IDENTIFICAÇÃO DA NATUREZA DO RUÍDO COM
APLICAÇÃO EM RECONHECIMENTO ROBUSTO DE VOZ

Denilson da Cruz da Silva

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. Fernando Gil Vianna Resende Junior, Ph.D.

Prof. Mariane Rembold Petraglia, Ph.D.

Prof. Márcio Nogueira de Souza, D.Sc.

Prof. Abraham Alcaim, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
FEVEREIRO DE 2005

SILVA, DENILSON DA CRUZ DA

Identificação da natureza do ruído com aplicação em reconhecimento robusto de voz [Rio de Janeiro] 2005

xv,126 pp 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2005)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1.Identificação da natureza do ruído

2.Reconhecimento Robusto de Voz

3.Detecção de Extremos

4.Banco de Filtros

5.Processamento em Sub-bandas

I.COPPE/UFRJ II.Título (série)

Dedicatória

Gostaria de dedicar esta tese:

- à minha amada esposa, Silvia, que com bastante sabedoria, dirigiu-me as palavras corretas, nos momentos mais difíceis de desenvolvimento deste trabalho.
- às minhas queridas filhas, Silvia Valéria, Denise Vitória e Desirèe, que foram bastante compreensivas nos momentos que abdicamos do lazer para eu poder me dedicar ao desenvolvimento das pesquisas.
- à minha querida mãe, irmãos e demais familiares, que sempre me deram total apoio e incentivo.
- ao amigo e orientador, prof. Fernando Gil, que sempre me indicou as direções certas a serem tomadas, para que os resultados pudessem ser obtidos de forma significativa.
- sobretudo e todos, a Deus, que me deu condições de chegar até aqui, vencendo todas as lutas e dificuldades, com total confiança Nele.

"O cavalo se prepara para a batalha, mas do Senhor vem a vitória"

Provérbios 21:31

Denilson da Cruz da Silva

Agradecimentos

Eu gostaria de agradecer à minha família, colegas e amigos, por todo o apoio dispensado ao longo dos anos de preparação desta tese. À minha esposa Silvia, e filhos, externo um agradecimento especial pela compreensão que sempre tiveram.

Quero agradecer ao meu orientador, prof. Fernando Gil, pela confiança e oportunidade que me foram dispensadas no desevolvimento deste tema e por todo incentivo e apoio irrestrito em todos os momentos, até mesmo nos dedicados ao seu descanso no lar.

Gostaria de agradecer à prof[a]. Mariane Petraglia, ao prof. Márcio Souza e ao prof. Abraham Alcaim, pela participação na banca de defesa de tese.

Agradeço aos meus colegas do LPS, que de alguma forma contribuíram para a realização deste trabalho e pelos bons momentos passados juntos. A todos os funcionários, alunos e professores do PEE e da COPPE.

Agradeço também ao Eng. Maurício Moraes, chefe da Subdivisão de Engenharia do PAMAGL, pelo apoio dispensado e a todos os companheiros do Comando da Aeronáutica que contribuiram para a realização deste trabalho.

E, finalmente, quero expressar o meu principal agradecimento, que não poderia ser direcionado a outro, senão ao meu grandioso Deus, que me deu saúde, força e fé para prosseguir sempre, principalmente nos momentos de maior dificuldade.

"Porque eu conheço que o Senhor é grande e que o nosso Senhor está acima de todos os deuses." Salmos 135:5

Obrigado a todos!

Denilson da Cruz da Silva

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

IDENTIFICAÇÃO DA NATUREZA DO RUÍDO COM APLICAÇÃO EM RECONHECIMENTO ROBUSTO DE VOZ

Denilson da Cruz da Silva

Fevereiro/2005

Orientador: Fernando Gil Vianna Resende Junior

Programa: Engenharia Elétrica

Esta tese apresenta um conjunto de algoritmos para reconhecimento robusto de voz, que possibilitam realizar o reconhecimento da fala de forma mais adequada, combinando o processamento em banda única e o processamento em múltiplas bandas, e seus parâmetros tais como: tipo de banco de filtros, número de sub-bandas, etc. A base é o algoritmo para identificar a natureza do ruído, que utilizado em conjunto com a estimação da sua potência, permite conduzir apropriadamente o reconhecimento.

O algoritmo para identificação da natureza do ruído é realizado através de Hidden Markov Model (HMM) e os parâmetros extraídos são entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Com o método proposto, para quatro possíveis tipos de ruído (branco, rosa, falatório e interior de carro), obtivemos $97,22\%$ de correta identificação da natureza do ruído quando misturados com comandos de voz, e 100% quando o sinal é composto apenas por ruído.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

IDENTIFICATION OF THE NOISE NATURE WITH
APPLICATION IN ROBUST SPEECH RECOGNITION

Denilson da Cruz da Silva

February/2005

Advisor: Fernando Gil Vianna Resende Junior

Department: Electrical Engineering

This thesis presents a collection of algorithms for robust speech recognition, that make possible to carry out the speech recognition in a more appropriate way, combining the single band processing with the multiband processing, and its parameters such as: type of filterbank, number of subbands, etc. The foundation is the algorithm to identify the noise nature, that used together with the estimate of its power, allows to lead the recognition adequately.

The algorithm for identification of the noise nature is carried out by HMM and the extracted parameters are spectral entropy, zero crossing rate and log-energy from 16 subbands. With the proposed method, for four possible types of noise (white, pink, babble and car interior), we got 97.22% of correct identification of the noise nature when mixed with voice commands, and 100% when the signal is composed only of noise.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

BFD - Banco de Filtros Digitais

CMN - Normalização da média cepstral (*Cepstral Mean Normalization*)

DFT - Transformada de Fourier Discreta (*Discrete Fourier Transform*)

FFT - Transformada Rápida de Fourier (*Fast Fourier Transform*)

FIR - Resposta ao Impulso Finita (*Finite Impulse Response*)

HMMs - Modelos Escondidos de Markov (*Hidden Markov Models*)

IFFT - Transformada Rápida de Fourier Inversa (*Inverse Fast Fourier Transform*)

IIR - Resposta ao Impulso Infinita (*Infinite Impulse Response*)

LBG - *Linde-Buzo-Gray*

MLLR - Regressão Linear de Máxima Verossimilhança (*Maximum Likelihood Linear Regression*)

PSD - Densidade Espectral de Potência (*Power Spectral Density*)

QMF - Filtro de Quadratura de Espelho (*Quadrature Mirror Filter*)

SNR - Relação Sinal-Ruído (*Signal-to-Noise Ratio*)

SPIB - *Signal Processing Information Base*

SRRV Sistemas de Reconhecimento Robusto de Voz

SRV Sistemas de Reconhecimento de Voz

Capítulo 1

Introdução

1.1 Motivação

Vários Sistemas de Reconhecimento de Voz (SRV) têm sido desenvolvidos recentemente [1, 2, 3] visando tornar a interação homem-máquina uma realidade com a máxima eficiência possível.

Apesar de todo o desenvolvimento no âmbito da tecnologia de reconhecimento de voz, uma questão muito importante precisa ser considerada: o ruído ambiente (Figura 1.1). A precisão dos SRV é bastante afetada quando eles são operados em ambientes acústicos impregnados por ruído.

A inserção de algoritmos que tornem os SRV mais robustos ao ruído é imprescindível e, nos últimos anos, a área de reconhecimento robusto de voz [4, 5, 6] tem se tornado uma das mais promissoras.

Na Figura 1.2 temos amostras de sinais limpo e contaminado por falatório com Relação Sinal-Ruído (SNR - *Signal-to-Noise Ratio*) de 5dB. Podemos observar como o ruído se sobrepõe aos trechos de baixa energia da locução, o que afeta diretamente os resultados, já que num processo tradicional de reconhecimento de voz, o trecho de voz sobreposto pelo ruído pode ser processado pelo sistema como sendo apenas ruído.

Desta forma, sistemas que não sejam robustos ao ruído não têm uma aplicação prática muito expressiva, já que o ruído ambiente está presente todo tempo. Nesta

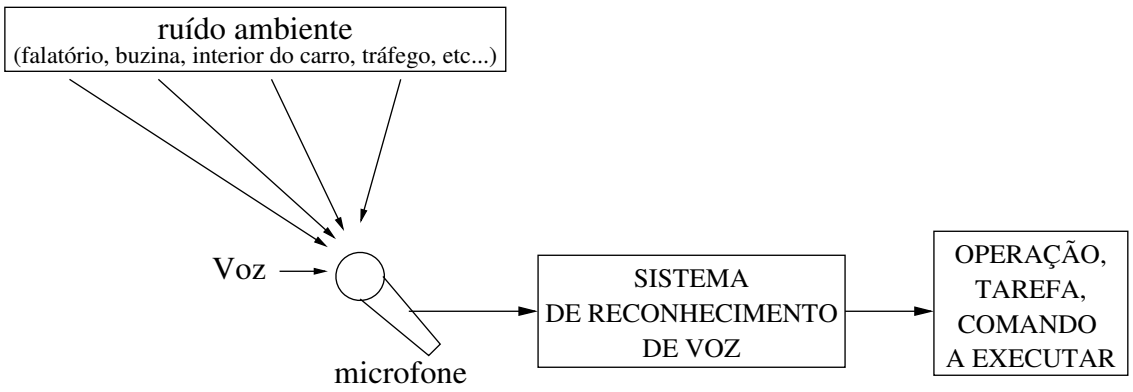


Figura 1.1: Reconhecimento de voz em ambiente ruidoso.

tese, exploramos a robustez nos SRV, como forma de desenvolver pesquisas em sistemas e algoritmos aplicáveis ao mundo real, ou seja, eficientes em condições reais de campo.

Em [7] temos uma análise comparativa de três técnicas para melhorar os índices de acerto no reconhecimento de voz em presença de ruído, aplicadas individualmente e duas a duas: Regressão Linear de Máxima Verossimilhança (MLLR - *Maximum Likelihood Linear Regression*), Subtração Espectral e Normalização da média cepstral (CMN - *Cepstral Mean Normalization*). Cada uma destas técnicas pertencem a diferentes categorias, que atuam, respectivamente, na compensação dos modelos, no realce da fala e nas características robustas ao ruído. Os resultados apresentados são significativos e a combinação da subtração espectral com a MLLR é a que apresenta melhor performance, no caso de reconhecimento de dígitos isolados dependente do locutor.

Um método que vem sendo utilizado, tanto em SRV [8] como em Sistemas de Reconhecimento Robusto de Voz (SRRV) [9, 10], é o processamento em sub-bandas. Neste método, o espectro de freqüência é dividido em sub-bandas através de um Banco de Filtros Digitais (BFD). Em seguida, o sinal é processado dentro da sistemática de reconhecimento de voz.

A proposta desta tese é melhorar a performance no reconhecimento através da identificação da natureza do ruído ambiente e da sua potência. Isto pode proporcionar a vantagem de conduzir apropriadamente o reconhecimento, optando por

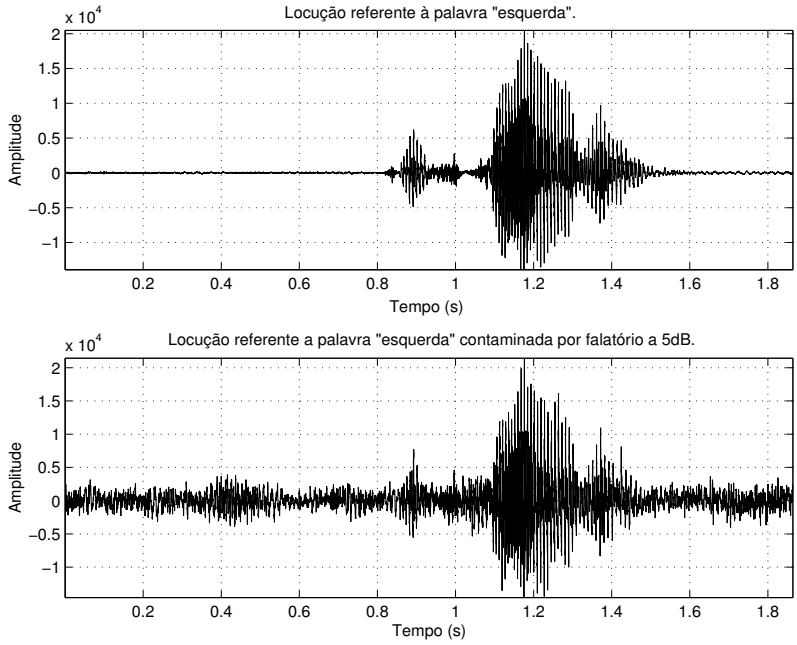


Figura 1.2: Locução referente à palavra "esquerda" na condição limpa e na condição contaminada por falatório com SNR de 5dB.

método em banda única ou em múltiplas bandas, e pelos parâmetros mais adequados para um maior índice de acertos, tais como tipo de banco de filtros, número de sub-bandas, etc.

A Figura 1.3 mostra, esquematicamente, onde o algoritmo para condução apropriada do reconhecimento atua. No algoritmo proposto, os parâmetros extraídos do sinal ruidoso configuram outros parâmetros do SRRV de forma que o reconhecimento seja mais eficiente.

1.2 Resultados obtidos

- Um algoritmo para identificação da natureza do ruído é apresentado. O processo de identificação é realizado através de Modelos Escondidos de Markov (HMMs - *Hidden Markov Models*) e os parâmetros extraídos são entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Com este método, para quatro possíveis tipos de ruído (branco, rosa, falatório e

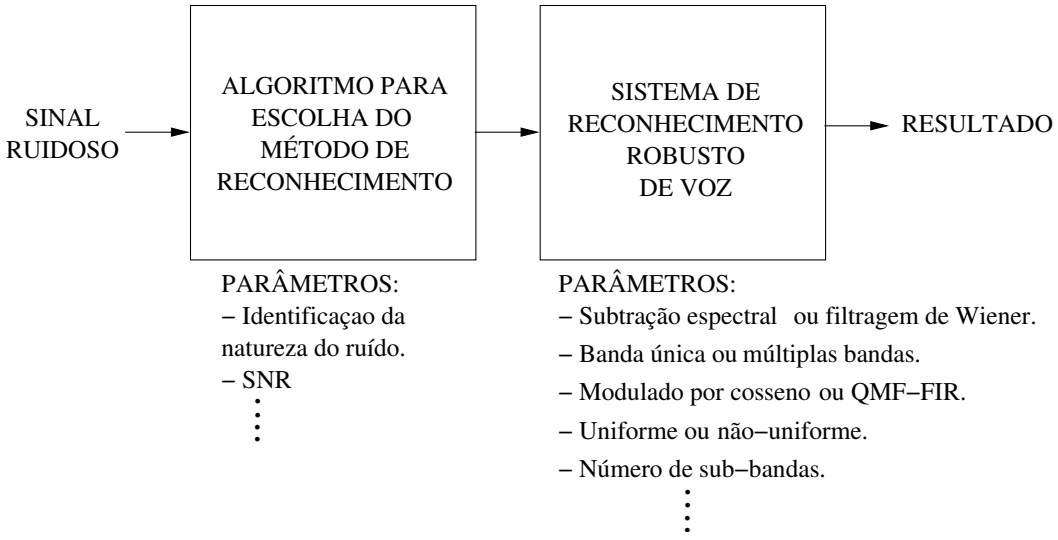


Figura 1.3: Proposta da tese com o algoritmo para configuração do SRRV.

interior de carro), obtivemos $97,22\%$ de correta identificação da natureza do ruído quando misturados com comandos de voz, e $100,00\%$ quando o sinal é composto apenas por ruído.

- Uma aplicação para o método de identificação da natureza do ruído através de HMM, associado à SNR e à distância euclidiana da log-energia calculados em cada *frame*, é apresentado em um detector de extremos robusto. Com este método a redução média da taxa de erro na detecção de início foi de até $12,23\%$ e na detecção de fim foi de até $26,24\%$, para o caso de ruído rosa com SNR de 0 a 20dB, em relação ao método tradicional baseado na energia e na taxa de cruzamentos por zero conforme implementado em [11].

- O algoritmo de decisão em sub-bandas e o algoritmo para a detecção de extremos robusta apresentam uma melhoria de até 32% na taxa de reconhecimento em múltiplas bandas [12], para o caso de ruído rosa, em condições críticas de contaminação, quando comparado com o método tradicional de detecção de extremos baseado na energia e taxa de cruzamentos por zero conforme implementado em [11].

- A partir da identificação da natureza do ruído e estimação da SNR, implementou-

se um algoritmo que seleciona os parâmetros mais adequados para um SRRV (BFD, número de bandas, etc).

1.3 Estrutura da tese

No Capítulo 2, descrevemos brevemente alguns aspectos teóricos sobre ruído inseridos no contexto desta tese, além de alguns dos seus tipos e algumas formas de caracterizá-lo.

No Capítulo 3 são apresentadas duas técnicas de realce da fala utilizadas, subtração espectral e filtragem de Wiener, e introduzimos o algoritmo para estimar a potência do ruído, bem como a SNR do sinal.

O método proposto de identificação da natureza do ruído é apresentado no Capítulo 4, juntamente com os detalhes do algoritmo. Uma aplicação para o método proposto é introduzida através de um detector de extremos robusto baseado na identificação da natureza do ruído, na SNR dos *frames* e na distância euclidiana da log-energia de 16 sub-bandas. Os resultados de um teste com 121 locuções contaminadas por quatro tipos de ruídos com SNR de 0 a 20dB são apresentados.

No Capítulo 5 descrevemos a forma como o processamento em múltiplas bandas é realizado. Fazemos um breve comentário sobre a teoria de BFD utilizada no trabalho, bem como a descrição do sistema de reconhecimento em múltiplas bandas e seus principais blocos, incluindo o algoritmo de detecção de extremos robusto e o algoritmo de decisão em sub-bandas.

O Capítulo 6 versa sobre a base de dados utilizada, tanto de voz como de ruído, bem como apresenta os resultados dos testes realizados de identificação da natureza do ruído e de comparação do processamento em banda única com o processamento em múltiplas bandas. Uma análise do reconhecimento mais adequado é realizada, mediante o conhecimento da natureza do ruído e da sua potência.

No Capítulo 7, temos as conclusões e as propostas para realização de trabalhos futuros.

No Apêndice A descrevemos o algoritmo para confecção do sinal ruidoso.

A resposta em freqüência dos bancos de filtros utilizados nesta tese estão ilustradas no Apêndice B, juntamente com as tabelas de coeficientes dos filtros protótipos.

O Apêndice C contém as tabelas com os resultados dos testes realizados.

E os Apêndices D, E e F, apresentam os artigos publicados ao longo do desenvolvimento desta tese.

Capítulo 2

Aspectos teóricos sobre ruídos

Neste capítulo, apresentamos uma breve descrição de alguns aspectos teóricos sobre ruídos. Na Seção 2.1 temos o conceito de ruído e o seu impacto no processo de reconhecimento de voz. Alguns tipos de ruído são mencionados na Seção 2.2. E finalmente, na Seção 2.3 são citadas algumas formas de caracterização dos ruídos utilizados nesta tese, tanto no domínio do tempo como no domínio da freqüência.

2.1 O ruído

Por definição, ruído é qualquer sinal que tem a capacidade de reduzir a inteligibilidade da informação contida em um sinal de voz, aúdio, imagem ou dados [13]. Os ruídos são sinais indesejáveis e provêm de uma variedade de fontes.

Em sistemas de reconhecimento de voz, a inteligibilidade da informação precisa ser preservada, de forma que os índices de acerto atendam aos níveis aceitáveis de eficiência e boa performance. Em ambiente ruidoso, a tarefa de manutenção da inteligibilidade é difícil, já que a degradação do sinal ocorre naturalmente e de forma ascendente, à medida que o nível de ruído também se eleva, caso nenhum tipo de medida seja tomada. Podemos observar na Figura 1.2 como o ruído ambiente mascara o sinal de voz, onde partes do sinal (regiões da fala com uma menor quantidade de energia) são completamente "imersas" no ruído.

Atualmente temos aplicações de reconhecimento de voz em vários dispositivos

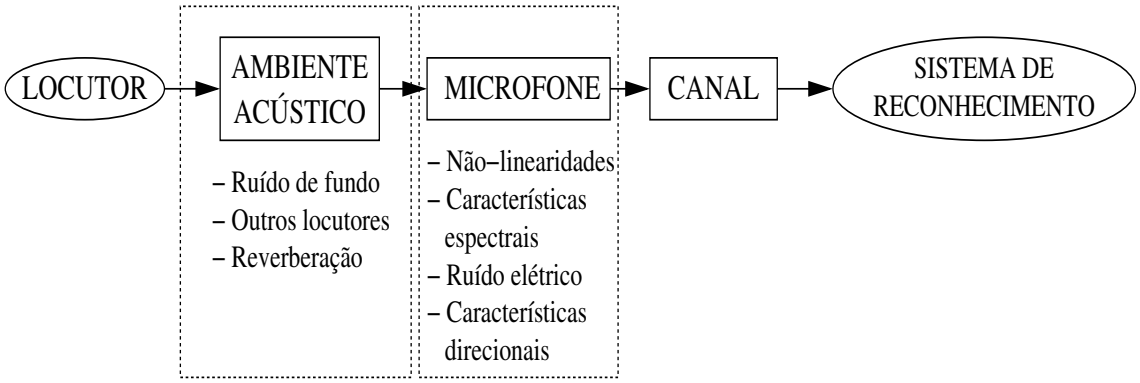


Figura 2.1: Condições prejudiciais ao sinal de voz no percurso entre o locutor e o SRV.

e equipamentos, com ambientes de operação diversificados. Por exemplo, aparelhos de telefonia móvel podem estar sendo utilizados em ambientes com ruído proveniente de uma rua com tráfego intenso, do falatório em um ambiente público ou até mesmo no interior de um veículo em movimento, susceptível ao ruído do seu interior como também o do meio externo (motor, atrito com a pista, vento relativo, chuva, buzina, etc). A Figura 2.1 representa o percurso desde a fonte do sinal de voz (locutor) até o SRV. Ao longo do percurso podemos observar que existem várias condições que podem vir a comprometer o reconhecimento. Sendo assim, os SRV devem possuir algoritmos capazes de tornar a tarefa de reconhecimento mais resistente à ação do ruído, ou seja, devem ser mais robustos.

2.2 Tipos de ruídos

Dois tipos de ruídos, em geral, vão fazer a composição do sinal ruidoso, juntamente com a voz limpa: o ruído aditivo e o ruído convolucional (Figura 2.2). Nesta tese nos detemos na degradação do sinal de voz por ruído aditivo.

2.2.1 Ruído Aditivo

O ruído aditivo é caracterizado como qualquer fonte de ruído descorrelatado, ou seja, estatisticamente independente do sinal de voz, sobreposto ao sinal dese-

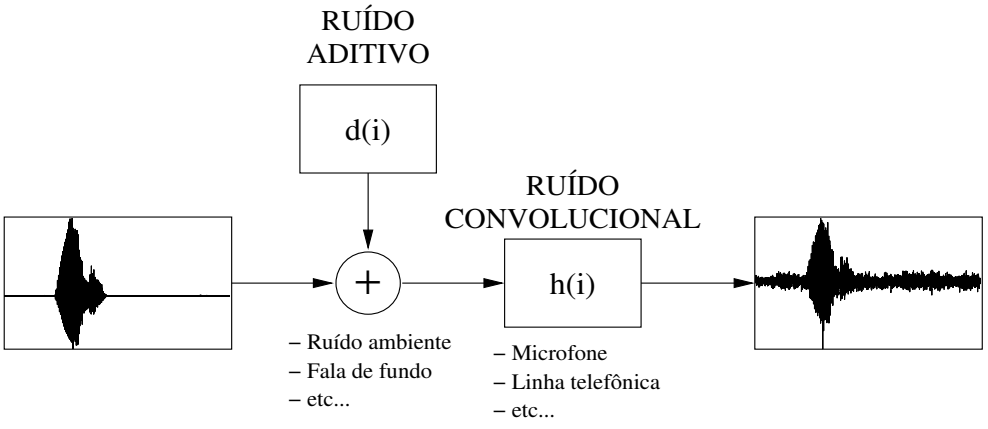


Figura 2.2: Composição do sinal ruidoso.

jado [14]. Por exemplo falatório, motores, carros, helicópteros, etc. Ele é às vezes chamado de "ruído de fundo" ou "ruído ambiente". Ele está adicionado ao sinal desejado, mas não carrega qualquer informação de interesse, pelo contrário, degrada a informação contida no sinal de voz.

Seja o sinal de voz $x(i)$, uma sobreposição do sinal de voz $s(i)$, no domínio do tempo, com o ruído $d(i)$:

$$x(i) = s(i) + d(i) \tag{2.1}$$

onde i é a i-ésima amostra do sinal.

A mesma sobreposição acontece no domínio da freqüência:

$$X(z) = S(z) + D(z) \tag{2.2}$$

onde $X(z)$, $S(z)$ e $D(z)$ são, respectivamente, as transformadas Z das seqüências discretas $x(i)$, $s(i)$ e $d(i)$. A tranformada Z da seqüência $x(i)$ é dada por:

$$X(z) = \sum_{i=-\infty}^{\infty} x(i) z^{-i} \tag{2.3}$$

Podemos observar que a transformada Z é uma transformação linear, assim sendo o resultado da sua aplicação sobre (2.1) mantem a propriedade de linearidade.

2.3 Caracterização do ruído

Nesta seção realizamos uma breve descrição de algumas características temporais e espectrais dos seguintes tipos de ruídos: branco, rosa, falatório e interior do carro, considerando o princípio de ergodicidade [15].

2.3.1 Caracterização temporal

No domínio do tempo, a caracterização pode ser realizada por suas propriedades estatísticas, tais como a média μ e a variância σ^2 (ou o desvio padrão σ) [16, 17, 18, 19], considerando ruído gaussiano.

Seja uma seqüência $\mathbf{d} = [d(0) \dots d(N-1)]^T$, a média é definida nesta tese como:

$$\mu_d = \frac{1}{N} \sum_{i=0}^{N-1} d(i) \tag{2.4}$$

A variância da seqüência é definida como:

$$\sigma_d^2 = \frac{1}{N} \sum_{i=0}^{N-1} [d(i) - \mu_d]^2 \tag{2.5}$$

Alguns tipos de ruídos também podem ser caracterizados pela sua distribuição ao longo do tempo, segundo uma Função Densidade de Probabilidade, fdp, que pode seguir diversas leis de probabilidade descritas em [15]. Podemos citar como exemplo, os ruídos branco e rosa que possuem distribução gaussiana, ou seja, a distribuição desses sinais segue a curva de Gauss (Figura 2.3).

A Figura 2.4 apresenta, respectivamente, histogramas com a distribuição do ruído branco e do ruído rosa. Nesta figura, podemos constatar a distribuição gaussiana dos ruídos apresentados. Ambos estão concentrados, aproximadamente, em torno da média zero.

Outra forma de caracterizarmos o ruído é através da função autocorrelação $R_{dd}(m)$, na qual temos uma medida da similaridade de um sinal com sua versão atrasada no tempo [20, 21].

$$R_{dd}(m) \quad = \quad \begin{cases} \sum_{i=0}^{N-m-1} d(i+m)d(i), & \text{para } m \geq 0 \\ R_{dd}(-m), & \text{para } m < 0 \end{cases} \tag{2.6}$$

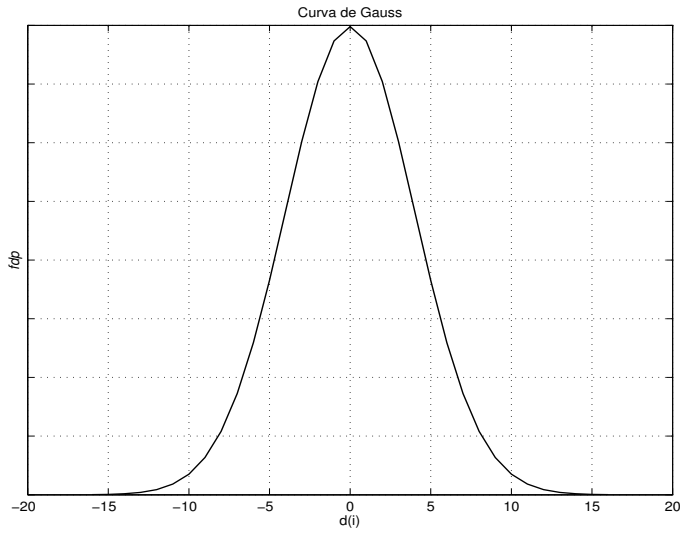


Figura 2.3: Curva de Gauss.

No exemplo mostrado nas Figuras 2.5, 2.6 e 2.7 temos a autocorrelação de um trecho de fala limpa e a autocorrelação de segmentos de ruído branco e ruído rosa, falatótrio e ruído no interior do carro, respectivamente. Podemos observar como a autocorrelação dos sinais apresentados possuem características distintas.

Associada com a autocorrelação, podemos citar também a autocovariância, $C_{dd}(m)$, utilizada também na análise temporal de processos aleatórios.

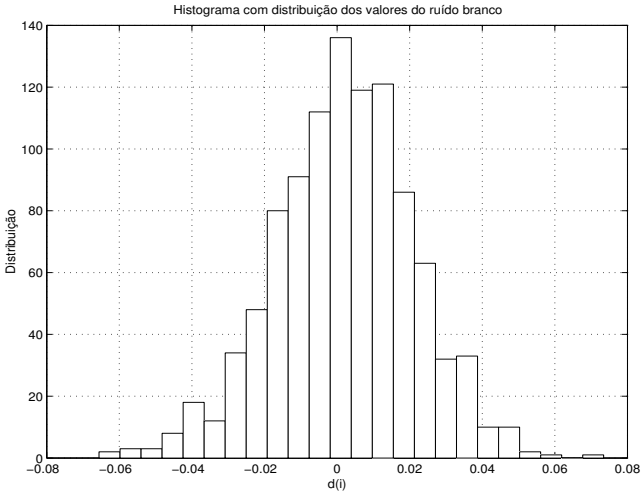
$$C_{dd}(m) = R_{dd}(m) - \mu_d^2 \tag{2.7}$$

Tanto a autocorrelação como a autocovariância, conforme descritas em (2.6) e (2.7), devem ser tidas como estimativas, porque apenas podemos ter acesso a um segmento finito do processo aleatório **d** em análise.

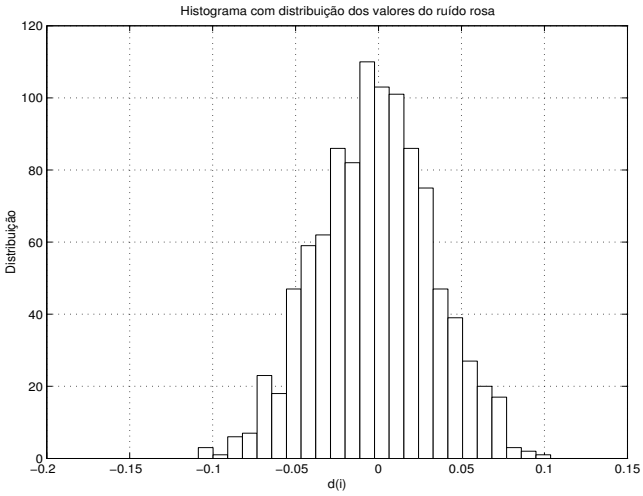
2.3.2 Caracterização espectral

No domínio da freqüência, algumas das formas de caracterização do ruído são a entropia espectral [22] e a Densidade Espectral de Potência (PSD - *Power Spectral Density*) [20, 21].

A entropia espectral, H_d, que está relacionada com a idéia de que um sinal tem tanto mais informação quanto maior for o seu grau de imprevisibilidade, é



(a) Ruído branco.



(b) Ruído rosa.

Figura 2.4: Histogramas com a distribuição temporal.

definida na equação a seguir:

$$H_d = -\sum_{k=0}^{N-1} p_k log(p_k) \tag{2.8}$$

onde

$$p_j = \frac{D(f_j)}{\sum_{k=0}^{n-1} D(f_k)}, \text{ para } j = 0, ..., n-1 \tag{2.9}$$

na qual p_j é uma função densidade de probabilidade do espectro na freqüência de

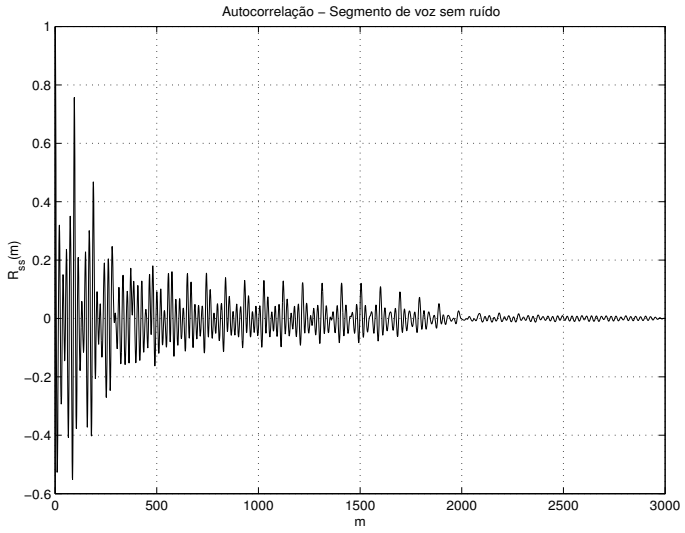


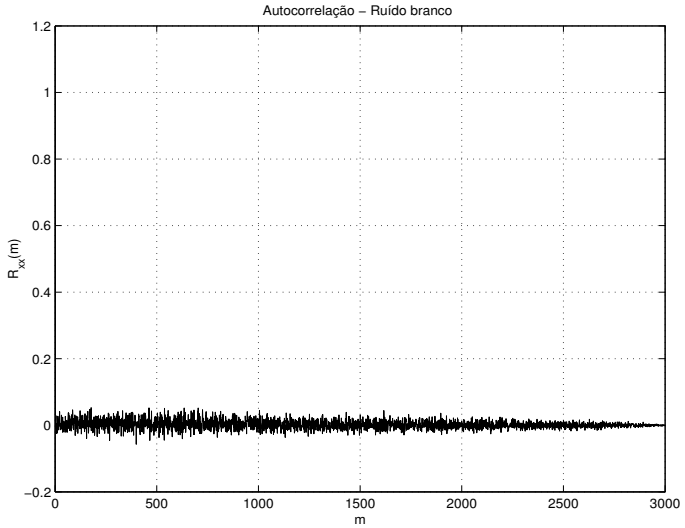
Figura 2.5: Autocorrelação da voz sem ruído.

índice j, estimada pela normalização sobre todas as componentes de freqüência e n é o número de componentes de freqüência.

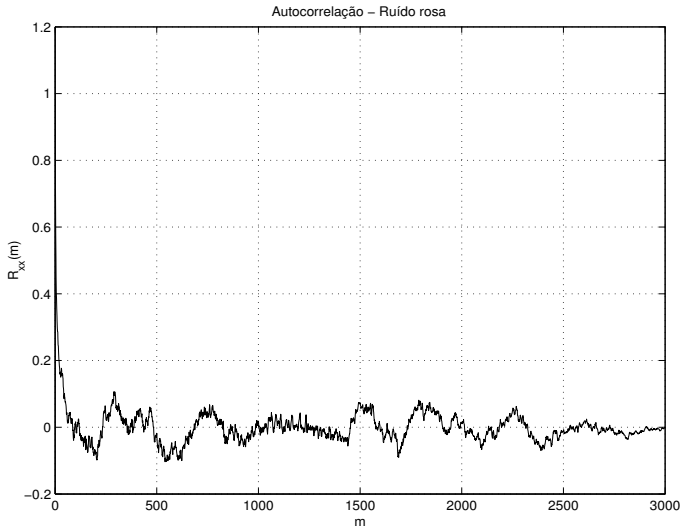
Uma outra forma de caracterizar o ruído é através da densidade espectral de potência, $P_{dd}(\omega)$. A PSD, do processo aleatório estacionário **d** está relacionada com a seqüência de autocorrelação através da Transformada de Fourier Discreta (*DFT - Discrete Fourier Transform*) [20]:

$$P_{dd}(\omega) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} R_{dd}(m) e^{-j\omega m} \tag{2.10}$$

A Figura 2.8 mostra a PSD, expressa em dB/Hz, de um segmento do ruído gravado no interior de um carro, como descrito na Subseção 6.1.2.

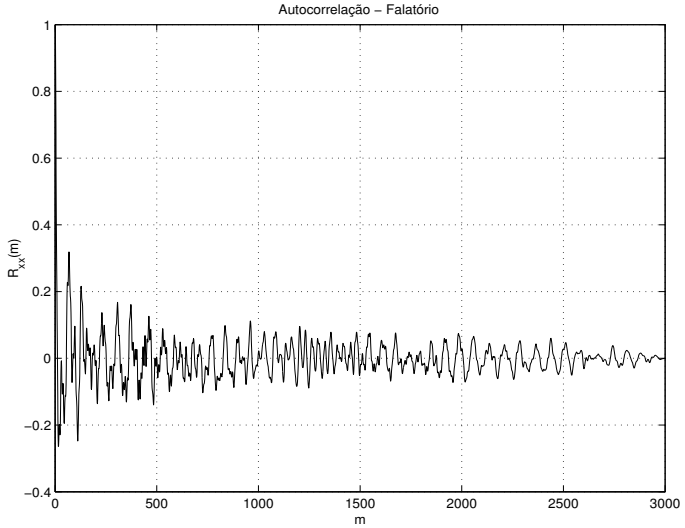


(a) Autocorrelação do ruído branco.

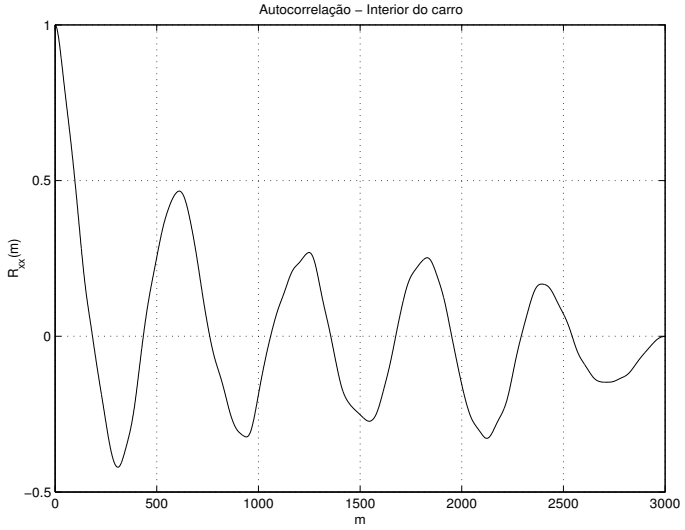


(b) Autocorrelação do ruído rosa.

Figura 2.6: Autocorrelação dos ruídos branco e rosa.



(a) Autocorrelação do falatório.



(b) Autocorrelação do ruído no interior do carro.

Figura 2.7: Autocorrelação do falatório e do ruído no interior de um carro.

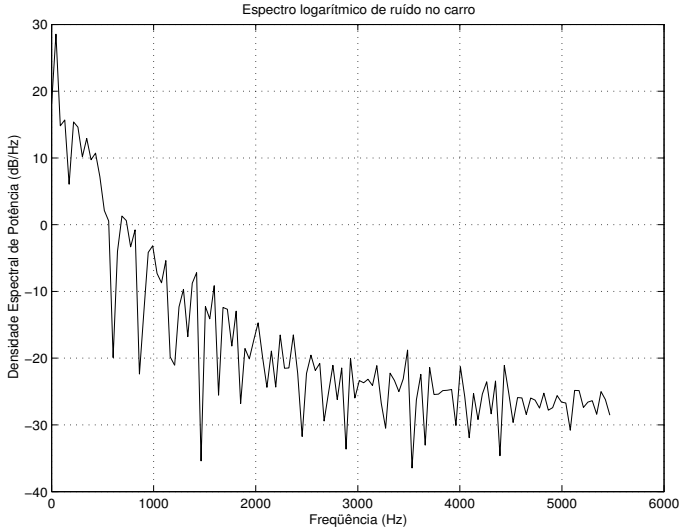


Figura 2.8: Densidade espectral de potência do ruído no interior do carro.

Capítulo 3

Algoritmos para pré-processamento robusto

Neste capítulo apresentamos alguns algoritmos utilizados no pré-processamento robusto dos sinais de voz. Na Seção 3.1 descrevemos as técnicas robustas de subtração espectral e filtragem de Wiener. O algoritmo utilizado para estimar a potência do ruído e calcular a SNR é descrito na Seção 3.2.

3.1 Técnicas de realce da fala

Duas técnicas de realce da fala são mencionadas nesta seção: subtração espectral e filtragem de Wiener.

3.1.1 Subtração espectral

A técnica de subtração espectral é utilizada para realçar a fala na presença de ruído aditivo [23, 24]. Especificamente, o espectro do ruído é estimado das pausas que ocorrem no sinal de voz, por exemplo, em alguns milisegundos iniciais do sinal. A Figura 3.1 mostra uma estrutura simplificada do método de subtração espectral.

Como apresentado na Seção 2.2.1, a fala contaminada por ruído aditivo pode ser descrita como:

$$x(i) = s(i) + d(i) \tag{3.1}$$

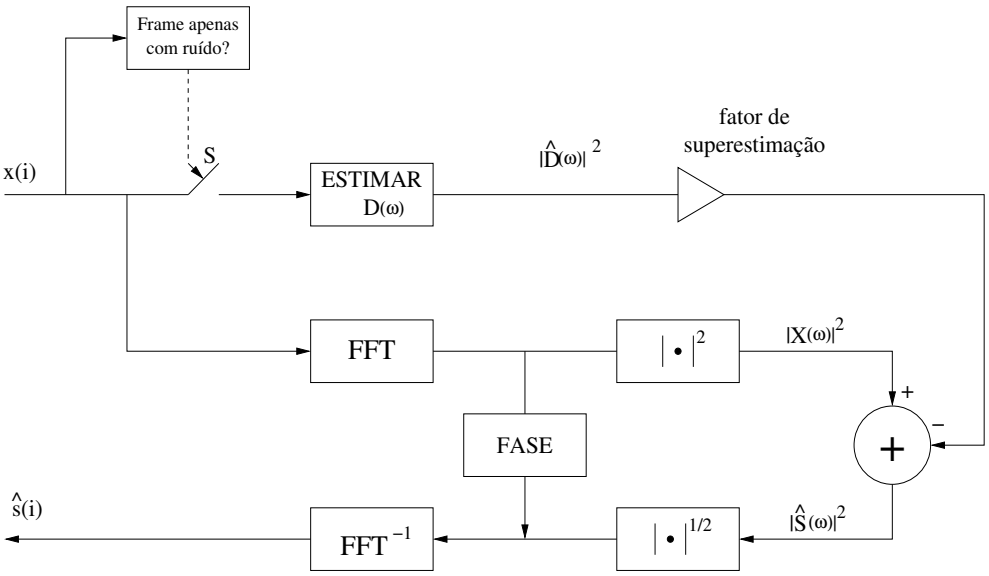


Figura 3.1: Método de redução de ruído baseado na subtração espectral.

onde $x(i)$ é a fala contaminada, $s(i)$ é a fala limpa e $d(i)$, o ruído aditivo, discretos no domínio do tempo. i é o índice da i-ésima amostra dos sinais discretos no tempo.

A Subtração Espectral estima $s(i)$ a partir de $x(i)$. Como $d(i)$ é um processo aleatório, podemos fazer certas suposições:

- o ruído é um processo estacionário de curta duração, ou seja, é assumido que o espectro de potência do ruído permanece constante durante curtos segmentos do sinal;
- o ruído é considerado descorrelatado do sinal de voz;
- a audição humana é insensível às variações de fase, de forma que o efeito do ruído na fase de $s + d$ pode ser ignorado [20, 25].

Considerando uma análise em cada *frame*, $x(k, i)$, o espectro é representado por $|X(k, \omega)|^2$, onde a partir daqui, omitimos o índice k do *frame* no espectro.

Se o ruído é representado pelo seu espectro de potência estimado $|\hat{D}(\omega)|^2$, a partir de períodos do sinal com ausência de voz, o espectro de potência estimado da voz, $|\hat{S}(\omega)|^2$, pode ser escrito como:

$$|\hat{S}(\omega)|^2 = |X(\omega)|^2 - |\hat{D}(\omega)|^2 \tag{3.2}$$

onde $|X(\omega)|^2$ é o espectro da fala ruidosa.

Generalizando a forma de representação, temos a expressão da subtração espectral tradicional [23]:

$$|\hat{S}(\omega)|^a = |X(\omega)|^a - |\hat{D}(\omega)|^a \tag{3.3}$$

onde a é igual a 1 para subtração espectral em magnitude e a igual a 2 para subtração espectral em potência.

A fase da voz $\phi_{\hat{S}}(\omega)$ é estimada diretamente da fase do sinal ruidoso $\phi_X(\omega)$ [20, 25].

$$\phi_{\hat{S}}(\omega) = \phi_X(\omega) \tag{3.4}$$

Pelo apresentado em (3.2), não temos nenhuma garantia que $|\hat{S}(\omega)|^2$ será positivo. Valores negativos fazem com que sejam introduzidos no sinal efeitos conhecidos como ruído musical. Ele é provocado pela variação de amplitude do espectro do sinal aleatório (ruído). Caracteriza-se por curtas explosões de tons isolados distribuídos aleatoriamente ao longo da freqüência.

Uma forma para a estimativa da fala no domínio da freqüência, que reduz razoavelmente o ruído musical, pela diminuição das não-linearidades no espectro, foi proposta por [26] como:

$$|\hat{S}(\omega)| = \left(\max(|X(\omega)|^a - \xi|\hat{D}(\omega)|^a, \varphi|\hat{D}(\omega)|^a)\right)^{\frac{1}{a}} \tag{3.5}$$

onde o fator de subtração ξ, que deve ser maior que 1, é usado na superestimação do ruído considerando o espectro estimado, ou seja, ξ controla a quantidade de subtração. Em [26] é sugerido um valor entre 3 e 6, mas que também pode ser dependente da SNR. φ, compreendido entre $0,005$ e $0,1$, define o valor espectral mínimo após a subtração.

O algoritmo para a subatração espectral utilizado nesta tese, conforme diagrama em bloco apresentado na Figura 3.1, é descrito na Tabela 3.1.

3.1.2 Filtragem de Wiener

Filtragem de Wiener é uma técnica derivada da teoria do filtro ótimo [27]:

$$E|s(i) - \hat{s}(i)| = \min \tag{3.6}$$

Tabela 3.1: Algoritmo para Subtração Espectral.

Passo 1 - Segmentar o sinal ruidoso, $x(i)$;

Passo 2 - Aplicar a janela de Hamming, $W(i)$, sobre cada *frame*, $x(k,i)$;

Passo 3 - Calcular o espectro, $|X(k,\omega)|^2$, e a fase, $\phi_X(k,\omega)$, via Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*), sendo $\omega = \frac{2\pi i}{N}$;

for $k = 0$ to $K-1$ **do**

  for $i = 0$ to $N-1$ **do**

    $x^{(w)}(k,i) = x(k,i)W(i)$

    $|X(k,\omega)|^2 = (Re\{FFT[x^{(w)}(k,i)]\})^2 + (Im\{FFT[x^{(w)}(k,i)]\})^2$

    $\phi_X(k,\omega) = \arctan\left(\frac{Im\{FFT[x^{(w)}(k,i)]\}}{Re\{FFT[x^{(w)}(k,i)]\}}\right)$

  end for

end for

onde $W(i) = 0,54 - 0,46\cos(2\pi i/(N-1))$

Passo 4 - Estimar o espectro médio do ruído, $|\bar{D}(\omega)|^2$, em m *frames* iniciais;

$|\bar{D}(\omega)|^2 = \sum_{k=0}^{m-1} |X(k,\omega)|^2$

Passo 5 - Realizar a subtração do espectro;

Passo 6 - Aplicar a Transformada Rápida de Fourier Inversa (IFFT - *Inverse Fast Fourier Transform*) para recompor o sinal;

for $k = 0$ to $K-1$ **do**

  for $i = 0$ to $N-1$ **do**

    $\hat{S}(k,\omega) = \left(\left(\max(|X(k,\omega)|^2 - \xi|\bar{D}(\omega)|^2, \varphi|\bar{D}(\omega)|^2)\right)^{\frac{1}{2}}\right)$

    $\hat{s}(k,i) = Re\left(IFFT\{\hat{S}(k,\omega)e^{\jmath\phi_X(k,\omega)}\}\right)$

  end for

end for

Da mesma forma que na subtração espectral, a aplicação do filtro de Wiener requer que as estatísticas do ruído sejam obtidas a partir de pausas do sinal corrompido.

Adotando as mesmas considerações usadas na aplicação da subtração espectral quanto a composição do sinal corrompido, $x(i)$ (equação (3.1)), o processo representado por $d(i)$ é considerado descorrelatado.

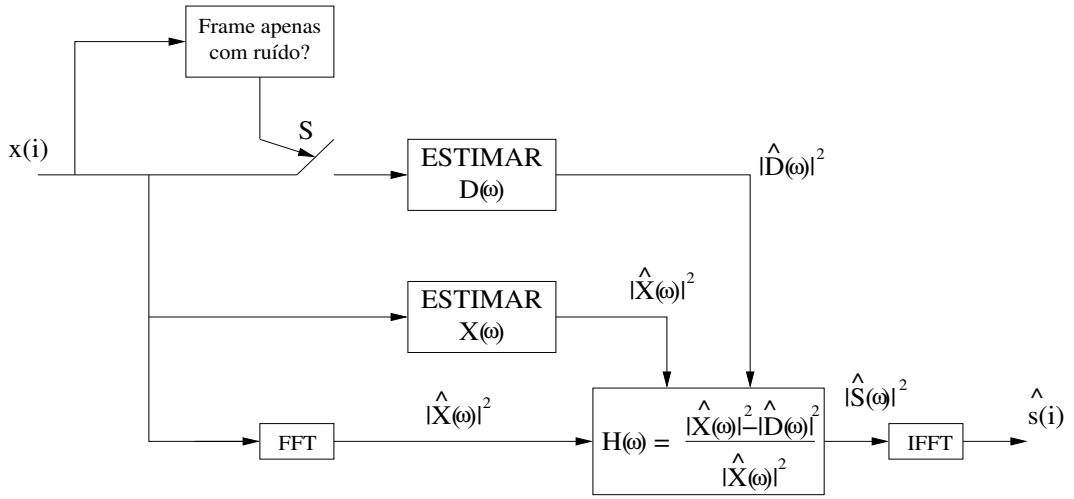


Figura 3.2: Método de redução de ruído baseado na filtragem de Wiener.

O filtro de Wiener, $H(\omega)$, é aplicado sobre o sinal, desde que $x(i)$ seja estacionário a curtos segmentos. Uma forma de realizar a filtragem de Wiener é adotar uma análise em cada *frame*.

Considerando $\Gamma(\omega)$ como uma relação sinal-ruído na componente de freqüência ω,

$$\Gamma(\omega) = \frac{|\hat{S}(\omega)|^2}{|\hat{D}(\omega)|^2} \tag{3.7}$$

sendo a forma usual do filtro de Wiener:

$$H(\omega) = \frac{\Gamma(\omega)}{\Gamma(\omega) + 1} \tag{3.8}$$

onde temos que:

$$\Gamma(\omega) \;=\; \begin{cases} \frac{|X(\omega)|^2 - |\hat{D}(\omega)|^2}{|\hat{D}(\omega)|^2}, & |X(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0, & |X(\omega)|^2 \leq |\hat{D}(\omega)|^2 \end{cases} \tag{3.9}$$

substituindo $\Gamma(\omega)$ na equação 3.9, temos que:

$$H(\omega) \;=\; \begin{cases} 1 - \frac{|\hat{D}(\omega)|^2}{|X(\omega)|^2}, & |X(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0, & |X(\omega)|^2 \leq |\hat{D}(\omega)|^2 \end{cases} \tag{3.10}$$

Dada a resposta do filtro de Wiener $H(\omega)$, a estimativa do espectro do sinal de voz, $|\hat{S}(\omega)|^2$, é então obtida filtrando-se o espectro da fala ruidosa $|X(\omega)|^2$, o qual por ser estimado via FFT, passa a ser grafado com $|\hat{X}(\omega)|^2$.

$$|\hat{S}(\omega)|^2 = H(\omega)|\hat{X}(\omega)|^2 \tag{3.11}$$

Não necessariamente o filtro deve fixar seu limiar inferior em zero, podendo ser realizado o mesmo procedimento adotado na subtração espectral, definindo um limiar inferior como $\varphi|\hat{D}(\omega)|^2$.

O algoritmo para a filtragem de Wiener, conforme diagrama em bloco apresentado na Figura 3.2, é descrito na Tabela 3.2.

3.2 Estimação da potência do ruído e da SNR

Nem sempre temos disponíveis no sinal em análise uma quantidade suficiente de intervalos ou pausas para realizar a estimação do espectro do ruído. Nessas condições, em que praticamente temos trechos com fala ruidosa, o espectro do ruído, que não é exatamente conhecido, pode ser estimado. Basta que tenhamos um trecho mínimo somente com ruído no início do sinal para que uma atualização adaptativa do espectro possa ser realizada.

Uma técnica que vem sendo utilizada é a aplicação de um filtro com Resposta ao Impulso Infinita (IIR - *Infinite Impulse Response*) de um pólo, sobre o sinal ruidoso [28, 29, 30, 31].

Considerando o sinal ruidoso, $x(i)$, descrito na Seção 2.2.1, a tarefa de estimação da potência do ruído é realizada baseada em uma análise em cada *frame*. Uma divisão do sinal, $x(i)$, em K *frames* com tamanho N é realizada para uma posterior identificação global da natureza do ruído.

$$x(k,i) = x(kN+i), \text{ para } 0 \leq k \leq K-1 \text{ e } 0 \leq i \leq N-1 \tag{3.12}$$

onde k é o índice do *frame*.

Estimar a SNR do sinal capturado pelo SRRV é uma iniciativa importante para mensurar quão degradado está o sinal e qual processamento a ser realizado é mais adequado.

Tabela 3.2: Algoritmo para filtragem de Wiener.

Passo 1 - Segmentar o sinal ruidoso, $x(i)$;

Passo 2 - Aplicar a janela de Hamming, $W(i)$, sobre cada *frame*, $x(k,i)$;

Passo 3 - Calcular o espectro, $|X(k,\omega)|^2$, e a fase, $\phi_X(k,\omega)$, via FFT, sendo $\omega = \frac{2\pi i}{N}$;

for $k = 0$ to $K-1$ **do**

  for $i = 0$ to $N-1$ **do**

    $x^{(w)}(k,i) = x(k,i)W(i)$

    $|X(k,\omega)|^2 = (Re\{FFT[x^{(w)}(k,i)]\})^2 + (Im\{FFT[x^{(w)}(k,i)]\})^2$

    $\phi_X(k,\omega) = \arctan\left(\frac{Im\{FFT[x^{(w)}(k,i)]\}}{Re\{FFT[x^{(w)}(k,i)]\}}\right)$

  end for

end for

onde $W(i) = 0,54 - 0,46\cos(2\pi i/(N-1))$

Passo 4 - Estimar o espectro médio do ruído, $|\bar{D}(\omega)|^2$, em m *frames* iniciais;

$|\bar{D}(\omega)|^2 = \sum_{k=0}^{m-1} |X(k,\omega)|^2$

Passo 5 - Calcular o filtro $H(k,\omega)$ e aplicar sobre cada *frame*;

Passo 6 - Aplicar a IFFT para recompor o sinal;

for $k = 0$ to $K-1$ **do**

  for $i = 0$ to $N-1$ **do**

    $H(k,\omega) = 1 - \frac{|\bar{D}(\omega)|^2}{|X(k,\omega)|^2}$

    if $\frac{|X(k,\omega)|^2}{|\bar{D}(\omega)|^2} \geq 1.0$ **then**

      $|\hat{S}(k,\omega)|^2 = |X(k,\omega)|^2 H(k,\omega)$

    else

      $|\hat{S}(k,\omega)|^2 = 0$

    end if

  end for

  for $i = 0$ to $N-1$ **do**

    $\hat{s}(k,i) = Re\left(IFFT\left\{\left(\sqrt{|\hat{S}(k,\omega)|^2}\right) e^{\jmath\phi_X(k,\omega)}\right\}\right)$

  end for

end for

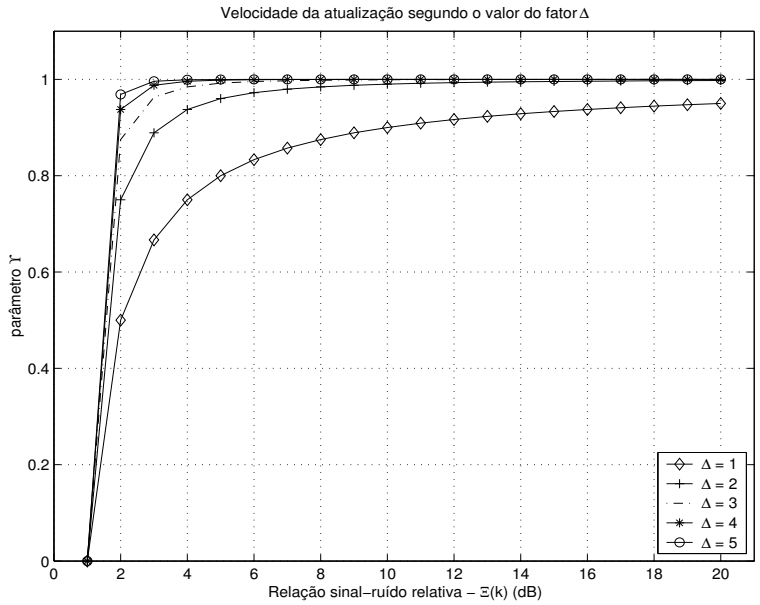


Figura 3.3: Análise da variação de $\Upsilon(k)$ de acordo com o fator Δ escolhido.

Definimos a potência do sinal ruidoso no k-ésimo *frame* ($\hat{\sigma}_x^2(k)$), como sendo a variância do sinal no *frame* em análise.

$$\hat{\sigma}_x^2(k) = \frac{1}{N}\sum_{i=0}^{N-1}[x(k,i) - \mu_x(k)]^2 \tag{3.13}$$

Definindo também $\hat{\sigma}_d^2(k)$ como a potência do ruído no k-ésimo *frame* e assumindo que ela varia de forma mais lenta que a potência do sinal ruidoso, $\hat{\sigma}_x^2(k)$, uma estimativa da potência do ruído é realizada, recursivamente, através do filtro IIR [28]:

$$\hat{\sigma}_d^2(k) = \Upsilon(k)\hat{\sigma}_d^2(k-1) + (1 - \Upsilon(k))\hat{\sigma}_x^2(k) \tag{3.14}$$

onde:

$$\Upsilon(k) = 1 - \min\left\{1, \Xi(k)^{-\Delta}\right\} \tag{3.15}$$

e o operador $\min\{\centerdot\}$ foi utilizado para eliminar possíveis valores negativos de $\Upsilon(k)$. O fator Δ define a velocidade de atualização da estimativa. Na Figura 3.3, mostramos a variação do valor de $\Upsilon(k)$ conforme o valor do fator Δ escolhido, que quanto maior, define mais rapidamente a saturação do valor de $\Upsilon(k)$ em torno de 1. Nesta tese consideramos $\Delta = 5$.

O parâmetro $\Upsilon(k)$ é dinâmico, escolhido de acordo com uma SNR relativa estimada ($\Xi(k)$). Essa SNR está baseada nos primeiros milissegundos do sinal, a fim de tomarmos uma referência para o ruído ambiente e assim termos uma estimação local bem próxima do real, já que a partir deste trecho inicial a atualização adaptativa é realizada pelo filtro. A SNR relativa no k-ésimo *frame*, $\Xi(k)$ [28], é:

$$\Xi(k) = \frac{\hat{\sigma}_x(k)}{\frac{1}{m}\sum_{n=0}^{m-1}\hat{\sigma}_x(n)} \tag{3.16}$$

sendo m um número de *frames* dentro do intervalo inicial, supostamente sem a presença do sinal de voz $s(i)$ e $0 \leq k \leq K-1$.

O parâmetro $\Upsilon(k)$ é o elemento que conduz a atualização da estimativa sob dois aspectos: ou em direção à potência do sinal ruidoso no *frame* k, $\hat{\sigma}_x^2(k)$, ou em direção à estimativa da potência do ruído no *frame* anterior, $\hat{\sigma}_d^2(k-1)$. Assim, se um dado *frame* possui apenas ruído, a sua $\Xi(k)$ ficará bem próxima de 1, resultando num $\Upsilon(k)$ pequeno (em torno de 0). Desta forma, a atualização do ruído segue o sinal ruidoso no *frame* k. Caso o *frame* possua voz, a $\Xi(k)$ é bem mais elevada e $\Upsilon(k)$ fica mais próximo de 1. Neste caso, a estimativa do ruído segue a atualização do *frame* anterior.

A Figura 3.4 mostra a estimativa da potência do ruído $d(i)$ a partir do sinal ruidoso $x(i)$, para o caso de uma locução contaminada por ruído rosa com SNR de 5dB. Podemos observar nesta figura que a estimativa da potência do ruído até o *frame* dois e após o *frame* sete, segue perfeitamente a potência do sinal ruidoso $x(i)$, por se tratar de trechos onde a variação de $\hat{\sigma}_x(k)$ é lenta ao longo dos *frames*. Fora deste trecho a variação entre *frames* de $\hat{\sigma}_x(k)$ é brusca, denotando a presença de $s(i)$ nesta região e a potência do ruído $\hat{\sigma}_d(k)$ passa a seguir as estimativas anteriores.

Com o conhecimento de $\hat{\sigma}_x(k)$ e $\hat{\sigma}_d(k)$, uma estimativa da SNR é feita segundo a equação a seguir:

$$SNR = 10\log\left(\frac{\sum_{k=0}^{K-1}\hat{\sigma}_x(k) - \sum_{k=0}^{K-1}\hat{\sigma}_d(k)}{\sum_{k=0}^{K-1}\hat{\sigma}_d(k)}\right) \tag{3.17}$$

onde K é o número total de *frames* do sinal.

O algoritmo para estimar a potência do ruído, bem como a SNR do sinal, é descrito na Tabela 3.3.

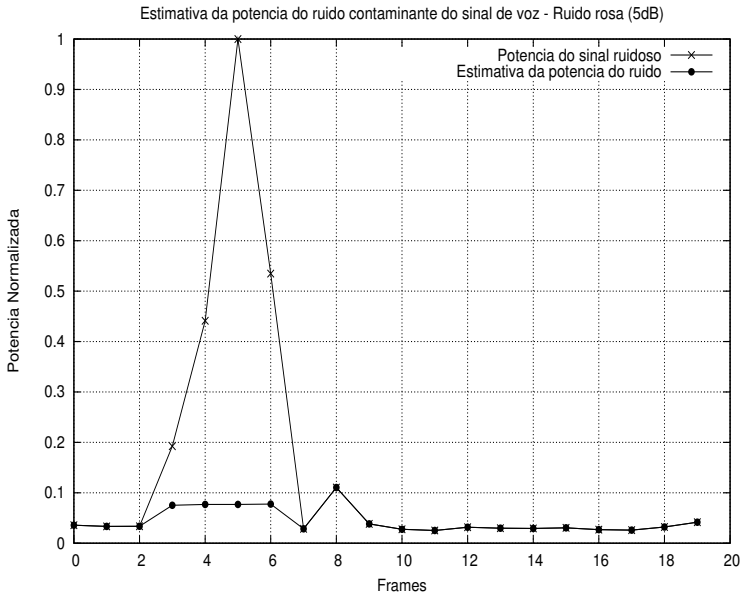


Figura 3.4: Estimativa da potência do ruído num sinal contaminado por ruído rosa.

Tabela 3.3: Algoritmo para determinação da potência do ruído e da SNR.

Passo 1 - Segmentar $x(i)$ em *frames*, $x(k,i)$;
Passo 2 - Calcular a potência do sinal ruidoso $x(k,i)$, $\sigma_x^2(k)$.
Passo 3 - Estimar a potência do ruído, $\sigma_d^2(k)$, através do filtro IIR.
a. Calcular a SNR relativa, $\Xi(k)$.
b. Calcular o parâmetro $\Upsilon(k)$.
c. Estimar a potência do ruído.
Passo 4 - Calcular a SNR do sinal.

Capítulo 4

Identificação da natureza do ruído

O processamento em sub-bandas da voz é uma técnica que vem sendo utilizada com freqüência [8, 9, 28]. Aplicado em reconhecimento, podemos observar em [9] que o processamento sub-bandas é eficiente para alguns tipos de ruídos e ineficientes para outros, dependendo também da SNR. Neste sentido, buscamos tornar o reconhecimento mais eficiente, realizando uma identificação prévia da natureza do ruído [31], bem como da sua potência (definida no Capítulo 3), tentando com isso conduzir de forma mais adequada o reconhecimento, configurando os parâmetros do SRRV, tais como: banda única ou múltibandas, banco de filtros, número de bandas, etc.

Neste capítulo apresentamos o algoritmo para identificar a natureza do ruído que contamina o sinal, baseado em HMMs discretos [32].

4.1 Metodologia

Esta técnica foi introduzida em [31], e consiste na identificação da natureza espectral do ruído através de uma classificação dos *frames* do sinal ruidoso ($x(k,i)$), por meio de HMM [1, 32, 33], entre os tipos de ruídos envolvidos no treinamento. O processo de identificação toma cada um dos *frames* e os divide em *subframes* para a extração dos parâmetros de entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Um detector de extremos robusto é apresentado em

[34] com a aplicação desta técnica.

O conceito de natureza do ruído foi introduzido em [31] e é retratado como um mapeamento espectral do ruído ao longo da freqüência.

4.2 Extração de parâmetros

A entropia espectral, $H(k)$, que está relacionada com a idéia de que um sinal tem tanto mais informação quanto maior for o seu grau de imprevisibilidade [22], é definida na equação a seguir:

$$H(k) = -\sum_{j=0}^{F-1} p(k,j) log(p(k,j)) \tag{4.1}$$

onde

$$p(k,j) = \frac{S(k,j)}{\sum_{n=0}^{F-1} S(k,n)}, \text{ para } j = 0, ..., F-1 \tag{4.2}$$

na qual $p(k,j)$ é uma função densidade de probabilidade do espectro na freqüência de índice j, estimada pela normalização sobre todas as componentes de freqüência e F é o número de componentes de freqüência. O espectro $S(k,j)$, para cada freqüência j, é obtido através da FFT.

A taxa de cruzamentos por zero para cada *frame* k, $ZC(k)$, representa uma média do número de vezes que o sinal cruza o eixo onde o valor de $x(k,i)$ é zero [1, 11]. Ela é definida na equação a seguir:

$$ZC(k) = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|sgn\{x(k,i)\} - sgn\{x(k,i-1)\}|}{2} \tag{4.3}$$

onde

$$sgn\{x(k,i)\} \quad = \quad \begin{cases} +1, & x(k,i) \geq 0 \\ -1, & x(k,i) < 0 \end{cases}$$

Os parâmetros de log-energia foram extraídos de cada *subframe* segundo um critério estabelecido, onde o espectro logarítmico do sinal é dividido em sub-bandas, e o mapeamento espectral é realizado.

O algoritmo descrito na Tabela 4.1 mostra o procedimento realizado para mapear o espectro do sinal.

Tabela 4.1: Algoritmo para extração dos parâmetros de log-energia.

Passo 1 - Calcular o espectro do sinal ruidoso, $X(k, f)$, via FFT;

Passo 2 - Calcular o espectro logarítmico, $\Psi(k, f)$;

$\Psi(k, f) = 10 \log 10 \left(\frac{X(k,f)}{f_s m} \right)$, onde $0 < f < F-1$, F é o número de componentes de freqüência, f_s é a freqüência de amostragem e m é o número de pontos usados na FFT;

Passo 3 - Estimar a envoltória do espectro logarítmico ($\Phi(k, f)$):

a. Inicialização:

$\Phi(k, 0) = \Psi(k, 0)$

$\Phi(k, F-1) = \Psi(k, F-1)$

b. Interação:

for $f = 1$ to $F - 2$ **do**

    if $\Psi(k, f) > \Psi(k, f-1)$ and

    $\Psi(k, f) > \Psi(k, f+1)$ **then**

        $\Phi(k, f) = \Psi(k, f)$

    else

        $\Phi(k, f) = 0$

    end if

end for

c. Interpolar os valores entre cada par de valores adjacentes não-nulos através do método de Newton das diferenças divididas [35].

Passo 4 - Dividir a envoltória log-espectral em P sub-bandas com F' freqüências em cada sub-banda.

Passo 5 - Extrair P parâmetros de log-energia ($LogE(k, p)$).

$LogE(k, p) = \sum_{j=0}^{F'-1} \Phi(k, (pF' + j))$, onde $0 < p < P$

Na Figura 4.1 apresentamos o espectro de potência logarítmico, juntamente com a envoltória, para um *frame* com ruído no interior do carro. Os intervalos das sub-bandas utilizadas na extração dos parâmetros de log-energia são apresentados

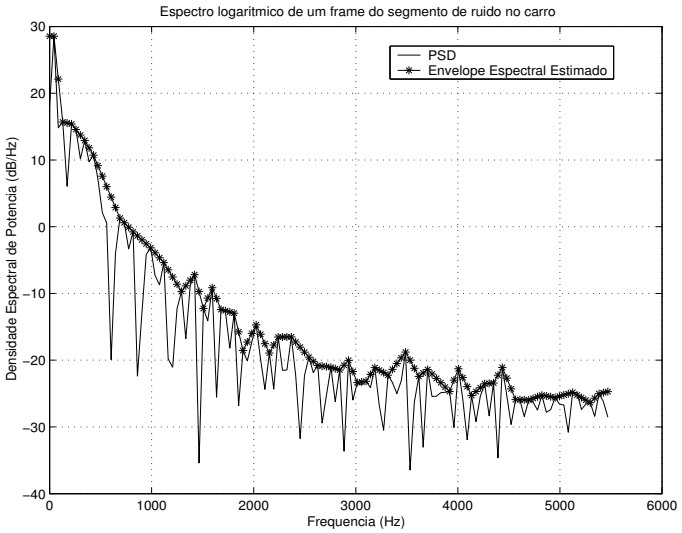


Figura 4.1: Envoltória do espectro e PSD de um *frame* com ruído no carro.

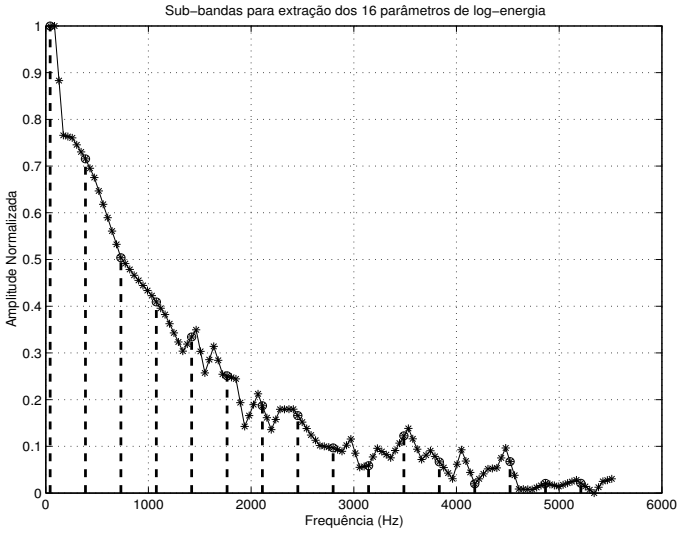


Figura 4.2: Divisão do espectro logarítmico em sub-bandas.

na Tabela 4.2 no caso de utilizarmos 16 sub-bandas ($P = 16$). Na Figura 4.2 temos uma divisão do espectro logarítmico em sub-bandas para extração dos parâmetros de log-energia. O p-ésimo parâmetro de log-energia, $LogE(k, p)$, é definido como a energia contida na p-ésima sub-banda.

Tabela 4.2: Tabela com os intervalos de freqüências para extração dos parâmetros de log-energia.

SUB-BANDA	FREQÜÊNCIA
1	$0 - 344.5$Hz
2	$344.5 - 689.0$Hz
3	$689 - 1033.5$Hz
4	$1033.5 - 1378$Hz
5	$1378 - 1722.5$Hz
6	$1722.5 - 2067$Hz
7	$2067 - 2411.5$Hz
8	$2411.5 - 2756$Hz
9	$2756 - 3100.5$Hz
10	$3100.5 - 3445$Hz
11	$3445 - 3789.5$Hz
12	$3789.5 - 4134$Hz
13	$4134 - 4478.5$Hz
14	$4478.5 - 4823$Hz
15	$4823 - 5167.5$Hz
16	$5167.5 - 5512$Hz

4.3 Configuração dos HMMs

O processo de treinamento dos HMMs na identificação da natureza do ruído segue uma extração de parâmetros, geração do codebook, quantização vetorial e treinamento dos modelos HMM. O algoritmo *Linde-Buzo-Gray* (LBG) [20] é usado para gerar o codebook, a partir dos parâmetros da base de treino, através do método "centroid splitting". A geração dos HMM é realizada por meio do algoritmo de Baum-Welch, sobre os dados oriundos de uma quantização vetorial [20].

No processo de reconhecimento existe apenas um modelo para cada tipo de ruído a ser identificado. Assim, para cada seqüência de observação O, iremos cal-

cular a probabilidade de cada modelo gerar a observação O ($P[\lambda/O]$), vencendo o modelo que apresentar a maior probabilidade. Maximizar $P[\lambda/O]$ é o mesmo que maximizar $P[O/\lambda]$, de acordo com o teorema de Bayes [1]. Dessa forma, com os parâmetros de cada segmento, sobre os quais realizamos uma quantização vetorial, onde cada conjunto de parâmetros é quantizado pelo seu índice no codebook treinado anteriormente. Essa seqüência de índices é utilizada no identificador HMM, definindo qual natureza de ruído está presente em cada *frame*.

O processo de identificação da natureza do ruído é realizado por HMM, que precisam ser treinados com amostras dos tipos de ruídos a serem identificados. Essas amostras podem ser gravadas nos locais de interesse e posteriormente segmentadas em trechos menores com tamanho fixo. As gravações podem ser tomadas de falatórios em lugares públicos, como cantinas ou restaurantes, no interior de fábricas, próximos a máquinas ou no interior de carros em movimento e com condições externas diversas. Cada um dos tipos de ruído terá os seus parâmetros usados no treinamento dos HMM's discretos de [32], com a seguinte configuração: 18 parâmetros por *subframe* de 23ms com 50% de superposição (considerando o uso de 16 sub-bandas) e aplicação da janela de Hamming, modelo do tipo esquerda-direita com possibilidade de *salto* do estado i para o estado $i+2$, quatro estados e 128 centróides.

A perturbação para determinação de novas palavras códigos no processo de "splitting" é de 0,0001 e o limiar de parada do algoritmo LBG é de 0,0000001.

4.4 Critérios de decisão

A decisão acerca da natureza do ruído é realizada partindo de locuções contaminadas e com tamanho fixo de aproximadamente dois segundos. Elas são segmentadas em trechos de 100ms com superposição de 50%. Cada um desses segmentos é submetido ao classificador e a identificação do mesmo é registrada. Em seguida, uma contagem é realizada apenas sobre os segmentos de 100 ms que pertencem à região do sinal $x(i)$, sem a presença de $s(i)$. Para isto, levamos em consideração a potência estimada do ruído no *frame* k, $\hat{\sigma}_d(k)$, realizada no Capítulo 3. Ou seja, es-

Tabela 4.3: Algoritmo para identificação da natureza do ruído.

Treinamento

Passo 1 - Montar uma sub-base de dados com K' amostras de aproximadamente 100ms de cada tipo de ruído a ser envolvido no treinamento.

Passo 2 - Cada amostra é recebida e segmentada em *subframes* de 23ms com overlap de 50%.

Passo 3 - De cada *subframe* são extraídos os 18 parâmetros para o treinamento dos HMM. Um para entropia espectral (Equação (4.1)), um da taxa de cruzamentos por zero (Equação (4.3)) e os 16 de log-energia conforme o algoritmo da Tabela 4.1.

Passo 4 - O treinamento é realizado considerando a seguinte configuração dos HMM: 500 amostras de ruído, quatro modelos, 18 parâmetros por *subframe* , modelo do tipo *esquerda-direita* com possibilidade de *salto* do estado i para o estado $i+2$, quatro estados e 128 centróides. A perturbação para determinação de novas palavras códigos no processo de "splitting" é de 0,0001 e o limiar de parada do algoritmo LBG é de 0,0000001 [32].

Identificação

Passo 1 - O algoritmo da Tabela 3.3 é utilizado até o passo 3, em cada locução de entrada, para se estabelecer a condição descrita na Subseção 4.4, ou seja, os *frames* com $\frac{\hat{\sigma}_x(k)-\hat{\sigma}_d(k)}{\hat{\sigma}_d(k)} > 2$ ficam de fora do processo de identificação.

Passo 2 - A partir deste passo, os *frames* são reorganizados de forma que passem a ser de 100ms com overlap de 50%. Os *frames* válidos são submetidos ao processo de identificação por HMM segundo o treinamento anteriormente estabelecido.

Passo 3 - Cada *frame* é submetido ao classificador. A natureza é estabelecida de acordo com aquela que recebeu o maior número de classificações entre os *frames* .

tabelecemos empiricamente uma condição para esta análise, na qual são descartados os *frames* com SNR acima de 2 ($\frac{\hat{\sigma}_x(k)-\hat{\sigma}_d(k)}{\hat{\sigma}_d(k)} > 2$).

Essa medida faz com que os *frames* localizados na região do sinal com energia mais elevada não seja computados, já que neste trecho temos $x(i)$, e não apenas o sinal de interesse $d(i)$. Este procedimento é adotado para evitar erros de classificação no caso de ocorrerem locuções muito curtas com intervalos de voz muito longos. No exemplo da Figura 3.4, os *frames* submetidos à análise são os que vão até o *frame* 2 e os que estão após o *frame* 7.

4.5 Uma aplicação em detecção de extremos.

A detecção dos extremos da voz ainda é um grande problema, principalmente em situações de reconhecimento de voz em ambiente impregnado por ruído. Enquanto os métodos tradicionais procuram detectar o início e o fim dos trechos de voz numa locução, procuramos abordar o assunto buscando delimitar os trechos onde temos apenas ruído. O método que propomos está baseado na identificação da natureza do ruído por HMM, associado à SNR e à distância euclidiana da log-energia calculados em cada *frame* . Com o método proposto a redução média da taxa de erro na detecção de início foi de até 12,23% e na detecção de fim foi de até 26,24%, para o caso de ruído rosa com SNR de 0 a 20dB, em relação ao método tradicional baseado na energia e na taxa de cruzamentos por zero [1].

Definimos que um *frame* com *flag* positiva (+1) possui somente ruído, enquanto que um *frame* com *flag* negativa (−1) possui voz misturada com ruído.

Seja o sinal ruidoso $x(i)$ uma composição do sinal de voz limpo $s(i)$ e do ruído $d(i)$, conforme descrito na Subseção 2.2.1. O sinal $x(i)$ é dividido em K *frames* com tamanho N e superposição de L amostras:

$$x(k,i) = x(k(N-L)+i) \tag{4.4}$$

onde $0 \leq k \leq K-1$ e $0 \leq i \leq N-1$.

4.5.1 Identificação da natureza do ruído

O processo de identificação da natureza do ruído é realizado através de uma classificação dos *frames* do sinal ruidoso ($x(k,i)$), por meio de HMM, entre os tipos de ruídos envolvidos no treinamento. As etapas da identificação são: extração de parâmetros (Seção 4.2), configuração do HMM (Seção 4.3), critérios de decisão (Seção 4.4).

A decisão acerca da natureza do ruído é realizada a partir de locuções contaminadas e com tamanho fixo de dois segundos. As locuções são segmentadas em trechos de 512 amostras ($\approx$ 46ms), com superposição de 80%. Os *frames* são submetidos ao classificador e a identificação dos mesmos é registrada (ruído do *frame*), de acordo com os ruídos treinados. Em seguida, uma contagem é realizada para verificar qual tipo de ruído recebeu o maior número de classificações, ou seja, qual a natureza do ruído mais presente em todo sinal (ruído do sinal). Uma comparação é realizada da seguinte forma:

$$flag_{HMM}(k) \;=\; \begin{cases} -1, & \text{se ruído do } \textit{frame} \neq \text{ruído do sinal} \\ +1, & \text{se ruído do } \textit{frame} = \text{ruído do sinal} \end{cases}$$

onde $flag_{HMM}(k)$ é a *flag* atribuída ao k-ésimo *frame* através do critério de classificação por HMM.

4.5.2 Determinação da SNR dos *frames*

Através do algoritmo da Tabela 3.3, calculamos a SNR dos *frames* e definimos as seguintes condições:

$$flag_{SNR}(k) \;=\; \begin{cases} -1, & \text{se } \sigma_x^2(k)/\sigma_d^2(k) \geq th1 \\ +1, & \text{se } \sigma_x^2(k)/\sigma_d^2(k) < th1 \end{cases}$$

onde $th1$ recebeu empiricamente o valor de 1,25 se $SNR_{est} < 10$dB e 2,50 se $SNR_{est} \geq$ 10dB. SNR_{est} é a estimativa da SNR de todo o sinal, calculada como:

$$SNR_{est} = 10 \log 10 \left(\frac{\sum_{k=0}^{K'-1} \hat{\sigma}_x^2(k) - \sum_{k=0}^{K'-1} \hat{\sigma}_d^2(k)}{\sum_{k=0}^{K'-1} \hat{\sigma}_d^2(k)} \right) \tag{4.5}$$

onde $\hat{\sigma}_x^2(k)$ e $\hat{\sigma}_d^2(k)$ são calculados de forma semelhante às equações (3.13) e (3.14), respectivamente, porém sem superposição dos K' *frames*, ou seja, $x(k,i) = x(kN+i)$, para $0 \leq k \leq K'-1$.

$flag_{SNR}(k)$ é a *flag* atribuída ao k-ésimo *frame* através do critério da SNR.

4.5.3 Determinação da relação euclidiana da log-energia

A relação euclidiana é estabelecida entre as distâncias euclidianas do sinal e do ruído. A distância euclidiana do sinal $x(i)$, $ED_x(k)$, é determinada da seguinte forma:

$$ED_x(k) = \sqrt{\sum_{p=0}^{P-1}(LogE(k,p) - LogE_{ref}(p))^2} \tag{4.6}$$

onde:

$$LogE_{ref}(p) = \frac{1}{m}\sum_{k=0}^{m-1} LogE(k,p) \tag{4.7}$$

A distância euclidiana da log-energia do ruído, ED_d, é tomada como a $ED_x(k)$ máxima nos m *frames* iniciais do sinal, que neste trabalho foi igual a 15:

$$ED_d = \max_{0 \leq k \leq m-1}\{ED_x(k)\} \tag{4.8}$$

Uma relação entre as distâncias euclidianas em cada *frame* é estabelecida como $\frac{ED_x(k)}{ED_d}$, onde:

$$flag_{ED}(k) \quad = \quad \begin{cases} -1, & \text{se } ED_x(k)/ED_d \geq th2 \\ +1, & \text{se } ED_x(k)/ED_d < th2 \end{cases}$$

onde $th2$ recebeu empiricamente o valor de 1,4. $flag_{ED}(k)$ é a *flag* atribuída ao k-ésimo *frame* através do critério da relação euclidiana.

4.5.4 Delimitação do sinal

Após todos os *frames* do sinal receberem três *flags* cada um (por tipo de ruído, por razão de variâncias e por razão de distâncias euclidianas da log-energia), os *frames* são novamente rotulados, recebendo uma única *flag*, $flag_{global}(k)$, baseada nas seguintes condições:

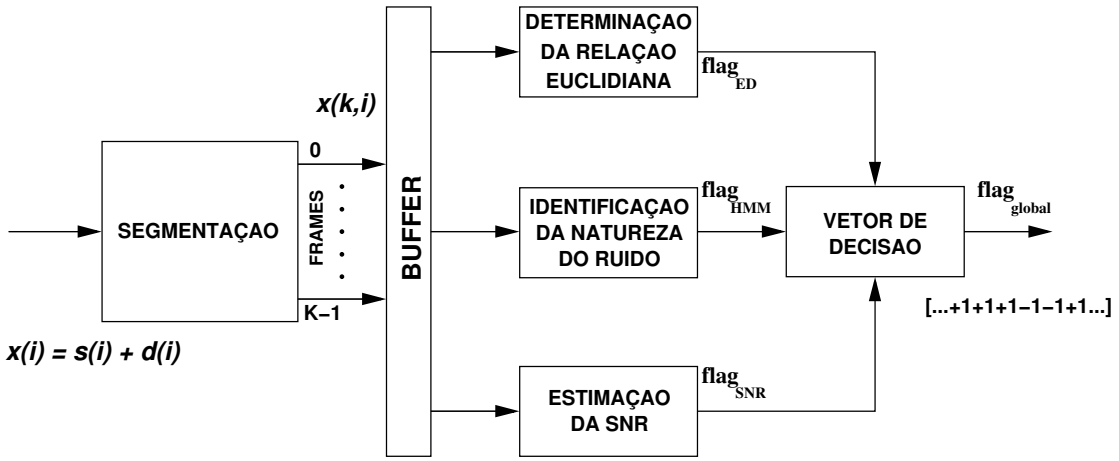


Figura 4.3: Método de detecção de extremos proposto com aplicação da identificação da natureza do ruído.

if $flag_{HMM}(k) = flag_{SNR}(k) = flag_{ED}(k) = +1$ **then**

$flag_{global}(k) = +1$

else

$flag_{global}(k) = -1$

end if

Em seguida, as *flags* dos *frames* são analisadas a partir do *frame* 0 em direção ao último, buscando a primeira seqüência de 15 *frames* com *flag* negativa, onde o primeiro *frame* da seqüência caracteriza o término de um possível trecho inicial contendo apenas ruído. De forma semelhante, partindo do último *frame* em direção ao *frame* 0, uma seqüência de 15 *frames* com *flag* negativa também é pesquisada, onde o último *frame* da seqüência caracteriza o início de um possível trecho final contendo apenas ruído.

4.5.5 Resultados obtidos

A detecção de extremos pode ter a sua performance avaliada de duas formas: uma delas é comparar os resultados obtidos na detecção com valores de referência obtidos de um recorte manual. A outra é passar as locuções recortadas, após a contaminação, em um sistema de reconhecimento de voz, comparando os seus resultados. Optamos pela primeira forma por termos condições de avaliar de forma direta os

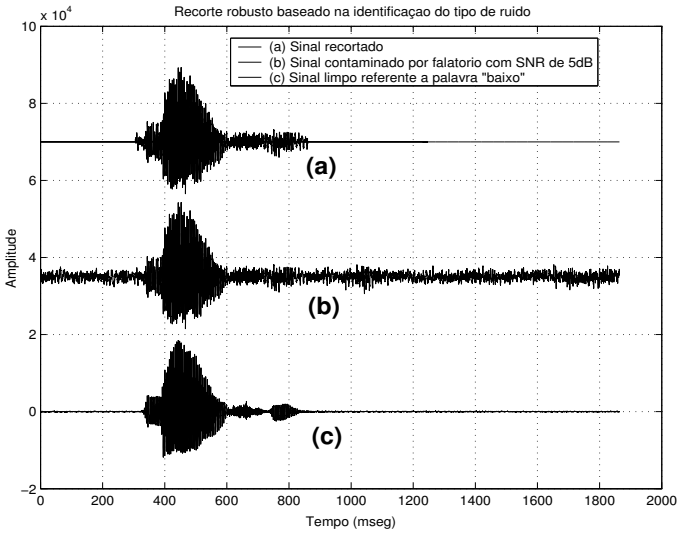


Figura 4.4: Recorte robusto baseado na identificação do tipo de ruído. (a) Sinal recortado pelo método proposto. (b) Sinal contaminado por falatório com SNR de 5dB e (c) Sinal limpo referente a palavra "baixo".

resultados, fazendo uma comparação com outros métodos.

Em relação ao sinal da Figura 4.4, a Figura 4.5 mostra as relações de variâncias e distâncias euclidianas para cada *frame*. Observamos na Figura 4.4(c), trechos de voz com baixa energia. Podemos notar na Figura 4.5, que se apenas a SNR for considerada, aqueles trechos não seriam incluídos na delimitação. Com o uso da distância euclidiana da log-energia, esse trecho é identificado como voz misturada com ruído. O trecho pode ser observado na Figura 4.4 entre 600ms e 800ms, correspondendo, aproximadamente, ao intervalo entre os *frames* 62 e 80 da Figura 4.5.

Inicialmente foi feito um treinamento com 500 segmentos de 100ms de cada um dos quatro tipos de ruídos modelados para realizarmos a identificação da natureza do ruído. Em seguida foi realizado um teste funcional com sinais de voz contaminados pelos ruídos selecionados com uma SNR que vai de 0 a 20dB. O procedimento foi fixar uma SNR e o ruído, observando o percentual de erro na detecção ao longo dos vários valores de SNR, tanto para detecção de início como para detecção de fim. Para as 121 locuções inseridas no detector, calculamos o erro percentual na detecção de início (ε_i) e o erro percentual na detecção de fim (ε_f) em relação ao

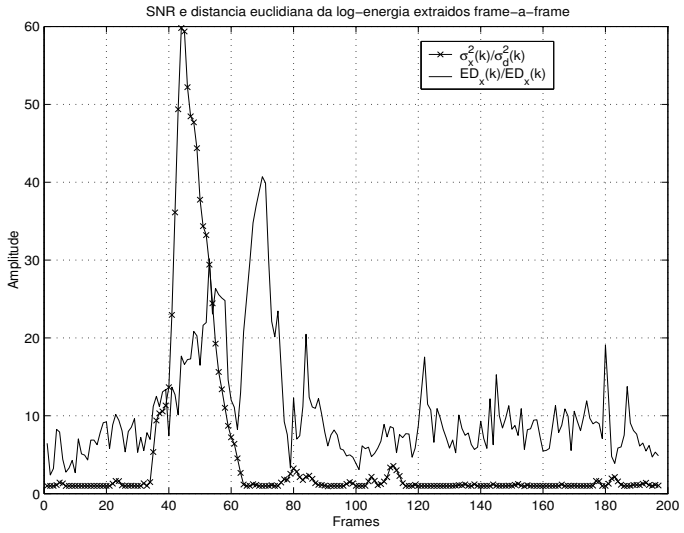


Figura 4.5: Razão das variâncias e das distâncias euclidianas para cada *frame* da palavra "baixo", contaminada por falatório com SNR de 5dB, ilustrada na Figura 4.4.

recorte manual previamente realizado, da seguinte forma:

$$\varepsilon_i = \frac{|I - e_i|}{F - I} \times 100\% \tag{4.9}$$

$$\varepsilon_f = \frac{|F - e_f|}{F - I} \times 100\% \tag{4.10}$$

onde I e F são, respectivamente, os pontos de início e fim do recorte manual, e_i e e_f são, respectivamente, os pontos detectados.

Para os quatro tipos de ruídos, a redução média da taxa de erro na detecção, comparativamente ao método de [11], é mostrada na Tabela 4.4. Os valores calculados são uma média das taxas de erro de detecção de início e fim, para cada tipo de ruído, com os cinco valores de SNR simulados neste trabalho. Podemos observar que não houve uma redução efetiva da taxa de erro no caso de ruído no interior do carro, justificado pelo resultado negativo, possivelmente por se tratar de um tipo de ruído com potência concentrada em freqüências muito baixas. Para as demais situações, a redução foi considerável.

Os resultados individuais, que foram alcançados nos testes, são comparados com o método de [11] e mostrados nos gráficos das Figuras 4.6, 4.7, 4.8 e 4.9, respectivamente, para falatório, ruído rosa, ruído no interior do carro e ruído branco. É notado também que a melhor performance do detector ocorre nas faixas de baixa

Tabela 4.4: Tabela com os percentuais médios de redução da taxa de erro obtidos com o detector robusto em locuções contaminadas por ruído com SNR variando entre 0 e 20dB.

Redução média do erro de detecção				
	BABBLE	PINK	VOLVO	WHITE
Início	6,82%	12,23%	−2,65%	12,05%
Fim	20,81%	26,24%	0,70%	24,88%

SNR, situação em que a contaminação mais compromete o processamento do sinal.

4.5.6 Conclusões

A aplicação proposta é um método de recorte robusto baseado na identificação da natureza do ruído por HMM, associado à SNR e à distância euclidiana da log-energia, que foram utilizadas para ajustar a detecção em trechos de voz com baixa energia. Com este método buscamos detectar os *frames* que possuem somente ruído e separá-los da informação útil do sinal. Os resultados de um teste com 121 locuções contaminadas por quatro tipos de ruídos (branco, falatório, rosa e interior do carro) com SNR de 0 a 20dB foi apresentado. A redução média da taxa de erro no recorte em relação ao método de [11], foi de até 12,23% na detecção de início e de até 26,24% na detecção de fim, no caso de ruído rosa. Nas demais situações de contaminação os resultados foram mais modestos, apresentando grandes oscilações na performance, apenas quando processado com ruído no interior do carro. Como um trabalho futuro, pretendemos tornar a detecção de extremos mais eficiente em ambientes com ruídos de baixa freqüência, semelhantes ao ruído no interior do carro.

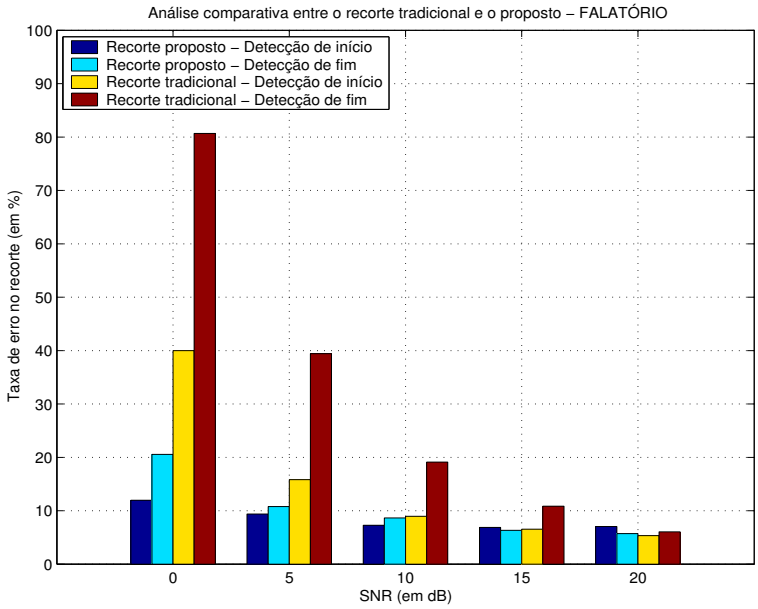


Figura 4.6: Performance do detector de extremos: análise na condição de contaminação por falatório.

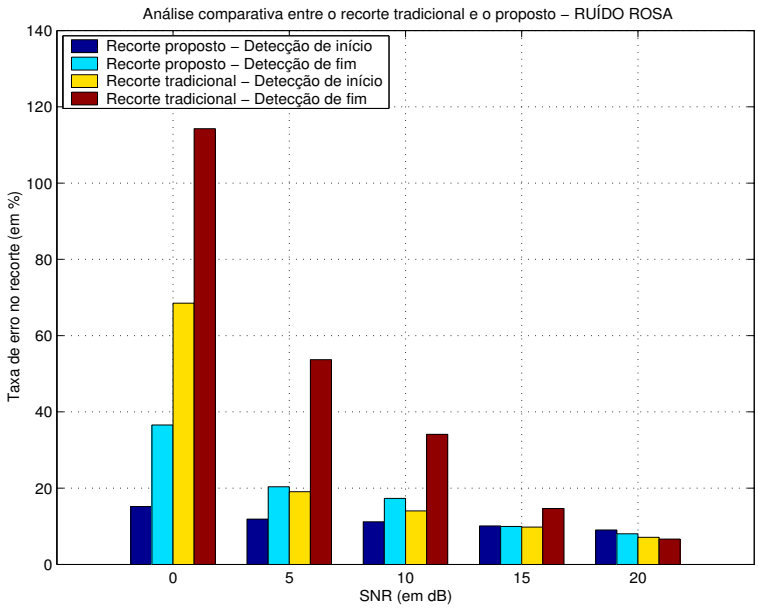


Figura 4.7: Performance do detector de extremos: análise na condição de contaminação por ruído rosa.

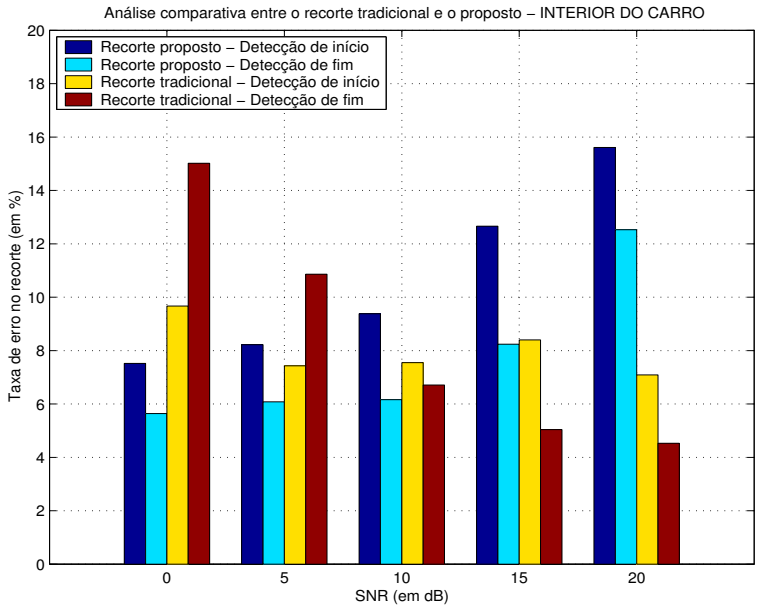


Figura 4.8: Performance do detector de extremos: análise na condição de contaminação por ruído no interior do carro.

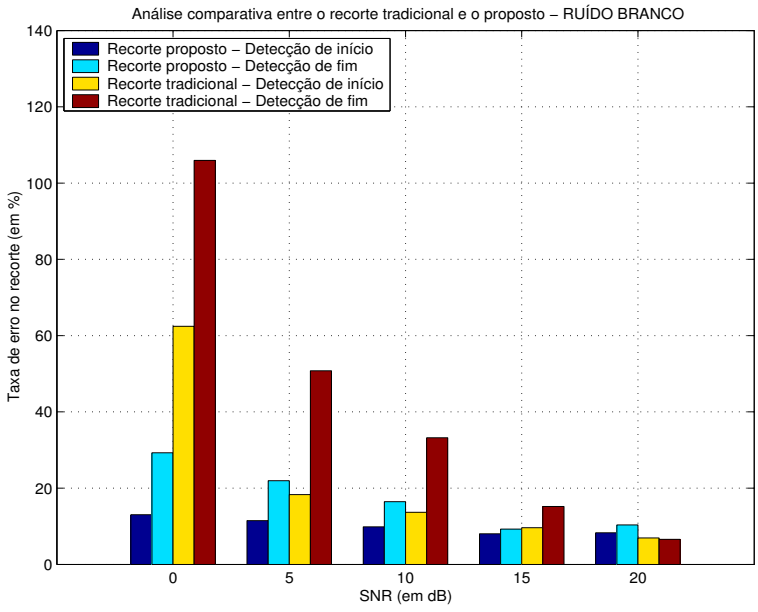


Figura 4.9: Performance do detector de extremos: análise na condição de contaminação por ruído branco.

Capítulo 5

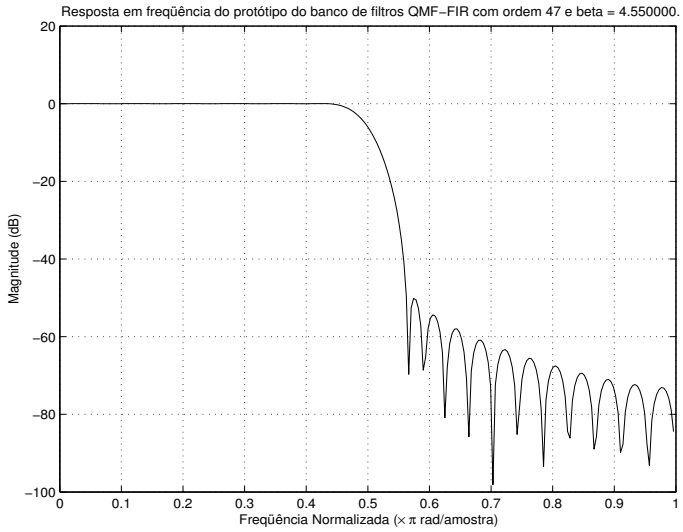
Reconhecimento robusto de voz usando multi-bandas

Neste capítulo descrevemos o processamento em múltiplas bandas, base para a tarefa de reconhecimento. Na Seção 5.1, fazemos uma breve descrição da teoria dos bancos de filtros utilizados nesta tese. Na Seção 5.2, apresentamos o sistema multi-bandas utilizado no processamento, bem como descrevemos os seus blocos, tais como redução do nível de ruído (Subseção 5.2.1), detecção de extremos robusta (5.2.2), divisão em sub-bandas (5.2.3) e, por fim, um algoritmo de decisão em sub-bandas (5.2.4).

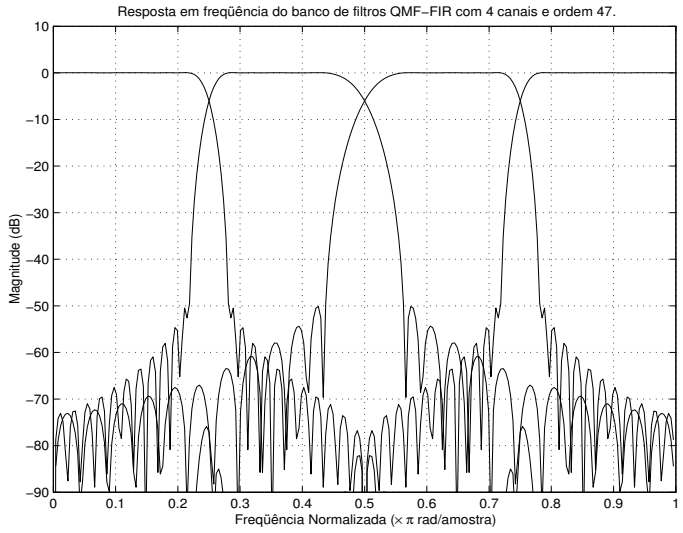
5.1 Banco de filtros digitais

Este trabalho está focado na divisão em multi-bandas através de BFD com Filtro de Quadratura de Espelho (QMF - *Quadrature Mirror Filter*) Resposta ao Impulso Finita (FIR - *Finite Impulse Response*) (com 4 e 8 bandas uniformes e 3 e 4 bandas não-uniformes) e modulado por cosseno com 4 e 8 bandas uniformes, ilustrados no Apêndices B.

Os bancos de filtros foram construídos a partir de um filtro protótipo FIR implementado pelo método do janelamento, usando a janela de Kaiser. Fixamos uma atenuação da banda de rejeição, A_r, de $\approx$ 50dB e um número N de coeficientes



(a) Protótipo QMF-FIR.



(b) Banco de filtros QMF-FIR de 4 canais

Figura 5.1: Banco de filtros QMF-FIR com 4 canais com protótipo de 48 coeficientes e $\beta = 4.55$.

igual a 48 (Figura 5.1) [36].

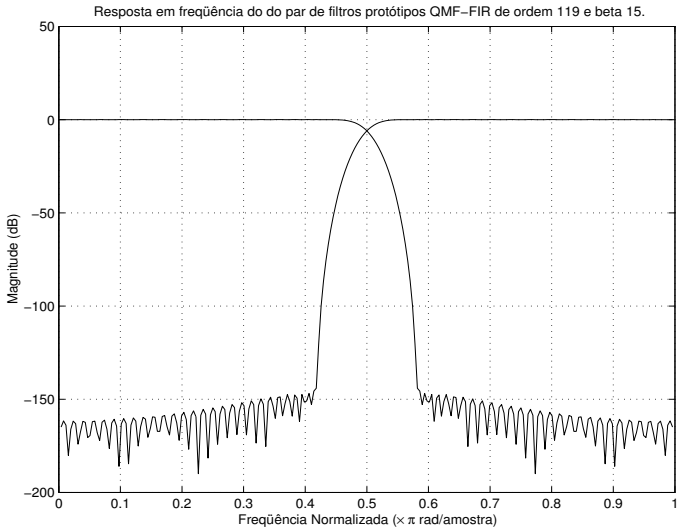


Figura 5.2: Par de filtros protótipos FIR.

5.1.1 QMF-FIR

O par de filtros QMF-FIR, como o mostrado na Figura 5.2, possui a consideração de linearidade de fase e comprimento N par. Para $H_0(z)$ e $H_1(z)$ da Figura 5.3, componentes da análise do sinal, temos:

$$h_0(i) = h_0(N - i - 1) \text{ para } 0 \leq i \leq N/2 - 1 \tag{5.1}$$

$$h_1(i) = (-1)^i h_0(i) \tag{5.2}$$

$$h_1(i) = -h_1(N - i - 1) \text{ para } 0 \leq i \leq N/2 - 1 \tag{5.3}$$

$$|H_0(\omega)|^2 + |H_1(\omega)|^2 = 1 \text{ para } 0 \leq \omega \leq \pi \tag{5.4}$$

$$|H_0(\omega)|^2 + |H_0(\pi - \omega)|^2 = 1 \tag{5.5}$$

Cada sub-banda do sinal $x(i)$ é gerada pela aplicação do BFD QMF-FIR. Considerando M canais, a estrutura equivalente para cada ramo é convertida em um único filtro, resultante da aplicação das operações de identidade $H_1^{(k)}(z) = H_0^{(k)}(-z)$ [37, 38].

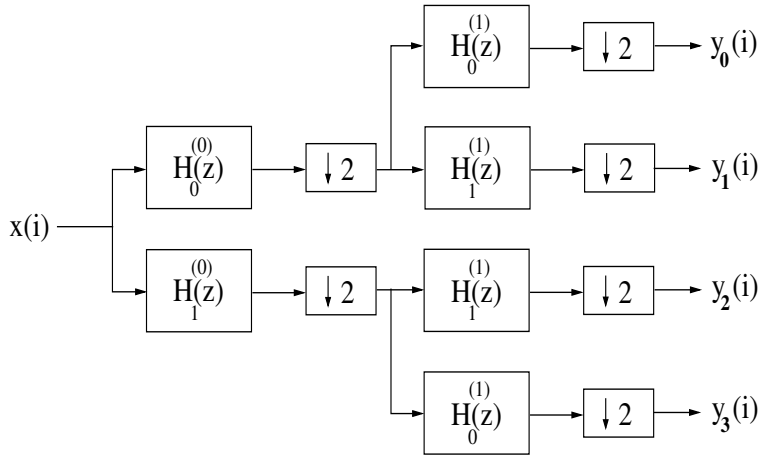


Figura 5.3: Banco de filtros QMF-FIR.

5.1.2 Modulado por cosseno

O BFD modulado por cosseno é uma boa opção, pois possui a vantagem de um projeto simples, baseado apenas na geração de um filtro protótipo passa-baixa $P_0(z)$ e, também possui um baixo custo de implementação em termos de multiplicações na modulação do protótipo, para a geração de cada sub-filtro. Ou seja, temos aqui um banco de filtros baseado na modulação por cosseno, onde todos os M canais são derivados da modulação do filtro protótipo $P_0(z)$ [37].

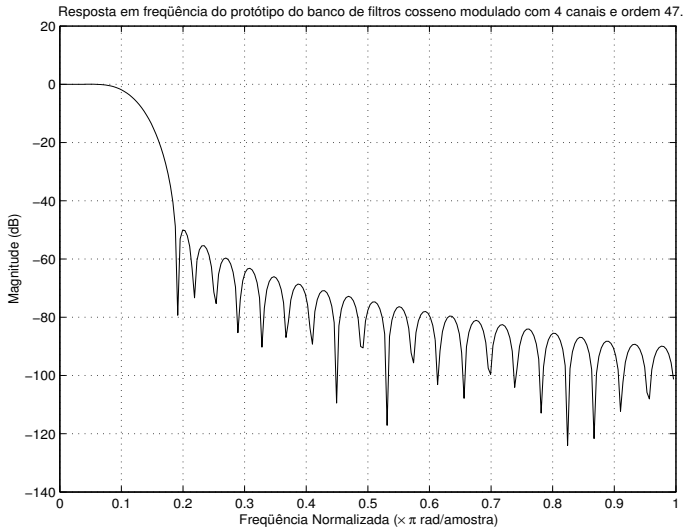
O desenvolvimento é baseado na busca de um filtro protótipo passa-baixa $P_0(z)$ de ordem $N-1$ e freqüência de corte estabelecida em $\omega_c = \frac{\pi}{2M}$. Cada sub-filtro é dado por:

$$h_k(i) = 2p_0(i)cos\left((2k+1)\frac{\pi}{2M}\left(i - \frac{N}{2}\right) + (-1)^k\frac{\pi}{4}\right) \tag{5.6}$$

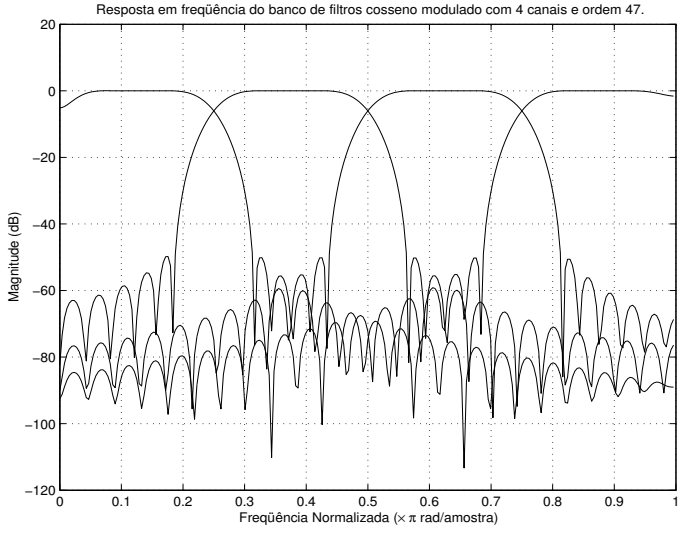
onde N é o número de coeficientes do filtro protótipo $(p_0(i))$, M é o número de sub-bandas e $k = 0, ..., M-1$ e $i = 0, ..., N-1$.

5.2 Sistema de reconhecimento em multi-bandas

A seguir descrevemos como o reconhecimento em múltiplas bandas (Figura 5.5) é realizado, procurando detalhar alguns blocos que compõem o processo.



(a) Protótipo modulado por cosseno.



(b) Banco de filtros modulado por cosseno.

Figura 5.4: Banco de filtros modulado por cosseno de 4 canais com protótipo de 48 coeficientes e $\beta = 4.55$.

5.2.1 Redução do nível de ruído

Antes do sinal ruidoso $x(i)$ ser submetido ao processo de reconhecimento através de múltiplas bandas, realizamos uma redução do nível de ruído do sinal aplicando uma das técnicas apresentadas no Capítulo 3, subtração espectral

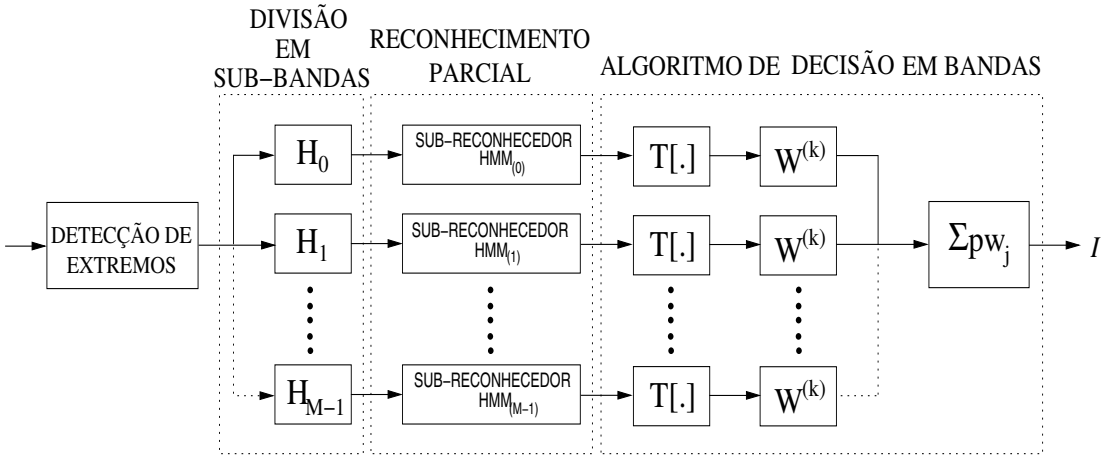


Figura 5.5: Reconhecimento de voz usando multi-bandas.

(Subseção 3.1.1) e filtragem de Wiener (Subseção 3.1.2).

5.2.2 Detecção de extremos

A detecção de extremos utilizada é apresentada em [12] e é aplicada tanto no processamento em banda única como no processamento em múltiplas bandas.

Este método é adotado no processamento do sinal, apesar do método proposto na Seção 4.5 atingir níveis consideráveis na sua performance. Esta decisão é tomada devido a existência de várias locuções na base de dados originalmente contaminadas com um tipo de ruído convolucional, inserido no ato da gravação por um microfone de baixa qualidade. Para manter a performance daquele detector de extremos, seria preciso um número considerável de amostras daquele tipo de ruído para treinar os HMMs, o que não estava disponível.

Baseado no conhecimento de que a base de dados foi gravada com um intervalo inicial sem a presença da voz e assumindo aqui o estado da fala já com ruído adicionado, foi feita uma segmentação do sinal em *frames* com uma duração de aproximadamente 23ms e com overlap de 30%. Foram usados os 150 ms iniciais como referência para o intervalo de silêncio.

São calculados sobre todos os *frames* do sinal o seguinte grupo de estatísticas: a energia do k-ésimo *frame* ($\mathcal{E}_{sp}^{(k)}$) e a distância cepstral do k-ésimo *frame* em relação

à média ($d_{sp}^{(k)}$).

$$\mathcal{E}_{sp}^{(k)} = \frac{1}{N}\sum_{i=0}^{N-1}[x(k,i)]^2 \tag{5.7}$$

$$C(n) = \frac{1}{K}\sum_{k=0}^{K-1} c^{(k)}(n) \tag{5.8}$$

$$d_{sp}^{(k)} = \sum_{n=0}^{p-1}(c^{(k)}(n) - C(n))^2 \tag{5.9}$$

onde N é o tamanho do *frame*, K é o número de *frames*, $c^{(k)}(n)$ é o n-ésimo parâmetro cepstral do k-ésimo *frame*, $n = 0, ..., p-1$ e p é a dimensão da análise cepstral.

Um outro grupo de grandezas estatísticas é também calculado, porém sobre o intervalo assumido como silêncio: a energia média ($\mathcal{E}_{sil}$), o vetor cepstral médio (C_{sil}) e a distância cepstral média (d_{sil}).

$$\mathcal{E}_{sil} = \frac{1}{m}\sum_{k=0}^{m-1}\mathcal{E}_{sp}^{(k)} \tag{5.10}$$

$$C_{sil}(n) = \frac{1}{m}\sum_{j=0}^{m-1} c^{(j)}(n) \tag{5.11}$$

$$d_{sil} = \sum_{n=0}^{p-1}(c^{(j)}(n) - C_{sil}(n))^2 \tag{5.12}$$

para $j = 0, ..., m-1$

O *frame* é considerado com voz se pelo menos uma das condições a seguir ocorrerem [39]:

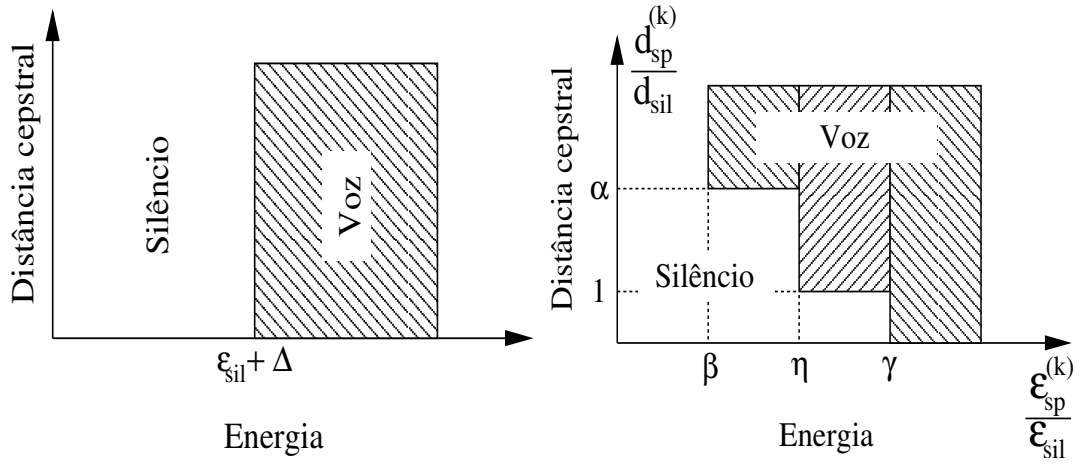
$$\text{Condição I: } \frac{d_{sp}^{(k)}}{d_{sil}} > \alpha \text{ e } \frac{\mathcal{E}_{sp}^{(k)}}{\mathcal{E}_{sil}} > \beta \tag{5.13}$$

$$\text{Condição II: } \frac{d_{sp}^{(k)}}{d_{sil}} > 1 \text{ e } \frac{\mathcal{E}_{sp}^{(k)}}{\mathcal{E}_{sil}} > \eta \tag{5.14}$$

$$\text{Condição III: } \frac{\mathcal{E}_{sp}^{(k)}}{\mathcal{E}_{sil}} > \gamma \tag{5.15}$$

A Condição I é direcionada ao problema da fala com baixa energia na presença de ruído. A condição estabelecida requer que haja uma elevada relação da distância cepstral para o *frame* ser classificado como contendo voz.

A Condição II é direcionada aos erros de classificação dos sons com energia elevada que não são voz. De forma semelhante à condição anterior, é requerida



(a) Detecção de extremos não-robusta (b) Detecção de extremos robusta

Figura 5.6: Detecção de extremos robusta considerando a relação de energia e distância cepstral.

uma relação da distância cepstral acima da unidade, mesmo que o trecho contenha energia elevada, para o *frame* ser classificado como contendo voz.

Por fim, a Condição III, permite que sons com energia muito elevada sejam classificados como contendo voz, para qualquer valor da sua relação da distância cepstral.

A Figura 5.6, mostra estas condições através de regiões hachuradas delimitadas pelos parâmetros α, β, η e γ, que foram determinados empiricamente como 4.0, 0.9, 1.25 e 2.25, respectivamente. O número de coeficientes cepstrais por *frame* utilizado foi de 14. A detecção de início e fim foi experimentalmente definida quando uma seqüência de no mínimo 10 *frames* classificados como contendo fala for encontrada, seja a partir do início do sinal, para detectar o início de voz, como a partir do final, para detecção do término da locução.

Resultados comparativos deste método com o método tradicional da energia e da taxa de cruzamentos por zero implementado em [11], são apresentados no Capítulo 6 (gráficos) e no Apêndice C (tabelas).

Tabela 5.1: Algoritmo para detecção de extremos robusta.

Passo 1 - Segmentar o sinal ruidoso $x(i)$;
Passo 2 - Calcular as estatísticas do sinal: a energia ($\mathcal{E}_{sp}^{(k)}$) e a distância cepstral ($d_{sp}^{(k)}$);
Passo 3 - Calcular as estatísticas do intervalo assumido como silêncio: a energia média ($\mathcal{E}_{sil}$), o vetor cepstral médio (C_{sil}) e a distância cepstral média (d_{sil});
Passo 4 - Verificar as condições definidas nas Equações 5.13, 5.14 e 5.15;
Passo 5 - Definir os extremos.

5.2.3 Processo de divisão em sub-bandas

O processo de divisão em sub-bandas é baseado na aplicação dos bancos de filtros apresentados no Apêndice B, segundo a suas variáveis. Cada sub-banda é treinada e testada independentemente, não necessitando portanto de reconstrução do sinal através de banco de síntese.

Os sub-reconhecedores HMMs ($\text{HMM}_{(0)}$, $\text{HMM}_{(1)}$, ..., $\text{HMM}_{(M-1)}$), ilustrados na Figura 5.5, são configurados para promoverem o treinamento das sub-bandas das locuções de forma semelhante ao apresentado na Seção 4.3, exceto pela extração de 28 parâmetros por *frame* (1 energia, 13 rasta-plp, 1 delta-energia e 13 delta-rasta) [1, 40] de 20ms com 65% de superposição e 10 estados.

5.2.4 Algoritmo de decisão em sub-bandas

Cada sub-reconhecedor promove um reconhecimento parcial ao longo do processamento em sub-bandas, mas faz-se necessário que haja uma forma de extrairmos um resultado global do reconhecimento. Para isto, apresentamos a seguir um algoritmo para decisão em sub-bandas que combina as verossimilhanças parciais dos sub-reconhecedores para a identificação de um modelo global [12].

Baseado nos resultados do algoritmo de Viterbi em cada sub-reconhecedor $\text{HMM}_{(j)}$, temos na j-ésima sub-banda, a probabilidade de se gerar a seqüência de observações (O) dado o i-ésimo modelo ($\lambda^{(i)}$), ou seja, $P\left(O/\lambda^{(i)}\right)_j$, onde $i = 0, \ldots, Q-1$

e $j = 0, ..., M-1$. Q é o número de modelos e M é o número de sub-bandas.

Seja

$$\vec{P}_j = \left[P\left(O/\lambda^{(0)}\right)_j \;\; P\left(O/\lambda^{(1)}\right)_j \ldots P\left(O/\lambda^{(Q-1)}\right)_j \right]^T, \tag{5.16}$$

o vetor contendo todas as probabilidades dos Q modelos na banda j e a matriz diagonal $W = diag[w_0 \; w_1 \ldots w_{Q-1}]$, onde $w_k = \frac{Q-k}{Q\sum_{n=0}^{Q-1} w_n}$, para $k = 0, \ldots, Q-1$.

Aplicando-se uma transformação $T[\cdot]$ sobre $\vec{P}_j$ de forma que os seus elementos sejam ordenados segundo os seus valores de probabilidade, chegamos ao vetor

$$\vec{P}_j^{(k)} = T\left[\vec{P}_j\right] \tag{5.17}$$

onde

$$\vec{P}_j^{(k)} = \left[p_j^{(k)}\right] \tag{5.18}$$

e

$$p_j^{(k)} = P\left(O/\lambda^{(i)}\right)_j^{(k)} \tag{5.19}$$

onde $k = 0, ..., Q-1$. Assim,

$$\vec{P}_j^{(k)} = \left[\min\left\{P(O/\lambda^{(i)})_j\right\} \;\; \ldots \;\; \max\left\{P(O/\lambda^{(i)})_j\right\}\right]^T \tag{5.20}$$

para $0 \leq i \leq Q-1$ e $0 \leq j \leq M-1$. O índice k guarda a nova posição da probabilidade $P\left(O/\lambda^{(i)}\right)_j^{(k)}$ no vetor $\vec{P}_j^{(k)}$ para cada modelo i na sub-banda j.

Aplicando-se W sobre $\vec{P}_j^{(k)}$, estaremos dando mais ênfase às probabilidades parciais com maior verossimilhança em cada sub-banda. Sendo assim, temos:

$$\vec{Pw}_j = \left[\vec{P}_j^{(k)}\right]^T W = \left[pw_j^{(i)}\right] \tag{5.21}$$

onde $pw_j^{(i)}$ é a probabilidade do i-ésimo modelo na banda j após a ponderação realizada por W.

A decisão é feita através de:

$$I = \max_{0 \leq i \leq Q-1} \left\{ \sum_{j=0}^{M-1} pw_j^{(i)} \right\} \tag{5.22}$$

sendo I o modelo selecionado a partir das parciais em cada sub-banda.

Tabela 5.2: Algoritmo de decisão em sub-bandas.

Passo 1 - Formar o vetor $\vec{P}_j$ com todas as probabilidades dos Q modelos $\lambda^{(i)}$ na banda j.
Passo 2 - Gerar a matriz $W = diag[w_0\ w_1 \ldots w_{Q-1}]$, onde $w_k = \frac{Q-k}{Q\sum_{n=0}^{Q-1} w_n}$, para $k = 0, \ldots, Q-1$.
Passo 3 - Ordenar de forma crescente as probabilidades de $\vec{P}_j$ em cada sub-banda, aplicando $T[\cdot]$.
Passo 4 - Aplicar W sobre $\vec{P}_j^{(k)}$, gerando $\vec{Pw}_j$.
Passo 5 - Decidir pelo modelo, sendo I o modelo selecionado a partir das parciais em cada sub-banda, onde $I = \max_{0 \leq i \leq Q-1} \left\{ \sum_j pw_j^{(i)} \right\}$.

A Tabela 5.2 descreve o algoritmo para realizar a decisão global das sub-bandas.

O procedimento apresentado no algoritmo da Tabela 5.2 proporciona o fortalecimento das verossimilhanças maiores das sub-bandas, já que, com exceção do ruído branco, todos os outros apresentam uma sub-banda com uma quantidade reduzida de ruído e conseqüentemente uma verossimilhança maior com o modelo correspondente.

Capítulo 6

Resultados experimentais

Neste capítulo descrevemos as bases de dados utilizadas e apresentamos os resultados experimentais dos métodos de reconhecimento em banda única e em múltiplas bandas e do teste de identificação da natureza do ruído. Na Seção 6.1 apresentamos as bases de dados utilizadas, tanto de voz como de ruído, especificando os ruídos que foram selecionados para simulação dos ambientes ruidosos. Em seguida, na Seção 6.2, apresentamos os resultados obtidos nos testes de verificação da identificação da natureza do ruído, bem como os testes de performance do reconhecimento em banda única e em múltiplas bandas.

6.1 Base de dados

Para este trabalho foram simulados quatro ambientes ruidosos onde a locução a ser reconhecida estava sendo pronunciada. A locução ruidosa foi gerada a partir de locuções limpas, sobre as quais o ruído é adicionado, com uma SNR especificada, segundo o algoritmo descrito no Apêndice A.

6.1.1 Locuções

As locuções utilizadas nesta tese são as mesmas da base de dados mencionada em [32], onde temos 10 palavras isoladas, pronunciadas em português do Brasil (ANDA, BAIXO, CIMA, DESLIGA, DIREITA, ESQUERDA, FRENTE, MÃO,

OLHA e TRÁS), repetidas 5 vezes cada uma por 20 locutores diferentes, sendo 15 utilizadas no treinamento e 5 utilizadas no teste, dando um total de 1000 locuções. A taxa de amostragem é de 11025Hz.

6.1.2 Ruídos

Os ruídos utilizados foram adquiridos da *Signal Processing Information Base* (SPIB) [41], para formar distintos ambientes ruidosos. A citada base de dados, disponibilizada pela Universidade de Rice, através da internet, na página do grupo de Processamento de Sinais Digitais, é indicada em [42], juntamente com outras origens que possuem exemplos da base de dados NOISEX-92.

A SPIB disponibiliza exemplos dos seguintes tipos de ruídos:

- Branco (WHITE);
- Rosa (PINK);
- Canal de rádio de alta freqüência (HFCHANNEL);
- Falatório (BABBLE);
- Ruído de fábrica 1 (FACTORY1);
- Ruído de fábrica 2 (FACTORY2);
- Cockpit de caça 1 (BUCCANEER1);
- Cockpit de caça 2 (BUCCANEER2);
- Sala de máquinas do Destroyer (DESTROYERENGINE);
- Sala de operações do Destroyer (DESTROYEROPS);
- Cockpit do f-16 (F-16);
- Veículo militar (LEOPARD);
- Tanque militar (M109);

- Metralhadora (MACHINEGUN);
- Interior do carro (VOLVO)

Para o presente trabalho foram selecionados quatro tipos de ruídos para formação dos ambientes ruidosos de simulação: WHITE, PINK, BABBLE e VOLVO.

1. WHITE - Foi obtido de um gerador analógico de ruídos.
2. PINK - Também adqüirido de um gerador de ruídos, porém com uma distribuição desigual da potência, ou seja, existe uma atenuação de -3dB por oitava.
3. BABBLE - É um falatório gravado em uma cantina com 100 pessoas conversando. Pelo fato do raio do local ser superior a dois metros, a percepção de vozes individuais fica inaudível.
4. VOLVO - É uma gravação feita no interior de um Volvo 340 a 120 km/h, em quarta marcha, numa pista de asfalto, em condições de chuva.

Os dois primeiros tipos de ruído foram selecionados por serem tradicionais em tarefas de processamento robusto. Os demais foram selecionados por representarem situações reais de operação em condições adversas.

As locuções ruidosas foram formadas através da adição dos ruídos selecionados às locuções "limpas" de acordo com SNRs estabelecidas, que vão de -5 a 25dB, conforme descrito no Apêndice A.

A base de dados de ruído utilizada possui sinais com 235s e freqüência de amostragem de 19980Hz. Todos foram sub-amostrados para 11025Hz, sendo ajustados para composição do sinal ruidoso nos testes. Outro ponto é que as locuções possuem em média 2s de duração. Para compatibilizar a duração do ruído com as locuções no ato da confecção do sinal ruidoso, o sinal de ruído sofreu recortes com o mesmo tamanho das locuções, porém em pontos aleatórios ao longo de toda a sua duração, tornando cada contaminação diferente, aproximando-se de um ambiente de operação real.

Exemplos com a PSD de segmentos dos ruídos utilizados nesta tese são apresentados nas Figuras 6.1, 6.2, 6.3 e 6.4.

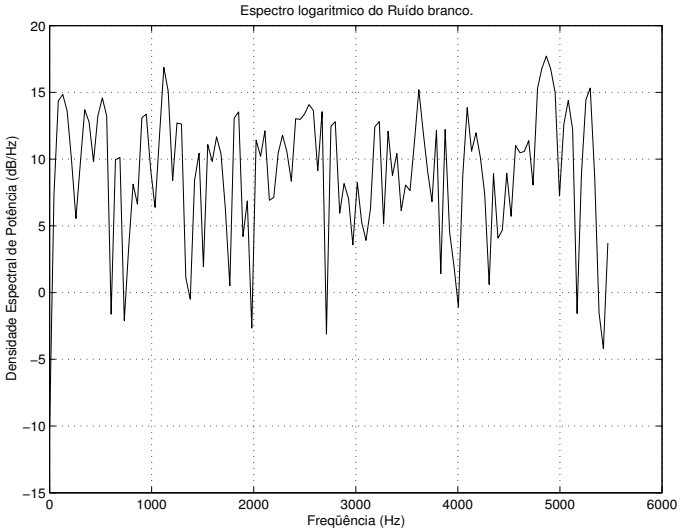


Figura 6.1: Densidade Espectral de Potência de segmento do ruído branco usado neste trabalho.

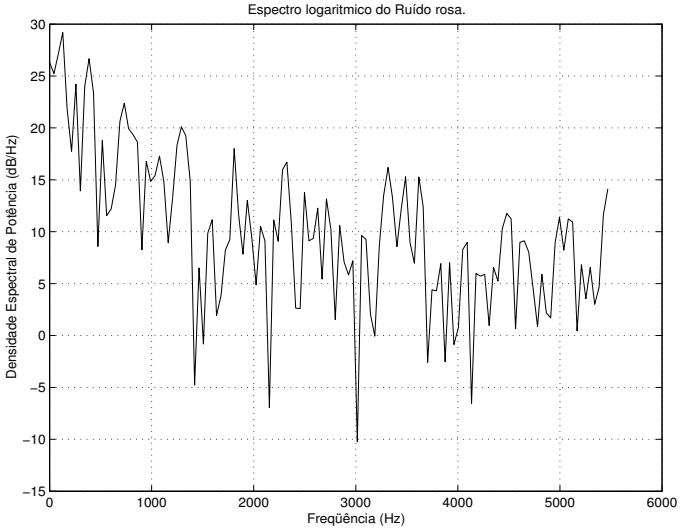


Figura 6.2: Densidade Espectral de Potência de segmento do ruído rosa usado neste trabalho.

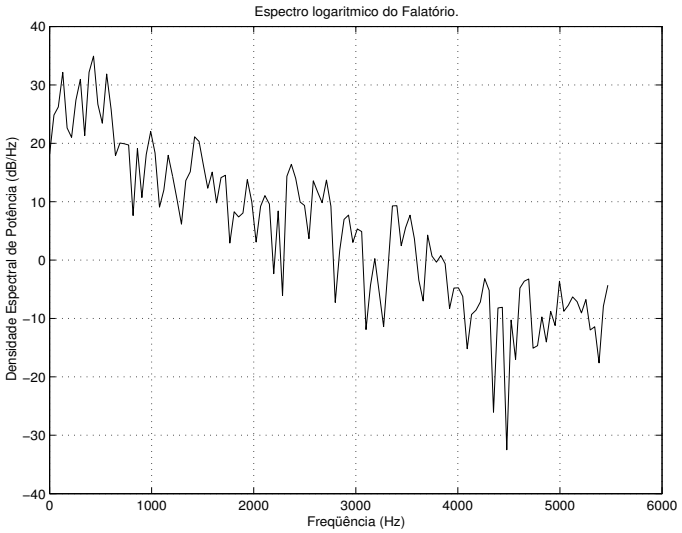


Figura 6.3: Densidade Espectral de Potência de segmento do falatório usado neste trabalho.

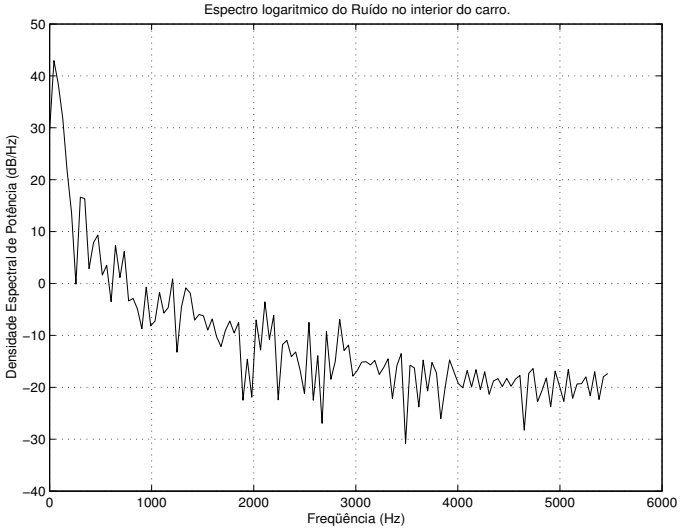


Figura 6.4: Densidade Espectral de Potência de segmento do ruído no interior do carro usado neste trabalho.

6.2 Resultados

Nesta seção apresentamos os resultados obtidos nos testes dos algoritmos utilizados nesta tese. Como uma análise estatística dos resultados não foi realizada,

os valores podem apresentar alguma imprecisão. Mesmo assim, podemos ter uma idéia das performances dos algoritmos através das análises realizadas.

6.2.1 Teste de identificação da natureza do ruído

Inicialmente foi feito um treinamento com 500 segmentos de 100ms de cada um dos quatro tipos de ruídos utilizados neste trabalho, conforme o algoritmo descrito na Seção 4.1. Em seguida, foram feitos dois testes de performance, buscando medir a eficiência na identificação da natureza dos ruídos. O primeiro teste foi realizado com ruído, onde foram inseridos 100 segmentos de cada um dos tipos envolvidos, diferentes dos usados no treino, com 100ms cada. Foram extraídos 18 parâmetros de cada sub-frame, com 23ms e superposição de 50%. Neste teste a performance atingiu $100,00\%$. Os demais resultados são apresentados na Tabela C.5 (Apêndice C).

Em seguida foi realizado um teste funcional com sinais de voz contaminados pelos ruídos selecionados com SNRs que vão de 0dB a 20dB. O procedimento foi fixar uma SNR e o ruído, observando o acerto ao longo das várias SNRs. Foram inseridas 121 locuções no sistema, conforme Tabela 6.1. Após contaminadas pelos ruídos, as locuções foram segmentadas em trechos de 100ms e analisadas através de HMM. A performance atingiu $97,22\%$. Os resultados obtidos são mostrados na Tabela C.6 (Apêndice C). A Figura 6.5 mostra esses resultados de forma mais clara, onde observamos o êxito maior da identificação, nas condições de contaminação crítica, onde um processamento robusto é mais necessário.

6.2.2 Teste comparativo da detecção de extremos

O detector de extremos descrito no Capítulo 5 é comparado com o detector implementado em [11], num reconhecedor em múltiplas bandas conforme descrito no Apêndice D. A Figura 6.6 mostra as diferenças na taxa de acerto ao utilizarmos o recorte robusto e o recorte tradicional baseado na energia e na taxa de cruzamentos por zeros. Pode ser observado que o detector robusto apresenta uma vantagem razoável de cerca de 32% acima dos resultados obtidos com a detecção tradicional,

Tabela 6.1: Locuções utilizadas no teste de verificação da natureza do ruído.

LOCUÇÃO	REPETIÇÕES
ANDA	16
BAIXO	19
CIMA	10
DESLIGA	13
DIREITA	6
ESQUERDA	10
FRENTE	7
MÃO	11
OLHA	14
TRÁS	15

para o caso de ruído rosa, na faixa de 5dB. As maiores diferenças foram registradas nas SNRs mais baixas, onde o processamento robusto é ainda mais necessário. Alguns pontos apresentaram resultados negativos, mostrando que a detecção robusta não prevaleceu sobre o detecção tradicional.

6.2.3 Teste de verificação do desempenho

Nesta subseção apresentamos os resultados oriundos de uma análise comparativa entre o processamento em banda única e o processamento em múltiplas bandas. A idéia é verificar, para cada tipo de ruído, qual forma de processamento é mais eficiente e quais os parâmetros que utilizados resultam numa melhor performance.

Foram realizados treinamentos e testes nas seguintes situações:

1. **BUSR** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em banda única, sem realce de voz;

2. **BUSE** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em banda única, com subtração espectral;

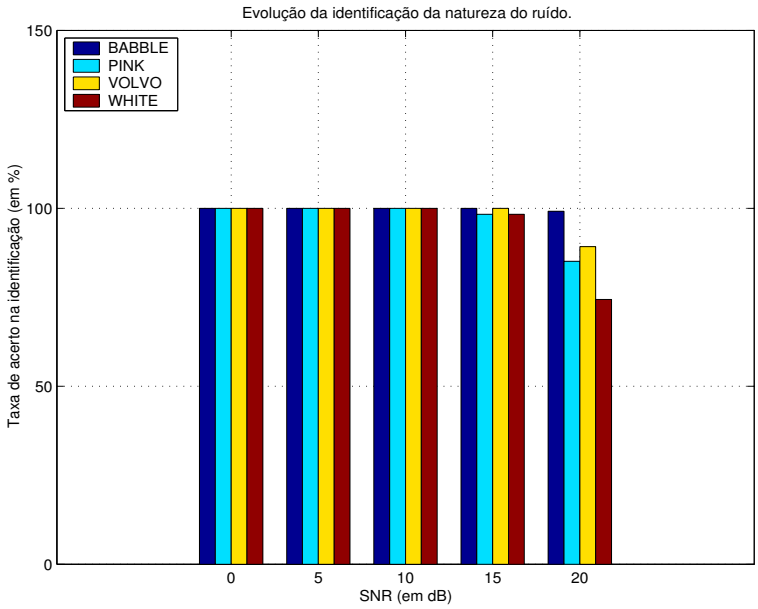


Figura 6.5: Resultados do teste de identificação da natureza do ruído em locuções.

3. **BUFW** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em banda única, com filtragem de Wiener;

4. **MBSECOS4U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros modulado por cosseno com 4 bandas uniformes, com subtração espectral;

5. **MBSECOS8U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros modulado por cosseno com 8 bandas uniformes, com subtração espectral;

6. **MBSEQMF4U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 4 bandas uniformes, com subtração espectral;

7. **MBSEQMF8U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 8 bandas uniformes, com subtração espectral;

8. **MBSEQMF3N** - Treinamento com locuções limpas e teste com locuções

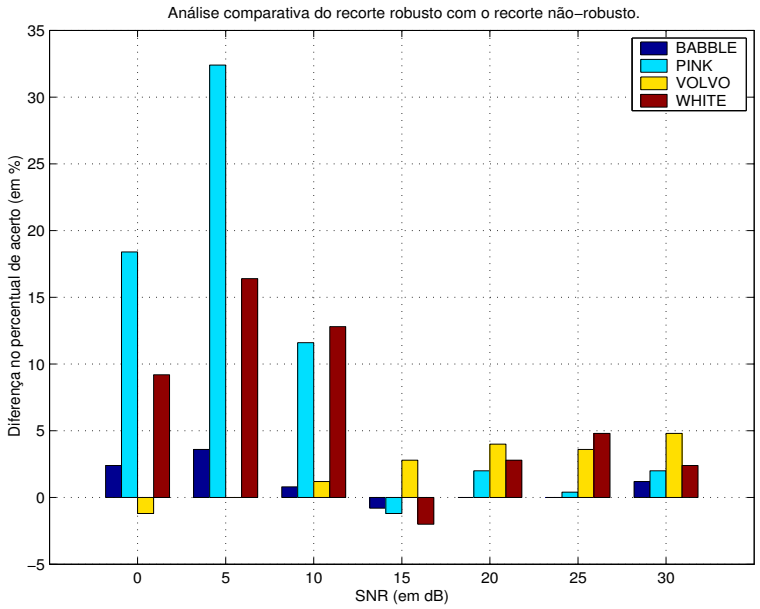


Figura 6.6: Análise comparativa do recorte robusto com o recorte tradicional.

ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 3 bandas não-uniformes, com subtração espectral;

9. **MBSEQMF4N** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 4 bandas não-uniformes, com subtração espectral;

10. **MBFWCOS4U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros modulado por cosseno com 4 bandas uniformes, com filtragem de Wiener;

11. **MBFWCOS8U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros modulado por cosseno com 8 bandas uniformes, com filtragem de Wiener;

12. **MBFWQMF4U** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 4 bandas uniformes, com filtragem de Wiener;

13. **MBFWQMF8U** - Treinamento com locuções limpas e teste com locuções

ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 8 bandas uniformes, com filtragem de Wiener;

14. **MBFWQMF3N** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 3 bandas não-uniformes, com filtragem de Wiener;

15. **MBFWQMF4N** - Treinamento com locuções limpas e teste com locuções ruidosas, nas SNRs especificadas, em múltiplas bandas com banco de filtros QMF-FIR com 4 bandas não-uniformes, com filtragem de Wiener;

Os resultados referentes aos ruídos simulados são apresentados nas Tabelas do Apêndice C e ilustrados nas Figuras 6.7 até 6.10, para os casos de aplicação da subtração espectral e filtragem de Wiener em cada ambiente ruidoso simulado.

Para o caso de contaminação por ruído branco, a subtração espectral apresentou uma melhor performance comparada com a filtragem de Wiener, tanto para banda única como para múltiplas bandas. Comparadas em processamentos distintos o método em múltiplas bandas com subtração espectral superou o método em banda única em até $8,4\%$ e o método em múltiplas bandas com filtragem de Wiener em até $21,2\%$, com o banco de filtros **MBSECOS8U**. O método em banda única prevaleceu ligeiramente sobre o método em múltiplas bandas apenas na faixa de 10dB.

Em contaminação por ruído rosa, a subtração espectral também apresentou um melhor desempenho comparada com a filtragem de Wiener, em banda única e em múltiplas bandas. O método em múltiplas bandas com subtração espectral superou o método em banda única em até $12,4\%$ e o método em múltiplas bandas com filtragem de Wiener em até $24,8\%$, também com o banco de filtros **MBSECOS8U**. Nesta condição de contaminação o método em múltiplas bandas prevaleceu sobre banda única sempre.

Em ambiente impregnado por ruído no carro, os índices foram os mais elevados obtidos até então, porém apresentaram uma oscilação que foi acompanhada em forma tanto por banda única como por múltiplas bandas. A subtração espec-

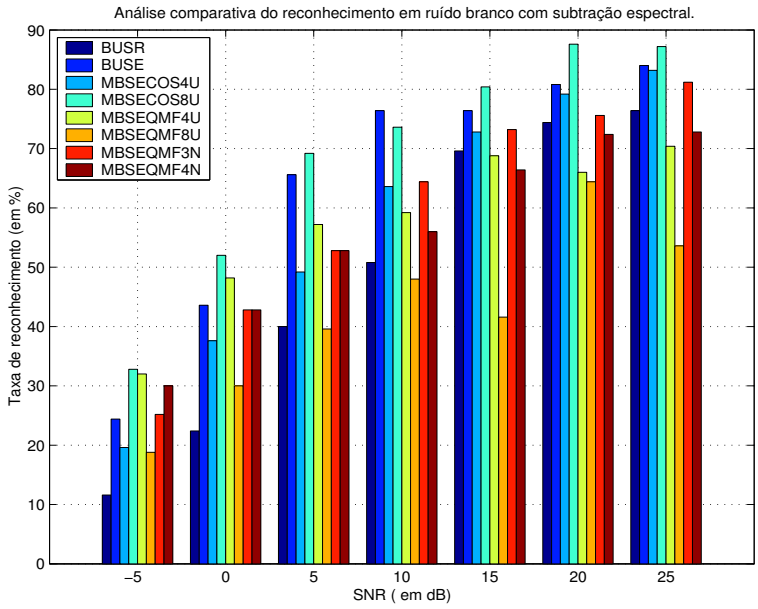
tral apresentou uma melhor desempenho comparada com a filtragem de Wiener, em banda única, apenas em algumas faixas de contaminação e em múltiplas bandas até a faixa de 15dB. O método em múltiplas bandas com subtração espectral superou o método em banda única em até $7,2\%$ e o método em múltiplas bandas com filtragem de Wiener em até $23,2\%$, igualmente para o caso do banco de filtros **MBSECOS8U**. Nesta condição de contaminação o método em múltiplas bandas também prevaleceu sobre o método em banda única sempre.

Por fim, em ambiente contaminado por falatório, a subtração espectral apresentou uma melhor desempenho comparada com a filtragem de Wiener, em banda única e em múltiplas bandas até a faixa de 20dB. O método em múltiplas bandas com subtração espectral superou o método em banda única em até $14,4\%$ e o método em múltiplas bandas com filtragem de Wiener em até $15,6\%$, novamente para o caso do banco de filtros **MBSECOS8U**. Nesta condição de contaminação o método em múltiplas bandas também prevaleceu sobre o método em bandas única sempre.

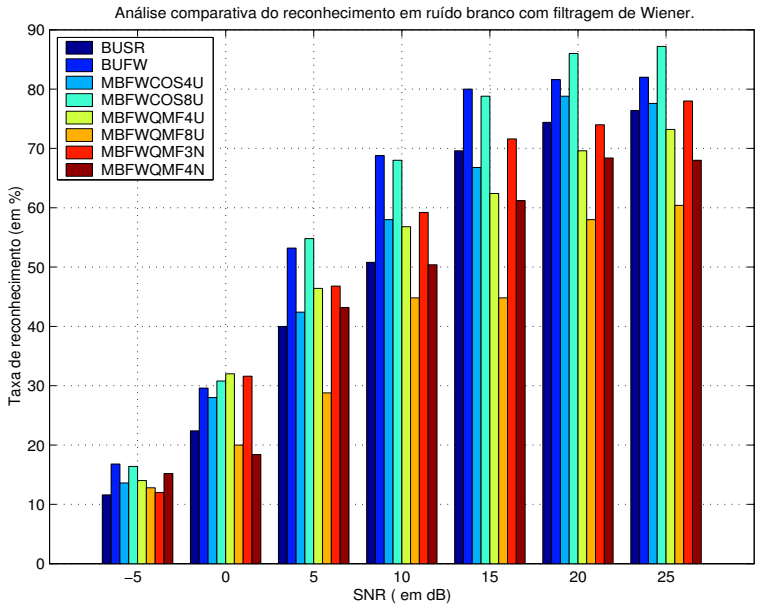
6.2.4 Performance com algoritmo proposto

Neste teste, antes de realizarmos o reconhecimento, extraímos de cada sinal de voz contaminado, dois parâmetros que configuram o SRRV: a natureza do ruído e a SNR do sinal. Através destes parâmetros o SRRV será configurado de forma a alcançar melhores índices de acerto. Desta forma, o reconhecimento pode ser em banda única ou múltiplas bandas, a redução de ruído pode ser por subtração espectral ou por filtragem de Wiener, além de podermos variar o número de subbandas e o tipo de banco de filtros, caso o reconhecimento seja realizado em múltiplas bandas.

Analisando a performance com o algoritmo proposto, podemos conduzir adequadamente o processamento, com o conhecimento da natureza do ruído e sua potência, como ilustrado na Figura 6.11. Isto pode ser observado nas Figuras 6.12 e 6.13, onde a linha cheia do gráfico segue pelo melhor processamento em cada SNR e no respectivo ambiente ruidoso. Optando, assim, por banda única ou múltiplas bandas, além dos seus parâmetros. No caso de contaminação por ruído branco, por



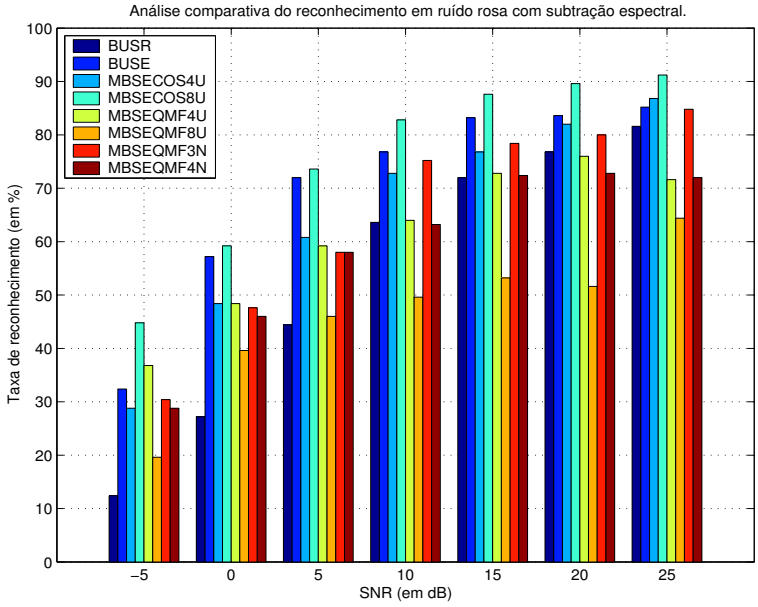
(a) Subtração espectral.



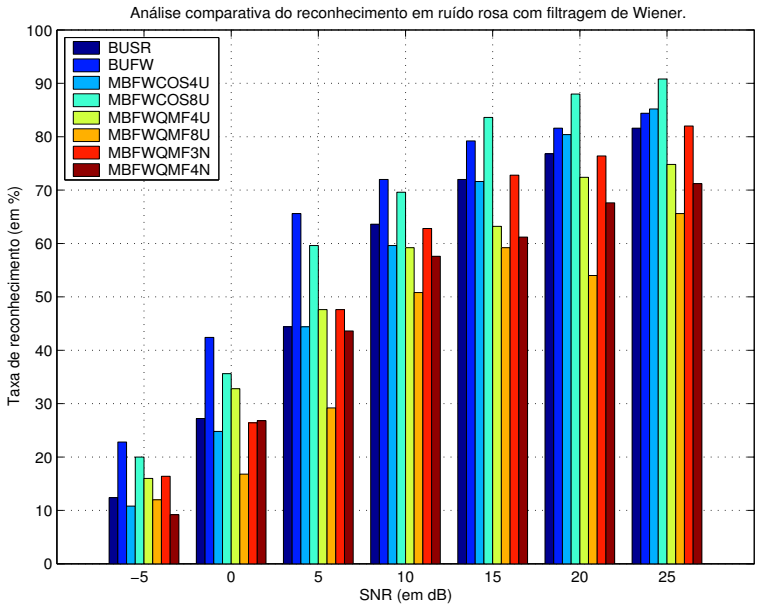
(b) Filtragem de Wiener.

Figura 6.7: Resultados do teste de verificação de desempenho em contaminação por ruído branco.

exemplo, temos uma inversão de método em 10dB, onde o reconhecimento passa a ser favorecido por banda única.

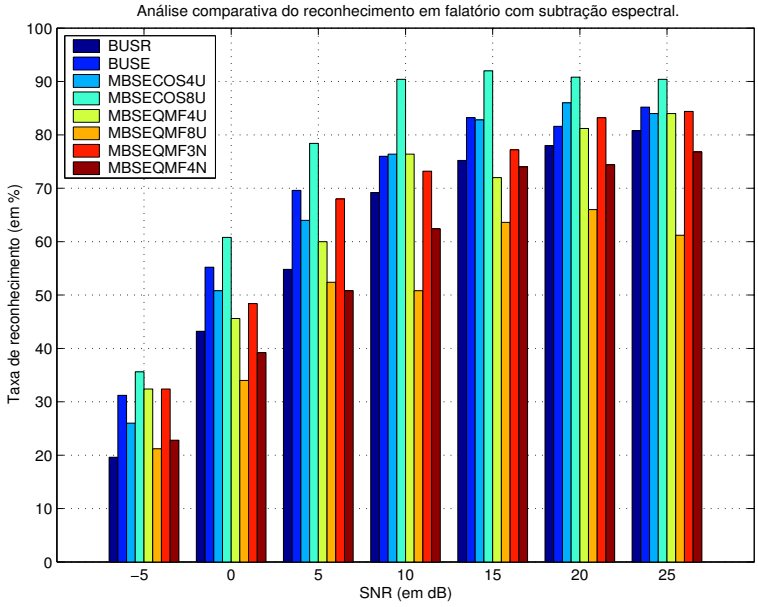


(a) Subtração espectral.

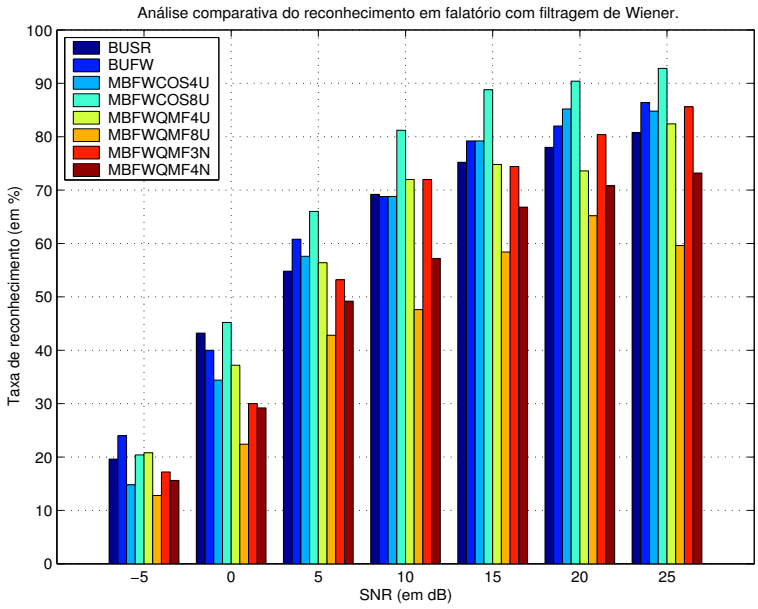


(b) Filtragem de Wiener.

Figura 6.8: Resultados do teste de verificação de desempenho em contaminação por ruído rosa.

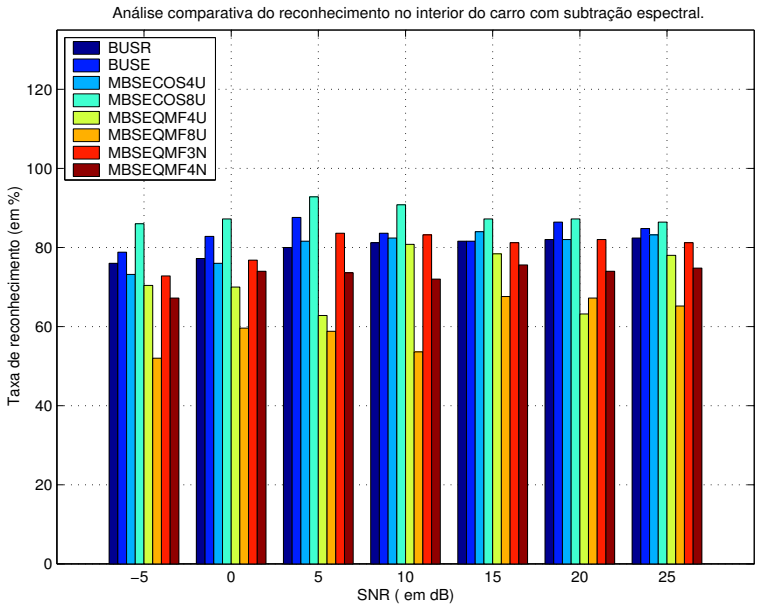


(a) Subtração espectral.

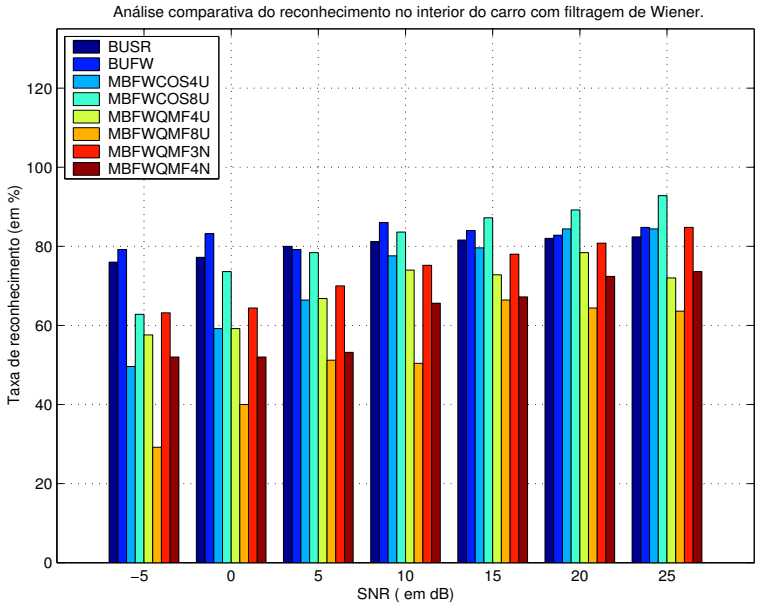


(b) Filtragem de Wiener.

Figura 6.9: Resultados do teste de verificação de desempenho em contaminação por falatório.

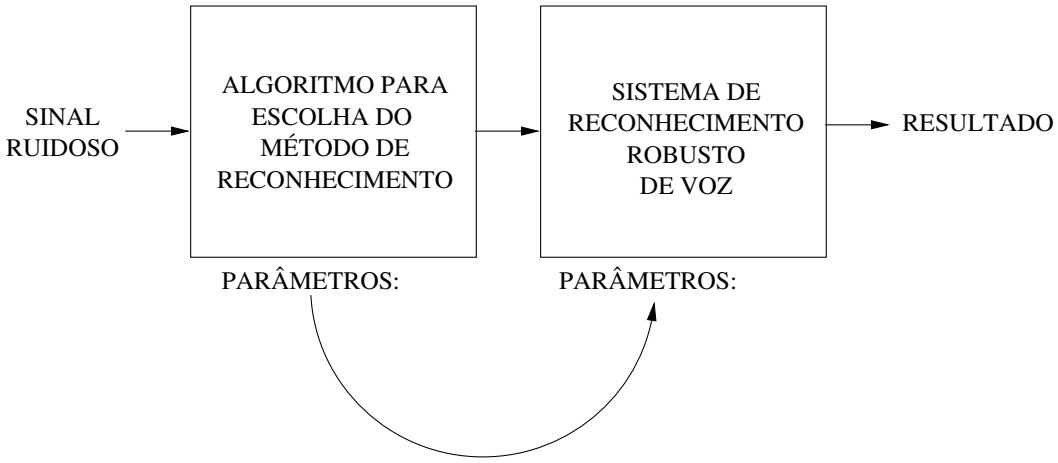


(a) Subtração espectral.

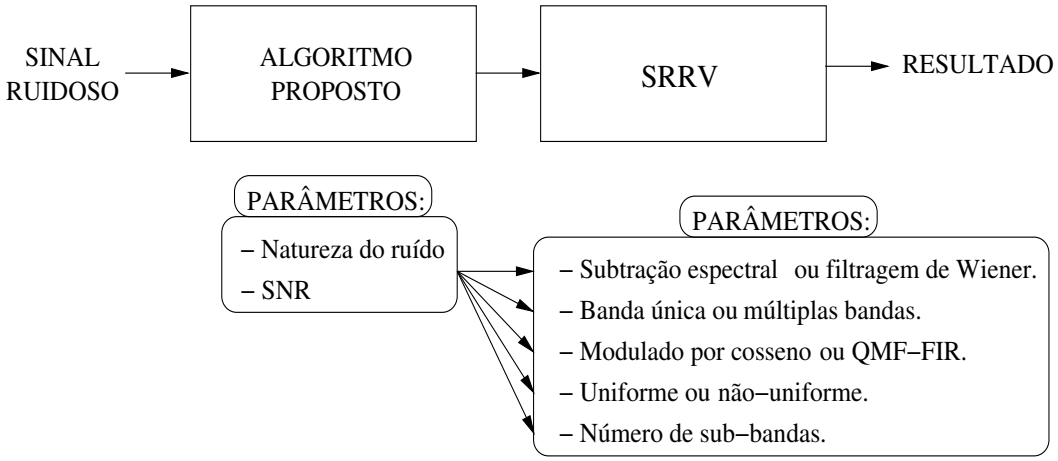


(b) Filtragem de Wiener.

Figura 6.10: Resultados do teste de verificação de desempenho em contaminação por ruído no interior do carro.

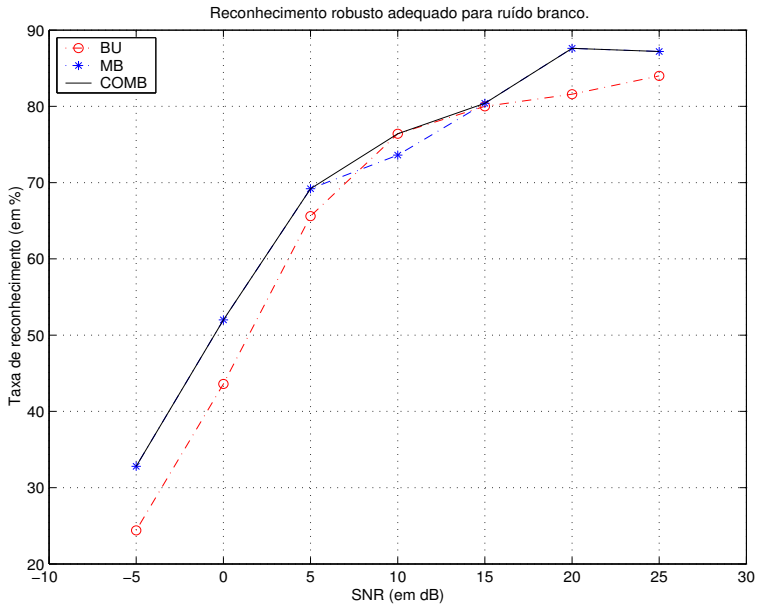


(a) Esquema geral.

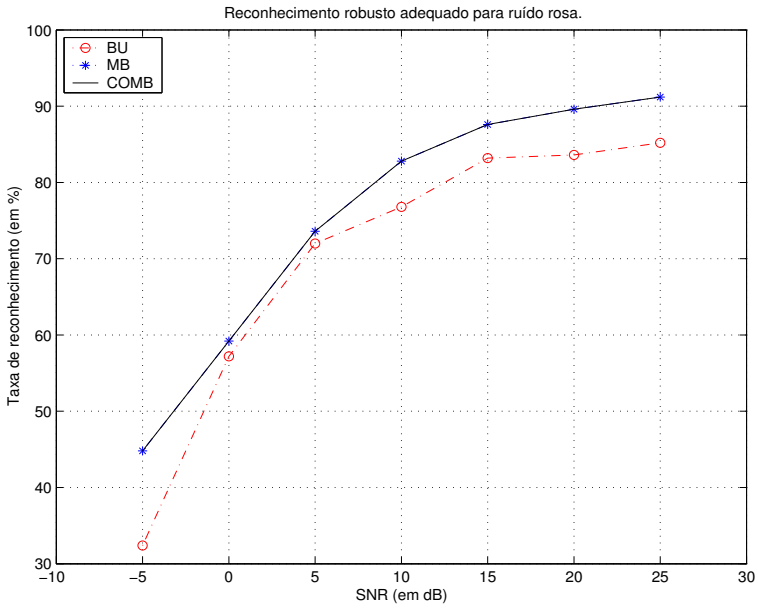


(b) Esquema detalhado.

Figura 6.11: Configuração do SRRV a partir de parâmetros extraídos do sinal ruidoso.

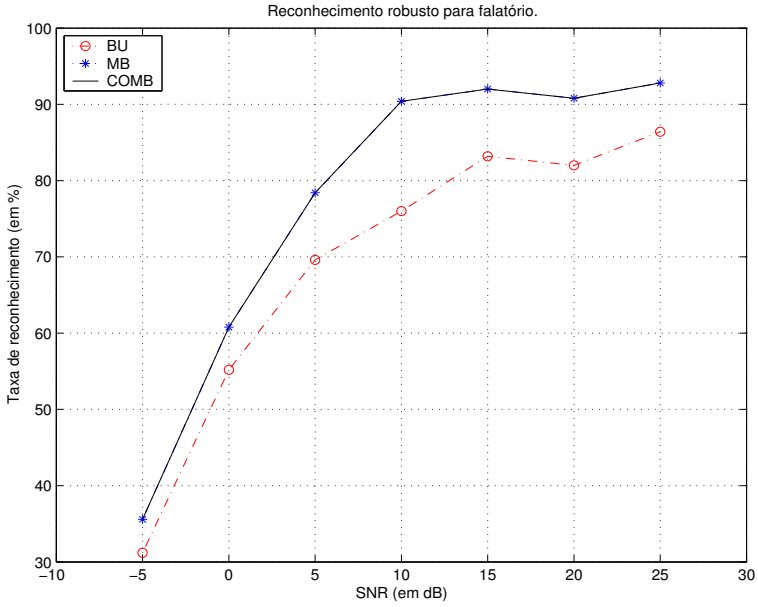


(a) Ruído branco.

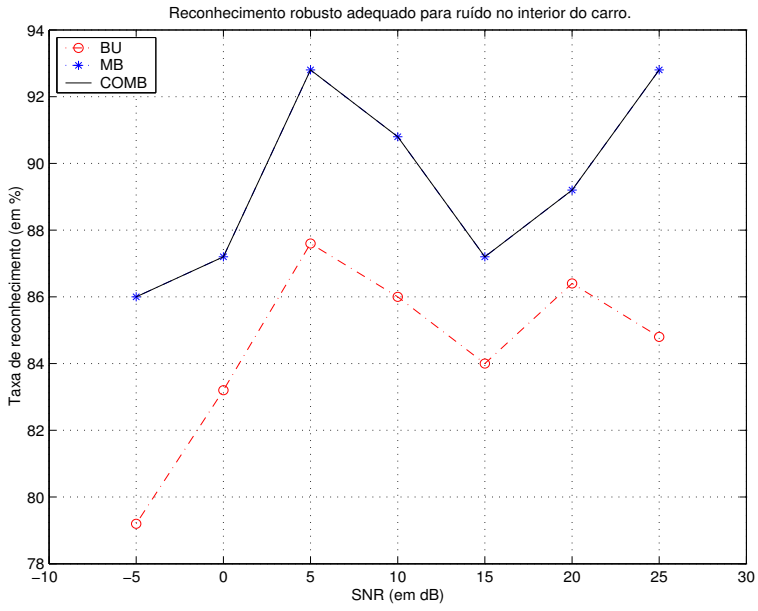


(b) Ruído rosa.

Figura 6.12: Reconhecimento robusto de voz conduzido pelo algoritmo proposto em ruído branco e ruído rosa.



(a) Falatório.



(b) Ruído no interior do carro.

Figura 6.13: Reconhecimento robusto de voz conduzido pelo algoritmo proposto em falatório e ruído no interior do carro.

Capítulo 7

Conclusões e Trabalhos Futuros

Na Seção 7.1 apresentamos as conclusões e na Seção 7.2 citamos trabalhos futuros que podem ser realizados.

7.1 Conclusões

Foram realizados testes de identificação da natureza do ruído com o propósito de conduzir apropriadamente o reconhecimento. O processo é realizado usando HMM e os parâmetros extraídos são entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Com este método, para quatro possíveis tipos de ruído simulados (branco, rosa, falatório e interior de carro), obtivemos $97,22\%$ de correta identificação da natureza do ruído quando misturados com comandos de voz, e $100,00\%$ quando o sinal é composto apenas por ruído. Os melhores índices de acerto ocorreram em condições críticas de contaminação, atingindo $100,00\%$ em todos os ambientes ruidosos simulados.

Uma aplicação para o método de identificação da natureza do ruído através de HMM, associado à SNR e à distância euclidiana da log-energia calculados em cada frame, foi apresentado em um detector de extremos robusto. Com este método a redução média da taxa de erro na detecção de início foi de até $12,23\%$ e na detecção de fim foi de até $26,24\%$, para o caso de ruído rosa com SNR de 0 a 20dB, em relação ao método tradicional baseado na energia e na taxa de cruzamentos por zero.

O algoritmo de decisão em sub-bandas e o algoritmo para detecção de extremos robusta apresentam uma melhoria de até 32% na taxa de reconhecimento em múltipplas bandas, para o caso de ruído rosa, em condições críticas de contaminação, quando comparado com o método de detecção de extremos baseado na energia e taxa de cruzamentos por zero implementado em [11].

A tarefa de reconhecimento em múltiplas bandas superou o processamento em banda única em até $14,4\%$ no caso de contaminação por falatório. E dentro do processamento em múltiplas bandas, a subtração espectral apresentou melhor performance que a filtragem de Wiener em até $24,8\%$, em caso de contaminação por ruído rosa. O banco de filtros que apresentou melhor eficiência foi o modulado por cosseno com oito bandas uniformes, superando os outros bancos como também o processamento em banda única.

7.2 Trabalhos Futuros

Os algoritmos propostos nesta tese foram testado com quatro tipos de ruído. Ruídos de outras naturezas podem ser aplicados.

Um aperfeiçoamento da análise do sinal ruidoso poderá ser verificada usando outros parâmetros, além da identificação da natureza do ruído e sua SNR.

Outros parâmetros podem ser variados para avaliar a performance do reconhecimento em múltiplas bandas frente o reconhecimento em banda única, tais como:

- as técnicas de redução de ruído;
- os tipos de bancos de filtros;
- o número de sub-bandas.

Por fim, uma questão importante a ser desenvolvida é a realização do processamento em múltiplas bandas de forma paralela, já que a complexidade computacional ficou multiplicada pelo número de sub-bandas.

Referências Bibliográficas

[1] RABINER, L. R., JUANG, B., *Fundamentals on Speech Recognition*. New Jersey, Prentice Hall, 1996.

[2] YOUNG, S., EVERMANN, G., KERSHAW, D., *et al.*, *The HTK Book (for HTK Version 3.1)*. Cambridge University Engineering Department, 2001.

[3] WALKER, W., LAMERE, P., KWOK, P., *et al.*, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition", http://cmusphinx.sourceforge.net/sphinx4/doc/Sphinx4Whitepaper.pdf, White paper - SUN MICROSYSTEMS INC., 2004.

[4] XU, H., TAN, Z., DALSGAARD, P., *et al.*, "Spectral subtraction with full-wave rectification and likelihood controlled instantaneous noise estimation for robust speech recognition", *8th International Conference on Spoken Language Processing*, pp. 2085–2088, 2004.

[5] CHEN, C.-P., BILMES, J., ELLIS, D., "Speech feature smoothing for robust automatic speech recognition", November 2004. Aceito para ICASSP-05.

[6] JUNQUA, J.-C., *Robust Speech Recognition in Embedded Systems and PC Applications*. 1 ed. Kluwer, 2000.

[7] SANTOS, D. A. O., *Reconhecimento de Voz em presença de ruído*. M.Sc. dissertation, PUC-RJ, Julho 2001.

[8] BOURLARD, H., DUPONT, S., "Subband-based speech recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 2, pp. 1251–1254, 1997.

[9] TIBREWALA, S., HERMANSKY, H., "Sub-band based recognition of noisy speech", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 2, pp. 1255–1258, 1997.

[10] PARK, H. M., JUNG, H. Y., LEE, T. W., *et al.*, "On subband-based blind signal separation for noisy speech recognition", *Electronic Letters*, v. 35(23), pp. 2011–2012, 1999.

[11] TERUSZKIN, R., CONSORT, T. A., RESENDE JUNIOR, F. G. V., "Endpoint detection analysis for a implementation of a speech recognition system applied to robot control", *Preceedings of SAWCAS*, , 2001.

[12] SILVA, D. C., RESENDE JUNIOR, F. G. V., "Reconhecimento robusto de palavras isoladas usando multi-bandas", *Anais da I Semana da Eletrônica*, , 2003.

[13] FILHO, S. N., "Fundamentos sobre ruídos", Janeiro 2003. Universidade Federal de Santa Catarina - Dept. de Engenharia Elétrica , http://www.linse.ufsc.br/∼sidnei/.

[14] KAMATH, S. D., *A multi-band spectral subtraction method for enhancing speech corrupted by colored noise*. M.Sc. dissertation, University of Texas, December 2001.

[15] PAPOULIS, A., *Probability, Random Variables, and Stochastic Processes*. International Edition, McGraw-Hill, 1991.

[16] MA, N., BOUCHARD, M., GOUBRAN, R. A., "Perceptual Kalman filtering for speech enhancement in colored noise", 2004.

[17] DONOHO, D. L., "Denoising by soft thresholding", *IEEE Transactions on Information Theory*, v. 41, pp. 613–627, 1995.

[18] EPHRAIM, Y., MALAH, D., "Speech enhencement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 33, pp. 443–445, 1985.

[19] EPHRAIM, Y., TREES, H. L. V., "A signal subspace approach for speech enhancement", *IEEE Transactions Speech and Audio Processing*, v. 3, pp. 251–266, 1995.

[20] DELLER, J. R., HANSEN, J. H. L., PROAKIS, J. G., *Discrete-Time Processing of Speech Signals*. New York, IEEE Press, 2000.

[21] OPPENHEIM, A. L., SCHAFER, R. W., *Discrete-Time Signal Processing*. New Jersey, Prentice-Hall, 1989.

[22] SHEN, J.-L., HUNG, J.-W., LEE, L.-S., "Robust entropy-based endpoint detection for speech recognition in noisy environments", *International Conference on Spoken Language Processing*, , 1998.

[23] BOLL, S. F., "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 27, pp. 113–120, 1979.

[24] BOLL, S. F., "A spectral subtraction algorithm for suppression of acoustic noise in speech", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 200–203, 1979.

[25] WANG, D. L., LIM, J. S., "The inimportance of phase in speech enhancement", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 30, pp. 679–681, 1982.

[26] BEROUTI, N., SCHWARTZ, R., MAKHOUL, J., "Enhancement of speech corrupted by accoustic noise", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 208–211, 1979.

[27] VASEGHI, S. V., *Advanced Signal Processing and Digital Noise Reduction*. 1 ed. John Wiley and Teubner, 1996.

[28] LIN, L., HOLMES, W. H., AMBIKAIRAJAH, E., "Subband noise estimation for speech enhancement using a perceptual Wiener filter", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 1, pp. 80–83, 2003.

[29] COHEN, I., "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging", *IEEE Transactions Speech and Audio Processing*, v. 11, pp. 466–475, 2003.

[30] POPESCU, D. C., ZELIJKOVIC, I., "Kalman filtering of colored noise for speech enhancement", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 2, pp. 997–1000, 1998.

[31] SILVA, D. C., RESENDE JUNIOR, F. G. V., "Identificação da natureza do ruído baseada em HMM.", *Anais do* XXI *Simpósio Brasileiro de Telecomunicações*, , 2004.

[32] TERUSZKIN, R., RESENDE JUNIOR, F. G. V., VILLAS-BOAS, S. B., *et al.*, "Biblioteca orientada a objeto para reconhecimento de voz e aplicação em controle de robô", *Anais do* XIV *Congresso Brasileiro de Automática*, , 2002.

[33] RABINER, L. R., "A tutorial on hidden Markov models and selected applications in speech recognition", *Preceedings of IEEE*, v. 77, 1989.

[34] SILVA, D. C., RESENDE JUNIOR, F. G. V., "Detecção de extremos robusta baseada na identificação da natureza do ruído", *Anais da II Semana da Eletrônica*, , 2004.

[35] HILDEBRAND, F. B., *Introduction to Numerical Analysis*. 2 ed. McGraw-Hill, 1974.

[36] DINIZ, P. S. R., SILVA, E. A. B. D., NETTO, S. L., "Digital Signal Processing: Systems Analysis and Design", Março 2001. Apostila do curso de Filtros Digitais, EE/COPPE/UFRJ.

[37] VAIDYANATHAN, P. P., *Multirate Systems and Filter Banks*. New Jersey, Prentice-Hall, 1993.

[38] CROCHIERE, R. E., RABINER, L. R., *Multirate Digital Signal Processing*. New Jersey, Prentice-Hall, 1996.

[39] BOU-GHAZALE, S. E., ASSALEH, K., "A robust endpoint detection of speech for noisy environment with application to automatic speech recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 4, pp. 3808–3811, 2002.

[40] HERMANSKY, H., MORGAN, N., BAYYA, A., *et al.*, "RASTA-PLP speech analisys technique", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 1, pp. 121–124, 1992.

[41] http://spib.rice.edu/pib/select_noise.html, "Signal Processing Information Base (SPIB)", September 2002.

[42] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html, "Carnegie Mellon University - Página do Grupo de tecnolgia de voz", August 1996.

Apêndice A

Adição sinal-ruído

Todos os testes realizados neste trabalho tiveram suas amostras do sinais ruidosos formadas mediante a adição do ruído selecionado ao sinal limpo, de acordo com uma SNR previamente estabelecida, conforme a composição descrita em 2.2.1. Dessa forma, puderam ser criados diversos ambientes de simulação, a fim de analisar a performance do sistema em condições adversas.

Considerando o sinal de voz isento de ruído, $s(i)$, a composição do sinal de voz ruidoso, $x(i)$, é realizada através da adição do ruído original $\hat{d}(i)$, escalonado por um fator multiplicativo.

$$d(i) = \delta \times \hat{d}(i) \tag{A.1}$$

onde δ é o fator de multiplicação utilizado para atender a SNR desejada, a partir do ruído oriundo da base de dados NOISEX-92.

Por definição, a SNR pode ser expressa como:

$$SNR = 10 \log \left(\frac{\sigma_s^2}{\sigma_d^2} \right) \tag{A.2}$$

onde σ_s^2 é a energia do sinal de voz limpo e σ_d^2 é a energia do ruído.

Realizando uma manipulação matemática simples, temos que:

$$\frac{SNR}{10} = \log \left(\frac{\sigma_s^2}{\sigma_d^2} \right) \tag{A.3}$$

$$10^{\frac{SNR}{10}} = \frac{\sigma_s^2}{\sigma_d^2} \tag{A.4}$$

$$\sigma_d^2 = \frac{\sigma_s^2}{10^{\frac{SNR}{10}}} \tag{A.5}$$

Considerando que a energia média de um determinado sinal, $d(i)$, pode ser dada, pela relação de Parseval [20, 21], da seguinte forma:

$$\sigma_d^2 = \frac{1}{N}\sum_{i=0}^{N-1} |d(i)|^2 \tag{A.6}$$

Assim, substituindo A.1 em A.6, temos que:

$$\sigma_x^2 = \frac{1}{N}\sum_{i=0}^{N-1} |\delta \times \hat{d}(i)|^2 \tag{A.7}$$

Em seguida, promovendo a igualdade entre A.5 e A.7, teremos:

$$\frac{\sigma_s^2}{10^{\frac{SNR}{10}}} = \delta^2 \times \left(\frac{1}{N}\sum_{i=0}^{N-1} |\hat{d}(i)|^2\right) \tag{A.8}$$

$$\frac{\sigma_s^2}{10^{\frac{SNR}{10}}} = \delta^2 \times \hat{\sigma}_d^2 \tag{A.9}$$

$$\delta = \sqrt{\frac{\sigma_s^2}{\hat{\sigma}_d^2} \times \frac{1}{10^{\frac{SNR}{10}}}} \tag{A.10}$$

Portanto, considerando que os segmentos dos sinais $s(i)$ e $d(i)$ possuem o mesmo comprimento, a adição ocorre da seguinte forma:

$$x(i) = s(i) + \delta \times \hat{d}(i) \tag{A.11}$$

Apêndice B

Família de banco de filtros utilizadas

Tabela B.1: Coeficientes do filtro protótipo QMF-FIR (N=48 e $\beta = 4.55$).

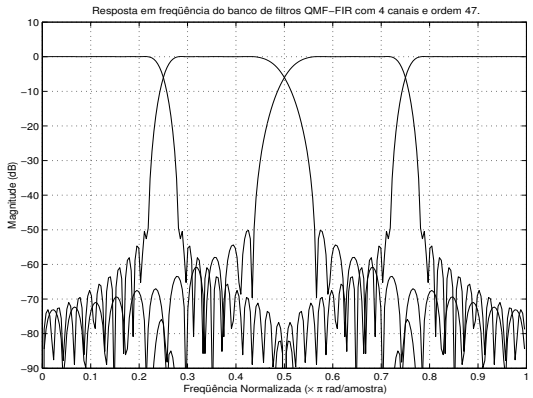
h(0) = -5.2468428E-004	h(8) = -5.4880831E-003	h(16) = -2.4369835E-002
h(1) = -8.1096679E-004	h(9) = -6.7152999E-003	h(17) = -2.9646835E-002
h(2) = 1.1692978E-003	h(10) = 8.1481653E-003	h(18) = 3.6644890E-002
h(3) = 1.6098578E-003	h(11) = 9.8228928E-003	h(19) = 4.6478951E-002
h(4) = -2.1438117E-003	h(12) = -1.1787354E-002	h(20) = -6.1543589E-002
h(5) = -2.7835446E-003	h(13) = -1.4106957E-002	h(21) = -8.8073378E-002
h(6) = 3.5430100E-003	h(14) = 1.6874271E-002	h(22) = 1.4894452E-001
h(7) = 4.4382392E-003	h(15) = 2.0225451E-002	h(23) = 4.5009479E-001

Tabela B.2: Coeficentes do filtro protótipo do banco de filtros modulado por cosseno com 4 bandas uniformes (N=48 e $\beta = 4.55$).

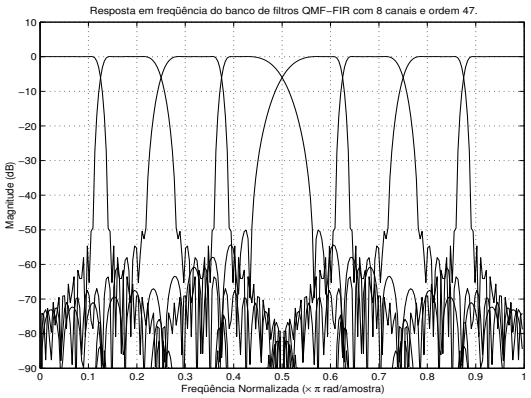
h(0) = 1.4459229e-004	h(8) = -1.5124038e-003	h(16) = 6.7158295e-003
h(1) = 6.3643403e-004	h(9) = -5.2700621e-003	h(17) = 2.3266372e-002
h(2) = 1.3733552e-003	h(10) = -9.5701240e-003	h(18) = 4.3039891e-002
h(3) = 2.2303490e-003	h(11) = -1.3608952e-002	h(19) = 6.4393438e-002
h(4) = 2.9701059e-003	h(12) = -1.6330580e-002	h(20) = 8.5264472e-002
h(5) = 3.2693086e-003	h(13) = -1.6568801e-002	h(21) = 1.0344330e-001
h(6) = 2.7804987e-003	h(14) = -1.3242663e-002	h(22) = 1.1688932e-001
h(7) = 1.2230882e-003	h(15) = -5.5737218e-003	h(23) = 1.2403695e-001

Tabela B.3: Coeficentes do filtro protótipo do banco de filtros modulado por cosseno com 8 bandas uniformes (N=48 e $\beta = 4.55$).

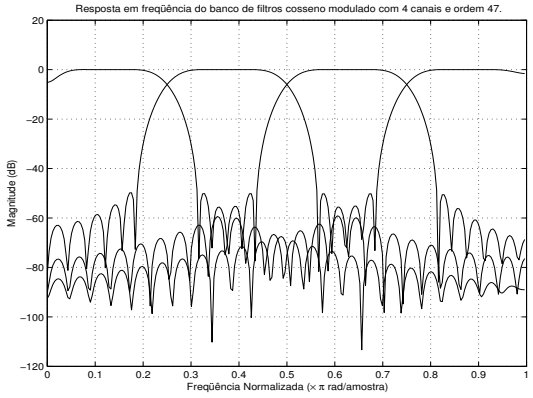
h(0) = -7.4283665e-004	h(8) = 7.6526927e-004	h(16) = 3.4502285e-002
h(1) = -1.1040265e-003	h(9) = 2.7731994e-003	h(17) = 4.0360336e-002
h(2) = -1.4670554e-003	h(10) = 5.4643433e-003	h(18) = 4.5976384e-002
h(3) = -1.7703717e-003	h(11) = 8.8652207e-003	h(19) = 5.1113222e-002
h(4) = -1.9348032e-003	h(12) = 1.2962634e-002	h(20) = 5.5543464e-002
h(5) = -1.8667077e-003	h(13) = 1.7699245e-002	h(21) = 5.9063992e-002
h(6) = -1.4631474e-003	h(14) = 2.2972139e-002	h(22) = 6.1509218e-002
h(7) = -6.1887695e-004	h(15) = 2.8634756e-002	h(23) = 6.2762118e-002



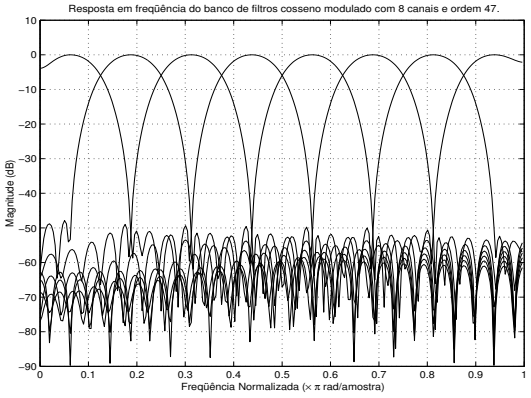
(a) QMF-FIR com 4 bandas uniformes.



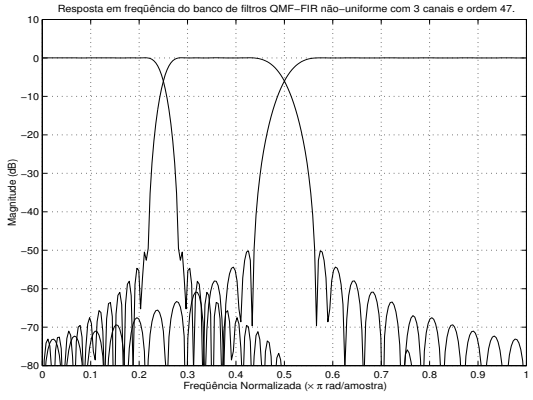
(b) QMF-FIR com 8 bandas uniformes.



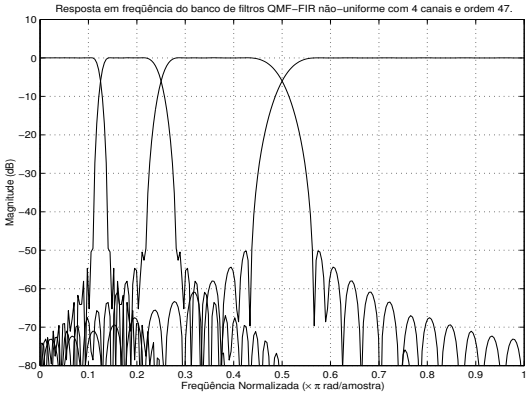
(c) Modulado por cosseno com 4 bandas uniformes.



(d) Modulado por cosseno com 8 bandas uniformes.



(e) QMF-FIR com 3 bandas não-uniformes.



(f) QMF-FIR com 4 bandas não-uniformes.

Figura B.1: BFD uniformes e não-uniformes.

Apêndice C

Tabelas de resultados

Este Apêndice contém as seguintes tabelas:

- Os resultados comparativos do recorte não-robusto e robusto vão da Tabela C.1 até a Tabela C.4.
- Resultados dos testes de identificação da natureza do ruído: da Tabela C.5 até a Tabela C.6.
- Os resultados comparativos do processamento em banda única e em múltiplas bandas vão da Tabela C.7 até a Tabela C.14.

Tabela C.1: Resultados obtidos com sinal contaminado por falatório e processado por multi-bandas usando recorte não-robusto e robusto.

SNR	Recorte não-robusto	Recorte robusto
0dB	$46,0\%$	$48,4\%$
5dB	$70,8\%$	$74,4\%$
10dB	$87,6\%$	$88,4\%$
15dB	$92,8\%$	$92,0\%$
20dB	$94,0\%$	$94,0\%$
25dB	$94,0\%$	$94,0\%$
30dB	$93,6\%$	$94,8\%$

Tabela C.2: Resultados obtidos com sinal contaminado por ruído rosa e processado por multi-bandas usando recorte não-robusto e robusto.

SNR	Recorte não-robusto	Recorte robusto
0dB	$18,4\%$	$36,8\%$
5dB	$40,8\%$	$73,2\%$
10dB	$73,2\%$	$84,8\%$
15dB	$91,6\%$	$90,4\%$
20dB	$92,8\%$	$94,8\%$
25dB	$95,2\%$	$95,6\%$
30dB	$94,4\%$	$96,4\%$

Tabela C.3: Resultados obtidos com sinal contaminado por ruído no interior do carro e processado por multi-bandas usando recorte não-robusto e robusto.

SNR	Recorte não-robusto	Recorte robusto
0dB	$89,6\%$	$88,4\%$
5dB	$90,0\%$	$90,0\%$
10dB	$92,0\%$	$93,2\%$
15dB	$92,8\%$	$95,6\%$
20dB	$92,0\%$	$96,0\%$
25dB	$92,4\%$	$96,0\%$
30dB	$91,2\%$	$96,0\%$

Tabela C.4: Resultados obtidos com sinal contaminado por ruído branco e processado por multi-bandas usando recorte não-robusto e robusto.

SNR	Recorte não-robusto	Recorte robusto
0dB	$17,6\%$	$26,8\%$
5dB	$38,4\%$	$54,8\%$
10dB	$65,6\%$	$78,4\%$
15dB	$87,6\%$	$85,6\%$
20dB	$92,4\%$	$95,2\%$
25dB	$92,8\%$	$97,6\%$
30dB	$94,4\%$	$96,8\%$

Tabela C.5: Tabela de confusão do teste de identificação dos segmentos de ruído

	BABBLE	PINK	VOLVO	WHITE
BABBLE	100%	0%	0%	0%
PINK	0%	100%	0%	0%
VOLVO	0%	0%	100%	0%
WHITE	0%	0%	0%	100%
Taxa de acerto média: 100%				

Tabela C.6: Tabela com resultados da identificação da natureza do ruído em locuções contaminadas.

SNR	BABBLE	PINK	VOLVO	WHITE
0dB	100,00%	100,00%	100,00%	100,00%
5dB	100,00%	100,00%	100,00%	100,00%
10dB	100,00%	100,00%	100,00%	100,00%
15dB	100,00%	98,34%	100,00%	98,34%
20dB	99,17%	85,12%	89,25%	74,38%
Taxa de acerto média: 97,22%				

Tabela C.7: Reconhecimento em banda única - Ruído ambiente: BABBLE.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
BUSR	19,6%	43,2%	54,8%	69,2%	75,2%	78,0%	80,8%
BUSE	31,2%	55,2%	69,6%	76,0%	83,2%	81,6%	85,2%
BUFW	24,0%	40,0%	60,8%	68,8%	79,2%	82,0%	86,4%

Tabela C.8: Reconhecimento em múltiplas bandas - Ruído ambiente: BABBLE.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
MBSECOS4U	26,0%	50,8%	64,0%	76,4%	82,8%	86,0%	84,0%
MBSECOS8U	**35,6%**	**60,8%**	**78,4%**	**90,4%**	**92,0%**	**90,8%**	**90,4%**
MBSEQMF4U	32,4%	45,6%	60,0%	76,4%	72,0%	81,2%	84,0%
MBSEQMF8U	21,2%	34,0%	52,4%	50,8%	63,6%	66,0%	61,2%
MBSEQMF3N	32,4%	48,4%	68,0%	73,2%	77,2%	83,2%	84,4%
MBSEQMF4N	22,8%	39,2%	50,8%	62,4%	74,0%	74,4%	76,8%
MBFWCOS4U	14,8%	34,4%	57,6%	68,8%	79,2%	85,2%	84,8%
MBFWCOS8U	**20,4%**	**45,2%**	**66,0%**	**81,2%**	**88,8%**	**90,4%**	**92,8%**
MBFWQMF4U	20,8%	37,2%	56,4%	72,0%	74,8%	73,6%	82,4%
MBFWQMF8U	12,8%	22,4%	42,8%	47,6%	58,4%	65,2%	59,6%
MBFWQMF3N	17,2%	30,0%	53,2%	72,0%	74,4%	80,4%	85,6%
MBFWQMF4N	15,6%	29,2%	49,2%	57,2%	66,8%	70,8%	73,2%

Tabela C.9: Reconhecimento em banda única - Ruído ambiente: PINK.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
BUSR	12,4%	27,2%	44,4%	63,6%	72,0%	76,8%	81,6%
BUSE	32,4%	57,2%	72,0%	76,8%	83,2%	83,6%	85,2%
BUFW	22,8%	42,4%	65,6%	72,0%	79,2%	81,6%	84,4%

Tabela C.10: Reconhecimento em múltiplas bandas - Ruído ambiente: PINK.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
MBSECOS4U	28,8%	48,4%	60,8%	72,8%	76,8%	82,0%	86,8%
MBSECOS8U	**44,8%**	**59,2%**	**73,6%**	**82,8%**	**87,6%**	**89,6%**	**91,2%**
MBSEQMF4U	36,8%	48,4%	59,2%	64,0%	72,8%	76,0%	71,6%
MBSEQMF8U	19,6%	39,6%	46,0%	49,6%	53,2%	51,6%	64,4%
MBSEQMF3N	30,4%	47,6%	58,0%	75,2%	78,4%	80,0%	84,8%
MBSEQMF4N	28,8%	46,0%	58,0%	63,2%	72,4%	72,8%	72,0%
MBFWCOS4U	10,8%	24,8%	44,4%	59,6%	71,6%	80,4%	85,2%
MBFWCOS8U	**20,0%**	**35,6%**	**59,6%**	**69,6%**	**83,6%**	**88,0%**	**90,8%**
MBFWQMF4U	16,0%	32,8%	47,6%	59,2%	63,2%	72,4%	74,8%
MBFWQMF8U	12,0%	16,8%	29,2%	50,8%	59,2%	54,0%	65,6%
MBFWQMF3N	16,4%	26,4%	47,6%	62,8%	72,8%	76,4%	82,0%
MBFWQMF4N	9,2%	26,8%	43,6%	57,6%	61,2%	67,6%	71,2%

Tabela C.11: Reconhecimento em banda única - Ruído ambiente: VOLVO.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
BUSR	76,0%	77,2%	80,0%	81,2%	81,6%	82,0%	82,4%
BUSE	78,8%	82,8%	87,6%	83,6%	81,6%	86,4%	84,8%
BUFW	79,2%	83,2%	79,2%	86,0%	84,0%	82,8%	84,8%

Tabela C.12: Reconhecimento em múltiplas bandas - Ruído ambiente: VOLVO.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
MBSECOS4U	73,2%	76,0%	81,6%	82,4%	84,0%	82,0%	83,2%
MBSECOS8U	**86,0%**	**87,2%**	**92,8%**	**90,8%**	**87,2%**	**87,2%**	**86,4%**
MBSEQMF4U	70,4%	70,0%	62,8%	80,8%	78,4%	63,2%	78,0%
MBSEQMF8U	52,0%	59,6%	58,8%	53,6%	67,6%	67,2%	65,2%
MBSEQMF3N	72,8%	76,8%	83,6%	83,2%	81,2%	82,0%	81,2%
MBSEQMF4N	67,2%	74,0%	73,6%	72,0%	75,6%	74,0%	74,8%
MBFWCOS4U	49,6%	59,2%	66,4%	77,6%	79,6%	84,4%	84,4%
MBFWCOS8U	**62,8%**	**73,6%**	**78,4%**	**83,6%**	**87,2%**	**89,2%**	**92,8%**
MBFWQMF4U	57,6%	59,2%	66,8%	74,0%	72,8%	78,4%	72,0%
MBFWQMF8U	29,2%	40,0%	51,2%	50,4%	66,4%	64,4%	63,6%
MBFWQMF3N	63,2%	64,4%	70,0%	75,2%	78,0%	80,8%	84,8%
MBFWQMF4N	52,0%	52,0%	53,2%	65,6%	67,2%	72,4%	73,6%

Tabela C.13: Reconhecimento em banda única - Ruído ambiente: WHITE.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
BUSR	11,6%	22,4%	40,0%	50,8%	69,6%	74,4%	76,4%
BUSE	24,4%	43,6%	65,6%	76,4%	76,4%	80,8%	84,0%
BUFW	16,8%	29,6%	53,2%	68,8%	80,0%	81,6%	82,0%

Tabela C.14: Reconhecimento em múltiplas bandas - Ruído ambiente: WHITE.

	SNR						
	-5dB	0dB	5dB	10dB	15dB	20dB	25dB
MBSECOS4U	19,6%	37,6%	49,2%	63,6%	72,8%	79,2%	83,2%
MBSECOS8U	**32,8%**	**52,0%**	**69,2%**	**73,6%**	**80,4%**	**87,6%**	**87,2%**
MBSEQMF4U	32,0%	48,2%	57,2%	59,2%	68,8%	66,0%	70,4%
MBSEQMF8U	18,8%	30,0%	39,6%	48,0%	41,6%	64,4%	53,6%
MBSEQMF3N	25,2%	42,8%	52,8%	64,4%	73,2%	75,6%	81,2%
MBSEQMF4N	30,0%	42,8%	52,8%	56,0%	66,4%	72,4%	72,8%
MBFWCOS4U	13,6%	28,0%	42,4%	58,0%	66,8%	78,8%	77,6%
MBFWCOS8U	**16,4%**	**30,8%**	**54,8%**	**68,0%**	**78,8%**	**86,0%**	**87,2%**
MBFWQMF4U	14,0%	32,0%	46,4%	56,8%	62,4%	69,6%	73,2%
MBFWQMF8U	12,8%	20,0%	28,8%	44,8%	44,8%	58,0%	60,4%
MBFWQMF3N	12,0%	31,6%	46,8%	59,2%	71,6%	74,0%	78,0%
MBFWQMF4N	15,2%	18,4%	43,2%	50,4%	61,2%	68,4%	68,0%

Apêndice D

Artigo SE 2003

Este artigo, de autoria de D. C. Silva e F. G. V. Resende Jr., foi apresentado na I Semana da Eletrônica, em setembro de 2003.

RECONHECIMENTO ROBUSTO DE PALAVRAS ISOLADAS USANDO MULTI-BANDAS

Denilson C. Silva, Eng. [1], Fernando Gil V. Resende Jr., Ph.D. [2]
[1] Programa de Engenharia Elétrica / COPPE
[2] Departamento de Eletrônica e de Computação / Universidade Federal do Rio de Janeiro
Cidade Universitária – Ilha do Fundão
21.945-970 – Rio de Janeiro, RJ

Palavras-chave: reconhecimento robusto de voz, ruído, detecção de extremos, banco de filtros.

Este trabalho visa mostrar os resultados obtidos no sistema implementado de reconhecimento robusto de palavras isoladas usando multi-bandas. O enfoque é a técnica utilizada no recorte de palavras, que proporciona uma melhora significativa na performance do sistema. As simulações foram realizadas em sete ambientes ruidosos distintos com SNR de 0, 5, 10, 15, 20, 25 e 30 dB usando um banco de filtros complementares em potência com quatro bandas uniformes.

Keywords: robust speech recognition, noise, endpoint detection, filter bank.

This work shows the results obtained with the implemented robust isolated-word recognition system using multi-bands. The focus is the approach used in the endpoint detection of words, which provides a significant improvement in the performance of the system. The simulations were accomplished in seven distinct noisy environments with SNR of 0, 5, 10, 15, 20, 25 and 30 dB using a power complementary filter bank with four uniform bands.

1. INTRODUÇÃO

Os sistemas de reconhecimento de voz, independentemente da finalidade para qual foram projetados, precisam operar em ambientes com condições reais, os quais, na maioria dos casos, são bastante diferentes daqueles onde foram treinados. Algoritmos visando o reconhecimento robusto da voz tem sido alvo crescente do interesse de pesquisadores do mundo acadêmico e comercial e tem por objetivo reduzir o impacto negativo que ambientes ruidosos causam na performance de sistemas de reconhecimento.

A detecção de extremos, ou recorte, exerce um papel muito importante no processo de reconhecimento de palavras isoladas, já que delimita o início e fim da locução para posterior processamento da fala. Caso não seja bem realizada, todo o processo fica comprometido. Essa tarefa torna-se ainda mais problemática quando lidamos com a situação de sinal contaminado por ruído. Um esquema de recorte robusto aumenta a consistência da detecção de início e fim, ajudando a manter a eficiência do sistema.

As Figuras 1, 2, 3 e 4 podem dar uma idéia de como seria realizar o recorte por solução robusta e não-robusta. Pode-se observar que quando um algoritmo robusto para recorte não é utilizado o trecho de voz extraído muitas vezes captura somente parte da palavra

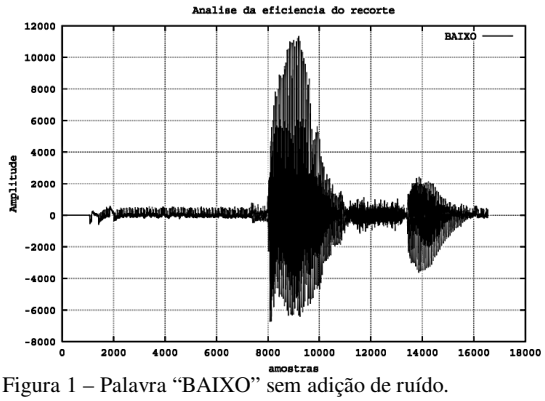


Figura 1 – Palavra "BAIXO" sem adição de ruído.

falada, o que tem grande impacto na performance do sistema de reconhecimento.

Neste artigo apresentamos a implementação de um sistema de detecção de extremos robusto e o aplicamos em um sistema de reconhecimento de palavras isoladas baseado em multi-bandas com adição de sete tipos diferentes de ruídos com distintos SNR (signal noise ratio). Os resultados confirmam a importância do recorte, particularmente em ambientes ruidosos.

A organização do artigo está descrita a seguir. Na Seção 2, as bases de dados utilizadas nos experimentos são descritas. Na Seção 3, o sistema implementado é

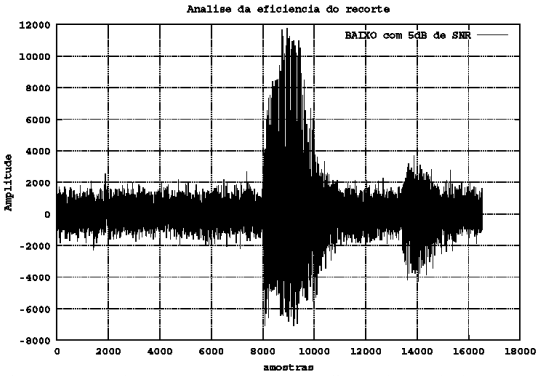


Figura 2 – Palavra "BAIXO" com adição de ruído branco a 5dB.

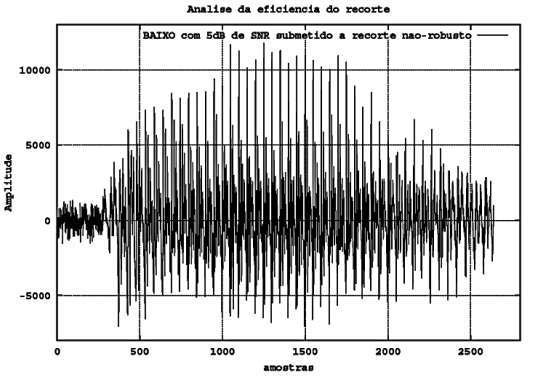


Figura 3 – Palavra "BAIXO" com adição de ruído branco a 5dB, submetido a recorte não-robusto.

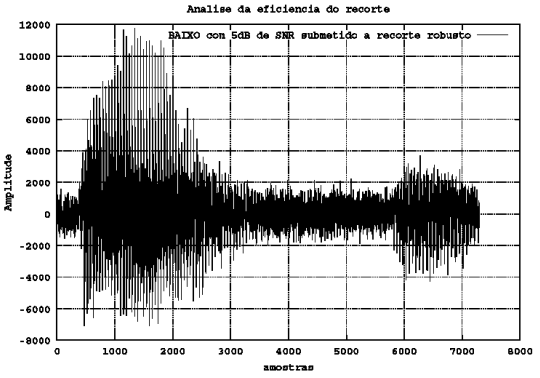


Figura 4 – Palavra "BAIXO" com adição de ruído branco a 5dB, submetido a recorte robusto.

apresentado. Na Seção 4, resultados são mostrados e comentados, e na Seção 5 temos as conclusões.

2. BASE DE DADOS

2.1. Locuções

As locuções utilizadas neste artigo são as mesmas de (TERUSZKIN, 2002), onde existem 10 palavras (ANDA, BAIXO, CIMA, DESLIGA, DIREITA, ESQUERDA, FRENTE, MÃO, OLHA e TRAS), repetidas 5 vezes cada uma por 20 locutores diferentes, sendo 15 no treino e 5 no teste (1000 locuções). A taxa de amostragem é de 11025Hz.

2.2. Ruídos

Os ruídos utilizados foram adquiridos de (RICE UNIVERSITY, 2003), para formar 7 ambientes distintos: WHITE, PINK, BABBLE, FACTORY, DESTROYER-ENGINE, VOLVO e HF CHANNEL. Eles foram adicionados aos sinais de voz para formar a fala ruidosa nas SNR desejadas, segundo o prescrito em (SANTOS, 2001). Os espectros dos ruídos citados anteriormente são apresentados a seguir, nas Figuras 5, 6, 7, 8, 9, 10 e 11.

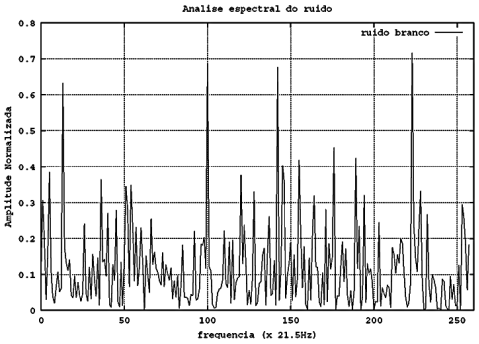


Figura 5 – Espectro do ruído branco.

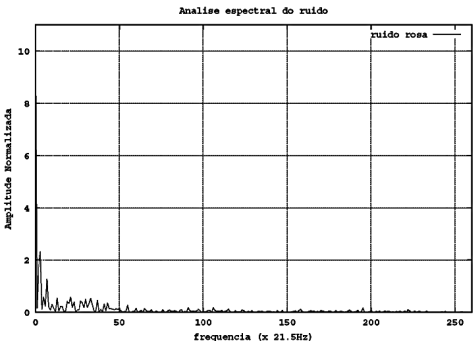


Figura 6 – Espectro do ruído rosa.

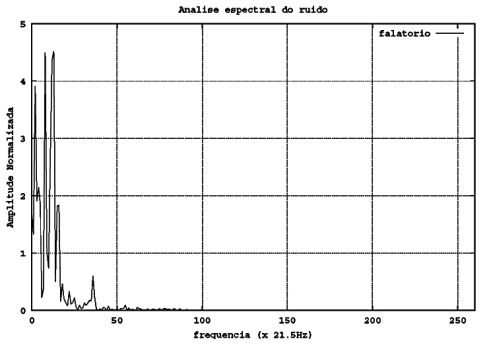


Figura 7 – Espectro do falatório.

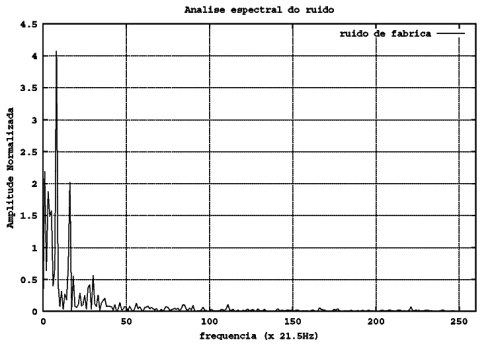


Figura 8 – Espectro do ruído de fábrica.

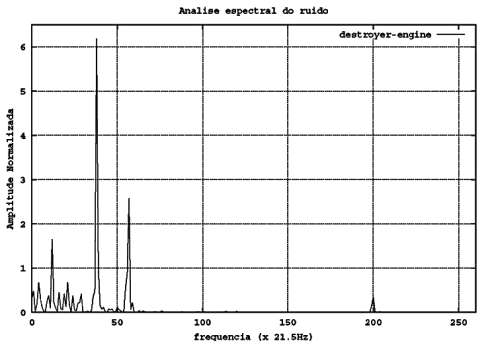


Figura 9 – Espectro do ruído "destroyer-engine".

3. CARACTERÍSTICAS DO SISTEMA

O sistema implementado de reconhecimento robusto de palavras isoladas usando múltiplas bandas é composto de quatro "sub-reconhecedores" HMM independentes,

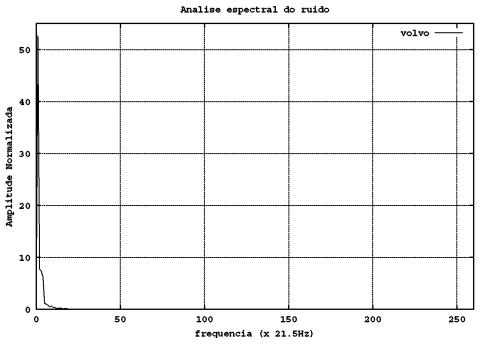


Figura 10 – Espectro do ruído "volvo".

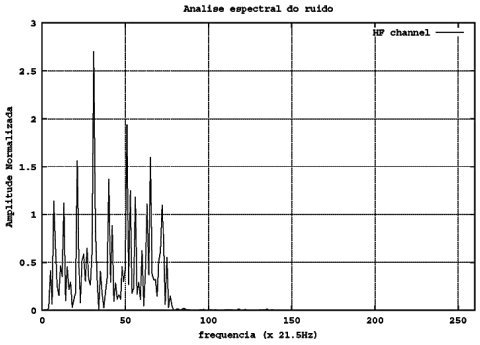


Figura 11 – Espectro do ruído "HF channel".

montados a partir da saída de cada banda de um banco de filtros, conforme mostrado na Figura 12.

O banco de filtros é do tipo QMF (*quadrature mirror filters*), complementares em potência, com 16 coeficientes, definido a partir de dados extraídos de (CROCHIERE, 1983) e suas saídas vão alimentar os sub-reconhecedores HMM. Cada sub-reconhecedor é estruturado com HMM de 10 estados e parametrização com 30 parâmetros por frame, sendo um de energia, 14 melcepstrais, um de delta-energia e 14 delta-melcepstrais.

A ação de cada sub-reconhecedor é independente, de forma tal que possamos ter resultados com probabilidades parciais independentes nas bandas e a partir daí usarmos o algoritmo descrito na Seção 3.2 para obtermos um resultado global para o sistema.

3.1. Detecção de início e fim de palavra

A detecção de início e fim de palavra utilizada no nosso trabalho é semelhante à usada em (BOU-GHAZALE,

2002), com algumas alterações, tais como o número de frames para referência de silêncio, onde utilizamos 10 frames e os valores dos parâmetros de decisão (α, β, η e γ) foram otimizados para conseguirmos melhores resultados.

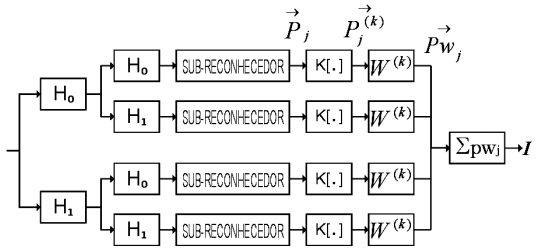


Figura 12 - Sistema de reconhecimento de voz multi-bandas.

Baseado no conhecimento de que a base de dados foi gravada com um silêncio inicial e assumindo aqui o estado da fala já com ruído adicionado, foi feita uma segmentação do sinal em frames com uma duração de aproximadamente 23ms e com overlap de 30%. Foram usados 10 frames como referência para o intervalo de silêncio. Algumas amostras iniciais do sinal foram descartadas devido a algumas descontinuidades existentes em certas locuções da base de dados. Na Figura 1 pode ser visto o efeito citado anteriormente.

São calculados sobre o intervalo assumido como silêncio as seguintes estatísticas: a energia média $\left(E_{sil}\right)$, o vetor cepstral médio $\left(\vec{c}_{sil}\right)$ e a distância cepstral média $\left(d_{sil}\right)$.

Um outro grupo de grandezas estatísticas é também calculado, porém para todo o sinal: a energia do i-ésimo frame $\left(E_{sp}^{i}\right)$ e a distância cepstral do i-ésimo frame em relação à média $\left(d_{sp}^{i}\right)$. O frame é considerado com voz se pelo menos uma das condições a seguir ocorrer:

$$(1) \quad \frac{d_{sp}^{i}}{d_{sil}} > \alpha \qquad e \qquad \frac{E_{sp}^{i}}{E_{sil}} > \beta$$

$$(2) \quad \frac{d_{sp}^{i}}{d_{sil}} > 1 \qquad e \qquad \frac{E_{sp}^{i}}{E_{sil}} > \eta$$

$$(3) \quad \frac{E_{sp}^{i}}{E_{sil}} > \gamma$$

Onde os parâmetros α, β, η e γ foram determinados empiricamente como 4.0, 0.95, 1.25 e 1.85 respectivamente.

O número de coeficientes cepstrais por frame utilizado foi de 14, considerado ideal para representar bem as características da fala. Ressaltando que estes coeficientes são completamente distintos daqueles utilizados na parametrização do sinal nos "sub-reconhecedores".

A detecção de início e fim foi experimentalmente definida quando uma seqüência de, no mínimo, 20 frames classificados como contendo fala for encontrada, seja a partir do início do sinal, para detectar o início de voz, como a partir do final, para detecção do término da locução.

3.2. Algoritmo de decisão

Aqui descrevemos o nosso algoritmo de decisão por bandas, a fim de aproveitar a características espectrais dos ruídos utilizados neste artigo e plotadas nas figuras 5, 6, 7, 8, 9, 10 e 11 da Seção 2.2.

Baseado nos resultados dados pelo algoritmo de Viterbi em cada sub-reconhecedor HMM, temos na j-ésima banda, a probabilidade de se gerar a seqüência de observações (O) dado o i-ésimo modelo $\left(\lambda^{(i)}\right)$, ou seja, $\left(P(O/\lambda^{(i)})\right)_j$, onde $i = 0,...,N-1$ e $j = 0,...,M-1$. N é o número de modelos e M é o número de bandas.

Seja $\vec{P}_j$ o vetor contendo todas as probabilidades dos i-ésimos modelos na banda j, $\vec{P}_j = \left[P(O/\lambda^{(i)})_j\right]$, e ainda, a matriz diagonal $W^{(k)} = diag\left[w_k\right]$, rotulada por $k = 0,...,N\text{-}1$, onde $\sum_k w_k \cong 1$ e $w_0 < w_1 < ... < w_{N-1}$.

Aplicando-se uma transformação K sobre $\vec{P}_j$ de forma que os seus elementos sejam ordenados segundo os seus valores, chegamos ao vetor

$$P_j^{(k)} = K\left[\vec{P}_j\right] \qquad (1)$$

onde:

$\vec{P_j^{(k)}} = \left[p_j^{(k)} \right]$ e $p_j^{(k)} = P(O/\lambda^{(i)})_j^{(k)}$, para $0 \le i \le N-1$. Assim,

$\vec{P_j^{(k)}} = \left[\min\left\{P(O/\lambda^{(i)})_j\right\}, ..., \max\left\{P(O/\lambda^{(i)})_j\right\} \right]$.

Aplicando-se $W^{(k)}$ sobre $\vec{P_j^{(k)}}$, estaremos dando mais ênfase às probabilidades parciais com maior verossimilhança em cada banda. Sendo assim, temos:

$$\vec{Pw_j} = \vec{P_j^{(k)}} W^{(k)} = \left[pw_j^{(i)} \right] \quad (2)$$

onde $pw_j^{(i)}$ é a probabilidade do i-ésimo modelo na banda *j* após a pesagem. A decisão é feita através de:

$$I = \max_{0 \le i \le N-1} \left\{ \sum_j pw_j^{(i)} \right\} \quad (3)$$

Sendo I o modelo selecionado a partir das parciais em cada banda.

Esse procedimento garante o fortalecimento das verossimilhanças maiores numa dada banda, já que, com exceção do ruído branco, todos os outros vão apresentar uma banda com uma quantidade reduzida de ruído e conseqüentemente, uma verossimilhança maior com o modelo correspondente.

4. RESULTADOS

Aqui podemos fazer uma análise dos resultados obtidos nos testes e observar como o recorte e uma decisão em bandas pôde dar uma boa robustez ao sistema de reconhecimento apresentado. Nossa análise busca comparar as taxas de acertos obtidos pelo reconhecedor usando o recorte não-robusto de (TERUSZKIN, 2001) e recorte robusto implementado e inserido no nosso sistema com base em (BOU-GHAZALE, 2002). As tabelas com as taxas de acertos são mostradas junto aos gráficos para os dois casos (robusto e não-robusto). Os testes foram realizados com todos os critérios definidos anteriormente e os parâmetros aqui apresentados nas Figuras 13, 14, 15 16, 17, 18 e 19 são os que até então nos deram os melhores resultados, mostrados nas Tabelas 1, 2, 3, 4, 5, 6 e 7, correspondentemente aos gráficos. Podemos notar uma melhoria bastante significativa no reconhecimento das locuções em praticamente todos os ambientes simulados, principalmente nas regiões de baixa SNR, onde os casos são mais críticos devido a grande quantidade de ruído, o que pode vir a comprometer o reconhecimento, caso o sistema não seja robusto. Algumas pequenas oscilações de eficiência foram constatadas, mas na maioria dos casos o sistema apresentado tem sua performance melhorada com a aplicação de um esquema de recorte robusto. Isto torna o sistema prático para aplicações diversas.

Tabela 1 – Resultados obtidos com sinal contaminado por ruído branco e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	17.6%	26.8%
5dB	38.4%	54.8%
10dB	65.6%	78.4%
15dB	87.6%	85.6%
20dB	92.4%	95.2%
25dB	92.8%	97.6%
30dB	94.4%	96.8%

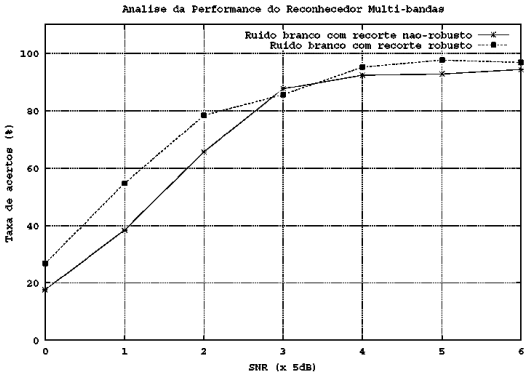


Figura 13 – Gráfico contendo resultados mostrados na Tabela 1, caso de ruído branco.

Tabela 2 – Resultados obtidos com sinal contaminado por ruído rosa e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	18.4%	36.8%
5dB	40.8%	73.2%
10dB	73.2%	84.8%
15dB	91.6%	90.4%
20dB	92.8%	94.8%
25dB	95.2%	95.6%
30dB	94.4%	96.4%

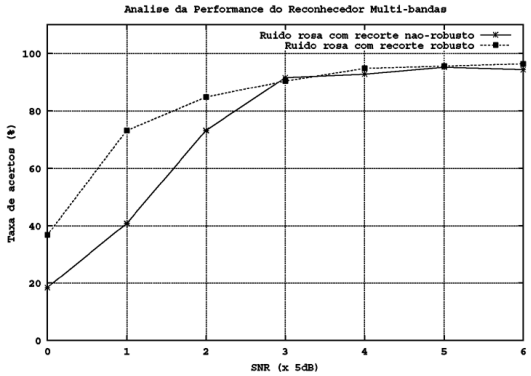


Figura 14 – Gráfico contendo resultados mostrados na Tabela 2, caso de ruído rosa.

Tabela 3 – Resultados obtidos com sinal contaminado por falatório e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	46.0%	48.4%
5dB	70.8%	74.4%
10dB	87.6%	88.4%
15dB	92.8%	92.0%
20dB	94.0%	94.0%
25dB	94.0%	94.0%
30dB	93.6%	94.8%

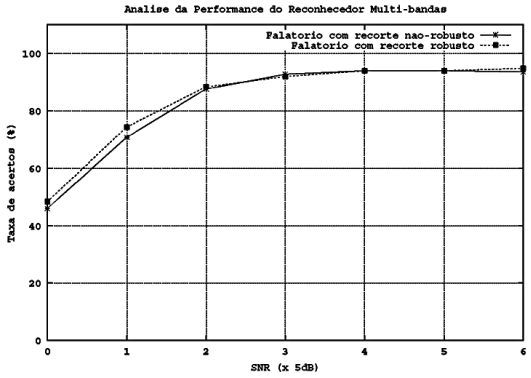


Figura 15 – Gráfico contendo resultados mostrados na Tabela 3, caso de falatório.

Tabela 4 – Resultados obtidos com sinal contaminado por ruído de fábrica e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	25.2%	40.0%
5dB	55.6%	75.2%
10dB	77.6%	84.4%
15dB	92.4%	91.2%
20dB	94.8%	93.6%
25dB	94.4%	95.2%
30dB	93.6%	94.8%

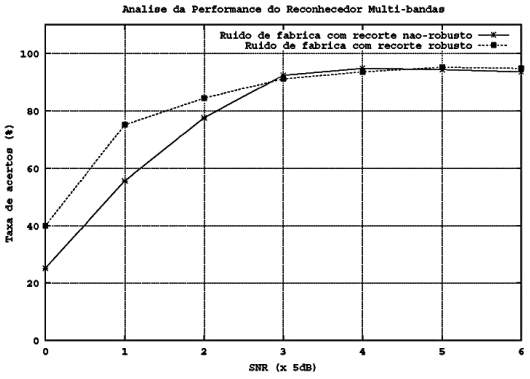


Figura 16 – Gráfico contendo resultados mostrados na Tabela 4, caso de ruído de fábrica.

Tabela 5 – Resultados obtidos com sinal contaminado por ruído "volvo" e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	89.6%	88.4%
5dB	90.0%	90.0%
10dB	92.0%	93.2%
15dB	92.8%	95.6%
20dB	92.0%	96.0%
25dB	92.4%	96.0%
30dB	91.2%	96.0%

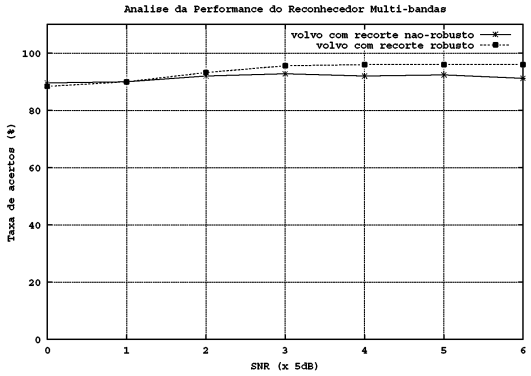


Figura 17 – Gráfico contendo resultados mostrados na Tabela 5, caso de ruído "volvo".

Tabela 6 – Resultados obtidos com sinal contaminado por ruído "destroyer-engine" e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	25.2%	37.6%
5dB	48.0%	65.2%
10dB	78.4%	76.4%
15dB	90.4%	88.4%
20dB	94.8%	94.4%
25dB	96.4%	96.4%
30dB	95.2%	96.8%

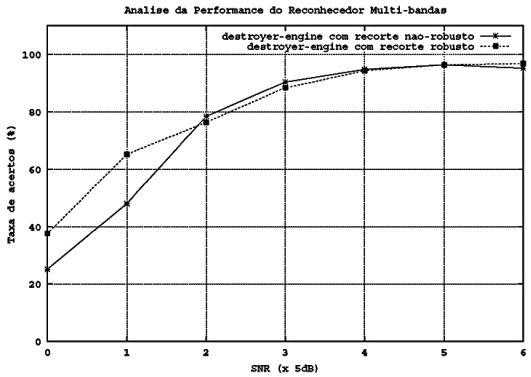


Figura 18 – Gráfico contendo os resultados mostrados na Tabela 6, caso de ruído "destroyer-engine".

Tabela 7 – Resultados obtidos com sinal contaminado por ruído "HF channel" e processado por multi-bandas usando recorte não-robusto e robusto.

SNR \ Recorte	Não-robusto	Robusto
0dB	23.2%	31.6%
5dB	44.8%	56.0%
10dB	76.0%	73.6%
15dB	87.6%	86.8%
20dB	91.6%	90.8%
25dB	93.2%	94.4%
30dB	94.8%	94.4%

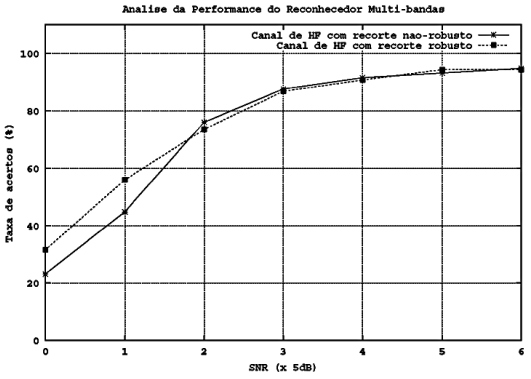


Figura 19 – Gráfico contendo resultados mostrados na Tabela 7, caso de ruído "HF channel".

5. CONCLUSÕES E TRABALHOS FUTUROS

Neste artigo foram apresentados resultados de um sistema de reconhecimento de palavras isoladas multi-bandas com recorte robusto e convencional para diversos tipos e intensidades diferentes de ruídos.

Pode-se constatar a importância do algoritmo de recorte pelo impacto significativo na performance do sistema, particularmente quando a SNR é baixa. Devido à localização espectral dos ruídos aqui utilizados, um trabalho em andamento é a utilização de um banco de filtros não-uniforme com um número maior de bandas, buscando uma resolução maior em freqüência.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BOU-GHAZALE, S., ASSALEH, K. A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition. In: IEEE INT. CONF. ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, April 2002. Anais... Orlando, Florida, USA.

CROCHIERE, R. R., RABINER, L. R. *Multirate Digital Signal Processing*. Englewood Clifs: Prentice-Hall, Inc., 1983. 411p.

RICE UNIVERSITY. Signal Processing Information Base (SPIB). (World Wide Web, 23/01/03, http://spib.rice.edu/spib/select_noise.html).

SANTOS, D. A. O. Reconhecimento de Voz em Presença de Ruído. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, 2001.

TERUSZKIN, R., CONSORT, T. A., RESENDE JR., F. G. V., Endpoint Detection Analysis for a Implementation of a Speech Recognition System Applied to Robot Control. In: PROCEEDINGS OF SAWCAS, Novembro 2001. Anais... Rio de Janeiro, Brasil.

TERUSZKIN, R., RESENDE JR., F. G. V., VILLAS-BOAS, S. B., LIZARRALDE, F. Biblioteca Orientada a Objeto para Reconhecimento de Voz e Aplicação em Controle de Robô. In: CONGRESSO BRASILEIRO DE AUTOMÁTICA, Setembro 2002. Anais... Rio Grande do Norte, Brasil.

Apêndice E

Artigo SBT 2004

Este artigo, de autoria de D. C. Silva e F. G. V. Resende Jr., apareceu no XXI Simpósio Brasileiro de Telecomunicações, em setembro de 2004.

Identificação da natureza do ruído baseada em HMM

Denilson C. Silva e Fernando G. V. Resende Jr.

***Resumo*— Este artigo apresenta um método eficiente para identificar tipos de ruído em sinais de voz com baixa SNR. De acordo com a natureza do ruído e sua potência, a técnica mais apropriada de tratamento robusto pode ser escolhida. O processo de identificação é realizado através de HMM e os parâmetros extraídos são entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Com o método proposto, para quatro possíveis tipos de ruído (branco, rosa, falatório e interior de carro), obtivemos 97,22% de correta identificação da natureza do ruído quando misturados com comandos de voz, e 100% quando o sinal é composto apenas por ruído.**

***Palavras-Chave*— Ruído, Processamento Robusto de Voz, HMM, Espectro de Potência.**

***Abstract*— This article presents an efficient method to identify types of noise in speech signals with low SNR. In accordance with the type of noise and its power, the most suitable technique for robust treatment can be selected. The process is carried out by HMM and the extracted parameters are spectral entropy, zero crossing rate and log-energy from 16 sub-bands. With the considered method, for four possible types of noise (white, pink, babble and car interior), we got 97.22% of correct identification of the nature of the noise when mixed with commands of voice, and 100% when the signal is composed only for noise.**

***Keywords*— Noise, Robust Speech Processing, HMM, Power Spectrum.**

I. INTRODUÇÃO

Pesquisas na área de reconhecimento robusto têm sido intensamente realizadas no sentido de viabilizar aplicações práticas de reconhecimento de voz. Em atividades recentes de processamento robusto de voz, estimativas do espectro do ruído são realizadas, a partir das amostras do sinal ruidoso [1], [2].

Neste trabalho, não só a potência do ruído é estimada, mas também a sua natureza, o que permite a escolha da técnica de processamento robusto mais apropriada. Isso pode ser observado em [3], onde temos uma técnica eficiente para alguns tipos de ruídos e ineficiente para outros.

O algoritmo que utilizamos para identificar a natureza dos ruídos é baseado na extração da entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas na classificação utilizando modelos de Markov escondidos (hidden Markov models - HMM). Com o método proposto obtivemos 97,22% de acerto na identificação da natureza do ruído em locuções contaminadas e 100% em sinal composto somente por ruído.

Este artigo está organizado da seguinte forma. A Seção II descreve o procedimento usado para estimativa da relação sinal-ruído (signal noise rate - SNR). Na Seção III, apresentamos o método de identificação do tipo de ruído. A Seção IV apresenta a origem dos sinais utilizados, tanto os ruídos como as locuções. Por fim, na Seção V temos os resultados obtidos e na Seção VI, as conclusões.

Denilson C. Silva, Eng., Fernando G. V. Resende Jr., Ph.D., Departamento de Engenharia Elétrica, Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, Brasil, E-mails: denilson@lps.ufrj.br, gil@lps.ufrj.br.

II. ESTIMATIVA DA *SNR*

Em termos práticos, quanto menor for a *SNR*, maior é a necessidade de que um processamento robusto seja realizado. Para a estimativa da *SNR*, uma atualização adaptativa da potência do ruído é realizada, onde um filtro IIR de um pólo é utilizado [1], [2], [4].

Seja o sinal ruidoso $x(i)$ uma composição do sinal de voz limpo $s(i)$ e do ruído $d(i)$:

$$x(i) = s(i) + d(i) \tag{1}$$

Definimos como $\sigma_x(k)$ a potência do sinal ruidoso no k-ésimo frame.

$$\sigma_x(k) = \sum_{i=0}^{L-1} x^2(kL + i) \tag{2}$$

onde k é o índice do frame e L é o comprimento do frame.

Definindo também $\sigma_d(k)$ como a potência do ruído no k-ésimo frame e assumindo que ela varia de forma mais lenta que a potência do sinal ruidoso, $\sigma_x(k)$, uma estimativa da potência do ruído é realizada, recursivamente, através do filtro IIR:

$$\sigma_d(k) = \alpha(k)\sigma_d(k-1) + (1 - \alpha(k))\sigma_x(k) \tag{3}$$

O parâmetro $\alpha(k)$ é o elemento que conduz a atualização da estimativa em direção ao sinal ruidoso no frame k ou em direção à estimativa da potência do ruído no frame anterior:

$$\alpha(k) = 1 - min(1, RSNR(k)^{-Q}) \tag{4}$$

onde $Q = 5$ e a *SNR* relativa no k-ésimo frame, $RSNR(k)$, é:

$$RSNR(k) = \frac{\sigma_x(k)}{\frac{1}{M}\sum_{n=0}^{M-1}\sigma_x(n)} \tag{5}$$

sendo M um número de frames dentro do intervalo inicial, sem a presença do sinal de voz $s(i)$. Neste trabalho $M = 5$.

Assim, se um dado frame possui apenas ruído, a sua $RSNR$ fica bem próxima de 1, resultando num $\alpha(k)$ pequeno (em torno de 0). Desta forma, a atualização do ruído segue o sinal ruidoso no frame k. Caso o frame possua voz, a $RSNR$ é bem mais elevada e $\alpha(k)$ fica mais próximo de 1. Neste caso, a estimativa do ruído segue a atualização do frame anterior.

A Figura 1 mostra a estimativa da potência do ruído $d(i)$ a partir do sinal ruidoso $x(i)$, para o caso de uma locução contaminada por ruído rosa com *SNR* de 5dB. Podemos

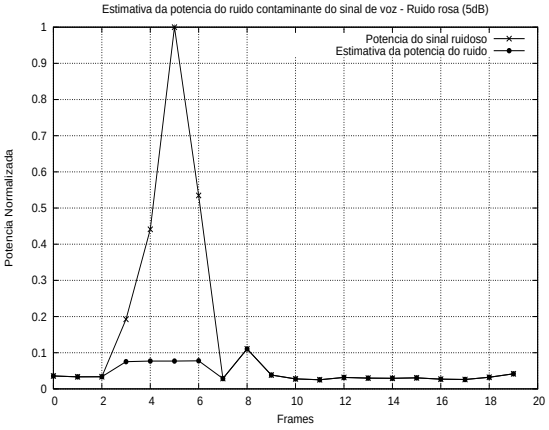


Fig. 1. Estimativa da potência do ruído num sinal contaminado por ruído rosa.

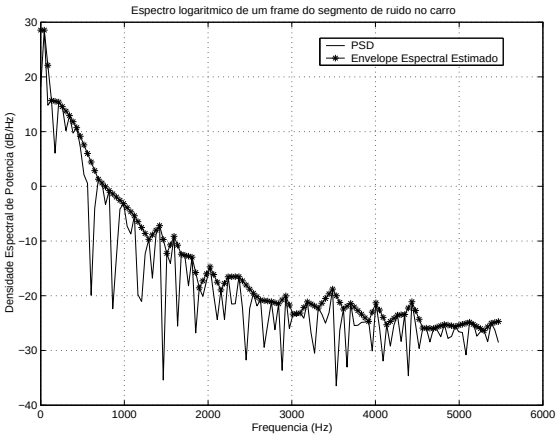


Fig. 2. Envoltória do espectro de um frame com ruído no carro.

observar nesta figura que a estimativa da potência do ruído até o frame dois e após o frame sete, segue perfeitamente a potência do sinal ruidoso $x(i)$, por se tratar de trechos onde a variação de $\sigma_x(k)$ é lenta ao longo dos frames. Fora deste trecho a variação entre frames de $\sigma_x(k)$ é brusca, denotando a presença de $s(i)$ nesta região e a potência do ruído $\sigma_d(k)$ passa a seguir as estimativas anteriores.

Com o conhecimento de $\sigma_x(k)$ e $\sigma_d(k)$, uma estimativa da *SNR* é feita segundo a equação a seguir:

$$SNR = 10log\left(\frac{\sum_{k=th1}^{th2}\sigma_x(k) - \sum_{k=th1}^{th2}\sigma_d(k)}{\sum_{k=th1}^{th2}\sigma_d(k)}\right) \quad (6)$$

onde $th1$ e $th2$ são os limites inferior e superior, respectivamente, do conjunto de frames onde temos variações mais bruscas de σ_x e, conseqüentemente, presença de $s(i)$. Assim, para o cálculo da *SNR*, serão utilizados apenas os k-ésimos frames onde a relação de potências, $\frac{\sigma_x(k)}{\sigma_d(k)}$, ultrapassar um limiar $TH = 2$. No exemplo da Figura 1, $th1 = 2$ e $th2 = 7$.

Com a estimativa da *SNR* do sinal, temos condições de dizer se o nível de ruído é tolerável ou não, dependendo da aplicação do processamento.

III. Identificação do tipo ruído

A. Extração de parâmetros

No processo de identificação do tipo de ruído, para cada frame k são extraídos 18 coeficientes:

- Entropia espectral (01);
- Taxa de cruzamentos por zero (01);
- Log-energia (16);

A entropia espectral, H, é definida na equação a seguir:

$$H = -\sum_{k=0}^{P-1} p_k log(p_k) \quad (7)$$

onde p_k é uma função densidade de probabilidade do espectro na freqüência de índice k, estimada pela normalização sobre todas as componentes de freqüência e P é o número de componentes de freqüência. O espectro $S(f_j)$, para cada freqüência f_j, é obtido em cada frame através da transformada rápida de Fourier (fast Fourier transform - FFT).

$$p_j = \frac{S(f_j)}{\sum_{k=0}^{P-1} S(f_k)}, \text{ para } j = 0, ..., P-1 \quad (8)$$

A entropia espectral está relacionada com a idéia de que um sinal tem tanto mais informação quanto maior for o seu grau de imprevisibilidade [5], [6].

A taxa de cruzamentos por zero, ZCR, é definida na equação a seguir:

$$ZCR = \frac{1}{L-1}\sum_{i=1}^{L-1}\frac{|sgn\{x(i)\} - sgn\{x(i-1)\}|}{2} \quad (9)$$

onde

$$sgn\{x(i)\} = \begin{cases} +1, & x(i) \geq 0 \\ -1, & x(i) < 0 \end{cases}$$

A taxa de cruzamentos por zero, muito utilizada em métodos de detecção de extremos [7], representa do número de vezes que o sinal cruza o eixo onde o valor de $x(i)$ é zero.

Considerando, agora, o espectro do sinal $x(i)$, definimos como $\hat{S}(f)$ uma estimativa da envoltória do espectro logarítmico. Uma divisão do espectro em K sub-bandas com P' freqüências em cada sub-banda é realizada, para a extração de K parâmetros. O k-ésimo parâmetro de log-energia, $LogE(k)$, é definido como a energia contida na k-ésima sub-banda, limitada pela envoltória, conforme a seguir:

$$LogE(k) = \sum_{j=0}^{P'-1}\hat{S}^{(k)}\{F(kP'+j)\} \quad (10)$$

onde $k = 0, ..., K-1$ e $F = \frac{f_s}{N}$, sendo f_s a freqüência de amostragem e N o número de pontos usados na FFT.

Na Figura 2 apresentamos o espectro de potência logarítmico, juntamente com a estimativa da envoltória, para

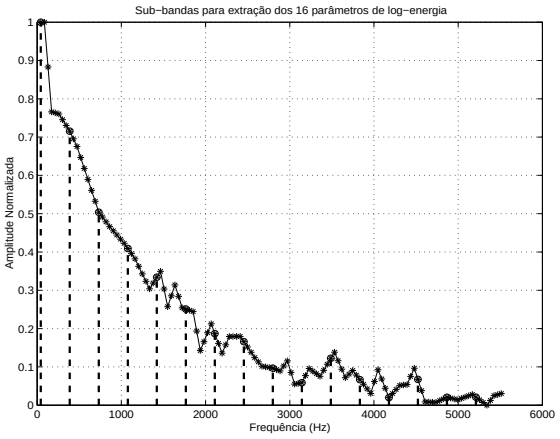


Fig. 3. Sub-bandas para extração de 16 parâmetros de log-energia

o caso de ruído no interior do carro. A envoltória foi estimada através de interpolações entre os picos do espectro logarítmico. Partindo deste espectro, foi feita a divisão em sub-bandas para a extração dos coeficientes. Cada frame foi dividido em 16 sub-bandas, gerando 16 coeficientes de log-enegria, $LogE$. Na Figura 3 temos a concentração de energia em cada uma das 16 sub-bandas, para o caso do ruído no interior do carro, mostrado na Figura 2. Os intervalos das sub-bandas, delimitados pelas linhas tracejadas da Figura 3, são apresentados na Tabela I.

TABELA I
TABELA COM OS INTERVALOS DE FREQÜÊNCIAS PARA EXTRAÇÃO DOS PARÂMETROS DE LOG-ENERGIA

BANDA	FREQÜÊNCIA
1	$0 - 344.5Hz$
2	$344.5 - 689.0Hz$
3	$689 - 1033.5Hz$
4	$1033.5 - 1378Hz$
5	$1378 - 1722.5Hz$
6	$1722.5 - 2067Hz$
7	$2067 - 2411.5Hz$
8	$2411.5 - 2756Hz$
9	$2756 - 3100.5Hz$
10	$3100.5 - 3445Hz$
11	$3445 - 3789.5Hz$
12	$3789.5 - 4134Hz$
13	$4134 - 4478.5Hz$
14	$4478.5 - 4823Hz$
15	$4823 - 5167.5Hz$
16	$5167.5 - 5512Hz$

B. Configuração do HMM

Cada um dos 4 tipos de ruídos têm os seus parâmetros usados no treinamento dos HMM discretos de [8], os quais possuem as seguintes características:

- 50% de superposição;
- 04 modelos;
- 04 estados;
- 128 centróides;
- 18 parâmetros por frame, sendo 01 de entropia espectral, 01 de taxa de cruzamentos por zero e 16 de log-energia espectral;

C. Critérios de decisão

A decisão acerca da natureza do ruído é realizada partindo de locuções contaminadas e com tamanho fixo de dois segundos de duração. Elas são segmentadas em trechos de 100 ms com superposição de 50%. Cada um desses segmentos é submetido ao classificador e a identificação do mesmo é registrada. Em seguida, uma contagem é realizada apenas sobre os segmentos de 100 ms que pertencem à região do sinal $x(i)$, sem a presença de $s(i)$. Para isto, levamos em consideração a potência estimada do ruído no frame k, $\sigma_d(k)$, realizada na Seção II. Os parâmetros $th1$ e $th2$ são os delimitadores das regiões de $x(i)$ com potência mais baixa. Este proced imento é adotado para evitar erros de classificação no caso de ocorrerem locuções muito longas com intervalos de voz muito curtos. No exemplo da Figura 1, os frames submetidos à análise são os que vão até $th1 = 2$ e estão após $th2 = 7$.

IV. BASE DE DADOS

As locuções utilizadas neste artigo foram coletadas da base de dados descrita em [8], onde temos 10 palavras isoladas (ANDA, BAIXO, CIMA, DESLIGA, DIREITA, ESQUERDA, FRENTE, MÃO, OLHA e TRAS), repetidas por diferentes locutores. A taxa de amostragem é de 11025Hz.

A base de dados de ruídos foi utilizada de [9], em formato "wave", onde cada ruído possui 9.0Mb, 235 segundos de duração, taxa de amostragem original de 19980Hz subamostrada para 11025Hz e tomadas da NOISEX-92. Foram selecionados quatro tipos de ruídos:

- Ruído branco (WHITE);
- Ruído rosa (PINK);
- Ruído no interior de um carro (VOLVO); e
- Falatório (BABBLE).

Os dois primeiros foram selecionados por serem tradicionais em tarefas de processamento robusto e os dois últimos por representarem situações reais de operação em condições adversas.

Os ruídos foram segmentados em trechos de 100 ms com superposição de 50% a fim de formar uma sub-base de dados para treinar os HMM. Com cada um dos tipos de ruídos foram obtidos 9178 segmentos com 100 ms de ruído.

As locuções ruidosas foram formadas através da adição dos ruídos selecionados às locuções limpas de acordo com *SNR*'s estabelecidas, que vão de 0 a 20dB.

V. RESULTADOS OBTIDOS

Inicialmente foi feito um treinamento com 500 segmentos de 100 ms de cada um dos quatro tipos de ruídos envolvidos. Em seguida, foram feitos dois testes com o sistema, buscando-se medir sua eficiência na identificação dos tipos de ruídos. O primeiro teste foi realizado com ruído, onde foram inseridos 100 segmentos de cada um dos tipos envolvidos, diferentes

TABELA II

TABELA DE CONFUSÃO DO TESTE COM SEGMENTOS DE RUÍDO

	BABBLE	PINK	VOLVO	WHITE
BABBLE	100%	0%	0%	0%
PINK	0%	100%	0%	0%
VOLVO	0%	0%	100%	0%
WHITE	0%	0%	0%	100%
Taxa de acerto média: 100%				

TABELA III

TABELA COM RESULTADOS DO TESTE COM LOCUÇÕES CONTAMINADAS

	BABBLE	PINK	VOLVO	WHITE
0dB	100%	100%	100%	100%
5dB	100%	100%	100%	100%
10dB	100%	100%	100%	100%
15dB	100%	98.34%	100%	98.34%
20dB	99.17%	85.12%	89.25%	74.38%
Taxa de acerto média: 97.22%				

dos usados no treino, com 100 ms cada. Através de HMM, como mencionado na Sub-seção III-B, foram extraídos 18 parâmetros de cada frame, com 20 ms e superposição de 50%. Os resultados são apresentados na Tabela II.

Em seguida foi realizado um teste funcional com sinais de voz contaminados pelos ruídos selecionados a *SNR*'s que vão de 0dB a 20dB. O procedimento foi fixar uma *SNR* e o ruído, observando o acerto ao longo das várias *SNR*. Foram inseridas 121 locuções no sistema. Após contaminadas pelos ruídos, as locuções foram segmentadas em trechos de 100 ms e analisadas através de HMM. Como mencionado na Sub-seção III-C, os segmentos com relação de potências acima do limiar TH não foram considerados na classificação. Os resultados obtidos são mostrados na Tabela III.

VI. CONCLUSÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho apresentamos um método de identificação da natureza do ruído baseado em HMM para diferentes *SNR*'s. Foram extraídos como parâmetros entropia espectral, taxa de cruzamentos por zero e log-energia de 16 sub-bandas. Os testes foram realizados somente com ruídos e, também, com locuções contaminadas por quatro tipos de ruídos (branco, rosa, falatório e interior do carro). Os resultados obtidos atingiram 100% no teste com ruídos e 97,22% na identificação em locuções contaminadas.

Trabalhos futuros estão sendo desenvolvidos, baseados na proposta aqui apresentada, visando atividades de processamento robusto através de múltiplas bandas, onde, dependendo do tipo de ruído e da *SNR*, podemos conduzir o nosso processamento de forma apropriada.

REFERÊNCIAS

[1] L. Lin, W. H. Holmes and E. Ambikairajah, *Subband noise estimation for enhancement using a perceptual Wiener filter*. In ICASSP, v.I, p.80-83, April 2003.

[2] I. Cohen, *Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging*. In IEEE Trans. on Speech and Audio Processing, v.11, p.466-475, September 2003.

[3] S. Tibrewala and H. Hermansky, *Sub-band based recognition of noisy speech*. In Proc. ICASSP, v.II, p.1255-1258, April 1997.

[4] D. C. Popescu and I. Zeljkovic, *Kalman filtering of colored noise for speech enhancement*. In ICASSP, v.2, p.997-1000, May 1998.

[5] J. G. Proakis, *Digital communications*. McGraw-Hill, New York, Fourth Edition, 2000.

[6] A. Papoulis, *Probability, random variables, and stochastic processes*. McGraw-Hill, New York, Second Edition, 1984.

[7] R. Teruszkin, T. A. Consort and F. G. V. Resende Jr., *Endpoint detection analysis for an implementation of a speech recognition system applied to robot control*. In Proc. SAWCAS, November 2001.

[8] R. Teruszkin, F. G. V. Resende Jr., S. B. Villas-Boas and F. Lizarralde, *Biblioteca orientada a objeto para reconhecimento de voz e aplicação em controle de robô*. In CBA, September 2002.

[9] Rice University, *Signal Processing Information Base (SPIB)*. http://spib.rice.edu/spib/select_noise.html.

Apêndice F

Artigo SE 2004

Este artigo, de autoria de D. C. Silva e F. G. V. Resende Jr., foi apresentado na II Semana da Eletrônica, em setembro de 2004.

DETECÇÃO DE EXTREMOS ROBUSTA BASEADA NA IDENTIFICAÇÃO DA NATUREZA DO RUíDO

Denilson C. Silva, Eng.[1] *e Fernando G. V. Resende Jr., Ph.D.*[1,2]

[1]Programa de Engenharia Elétrica - COPPE
[2]Departamento de Eletrônica e de Computação - UFRJ
Cidade Universitária - Ilha do Fundão - 21945-970
Rio de Janeiro - RJ - Brasil
{denilson, gil}@lps.ufrj.br

RESUMO

A detecção dos extremos da voz ainda é um grande problema, principalmente em situações de reconhecimento de voz em ambiente impregnado por ruído. Enquanto os métodos tradicionais procuram detectar o início e o fim dos trechos de voz numa locução, procuramos abordar o assunto buscando delimitar os trechos onde temos apenas ruído. O método que propomos está baseado na identificação da natureza do ruído por HMM, associado à SNR e à distância euclidiana da log-energia calculados em cada frame. Com o método proposto a redução média da taxa de erro na detecção de início foi de até 12,23% e na detecção de fim foi de até 26,24%, para o caso de ruído rosa com SNR de 0 a 20dB, em relação ao método tradicional baseado na energia e na taxa de cruzamentos por zero.

1. INTRODUÇÃO

Em diversas aplicações de processamento de sinais da fala, a determinação dos extremos das locuções é necessária. Os métodos tradicionais de detecção de extremos baseados na energia e na taxa de cruzamentos por zero funcionam muito bem com a fala limpa [1]. Quando temos locuções com fricativas, por exemplo, a detecção dos extremos fica bastante comprometida se o processo de delimitação ocorrer em ambiente ruidoso.

Vários trabalhos têm buscado solucionar a questão da detecção de extremos em condições de ambientes ruidosos [2, 3, 4], mas os resultados obtidos estão diretamente relacionados com a relação sinal-ruído (signal-to-noise ratio - *SNR*). Ou seja, quanto maior a carga de contaminação do sinal, menor é a eficiência da detecção dos extremos, principalmente se fricativas estão presentes.

Neste artigo é proposto um método para detecção dos extremos em locuções para reconhecimento de voz baseado no princípio de identificação da natureza do ruído que contamina o sinal [5], através de uma classificação de cada frame por modelos escondidos de Markov (hidden Markov models - HMMs), delimitando os trechos com voz a partir da identificação de frames apenas com ruído. A *SNR* e a distância euclidiana da log-energia de cada frame são também utilizadas para realizar um refinamento na detecção. O método proposto resultou, por exemplo, em uma redução média da taxa de erro em até 12,23% na detecção de início e em até 26,24% na detecção de fim, no caso de ruído rosa, comparativamente ao método descrito em [1], com *SNR* entre 0 e 20 dB.

Este artigo está organizado da seguinte forma. Na Seção 2 apresentamos o método de detecção proposto e seus detalhes. Na Seção 3 descrevemos a origem dos dados utilizados. Na Seção 4 mostramos os resultados dos testes realizados e as comparações com o método da energia e taxa de cruzamentos por zero de [1]. Finalmente, na Seção 5 apresentamos as conclusões.

2. MÉTODO PROPOSTO DE DETECÇÃO DOS EXTREMOS

O método de detecção proposto é realizado através de três processos decisórios a partir de uma análise de cada frame: a identificação da natureza do ruído, a *SNR* e a relação euclidiana da log-energia. Definimos que um frame com "flag" positiva ($+1$) possui somente ruído, enquanto que um frame com "flag" negativa (-1) possui voz misturada com ruído.

2.1. Identificação da natureza do ruído

Seja o sinal ruidoso $x(i)$ uma composição do sinal de voz limpo $s(i)$ e do ruído $d(i)$:

$$x(i) = s(i) + d(i) \tag{1}$$

O sinal $x(i)$ é dividido em K frames com tamanho N e superposição de L amostras:

$$x(k,i) = x(k(N-L)+i) \tag{2}$$

onde $0 \leq k \leq K-1$ e $0 \leq i \leq N-1$.

O processo de identificação da natureza do ruído é realizado através de uma classificação dos frames do sinal ruidoso ($x(k,i)$), por meio de HMM, entre os tipos de ruídos envolvidos no treinamento. As etapas da identificação são descritas a seguir.

2.1.1. Extração de parâmetros

Para cada frame são extraídos 18 coeficientes, sendo um de entropia espectral, um de taxa de cruzamentos por zero e 16 de log-energia.

2.1.1.a. **Entropia Espectral**: A entropia espectral, $H(k)$, que está relacionada com a idéia de que um sinal tem tanto mais informação quanto maior for o seu grau de imprevisibilidade [4], é definida na equação a seguir:

$$H(k) = -\sum_{j=0}^{F-1} p(k,j) log(p(k,j)) \tag{3}$$

onde $0 \leq k \leq K-1$ e

$$p(k,j) = \frac{S(k,j)}{\sum_{n=0}^{F-1} S(k,n)}, \text{ para } j = 0, ..., F-1 \qquad (4)$$

na qual $p(k,j)$ é uma função densidade de probabilidade do espectro na freqüência de índice j, estimada pela normalização sobre todas as componentes de freqüência e F é o número de componentes de freqüência. O espectro $S(k,j)$, para cada freqüência j, é obtido através da transformada rápida de Fourier (fast Fourier transform - FFT).

2.1.1.b. **Taxa de cruzamentos por zero**: A taxa de cruzamentos por zero para cada frame k, $ZC(k)$, representa uma média do número de vezes que o sinal cruza o eixo onde o valor de $x(k,i)$ é zero. Ela é definida na equação a seguir:

$$ZC(k) = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|sgn\{x(k,i)\} - sgn\{x(k,i-1)\}|}{2} \qquad (5)$$

onde

$$sgn\{x(k,i)\} \quad = \quad \begin{cases} +1, & x(k,i) \geq 0 \\ -1, & x(k,i) < 0 \end{cases}$$

2.1.1.c. **Log-energia**: Considerando o espectro do sinal no frame k, $S(k,j)$, o espectro de potência logarítmico é calculado como:

$$LS(k,j) = 10 \log 10 \left(\frac{S(k,j)}{f_s m} \right) \qquad (6)$$

onde f_s é a freqüência de amostragem e m o número de pontos usados na FFT.

Em seguida calculamos a envoltória do espectro logarítmico na freqüência j, $ENV(k,j)$, da seguinte forma:

Inicialização:
$ENV(k,0) = LS(k,0)$
$ENV(k,F-1) = LS(k,F-1)$
Interação: para j = 1,...,$F-2$;
se $(LS(k,j) > LS(k,j-1)$ e $LS(k,j) > LS(k,j+1))$
$ENV(k,j) = LS(k,j)$
outro
$ENV(k,j) = 0$;

Em seguida, os valores entre dois $ENV(k,j)$ não-nulos são interpolados.

Seja $ENV(k,n)$ e $ENV(k,n+l)$ dois valores não-nulos com $l-1$ nulos entre eles, temos que o incremento (β) da interpolação será dado por:

$$\beta = \frac{ENV(k,n+l) - ENV(k,n)}{l} \qquad (7)$$

Os resultados das interpolações na envoltória para o referido trecho serão:

$$ENV(k,j) = ENV(k,n) + j\beta, \text{ para } j = 1,...,l-1 \qquad (8)$$

Uma divisão do espectro, delimitado pela envoltória, é realizada em P sub-bandas com F' freqüências em cada sub-banda, para a extração de P parâmetros de log-energia. O p-ésimo parâmetro de log-energia no *k*-ésimo frame, $LogE(k,p)$, é definido como a energia contida na p-ésima sub-banda:

$$LogE(k,p) = \sum_{j=0}^{F'-1} ENV(k,(pF'+j)) \qquad (9)$$

onde $p = 0, ..., P-1$ e $F' = F/P$.

Na Figura 1 apresentamos o espectro de potência logarítmico, juntamente com a envoltória, para um frame com ruído no interior do carro. Os intervalos das sub-bandas utilizadas na extração dos parâmetros de log-energia são apresentados na Tabela 1. Foram utilizadas 16 sub-bandas ($P = 16$).

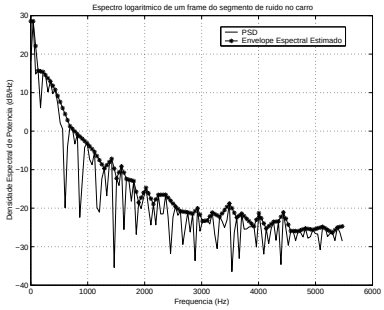


Fig. 1. Envoltória do espectro de um frame com ruído no carro.

Tabela 1. Tabela com os intervalos de freqüências para extração dos parâmetros de log-energia

SUB-BANDA	FREQÜÊNCIA
1	$0 - 344.5$Hz
2	$344.5 - 689.0$Hz
3	$689 - 1033.5$Hz
4	$1033.5 - 1378$Hz
5	$1378 - 1722.5$Hz
6	$1722.5 - 2067$Hz
7	$2067 - 2411.5$Hz
8	$2411.5 - 2756$Hz
9	$2756 - 3100.5$Hz
10	$3100.5 - 3445$Hz
11	$3445 - 3789.5$Hz
12	$3789.5 - 4134$Hz
13	$4134 - 4478.5$Hz
14	$4478.5 - 4823$Hz
15	$4823 - 5167.5$Hz
16	$5167.5 - 5512$Hz

2.1.2. Configuração do HMM

O sistema implementado para as simulações deste artigo foi treinado com quatro tipos de ruído: branco, rosa, falatório (babble) e interior do carro (volvo). Cada um dos quatro tipos de ruído tem os seus parâmetros usados no treinamento dos HMMs discretos de [6], com a seguinte configuração: segmentos de 100ms, 50% de superposição, quatro estados, 128 centróides e 18 parâmetros por frame.

2.1.3. Critérios de decisão

A decisão acerca da natureza do ruído é realizada a partir de locuções contaminadas e com tamanho fixo de dois segundos. As locuções são segmentadas em trechos de 512 amostras ($\approx$ 46ms), com superposição de 80%. Os frames são submetidos ao classificador e a identificação dos mesmos é registrada (ruído do frame), de acordo com os ruídos treinados. Em seguida, uma contagem é realizada para verificar qual tipo de ruído recebeu o maior número de classificações, ou seja, qual a natureza do ruído mais presente em todo sinal (ruído do sinal). Uma comparação é realizada da seguinte forma:

$$flag_{HMM}(k) \quad = \quad \begin{cases} -1, & \text{se ruído do frame} \neq \text{ruído do sinal} \\ +1, & \text{se ruído do frame} = \text{ruído do sinal} \end{cases}$$

onde $flag_{HMM}(k)$ é a "flag" atribuída ao *k*-ésimo frame através do critério de classificação por HMM, como definida no início da Seção 2.

2.2. Determinação da SNR dos frames

Definimos como $\sigma_x^2(k)$ a variânicia no *k*-ésimo frame do sinal ruidoso $x(k,i)$:

$$\sigma_x^2(k) = \frac{1}{N}\sum_{i=0}^{N-1}[x(k,i) - \mu(k)]^2 \tag{10}$$

onde $\mu(k)$ é a média no *k*-ésimo frame dada por:

$$\mu(k) = \frac{1}{N}\sum_{i=0}^{N-1} x(k,i) \tag{11}$$

Definimos, também $\sigma_d^2(k)$ como a estimativa da variância do ruído no *k*-ésimo frame realizada recursivamente através de um filtro IIR [7]:

$$\sigma_d^2(k) = \alpha(k)\sigma_d^2(k-1) + (1-\alpha(k))\sigma_x^2(k) \tag{12}$$

O parâmetro $\alpha(k)$ é o elemento que conduz a atualização da estimativa em direção à variância do sinal ruidoso no frame k ($\sigma_x^2(k)$) ou em direção à estimativa da variância do ruído no frame anterior ($\sigma_d^2(k-1)$):

$$\alpha(k) = 1 - min(1, RSNR(k)^{-Q}) \tag{13}$$

onde $Q = 5$ e a *SNR* relativa no *k*-ésimo frame, $RSNR(k)$, é:

$$RSNR(k) = \frac{\sigma_x^2(k)}{\frac{1}{M}\sum_{n=0}^{M-1}\sigma_x^2(n)} \tag{14}$$

sendo M um número de frames dentro do intervalo inicial da locução, sem a presença do sinal de voz $s(i)$. Neste trabalho utilizamos $M = 15$. Para cada frame k, a *SNR* é estabelecida através da relação $\frac{\sigma_x^2(k)}{\sigma_d^2(k)}$, onde:

$$flag_{SNR}(k) \quad = \quad \begin{cases} -1, & \text{se } \sigma_x^2(k)/\sigma_d^2(k) \geq \gamma th1 \\ +1, & \text{se } \sigma_x^2(k)/\sigma_d^2(k) < \gamma th1 \end{cases}$$

onde $th1$ recebeu empiricamente o valor de 1,25. $flag_{SNR}(k)$ é a "flag" atribuída ao *k*-ésimo frame através do critério da *SNR*. γ é um fator de ajuste do limiar, utilizado para manter a performance do detector nas *SNRs* mais elevadas, independente da natureza do ruído.

$$\gamma \quad = \quad \begin{cases} 1, & \text{se } SNR_{est} < 10dB \\ 2, & \text{se } SNR_{est} \geq 10dB \end{cases}$$

onde SNR_{est} é a estimativa da *SNR* de todo o sinal, calculada como:

$$SNR_{est} = 10\log 10\left(\frac{\sum_{k=0}^{K'}\hat{\sigma}_x^2(k) - \sum_{k=0}^{K'}\hat{\sigma}_d^2(k)}{\sum_{k=0}^{K'}\hat{\sigma}_d^2(k)}\right) \tag{15}$$

onde $\hat{\sigma}_x^2(k)$ e $\hat{\sigma}_d^2(k)$ são calculados de forma semelhante as equações (10) e (12), respectivamente, porém sem superposição dos K' frames, ou seja, $x(k,i) = x(kN+i)$, para $0 \leq k \leq K'-1$.

2.3. Determinação da relação euclidiana da log-energia

A relação euclidiana é estabelecida entre as distâncias euclidianas do sinal e do ruído. A distância euclidiana do sinal $x(i)$, ($ED_x(k)$), é determinada da seguinte forma:

$$ED_x(k) = \sqrt{\sum_{p=0}^{P-1}(LogE(k,p) - LogE_{ref}(p))^2} \tag{16}$$

onde:

$$LogE_{ref}(p) = \frac{1}{M}\sum_{k=0}^{M-1} LogE(k,p) \tag{17}$$

A distância euclidiana da log-energia do ruído, ED_d, é tomada como a $ED_x(k)$ máxima nos M frames iniciais do sinal:

$$ED_d = \max_{0 \leq k \leq M-1}\{ED_x(k)\} \tag{18}$$

Uma relação entre as distâncias euclidianas em cada frame é estabelecida como $\frac{ED_x(k)}{ED_d}$, onde:

$$flag_{ED}(k) \quad = \quad \begin{cases} -1, & \text{se } ED_x(k)/ED_d \geq th2 \\ +1, & \text{se } ED_x(k)/ED_d < th2 \end{cases}$$

onde $th2$ recebeu empiricamente o valor de 1,4. $flag_{ED}(k)$ é a "flag" atribuída ao *k*-ésimo frame através do critério da relação euclidiana.

2.4. Delimitação do sinal

Após todos os frames do sinal receberem três "flags" cada um (por tipo de ruído, por razão de variâncias e por razão de distâncias euclidianas da log-energia), os frames são novamente rotulados, recebendo uma única "flag", ($flag_{global}(k)$), baseada nas seguintes condições:

se $flag_{HMM}(k) = flag_{SNR}(k) = flag_{ED}(k) = +1$,
$flag_{global}(k) = +1$
outro,
$flag_{global}(k) = -1$

Em seguida, as "flags" dos frames são analisadas a partir do frame 0 em direção ao último, buscando a primeira seqüência de 15 frames com "flag" negativa, onde o primeiro frame da seqüência caracteriza o término de um possível trecho inicial contendo apenas ruído. De forma semelhante, partindo do último frame em direção ao frame 0, uma seqüência de 15 frames com "flag" negativa também é pesquisada, onde o último frame da seqüência caracteriza o início de um possível trecho final contendo apenas ruído.

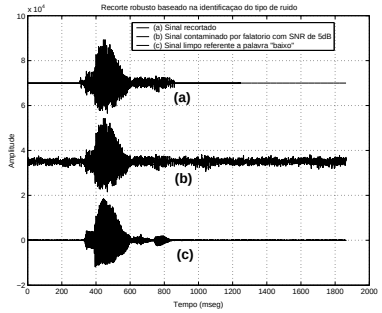


Fig. 2. Recorte robusto baseado na identificação do tipo de ruído. (a) Sinal recortado pelo método proposto. (b) Sinal contaminado por falatório com *SNR* de 5dB e (c) Sinal limpo referente a palavra "baixo".

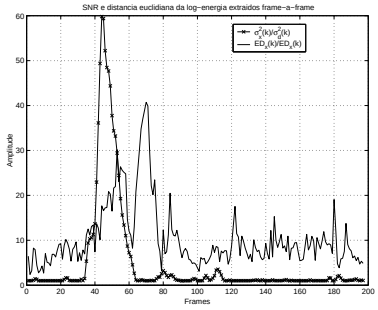


Fig. 3. Razão das variâncias e das distâncias euclidianas para cada frame da palavra "baixo", contaminada por falatório com *SNR* de 5dB, plotada na Figura 2.

3. BASE DE DADOS

As locuções utilizadas neste artigo foram coletadas da base de dados descrita em [6], onde temos 10 palavras isoladas (ANDA, BAIXO, CIMA, DESLIGA, DIREITA, ESQUERDA, FRENTE, MÃO, OLHA e TRÁS), repetidas por diferentes locutores. A taxa de amostragem é de 11025Hz.

A base de dados de ruídos foi utilizada de [8], com 235s de duração em cada tipo de ruído, taxa de amostragem original de 19980Hz sub-amostrada para 11025Hz. Foram selecionados quatro tipos de ruídos: ruído branco (WHITE), ruído rosa (PINK), ruído no interior de um carro (VOLVO) e falatório (BABBLE). Os dois primeiros foram selecionados por serem tradicionais em tarefas de processamento robusto e os dois últimos por representarem situações reais de operação em condições adversas.

As locuções ruidosas foram formadas através da adição dos ruídos selecionados às locuções limpas de acordo com a *SNR* estabelecida, que vai de 0 a 20dB.

4. RESULTADOS OBTIDOS

A detecção de extremos pode ter a sua performance avaliada de duas formas: uma delas é comparar os resultados obtidos na detecção com valores de referência obtidos de um recorte manual. A outra é passar as locuções recortadas, após a contaminação, em um sistema de reconhecimento de voz, comparando os seus resultados. Optamos pela primeira forma por termos condições de avaliar de forma direta os resultados, fazendo uma comparação com outros métodos.

Com relação ao sinal plotado na Figura 2, a Figura 3 mostra as relações de variâncias e distâncias euclidianas para cada frame. Podemos notar que a Figura $2(c)$ possui trechos de voz com baixa energia. Observando a Figura 3, se apenas a *SNR* for considerada, aqueles trechos não seriam incluídos na delimitação. Com o uso da distância euclidiana da log-energia, esse trecho é identificado como voz misturada com ruído. O trecho pode ser observado na Figura 2 entre 600ms e 800ms, correspondendo, aproximadamente, ao intervalo entre os frames 62 e 80 da Figura 3.

Inicialmente foi feito um treinamento com 500 segmentos de 100ms de cada um dos quatro tipos de ruídos modelados para realizarmos a identificação da natureza do ruído. Em seguida foi realizado um teste funcional com sinais de voz contaminados pelos ruídos selecionados com uma *SNR* que vai de 0 a 20dB. O procedimento foi fixar uma *SNR* e o ruído, observando o percentual de erro na detecção ao longo dos vários valores de *SNR*, tanto para detecção de início como para detecção de fim. Para as 121 locuções inseridas no detector, calculamos o erro percentual na detecção de início (ε_i) e o erro percentual na detecção de fim (ε_f) em relação ao recorte manual previamente realizado, da seguinte forma:

$$\varepsilon_i = \frac{|I - e_i|}{F - I} \times 100\% \tag{19}$$

$$\varepsilon_f = \frac{|F - e_f|}{F - I} \times 100\% \tag{20}$$

onde I e F são, respectivamente, os pontos de início e fim do recorte manual, e_i e e_f são, respectivamente, os pontos detectados.

Tabela 2. Tabela com os percentuais médios de redução da taxa de erro obtidos com o detector robusto em locuções contaminadas por ruído com SNR variando entre 0 e 20dB.

Redução média do erro de detecção				
	BABBLE	PINK	VOLVO	WHITE
Início	6,82%	12,23%	−2,65%	12,05%
Fim	20,81%	26,24%	0,70%	24,88%

Para os quatro tipos de ruídos, a redução média da taxa de erro na detecção, comparativamente ao método de [1], é mostrada na Tabela 2. Os valores calculados são uma média das taxas de erro de detecção de início e fim, para cada tipo de ruído, com os cinco valores de *SNR* simulados neste trabalho. Podemos observar que não houve uma redução efetiva da taxa de erro no caso de ruído no interior do carro. Para as demais situações, a redução foi considerável.

Os resultados individuais, que foram alcançados nos testes, são comparados com o método de [1] e mostrados nos gráficos das Figuras 4, 5, 6 e 7. É notado também que a melhor performance do detector ocorre nas faixas de baixa *SNR*, situação em que a contaminação mais compromete o processamento do sinal.

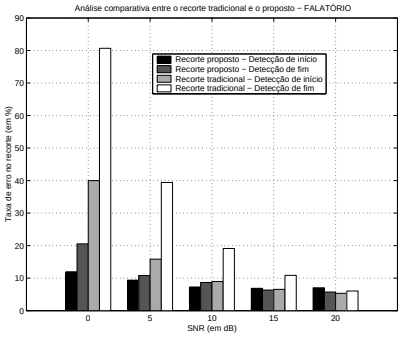


Fig. 4. Análise na condição de contaminação por falatório.

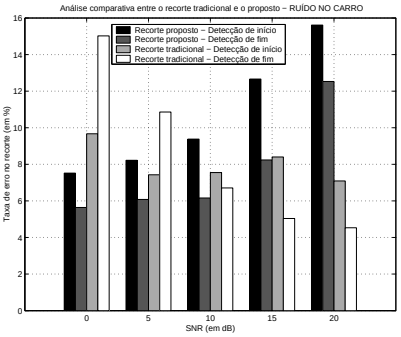


Fig. 6. Análise na condição de contaminação por ruído no carro.

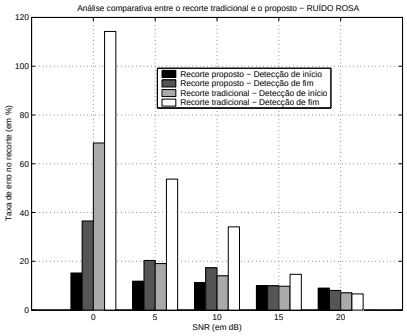


Fig. 5. Análise na condição de contaminação por ruído rosa.

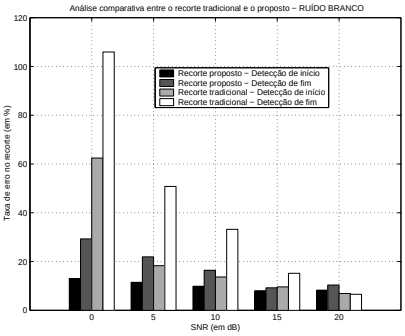


Fig. 7. Análise na condição de contaminação por ruído branco.

5. CONCLUSÕES

Neste trabalho foi proposto um método de recorte robusto baseado na identificação da natureza do ruído por HMM, associado à *SNR* e à distância euclidiana da log-energia, que foram utilizadas para ajustar a detecção em trechos de voz com baixa energia. Com este método buscamos detectar os frames que possuem somente ruído e separá-los da informação útil do sinal. Os resultados de um teste com 121 locuções contaminadas pelos quatro tipos de ruídos com *SNR* de 0 a 20dB foi apresentado. A redução média da taxa de erro no recorte em relação ao método de [1], foi de até 12,23% na detecção de início e de até 26,24% na detecção de fim, no caso de ruído rosa. Nas demais situações de contaminação os resultados foram mais modestos, apresentando grandes oscilações na performance, apenas quando processado com ruído no interior do carro. Como trabalho futuro, pretendemos tornar a detecção de extremos mais eficiente em ambientes com ruídos de baixa freqüência, semelhantes ao ruído no interior do carro.

6. REFERÊNCIAS BIBLIGRÁFICAS

[1] R. Teruszkin, T. A. Consort, and F. G. V. Resende Jr., "Endpoint detection analysis for an implementation of a speech recognition system applied to robot control," *In Proc. SAWCAS*, November 2001.

[2] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environment with application to automatic speech recognition," *In ICASSP*, May 2002.

[3] L. Gu, J. Gao, and J. G. Harris, "Endpoint detection in noisy environment using a poincaré recurrence metric," *In ICASSP*, April 2003.

[4] Jia-Lin Shen, Jeih-Weih Hung, and Lin-Shan Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," *In ICSLP*, 1998.

[5] D. C. Silva and F. G. V. Resende Jr., "Identificação da natureza do ruído baseada em HMM," *In SBrT*, September 2004.

[6] R. Teruszkin, F. G. V. Resende Jr., S. B. Villas-Boas, and F. Lizarralde, "Biblioteca orientada a objeto para reconhecimento de voz e aplicação em controle de robô," *In CBA*, September 2002.

[7] L. Lin, W. H. Holmes, and E. Ambikairajah, "Subband noise estimation for enhancement using a perceptual Wiener filter," *In ICASSP*, April 2003.

[8] http://spib.rice.edu/pib/select_noise.html, "Signal processing information base (SPIB)," September 2002.