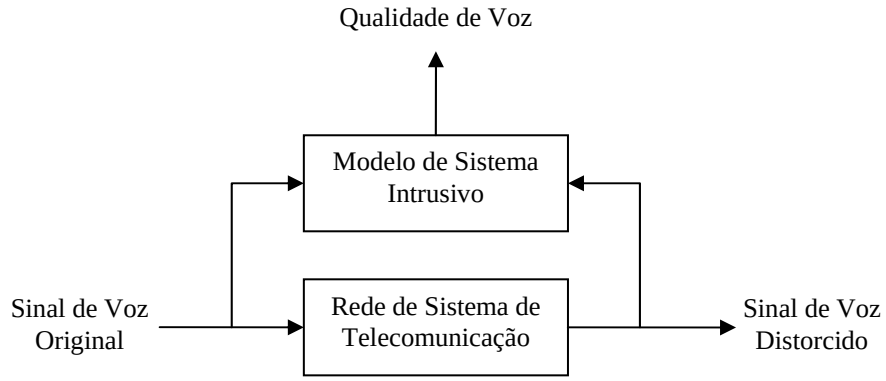
Considerando que a natureza da qualidade de sinais de voz é uma sensação subjetiva para ouvintes humanos, o caminho mais confiável para se avaliar a qualidade de um sinal é executar um teste de escuta subjetiva. Historicamente, testes formais de escuta subjetiva têm sido usados em avaliações de desempenho de sistemas de processamento de sinais de voz e áudio, como os codificadores de sinais, por exemplo. Um dos mais usados em testes de escuta é o teste Absolute Category Rating (ACR). Nesses testes, é pedido para um número de ouvintes classificarem a qualidade de pequenas sentenças de sinais processado pelo sistema em teste em uma escala de 5 pontos (5: excelente, 4: bom, 3: regular, 2: ruim, 1: péssimo). A média de classificação é comumente referida como Mean Opinion Score (MOS) (ITU-T Rec P.800, 1996).

Em geral, testes de escutas subjetivas requerem o controle absoluto de sons externos à sala de teste para obter resultados confiáveis relacionados somente com a qualidade do sinal em teste. Esses testes são caros e demorados. Durante o desenvolvimento de sistemas de redes de comunicação e suas posições estratégicas de consumo, é altamente necessário investigar o impacto de componentes específicos do sistema, suas combinações e conjuntos de parâmetros do sistema na percepção de qualidade do sinal. Dada a dificuldade em se obter esses resultados rapidamente e constantemente por testes subjetivos, é desejável ter um modelo computacional que possa avaliar e classificar um sinal de voz de uma maneira confiável. Durante décadas, vários modelos objetivos de estimação de qualidade de sinais de voz têm sido propostos. Dentre vários, os destaques são a Measuring Normalizing Block (MNB) (Voran, 1999), a Perceptual Speech Quality Measure (PSQM) (ITU-T Rec. P.861, 1996). e a Perceptual Evaluation of Speech Quality (PESQ) (ITU-T Rec. P.862, 2001), sendo essas duas útlimaso adotados pela International Telecommunication Union Telecommunication Standardization Sector (ITU-T) como padrão de recomendação para modelos objetivos de estimação de qualidade de voz dentro da faixa de telefonia (300 a 3400 Hz) (BEERENDS; STEMERDINK, 1994).

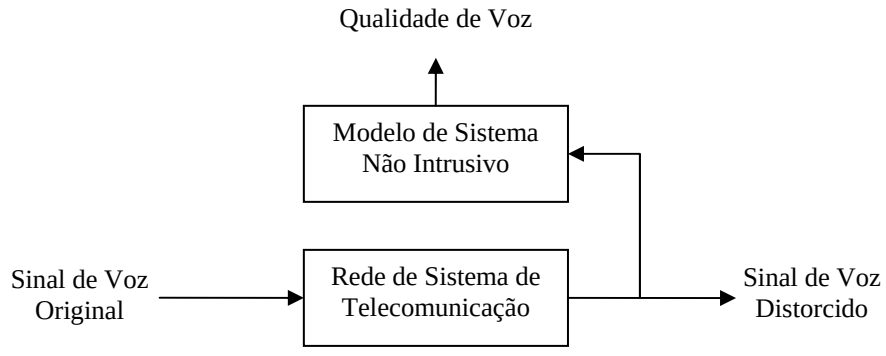
Para avaliar a qualidade subjetiva do sinal de voz degradado, os métodos convencionais requerem uma fonte de sinal de voz não degradado para referência que, juntamente com o sinal degradado, são aplicados na entrada do sistema. O maior inconveniente desses métodos é que na maioria das aplicações reais não se tem um sinal de referência. A alternativa seria uma metodologia que não dependesse do sinal de referência, que tem sido mencionada na literatura como modelo não intrusivo de estimativa da qualidade de sinais de voz. O maior desafio para esta metodologia é conseguir modelos que representem adequadamente os sinais de fala do ponto de vista da percepção auditiva.

4.2 Modelos Intrusivos e Não-Intrusivos

Atualmente existem várias pesquisas visando obter técnicas de avaliações objetivas de qualidade utilizando sistemas intrusivos e não intrusivos. Modelos intrusivos são processos onde é necessário utilizar um sinal de voz de referência de alta qualidade para avaliar o sinal de voz degradado, como apresentado na Figura 14(a). Em contraste aos modelos intrusivos, o método não intrusivo é um modelo desafiador no sentido de chegar ao objetivo de avaliar a qualidade do sinal, conforme apresentado Figura 14(b), sem qualquer sinal de referência.



(a) - Sistema de Avaliação Intrusiva.



(b) - Sistema de Avaliação Não Intrusiva.

Figura 14 – Diagrama de blocos de sistemas (a) intrusivos de avaliação de qualidade de voz e (b) e sistemas não intrusivos de avaliação de qualidade de voz.

Métodos não intrusivos (algumas vezes chamado, single-ended ou output based methods) possuem grande potencial para aplicações reais como, por exemplo, o monitoramento de qualidade de sinais de rede de comunicação em serviço (in-services networks), onde não existe um sinal de voz para ser usado como referência. Já os modelos intrusivos só podem ser usados em testes específicos, já que o sinal original é necessário no processo de avaliação.

4.3 ANIQUE

A técnica *ANIQUE* é um modelo não intrusivo de estimação de qualidade de sinais de voz (Kim, 2005) e foi um dos modelos dos candidatos para a padronização P.SEAM (Single-Ended Assessiment Models) pela ITU-T (Kim; Tarraf, 2004). Nessa técnica, um dos pontos básicos usados para a estimação da qualidade de sinais de voz é a representação da envoltória temporal do sinal. O modelo proposto é baseado no princípio de funcionamento do sistema de audição e articulação doe ser humano. Avaliações experimentais em 35 diferentes testes demonstraram eficiência do modelo proposto por Kim (2005).

Na Figura 15 mostra-se o diagrama de blocos completo do modelo *ANIQUE*. O sinal de voz é inicialmente processado para ma normalização do nível de amplitude e adequação à faixa básica de frequência. No bloco da Figura 15 denominado de "*Estimação de Qualidade*" e apresentado em detalhes na Figura 16, o sinal de voz pré-processado é dividido em uma sequência de quadros (*frames*) no tempo e a qualidade $v_s(m)$ de cada quadro *m* é estimada. Na sequência, distorções temporais de descontinuidade no sinal são detectadas e o quadro de qualidade é modificado, gerando um quadro atualizado de qualidade $\tilde{v}_s(m)$ que é usado para estimar a qualidade *Qs*. O bloco "*Compensação de Expressão*" da estimação de qualidade é compensado pelo processamento do sinal no caminho inferior da Figura 15. Para realização deste trabalho, somente o bloco "*Estimação de Qualidade*" com suas técnicas psicoacústicas será analisado e incorporado às técnicas de redução de ruído em sinais de voz, pois neste bloco está o principal sistema da *ANIQUE* que permite um aprimoramento das técnicas de *MSV*.

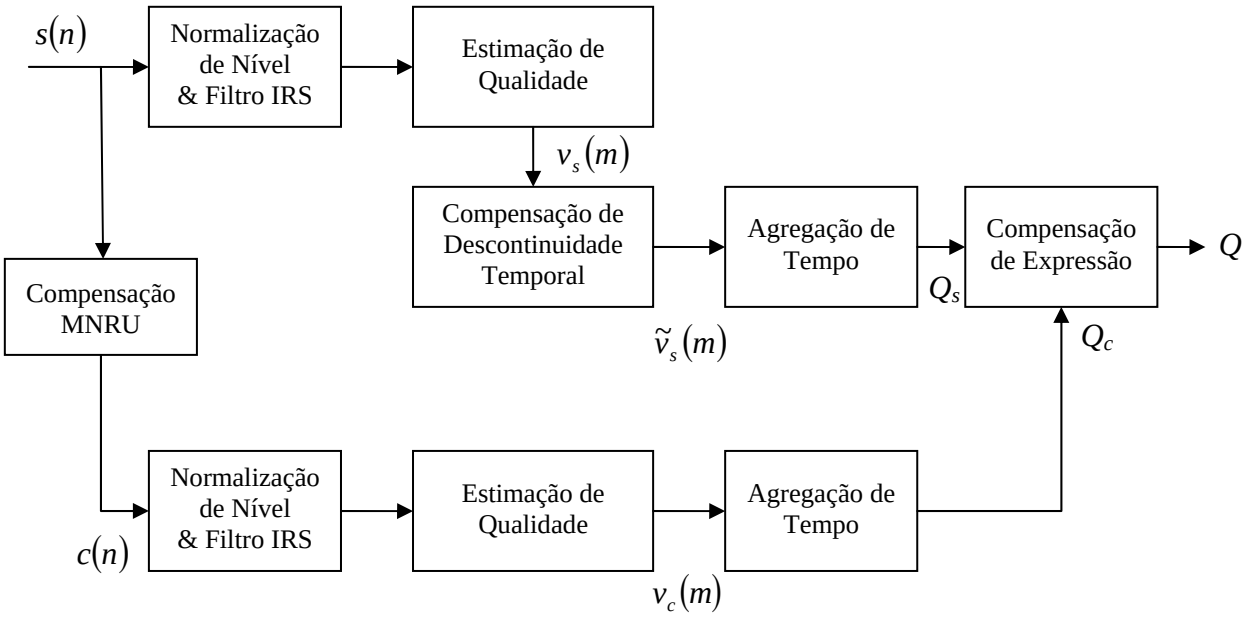


Figura 15 – Diagrama de blocos da técnica *ANIQUE*.

4.4 Banco de Filtros Cocleares e Envoltória Temporal

Na Figura 16 mostra-se em detalhes o diagrama do bloco "*Estimação de Qualidade"*, onde é simulado o primeiro estágio do sistema de audição. Após a normalização de amplitude e a filtragem do sinal de voz pelo IRS, o sinal é filtrado por um banco de filtros de banda crítica gammatone, onde são usados 23 filtros para simular o processo do desempenho da cóclea (Slaney, 1993). Então, calcula-se a envoltória e a fase instantânea de cada sinal filtrado pelo banco de filtros de banda crítica.

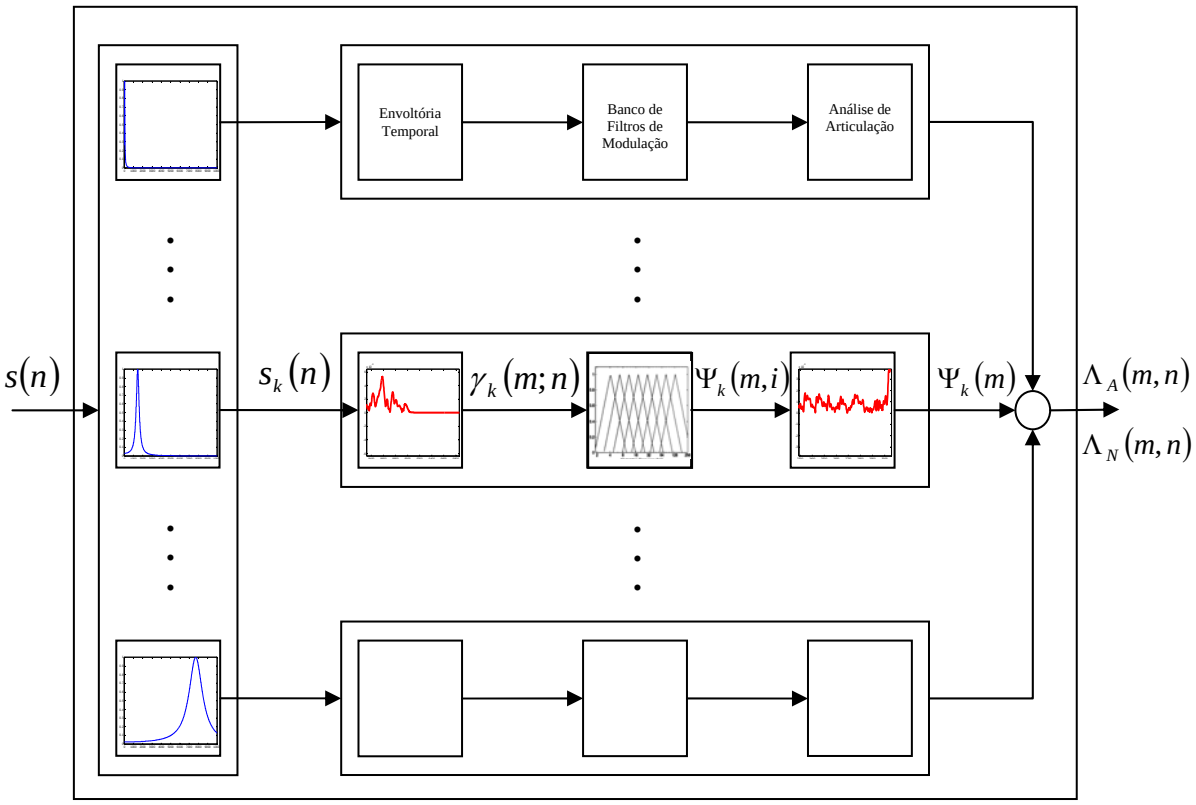


Figura 16 – Diagrama de blocos do bloco *"Estimação de Qualidade"*.

A decomposição do sinal de voz em sua envoltória e portadora provê uma percepção útil nos sinais de voz, pois a envoltória é conhecida por ser relevante em vários atributos de percepção de voz, como a inteligibilidade e qualidade. Drullman, Festen e Plomp (1994) mostraram o quanto de informações pode ser obtido da envoltória sem afetar o desempenho humano em tarefas de identificações de fonemas. Em termos de qualidade, Ghitza (2001) investigou a relação entre a envoltória e a qualidade de sinais de voz e propôs um novo método para codificar sinais de áudio. O uso da envoltória em avaliação objetiva de qualidade de sinais de voz foi proposto por Kim e Tarraf (2004).

4.5 Banco de Filtro de Modulação

A sensibilidade humana para a envoltória é um interessante tópico em física acústica. Experimentos de detecção mostram que a sensibilidade humana para modulação pode ser representado por um filtro passa-baixas com frequência de corte em aproximadamente 50 Hz (VIEMEISTER, 1997). Dau, Puschel e Kohlrausch (1997a, 1997b) propuseram um modelo de audição no qual um banco de detectores de modulação é empregado para explicar a detecção de modulação e o mascaramento de modulação de dados obtidos em experimentos físico acústicos. Estudos neurofisiológicos sustentam essa idéia e mostram que a decomposição da envoltória funciona mais no nível central do sistema de audição do que nos níveis periféricos. Por exemplo, em (GIRAUD et al., 2000), a representação cortical da envoltória de sons é investigada usando *Functional Magnetic Resonance Imaging* (*F-MRI*), e foi mostrado que o caminho da audição é organizado como um banco de filtros hierárquico, onde cada nível de processamento é ajustado para uma determinada frequência modulada em amplitude (AM): 256 Hz para o núcleo da cóclea, 32-256 Hz para o colliculus inferior, 16 Hz para o corpo geniculado medial, 8 Hz para o córtex primário e 4-8 Hz para a região secundária.

No modelo *ANIQUE,* o caminho da audição é modelado por um banco de filtros de modulação. Para cada envoltória mencionada anteriormente, calcula-se a transformada de Fourier e, com o banco de filtros de modulação, calcula-se $\Psi_k(m,i)$, que será utilizada para o cálculo do bloco de análise de articulação.

4.6 Análise de Articulação

O mecanismo de percepção de qualidade de sinal de voz no sistema de audição humano ainda não é claro. No modelo *ANIQUE* é considerada a hipótese de que o sistema auditivo utiliza um tipo de modulação espectral ao longo do caminho da audição na determinação da percepção de qualidade de sinal, separando fatores que contribuem para a naturalidade do sinal de voz de sons indesejáveis, que não podem ser produzidos pelo sistema de articulação humana. Assim, tem-se uma separação entre a energia média dos sons indesejáveis produzidos na taxa além da velocidade do sistema de articulação humano e a energia média dos sinais que não podem ser produzidos pela articulação humana.

CAPÍTULO 5

Técnicas de Redução de Ruído Baseadas na Relação Sinal/Ruído a Priori

A redução de ruído em sinais de voz, em seu vasto campo de aplicação, começou a ser explorada com mais intensidade na década de 70. No entanto, nos anos 40 Egan e Wiener (1946) já publicavam trabalhos científicos enfocando o problema da inteligibilidade dos sinais de voz. Na época, com aplicações voltadas às telecomunicações, eles já usavam o efeito de mascaramento auditivo do ruído. Com a continuação desses estudos, as técnicas de melhoramento de sinais de voz (MSV) evoluíram consideravelmente. Hoje, em muitas dessas técnicas usa-se a estimação da amplitude espectral de curto prazo, onde a principal vantagem é a facilidade de implementação, tendo-se como ferramenta básica a transformada de Fourier de curto prazo (STFT).

Neste capítulo apresenta-se inicialmente um estudo de duas técnicas importantes de redução de ruído em sinais de voz: a *subtração espectral* e a *minimização do erro quadrático médio*. Na sequência são apresentados os dois principais parâmetros que são fundamentais na melhoria das técnicas clássicas de redução de ruído, que são a *SNR_Prio* e a *SNR_Post*.

5.1 Subtração Espectral

O ruído aditivo pode degradar a qualidade e a inteligibilidade dos sinais de voz na maioria das aplicações de telecomunicações. Esse ruído pode ser proveniente de diferentes fontes, tais como ruído ambiente, perdas devidas às codificações de enlaces digitais, etc.

Com o objetivo de melhorar a qualidade do sinal de voz, muitas pesquisas têm sido realizadas nesta área e muitas metodologias têm sido propostas. Dentre essas várias metodologias tem-se a subtração espectral, que é uma técnica me que a melhoria na qualidade do sinal de voz é obtida por meio de uma subtração entre o espectro do sinal de voz contaminado por ruído aditivo e o espectro da estimativa média do ruído, avaliada em instantes de silêncio.

A técnica baseada na subtração espectral foi proposta por Boll (1979) e foi uma das primeiras a oferecer possibilidades reais de implementação prática, pois considera que os sinais de voz e de ruído são processos aleatórios estacionários e independentes.

A Subtração Espectral pode ser aplicada somente para sinais ruidosos estacionários (OPPENHEIM; SCHAFER, 1989). Porém, sabe-se que um sinal de voz apresenta uma característica de não-estacionariedade extremamente forte. No entanto, estudos mostram que para pequenos intervalos de tempo, normalmente com duração de até 40 ms (OPPENHEIM; SCHAFER, 1989), o sinal de voz pode ser considerado aproximadamente estacionário. Com isso, pode-se aplicar de forma direta a Transformada de Fourier de Curto Prazo (STFT) (RABINER; SCHAFER, 1988).

Baseada nas premissas de que o ruído é aditivo e que seu espectro de potência é conhecido, a técnica de subtração espectral busca subtrair, do sinal degradado, a informação referente ao espectro do ruído.

Considere um sinal de voz puro $v(t)$, degradado por um ruído aditivo $r(t)$, formando um sinal ruidoso $y(t)$ como apresentado na equação (5.1):

$$y(t) = v(t) + r(t) \tag{5.1}$$

Após um processo de amostragem do sinal ruidoso, pode-se reescrever a equação (5.1) como segue:

$$y(n) = v\ (n) + r\ (n) \tag{5.2}$$

No domínio da frequência tem-se:

$$Y(\omega) = V\ (\omega) + R(\omega) \tag{5.3}$$

Tomando-se o quadrado na equação (5.3) e usando a hipótese de que o ruído é aditivo e descorrelacionado com o sinal de voz, obtêm-se:

$$\left|V\ (\omega)\right|^2 = \left|Y\ (\omega)\right|^2 - \left|R\ (\omega)\right|^2 \tag{5.4}$$

A partir da equação (5.4) nota-se que o propósito da subtração espectral, portanto, é a obtenção de uma estimativa do sinal não degradado a partir do sinal degradado e de um conhecimento prévio da estatística do ruído adicionado ao sinal. Além disso, observa-se que não existe uma recuperação da fase do sinal. Isto ocorre porque, além de não existir método que permita uma estimação da fase original, estudos mostram que o ouvido humano é pouco sensível às variações de fase nestas condições (sinal puro/sinal ruidoso) (FLANAGAN, 1972). Portanto, a subtração espectral é aplicada somente para o espectro de potência do sinal, ou mesmo do espectro de amplitude, preservando-se a fase do sinal ruidoso.

Assim, estimando-se a potência do ruído ($\mathrm{E}\{|\ R\ (\omega)\ |^2\}$) e aplicando-se o valor estimado na equação (5.4), obtém-se:

$$\mathrm{E}\{|\ V\ (\omega)\ |\} = \sqrt{\ |\ Y\ (\omega)\ |^2\ -\ \mathrm{E}\{|\ R\ (\omega)\ |^2\}} \tag{5.5}$$

A partir da equação anterior pode-se definir a função de transferência do filtro redutor de ruído como sendo:

$$|\ H\ (\omega)\ | = \frac{\sqrt{\ |\ Y\ (\omega)\ |^2\ -\ \mathrm{E}\{|\ R(\omega)\ |^2\}}}{|\ Y\ (\omega)\ |} \tag{5.6}$$

Da equação (5.6), verifica-se que o filtro obtido pelo método da *subtração espectral* só é realizável se a potência estimada do ruído for menor ou igual à potência do sinal ruidoso. No entanto, considerando o fato de os sinais terem fases aleatórias, isto não é garantido.

Esta aleatoriedade da fase pode, no processo de adição dado na equação (5.3), resultar na formação de um sinal ruidoso de potência inferior à do ruído. Uma solução para este problema é fazer uma "retificação de meia-onda", que resultaria num filtro dado por:

$$|H_1(\omega)| = \begin{cases} \dfrac{\sqrt{|Y(\omega)|^2 - E\{|R(\omega)|^2\}}}{|Y(\omega)|}, & \textit{se} \quad |Y(\omega)|^2 \geq E\{|R(\omega)|^2\} \\ 0, & \textit{caso contrário} \end{cases} \tag{5.7}$$

Uma outra solução seria tomar o valor absoluto obtido a partir da equação (5.6), definindo-se uma outra função de transferência, dada por:

$$|H_2(\omega)| = \frac{\sqrt{\;||Y(\omega)|^2 - \mathrm{E}\{|R(\omega)|^2\}|\;}}{|Y(\omega)|} \tag{5.8}$$

Entretanto, qualquer que seja a solução adotada, o resultado será uma alteração aleatória das amplitudes nas frequências onde isto ocorre, acarretando a geração de tons indesejáveis no sinal processado. Esses tons são denominados na literatura de ruído musical e são os maiores inconvenientes na aplicação da subtração espectral e de outras técnicas (CAPPÉ, 1994; EPHRAIM, 1992).

A Figura 17 mostra o esquema básico da *subtração espectral*, destacando-se a reutilização da fase do sinal ruidoso na reconstrução do sinal processado.

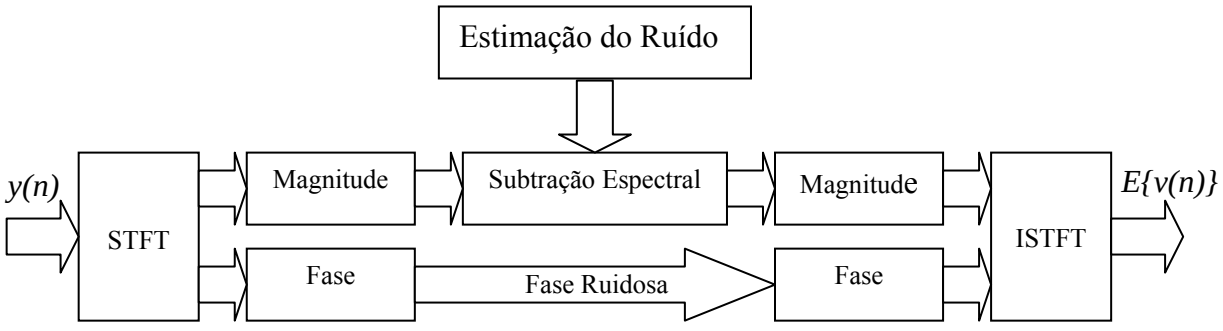


Figura 17 – Esquema básico da subtração espectral.

5.2 Minimização do Erro Quadrático Médio

Nesta técnica, desenvolvida por Ephraim e Malah (1984), os sinais de voz e ruído são modelados estatisticamente como processos aleatórios estacionários e independentes. Eles consideram que os coeficientes da expansão de Fourier são variáveis aleatórias gaussianas estatisticamente independentes. A base matemática do modelamento é o teorema do limite central, considerando que cada coeficiente de Fourier é, no final, uma soma ponderada (ou integral) de variáveis aleatórias resultantes de um grande número de amostras do processo. Assume-se que os processos possuem média igual à zero.

Os sinais de voz puro e ruidoso podem ser escritos como

$$V(\omega) = |V(\omega)| \cdot e^{j\,\alpha\,(\omega)} \tag{5.9}$$

$$Y(\omega) = |Y(\omega)| \cdot e^{j\,\theta\,(\omega)} \tag{5.10}$$

O objetivo é estimar o espectro de amplitude do sinal de voz $|V(\omega)|$ a partir do sinal ruidoso $y(t)$, dentro de um determinado intervalo de observação (0 ~ *T*).

Assumindo que as componentes espectrais são estaticamente independentes, obtém-se então o estimador *MMSE* diretamente do sinal ruidoso, ou seja,

$$|\hat{V}(\omega)| = \mathrm{E}\{|V(\omega)| \setminus y(t)\} \qquad , \; 0 \le t \le T \tag{5.11}$$

$$|\hat{V}(\omega)| = \mathrm{E}\{|V(\omega)| \setminus Y_0, Y_1, ...\} = \mathrm{E}\{|V(\omega)| \setminus Y_i\} \tag{5.12}$$

onde Y_i é a amostra de $|Y(\omega)|$ e k representa a frequência específica analisada dentro do intervalo observado.

Da equação (5.12) tem-se,

$$|\hat{V}(\omega)| = \frac{\int_0^{\infty}\int_0^{2\pi} v(\omega)\cdot p[Y(\omega)/v(\omega),\alpha(\omega)]\cdot p[v(\omega),\alpha(\omega)]\cdot d\alpha(\omega)dv(\omega)}{\int_0^{\infty}\int_0^{2\pi} p[Y(\omega)/v(\omega),\alpha(\omega)]\cdot p[v(\omega),\alpha(\omega)]\cdot d\alpha(\omega)dv(\omega)} \tag{5.13}$$

onde $p(.)$ representa a função densidade de probabilidade e $v(\omega)$ representa a variável aleatória do espectro de amplitude $V(\omega)$ do sinal de voz.

Assumindo o modelo gaussiano, tem-se:

$$p[Y(\omega)/v(\omega),\alpha(\omega)] = \frac{1}{\pi\sigma_R{}^2(\omega)}\cdot\exp\left\{-\frac{1}{\sigma_R{}^2(\omega)}|Y(\omega)-v(\omega)\cdot e^{j\alpha(\omega)}|^2\right\} \tag{5.14}$$

$$p[v(\omega),\alpha(\omega)] = \frac{v(\omega)}{\pi\sigma_v{}^2(\omega)}\cdot\exp\left\{-\frac{v^2(\omega)}{\sigma_v{}^2(\omega)}\right\} \tag{5.15}$$

Nas equações (5.14) e (5.15), $\sigma_R{}^2(\omega) = \mathrm{E}\{|R(\omega)|^2\}$ e $\sigma_v{}^2(\omega) = \mathrm{E}\{|V(\omega)|^2\}$, representam as variâncias do ruído e do sinal de voz, respectivamente.

Substituindo-se as equações (5.14) e (5.15) na equação (5.13) chega-se ao seguinte estimador:

$$|\hat{V}(\omega)| = \Gamma(1{,}5)\cdot\frac{\sqrt{\eta(\omega)}}{\gamma(\omega)}\cdot \mathrm{M}\left(-0{,}5;1{,}0;-\eta(\omega)\right)\cdot|R(\omega)| \tag{5.16}$$

onde $\Gamma(\cdot)$ representa a função gama e $\mathrm{M}(a;c;x)$ representa a função hipergeométrica.

Tem-se também que

$$\eta(\omega) = \frac{\xi(\omega)}{1+\xi(\omega)}\cdot\gamma(\omega) \tag{5.17}$$

sendo

$$\xi(\omega) = \frac{\sigma_v{}^2(\omega)}{\sigma_R{}^2(\omega)} \qquad e \qquad \gamma(\omega) = \frac{|Y(\omega)|^2}{\sigma_R{}^2(\omega)} \tag{5.18}$$

A equação (5.18) fornece os dois termos mais importantes do estimador desenvolvido por Ephraim e Malah (1984), isto é, $\xi(\omega)$ e $\gamma(\omega)$, que representam as relações sinal/ruído a *Priori* e a *Posteriori*, respectivamente. Vale lembrar que estes termos foram introduzidos primeiramente por McAulay e Malpass (1980).

A equação (5.16) pode ainda ser desenvolvida, pois a função hipergeométrica pode ser escrita em termos de funções de Bessel. Neste caso tem-se:

$$|\hat{V}(\omega)| = \Gamma(1{,}5)\frac{\sqrt{\eta(\omega)}}{\gamma(\omega)}\exp\left[\frac{-\eta(\omega)}{2}\right]\cdot\left\{[1+\eta(\omega)]I_0\left[\frac{-\eta(\omega)}{2}\right]+\eta(\omega)I_1\left[\frac{-\eta(\omega)}{2}\right]\right\}\cdot|Y(\omega)| \quad (5.19)$$

Na equação (5.19), $I_0[\cdot]$ e $I_1[\cdot]$ representam as funções de Bessel modificada do tipo um e de ordens zero e um, respectivamente.

O parâmetro chave deste estimador é a relação sinal/ruído a priori, que deve ser cuidadosamente calculada, como será visto posteriormente. No entanto, pode-se antecipar que em se tratando de uma estimação, visto que o sinal disponível já incorpora o ruído, uma relação sinal/ruído a priori superestimada poderá causar distorções no sinal de voz processado. E, ao contrário, uma subestimação deixará um ruído residual forte.

5.3 Relação Sinal/Ruído na Redução de Ruído em Sinais de Voz

No item anterior foram apresentadas duas técnicas de redução de ruído: a s*ubtração espectral* e a *minimização do erro quadrático médio*. Com a definição desses novos parâmetros, a relação sinal/ruído a posteriori (*SNR_post*) e a relação sinal/ruído a priori (*SNR_prio*), pode-se melhorar as técnicas clássicas de redução de ruído, como a subtração espectral. Somente o uso do parâmetro *SNR_post* não elimina eficientemente o problema do ruído musical, sendo necessária a utilização do parâmetro *SNR_prio*.

Voltando à discussão da subtração espectral, na equação (5.20) tem-se o equacionamento do espectro de potência do sinal limpo obtido a partir de uma subtração entre a potência do sinal degradado e a potência do ruído, usando a hipótese de que o ruído é aditivo e descorrelacionado com o sinal de voz. A partir de uma manipulação matemática simples obtêm-se as equações (5.21) e (5.22), que são exatamente a *SNR_post* e a *SNR_prio*.

$$|V(\omega)|^2 = |Y(\omega)|^2 - |R(\omega)|^2 \tag{5.20}$$

$$SNR\_post\ (\omega) = \frac{|Y(\omega)|^2}{|R(\omega)|^2} \tag{5.21}$$

$$SNR\_prio\ (\omega) = \frac{|V(\omega)|^2}{|R(\omega)|^2} \tag{5.22}$$

5.4 Técnicas de MSV Baseadas na Relação SNR a Posteriori

Nas seções 5.2 5.3 foi apresentado o equacionamento da função de transferência do filtro redutor de ruído. Esta função pode ser estimada como:

$$|\hat{H}_S(\omega)|^2 = \frac{E\{|Y(\omega)|^2 - |R(\omega)|^2\}}{E\{|Y(\omega)|^2\}} \tag{5.23}$$

Das equações (5.20) e (5.22), o filtro pode ser reescrito como:

$$|\hat{H}_S(\omega)| = \sqrt{1 - \frac{1}{\hat{SNR}\_post\ (\omega)}} \tag{5.24}$$

O filtro dado na equação (5.24) será definido apenas se $SNR\_post\ (\omega) \geq 1$, o que implica numa possível "retificação" para os casos onde a potência do sinal ruidoso esteja abaixo da potência do ruído estimada. Mesmo assim, o fato de escrever-se o filtro em função da SNR_post já pode resultar em um melhor desempenho do filtro, tendo em vista que a retificação será feita na relação sinal/ruído a posteriori e não diretamente como apresentado anteriormente. Isto significa que, dependendo do processo de estimação da SNR_post pode-se reduzir o efeito do ruído musical.

5.5 Técnicas de MSV Baseadas na Relação SNR a Priori

Devido aos problemas apresentados pelas técnicas clássicas baseadas na $SNR\_post(\omega)$, Ephraim e Malah (1984) desenvolveram uma técnica onde a principal vantagem é a ausência do ruído musical (RABINER; SCHAFER, 1988)

Este filtro, apresentado no tópico 5.3, pode ser escrito como:

$$|\hat{H}_{EM}(\omega)| = \Gamma(1{,}5)\frac{\sqrt{\eta(\omega)}}{\gamma(\omega)}\exp\left[\frac{-\eta(\omega)}{2}\right]\cdot\left\{[1+\eta(\omega)]I_0\left[\frac{-\eta(\omega)}{2}\right]+\eta(\omega)I_1\left[\frac{-\eta(\omega)}{2}\right]\right\}| \qquad (5.25)$$

onde:

$$\eta(\omega) = \frac{\hat{SNR}\_prio\ (\omega)}{1+\hat{SNR}\_prio\ (\omega)}\cdot \hat{SNR}\_post\ (\omega) \qquad (5.26)$$

Nas equações (5.25) e (5.26), verifica-se que o parâmetro principal do filtro redutor de ruído obtido por Ephraim e Malah é a *SNR_Prio*, sendo *SNR_Post* um parâmetro secundário.

5.6 Técnicas Clássicas de Redução de Ruído usando a SNR a Priori

Os estudos apresentados anteriormente mostraram a importância da *SNR_Prio*, que pode ser uma solução para a eliminação do ruído musical. Tendo em vista os problemas discutidos com relação à utilização da *SNR_Post*, torna-se importante explorar o uso da *SNR_Prio* nas técnicas clássicas.

A *SNR_Post* pode ser escrita como uma função da *SNR_Prio*. Dividindo a equação (5.20) por $|R(\omega)|^2$ obtém-se:

$$\frac{|Y(\omega)|^2}{|R(\omega)|^2} = \frac{|V(\omega)|^2}{|R(\omega)|^2} + 1 \qquad (5.27)$$

De acordo com as definições das equações (5.21) e (5.22) tem-se que:

$$SNR\_post(\omega) = SNR\_prio(\omega) + 1 \tag{5.28}$$

A estimação da *SNR_Post* com base na *SNR_Prio* elimina definitivamente a necessidade de se fazer uma retificação. Uma consequência imediata é a redefinição dos filtros clássicos de redução de ruído.

Considerando a equação (5.26), o filtro da subtração espectral definido na equação (5.24) pode ser escrito em função da *SNR_Prio* como segue:

$$|\hat{H}_S(\omega)| = \sqrt{\frac{\hat{SNR}\_prio\ (\omega)}{1 + \hat{SNR}\_prio\ (\omega)}} \tag{5.29}$$

Esse novo filtro diminui significativamente o problema do ruído musical, mantendo o mesmo nível de redução de ruído, pois o mesmo é baseado na relação entre as potências do sinal e do ruído, no lugar da diferença entre essas potências, como definidos pela subtração espectral clássica.

CAPÍTULO 6

Obtenção de um Filtro baseado na *SE* e em Critérios Psicoacústicos

---

Neste capítulo apresentam-se o sistema desenvolvido e os detalhes da implementação com os novos parâmetros extraídos da técnica *ANIQUE* para aprimoramento da *subtração espectral* baseada na *SNR a Priori*. Uma breve descrição de todo o sistema é discutido inicialmente, seguido de uma descrição mais detalhada de cada processo.

6.1 Incorporação de Técnicas Psicoacústicas na de Redução de Ruído

O desenvolvimento teórico de filtros de redução de ruído é apenas uma parte do trabalho envolvido no melhoramento de sinais de voz. Assim, para aplicá-los, é necessário um conjunto de técnicas de processamento de sinais tais como janelamento temporal do sinal amostrado, transformação para o domínio da freqüência usando a STFT, detecção de intervalos de silêncio, etc. Na prática um dos fatores mais desafiadores é a estimação da potência do ruído, que exige a identificação dos intervalos de voz e de silêncio (*DVS*) em sinais com baixas relações sinal/ruído.

A STFT é uma das ferramentas mais utilizadas no processamento de sinais de voz (RABINER; SCHAFER, 1988). Com sua aplicação é possível obter a estimação de pequenas parcelas do espectro do sinal de voz, consideradas estacionárias quando os intervalos de análise variam entre 20 e 40 ms.

Na Figura 18 apresenta-se o diagrama de blocos de todo o sistema baseado na *subtração espectral* e na *SNR a Priori*, cuja estimação passa a ser feita com a incorporação dos novos parâmetros *Potência de Articulação* e *Potência de Não-Articulação*. O sinal

amostrado passa inicialmente por uma janela temporal, com dois objetivos: garantir um sinal discreto de duração limitada, para permitir o uso da transformada discreta de Fourier, e assegurar trechos do sinal que sejam praticamente estacionários. Associado à janela, deve-se definir também o intervalo de sobreposição das janelas consecutivas. Na maioria das aplicações de sinais de voz, as janelas consecutivas são sobrepostas com uma repetição de metade das amostras, ou seja, uma sobpreposição de típico é 50%. Na figura Figura 19 apresenta-se a curva característica da janela de Hanning, usada na implementação.

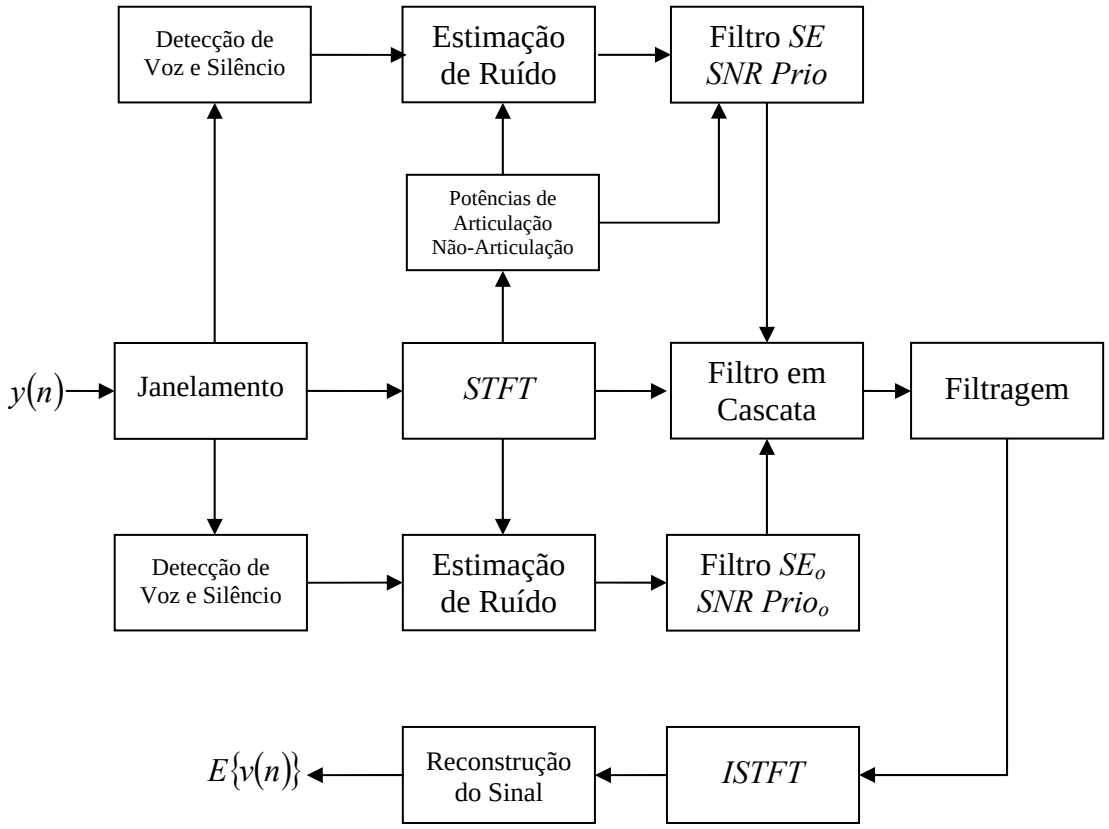


Figura 18 – Diagrama de blocos do sistema da técnica de redução de ruído baseado na *SNR a Priori* aprimorada com os parâmetros *Potências de Articulação* e *Não-Articulação*.

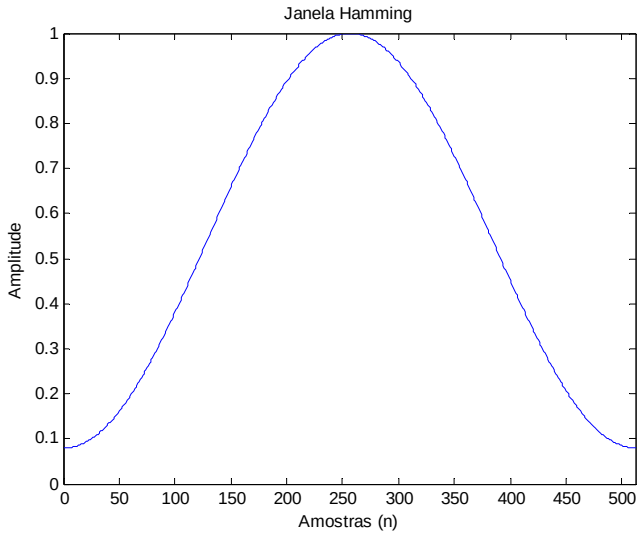


Figura 19 – Janela Hanning com largura de banda de 512 amostras.

No bloco Detecção de Voz e Silêncio (*DVS*) apresentado na Figura 18 o objetivo é determinar os intervalos de silêncio para se estimar a potência do ruído, que é suposto estacionário, fundamental para a implementação das técnicas estudadas. Porém, considerando que este bloco não faz parte do melhoramento das técnicas de *MSV* propostas neste trabalho, nas avaliações realizadas a *DSV* foi determinada de maneira ótima, ou seja, a partir da forma de onda dos sinais avaliados foram obtidas as separações entre voz e silêncio

Uma das principais alterações feitas nas técnicas estudadas foi incorporar algumas das técnicas psicoacústicas extraídas da *ANIQUE* para estimação das potências envolvidas nos sinais de voz e ruído. Isto é feito através dos novos parâmetros *Potência de Articulação* e *Potência de Não-Articulação*.

Na etapa seguinte tem-se uma outra alteração, que é a mais importante na nova proposta: a *SNR_post* e a *SNR_prio* são calculadas de acordo com as potências de *Articulação* e *Não-Articulação* para gerar um novo filtro, representado na Figura 18 pelo bloco "Filtro SE *SNR_prio".* Esse filtro é então colocado em cascata com o filtro original, gerando um novo filtro, representado no diagrama pelo bloco "Filtro em Cascata". É com este filtro que é feita a filtragem do sinal filtragem do sinal ruidoso janelado. Este processo não considera a estimação da fase. Portanto, na aplicação da transformada inversa de Fourier usa-se a fase do sinal ruidoso. Finalmente, para reconstruir o sinal estimado usa-se um método de síntese. O método mais adotado em *MSV*, quando baseado na STFT, é o Overlap Addition (*OLA*), podendo também utilizar-se um banco de filtros.

6.2 Potência de Articulação e Potência de Não-Articulação

No capítulo 4 foi apresentado o estudo da técnica *ANIQUE*, com detalhamento do bloco *"Estimação de Qualidade"*. Esse estudo permitiu a obtenção de novos parâmetros propostos, que foram denominados de *Potências de Articulação* $\Lambda_A(m,n)$ e *Potência de Não-Articulação* $\Lambda_N(m,n)$, que agora serão aplicados para os cálculos da estimação do ruído e dos parâmetros básicos *SNR_Post* e *SNR_Prio*.

6.3 Banco de Filtros de Banda Crítica e Envoltória Temporal

No bloco de estimação de qualidade, detalhado na Figura 20, o sinal de voz janelado $s(n)$ é filtrado por um banco de 23 filtros de banda crítica denominado de *gammatone,* para simular o processo da operação da cóclea (SLANEY, 1993). Cada filtro é representado por $h_k(t)$, $k = 1,2,...,N_{cb}$, onde $h_k(t)$ é a resposta ao impulso do *k*-ésimo filtro do canal e N_{cb} denota o número das bandas críticas (KATSIAMIS; DRAKAKIS, 2006). A resposta impulsiva $h_k(t)$ é dada por:

$$h_k(t) = At^{N-1}e^{-b_k t}\cos(\omega_0 t + \theta) \tag{6.1}$$

Na equação (6.1), A é uma constante usada para regular a amplitude da resposta impulsiva, $b_k = 2\pi 1{,}019 ERB_k$, ERB_k é a largura de banda retangular equivalente, N é a ordem do filtro, ω_0 é a frequência característica do filtro e θ é a fase.

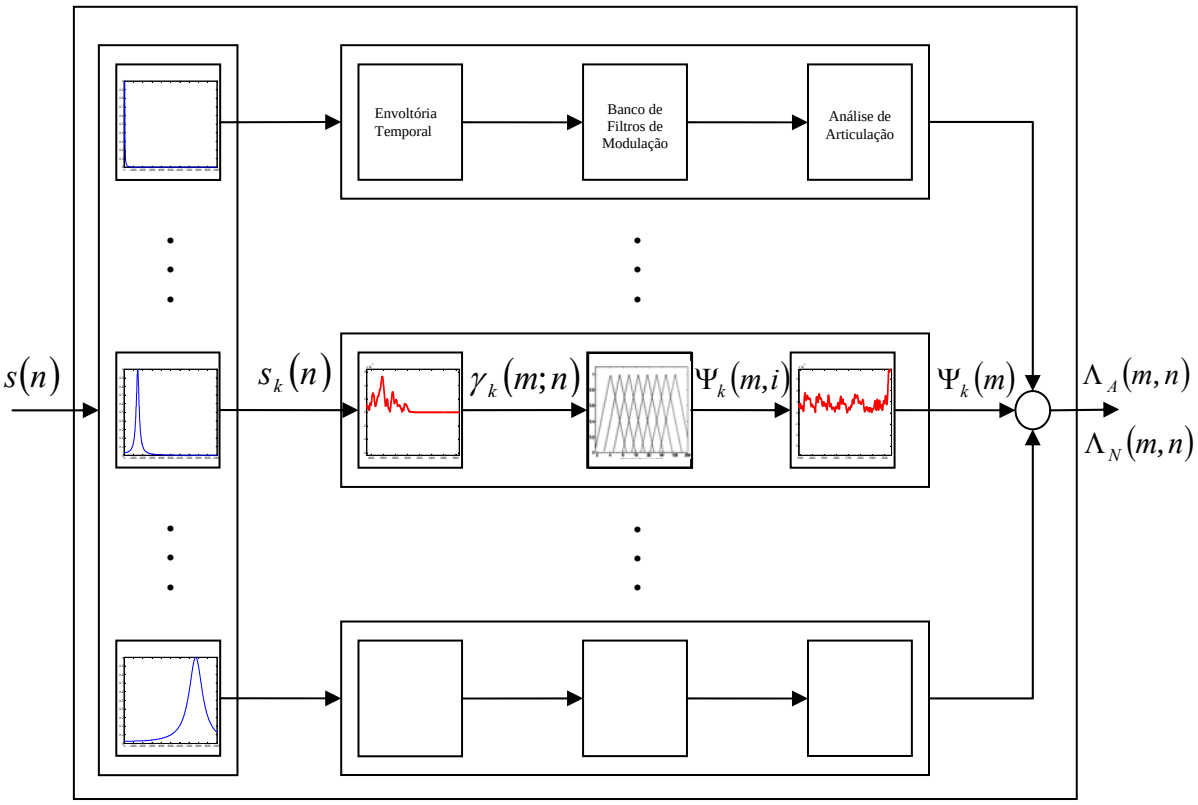


Figura 20 – Diagrama do bloco *"Estimação de Qualidade"*.

A resposta em frequência do banco de filtros cocleares pode ser obtida aplicando-se a Transformada de Laplace na equação (6.1). Assim, considerando o par de transformada de Laplace $t^{N-1}e^{at} \longleftrightarrow \Gamma(N)/(s-a)^N$, onde $\Gamma(N)$ é a função Gamma dada por $\Gamma(N)=(N-1)!$, tem-se que a resposta em frequência dos filtros será dada por:

$$H_k(s) = \frac{A\Gamma(N)}{2}\left[\frac{e^{j\theta}}{(s+b_k-j\omega_0)^N} + \frac{e^{-j\theta}}{(s+b_k+j\omega_0)^N}\right] \tag{6.2}$$

A constante A é arbitraria e na prática deve ser usado um valor tal que o ganho de pico da resposta em frequência seja unitário. A frequência ω_0 corresponde à frequência característica e os parâmetros b_k e N definem a largura de banda da resposta (KATSIAMIS; DRAKAKIS, 2006), sendo $b_k = 2\pi 1{,}019 ERB_k$.

Apenas como ilustração, nas Figuras 21 e 22 são apresentadas as respostas ao impulso e em freqüência, respectivamente, de um filtro *gammatone.*

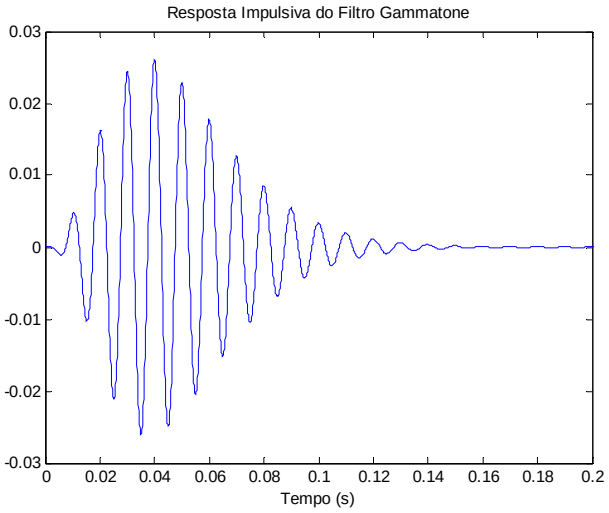


Figura 21 – Resposta ao impulso de um filtro Gammatone com $f_o = 1kHz$, $N = 4$, $ERB_k = 125Hz$, $\theta = 0°$.

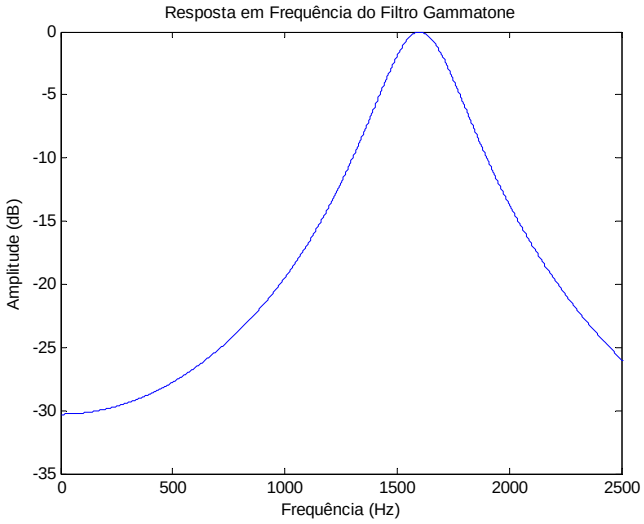


Figura 22 – Resposta em frequência de um filtro Gammatone com $f_o = 1{,}6kHz$, $N = 2$, $ERB_k = 198{,}4Hz$ e $\theta = 0°$.

A frequência característica do filtro no banco de filtro da cóclea abrange a faixa de 50 Hz a 10500 Hz, conforme a Quadro (3.2), e a largura de banda de cada filtro de banda crítica é caracterizada pela Largura de Banda Retangular Equivalente (*ERB*) (GLASBERG; MOORE, 1990).

$$ERB_k = \frac{F_k}{Q_{ear}} + B_{\min} \tag{6.3}$$

Na equação (6.3), F_k é a frequência característica do *k*-ésimo filtro de banda crítica em Hertz. De acordo com a sugestão de Glasberg e Moore (1990), os valores de Q_{ear} e $B_{\min}$ usados neste trabalho foram fixados em 9.26449 e 24.7, respectivamente. A Figura 23 mostra a resposta em frequência do banco de filtros cocleares que consiste em 23 filtros de banda crítica.

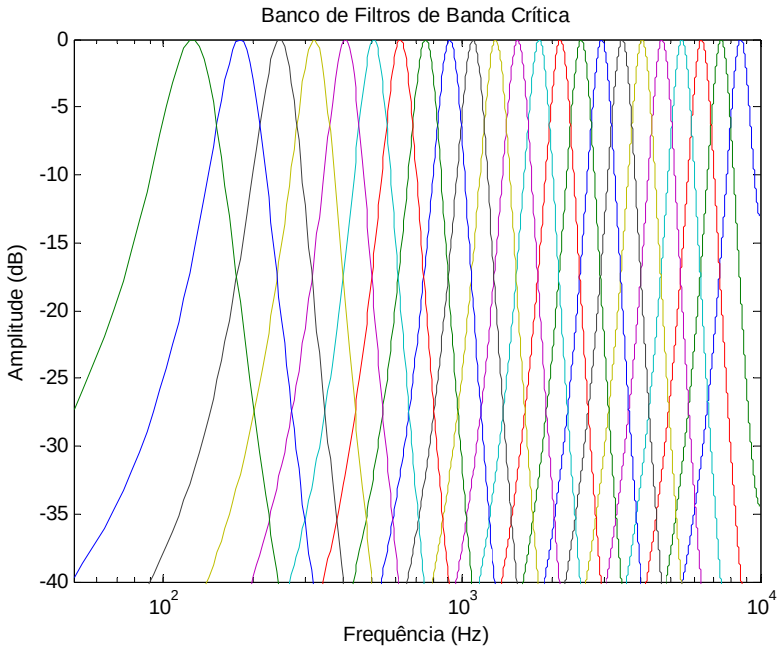


Figura 23 – Resposta em frequência do banco de filtros de banda crítica.

A filtragem do sinal no *k*-ésimo canal é representada com segue:

$$s_k(n) = s(n) * h_k(n) \tag{6.4}$$

Em geral, o sinal analítico de um sinal com banda limitada $s_k(n)$ é pode ser dado como segue:

$$z_k(n) = s_k(n) + j\,\tilde{s}_k(n) \tag{6.5}$$

onde $\tilde{s}_k(n)$ é a transformada de Hilbert de $s_k(n)$.

A envoltória de $s_k(n)$ é expressa como

$$\gamma_k(n) = \sqrt{s_k^2 + \tilde{s}_k^2(n)} \tag{6.6}$$

e a fase instantânea é dada por:

$$\phi_k(n) = \arctan \frac{\tilde{s}_k(n)}{s_k(n)} \tag{6.7}$$

A partir dessas equações pode-se expressar o sinal de banda crítica $s_k(n)$ em termos de sua envoltória e portadora, respectivamente, como segue:

$$s_k(n) = \Re\{z_k(n)\} \tag{6.8a}$$

$$s_k(n) = \gamma_k(n)\cos\phi_k(n) \tag{6.8b}$$

Na Figura 24(a) apresenta-se um trecho pequeno de um sinal de voz e na Figura 24(b) apresenta-se o sinal de envoltória equivalente. Observa-se que a envoltória mostra dois principais componentes: movimento muito lento do sistema de articulação humana, 2 ~ 30 Hz, e componentes de modulação causados pela excitação glótica em 184 Hz (pitch).

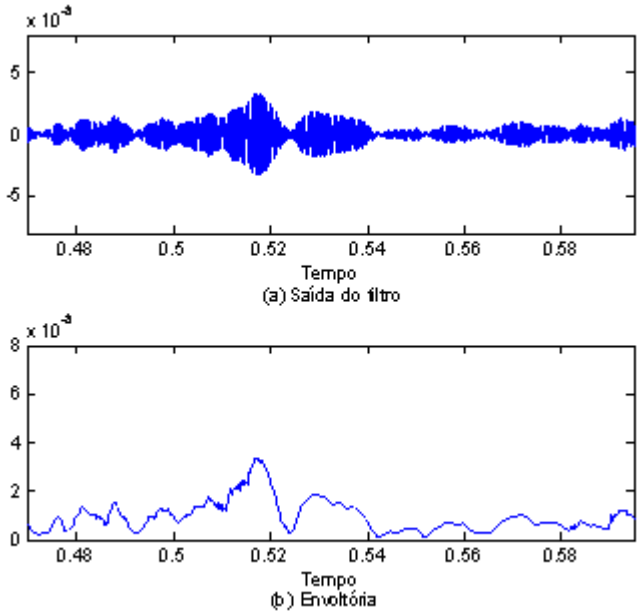


Figura 24 – Exemplo da envoltória do sinal e seu espectro de modulação: (a) 128 ms da saída do filtro de banda crítica centrado em 1600 Hz e (b) a envoltória de (a).

6.4 Banco de Filtros de Modulação

Após calcular $s_k(n)$, tem-se para cada banda crítica a envoltória $\gamma_k(n)$ que é multiplicada por 256 ms da janela de Hanning e é deslocada em 64 ms em todos os quadros, com o objetivo de obter-se $\gamma_k(m;n)$, que é a envoltória para a *k*-ésima banda crítica no *m*-ésimo quadro. Comparado com a largura típica da janela usada em processamento de sinais (na faixa 20-30 ms), uma janela relativamente longa é usada neste trabalho. Isto foi determinado empiricamente para obter uma resolução apropriada em termos de modulação em frequência e melhor desempenho do modelo. O espectro de modulação para cada banda crítica é então estimado pela transformada de Fourier como segue:

$$\Gamma_k(m,f) = \left|\Im\{\gamma_k(m;n)\}\right| \tag{6.9}$$

onde *f* representa a frequência.

O espectro de modulação é agrupado em *M* bandas pelo banco de filtros de modulação $\{W(i,f)\,|\,i=1,2,\ldots,M\}$, que são filtros passa-faixa de segunda ordem com fator de qualidade *Q=2* como proposto em (DAU; PUSCHEL; KOHLRAUSCH, 1996). As frequências características e as larguras de bandas estão na Quadro (3). Esses filtros são implementados e aplicados no domínio da modulação, obtendo-se o espectro da envoltória filtrada $\Psi_k(m,i)$ que estará sem os sons inconvenientes e perceptíveis produzidos numa taxa além da velocidade do sistema de articulação humano. A Figura 25 mostra a resposta em frequência do banco de filtros de modulação.

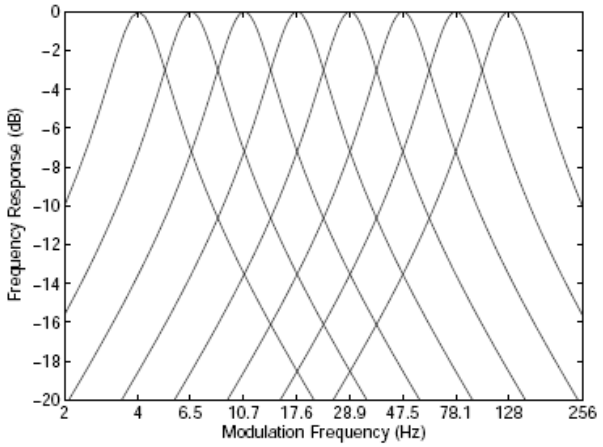


Figura 25 – Resposta em frequência do banco de filtros de modulação.

Quadro 3 – Frequências características e largura de banda dos filtros de modulação.

Índice das Bandas dos Filtros de Modulação								
	1	2	3	4	5	6	7	8
$f_0(Hz)$	4,0	6,5	10,7	17,6	28,9	47,5	78,1	128,0
$BW(Hz)$	2,4	3,9	6,5	11,0	18,2	29,1	47,6	78,8

6.5 Análise de Articulação

Após o sinal passar pelos filtros de modulação, obtém-se o espectro das envoltórias de articulação no *k*-ésimo canal cóclea que é dada pela equação:

$$\Psi_{k,A}(m) = \frac{1}{L_A}\sum_{i=1}^{L_A}\Psi_k(m,i) \tag{6.10}$$

onde $\Psi_{k,A}(m)$ é a energia média de articulação refletindo o componente relevante do sinal de voz natural humano.

O espectro das envoltórias de não-articulação no *k*-ésimo canal cóclea é dada por:

$$\Psi_{k,N}(m) = \frac{1}{L_N(k) - L_A}\sum_{i=L_A+1}^{L_N(k)}\Psi_k(m,i) \tag{6.11}$$

onde $\Psi_{k,N}(m)$ é a energia média de não articulação representando os sons inconvenientes perceptíveis produzidos numa taxa além da velocidade do sistema de articulação humano. Com objetivo de cobrir a extensão de frequência de 2-30 Hz, correspondendo à velocidade de movimento do sistema de articulação humana, L_A na equação (6.11) é assumido ser 4. Para o cálculo da energia média de não articulação $\Psi_{k,N}(m)$, a energia de banda de modulação do (L_A+1)-ésimo até a banda $L_N(k)$-ésimo é calculada como segue:

$$L_N(k) = \begin{cases} 5, & 0 \le k \le 13 \\ 6, & 14 \le k \le 18 \\ 7, & 19 \le k \le 23 \end{cases} \tag{6.12}$$

Isto significa que a maior frequência modulada na energia de não-articulação é escolhida de forma diferente para várias bandas críticas (note que L_N é uma função de k). A razão para isso é baseada na investigação realizada por Ghitza (2001). Neste experimento físico-acústico, foi mostrado que em um dado canal auditivo a largura mínima de banda da

envoltória de informação requerida para preservar a qualidade de sinais de voz é um pouco acima da metade da largura da banda crítica daquele canal. Isto significa que somente os componentes da modulação em frequência acima da metade da largura de banda crítica são relevantes para a percepção da qualidade de sinais de voz. Dessa forma, $L_N(k)$ é determinado assumindo que os canais dos filtros de modulação considerados para calcular $\Psi_{k,N}(m)$ cobrem, aproximadamente, um pouco mais da metade da largura de banda crítica.

As potências de articulação $\Lambda_A(m,n)$ e não-articulação $\Lambda_N(m,n)$, para cada quadro *m* (ou *frame*) com *n* pontos e N_{cb} bandas críticas são dadas, respectivamente, por:

$$\Lambda_A(m,n) = \left(\sum_{k=1}^{N_{cb}} \Psi_{k,A}(m,n) \right)^2 \tag{6.13}$$

$$\Lambda_N(m,n) = \left(\sum_{k=1}^{N_{cb}} \Psi_{k,N}(m,n) \right)^2 \tag{6.14}$$

Esses dois novos parâmetros serão utilizados posteriormente para os cálculos da estimação do ruído e a *SNR_Prio*.

6.6 Estimação da Potência do Ruído

A aplicação dos filtros redutores de ruído desenvolvidos neste trabalho passa por uma estimação da potência do ruído. Considerando os intervalos de silêncio definidos pela *DVS* e o ruído estacionário, pode-se estimar a potência do ruído usando uma filtragem recursiva de primeira ordem, como segue:

$$\sigma_{ro}{}^2(w_i, f_k) = \beta \cdot \sigma_{ro}{}^2(w_{i-1}, f_k) + (1-\beta) \cdot |Y(w_i, f_k)|^2 \tag{6.15}$$

$$\sigma_r{}^2(w_i, f_k) = \beta \cdot \sigma_r{}^2(w_{i-1}, f_k) + (1-\beta) \cdot |\Lambda_N(w_{i-1}, f_k)|^2 \tag{6.16}$$

onde w_i é o quadro (*frame*) *i* analisado, f_k a frequência processada, $\sigma_{ro}^{\ 2}(w_i, f_k)$ representa a estimação da potência do ruído utilizando a potência do sinal ruidoso, $|Y(w_i, f_k)|^2$, do quadro atual, $\sigma_r^{\ 2}(w_i, f_k)$ representa a estimação de ruído utilizando a potência de não articulação do sinal, $|\Lambda_N(w_{i-1}, f_k)|^2$, do quadro anterior e β é uma constante que define a dependência do ruído na janela atual de análise em relação às janelas anteriores.

Esta estimação é fundamental para um bom desempenho do filtro redutor de ruído. Por exemplo, uma variação brusca de potência de uma janela para outra pode contribuir fortemente para o surgimento do ruído musical. Nesses casos, um valor adequado do fator β passa a ser muito importante.

O cálculo de β é baseado no tamanho da janela de análise, na frequência de amostragem do sinal e na constante de tempo que garanta o esquecimento das informações passadas. Assim, tem-se que:

$$\beta = \exp\left(\frac{-L}{2 \cdot fa \cdot Tc}\right) \tag{6.17}$$

onde L é tamanho da janela, fa a frequência de amostragem e Tc o tempo associado ao fator de esquecimento. Um valor típico para Tc é 140 ms, o que equivale a várias janelas de sinal, considerando segmentos com duração variando de 16 ms a 32 ms.

6.7 Estimação da SNR a Posteriori

A relação sinal/ruído a posteriori pode ser obtida diretamente do sinal ruidoso captado e da potência do ruído estimada. Assim, considerando os parâmetros normais da subtração espectral e os novos parâmetros, obtém-se

$$\hat{SNR}\_post_o(w_i, f_k) = \frac{|Y(w_i, f_k)|^2}{\sigma_{ro}^{\ 2}(w_i, f_k)} \tag{6.18}$$

que é a *SNR_post* original, no sentido de ser baseada nos parâmetros normais dos sinais de voz e ruído, e

$$\hat{SNR}\_post\ (w_i, f_k) = \frac{|\Lambda_A(w_i, f_k)|^2}{\sigma_r{}^2(w_i, f_k)} \tag{6.19}$$

que é a nova *SNR_post*, obtida com os parâmetros psicoacústicos.

Verifica-se que a SNR_post_o e SNR_post acompanham de forma direta as variações do sinal ruidoso, o que confirma sua sensibilidade em relação às mudanças bruscas de fase do ruído.

6.8 Estimação da SNR a Priori

A estimação da relação sinal-ruído a priori depende do próprio sinal de voz estimado, o que caracteriza um sistema não-causal. Para resolver este problema, Ephraim e Malah (1984) propuseram um estimador que usa como potência do sinal de voz estimado na atual janela de análise a potência do sinal de voz estimada na janela imediatamente anterior. Isto é possível e gera resultados satisfatórios porque, em curtos intervalos de tempo, a variação de potência de uma janela para outra é desprezível.

Assim, e considerando a mesma análise feita para a *SNR_post*, a *SNR_prio* obtida com os parâmetros normais da subtração espectral é dada por:

$$\hat{SNR}\_prio_o(w_i, f_k) = \alpha \cdot \frac{|V(w_{i-1}, f_k)|^2}{\sigma_{ro}{}^2(w_i, f_k)} + (1-\alpha) \cdot T[\hat{SNR}\_post_o(w_i, f_k) - 1] \tag{6.20}$$

Já a *SNR_prio* considerando os parâmetros psicoacústicos é dada por:

$$\hat{SNR}\_prio\ (w_i, f_k) = \alpha \cdot \frac{|V(w_{i-1}, f_k)|^2}{\sigma_r{}^2(w_i, f_k)} + (1-\alpha) \cdot T[\hat{SNR}\_post\ (w_i, f_k) - 1] \tag{6.21}$$

O operador $T[]$ indica uma transformação sobre $[\hat{SNR}\_post\ (w_i, f_k) - 1])$ e α representa o grau de dependência da SNR_prio com relação a SNR_post. Lembrando-se que $SNR\_post(w_i, f_k) = SNR\_prio(w_i, f_k) + 1$, o uso da SNR_post na estimação da SNR_prio representa uma contribuição da SNR_prio na atual janela de análise, mas que deverá passar por uma transformação, visto que SNR_post pode ser menor que 1. Normalmente usa-se a retificação de meia-onda como transformação, como proposto por Boll (1979). Se for considerado apenas a SNR_post, isto significa uma potência nula para a voz naquela frequência, o que não é desejável. Assim, propõe-se nesses casos a utilização de um valor mínimo para SNR_post, aqui representado como δ, de forma que

$$se \quad \hat{SNR}\_post_o(w_i, f_k) - 1 \le \delta, \quad então \quad \hat{SNR}\_post_o(w_i, f_k) = \delta \tag{6.22}$$

Voltando à equação (6.22), tem-se que a contribuição da SNR_post na SNR_prio depende do valor de α. Nos estudos realizados, como também no trabalho original de Ephraim e Malah (1984), verificou-se que as fortes variações na SNR_post podem afetar o desempenho da SNR_prio. Entretanto, se for definido α=0, perde-se o ajuste fino na estimação da SNR_prio. Assim, o valor típico adotado para α varia entre 0.9 e 1. Normalmente é usado o valor de α=0.98.

6.9 Procedimento de Filtragem

Aplicando as equações (6.20) e (6.21) na resposta em frequência do filtro da *subtração espectral*, dada pela equação (5.29), têm-se as seguintes expressões:

$$|\hat{H_{So}}(\omega)| = \sqrt{\frac{\hat{SNR}\_prio_o(\omega)}{1 + \hat{SNR}\_prio_o(\omega)}} \tag{6.23}$$

$$|\hat{H_S}(\omega)| = \sqrt{\frac{\hat{SNR}\_prio\ (\omega)}{1+\hat{SNR}\_prio\ (\omega)}} \tag{6.24}$$

A proposta deste trabalho é criar um filtro baseado nos dois filtros calculados previamente. Assim, o novo filtro de redução de ruído é uma combinação dos dois filtros em cascatas, de modo que a nova resposta em frequência é dada por:

$$|\hat{H_{CS}}(\omega)| = \left[\left(|\hat{H_S}(\omega)|\right)^{An} . |\hat{H_{So}}(\omega)|\right]^{Ao} \tag{6.25}$$

O parâmetros $A_{\mathbf{n}}$ e A_o representam o grau de dependência da *subtração espectral* original com o filtro baseado nos critérios psicoacústicos obtidos a partir da *ANIQUE*. Testes experimentais exaustivos, considerando diferentes sinais ruidosos, mostraram que bons resultados podem ser obtidos quando $A_{\mathbf{n}}$ =0,1 e A_o =0,45. Naturalmente, se for considerada uma aplicação em tempo real, estes valores afetam significativamente o tempo de resposta do filtro.

Para ilustrar as características do novo filtro, na Figura 26 mostra-se um pequeno trecho dos filtros *subtração espectral* baseada na *SNR_prio* (SE+SNR_Prio, linha vermelha), *subtração espectral* baseada na *SNR_prio* com os novos parâmetros de *Potência de Articulação* (SE+SNR_Prio+P.Art, linha preta) e o *filtro proposto* (linha azul). Observa-se que o filtro obtido com os novos parâmetros *Potência de Articulação* e *Não-Articulação* que foram utilizados para o cálculo do filtro $H_{So}(\omega)$ proporciona uma correção nas amplitudes das frequências do filtro $H_S(\omega)$, que é baseado na *subtração espectral*. Portanto, o novo filtro em cascata $H_{CS}(\omega)$ permite uma filtragem que incorpora princípios psicoacústicos e melhora a qualidade do sinal filtrado

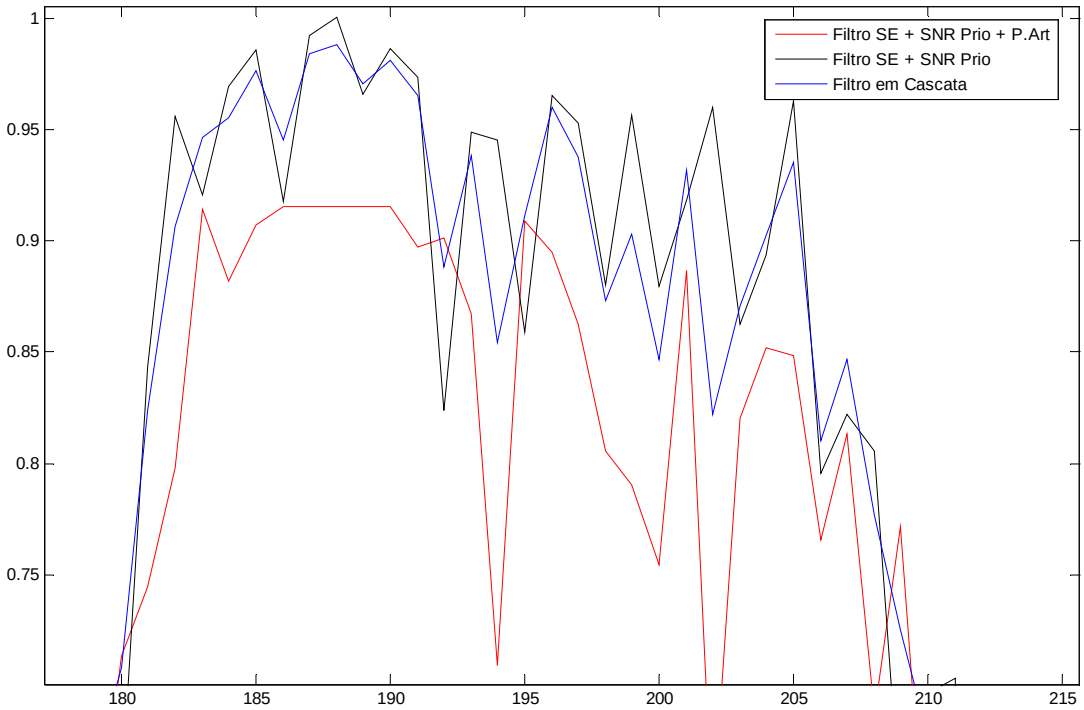


Figura 26 – Pequeno trecho das respostas em frequência dos filtros separadamente.

Após a filtragem do sinal com o *filtro em cascata*, é feito o retorno para o domínio do tempo através da transformada de Fourier utilizando inversa, onde a fase usada da reconstrução é a mesma fase do sinal ruidoso. O método de síntese "*Overlap Addition*" (*OLA*) permite uma reconstrução do sinal através de adições dos pequenos trechos do sinal janelado e filtrado.

CAPÍTULO 7

Simulações e Resultados

Neste capítulo são apresentados os resultados obtidos com as simulações das aplicações das técnicas *subtração espectral* baseados na relação *SNR a Priori* com a implementação dos novos parâmetros de *Articulação*. São apresentadas avaliações objetivas comparativas dos sinais processados usando a medida de avaliação objetiva *PESQ* (*Perceptual Evaluation of Speech Quality*). A *PESQ* é uma medida de qualidade de voz que pode ser classificada como intrusiva e foi padronizada pela União Internacional de Telecomunicações (ITU-Rec P.862, 2001). Os resultados da PESQ são dados em valores que representam notas e variam entre 0 (mais baixa qualidade) e 4,5 (mais alta qualidade).

O sistema de redução de ruído modificado foi implementado usando-se o software de simulação *MatLab*® . Todos os parâmetros necessários foram fixados de acordo com os valores apresentados nos Capítulos 4, 5 e 6.

Foram utilizados um sinal em português com as relações sinal/ruído de 0 dB, 5 dB, 10 dB e 15 dB, amostrados a uma taxa de 16 kHz, e quatro sinais em inglês com as mesmas relações sinal/ruído amostrados a uma taxa de 8 kHz. O ruído de fundo utilizado foi de um carro pequeno (Renault 25) com velocidade de 100 km/h (VIEIRA FILHO, 1996).

Todos os sinais foram processados com a mesma configuração para a STFT, ou seja, janelas de Hanning com intervalos de sobreposição de 50 %, 512 amostras por janela e FFT (Fast Fourier Transform) de 512 para os sinais em português e 256 amostras para os sinais em inglês.

Para possibilitar uma comparação, os sinais foram processados pelas diferentes técnicas de redução de ruído discutidas neste trabalho. Os resultados obtidos nas simulações foram avaliados em termos de inteligibilidade, nível de ruído musical, nível de redução de ruído e a avaliação obtida pelo *PESQ*.

7.1 Sinais Utilizados nas Simulações

O objetivo dos testes foi identificar o quanto a nova metodologia proposta foi melhor ou pior em relação à metodologia original e também em relação à técnica de redução de ruído proposta por Ephraim e Malah. Assim, entende-se que não é necessário um procedimento rigoroso baseado em banco de sinais ruidosos para identificar essas diferenças. Sendo assim, considera-se que os sinais apresentados a seguir são suficientes para se alcançar os objetivos. O procedimento básico adotado foi a utilização de sinais livres da presença de ruído que foram corrompidos adicionando-se um ruído com diferentes níveis. Isto permitiu a obtenção de diferentes níveis de relação sinal/ruído.

Os sinais usados nas simulações foram:

Ruído

Ruído constante de um carro pequeno (Renault 25) com velocidade de 100km/h.

Sinal 1 – voz masculina

Frase: “A bolsa ficara estável ou sofrerá uma pequena queda”.

Sinal 2 – voz masculina

Frase: “Her purse was full of useless trash”.

Sinal 3 – voz feminina

Frase: “Hedge apples may stain your hands green”.

Sinal 4: voz feminina

Frase: “The set of china hit the floor with a crash”.

Sinal 5: voz: masculina

Frase: “The club rented the rink for the fifth night”.

Apenas para dar uma idéia dos sinais usados, na Figura 27 apresenta-se a forma de onda do sinal 1 sem a adição de ruído.

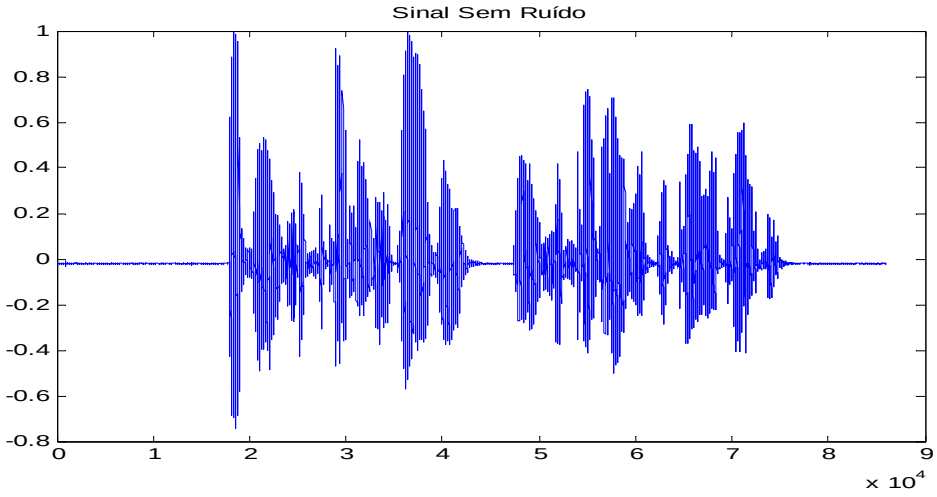


Figura 27 – Forma de onda do sinal 1 sem adição de ruído.

As Figuras 28, 29, 30 e 31 ilustram, respectivamente, as formas de onda desse mesmo sinal com *SNR* = 0dB, 5dB, 10dB e 15dB.

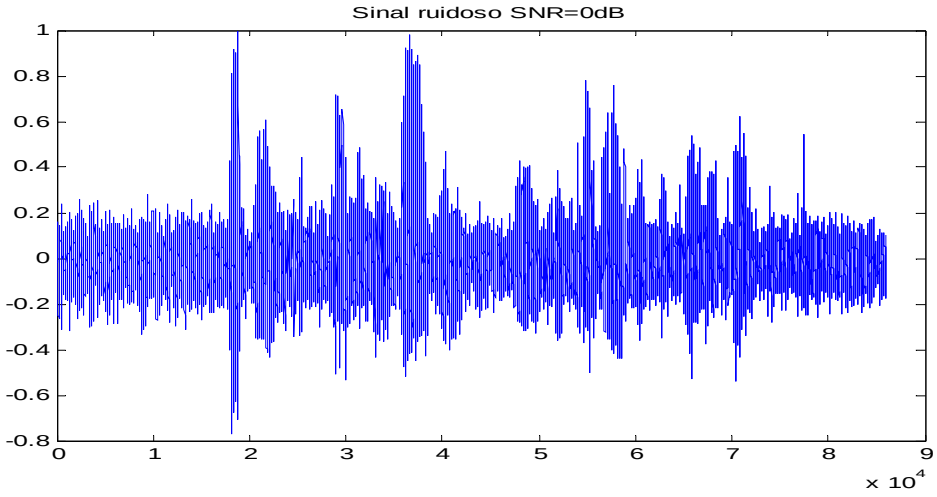


Figura 28 – Forma de onda do sinal 1com *SNR* = 0dB.

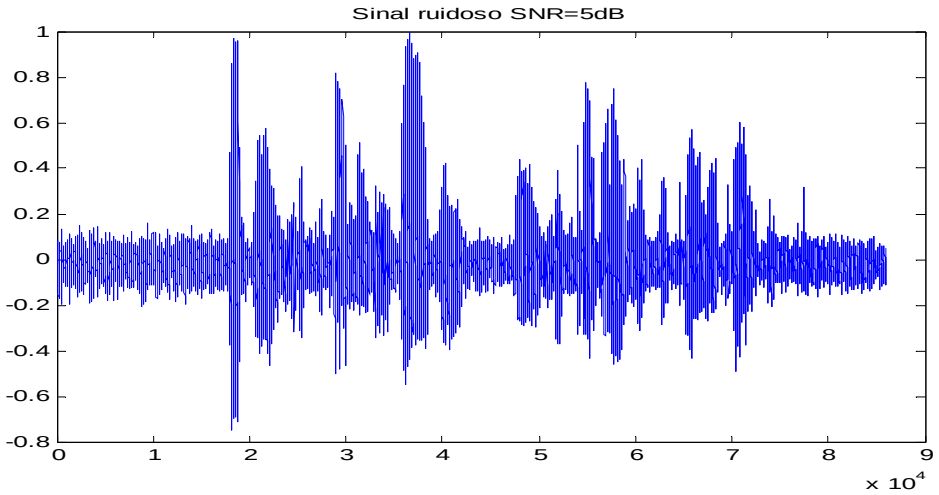


Figura 29 – Forma de onda do sinal 1 com *SNR* = 5dB.

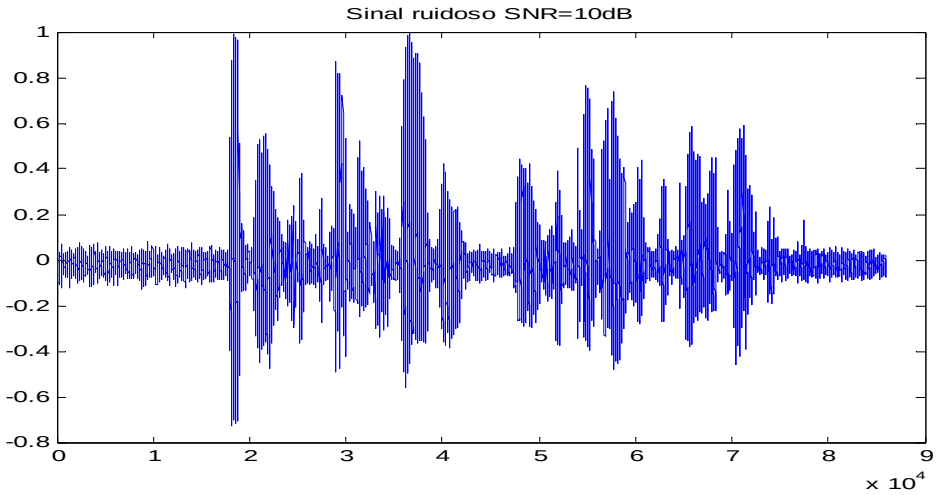


Figura 30 – Forma de onda do sinal 1 com *SNR* = 10dB.

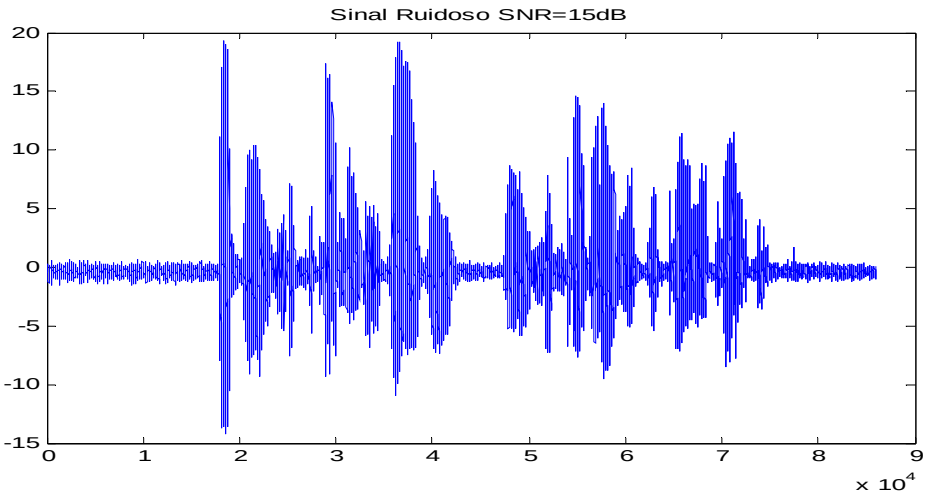


Figura 31 – Forma de onda do sinal 1 com *SNR* = 15dB.

7.2 Sinais Processados

Todos os sinais foram processados utilizando a *subtração espectral* baseada na *SNR a Priori* e nos parâmetros de *Articulação*, com avaliação posterior usando-se a *PESQ*. As Figuras 32 e 33 ilustram as formas de onda do sinal 1 com *SNR* = 10dB e seu respectivo sinal processado, obtendo-se uma nota PESQ 2,191 pontos.

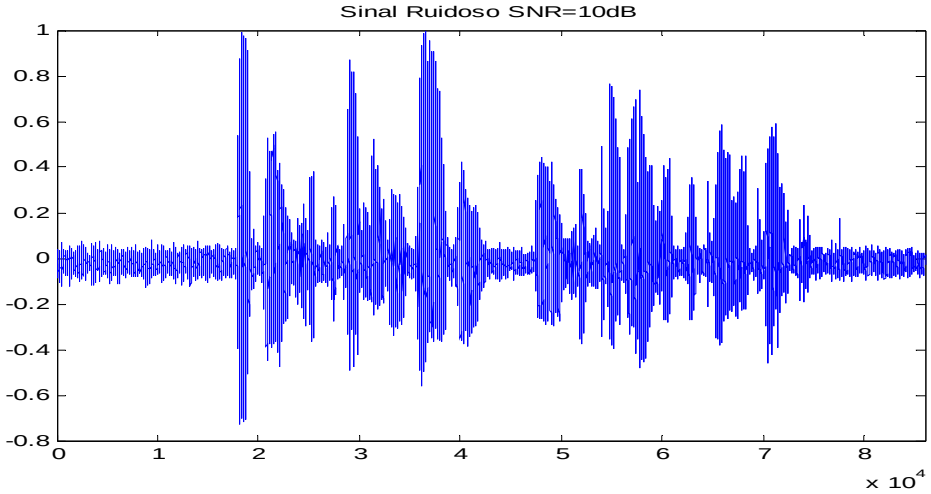


Figura 32 – Forma de onda do sinal 1 com a relação *SNR*=10dB.

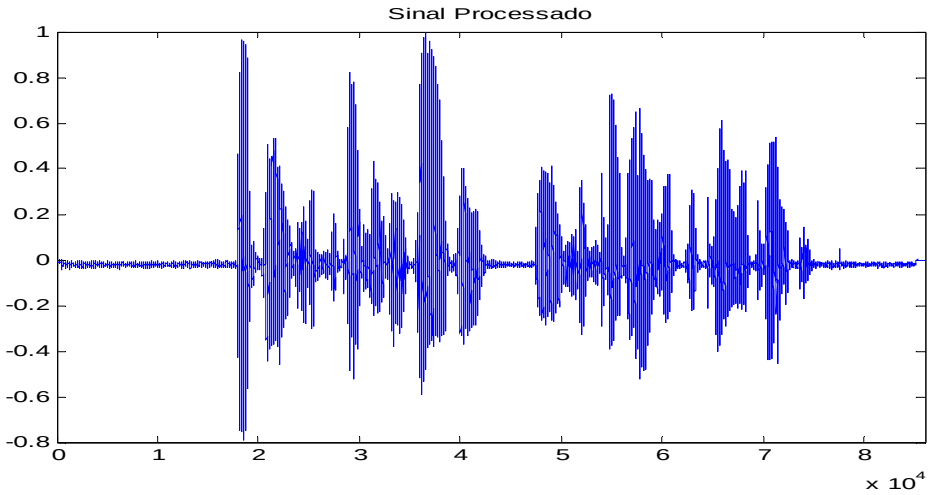


Figura 33 – Forma de onda do sinal 1 com a relação *SNR*=10dB processado utilizando a *subtração espectral* baseado na relação *SNR Prio* com os parâmetros de *Articulação*.

Nas Quadros (4) à (7) são apresentados os resultados das avaliações objetivas de todos os sinais processados e seus respectivos gráficos, onde *SE+SNR_Prio+P.Art* é a técnicas *subtração espectral* baseado na *SNR a Priori* com os parâmetros de *Articulação*, *SE+SNR_Prio* é a *subtração espectral* baseado na *SNR a Priori*, *MMSE+SNR Prio* é a *minimização do erro quadrático médio* baseado na relação *SNR a Priori* e o *Sinal Ruidoso* é o sinal sem qualquer filtragem.

Quadro 4 – Avaliações objetivas dos sinais processados com SNR=0dB.

SNR 0dB	Sinal 1	Sinal 2	Sinal 3	Sinal 4	Sinal 5
SE+SNR_Prio+P.Art	1,989	1,712	1,823	2,063	2,094
Sinal Ruidoso	1,517	1,396	1,868	1,599	1,628
SE+SNR_Prio	1,571	1,564	1,893	1,862	1,884
MMSE+SNR_Prio	1,684	1,604	1,927	2,015	2,023

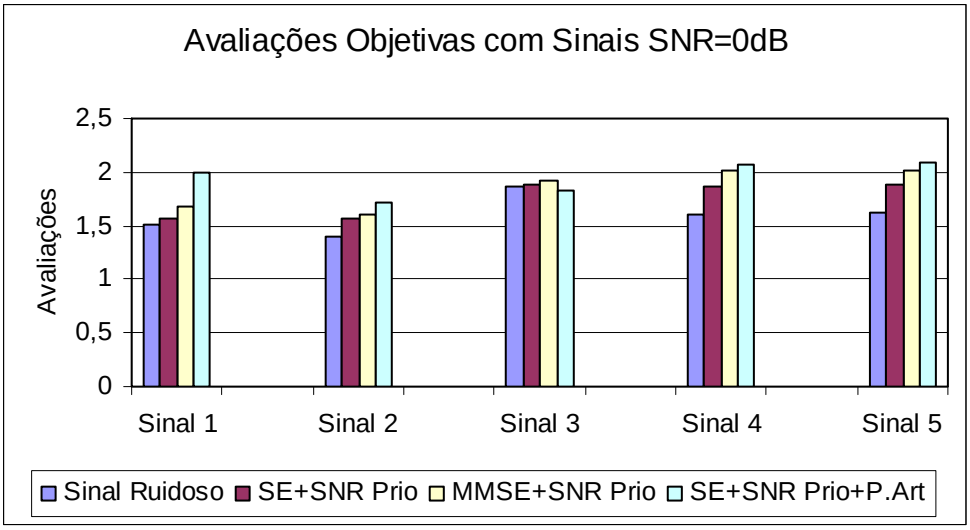


Figura 34 – Avaliações objetivas dos sinais processados com SNR=0dB.

Quadro 5 – Avaliações objetivas dos sinais processados com SNR=5dB.

SNR 5dB	Sinal 1	Sinal 2	Sinal 3	Sinal 4	Sinal 5
SE+SNR_Prio+P.Art	1,887	2,235	2,176	2,333	2,457
Sinal Ruidoso	1,577	1,829	2,017	1,918	1,913
SE+SNR_Prio	1,686	1,868	2,096	1,979	2,052
MMSE+SNR_Prio	1,745	2,125	2,143	2,123	2,191

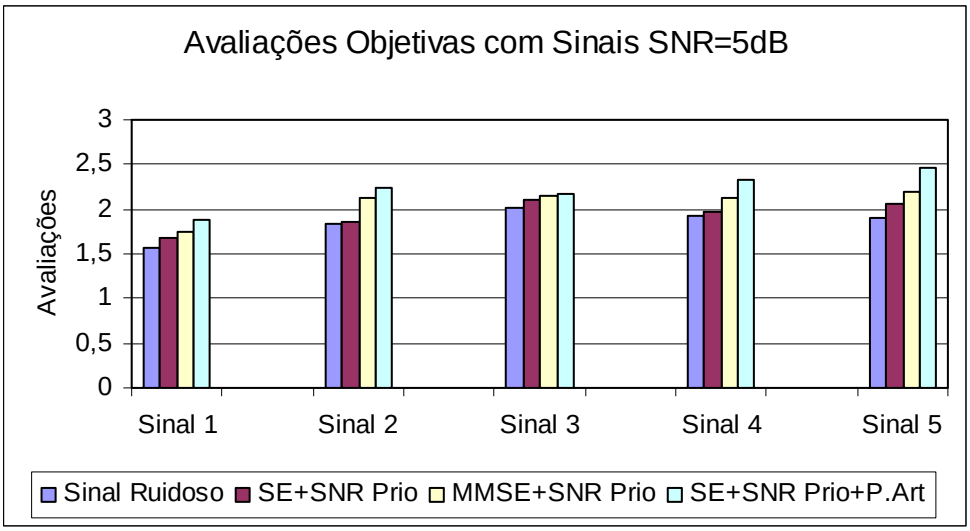


Figura 35 – Avaliações objetivas dos sinais processados com SNR=5dB.

Quadro 6 – Avaliações objetivas dos sinais processados com SNR=10dB.

SNR 10dB	Sinal 1	Sinal 2	Sinal 3	Sinal 4	Sinal 5
SE+SNR_Prio+P.Art	2,191	2,570	2,429	2,670	2,589
Sinal Ruidoso	1,742	2,169	2,211	2,243	2,227
SE+SNR_Prio	1,989	2,212	2,268	2,325	2,311
MMSE+SNR_Prio	2,056	2,371	2,321	2,481	2,504

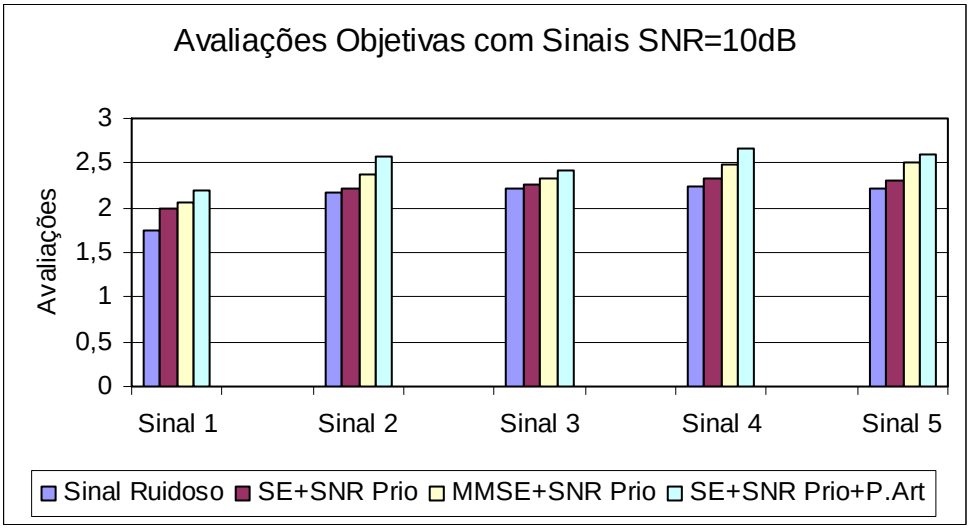


Figura 36 – Avaliações objetivas dos sinais processados com SNR=10dB.

Quadro 7 – Avaliações objetivas dos sinais processados com SNR=15dB.

SNR 15dB	Sinal 1	Sinal 2	Sinal 3	Sinal 4	Sinal 5
SE+SNR_Prio+P.Art	2,715	2,865	2,609	2,853	2,886
Sinal Ruidoso	1,945	2,511	2,504	2,500	2,268
SE+SNR_Prio	2,157	2,614	2,564	2,591	2,402
MMSE+SNR_Prio	2,369	2,671	2,628	2,698	2,548

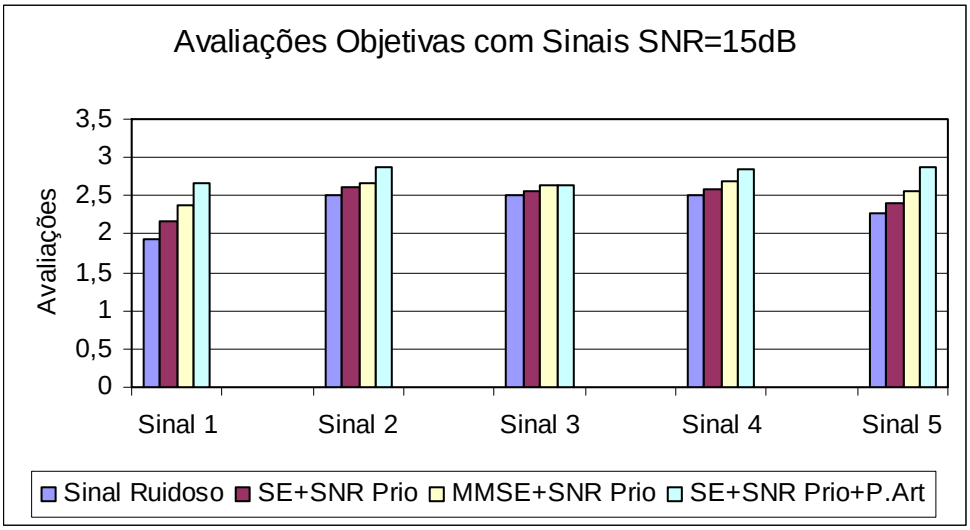


Figura 37 – Avaliações objetivas dos sinais processados com SNR=15dB.

Analisando os dados obtidos nas avaliações objetivas das Quadros (4) à (7), pode-se observar que os sinais processados com *SE+SNR_Prio+P.Art* tiveram uma nota de até 32,41% superior ao sinal *ruidoso*, indicando que houve o melhoramento do sinal processado. Observa-se também que houve o aprimoramento de até 25,05% nas avaliações comparados com os sinais processados com *SE+SNR_Prio+P.Art* e os sinais processados com *SE+SNR_Prio*. Além disso, comparando-se os sinais processados com *SE+SNR_Prio+P.Art* e os sinais processados com *MMSE+SNR_Prio* obteve-se avaliações de até 17,69% superiores. Foram avaliados e comparados os sinais processados pelo *MMSE+SNR_Prio*, pois de todas as técnicas de redução de ruído baseados na relação *SNR a Priori* a *MMSE* já foi demonstrado que esta é a técnica que gera melhores resultados (VIEIRA FILHO, 1996).

Em relação à audição dos sinais processados pela *SE+SNR_Prio+P.Art*, todos os sinais com SNR=0dB apresentaram uma pequena perda de inteligibilidade, ruído musical quase imperceptível e ótima atenuação de ruído de fundo. Os sinais com 5dB, 10dB e 15dB apresentaram ótima inteligibilidade, ruído musical imperceptível e ótima atenuação de ruído de fundo.

CAPÍTULO 8

Conclusões

A proposta deste trabalho foi melhorar o desempenho da técnica *subtração espectral* baseado *SNR a Priori* através da implementação dos parâmetros de *Articulação*, extraídas de técnicas psicoacústicas. Através dos estudos de algumas técnicas de redução de ruído em sinais de voz, foi possível observar que pode-se obter sinais processados eficientemente, mas com técnicas complexas em termos computacional como, por exemplo, a *minimização do erro quadrático médio (MMSE,)* ou técnicas simples computacionalmente, mas que não apresentam resultados satisfatórios, como é o caso da *subtração espectral* clássica, que deixa no sinal processado um ruído residual muito incomodo denominado de *ruído musical*.

Um estudo da relação Sinal/Ruído mostrou que o ruído musical presente nos sinais processados com a subtração espectral é originado da própria definição, pois uma análise mais detalhada mostra que *subtração espectral* é função direta da *SNR_Post*. Nos estudos realizados por Vieira Filho (1996) verificou-se que este parâmetro não apresenta uma boa performance quando a relação sinal/ruído é baixa (menor que 10dB). No entanto, definindo a *SNR_Post* em função da *SNR_Prio* diminui-se significativamente o inconveniente do ruído musical nos sinais processados com relação sinal/ruído menores que 10dB.

Para melhorar a performance da *subtração espectral* baseada na S*NR a Priori*, foi proposta a implementação dos parâmetros de *Articulação* obtidas através de algumas técnicas extraídas da *ANIQUE*, que é uma técnica não intrusiva de avaliação objetiva de qualidade de voz. Para realizar chegar a resultados satisfatórios, foi necessário estudar a técnica *ANIQUE* para identificar e extrair parâmetros que pudessem ser explorados na subtração espectral já baseada na *SNR_Pri.* Como a técnica *ANIQUE* tem como princípio o sistema de audição humana, foi necessário o estudo de todo o sistema auditivo humano para entender sua estrutura física, limitações e seus fenômenos auditivos, tais como mascaramento auditivo, bandas críticas e limiar de audibilidade.

Um dos novos parâmetros obtidos através da *ANIQUE* foi a *Potência de Articulação* do sinal analisado, que é a energia do sinal que o sistema de audição humana é capaz de ouvir e produzir através de seu sistema de articulações para reprodução da voz. O outro parâmetro obtido foi a *Potência de Não-Articulação* do sinal analisado que, ao contrário do parâmetro anterior, representa a energia que é perceptível ao sistema de audição humano, mas não pode ser reproduzido através do sistema de reprodução da voz humana.

A utilização dos parâmetros *Potência de Articulação* e *Não-Articulação* nos cálculos do filtro de subtração espectral $H_{So}(\omega)$, permitiu uma correção nas amplitudes das frequências do filtro $H_S(\omega)$ que é baseado na *subtração espectral*. Portanto, o novo filtro em cascata $H_{CS}(\omega)$ permite uma filtragem de ruídos mais apuradas que a *subtração espectral* baseada na relação *SNR a Priori*. Isto vem do fato que os parâmetros extraídos da *ANIQUE* utilizam modelos de baixa frequência, como a envoltória do sinal, e os filtros de modulação, onde se permitiu corrigir variações inadequadas de alterações de frequência causadas pelo filtro baseado na *subtração espectral*.

As avaliações objetivas apresentadas no capítulo 7 foram realizadas através do software *PESQ*, um modelo de avaliação intrusiva de sinais de voz. Os resultados obtidos através destas avaliações objetivas mostram que foi possível obter o aprimoramento da técnica *subtração espectral* através da implementação de técnicas psicoacústicas. As avaliações dos sinais processados com a *SE+SNR_Prio+P.Art* mostraram-se em até 32,41% superiores quando comparados com a avaliação do sinal sem qualquer processamento. Também foram obtidas avaliações 25,05% superiores em relação às avaliações que utilizaram as técnicas de *SE+SNR_Prio* e 17,69% para as avaliações em relação à *MMSE+SNR_Prio*.

Portanto, com os resultados apresentados neste trabalho pode-se afirmar que foi possível melhorar de modo satisfatório o desempenho da técnica *subtração espectral* baseada na *SNR_Prio* implementando os novos parâmetros de *Potência de Articulação* e *Potência de Não-Articulação*.

Referências

BACKUS, J. **The acoustical foundation of music.** New York: W.W. Norton, 1969.

BEERENDS, J. G.; STEMERDINK, J. A. A perceptual speech-quality measure based on psychoacoustic sound representation. **J. Audio Eng. Soc.**, New York, v. 42, n. 3, p. 115–123, 1994.

BERNE, R. M.; LEVY, M. N. (Eds.). **Fisiologia.** Rio de Janeiro: Guanabara-Koogan, 2000. p. 148-169.

BERANEK, L. L. **Acústica.** Buenos Aires: Editorial HASA, 1961.

BOLL S. F. Suppression of acoustic noise in speech using spectral subtraction. **IEEE Trans. Acoust. Speech Signal Process.,** New York, v. 27, p. 113-120, April 1979.

CAPPÉ, O. Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor. **IEEE Trans. Acoust. Speech Signal Process.,** New York, April 1994.

CAVE, C. R. **Perceptual modeling for low-rate audio coding.** 2002. 86 f. Thesis (M. Eng.) – Departament of Electrical and Computer Engineering, McGill University, Montreal, 2002.

DAU, T.; KOLLMEIER, B.; KOHLRAUSCH, A. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. **J. Acoust. Soc. Amer.**, New York, v. 102, p. 2892–2905, 1997a.

DAU, T., KOLLMEIER, B.; KOHLRAUSCH, A., Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. **J. Acoust. Soc. Amer.**, New York, v. 102, p. 2906–2919, 1997b.

DAU, T.; PUSCHEL, D.; KOHLRAUSCH, A. A quantitative model of the effective signal processing in the auditory system. I - model structure. **J. Acoust. Soc. Amer.**, New York, v. 99, n. 6, p. 3615-3622, 1996.

DRULLMAN, R.; FESTEN, J. M.;PLOMP, R. Effect of temporal envelope smearing on speech reception. **J. Acoust. Soc. Amer.**, New York, v. 95, p. 1053–1064, 1994.

EGAN, J.P.; WIENER, F. M. On the intelligibility of Bands of Speech in Noise. **J. Acoust. Soc. Amer.**, New York, v. 18, n. 2, Oct. 1946.

EPHRAIM, Y. Statistical-Model-Based Speech Enhancement Systems. **Proc. IEEE**, New York, v. 80, n. 10, p. 1526-1555, Oct. 1992.

EPHRAIM, Y.; MALAH, D. Speech enhancement using minimum mean square error short-time spectral amplitude estimator. **IEEE Trans. Acoust. Speech Signal Process.,** New York, v. 32, n. 6, Dec. 1984.

FLANAGAN, J. L. **Speech analysis synthesis and perception.** Berlin: Springer-Verlag, 1972.

GHITZA, O. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. **J. Acoust. Soc. Amer.**, New York, v. 110, n. 3, p. 1628–1640, Sep 2001.

GIGUERE, C.; WOODLAND, P. A computation model of the auditory periphery for speech and hearing science. **J. Acoust. Soc. Amer.**, New York, v. 101, p. 679-688, Mar. 1982.

GIRAUD, A.; LORENZI, C.; ASHBURNER, J.; WABLE, J.; JOHNSUDE, I.; FRACKOWIAK, R.; KLEINSCHMIDT, A. Representation of the temporal envelope of sounds in the human brain. **J. Physiol.**, Cambridge, p. 1588–1598, 2000.

GLASBERG, B. R.; MOORE, B. R. Derivation of auditory filter shapes from notched-noise data. **Hearing Res.**, Amsterdam, v. 47, p. 103–108, 1990.

HARTMANN, W. M. Pitch, periodicity, and auditory organization. **J. Acoust. Soc. Amer.,** New York, v. 100, p. 3491-3502, 1996.

IDSON, W. L.; MASSARO, D.W. A bidimensional model of pitch in the recognition of melodies. **Perception and Psychophysics**, Austin, v. 24, p. 551-565, 1978.

ITU-T Rec. P.800. **Methods for Objective and Subjective Assessment of Quality.** Place des Nations, Geneva, 1996.

ITU-T Rec. P.861. **Objective Quality Measurement of Telephone-Band (300–3400 Hz) Speech Codecs.** Place des Nations, Geneva, 1996.

ITU-T Rec. P.862. **Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs.** Place des Nations, Geneva, 2001.

ITU-T Rec. P.830. **Subjective performance assessment of telephoneband and wideband digital codecs.** Place des Nations, Geneva, 1996.

KATSIAMIS, A. G.; DRAKAKIS, E. M. Introducing the Differentiated All-Pole and One-Zero Gammatone Filter Responses and their Analog VLSI Log-Domain Implamentation. **Imperial College London,** London, v. 1, p. 561–565, 2006.

KIM, D. S. ANIQUE an auditory model for single-ended speech quality estimation, **IEEE**, New York, v. 13, n. 5, p. 821-831, 2005.

KIM, D. S.;TARRAF, A. Perceptual model for non-intrusive speech quality assessment. In: INTERNATIONAL CONFERENCE ON ACOUSTIC SPEECH, SGINAL PROCESSING, 2004, Montreal, QC, Canada. **Proceedings…** Montreal: [s.n.], 2004. p. 1060–1063.

KIM, D. S. A cue for objective speech quality estimation in temporal envelope representations. **IEEE Signal Processing Lett.**, New York, v. 11, n. 10, p. 849–852, Oct. 2004.

LEITE, S. B. **Melhoria do codificador de fala G.722.1 através do uso de um modelo perceptual.** 2003. 94 f. Dissertação (Mestrado) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2003.

MCAULAY, J. R.; MALPASS, M. Speech enhancement using a soft-decision noise suppression filter. **IEEE Trans. Acoust**., **Speech Signal Process.,** New York, v. 28, n. 2, April 1980.

MOORE, B. C. J. **An introduction to the psychology of hearing.** San Diego: Academic Press, 1997.

MOORE, B. C. J. **Hearing**. Sam Diego: Academic Press, 1995.

OPPENHEIM, L. R.; SCHAFER, R. W. **Discrete-time signal processing.** Uper Saddle River: Prentice-Hall, 1989.

PAINTER, T.; SPANIAS, A. Perceptual coding of digital audio, **Proceedings of the IEEE**, New York, v. 88, n. 4, p. 451-513, 2000.

PELLOM, B. L; HANSEN, J. H. L. An improved (Auto:I, LSP:T) constrained iterative speech enhancement for colored noise environments. **IEEE Trans. Acoust., Speech Signal Process.**, New York, v. 6, p. 573-579, Nov. 1998.

POHLMANN, K. **Principles of digital audio.** New York: McGraw Hill, 1995.

RABINER, L. R.; SCHAFER, R. W. Digital processing of speech signals. Uper Saddle River: Prentice-Hall, 1988.

SCHROEDER, M.; ATAL, B. S.; HALL J. L Optimizing digital speech coders by exploiting masking properties of the human ear. **J. Acoust. Soc. Amer.**, New York, p. 1647-1652, 1979.

SLANEY, M. **An efficient implementation of the Patterson-Holdsworth auditory filterbank, Apple Computer.** Buenos Aires: Perception Group, Tech. Rep., 1993.

TERHARDT E. Akustische Kommunikation - Grundlagen mit Hörbeispielen. Berlin: Springer, 1998.

VIEIRA FILHO, J. **Redução de ruído em sinais de voz nos sistemas rádio móveis veiculares.** 1996. 113 f. Tese (Doutorado) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 1996.

VIEMEISTER, N. F. Temporal modulation transfer functions based upon modulation thresholds. **J. Acoust. Soc. Amer.**, New York, v. 66, p. 1364–1380, 1997.

VORAN, S. Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique. **IEEE Trans. Speech Audio Process.**, New York, v. 7, n.4, p. 371–382, 1999.

ZWICKER, E.; FASTL, H. **Psychoacoustics:** facts and models. 2. ed. Berlin: Springer-Verlag, 1999.

Apêndice A

Definições Complementares

A.1 Pitch

O termo pitch tem sido usado com dois sentidos diferentes: na área de processamento de voz, o termo é freqüentemente utilizado para designar a freqüência de oscilação da glote (vibração das cordas vocais); e em psicoacústica, é usado como um atributo da sensação auditiva, segundo a definição encontrada na ANSI (American National Standards Institute), a qual estabelece que pitch é o atributo auditivo do som, de acordo com o qual os sons podem ser ordenados, em uma escala de freqüência, de baixo a alto. Este é o sentido adotado neste trabalho. Os estudos da percepção humana do pitch são complexos. Maiores informações podem ser encontradas em (BERANECK, 1961; HARTMANN, 1996; IDSON, 1978; TEHARDT, 1998).

A.2 SPL

Definição de SPL – Sound Pressure Level (Nível de Pressão Sonora) (Beraneck, 1961): tem como unidade o decibel SPL (dB_{SPL}), e é dado pela expressão $20\log (P/P_{ref})$, onde P é a pressão sonora do sinal que se está medindo e P_{ref} é a pressão sonora de referência, a qual pode assumir dois valores:

a) $P_{ref} = 0{,}0002 \mu B \quad \left(2 \times 10^5 \ N/m^2\right)$

b) $P_{ref} = 0{,}1 \mu B \quad \left(0{,}1 \ N/m^2\right)$

onde μB é a pressão em microbars.

É importante observar que a pressão sonora de referência dada no item (a) é mais utilizada nas medições relacionadas com a audição e nas medições de nível sonoro, o ar e nos líquidos, enquanto que aquela dada no item (b) tem maior aplicação na calibração de transdutores e certos tipos de medição de nível sonoro em líquidos. Os dois níveis de medição diferem um do outro em aproximadamente 74 dB. Por essa razão, é necessário indicar explicitamente o nível de referência adotado, neste trabalho usou-se o primeiro.

A.3 dBov

dBov é o nível relativo ao limiar de saturação (overload) de um sinal em um computador ou codec digital (ITU-T Rec. P.830, 2005). Por exemplo, para um sistema de 16 bits, o nível máximo 0 dBov corresponde a um nível DC igual a 32767. Este decibel é largamente utilizado em implementações digitais.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo