MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
SECRETARIA DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM ENGENHARIA ELÉTRICA

César Augusto Medina Sotomayor

Realce de Voz Aplicado à Verificação Automática de Locutor

Rio de Janeiro
2003

INSTITUTO MILITAR DE ENGENHARIA

CÉSAR AUGUSTO MEDINA SOTOMAYOR

REALCE DE VOZ APLICADO À VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

Dissertação de Mestrado apresentada ao Curso de Mestrado em Engenharia Elétrica do Instituto Militar de Engenheria, como requisito parcial para obtenção do título de Mestre em Ciências em Engenharia Elétrica.

Orientador: Prof. José Antonio Apolinário Jr. - D. Sc.
Co-orientador: Prof. Abraham Alcaim - Ph. D.

Rio de Janeiro
2003

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

M489 Medina S., César A.
Realce de Voz Aplicado à Verificação Automática de Locutor, César Augusto Medina Sotomayor, Rio de Janeiro, Instituto Militar de Engenharia, 2003.
91 p.:il, graf., tab.

Dissertação (mestrado) – Instituto Militar de Engenharia, Rio de Janeiro, 2003.

1. Verificação de Locutor, Realce de Voz, Denoising. I. Instituto Militar de Engenharia. II. Título.

CDD 006.454

INSTITUTO MILITAR DE ENGENHARIA

CÉSAR AUGUSTO MEDINA SOTOMAYOR

REALCE DE VOZ APLICADO À VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

Dissertação de Mestrado apresentada ao Curso de Mestrado em Engenharia Elétrica do Instituto Militar de Engenheria, como requisito parcial para obtenção do título de Mestre em Ciências em Engenharia Elétrica.

Orientador: Prof. José Antonio Apolinário Jr. - D. Sc.

Co-orientador: Prof. Abraham Alcaim - Ph. D.

Aprovada em 27 de Junho de 2003 pela seguinte Banca Examinadora:

Prof. José Antonio Apolinário Jr. - D. Sc. (COPPE/UFRJ) do IME - Presidente

---

Prof. Abraham Alcaim - Ph. D. (Imperial College London) do CETUC/PUC–Rio

---

Prof. Sergio Lima Netto - Ph. D. (University of Victoria) da COPPE/UFRJ

Rio de Janeiro
2003

A Deus, aos meus pais e aos meus irmãos;
fonte da minha vida.

AGRADECIMENTOS

Ao meu orientador, TC Prof. José Antonio Apolinário Jr., pela grande amizade e incentivo para realizar com êxito meu trabalho.

Ao professor Abraham Alcaim da PUC–Rio, meu co-orientador, por sua valiosa e incondicional ajuda no desenvolvimento desta dissertação.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) que me facilitou esta oportunidade tão grande na minha vida.

Aos colegas de turma e pesquisa, por todo o apoio prestado e pelo convívio amigável.

A todos os familiares e amigos, em especial a Carlos, Bruno e Luciana, que fizeram do mestrado um eterno e agradável momento.

“O homem encontra a Deus atrás de cada porta que a ciência consegue abrir.”
Albert Einstein.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

ABREVIATURAS

ANC	-	Adaptive Noise Canceller
AR	-	Autoregressive Linear Random Processs
ARMA	-	Autoregressive Moving–average Linear Random Process
ASV	-	Automatic Speaker Verification
CAP	-	Constrained Affine Projection
CBNDR-LMS	-	Constrained Bi–Normalized Data Reusing Least Mean Square
CCG	-	Constrained Conjugate Gradient
CLMS	-	Constrained Least Mean Square
CMS	-	Cepstral Mean Subtraction
CNLMS	-	Constrained Normalized Least Mean Square
CQN	-	Constrained Quasi–Newton
CRLS	-	Constrained RLS
DCT	-	Discrete Cosine Transform
DET	-	Detection Error Tradeoff
DFT	-	Discrete Fourier Transform
EM	-	Expectation Maximization
ERR	-	Equal Error Rate
FC	-	Freqüência Central
GMM	-	Gaussian Mixture Model
GSC	-	Generalized Sidelobe Canceller
HMM	-	Hidden Markov Models
LCMV	-	Linearly Constrained Minimum Variance
LP	-	Linear Prediction
LPC	-	Linear Prediction Coefficients
MA	-	Moving–average Linear Random Process
MB	-	Membrana Basilar
NBTC	-	National Biometric Test Center
NIST	-	National Institute of Standards and Technology
RAL	-	Reconhecimento Automático de Locutor

ROC	-	Receiver Operating Characteristics
SNR	-	Relação Sinal–Ruído
SPL	-	Sound Pressure Level
UBM	-	Universal Background Model
VAL	-	Verificação Automatica de Locutor

RESUMO

Os sistemas de verificação automática de locutor (VAL) independente do texto têm experimentado um grande avanço tecnológico nos últimos anos. Entretanto, o seu desempenho em ambientes reais ainda é limitado pela degradação dos sinais de voz em presença de ruído aditivo. Embora seja um tópico importante de estudo, encontram-se poucos trabalhos na literatura científica referentes ao efeito do ruído nos sistemas de verificação de locutor. Nesta dissertação são apresentados vários algoritmos de realce de voz para serem usados na etapa de pré–processamento de um sistema VAL visando diminuir a taxa de erro presente em tais sistemas.

O sistema VAL que foi implementado para avaliar os algoritmos de realce de voz é o modelo de misturas gaussianas (GMM), já amplamente difundido e recomendado para este tipo de sistemas. Este sistema foi treinado com características $mel-cepestrais$. Estas características têm-se mostrado apropriadas para as tarefas de verificação/identificação de locutor pelo fato de que levam em consideração informação de percepção auditiva do ouvido humano. O sistema foi avaliado usando uma base de dados em língua portuguesa composta de 50 locutores masculinos.

Dois tipos de algoritmos de realce de voz são apresentados: os derivados da subtração espectral e os baseados em wavelets. Da análise destes últimos algoritmos, uma primeira proposta é apresentada: a de realizar o realce da voz com um método baseado em wavelets e redes neurais. Uma primeira análise de desempenho baseada numa medida objetiva, o ganho em decibéis de razão sinal-ruído, permitiu-nos afirmar que afirmar que este método é superior aos métodos tradicionais de realce de voz com wavelets. Da análise subjetiva, ressalta-se a ausência de ruído musical neste algoritmo; no caso dos algoritmos derivados da subtração espectral, esta é uma fonte de desconforto.

Após avaliado o desempenho dos algoritmos de realce de voz no sistema VAL, uma segunda proposta é apresentada. Ela é baseada na adição de ruído ao sinal de treinamento, este ruído é modelado a partir de uma estimativa do ruído presente no sinal de teste. As simulações foram realizadas com diferentes tipos de ruído e variando a relação sinal–ruído. Os resultados obtidos indicam taxas de erro que tornam o sistema implementável na prática. Os resultados obtidos apresentam uma notável redução na taxa de erros tornando assim o sistema VAL implementável em ambientes ruidosos.

ABSTRACT

In the last years, Automatic Speaker Verification (ASV) systems have experimented a great technological evolution. Nevertheless, their performance in real environments is limited by the degradation of the signals. Furthermore, there are, in the technical literature, only few works about the noise effects in ASV systems. This dissertation presents the classical speech enhancement algorithms to be used as a preprocessing stage of the ASV system aiming the reduction of the Equal Error Rate (EER) in this type of systems.

In order to evaluate the speech enhancement algorithms, the widely employed Gaussian Mixture Model (GMM) based ASV system was implemented. The mel-cepstrum characteristics were chosen. This feature incorporates human perception models and has shown better performance than others. The evaluation of the system was carried out with a Brazilian Portuguese data base composed by 50 male speakers.

Two types of algorithms were used: the spectral subtraction – based speech enhancement and the wavelet based denoising. By analyzing the performance of the last technique, a first proposal is introduced: a speech enhancement using wavelets and neural networks. Objective performance evaluation suggests the superiority of the proposed when compared to the traditional wavelet–based methods. A preliminary evaluation based on an objective measure, the gain in SNR_{dB}, has shown the superiority of the proposed method when compared to traditional wavelet–based speech enhancement methods. Subjective evaluation has shown the absence of musical noise, a typical and uncomfortable artifact always present in spectral subtraction techniques.

After the performance evaluation of the speech enhancement algorithms in the ASV systems, a second proposal is presented: it is based on the addition of noise to the training signal. This noise is modeled by an estimate of the noise present in the test signal. The simulations were performed with different types of noise and different signal-to-noise-ratios. The results present a remarkable reduction in the error rates thus making the ASV system implementable in noisy environments.

1 INTRODUÇÃO

1.1 INTRODUÇÃO

Segundo o "National Biometric Test Center" (NBTC), o termo "autenticação biométrica" refere-se à identificação automática ou verificação de identidade das pessoas usando características fisiológicas ou comportamentos típicos (NBTC, 2000). Os sistemas biométricos atendem duas finalidades diferentes, que são:

1) Provar que uma pessoa é quem disse ser;

2) Provar que uma pessoa não é quem disse ser.

No primeiro caso, o usuário pede uma identificação positiva de identidade e o sistema verifica automaticamente, mediante a comparação da amostra submetida e a amostra existente no sistema, se a pessoa é quem disse ser. O propósito deste tipo de sistema é o de prevenir o uso de uma identidade por múltiplos usuários. Chamaremos este processo de *Verificação*. No segundo caso, de identificação negativa, o sistema estabelece se o usuário não é alguém do sistema ou não pertence ao grupo de usuários do sistema. O propósito deste sistema é de evitar o uso de múltiplas identidades por uma mesma pessoa. Chamaremos este sistema de *Identificação*.

Um sistema de autenticação biométrica completo deve, além de apresentar limiares de segurança adequados, ser desenvolvido visando ao melhor atendimento dos seguintes parâmetros:

→ Robustez - Deve ser repetitivo, quer dizer, os resultados não devem variar muito perante mudanças próprias dos usuários (ex. idade) ou do ambiente de autenticação (ex. ambientes com ruído);

→ Distinção - Deve existir grandes diferenças entre as amostras da população;

→ Fácil Acesso - Os sensores devem ser colocados em lugares de fácil acesso;

→ Aceitável - Não deve tirar conforto ao usuário; não pode ser intrusivo;

→ Disponível - Todo usuário deve ter capacidade de prover ao sistema a informação solicitada.

No mercado atual, existem diversas tecnologias de autenticação; cada uma delas apresentando diferente desempenho a respeito dos parâmetros antes mencionados. Entre as tecnologias mais comuns estão: reconhecimento facial, impressões digitais, impressão palmar, geometria da mão, geometria do dedo, escaneamento da íris, escaneamento da retina, termografia facial, reconhecimento (verificação/identificação) de locutor e reconhecimento dinâmico de assinatura.

O uso de uma tecnologia particular depende da aplicação e do tipo de usuário que estamos tratando. A tecnologia de verificação/identificação de locutor tem sido muito usada ultimamente em diversas aplicações, dentre as quais podemos destacar: uso em aparelhos de comunicações móveis, e-commerce, sistemas de segurança, sistemas forenses e sistemas de auxílio a deficientes físicos. Por isso, é de nosso interesse o Reconhecimento Automático de Locutor (RAL) ou, mais especificamente, a Verificação Automática de Locutor (VAL).

1.2 OBJETIVO DA DISSERTAÇÃO

O desempenho de um sistema de reconhecimento de locutor é afetado por diversos fatores como (WOODLAND, 1998): variações intra-locutor, variações inter-locutor, estilo de falar e distorções acústicas. É importante que o RAL seja robusto aos diferentes fatores que alteram seu desempenho, como por exemplo as distorções acústicas causadas por ruído aditivo, tal como vozes de fundo ou outro ruído sonoro qualquer.

O objetivo desta dissertação é, precisamente, melhorar o desempenho do VAL em presença de ruído aditivo. Isso é realizado através do realce da voz na etapa de pré–processamento, como uma técnica para dar robustez ao sistema.

1.3 ESTADO DA ARTE

Nos últimos anos, o reconhecimento de locutor, principalmente a tarefa de verificação de locutor, tem experimentado um grande número de técnicas de extração de características, de técnicas de modelagem de locutor e de métodos de avaliação. Estas técnicas estão bem documentadas em vários tutoriais (ATAL, 1976; ROSEMBERG, 1976; DODDINGTON, 1985; CAMPBELL JR., 1997). Uma referência aceita pela comunidade científica para

conhecer a tecnologia utilizada e o seu desempenho no reconhecimento de locutor, é dada pelo *National Institute of Standards and Technology* (NIST), que anualmente faz um concurso de âmbito mundial, ditando as regras e fornecendo o banco de dados para avaliação dos sistemas. Atualmente o modelo de misturas gaussianas (GMM) tem sido uma ferramenta muito utilizada na verificação de locutor independente do texto (REYNOLDS, 1995, 2000b). Comparações de GMM com outros sistemas de reconhecimento de locutor podem ser vistos em (REYNOLDS, 1992; FURUI, 1996; LIMA, 2001).

Em (REYNOLDS, 2000a) é apresentado, para a tarefa de verificação de locutor, um estudo comparativo entre o desempenho do computador versus o desempenho do ser humano. A experiência mostra que o humano apresenta um erro 15% superior ao erro do computador, quando as condições de gravação dos sinais de treinamento e de teste foram as mesmas. Se existe um descasamento entre as condições de gravação e presença de ruído aditivo, o erro no homem e no computador aumenta, mas o erro do ser humano é 44% inferior ao erro do computador, indicando a falta de robustez do sistema em situações reais. Nesta experiência usaram-se 3 segundos de voz para teste e 1 minuto de voz para treinamento.

A robustez do sistema logra-se da compensação nos diferentes estágios de processamento. A compensação pode ser feita no pré–processamento, na extração de características, no sistema classificador ou na medida de similaridade. Para os sistemas robustos ao ruído, na fase do pré–processamento a compensação é realizada através de sistemas de realce da voz por meio de: subtração espectral, canceladores adaptativos de ruído (GABREA, 2001) e beamforming (MCCOWAN, 2001). Na fase da extração de características e do sistema classificador, a compensação pode ser realizada mediante o uso de modelos matemáticos que integram as características estatísticas da voz e do ruído (ROSE, 1994). Já para a medida de similaridade, existem compensações usando uma adequada combinação de várias medidas de similaridade (SOLEWICZ, 2001). No melhor dos casos, estes sistemas atingem um taxa de erro (Equal Error Rate) de até $18,1\%$ em verificação de locutor (SOLEWICZ, 2001) e de $71,4\%$ de acerto em sistemas de identificação de locutor (ROSE, 1994), para sinais de teste de $SNR = 10\,dB$ em presença de ruído branco.

1.4 CONTRIBUIÇÃO DESTA DISSERTAÇÃO

Esta dissertação apresenta uma introdução aos sistemas de reconhecimento de locutor e a cada um dos seus componentes. São apresentadas diferentes técnicas de realce de voz juntamente com suas avaliações objetivas.

Duas novas propostas são apresentadas. A primeira é um método para realizar o realce de voz baseado em wavelets e redes neurais. A segunda proposta é um esquema de modelagem do ruído a partir de uma estimativa do ruído presente no sinal de teste; este modelo do ruído permite a geração de um outro para ser somado ao sinal de treinamento e desta forma diminuir a taxa de erro atingida pelo sistema. É realizada uma comparação do desempenho das duas propostas em conjunto versus o desempenho atingido por outros algoritmos no reconhecimento de locutor usando sinais de treinamento e teste em diferentes condições.

1.5 ORGANIZAÇÃO DA DISSERTAÇÃO

No Capítulo 2 da dissertação serão apresentados os conceitos básicos do sistema de reconhecimento de locutor. Incluem-se neste capítulo, as fases de pré–processamento, extração de características e o sistema de classificação usado. No Capítulo 3, serão apresentados os algoritmos tradicionais de realce de voz –baseados em subtração espectral– assim como alguns dos mais recentes, baseados em Wavelets. No Capítulo 4 uma nova proposta de realce de voz, também baseada em wavelets, é apresentada bem como a comparação objetiva desta com as técnicas anteriores.

Os resultados das simulações, juntamente com uma nova proposta de modelagem de ruído para corromper os sinais de treinamento no sistema de VAL, estão apresentados no Capítulo 5. Finalmente, algumas conclusões e recomendações são resumidas no Capítulo 6.

2 O RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

2.1 INTRODUÇÃO

A habilidade de reconhecer pessoas através da voz é conhecida como reconhecimento de locutor. Em muitas áreas de processamento de voz é difícil atingir o desempenho do ser humano; contudo, existem evidências que sugerem que no reconhecimento de locutor realizado por computadores (reconhecimento automático de locutor), pode-se chegar a superar o reconhecimento realizado por uma pessoa (ATAL, 1976). Entre as aplicações deste tipo de problema, podemos citar as transações eficientes em negócios, o controle de acesso de informação a indivíduos selecionados e o seu uso como ferramenta para o melhor desempenho das leis (aplicações forenses).

O processo de reconhecimento de locutor envolve a comparação de características previamente armazenadas com as características extraídas da voz que queremos analisar. Esta comparação é realizada em várias etapas, como indica a FIG. 2.1.

Cada uma destas etapas será estudada neste capítulo.

2.2 PRÉ-PROCESSAMENTO

O pré-processamento é efetuado sobre o sinal de voz e serve para adequar o sinal para algum processamento posterior; pode ser a redução na taxa de amostragem, a eliminação de algum trecho inconveniente da gravação, uma filtragem ou normalização, entre outros. Nós usamos o janelamento, a pré-ênfase e a eliminação de silêncio nos sinais.

2.2.1 JANELAMENTO E RECONSTRUÇÃO PERFEITA

Entende-se como janelamento o processo que extrai um certo número de amostras do sinal $x(n)$ e as multiplica por uma função janela, $w(n)$, para depois serem analisadas. O janelamento é realizado devido ao fato de que a voz é um processo não estacionário, mas a intervalos pequenos, de $30ms$ ou $40ms$, pode ser tratada como um sinal localmente estacionário. Isto ocorre pois o trato vocal muda de forma lentamente com o passar do tempo (OPPENHEIM, 1989). O processo de janelamento é ilustrado na FIG. 2.2.

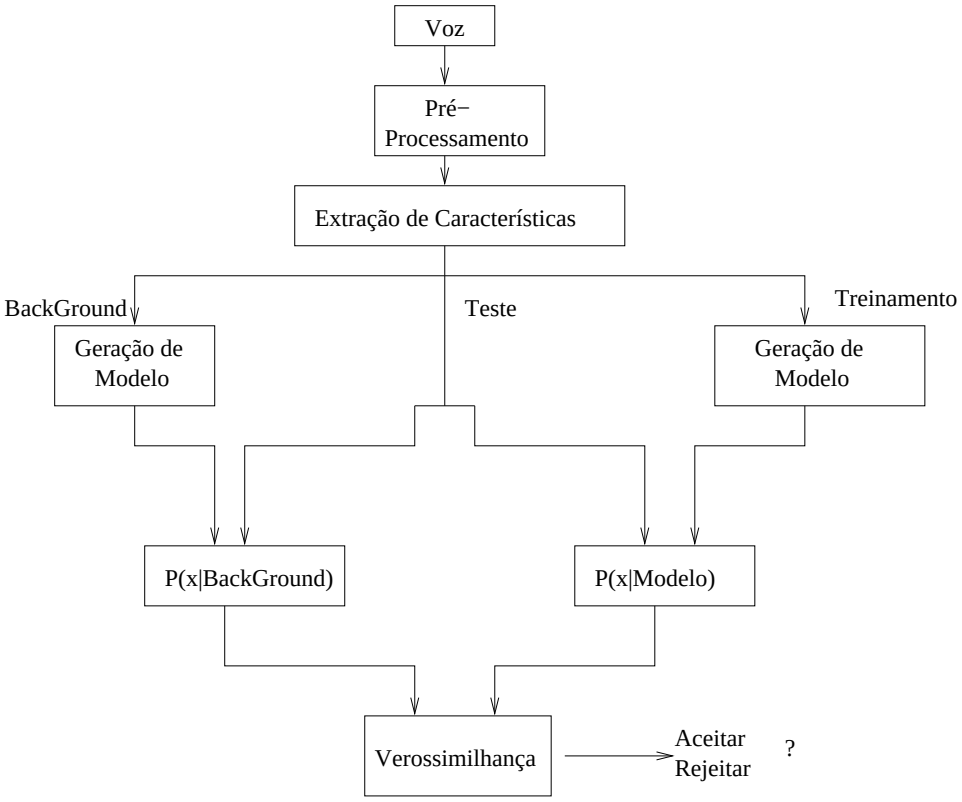

FIG. 2.1: Etapas Gerais no Reconhecimento Automático de Locutor.

Do janelamento depende a habilidade para diferenciar sinais senoidais que estão muito próximos na freqüência (OPPENHEIM, 1989). Para aumentar esta habilidade e diminuir o fenômeno de *Gibbs* devido ao truncamento do sinal de análise no domínio do tempo, deve-se utilizar janelas que possuam, no domínio da freqüência, um lóbulo principal o mais estreito possível e uma grande diferença de amplitude entre o lóbulo principal e o primeiro lóbulo lateral. No processamento de voz é comum o uso da janela de *Hamming* que apresenta um adequado compromisso na resolução tempo-freqüência. A janela de *Hamming* é dada pela seguinte função

$$w(n) = \begin{cases} 0,54 - 0,46\cos(2\pi n/N), & 0 \leq n \leq N; \\ 0, & \text{caso contrário.} \end{cases} \tag{2.1}$$

onde N é a ordem da janela (o tamanho é $N + 1$ amostras).

Normalmente uma superposição entre janelas é efetuada para assim aumentar a correlação entre amostras de janelas adjacentes. Na FIG. 2.2, a superposição está indicada com a letra R e normalmente é expressa como porcentagem do tamanho total da janela.

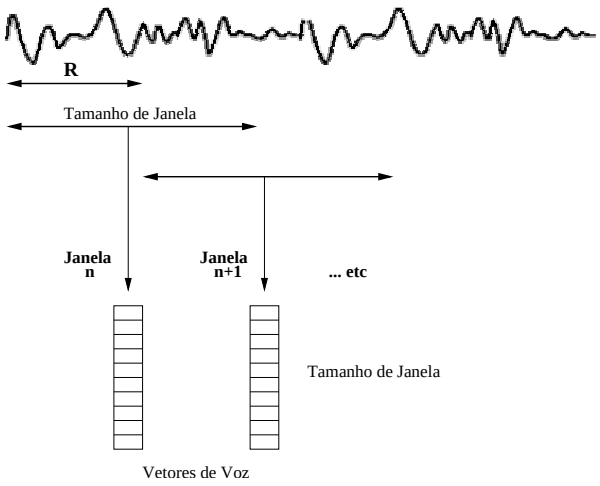

FIG. 2.2: Processo de janelamento de um sinal.

O processo inverso, ou de reconstrução do sinal é descrito a seguir e foi tomado de (RABINER, 1978). Considerando a seqüência depois do janelamento do sinal como $y_r(m) = x(m)w(rR-m)$, onde R é a superposição entre janelas e r é um inteiro; define-se $Y_r(e^{jw_k}) = X_{rR}(e^{jw_k})$, sendo que , $0 \leq k \leq N-1$.

O método de "overlap add" está baseado na seguinte equação

$$y(n) = \sum_{r=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=0}^{N-1} Y_r(e^{jw_k}) e^{jw_k n} \right] \tag{2.2}$$

Isto é, para reconstruir o sinal, a inversa de $Y_r(e^{jw_k})$ é calculada para cada valor de r obtendo-se as seqüencias

$$y_r(m) = x(m)w(rR-m) \quad -\infty < m < \infty \tag{2.3}$$

O sinal no tempo n é obtido somando os valores de todas as seqüencias $y_r(m)$ que se superpõem no tempo n. Isto é

$$y(n) = \sum_{r=-\infty}^{\infty} y_r(n) = x(n) \sum_{r=-\infty}^{\infty} w(rR-n) \tag{2.4}$$

Pode-se demonstrar que se $w(n)$ tem largura de banda limitada e $X_n(e^{jw_k})$ está apropriadamente amostrada no tempo, i.e., R é pequena o suficiente para evitar aliasing no tempo. Logo

$$\sum_{r=-\infty}^{\infty} w(rR-n) \approx W(e^{j0})/R \tag{2.5}$$

sem importar o valor de n. Para uma janela de *Hamming* de N-pontos, $R \leq N/4$.

Assim, a EQ. 2.4 pode ser escrita como

$$y(n) = x(n)W(e^{j0})/R \tag{2.6}$$

demonstrando que a regra de sínteses da EQ. 2.2 resulta em reconstrução perfeita de $x(n)$ (exceto por uma constante multiplicativa) somando as seções que se superpõem na forma de onda.

2.2.2 PRÉ-ÊNFASE

O processo de pré-ênfase corresponde a aplicar um filtro que incrementa a energia relativa das altas freqüências. Normalmente, o filtro usado é

$$P(z) = 1 - \mu z^{-1} \tag{2.7}$$

onde μ é o coeficiente de pré-ênfase.

O filtro é idêntico em forma ao filtro usado para modelar as perdas por radiação nos lábios (DELLER JR., 2000). Ele introduz um zero perto de $w = 0$ e um deslocamento de $6dB$ por oitava no espectro da voz.

Existem duas razões para usar o filtro de pré-ênfase (DELLER JR., 2000). Primeiro, tem-se argumentado que o componente de fase mínima do sinal na glote pode ser modelado com um filtro simples de dois pólos reais perto de $z = 1$. Portanto, dadas as características de perda por radiação nos lábios, com seus zeros perto de $z = 1$, as duas tendem a cancelar os efeitos espectrais de um dos pólos da glote. Introduzindo um segundo zero perto de $z = 1$, as contribuições da laringe e os lábios podem ser efetivamente eliminados e a análise da voz será realizada somente sobre os efeitos introduzidos pelo trato vocal.

A segunda razão para a pré-ênfase é para prevenir a instabilidade numérica no processo de obtenção das características da voz, especialmente, no caso dos coeficientes LPC pelo método de autocorrelação (DELLER JR., 2000).

Para sons sonoros, $0,9 \leq \mu \leq 1,0$. Para sons surdos, é melhor escolher μ próximo de 0. Normalmente, para todo tipo de som é escolhido um valor próximo de 1. O valor ótimo para μ é dado por (DELLER JR., 2000)

$$\mu = \frac{r_s(1;m)}{r_s(0;m)} \tag{2.8}$$

onde $r_s(\eta;m)$ é a autocorrelação do sinal s da janela m no *lag* η.

2.2.3 ELIMINAÇÃO DE SILÊNCIO

Na extração de características da voz, somente se deseja processar trechos que contêm informação. Isto significa que os trechos de silêncio devem ser eliminados, sem que isto altere os resultados do sistema de reconhecimento de locutor. Além disso, alguns algoritmos de realce da voz precisam desta ferramenta.

A extração de silêncios é realizada através um classificador de voz que pode diferenciar entre sons sonoros, surdos ou silêncio. Neste trabalho foi usado o seguinte classificador:

Classificador Baseado nas Características Temporais do Sinal

Um dos algoritmos mais simples é baseado nas seguintes características:

- Energia do sinal;
- Taxa de cruzamentos por zero.

O algoritmo mais conhecido é o proposto em (RABINER, 1974), porém, foi implementado uma modificação deste algoritmo, proposta em (LIMA, 2001), e que pode ser resumida como segue.

O algoritmo é baseado na estimação da amplitude média do sinal. Os 100 ms iniciais e 30 ms finais da locução são considerados como ruído de fundo. Desta estimação inicial e final, é calculada a média e o desvio padrão. Um limiar do valor da amplitude média é estipulado e todo sinal abaixo deste limiar é considerado como ruído de fundo. Considera-se sempre 3 janelas adjacentes para evitar ruídos espúrios.

O inconveniente deste método é que ele necessita ter no começo e no fim do sinal somente ruído, para daí extrair as características estatísticas dele. Sua vantagem é seu baixo custo computacional.

Os passos do algoritmo são:

a) Cálculo da amplitude média e desvio padrão das primeiras e últimas janelas da voz;

b) Com a média e o desvio padrão do passo anterior, se define o limiar: limiar = média + $0,5\cdot$ desvio padrão;

c) Comparação das amplitudes médias de três janelas consecutivas com o limiar estimado: se as amplitudes médias das três janelas estiverem acima do limiar, a primeira janela é marcada como trecho de voz; ao contrário, se as amplitudes médias de três janelas consecutivas não estiverem acima do limiar, a primeira janela é marcada como trecho de silêncio.

2.3 EXTRAÇÃO DE CARACTERÍSTICAS

As características usadas na análise da voz deverão atender, na medida do possível, às seguintes condições (ATAL, 1976):

a) Eficientes na representação da informação do locutor ou do texto;

b) Fáceis de medir;

c) Estáveis ao longo do tempo;

d) Ocorrerem naturalmente e freqüentemente na voz;

e) Mudarem pouco de um ambiente de gravação para outro;

f) Não serem vulneráveis à mímica ou imitação de voz de uma pessoa por outra (para o reconhecimento de locutor).

Na prática, a satisfação simultânea de todos os requisitos acima é aparentemente impossível de ser alcançada. No entanto, para determinadas aplicações como reconhecimento de locutor, é admissível o relaxamento parcial da exigência referente a algumas características (BEZERRA, 1994).

Neste trabalho, a modelagem do sinal de voz foi feita usando as características $mel-cepestrais$.

2.3.1 CEPESTRO

O cepestro de um sinal é uma transformação homomórfica. É uma operação que transforma a convolução em soma e pode ser calculada, como

$$c(n) = \mathcal{F}^{-1}\left[\log\left[\left|\mathcal{F}\left[x(n)\right]\right|\right]\right] \tag{2.9}$$

isto é, o cepestro é a transformada inversa de Fourier do logaritmo do módulo da transformada de Fourier de um sinal $x(n)$. Se o sinal $x(n)$ resultar de um processo de convolução, então a transformada de Fourier deste sinal representará uma multiplicação no domínio da freqüência, o logaritmo transformará este produto em uma soma, comprimindo de certa forma o sinal. A transformada inversa de Fourier retorna o cepestro ao domínio do tempo (domínio de freqüência do cepestro: Qüêfrência).

Na voz, a resposta ao impulso do trato vocal convoluida pela excitação glotal e a radiação nos lábios pode ser separada pelo uso do cepestro (RABINER, 1978), usando-se assim, para o reconhecimento de locutor, só a resposta em freqüência do trato vocal. O cepestro também pode ser utilizado para eliminar o ruído convolucional produzido por um filtro qualquer, para o cálculo dos coeficientes LPC, e para o cálculo do período fundamental de sons sonoros. Por tais motivos, o cepestro é uma característica muito utilizada no processamento de voz (DELLER JR., 2000).

2.3.2 $MEL - CEPESTRO$

Estudos psico–acústicos da percepção auditiva mostram que a escala de freqüências de percepção da voz humana é não–linear. Para cada tom com uma freqüência medida em $Hertz\,(Hz)$, há uma relação com uma freqüência de percepção medida na escala chamada mel. Stevens e Volkmans (STEVENS, 1940) arbitrariamente escolheram a freqüência $1000\,Hz$, $30\,dB$ acima da percepção auditiva, e a fizeram corresponder a $1000\,mels$.

Um mel é a unidade de medida da freqüência de um tom percebido pelo ser humano. Esta freqüência não corresponde linearmente à freqüência física de um tom, da mesma forma que o sistema auditivo humano não percebe um tom de modo linear. Trabalhos realizados por Stevens e Volkmans mostraram que a resolução de freqüência do ouvido é aproximadamente linear abaixo de $1000\,Hz$ e logarítmica acima deste. Um mapeamento da freqüência percebida versus a freqüência real resulta na escala mel, expressa por (PICONE, 1991):

$$mel = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.10}$$

onde f é a freqüência linear em Hz. Este mapeamento é mostrado na FIG. 2.3

A percepção de uma freqüência particular para o sistema humano é influenciada pela energia dentro de uma banda crítica centrada em torno da freqüência em questão. Por esse motivo usam-se filtros de bandas críticas -filtros passa faixas- para calcular o $mel - cepestro$. Alguns pesquisadores sugerem usar a log-energia total encontrada dentro das bandas críticas em torno desta freqüência em vez de usar a log-amplitude (DELLER JR., 2000). Além disso, a largura de banda dos filtros varia com a freqüência; começando por volta de $100\,Hz$ para freqüências abaixo de $1\,kHz$ e aumentando logaritmamente acima de $1\,kHz$. Para o cálculo dos coeficientes $mel - cepestrais$, costuma-se utilizar 20 filtros passa–banda triangulares (DELLER JR., 2000). A FIG. 2.4 mostra os filtro de banda

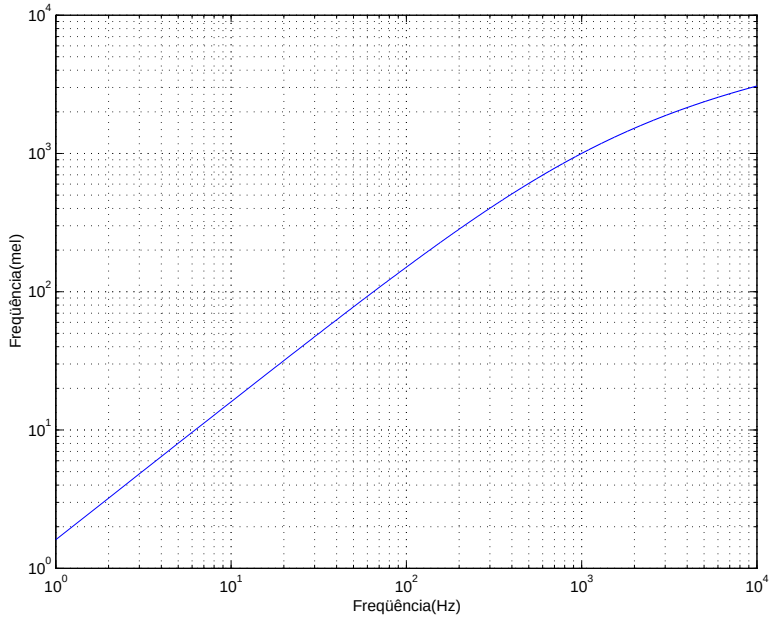

FIG. 2.3: Escala *mel* versus escala de freqüência real.

crítica para o cálculo do $mel - cepestro$ de um sinal amostrado a 8 kHz. Cada filtro é centrado em uma freqüência *mel*, que determina a largura de banda crítica do filtro.

Uma expressão para a largura de freqüência da cada banda crítica é dada por (PICONE, 1991):

$$BW_{crit} = 25 + 75\left[1 + 1,4\,(f/1000)^2\right]^{0,69} \quad Hz \tag{2.11}$$

onde f é a freqüência central de cada filtro (mel).

Segundo (DAVIS, 1980), os coeficientes $mel - cepestrais$, baseados num banco de filtros de banda crítica, podem ser calculados como:

$$MCC_i = \sum_{k=1}^{N} X_k \cos\left[i\left(k - \tfrac{1}{2}\right)\tfrac{\pi}{N}\right] \quad i = 1, 2, \ldots, M \tag{2.12}$$

onde M é o número de coeficientes $mel - cepestrais$, X_k, $k = 1, 2 \ldots, N$, representa a energia logarítmica do k-ésimo filtro e N é o número de filtros do banco de filtros.

O diagrama em blocos que ilustra a extração dos coeficientes $mel - cepestrais$ é apresentado na FIG. 2.5. Inicialmente extrai-se o espectro do sinal por meio da transformada discreta de Fourier (DFT). Em seguida, é calculada a potência espectral, que é filtrada por sua multiplicação por uma série de filtros triangulares espaçados segundo a escala

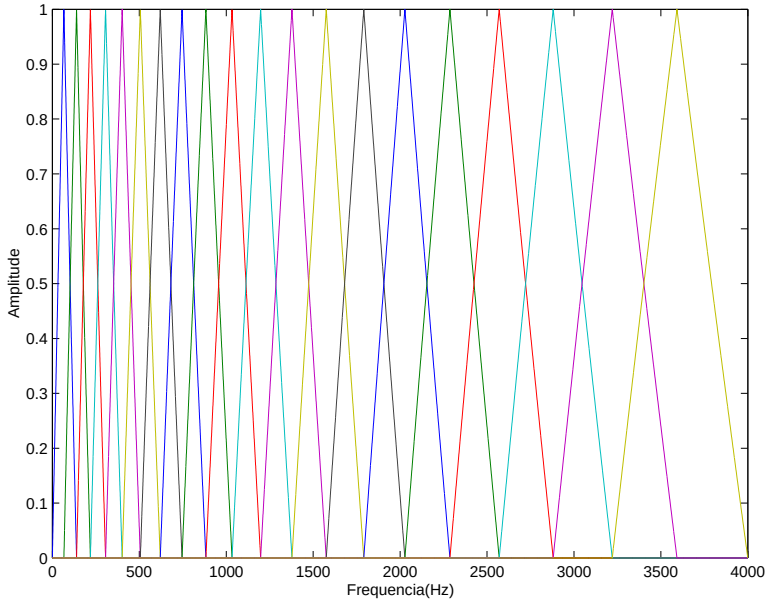

FIG. 2.4: Amplitude do espectro dos filtros de banda crítica utilizados na produção dos coeficientes $mel - cepestrais$.

mel. A energia resultante da filtragem é aplicada a uma função logarítmica e, finalmente, é utilizada a transformada discreta de coseno (DCT) para se obter os coeficientes no domínio do cepestro de freqüência (qüêfrência).

2.4 MODELAGEM ESTATÍSTICA

A modelagem estatística provê um sistema de classificação que resulta na comparação de dois valores para aceitar ou rejeitar um pretenso locutor. Nesta seção, assumimos que os parâmetros do sinal foram gerados por algum processo aleatório multi–variável. Assim, queremos aprender ou descobrir a natureza deste processo. Teremos, então, que impor um modelo aos dados e depois otimizar (ou treinar) o modelo; finalmente, temos que medir (testar) a qualidade da aproximação (medida de similaridade). Para a verificação de locutor, um sistema de classificação irá comparar valores do sinal de entrada com o modelo armazenado do pretenso locutor, aceitando-o ou rejeitando-o.

Existem dois tipos de modelos: o modelo estatístico ou estocástico e o modelo de casamento de padrões (CAMPBELL JR., 1997). No modelo estatístico, o sistema de

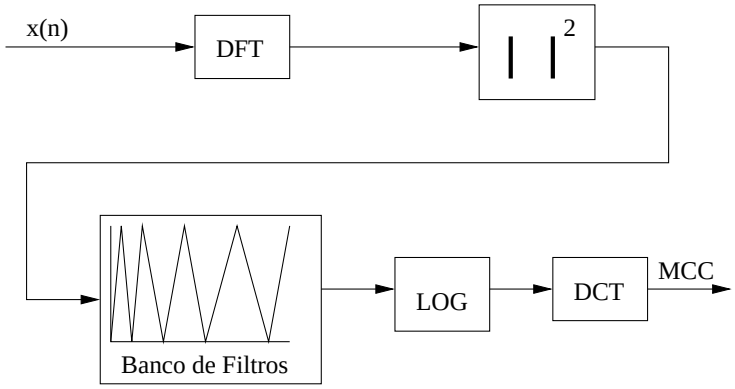

FIG. 2.5: Extração dos coeficientes mel-cepestrais.

classificação é probabilístico e resulta numa medida de verossimilhança ou probabilidade condicional, dada pela observação do modelo. Para modelos baseados em casamento de padrões característicos, o sistema de classificação faz comparações; assume-se que a observação é uma réplica imperfeita do modelo armazenado e, geralmente, é realizado o alinhamento dos quadros do modelo para os quadros observados para minimizar uma medida de distância d.

Neste trabalho, somente usaremos o modelo estatístico, mais especificamente o modelo de misturas gaussianas (GMM) como explicado e justificado a seguir.

2.4.1 O MODELO DE MISTURAS GAUSSIANAS

O modelo de misturas gaussianas (GMM) foi introduzido em (REYNOLDS, 1992) e pode ser visto como um modelo híbrido de dois modelos efetivos para o reconhecimento de locutor: um classificador uni-modal gaussiano e um quantizador vetorial (VQ), combinando a robustez e o amaciamento do modelo gaussiano com a modelagem arbitrária de um modelo VQ não–paramétrico.

As componentes gaussianas podem modelar um amplo conjunto de classes fonéticas, para caracterizar o som produzido por uma pessoa. As cadeias ocultas de Markov (HMM, conforme a sigla em inglês) não modelam somente classes acústicas desconhecidas, mas também a seqüência temporal entre estas classes. Embora a modelagem de estruturas temporais seja vantajosa para a tarefa de reconhecimento de locutor dependente do texto, ela pode limitar o desempenho do HMM em tarefas de reconhecimento de locutor

independente do texto (REYNOLDS, 1992).

O GMM vem sendo atualmente a ferramenta que apresenta uma das melhores respostas na tarefa de verificação de locutor independente do texto e sua utilização é amplamente justificada em termos físicos (modelagem de classes acústicas) e práticos (resultados).

O GMM no Reconhecimento de Locutor

Uma mistura de densidades de probabilidade gaussianas é uma soma ponderada de M densidades, vide FIG. 2.6, dada pela equação

$$p\left(\boldsymbol{x}|\lambda\right) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x}) \tag{2.13}$$

onde $\boldsymbol{x}$ é um vetor aleatório de dimensão N, $b_i(\boldsymbol{x})$, $i = 1, \ldots, M$, são as densidades componentes e p_i, $i = 1, \ldots M$, é a ponderação das misturas. Cada densidade componente é uma função gaussiana de dimensão N da forma:

$$b_i(\boldsymbol{x}) = \frac{e^{\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{K}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}}{(2\pi)^{\frac{N}{2}}\sqrt{|\boldsymbol{K}_i|}} \tag{2.14}$$

com vetor média $\boldsymbol{\mu}_i$ e matriz de covariância $\boldsymbol{K}_i$. A ponderação das misturas satisfaz à condição $\sum_{i=1}^{M} p_i = 1$.

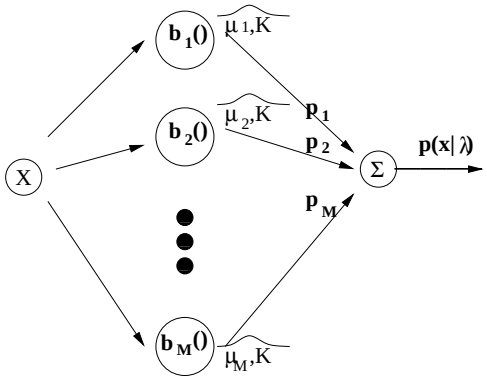

FIG. 2.6: Densidades de probabilidade formando um GMM.

A densidade de misturas gaussianas é parametrizada por um vetor de médias, por uma matriz de covariância e pela ponderação das componentes da mistura (modelo λ). Este parâmetros são representados em conjunto pela notação:

$$\lambda = \{p_i, \boldsymbol{\mu}_i, \boldsymbol{K}_i\} \quad i = 1, \ldots, M. \tag{2.15}$$

O GMM pode ter diferentes formas dependendo da escolha da matriz de covariância. O modelo pode ter uma matriz de covariância distribuída a cada componente gaussiana como indicado na EQ. 2.15 (matriz covariância nodal), uma matriz de covariância para todas as componentes gaussianas para um dado modelo (grande matriz de covariância), ou uma única matriz de covariância para todos os modelos (matriz de covariância global). A matriz de covariância também pode ser completa ou diagonal (REYNOLDS, 1992).

Neste trabalho foi usada, para a modelagem do GMM, uma grande matriz de covariância diagonal. O projeto e a implementação do GMM se faz com o algoritmo de máxima expectativa (*Expectation Maximization* – EM) (REYNOLDS, 1995; VUUREN, 1999).

Sistema de Verificação de Locutor com o GMM

A tarefa de verificação de locutor requer uma decisão binária, o sistema de classificação deve decidir se uma voz é ou não é pertencente a um determinado locutor, cujo modelo (modelo λ) já tenha sido determinado. Considerando uma seqüência de entrada (vetores de características $\boldsymbol{X}$) para verificação, a escolha deve ser feita entre H_0 e H_1, onde:

H_0: $\boldsymbol{X}$ pertencer ao locutor.

H_1: $\boldsymbol{X}$ não pertencer ao locutor.

Para obter uma razão de verossimilhança de teste, que decida entre H_0 e H_1 é usualmente empregado algum modelo do universo de possibilidades falsas, o denominado *background*, que é composto por um conjunto de locutores falsos, agregando assim, informação sobre possíveis impostores (REYNOLDS, 2000b).

Para vetores de características $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, extraídas da locução de um locutor de teste (que será aceitado ou rejeitado), com correspondente modelo λ_L e um modelo não pertencente ao pretenso locutor λ_B, a razão de verossimilhança é dada por:

$$\frac{P(\boldsymbol{X} \text{ pertence ao locutor})}{P(\boldsymbol{X} \text{ não pertence ao locutor})} = \frac{P(\lambda_L|\boldsymbol{X})}{P(\lambda_B|\boldsymbol{X})} \tag{2.16}$$

Aplicando a regra de Bayes e descartando as probabilidades constantes *a priori* para os locutores falso e verdadeiro, a razão de verossimilhança no domínio logaritmo é, então, igual a

$$\Lambda(\boldsymbol{X}) = \log p(\boldsymbol{X}|\lambda_L) - \log p(\boldsymbol{X}|\lambda_B) \tag{2.17}$$

O termo $p(\boldsymbol{X}|\lambda_L)$ é a verossimilhança da locução do pretenso locutor e $p(\boldsymbol{X}|\lambda_B)$ é a verossimilhança dada por um modelo não pertencente ao mesmo locutor (*background*). A razão de verossimilhança é comparada com um limiar θ e o pretenso locutor é aceito se

$\Lambda(\boldsymbol{X}) > \theta$ e rejeitado se $\Lambda(\boldsymbol{X}) \leq \theta$. Este limiar –estimado em testes de laboratório em condições controladas– pode ser: (1) global -estimado com os dados de todos os locutores (independente ao locutor)- usando o resultado de um grande número de testes disponíveis verdadeiros e falsos; (2) dependente do locutor, isto é, cada locutor possui um limiar próprio. Neste segundo caso existe a exigência de uma quantidade maior de informação do locutor (mais tempo de voz) para fornecer um limiar com significado estatístico. A FIG. 2.7, apresenta um diagrama em blocos para a verificação de locutor.

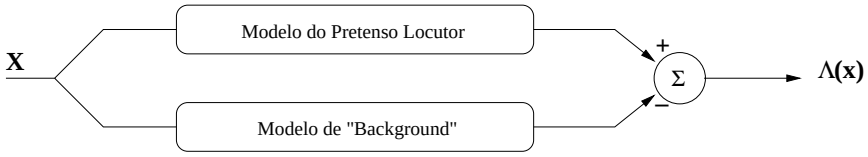

FIG. 2.7: Sistema de verificação de locutor usando GMM.

A verossimilhança para modelagem de um locutor verdadeiro é calculada diretamente através de

$$\log p(\boldsymbol{X}|\lambda_L) = \frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda_L) \tag{2.18}$$

A escala $\frac{1}{T}$ é usada para normalizar a verossimilhança de acordo com a duração da locução (número de vetores de características).

A verossimilhança das locuções não pertencentes ao locutor verdadeiro é formada usando-se o *background* que simula as condições de gravação das locuções, agregando também informações do ruído presentes nas mesmas. O *background* pode ser montado de duas formas: utilizando vários modelos individuais de locutores ou utilizando vários locutores em um único modelo. Este último é denominado *Universal Background Model* (UBM) e é montado utilizando as características de vários locutores modelados por um único GMM. Neste trabalho foi usado o *Universal Background Model*.

2.5 RESUMO

Neste capítulo, estudou-se o sistema completo de reconhecimento de locutor. Foi feita uma apresentação das etapas de pré-processamento, extração de características e do modelo de classificador que será usado nesta dissertação. Porém, o sistema aqui estudado apresenta

problemas quando trabalha com sinais reais. Estes problemas devem-se ao fato de que tais sinais podem estar corrompidos por sistemas convolucionais ou aditivos tais como sistemas com canais de comunicação e sistemas que somam ruído ao sinal original, respectivamente.

Na literatura técnica existem diversos estudos para melhorar o desempenho do reconhecimento de locutor perante o primeiro problema acima mencionado, por exemplo o uso de técnicas como o CMS e o RASTA (HERMANSKY, 1992; SILVA, 2002), que agregam robustez ao sistema.

Para o segundo caso, sinais corrompidos por ruído aditivo, existem poucos estudos e a maioria deles abordando o reconhecimento automático de voz e não de locutor. Nos capítulos seguintes, trataremos deste tipo de problema.

3 ALGORITMOS DE REALCE DE VOZ

3.1 INTRODUÇÃO

Quando o locutor e o ouvinte estão próximos um do outro, em um ambiente sem ruído, a comunicação é geralmente fácil e precisa; porém, quando eles estão distantes ou em ambientes ruidosos, a habilidade para ouvir uma mensagem torna-se difícil. Além disso, a comunicação pode não ser direta mas através de algum meio de transmissão como telefone, telefone móvel celular ou inter-comunicadores; estes meios introduzem diferentes tipos de ruído ou distorções no sinal de origem. Neste capítulo, abordaremos a eliminação de ruído aditivo ou realce de voz, entendendo-se este como o processamento efetuado sobre a voz para melhorar a perceptibilidade, aumentar a qualidade ou diminuir a fadiga auditiva. Os ruídos podem ser de diversos tipos; usaremos neste trabalho o ruído gaussiano branco e três tipos de ruído da base de dados Noisex-92 (VARGA, 1992): *speech like*, cabine de avião e ruído de fábrica.

3.2 CLASSIFICAÇÃO DOS ALGORITMOS DE REALCE DE VOZ

Existem várias formas de classificar os algoritmos de realce da voz. Uma classificação muito abrangente pode ser feita considerando a forma como o sinal de voz é modelado, classificando-os em métodos que modelam a voz como um processo estocástico e métodos que exploram as propriedades perceptíveis da voz.

Outro tipo de classificação pode ser feita em base ao número de fontes ou canais de informação disponíveis; assim, temos técnicas mono-canal e técnicas multi-canal. Nas primeiras só se dispõe do sinal obtido com um único microfone enquanto que para técnicas multi-canal, se dispõe de sinais obtidos de vários microfones estrategicamente localizados.

3.2.1 TÉCNICAS MONO-CANAL

Este tipo de técnicas são cegas no sentido de que, para estimar o sinal original, só conhecemos o sinal corrompido pelo ruído. Nesta dissertação usaremos as técnicas derivadas da subtração espectral e as técnicas baseadas em Wavelets. A facilidade e efetividade destes tipos de algoritmos provocaram um rápido crescimento de seu uso em diversas aplicações

de voz; estes algoritmos serão mais amplamente estudados neste capítulo.

3.2.2 TÉCNICAS MULTI-CANAL

Na literatura técnica encontram-se amplamente difundidos dois tipos de técnicas multi-canal usadas no realce da voz, a técnica de "cancelamento adaptativo de ruído" e a técnica de *beamforming*. Estas duas técnicas podem eliminar ruído estacionário assim como não-estacionário (DELLER JR., 2000).

A primeira técnica é formulada usando dois sinais obtidos simultaneamente de dois microfones e usa a adaptação temporal do filtro cancelador de ruído. Na FIG. 3.1 pode-se observar o esquema usado para este tipo de filtragem. O sinal de referência é gravado através de um microfone isolado física ou acusticamente do microfone principal. A principal desvantagem deste algoritmo é que o filtro precisa de cerca de 1500 *taps* (DELLER JR., 2000); além disso, tal filtro gera um desajuste (*misadjustment*) elevado, o que ocasiona eco na saída do filtro além do elevado custo computacional. O desajuste pode ser reduzido diminuindo o passo de adaptação; entretanto, isto resulta em diminuição do tempo de convergência do filtro (DINIZ, 1997).

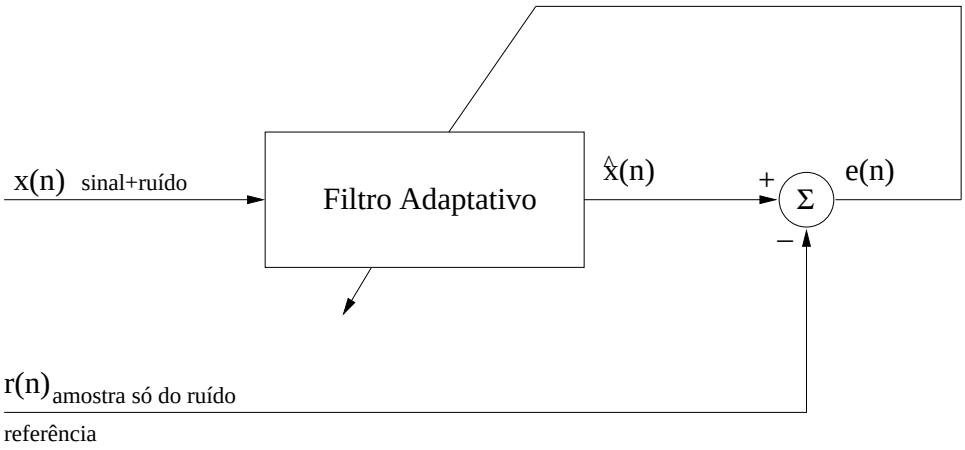

FIG. 3.1: Cancelamento Adaptativo de Ruído.

A segunda técnica empregada é conhecida como *beamforming* ou filtragem espacial. Esta técnica combina, de forma adequada, a saída de vários microfones para assim formar um só microfone altamente diretivo. Tradicionalmente o *beamforming* é uma técnica usada em vários tipos de problemas como (VAN VENN, 1988): radar (controle de tráfego),

sonar (localização de fonte e classificação), comunicações (transmissão direcional, broadcast setorizado em comunicação satelital), imagem (tomografia, ultrasom), exploração geofísica (exploração de petróleo, mapeamento), exploração astrofísica (imagem de alta resolução do universo) e biomedicina (monitoração do coração nos fetos, ajuda auditiva). Porém, os algoritmos desenvolvidos para estas aplicações não são diretamente aplicáveis ao processamento de voz devido a algumas diferenças, entre elas (RODRÍGUEZ, 2000):

- O sinal de voz é considerado de banda larga (6 ou 7 oitavas, considera-se 4 oitavas para canal telefônico), não estacionário e com grandes variações de nível em pouco tempo além de apresentar períodos de silêncios entre as palavras;
- As características espectrais do sinal desejado e o sinal de interferência podem ser muito similares ao longo do tempo;
- A distância da fonte aos sensores (microfones) normalmente é pequena (entre um e poucos metros) fazendo que as suposições de onda plana e fonte pontual não sejam muito reais;
- A reverberação pode ser um problema se a maior parte da energia provém de um caminho indireto;
- A relação sinal–ruído em qualquer canal pode ser positiva antes de qualquer processamento.

O *beamforming* é uma estrutura como indica a FIG. 3.2, onde cada microfone é a entrada de uma *tapped delay line*; a saída de cada linha de N retardos é somada para formar uma só saída que é a versão realçada da voz de entrada. Assumindo que um sinal chega aos microfones com direção θ, então a saída do i-ésimo dos M microfones apresentará um retardo $\Delta_i(\theta)$, ocasionado pela propagação da onda entre os microfones.

O retardo $\Delta_i(\theta)$ depende da geometria do arranjo de microfones, que deve ser escolhido especificamente para a aplicação desejada e para o número de direções ambíguas que queremos suprimir; por exemplo: existem os arranjos lineares que apresentam ambigüidade nas direções perpendiculares do arranjo, e arranjos bi-dimensional onde teremos só uma ambígüidade (FLIELLER, 1998). No arranjo linear, temos que levar em consideração que, para evitar o *aliasing* espacial, é preciso que a distância entre os microfones seja inferior a $\lambda/2$ (HAYKIN, 1986), onde λ é o cumprimento de onda.

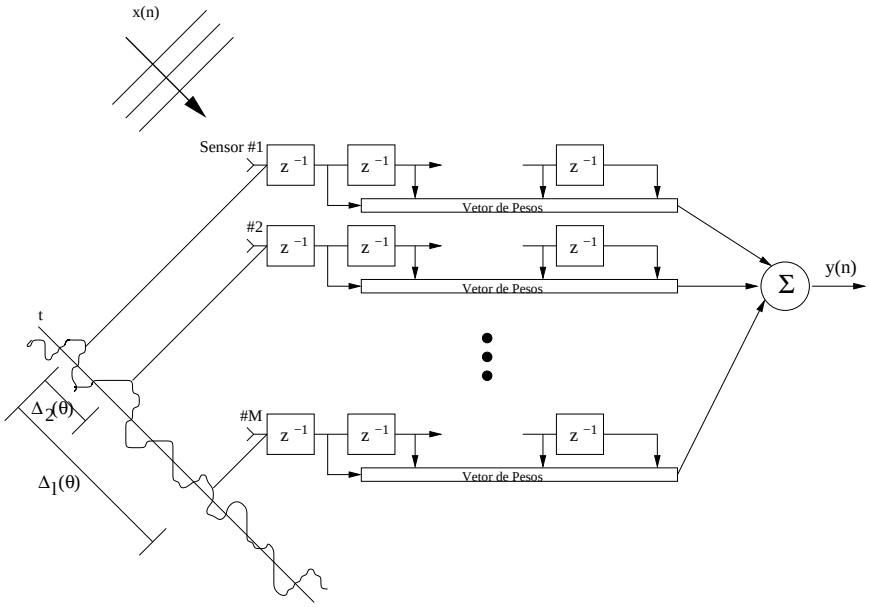

FIG. 3.2: Arranjo de Microfones usado para Realce da Voz.

Basicamente, existem dois tipos de algoritmos para implementar o *beamforming* (VAN VENN, 1988): independentes dos dados e estatísticamente ótimos. Entre estes últimos encontra-se o *Linearly Constrained Minimum Variance* (LCMV), que é estudado a seguir.

A idéia principal do LCMV é de restringir a saída a sinais de uma freqüência, ω, que vêm de uma direção especificada, θ, a um ganho e fase desejadas. Os pesos do filtro são escolhidos para minimizar a energia do sinal de saída, sujeito a um conjunto de restrições lineares. Matematicamente, pode-se formular o problema através da EQ. 3.1.

$$\min_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{R}_{xx} \boldsymbol{w} \quad \text{restringido a } \boldsymbol{C}^T \boldsymbol{w} = \boldsymbol{f} \tag{3.1}$$

onde $\boldsymbol{C}$ é a matriz de restrição e $\boldsymbol{f}$ é o vetor de resposta. Cada uma das colunas de $\boldsymbol{C}$ impõe uma restrição linear ao filtro. Para obter esta equação, foi assumido um sinal de referência nulo, $r(n) = 0$. Em (VAN VENN, 1988; BUCKLEY, 1987) encontra-se descritos métodos apropriados para escolher a matriz $\boldsymbol{C}$ e o vetor $\boldsymbol{f}$. O vetor de pesos ótimo (conhecido como filtro de Wiener), solução da EQ. 3.1, é dado por (BUCKLEY, 1987):

$$\boldsymbol{w}_{\text{opt}} = \boldsymbol{R}_{xx}^{-1} \boldsymbol{C} (\boldsymbol{C}^T \boldsymbol{R}_{xx}^{-1} \boldsymbol{C})^{-1} \boldsymbol{f} \tag{3.2}$$

A EQ. 3.2 é útil só no caso de sinais estacionários. Sendo a voz um sinal não esta-

cionário, é preciso o uso de técnicas de filtragem adaptativa com restrições ou técnicas alternativas como o uso da estrutura conhecida como *Generalized Sidelobe Canceller* (GSC).

Algoritmos com Restrições

Este tipo de algoritmo foi inicialmente proposto por Frost, em 1972, e implementam modificações dos algoritmos sem restrições, mas incorporando-as na solução. Entre os mais comuns estão: *Constrained Least Mean Square* (CLMS) (FROST, 1972), *Constrained Normalized Least Mean Square* (CNLMS) (APOLINÁRIO JR., 1998), *Constrained Bi–Normalized Data Reusing Least Mean Square* (CBNDR–LMS) (APOLINÁRIO JR., 1998), *Constrained Conjugate Gradient* (CCG) (APOLINÁRIO, 2001), *Constrained RLS* (CRLS) (RESENDE, 1996), *Constrained Affine Projection* (CAP) (CAMPOS, 2000) e *Constrained Quasi–Newton* (CQN) (CAMPOS, 1998).

Generalized Sidelobe Canceller (GSC)

Esta estrutura, proposta por (GRIFFITHS, 1982), oferece a vantagem de possibilitar que o algoritmo a ser utilizado não utilize restrições; ainda assim, devido a sua estrutura, as restrições continuam a ser satisfeitas. O algoritmo é explicado a seguir. Usamos uma matriz de transformação $T_{MN\times MN}$ definida como:

$$\boldsymbol{T} = [\boldsymbol{C}\boldsymbol{B}] \tag{3.3}$$

onde $\boldsymbol{B}$ é uma matriz $MN \times (MN-p)$ chamada *Blocking Matrix* que obedece a $\boldsymbol{C}^T\boldsymbol{B} = \boldsymbol{0}$, e p é o número de restrições lineares. O vetor de pesos pode ser escrito como:

$$\begin{aligned} \boldsymbol{w}(k) &= \boldsymbol{T}\overline{\boldsymbol{w}}(k) \\ &= \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \overline{\boldsymbol{w}}_U(k) \\ -\overline{\boldsymbol{w}}_L(k) \end{bmatrix} \\ &= \boldsymbol{C}\overline{\boldsymbol{w}}_U(k) - \boldsymbol{B}\overline{\boldsymbol{w}}_L(k) \end{aligned} \tag{3.4}$$

onde $\overline{\boldsymbol{w}}(k)$ é um vetor de coeficientes transformado. De EQ. 3.4, podemos reescrever a restrição em EQ. 3.1 como $\boldsymbol{C}^T\boldsymbol{w}(k) = \boldsymbol{C}^T\boldsymbol{C}\overline{\boldsymbol{w}}_U(k) - \boldsymbol{C}^T\boldsymbol{B}\overline{\boldsymbol{w}}_L(k) = \boldsymbol{f}$. Se levarmos em consideração que $\boldsymbol{C}^T\boldsymbol{B} = \boldsymbol{0}$, ficamos com $\overline{\boldsymbol{w}}_U(k) = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{f}$, que é a parte fixa do vetor de coeficientes $\overline{\boldsymbol{w}}(k)$, o qual não depende do sinal de entrada. Assim, a adaptação é feita apenas na parte inferior de $\overline{\boldsymbol{w}}(k)$, designada como $\boldsymbol{w}_{GSC}(k) = \boldsymbol{w}_L(k)$. A parte não adaptativa (fixa) de $\overline{\boldsymbol{w}}(k)$ é designada pelo vetor $\boldsymbol{F}$, tal que: $\boldsymbol{F} = \boldsymbol{C}\overline{\boldsymbol{w}}_U(k) = \boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{f}$, ver FIG. 3.3.

Para adaptar $\boldsymbol{w}_{GSC}$, podemos usar qualquer filtro sem restrições multi-canal como por exemplo algum da familia "Multichannel Fast QRD-RLS"; estudados em (**MEDINA**,

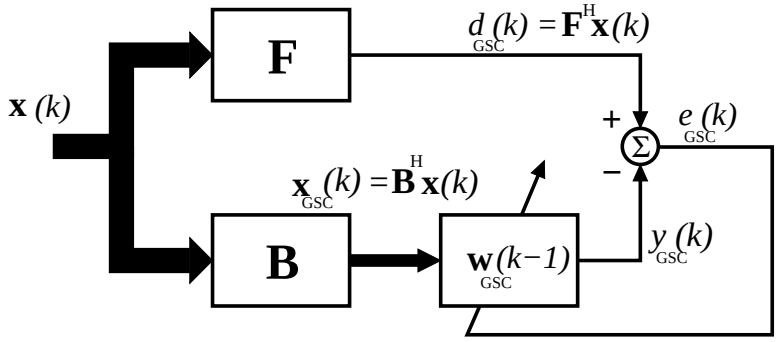

FIG. 3.3: Estrutura **GSC** mais usual (com sinal de referência nulo).

2002b) como extensões ao caso multi-canal da familia mono-canal "Fast QRD-RLS" abordados em (APOLINÁRIO JR., 2002) e (**MEDINA**, 2002a). Estas publicações foram realizadas durante o período de estudo do mestrado mas, pela extensão do tema, está sendo motivo de uma outra dissertação de mestrado do IME.

3.3 O REALCE DA VOZ USANDO TÉCNICAS MONO-CANAL DE SUBTRAÇÃO ESPECTRAL

O primeiro tratado detalhado da subtração espectral foi desenvolvido por Boll (BOLL, 1979b,a) e posteriormente expandido e generalizado por (MCAULAY, 1980; LIM, 1979). Seu fundamento é o seguinte: a voz na presença de ruído aditivo é expressa como:

$$x(n) = s(n) + w(n) \tag{3.5}$$

onde, $s(n)$ é o sinal limpo e $w(n)$ é o ruído adicionado; o problema é estimar $s(n)$ das observações de $x(n)$. Nestes tipos de algoritmos são usualmente feitas as seguintes suposições:

- O ruído, $w(n)$, é um processo aleatório, que na janela analisada, é considerado estacionário;
- O ruído e a voz são sinais descorrelatados;
- O ouvido humano é pouco sensível à fase de um sinal (DELLER JR., 2000).

Com estas condições, o módulo ao quadrado da transformada de Fourier[1] de $x(n)$ pode expressar-se como:

$$|\mathcal{X}(w)|^2 = |\mathcal{S}(w)|^2 + |\mathcal{W}(w)|^2 \tag{3.6}$$

Evidentemente, dado $|\mathcal{X}(w)|^2$ e uma estimativa $|\hat{\mathcal{W}}(w)|^2$ de $|\mathcal{W}(w)|^2$, podemos estimar o espectro de potência de $s(n)$ como sendo:

$$|\hat{\mathcal{S}}(w)|^2 = |\mathcal{X}(w)|^2 - |\hat{\mathcal{W}}(w)|^2 \tag{3.7}$$

Desde que só possuímos uma estimativa de $|\mathcal{W}(w)|$, a diferença dos termos na EQ. 3.7 pode ser negativa, neste caso utilizamos o valor de 0 como estimativa de $|\mathcal{S}(w)|^2$. Lembrando da terceira consideração feita inicialmente, a fase de $\hat{\mathcal{S}}(w)$ é tomada diretamente do sinal com ruído,

$$\phi_{\hat{\mathcal{S}}(w)} = \phi_{\mathcal{X}(w)}. \tag{3.8}$$

Assim, para estimar um sinal no domínio da freqüência usamos a seguinte fórmula:

$$\hat{\mathcal{S}}(w) = \begin{cases} \left(|\mathcal{X}(w)|^2 - |\hat{\mathcal{W}}(w)|^2\right)^{1/2} e^{\phi_{\mathcal{X}(w)}}, & |\mathcal{X}(w)|^2 \geq |\hat{\mathcal{W}}(w)|^2; \\ 0, & \text{caso contrário} \end{cases} \tag{3.9}$$

Normalmente a amplitude de $\hat{\mathcal{S}}(w)$ é expressa como a multiplicação de $|\mathcal{X}(w)|$ com uma função $G(w)$, contruida da EQ. 3.9 da seguinte forma:

$$\begin{aligned} |\hat{\mathcal{S}}(w)| &= \begin{cases} |\mathcal{X}(w)| \left(1 - \frac{|\hat{\mathcal{W}}(w)|^\gamma}{|\mathcal{X}(w)|^\gamma}\right)^{1/\gamma}, & |\mathcal{X}(w)|^{1/\gamma} \geq |\hat{\mathcal{W}}(w)|^{1/\gamma}; \\ 0, & \text{caso contrário} \end{cases} \\ &= |\mathcal{X}(w)| \begin{cases} \left(1 - \frac{|\hat{\mathcal{W}}(w)|^\gamma}{|\mathcal{X}(w)|^\gamma}\right)^{1/\gamma}, & |\mathcal{X}(w)|^{1/\gamma} \geq |\hat{\mathcal{W}}(w)|^{1/\gamma}; \\ 0, & \text{caso contrário} \end{cases} \\ &= |\mathcal{X}(w)| G(w) \end{aligned} \tag{3.10}$$

Onde $\gamma = 2$ para *Subtração Espectral de Potência* (como em EQ. 3.9), $\gamma = 1$ para *Subtração Espectral de Amplitude* (como sugerido por (BOLL, 1979b)) ou γ pode tomar outros valores menos comuns. Contudo, não é um parâmetro crítico do algoritmo (VIRAG, 1999). Nas simulações realizadas neste trabalho foi usado $\gamma = 2$.

[1]Será aqui denominado também de "Espectro de Potência" embora a definição correta deste conceito seja o quadrado da transformada de Fourier da função de autocorrelação do sinal (OPPENHEIM, 1989).

A estimativa $|\hat{\mathcal{W}}(w)|$ é obtida do espectro de potência do sinal corrompido, mas *somente é atualizada nas janelas de silêncio* usando a seguinte fórmula:

$$|\hat{\mathcal{W}}(w)|^2 = \lambda|\hat{\mathcal{W}}(w)|^2 + (1-\lambda)|\mathcal{X}(w)|^2 \tag{3.11}$$

onde $\lambda \approx 1$ é o fator de esquecimento e determina o compromisso entre a variância do espectro estimado e a resposta à mudança de condições estatísticas do ruído.

3.3.1 O ESTIMADOR DA AMPLITUDE ESPECTRAL USANDO MÍNIMO ERRO MÉDIO QUADRÁTICO

(EPHRAIM e MALAH, 1984) apresentaram um estimador da amplitude espectral que usa uma estimativa do SNR *a priori* e que permite reduzir o ruído musical e a distorção do sinal presentes na subtração espectral de potência. Este algoritmo estima a amplitude de cada componente espectral do sinal limpo. Para isso, assume que os coeficientes de Fourier (componentes espectrais) do sinal e do ruído podem ser modelados como sendo variáveis aleatórias gaussianas estatisticamente independentes. Os passos do algoritmo são detalhados a seguir. Nestas equações usa-se a notação $x_k(n-1)$ para indicar a k-ésima componente espectral de x na janela $n-1$.

Calculamos o SNR *a posteriori* e o SNR *a priori* da janela n, $\gamma_k(n)$ e $\xi_k(n)$, respectivamente, como:

$$\begin{aligned} \gamma_k(n) &= \frac{|\mathcal{X}_k(n)|^2}{|\hat{\mathcal{W}}_k(n)|^2} \\ \xi_k(n) &= \alpha G^2(\gamma_k(n-1))\gamma_k(n-1) + (1-\alpha)P[\gamma_k(n)-1] \end{aligned} \tag{3.12}$$

onde, $G(\gamma_k(n)) = \sqrt{1-1/\gamma_k(n)}P[\gamma_k(n)-1]$ é uma outra forma de expressar a subtração espectral, $|\hat{\mathcal{W}}_k(n)|^2$ é calculado como na subtração espectral e $P[\cdot]$ é usado para garantir que $\xi_k(n)$ seja sempre positiva, sendo definido como:

$$P[x] = \begin{cases} x, & \text{se } x \geq 0 \\ 0, & \text{caso contrário} \end{cases} \tag{3.13}$$

Definimos q_k como a probabilidade de ausência do sinal de voz na componente k. Para simplificar a notação, definimos também $\mu_k = (1-q_k)/q_k$, $\eta_k = \xi_k/(1-q_k)$ e

$$\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{3.14}$$

Com estas variáveis, a razão de verossimilhança generalizada é expressa por:

$$\Lambda(\eta_k, \gamma_k, q_k) = \mu_k \frac{e^{\nu_k}}{1+\eta_k} \tag{3.15}$$

Finalmente, a função do filtro é:

$$G^{D}_{MMSE}(\xi_k, \gamma_k, q_k) = \frac{\Lambda(\eta_k, \gamma_k, q_k)}{1 + \Lambda(\eta_k, \gamma_k, q_k)} G_{MMSE}(\xi_k, \gamma_k) \tag{3.16}$$

onde;

$$G_{MMSE}(\xi_k, \gamma_k) = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k} e^{\frac{-\nu_k}{2}} \left[(1+\nu_k) I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right] \tag{3.17}$$

e $\Gamma(\cdot)$ é a função gamma, $\Gamma(1,5) = \sqrt{\pi}/2$ e $I_0(\cdot)$ e $I_1(\cdot)$ são as funções modificadas de Bessel de ordem zero e de primeira ordem, respectivamente.

Nas simulações realizadas nesta dissertação, foram usados os valores $\alpha = 0,99$ e $q_k = 0,2$, que, segundo observados em (EPHRAIM, 1984) são valores que resultam no melhor desempenho se avaliado subjetivamente.

3.3.2 O MÉTODO DE VIRAG

A subtração espectral apresenta o problema de ruído musical, que é um ruído de estrutura não natural composto por tons a freqüências aleatórias, que por vezes pode chegar a ser menos confortável que o sinal com o ruído original. Para atenuar este problema, surgiram vários algoritmos dos quais o mais famoso é o proposto em (BEROUTI, 1979). Este, combinado com o esquema sugerido em (LIM, 1979), resulta na subtração espectral generalizada que modifica a função $G(w)$ da EQ. 3.10 com a seguinte função (VIRAG, 1999):

$$G(w) = \begin{cases} \left(1 - \alpha \left[\frac{|\hat{\mathcal{V}}(w)|}{|\mathcal{X}(w)|}\right]^{\gamma}\right)^{1/\gamma}, & \left[\frac{|\hat{\mathcal{V}}(w)|}{|\mathcal{X}(w)|}\right]^{\gamma} < \frac{1}{\alpha+\beta}; \\ \left(\beta \left[\frac{|\hat{\mathcal{V}}(w)|}{|\mathcal{X}(w)|}\right]^{\gamma}\right)^{1/\gamma}, & \text{caso contrário;} \end{cases} \tag{3.18}$$

onde α é o fator de sobre–subtração que diminui o ruído musical mas aumenta a distorção do sinal, estando tipicamente entre os valores de 1 e 6; β é o piso espectral que permite a diminuição de ruído musical mas aumenta o ruído de fundo, normalmente no intervalo $0 \leq \beta \ll 1$.

O método introduzido por (VIRAG, 1999) usa o limiar de mascaramento do sinal, $T(k)$, para adaptar de maneira perceptualmente ótima os coeficientes α e β. O valor de γ é fixado em 2. A adaptação é realizada em cada banda de cada janela analisada e é baseada nas seguintes funções:

$$\alpha_k = F_{\alpha}[\alpha_{min}, \alpha_{max}, T(k)] \tag{3.19}$$

$$\beta_k = F_{\beta}[\beta_{min}, \beta_{max}, T(k)] \tag{3.20}$$

onde α_{min}, α_{max}, β_{min} e β_{max} são os valores máximo e mínimo dos coeficientes de sobre–subtração e piso espectral. Virag observou que valores de $\alpha_{min} = 1$, $\alpha_{max} = 6$ e $\beta_{min} = 0$, $\beta_{max} = 0,01$ resultam em bom compromisso entre redução de ruído e distorção de voz. F_α e F_β são funções de interpolação linear tal que $F_\alpha = \alpha_{max}$ se $T(k) = T(k)_{min}$ e $F_\alpha = \alpha_{min}$ se $T(k) = T(k)_{max}$, onde $T(k)_{min}$ e $T(k)_{max}$ são os valores mínimo e máximo de $T(k)$ da janela analisada. Considerações semelhantes são usadas para F_β. Para evitar descontinuidades na função $G(k)$, devidas à adaptação, é realizada uma operação de amaciamento.

O cálculo do limiar de mascaramento, $T(k)$, é realizado em cada janela baseado no modelo de percepção auditiva de (JOHNSTON, 1988). Os seguintes passos são efetuados para o cálculo do limiar de mascaramento:

- Análise de bandas críticas do sinal;
- Consideração dos efeitos de espalhamento das bandas críticas do sinal;
- Cálculo do limiar de mascaramento espalhado;
- Renormalização do espectro de bandas críticas;
- Comparação com o limiar Absoluto Auditivo.

ANÁLISE DE BANDAS CRÍTICAS

O primeiro passo é calcular a energia presente em cada banda crítica. Isto é realizado somando as componentes espectrais do sinal que pertencem a uma mesma banda crítica:

$$B(i) = \sum_{k=b_l(i)}^{b_h(i)} |\mathcal{S}(k)|^2 \tag{3.21}$$

onde $b_l(i)$ e $b_h(i)$ são os limites inferior e superior da banda crítica i. Como não possuímos $|\mathcal{S}(k)|$, usamos uma estimativa obtida de $|\mathcal{X}(k)|$ através de algum método de subtração espectral, como os já indicados anteriormente.

FUNÇÃO DE ESPALHAMENTO

Para calcular os efeitos do espalhamento e mascaramento entre bandas críticas usamos a função dada por (PAINTER, 2000):

$$SF(x) = 15,81 + 7,5(x + 0,474) - 17,5\sqrt{1 + (x + 0,474)^2} \tag{3.22}$$

onde x tem medidas de $Barks$ e SF é expresso em dB.

A função é calculada para $abs(j-i) \leq 25$, onde i é a freqüência Bark do sinal mascarado e j é a freqüência Bark do sinal que está mascarando. Os valores são armazenados na matriz S_{ij}. A convolução de $B(i)$ com a função de espalhamento é implementada como uma multiplicação de matrizes, i.e., $C(i) = S_{ij} * B(i)$.

CÁLCULO DO LIMIAR DE MASCARAMENTO ESPALHADO

Primeiro é calculado o limiar de deslocamento relativo, $O(i)$. Isso é feito levando em consideração os tipos de mascaramento. Neste modelo usa-se dois tipos: o primeiro é um tom mascarando o ruído, que é estimado como $14,5+idB$ abaixo de $C(i)$; o segundo é ruído mascarando um tom, estimado como $5,5\,dB$ abaixo de $C(i)$. Para realizar estes cálculos, precisamos ter conhecimento da tonalidade do sinal. Com o objetivo de simplificar o cálculo usa-se uma aproximação proposta por (SINHA, 1993), baseada na idéia de que a voz nas baixas freqüências é de natureza tonal e nas altas freqüências é de natureza mais semelhante ao ruído. Os valores de deslocamento relativo usados por Virag são apresentados na FIG. 3.4.

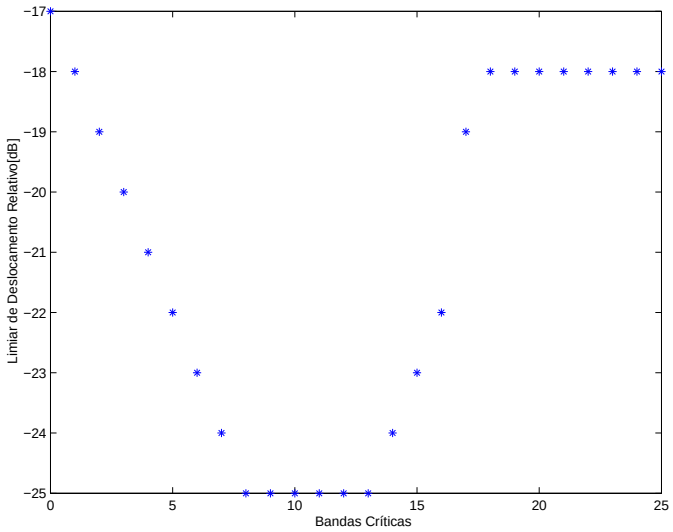

FIG. 3.4: Limiar de Deslocamento Relativo.

O limiar de mascaramento espalhado é $T(i) = 10^{\{\log C(i) - [O(i)/10]\}}$.

RENORMALIZAÇÃO DO ESPECTRO DE BANDAS CRÍTICAS

A convolução da função de espalhamento com $B(i)$ deve ser desfeita, ou seja, $T(i)$ deve ser deconvoluído para obter $T(k)$; entretanto, este processo é instável. Por isso, é usada a normalização que multiplica cada banda de $T(i)$ pelo inverso do ganho de energia, assumindo distribuição uniforme de 1 para cada banda.

COMPARAÇÃO COM O LIMIAR ABSOLUTO AUDITIVO

Desde que o limiar de mascaramento foi calculado sem referência a um valor absoluto, ele deve ser comparado para assegurar que o nível de ruído permitido não esteja abaixo do limiar auditivo absoluto; assim, sempre escolhe-se o maior dos dois. O limiar absoluto está dado pela seguinte função (PAINTER, 2000):

$$T_q(f) = 3,64(f)^{-0,8} - 6,5e^{-0,6(f-3,3)^2} + 10^{-3}(f)^4 \ (dB\,\text{SPL}) \tag{3.23}$$

onde f é a freqüência do sinal espressada em kHz .

3.4 O REALCE DA VOZ USANDO TÉCNICAS MONO-CANAL BASEADAS EM WAVELETS

Nas últimas décadas as transformadas wavelet têm sido muito pesquisadas e usadas em diferentes áreas. As várias aplicações incluem supressão de ruído (*denoising*) em sinais, bem como compressão, detecção e reconhecimento de padrões (BAHOURA, 2001). O processamento clássico em wavelets está ilustrado na FIG. 3.5, onde a transformada wavelet e a transformada wavelet inversa são operações lineares dinâmicas, ou seja, são operações lineares que dependem das amostras presentes e passadas do sinal de entrada. Já o bloco chamado de "processamento" é uma operação que pode ser linear ou não–linear mas que é algébrica, isto é, só depende das amostras presentes do sinal de entrada (BURRUS, 1998).

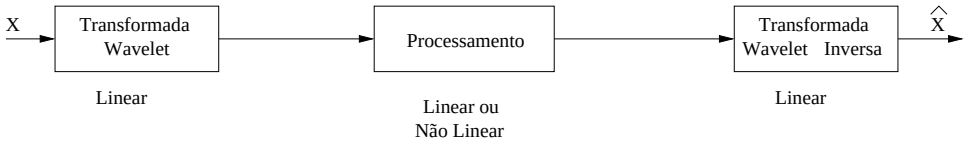

FIG. 3.5: Processamento de Sinais no Domínio da Transformada Wavelet.

A estrutura que separa as partes lineares dinâmicas das não–lineares algébricas do sistema, possibilita a obtenção de resultados muito difíceis, ou quase impossíveis de obter, através de sistemas dinâmicos não–lineares genéricos (BURRUS, 1998).

A capacidade de suprimir ruído presente nos sinais é possível por duas propriedades das transformadas wavelet. A primeira é que apenas alguns poucos coeficientes da transformada são não–nulos se as bases da transformada forem adequadamente selecionadas, assim se consegue uma alta concentração de energia nestes poucos coeficientes. A segunda propriedade é que, se o sinal apresentar distribuição gaussiana, os coeficientes wavelet também apresentarão tal distribuição. Neste sentido, a transformada wavelet de ruído gaussiano branco é ruído gaussiano branco e a energia total está espalhada em todos os coeficientes. Destas duas propriedades, observa-se que os coeficientes da transformada wavelet de um sinal terão amplitude comparativamente superior aos coeficientes da transformada do ruído. É esta diferença de amplitudes que faz possível uma operação de filtragem onde as componentes espectrais do sinal e do ruído podem estar superpostas em tempo e freqüência, o que é quase impossível com métodos baseados na transformada de Fourier (BURRUS, 1998; DONOHO, 1994a). Observa-se também que o sucesso dos algoritmos depende da base da transformada escolhida.

Seguindo as idéias propostas por (DONOHO, 1992, 1994a), o processo de supressão de ruído pode ser resumido como a seguir. Seja um sinal $x(n) = s(n) + w(n)$, onde $s(n)$ é o sinal limpo e $w(n)$ é ruído branco gaussiano de média 0 e variância σ_w^2. Usando a transformada wavelet, $s(n)$ e $x(n)$ podem ser representadas como:

$$s(n) = \sum_{k=-\infty}^{\infty} \alpha_{j_0 k} \varphi_{j_0 k}(n) + \sum_{j=j_0}^{j_1} \sum_{k=-\infty}^{\infty} \beta_{jk} \psi_{jk}(n) \tag{3.24}$$

e,

$$x(n) = \sum_{k=-\infty}^{\infty} Y_{j_0 k} \varphi_{j_0 k}(n) + \sum_{j=j_0}^{j_1} \sum_{k=-\infty}^{\infty} Z_{jk} \psi_{jk}(n) \tag{3.25}$$

onde, φ e ψ são as bases da transformada wavelet supostas ortonormais com suporte compacto, α_k's e Y_k's são os coeficientes de aproximação da transformada na banda k e β_{jk} e Z_{jk} são os coeficientes de detalhe da transformada na banda k e nível j.

Usando matrizes podemos escrever $\boldsymbol{s} = [s(n)\ s(n-1)\ s(n-2)\ \cdots\ s(n-N+1)]^T$, tal que:

$$\boldsymbol{S} = \boldsymbol{T}\boldsymbol{s} \tag{3.26}$$

sendo que N é o número de amostras, $\boldsymbol{T}$ é a matriz da transformada wavelet ($\boldsymbol{T}\boldsymbol{T}^{-1} = \boldsymbol{I}$) e $\boldsymbol{S}$ é o vetor que contém os coeficientes de aproximação e os coeficientes de detalhe da transformada wavelet de $\boldsymbol{s}$. Do mesmo modo, podemos escrever que $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{x}$ e $\boldsymbol{W} = \boldsymbol{T}\boldsymbol{w}$, tal que:

$$\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{W} \tag{3.27}$$

O estimador de $\boldsymbol{S}$, $\widehat{\boldsymbol{S}}$, é obtido simplesmente mantendo ou zerando individualmente os coeficientes da transformada, isto é

$$\widehat{\boldsymbol{S}} = \Delta \boldsymbol{X} \tag{3.28}$$

onde $\Delta = \text{diag}(\delta_1, \cdots, \delta_N)$, sendo que $\delta_i \in \{0, 1\}$.

Podemos, então, definir uma função risco do estimador dada por (DONOHO, 1994a; BURRUS, 1998):

$$\mathcal{R}(\hat{s}, s) = E[\|\hat{s} - s\|_2^2] = E[\|\boldsymbol{T}^{-1}(\widehat{\boldsymbol{S}} - \boldsymbol{S})\|_2^2] = E[\|\widehat{\boldsymbol{S}} - \boldsymbol{S}\|_2^2], \tag{3.29}$$

Quando $\delta_i = 0$, o erro quadrático entre o sinal original e o sinal estimado é $\boldsymbol{S}_i^2$; quando $\delta_i = 1$, o erro presente entre os dois sinais é $\sigma_{w_i}^2$. Isto pode-se obter somente quando $\delta_i = 1$ para $|S_i| > t$ e $\delta_i = 0$ caso contrário, onde $t = \sigma_w$ é o limiar ideal que depende do ruído, desconhecido na prática.

Se tivéssemos um algoritmo para obter o limiar ideal, o risco mínimo (ou ideal) atingido pelo processo de supressão de ruído é igual a:

$$\mathcal{R}_{id}(\hat{s}, s) = \sum_{i=1}^{N} \min(\boldsymbol{S}_i^2, \sigma_{w_i}^2) \tag{3.30}$$

É interessante notar que, quando permitimos um valor arbitrário para δ_i dentro do conjunto dos números reais, o risco ideal é melhorado por um fator de 2 (DONOHO, 1994a).

(DONOHO, 1992) propôs um algoritmo genérico para realizar a suppresão de ruído; ele é resumido em três passos:

(1) Aplicar um algoritmo wavelets rápido aos dados de entrada;

(2) Aplicar uma função de *thresholding* (limiar) aos coeficientes de detalhe da transformada, usando um limiar especialmente estimado;

(3) Inverter o algoritmo de wavelets para compor o sinal no domínio do tempo.

3.4.1 FUNÇÕES DE LIMIAR

Existem várias funções de limiar usadas na literatura científica (HÄRDLE, 1997). Aqui estão apresentadas duas delas, o *Soft–Thresholding* e o *Hard–Thresholding*.

No *soft–thresholding* os coeficientes de detalhe da transformada do sinal limpo, β_{jk}, são estimados como

$$\hat{\beta}_{jk} = \eta_S(Z_{jk}, t) = \begin{cases} \text{sign}(Z_{jk})(|Z_{jk}| - t), & |Z_{jk}| \leq t \\ 0, & \text{caso contrário.} \end{cases} \tag{3.31}$$

O índice j está relacionado com o nível da transformada e o índice k com a sub-banda.

No *hard–thresholding* os coeficientes de detalhe da transformada do sinal limpo, β_{jk}, são estimados como

$$\hat{\beta}_{jk} = \eta_H(Z_{jk}, t) = \begin{cases} Z_{jk}, & |Z_{jk}| \leq t \\ 0, & \text{caso contrário.} \end{cases} \tag{3.32}$$

Na FIG. 3.6 estão mostradas estas duas funções. O *soft–thresholding* apresenta propriedades matemáticas mais interessantes que o *hard–thresholding*, pois "encolhe" os coeficientes para evitar a descontinuidade abrupta presente no *hard–thresholding*. Contudo, como produto deste encolhimento, apresenta maior polarização na estimativa do sinal (SHU, 2002). Já o *hard–thresholding*, devido às descontinuidades que apresenta, gera estimativas de maior variância e pode causar instabilidade (SHU, 2002). Nesta dissertação foi usado o *soft–thresholding*.

3.4.2 CÁLCULO DO LIMIAR

Dependendo do tipo de ruído que tratamos de suprimir podemos escolher, entre outros, usar o limiar independente do nível ou o limiar dependente do nível.

O primeiro é calculado da observação dos coeficientes no nível de maior resolução da transformada e é usualmente usado para remover ruído branco. Como foi indicado anteriormente, a transformada wavelet do ruído gaussiano branco é ruído gaussiano branco e a sua energia encontra-se espalhada nos coeficientes de todos os níveis da transformada. Por outro lado, se a base da transformada foi apropriadamente escolhida para um sinal dado, sua energia encontra-se dividida, estando em sua maior parte nos coeficientes dos primeiros níveis da transformada. Por isto, no último nível (o de maior resolução) espera-se ter principalmente coeficientes relacionados ao ruído e não ao sinal; daí que o limiar é calculado neste nível e mantido constante para os outros.

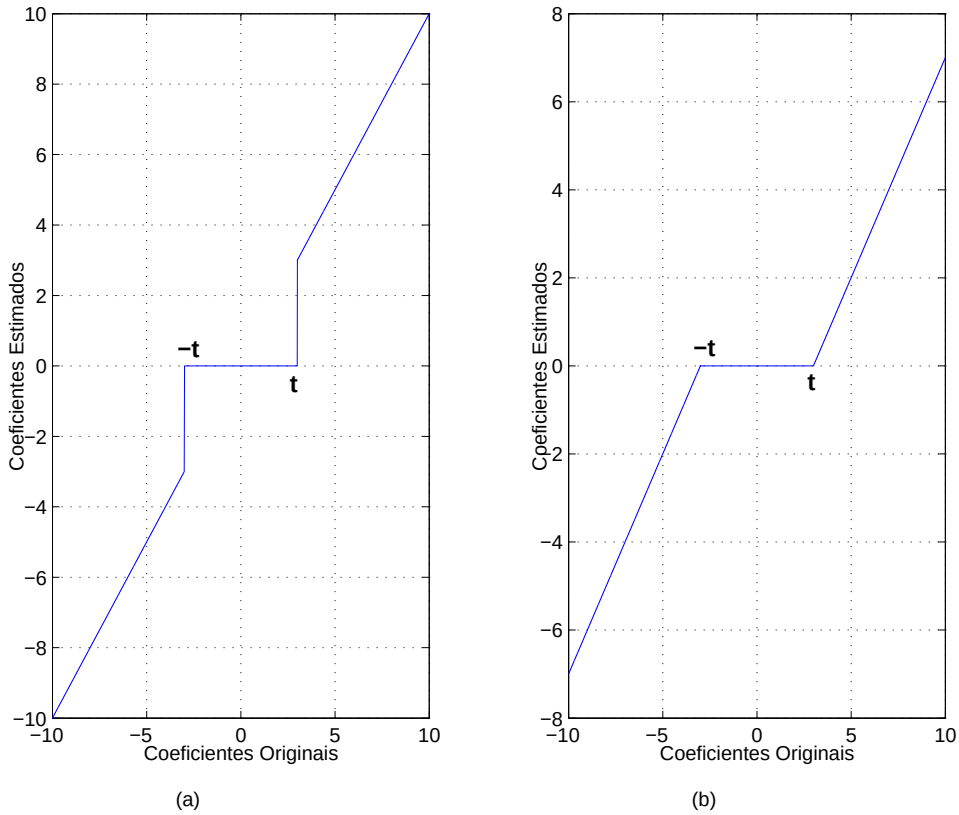

FIG. 3.6: Funções de Limiar; (a) *Hard–Thresholding* (b)*Soft–Thresholding*.

O uso do limiar dependente do nível é comum quando temos presença de ruído colorido. Neste tipo de ruído a energia no domínio da transformada é espalhada de maneira não uniforme, tendo que calcular o limiar apropriado a cada nível. Na FIG. 3.7 observa-se a diferença entre a transformada wavelet de 5 níveis para dois tipos de ruído: ruído branco gerado artificialmente e ruído colorido (cabine de avião da base de dados Noisex-92–no Capítulo 5 encontra-se uma descrição detalhada desta base), as linhas verticais indicam os limites entre os níveis da transformada. Observa-se nesta figura a diferença entre o espalhamento da energia do ruído em cada caso.

Neste trabalho foi, em todos os casos, usado o limiar dependente do nível. (DONOHO, 1994b) apresenta, entre outros, os seguintes métodos para calcular uma estimativa, $\hat{t}$, do limiar:

VisuShrink

Este método também é conhecido como "regra universal" e pode ser usado junto com

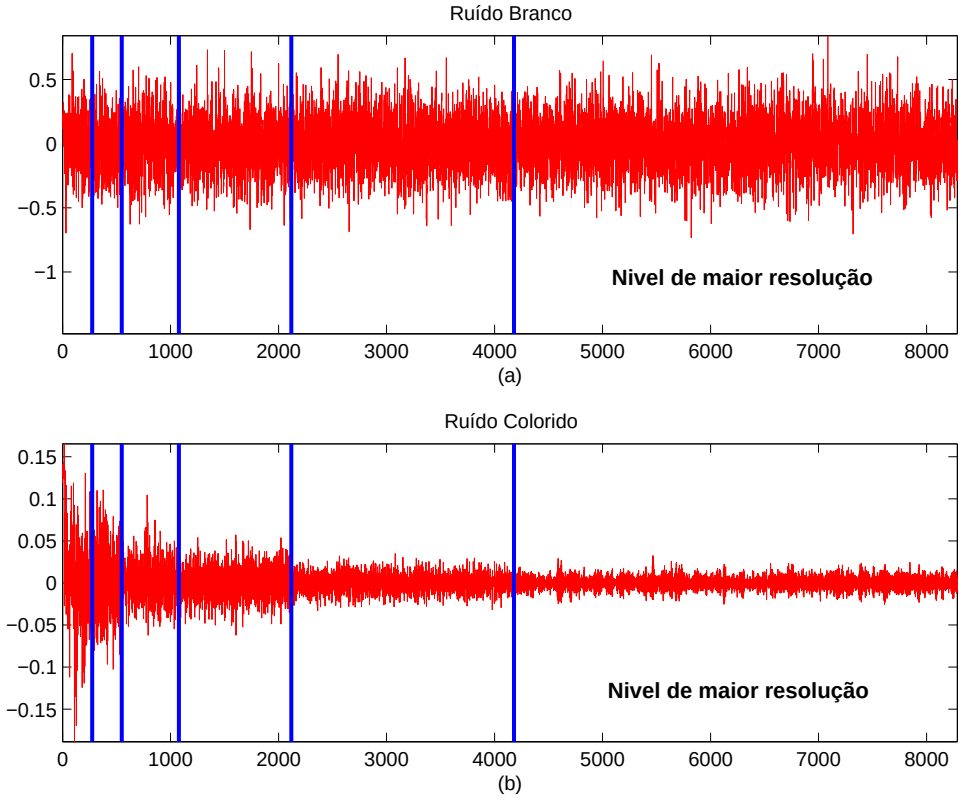

FIG. 3.7: Transformada wavelet de 5 níveis para (a) Ruído Branco (b) Ruído Colorido.

qualquer uma das funções de limiar antes vistas. Aplicado com o *soft–thresholding* e supondo ruído gaussiano branco, esté método apresenta reconstruções "livres de ruído" com o custo de eliminar também componentes do sinal original (DONOHO, 1994b).

A estimativa do limiar é obtida da minimização da função custo dada pela EQ. 3.29 e a imposição da restrição de, com alta probabilidade, o sinal estimado ter pelo menos a mesma suavidade do sinal original (DONOHO, 1995). A motivação para esta restrição foi a necessidade de uma melhor relação de compromisso entre variância e polarização (*bias*) da estimativa. Em subtração espectral, só minimizamos uma função custo –normalmente o erro médio quadrático– que resulta em um compromisso entre a polarização e a variância na mesma ordem de amplitude. O resultado deste processo de minimização é ótimo desde o ponto de vista do erro médio quadrático mas gera o aparecimento de ruído residual como o ruído musical, ondulações (*ripples*), manchas (*blips*) e oscilações (DONOHO, 1995). O VisuShrink, incorporando na solução esta restrição, livra-se dos ruídos espúrios.

(DONOHO, 1992) apresenta o uma estimativa do limiar como

$$\hat{t} = \hat{\sigma}\sqrt{2\log_{10} N} \tag{3.33}$$

onde N é o número de coeficientes, e $\hat{\sigma}$ é uma estimativa grosseira ($\hat{\sigma} = m/0,6745$) do nível de ruído presente e m é chamado de desvio da mediana do valor absoluto (*median absolute deviation*), calculado como a mediana do valor absoluto dos coeficientes da transformada. No caso do limiar ser independente do nível, m é calculado a partir dos coeficientes do nível de maior resolução.

SureShrink

Seja o k-ésimo coeficiente de detalhe da transformada wavelet de um sinal com ruído no nivel j definido por:

$$Z_{jk} = \beta_{jk} + \sigma_j \xi_{jk}, \quad k = 1, \cdots, N \tag{3.34}$$

onde $\sigma_j > 0$ são parâmetros desconhecidos e ξ_{jk} são variáveis aleatórias gaussianas independentes de média 0 e variância 1.

Define-se a função risco do estimador de β_{jk} como o erro médio quadrático,

$$\mathcal{R}_j(\hat{s}, s) = \sum_{k=1}^{N} E[(\hat{\beta}_{jk} - \beta_{jk})^2] \tag{3.35}$$

onde $\hat{\beta}_{jk} = Z_{jk} + H_t(Z_{jk})$ é o estimador e $H_t[\cdot]$ é uma função real diferenciável para qualquer valor de t_j.

O parâmetro t_j deve ser escolhido estatísticamente. Em outras palavras, a EQ. 3.35 define uma família de estimadores indexados por t_j e o problema é escolher o melhor deles, sendo o t_j ideal aquele que minimiza a função $\mathcal{R}_j(\hat{s}, s)$.

Na prática não conhecemos β_{jk}. Portanto, só poderemos escolher um $\hat{t}_j$ próximo de t_j que minimiza um estimador não polarizado, $\widehat{\mathcal{R}}_j(\hat{s}, s)$, do risco $\mathcal{R}_j(\hat{s}, s)$.

Para construir $\widehat{\mathcal{R}}_j(\hat{s}, s)$ deve-se notar que (HÄRDLE, 1997):

$$E[(\hat{\beta}_{jk} - \beta_{jk})^2] = \sigma_j^2 + 2\sigma_j E[\xi_{jk} H_t(Z_{jk})] + E[H_t^2(Z_{jk})] \tag{3.36}$$

por integração parcial:

$$\begin{aligned} E[\xi_{jk} H_t(\beta_{jk} + \sigma_j \xi_{jk})] &= \frac{1}{\sqrt{2\pi}} \int \xi_{jk} H_t(\beta_{jk} + \sigma_j \xi_{jk}) e^{-\frac{\xi_{jk}^2}{2}} d\xi_{jk} \\ &= \frac{1}{\sqrt{2\pi}} \int H_t(\eta_{jk}) \frac{\eta_{jk} - \beta_{jk}}{\sigma_j} e^{-\frac{(\eta_{jk} - \beta_{jk})^2}{2\sigma_j^2}} d\eta_{jk} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(\eta_{jk}-\beta_{jk})^2}{2\sigma_j^2}} \frac{dH_t(\eta_{jk})}{d\eta_{jk}} d\eta_{jk} \\
&= \sigma_j E\left[\left.\frac{dH_t(x)}{dx}\right|_{x=Z_{jk}}\right] \qquad (3.37)
\end{aligned}
$$

Desta última equação e da EQ. 3.36, segue-se:

$$E[(\hat{\beta}_{jk} - \beta_{jk})^2] = E[R_j(\sigma_j, Z_{jk}, t_j)] \qquad (3.38)$$

onde,

$$R_j(\sigma, x, t) = \sigma^2 + 2\sigma^2 \frac{\partial}{\partial x} H_t(x) + H_t^2(x). \qquad (3.39)$$

A EQ. 3.35 pode ser, então, expressa como $\mathcal{R}_j(\hat{s}, s) = E[\widehat{\mathcal{R}}_j(\hat{s}, s)]$, onde,

$$\widehat{\mathcal{R}}_j(\hat{s}, s) = \sum_{k=1}^{N} R_j(\sigma_j, Z_{jk}, t_j) \qquad (3.40)$$

é um estimador ou preditor de risco, e é chamado de *Stein's Unbiased Risk Estimator* (SURE).

O princípio de Stein é minimizar $\widehat{\mathcal{R}}_j(\hat{s}, s)$ com respeito a t_j e usar aquele minimizador como um estimador do limiar. Portanto, o estimador do limiar é dado por:

$$\hat{t}_j = \min_{t \geq 0} \sum_{k=1}^{N} R_j(\sigma_j, Z_{jk}, t) \qquad (3.41)$$

Este estimador depende da função de limiar, $H_t(k)$, escolhida. Para *soft–thresholding*, temos $H_t(x) = -xI\{|x| < t\} - tI\{|x| \geq t\}\text{sign}(x)$, onde $I\{\cdot\}$ é chamada de função indicadora e definida como:

$$I\{x\} = \begin{cases} 1, & x \text{ verdadeiro;} \\ 0, & x \text{ falso.} \end{cases} \qquad (3.42)$$

Reescrevendo a EQ. 3.39:

$$
\begin{aligned}
R_j(\sigma, x, t) &= (x^2 - \sigma^2) I\{|x| < t\} + (\sigma^2 + t^2) I\{|x| \geq t\} \\
&= [x^2 - \sigma^2] + (2\sigma^2 - x^2 + t^2) I\{|x| \geq t\}. \qquad (3.43)
\end{aligned}
$$

O termo entre colchetes não depende de t resultando que a EQ. 3.41, para *soft–thresholding*, pode ser expressa como:

$$\hat{t}_j = \min_{t \geq 0} \sum_{k=1}^{N} (2\sigma_j^2 + t^2 - Z_{jk}^2) I\{|Z_{jk}| \geq t\}. \qquad (3.44)$$

onde σ_j é estimado por $\hat{\sigma}_j = m_j/0,6745$, sendo que m é o desvio da mediana do valor absoluto calculado no nível j da transformada wavelet.

3.5 RESUMO

Neste capítulo, após uma breve introdução e classificação dos algoritmos de realce de voz, foram abordados com mais detalhe os algoritmos com processamento mono-canal. Dentre estes, ressaltou-se os relacionados à subtração espectral e os baseados em wavelets. No Capítulo seguinte, será proposto um novo método de cálculo de limiar (empregando redes neurais) a ser usado na técnica de realce de voz baseado em wavelets. Um estudo comparativo dos algoritmos será realizado, tendo como objetivo final a avaliação dos algoritmos de realce de voz na tarefa de verificação automática de locutor. Esta última fará parte do Capítulo 5 desta dissertação.

4 USO DE REDES NEURAIS EM REALCE DE VOZ BASEADO EM WAVELETS

4.1 INTRODUÇÃO

Neste capítulo é proposto um outro método para o cálculo do limiar em aplicações de supressão de ruído em sinais de voz. Este método estima um limiar que, experimentalmente e para o caso de sinais de voz, oferece uma melhor aproximação ao limiar ideal (aquele que minimiza a função custo expressa pela EQ. 3.29) que os métodos de cálculo VisuShrink e SureShrink.

4.2 O EMPREGO DE *DENOISING* PARA REALCE DE SINAIS DE VOZ

Como foi indicado no capítulo anterior, não é qualquer base que pode ser usada para representar um sinal; esta base deve atender duas condições. A primeira é que deve ser ortonormal, para assim poder realizar a reconstrução do sinal estimado. A segunda condição é o suporte compacto das funções base, que se refere à propriedade da decomposição wavelet do sinal possuir poucos coeficientes não nulos. Desta propriedade de suporte compacto depende o desempenho dos algoritmos de *denoising*. Uma transformada wavelet adequada deve concentrar mais de 90 % da energia do sinal nos primeiros $N/2$ coeficientes (AGBINYA, 1996). Para sinais de voz a transformada Daubechies 10 ($db10$) satisfaz esta condição (AGBINYA, 1996).

Para entender melhor o *denoising*, pode-se observar a FIG. 4.1. Nesta figura encontram-se a transformada wavelet do sinal *"heavy sine"* (DONOHO, 1995), a transformada wavelet do ruído gaussiano branco somado e o limiar calculado com o VisuShrink para cada nível da transformada wavelet $db10$ de 5 níveis. As linhas verticais indicam os limites entre níveis.

Como se esperava, a energia do sinal está concentrada em pouquíssimos coeficientes, enquanto que a energia do ruído encontra-se espalhada em todos os coeficientes, atendendo a propriedade de que a transformada wavelet do ruído gaussiano branco é ruído gaussiano branco. Observa-se também uma grande diferença entre as amplitudes dos coeficientes do sinal e do ruído, isto faz possível o processo de supressão do ruído. Na FIG. 4.2 pode-

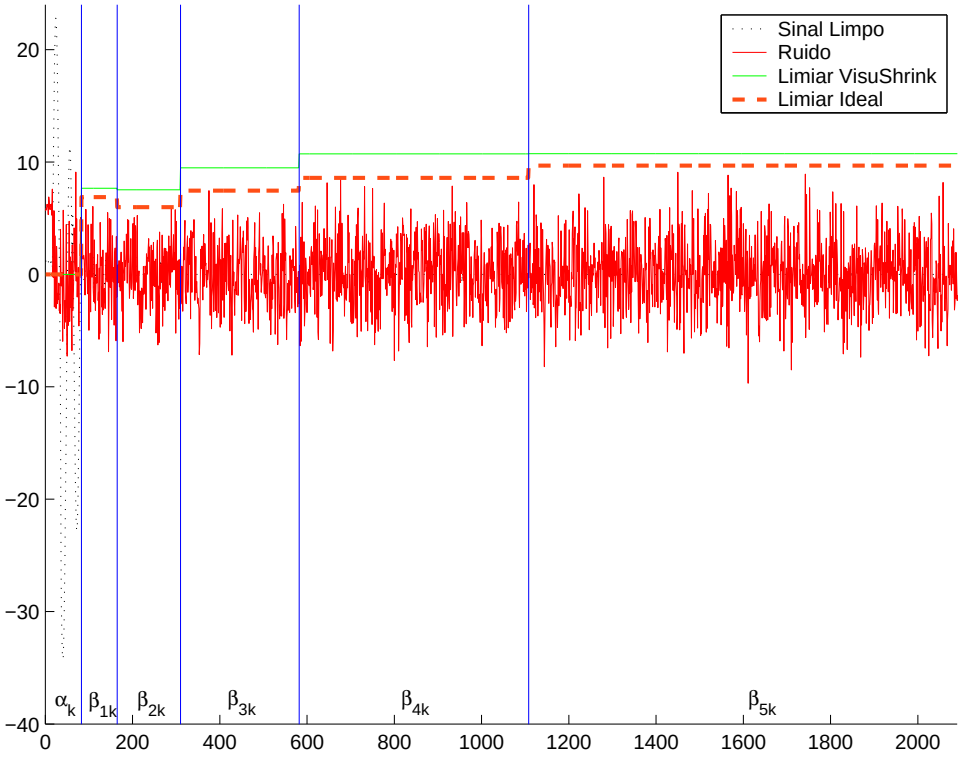

FIG. 4.1: Coeficientes wavelet do sinal *"Heavy Sine"* e limiar calculado com o método VisuShrink em presença de ruído branco.

se observar o sinal *"heavy sine"* limpo, o sinal corrompido com ruído gaussiano branco (SNR=0 dB) e o sinal estimado com o método VisuShrink.

Para sinais de voz, o processo de supressão de ruído é o mesmo, exceto que, antes de obter a transformada wavelet da voz, esta deve ser janelada. Nesta dissertação sempre foi usada a janela de *Hamming*. Como exemplo, na FIG. 4.3, pode-se ver a transformada wavelet de 5 níveis de um sinal de voz sonoro e de ruído gaussiano branco; está incluido também o limiar calculado com o método VisuShrink e o limiar ideal. Nesta figura observa-se que, devido ao janelamento, a transformada do ruído deixa de ser ruído gaussiano branco e a energia encontra-se espalhada não regularmente em todos os níveis da transformada. Isto sugere o uso do limiar dependente do nível.

Na FIG. 4.4 está apresentado o resultado do *denoising* com o método VisuShrink para o mesmo sinal de voz antes usado. Nesta figura, observa-se que o VisuShrink realiza uma estimativa "super-suavizada" do sinal; isto podería ocasionar a perda de características da voz úteis para algum processo posterior.

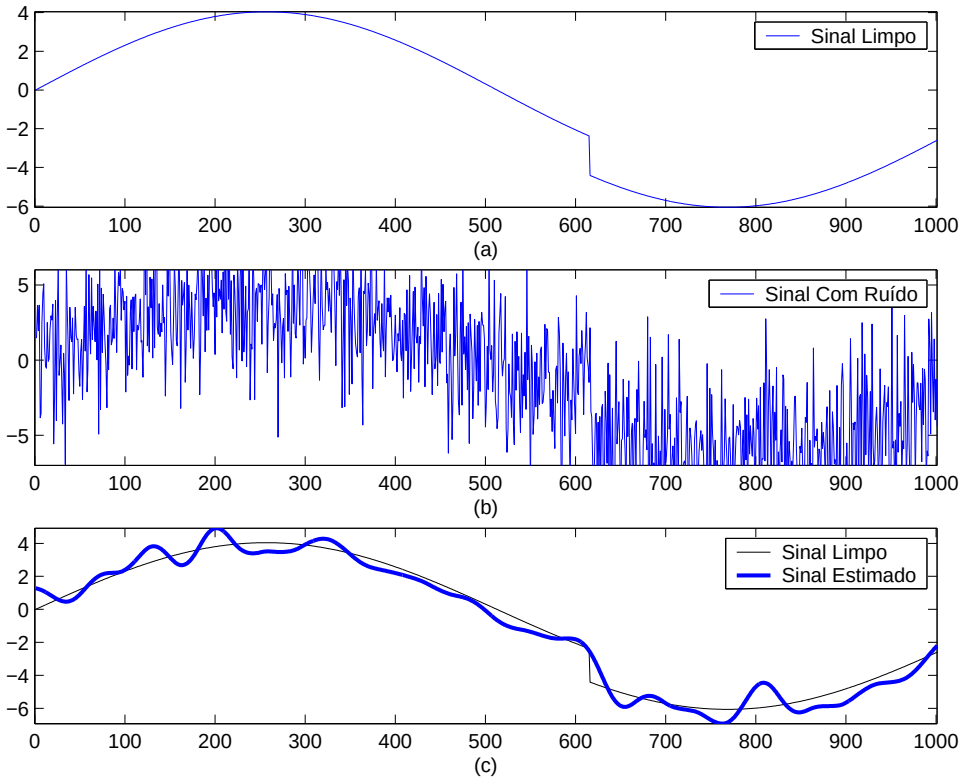

FIG. 4.2: Sinal *"Heavy Sine"* corrompida com ruído branco, $SNR = 0\ dB$; (a) Sinal limpo (b) Sinal com ruído (c) Sinal estimado com o método VisuShrink.

Se usarmos a estimativa calculada com o SureShrink neste mesmo exemplo, poderíamos observar que no terceiro nível a estimativa é mais aproximada ao limiar ideal mas, no último nível, a estimativa é muito menor que o limiar ideal; isto faz com que, em geral, o SureShrink ocasione menor distorção no sinal do que o VisuShrink, com a desvantagem de aumentar o ruído de fundo.

4.3 O USO DE REDES NEURAIS PARA O CÁLCULO DO LIMIAR

Na seção anterior, através de exemplos, foi mostrado que as estimativas de limiar oferecidas pelos métodos VisuShrink e SureShrink nem sempre se aproximam o suficiente do limiar ideal para o caso de sinais de voz. Nesta seção, é introduzido um método de cálculo do limiar fazendo uso de uma rede neural.

A rede neural que foi usada nesta implementação é a *backpropagation* de 1 camada

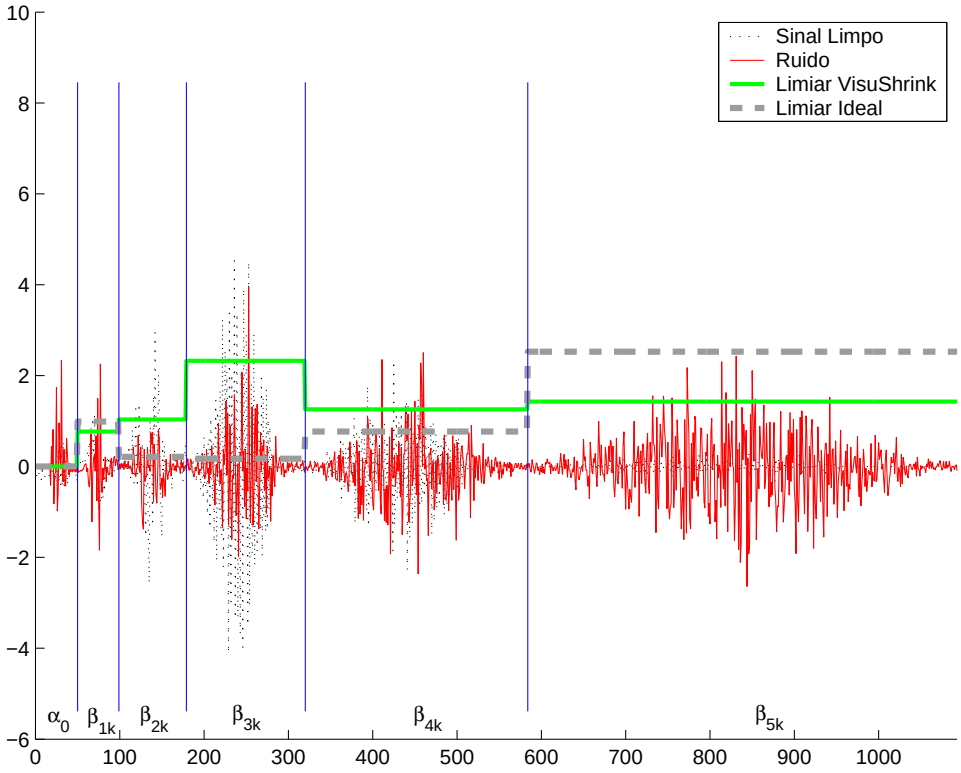

FIG. 4.3: Coeficientes wavelet de um sinal de voz sonoro e limiar calculado con método VisuShrink.

escondida de H neurônios. A saída desta rede neural é o uma estimativa do limiar que, no caso do treinamento da rede, é subtraído do limiar ideal previamente escalado pelo fator G, fornecendo a medida de erro. No caso do teste da rede neural, a saída é a estimativa do limiar que deve ser usado no *denoising*.

Foram escolhidos dois parâmetros de entrada que são calculados dos coeficientes da transformada wavelet do sinal: o primeiro é a mediana do valor absoluto, usado nos outros métodos de cálculo do limiar, e o segundo é a variância dos coeficientes. Sendo que o limiar calculado é dependente do nível, uma rede neural para cada nível é necessaria.

A configuração do sistema proposto é apresentada na FIG. 4.5. Nesta figura, m_j é a mediana do valor absoluto e v_j^2 é a variância da transformada wavelet no nível j, b representa o *bias* igual a 1 em todos os casos. Todos os neurônios usam a função de ativação log-sigmoidal.

Uma vez que o valor máximo do limiar ideal pode ser maior que 1 e o valor máximo

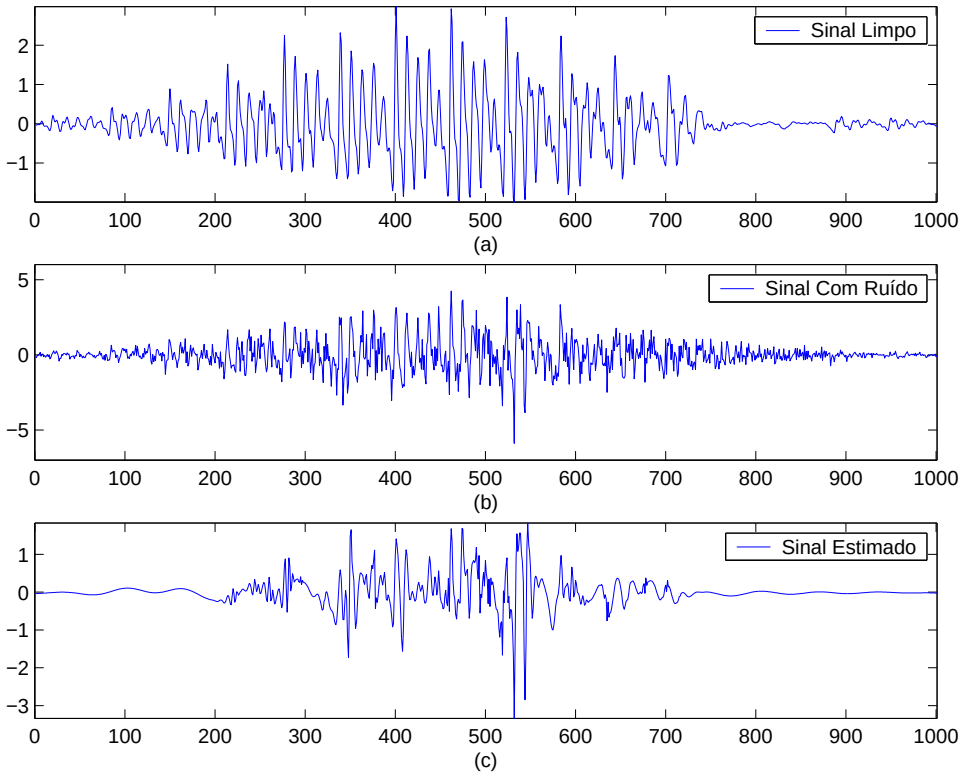

FIG. 4.4: Sinal de voz sonoro corrompido com ruído branco, $SNR = 0\,dB$; (a) Sinal limpo (b) Sinal com ruído (c) Sinal estimado com o método VisuShrink.

da função log-sigmoidal é 1, no treinamento da rede o limiar ideal é multiplicado por um ganho constante, G. No teste da rede o limiar obtido na saída é multiplicado pelo inverso de G.

Para o treinamento foi usado um sinal de 7 minutos de voz. Este sinal é formado por sons sonoros, sons surdos e silêncios. Ao sinal foi somado ruído gaussiano branco artificialmente gerado, de tal forma de obter uma SNR$=-5\,dB$. O sinal foi janelado com a janela de *Hamming* e decomposta com uma transformada wavelet $db10$ de 5 níveis; portanto, foram usadas 5 redes neurais, uma por cada nível. A entrada da rede, como já mencionado, é composta da mediana do valor absoluto e da variância dos coeficientes wavelet de cada nível.

O limiar ideal t_j para o nível j é aquele que minimiza a função risco, expressa pela EQ. 3.29 e reescrita a seguir.

$$\mathcal{R}_j(\hat{s}, s) = E[\|\hat{\beta}_{jk} - \beta_{jk}\|_2^2] \tag{4.1}$$

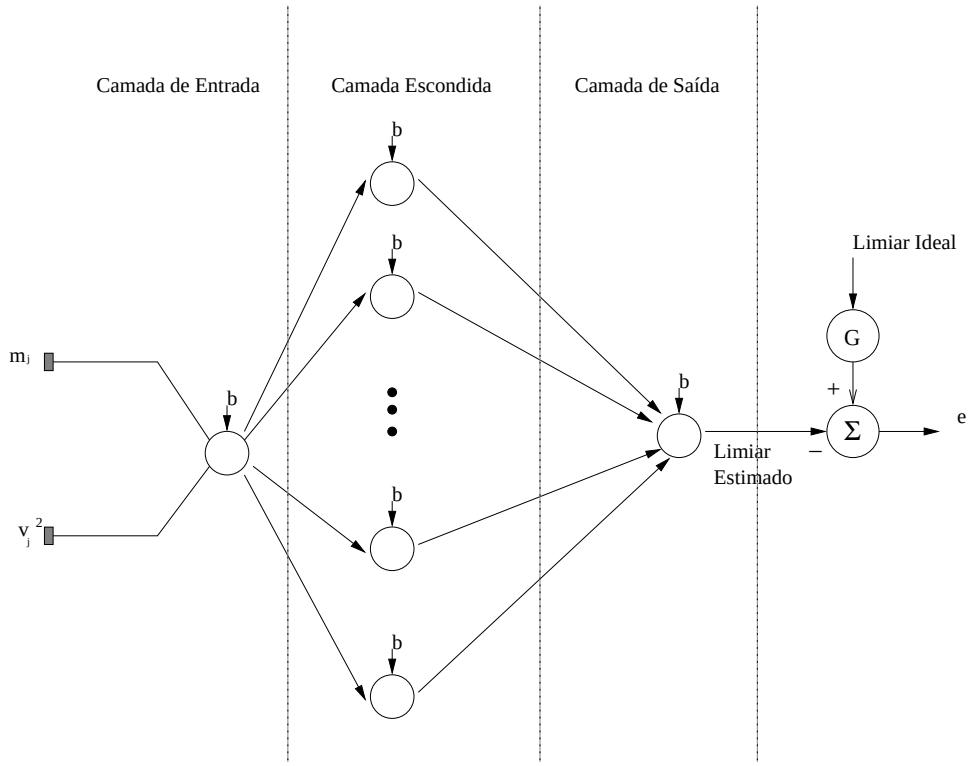

FIG. 4.5: Configuração da proposta usando uma Rede Neural para obter uma estimativa do limiar.

Este limiar só pode ser obtido quando temos o sinal limpo e portanto não é usado em aplicações práticas. Mas, para efeitos de treinamento possui-se o sinal limpo e o limiar ideal é calculado como a amplitude do l-ésimo coeficiente de detalhe da transformada wavelet do sinal com ruído, $|Z_{jl}|$. Isto é, $t_j = |Z_{jl}|$, onde

$$l = \min_{0 \leq l \leq N} \sum_{i=0}^{N} \left[\eta_S(Z_{ji}, |Z_{jl}|) - \beta_{ji}\right]^2 \tag{4.2}$$

$\eta_S(\cdot)$ é definido na EQ. 3.31.

O valor máximo do limiar ideal, obtido experimentalmente, é $6,50$. Portanto, o ganho G da rede deve ser menor ou igual ao inverso deste valor; nesta implementação foi escolhido $G = 1/20$.

O treinamento da rede neural foi do tipo *batch* supervisionado, com taxa de aprendizagem $\alpha = 0,99$. A rede foi treinada com 2, 8, 32 e 64 neurônios na camada escondida; em todos os casos, minimizou-se o erro médio quadrático em aproximadamente 40 iterações.

Como exemplo, na FIG. 4.6, está apresentada a transformada wavelet do mesmo sinal

de voz da FIG. 4.3. Observa-se que a rede neural aproxima bastante bem o limiar estimado ao limiar ideal em todos os níveis da transformada, diferentemente do limiar calculado com o VisuShrink–vide FIG. 4.3. Na FIG. 4.7 é apresentado o resultado da estimação do sinal de voz com o método proposto onde se observa um ganho considerável em relação ao método VisuShrink da FIG. 4.4. Estes resultados foram obtidos com a rede neural de 2 neurônios na camada escondida.

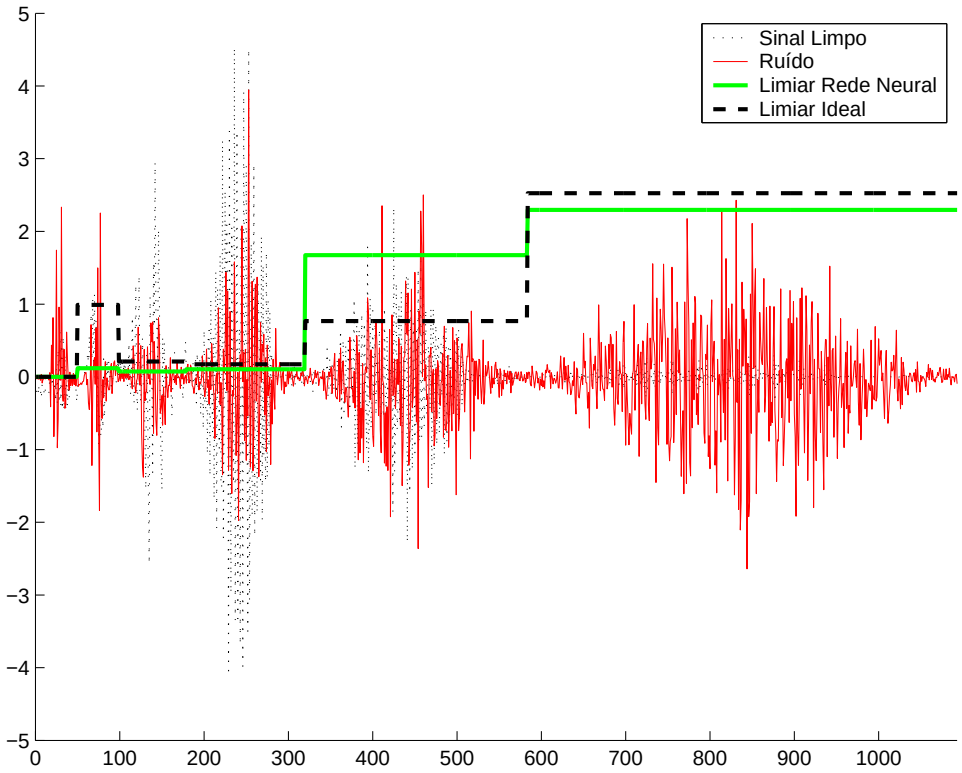

FIG. 4.6: Coeficientes wavelet de um sinal de voz sonoro e valores de limiar calculados com a rede neural e com o limiar ideal em presença de ruído branco.

Para o caso de ruído colorido, estão apresentados na FIG. 4.8 os valores dos limiares calculados com o método VisuShrink, com o método proposto e o limiar ideal para o mesmo sinal de voz. Da mesma forma que para o sinal corrompido com ruído branco, o método proposto estima o limiar com menor erro do que o limiar estimado com o método VisuShrink, exceto no caso do nível de maior resolução.

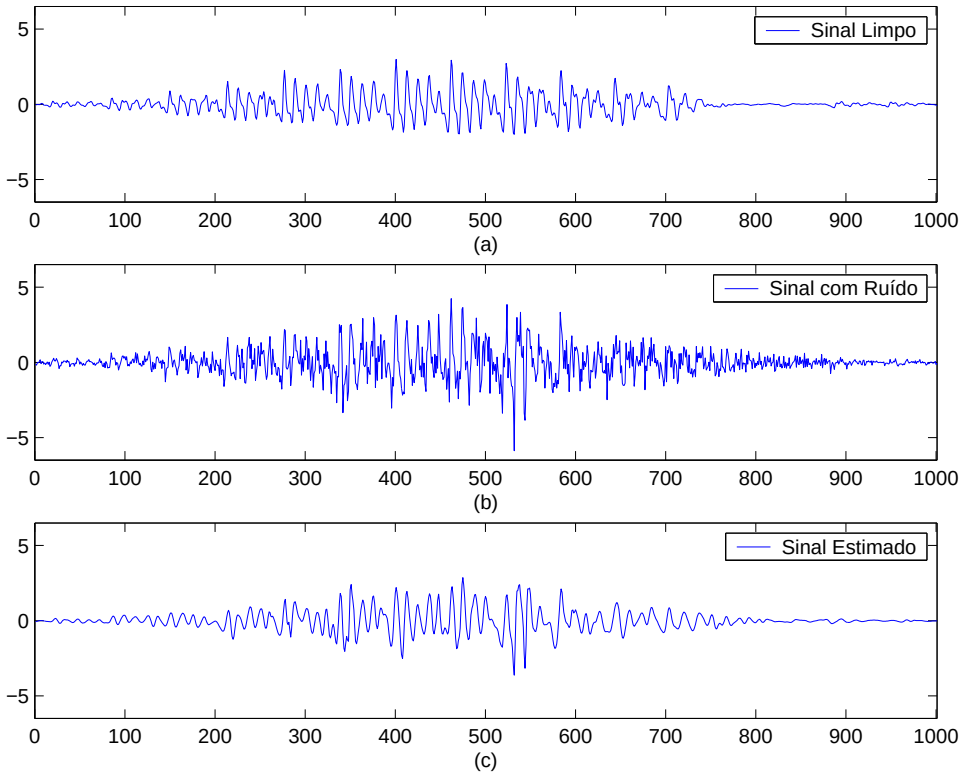

FIG. 4.7: Sinal de voz sonoro corrompido com ruído branco, $SNR = 0\ dB$; (a) Sinal limpo (b) Sinal com ruído (c) Sinal estimado com o método proposto.

4.4 ANÁLISE DE DESEMPENHO

Para realizar uma análise de desempenho do algoritmo proposto, com relação aos algoritmos apresentados no capítulo anterior, foi usada uma medida de desempenho objetiva. Esta é baseada em medidas objetivas de qualidade ou intelibilidade que tentam predizer qual é a preferência dos ouvintes para os sistemas de compressão de voz ou supressão de ruído, entre outros. O maior inconveniente destas medidas é que não sempre estão bem correlatadas com a percepção auditiva.

Sendo que o interesse principal desta dissertação é a avaliação dos algoritmos de realce de voz em verificação de locutor, aqui só é usada uma medida objetiva básica, o Ganho de Relação Sinal Ruído. Esta medida leva em consideração a distorção da voz e o ruído

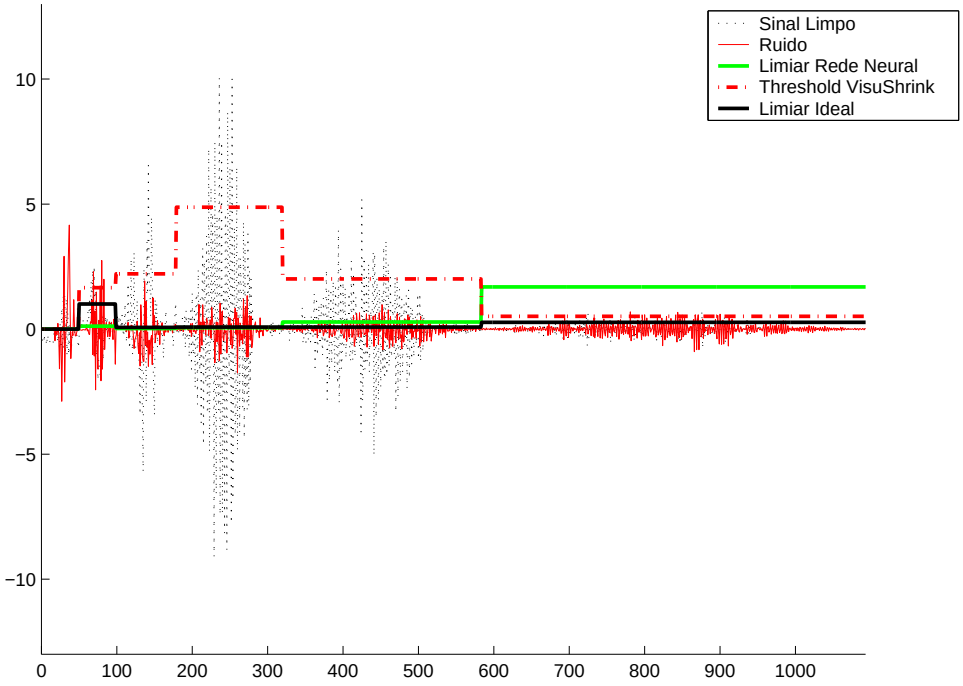

FIG. 4.8: Coeficientes wavelet de um sinal de voz sonoro e limiares calculados com o método VisuShrink, com a rede neural e o limiar ideal em presença de ruído cabine de avião.

residual (VIRAG, 1999) e está definida como (VIRAG, 1999):

$$\mathrm{G}_{SNR} = \frac{1}{L}\sum_{m=0}^{L-1} 10 \cdot \log_{10} \frac{\frac{1}{N}\sum_{n=0}^{N-1} w^2(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}\left[s(n+Nm) - \hat{s}(n+Nm)\right]^2}\ dB \tag{4.3}$$

onde L é o número de janelas no sinal testado, N é o número de amostras em cada janela e $w(n)$, $s(n)$ e $\hat{s}(n)$ são o ruído, o sinal original e o sinal estimado, respetivamente.

Para o teste foram usados o ruído gaussiano branco e 3 tipos de ruído da base de dados *NOISEX*-92: *speech like*, cabine de avião, e ruído de fábrica. Usaram-se para o teste 100 sinais de voz, tomados da base de dados *IME*-2001; no Capítulo 5 encontra-se uma descrição detalhada destas bases. O Ganho relação sinal–ruído apresentado nas seguintes tabelas é o valor médio do ganho obtido para cada sinal.

O primeiro teste realizado foi variando o número de camadas escondidas da rede neural. Os resultados estão listados na TAB. 4.1. Observa-se que a variação no número de neurônios não altera muito o desempenho do algoritmo. Por este motivo, foi usado somente 2 neurônios em todas as simulações seguintes.

Pode-se observar que, embora o sinal de treinamento tenha sido corrompido com ruído numa relação sinal–ruído de $-5\,dB$, a estimativa do limiar, calculada com o método proposto, faz que o G_{SNR} se aproxime do valor obtido com o limiar ideal para todos os casos de relação sinal–ruído, $-5\,dB$, $0\,dB$, $5\,dB$ e $10\,dB$.

No caso de sinais corrompidos com ruído branco o resultado do realce da voz é superior que no caso do sinal estar corrompido com outros tipos de ruído. Isto, possivelmente, deve-se ao fato do sinal de treinamento da rede ter sido corrompido também com ruído branco. Não se utilizou outro tipo de ruído no sinal de treinamento, para melhorar o desempenho no realce de voz, devido a que, como será visto capítulo seguinte, os resultados obtidos na verificação de locutor foram satisfatórios em todos os casos.

TAB. 4.1: $G_{SNR}(dB)$ para sinais corrompidos com diferentes tipos de ruído usando redes neurais.

SNR	Ideal	2 Neurônios	8 Neurônios	32 Neurônios	64 Neurônios
Ruído branco					
-5	11,97	11,16	11,16	11,24	11,14
0	10,84	10,14	10,15	10,20	10,19
5	9,56	8,25	8,24	8,36	8,49
10	8,05	5,63	5,61	5,82	6,11
Ruído *speech like*					
-5	7,09	2,44	2,47	2,32	2,07
0	6,16	2,78	2,80	2,74	2,59
5	5,31	3,02	3,04	3,11	3,06
10	4,48	2,33	2,34	2,50	2,61
Ruído cabine de avião					
-5	5,52	3,52	3,52	3,49	3,36
0	4,86	3,53	3,53	3,55	3,50
5	4,21	3,01	3,02	3,10	3,17
10	3,47	1,49	1,47	1,64	1,86
Ruído de fábrica					
-5	4,70	2,60	2,61	2,53	2,39
0	4,13	2,72	2,73	2,70	2,63
5	3,58	2,41	2,42	2,48	2,50
10	2,99	1,28	1,28	1,43	1,59

Complementando os resultados obtidos em (**MEDINA**, 2003a), listam-se na TAB. 4.2 os valores obtidos do G_{SNR} de todos os algoritmos apresentados neste trabalho: subtração espectral (SS), filtro de Eprahim–Malah (EMF), método de Virag, SureShrink, VisuShrink e o proposto com Redes Neurais (RN).

Uma análise comparativa destes resultados indica que o algoritmo proposto é, na maioria das vezes, superior aos outros algoritmos baseados em wavelets. Estes resultados objetivos são, quase em todos os casos, inferiores aos dos algoritmos baseados em subtração

TAB. 4.2: $G_{SNR}(dB)$ para sinais corrompidos com diferentes tipos de ruído.

SNR	SS	EMF	Virag	SureShrink	VisuShrink	RN
Ruído Branco						
-5	$7,36$	$14,85$	$9,34$	$2,53$	$7,32$	$11,16$
0	$6,83$	$12,40$	$8,39$	$1,74$	$5,37$	$10,14$
5	$6,21$	$9,88$	$7,07$	$0,43$	$2,99$	$8,25$
10	$5,46$	$7,24$	$5,48$	$-1,42$	$0,24$	$5,63$
Ruído *Speech Like*						
-5	$7,31$	$14,00$	$8,20$	$2,83$	$5,18$	$2,44$
0	$6,29$	$10,84$	$7,16$	$2,00$	$3,54$	$2,78$
5	$5,36$	$7,97$	$5,75$	$0,63$	$1,41$	$3,02$
10	$4,51$	$5,43$	$4,08$	$-1,27$	$-1,14$	$2,33$
Ruído de Cabine de Avião						
-5	$6,71$	$13,26$	$7,50$	$2,07$	$3,95$	$3,52$
0	$5,84$	$10,27$	$6,46$	$1,24$	$2,37$	$3,53$
5	$5,00$	$7,50$	$5,07$	$-0,14$	$0,30$	$3,01$
10	$4,15$	$4,88$	$3,42$	$-2,03$	$-2,19$	$1,49$
Ruído de Fábrica						
-5	$7,46$	$13,95$	$7,46$	$2,00$	$3,51$	$2,60$
0	$6,43$	$10,95$	$6,48$	$1,31$	$2,18$	$2,72$
5	$5,42$	$8,10$	$5,14$	$0,11$	$0,34$	$2,41$
10	$4,41$	$5,39$	$3,52$	$-1,62$	$-1,96$	$1,28$

espectral. Como foi falado antes, no caso do processamento com o algoritmo proposto, os resultados indicam um ganho relativamente alto para sinais corrompidos com ruído branco, isto não acontece com ruídos coloridos. Os algoritmos baseados em subtração espectral sempre apresentam um ganho considerável para qualquer tipo de ruído.

Uma análise perceptiva dos algoritmos aponta para conclusões um tanto diferentes. Em geral, os algoritmos baseados em subtração espectral apresentam um desconforto devido ao ruído musical. Este ruído musical não está presente nos algoritmos baseados em wavelets apresentados nesta dissertação.

Os algoritmos VisuShrink e SureShrink apresentam muito ruído de fundo e pouca distorção da voz. O algoritmo proposto apresenta um baixo nível de ruído de fundo, mas distorce o sinal de voz a níveis comparáveis com a distorção introduzida pelo algoritmo de subtração espectral, embora seja de diferente tipo. A escolha de um algoritmo obviamente dependerá da aplicação específica que está sendo desenvolvida.

Cabe ressaltar que o método de Virag, perceptualmente, apresenta um maior realce se comparado com o filtro de Eprahim–Malah, embora o resultado objetivo apresentado sugira o contrário.

4.5 RESUMO

Neste capítulo, foi abordado o uso de redes neurais em realce de voz baseado em wavelets. Ao implementar dois algoritmos clássicos de *denoising* baseados em wavelets, vistos no capítulo anterior, verificou-se que, para sinais de voz, estes algoritmos não apresentam resultados tão bons como os que têm sido obtidos com outros tipos de sinais (DONOHO, 1992, 1995).

Um novo método baseado em wavelets e redes neurais para o cálculo do limiar foi, então, proposto. Foi mostrado que este novo método apresenta resultados objetivos de desempenho na mesma amplitude que os resultados obtidos com o método de Virag para sinais corrompidos com ruído branco. Para outros tipos de ruído, os resultados sempre foram inferiores aos obtidos com o método de Virag, mas superiores aos obtidos com os outros dois métodos baseados em wavelets e estudados nesta dissertação.

Foram brevemente realizados alguns comentários sobre os resultados obtidos da avaliação subjetiva destes algoritmos. Concluiu-se que os métodos baseados em wavelets, a diferença dos métodos derivados da subtração espectral, não apresentam ruído musical. O algoritmo proposto apresenta um baixo nível de ruído de fundo, mas distorce o sinal de voz a níveis comparáveis com a distorção introduzida pelo algoritmo de subtração espectral, embora sendo de diferente tipo.

5 APLICAÇÃO DE REALCE DE VOZ NA VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

5.1 INTRODUÇÃO

Neste capítulo, apresentamos inicialmente as bases de dados usadas na verificação de locutor: a primeira usada como fonte de locuções para a verificação e a segunda usada como fonte de ruído. A seguir, os resultados obtidos das simulações em diferentes condições são apresentados. Finalmente, um novo esquema de verificação de locutor, útil em casos de ruído colorido, é proposto e os seus resultados são apresentados.

5.2 BASES DE DADOS

Para a avaliação do desempenho do sistema de verificação de locutor foram usadas duas bases de dados, descritas a seguir.

IME-2001

Esta base de dados está formada por locuções de 50 locutores, todos eles de sexo masculino; destes locutores, foram aleatoriamente escolhidos 10 para formar o *background*. Todos os locutores falaram 200 frases foneticamente balanceadas de português falado na cidade do Rio de Janeiro; extraídas de (ALCAIM, 1992). Estas locuções foram gravadas a uma taxa de amostragem de $22,05\,kHz$ com 16 bits. Os sinais da base de dados sofreram uma reamostragem para a freqüência de $8\,kHz$, valor mais empregado na comunidade científica para pesquisas em reconhecimento de locutor, e a freqüência de amostragem dos atuais sistemas telefônicos em geral.

Noisex-92

Esta base de dados contém 8 tipos diferentes de ruído: *speech like*, *machine gun*, "STITEL", "lynx" , ruído de cabine de avião, ruído de carro, ruído de fábrica e o ruído sala de operações. Em todas as simulações realizadas, foram usados somente os tipos: *speech like*, cabine de avião e ruído de fábrica, além do ruído gaussiano branco gerado artificialmente.

Como a taxa de amostragem destes sinais é de $16\,kHz$ e 16 bits; foi, então, necessário diminuir a taxa para $8\,kHz$ de modo a podermos efetuar a soma com os sinais de voz. Uma vez que os sinais de voz foram corrompidos por meio da soma destes com sinais

de ruído, nas simulações não foi considerado o efeito Lombard (tendência a aumentar a intensidade de uma vogal em ambientes com ruído) (PICK JR., 1989; WAKAO, 1996).

5.3 DESCRIÇÃO DO SISTEMA DE VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR

O sistema de verificação automática de locutor foi implementado em três etapas, descritas como segue:

Pré–Processamento

Na etapa de pré–processamento foram implementados:

1) O classificador de voz/silêncios - Uma vez que os sinais de voz da base usada contém silêncios no início e no fim da locução, foi possível o uso do classificador baseado nas características temporais do sinal introduzido no Capítulo 2;

2) Janelamento - Foi usada uma janela de 32 ms de duração e superposição de 50%;

3) Pré-ênfase - Foi usada a função mais comumente encontrada nos sistema de processamento de voz, $1 - 0,95z^{-1}$;

4) Realce de voz - Foram empregados os diferentes algoritmos listados neste trabalho: Subtração espectral, Filtro de Ephraim–Malah, Método de Virag, VisuShrink, SureShrink e Redes Neurais aplicada em Wavelets.

Extração de Características

Foram usados 15 coeficientes mel-cepestrais, obtidos de um banco de 23 filtros de banda crítica. Na TAB. 5.1 encontram-se resumidos os parâmetros usados no sistema.

TAB. 5.1: Condições de Análise Usadas na Verificação Automático de Locutor.

Parâmetro	Valor
Pré-ênfase	$1 - 0,95z^{-1}$
Tamanho de Janela	32 ms
Superposição de Janela	16 ms
Tipo de Janela	Hamming
Número de Coeficientes	15 $mel - cepestrais$
Freqüência de Amostragem	8 kHz

Sistema Classificador

O sistema classificador usado foi o Modelo de Misturas Gaussianas (GMM) com o Modelo

de Background Universal (UBM) de 32 misturas. Os sinais de treinamento e teste foram extraídas da base de dados IME-2001 da seguinte forma: de cada uma das 40 locuções existentes (1 para cada locutor não pertencente ao *background*), de duração média 495 segundos, foram extraídos os primeiros 120 segundos para treinamento e os restantes 375 segundos foram divididos em 15 locuções de teste (de 25 segundos cada).

Cada uma das 600 locuções de teste é comparada com cada uma das 40 locuções de treinamento, num total de 24000 testes diferentes, 600 dos quais são verdadeiros.

5.4 SOBRE A MEDIDA USADA PARA AVALIAR O SISTEMA

Nesta seção se introduz a medida do erro usada para avaliar o desempenho do sistema; a continuação uma breve introdução sobre a regra dos trinta, que garante a validade estatística dos resultados obtidos, é apresentada.

5.4.1 MEDIDA DE ERRO

Nos sistemas de verificação de locutor, podem ocorrer dois tipos de erro: o erro E_{FR} de falsa rejeição (FR), onde o sistema rejeita o locutor verdadeiro, e o erro E_{FA} de falsa aceitação (FA), no qual um locutor falso é aceito como verdadeiro.

O desempenho de um sistema é a medida do compromisso entre estes dois erros. O compromisso é controlado por um limiar θ (limiar de decisão). Na avaliação do desempenho do sistema são realizados N_V testes verdadeiros (locuções de teste pertencentes ao locutor de treinamento) e N_F testes falsos (locuções de teste não pertencentes ao locutor de treinamento). As probabilidades dos erros são calculadas como (REYNOLDS, 2000a):

$$P_{FR} = \frac{\#\text{ de testes verdadeiros} < \theta}{N_V} \tag{5.1}$$

$$P_{FA} = \frac{\#\text{ de testes falsos} > \theta}{N_F} \tag{5.2}$$

Na verificação de locutor a P_{FR} só pode ser diminuída se aumentarmos a P_{FA}, e vice versa. Por isto, na comparação deste tipo de sistema, é comum usar a curva DET (Detection Error Tradeoff) (MARTIN, 1997) produzida pelas coordenadas (P_{FA}, P_{FR}), que são calculadas variando o limiar de decisão do sistema de verificação. Esta curva é similar à curva ROC (Receiver Operating Characteristics) mas apresenta informação em diferente escala: a ROC usa a escala linear enquanto que a DET usa a escala do desvio padrão. A DET tem a vantagem de poder distinguir bem entre dois sistemas muito

semelhantes e apresentar comportamento linear quando a distribuição das probabilidades de erro é normal.

Na FIG. 5.1 encontra-se um exemplo de uma curva DET obtida dos resultados das simulações em três diferentes condições, todas com um sinal de teste corrompida com ruído branco tal que SNR= $-5\,dB$: sem ruído no sinal de treinamento, com ruído branco tal que SNR=$5\,dB$ no sinal de treinamento e com ruído branco tal que SNR=$-5\,dB$ no sinal de treinamento.

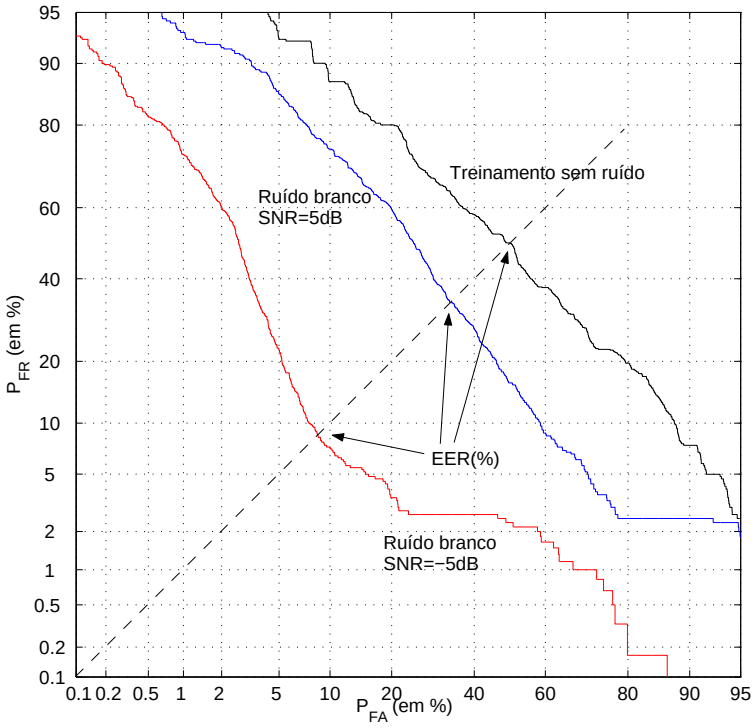

FIG. 5.1: Curva DET para três diferentes ambientes de gravação do sinal de treinamento (sinal de teste com ruído branco tal que SNR=$-5\,dB$).

Como se vê, a DET nos provê de uma representação bi-dimensional do sistema; mas, às vezes, é preciso uma representação uni-dimensional. Uma representação uni-dimensional muito usada é o *Equal Error Rate* (ERR), que é o ponto da DET onde as probabilidades de falsa aceitação e falsa rejeição são iguais.

5.4.2 REGRA DOS TRINTA

Em 1985, George Doddington recomendou que os dispositivos de reconhecimento de voz devem, como regra, ser testados até que pelo menos 30 erros aconteçam. Isto permite uma confiança de 90% de que o valor real do erro, quando pequeno, esteja dentro de ±30% do valor observado (REYNOLDS, 2000a)(NBTC, 2000, On the "30 error" Criterion).

Desenvolvimentos matemáticos posteriores mostraram que o valor real do erro, p, usualmente está dentro do intervalo (NBTC, 2000, On the "30 error" Criterion):

$$\hat{p} - k\sigma_{\hat{p}} \leq p \leq \hat{p} + k\sigma_{\hat{p}} \tag{5.3}$$

onde, $\hat{p}$ é o erro estimado, $\hat{p} = \frac{e}{N}$, sendo que e é o número de erros e N é o número de testes. $\sigma_{\hat{p}}$ é o desvio padrão do estimador e $k = 1$ para 68% de confiança ou $k = 2$ para 95% de confiança.

A confiança diz-nos a respeito do número de vezes que o erro cai dentro do intervalo estimado (intervalo de confiança). Para usar uma confiança diferente da indicada, k deve ser calculado como o número de desvios padrões que abrange a área desejada na função de densidade de probabilidade gaussiana. Vejamos um exemplo: a área que contém 68% do total da função densidade de probablidade gaussiana é a compreendida entre um desvio padrão a mais e a menos da média desta função, dai que $k = 1$ – vide FIG. 5.2.

O desvio padrão do estimador, $\sigma_{\hat{p}}$ é calculado como sendo (NBTC, 2000, On the "30 error" Criterion):

$$\sigma_{\hat{p}} = \frac{1}{N}\sqrt{e\left(1 - \frac{e}{N}\right)} \tag{5.4}$$

Quando o taxa de erro é pequena, o intervalo de confiança depende somente do número de erros observados. Isto pode ser verificado na seguinte fórmula:

$$\frac{\text{intervalo de confiaça}}{\hat{p}} = \frac{2k\sqrt{e\left(1 - \frac{e}{N}\right)}}{e} = \frac{2k}{\sqrt{e}} \tag{5.5}$$

A EQ. 5.5 permite satisfazer a regra dos trinta já que, se escolhermos $k = 1,65$ (confiança de 90%), podemos observar que com 30 erros, o intervalo de confiança será igual a $0,60 \cdot \hat{p}$, o que nos diz que $\hat{p} + 0,30 \cdot \hat{p} \leq p \leq \hat{p} + 0,30 \cdot \hat{p}$, isto é a regra dos trinta.

Os intervalos de confiança das probabilidades de erro, P_{FR} e P_{FA} , podem ser expressos como $\hat{P} - k\sigma_{\hat{P}} \leq P \leq \hat{P} + k\sigma_{\hat{P}}$, onde $k = 1,65$ para confiança de 90%, P é a probabilidade sendo analisada e, da EQ. 5.4,

$$\sigma_{\hat{P}} = \frac{1}{\sqrt{N}}\sqrt{\hat{P}(1 - \hat{P})} \tag{5.6}$$

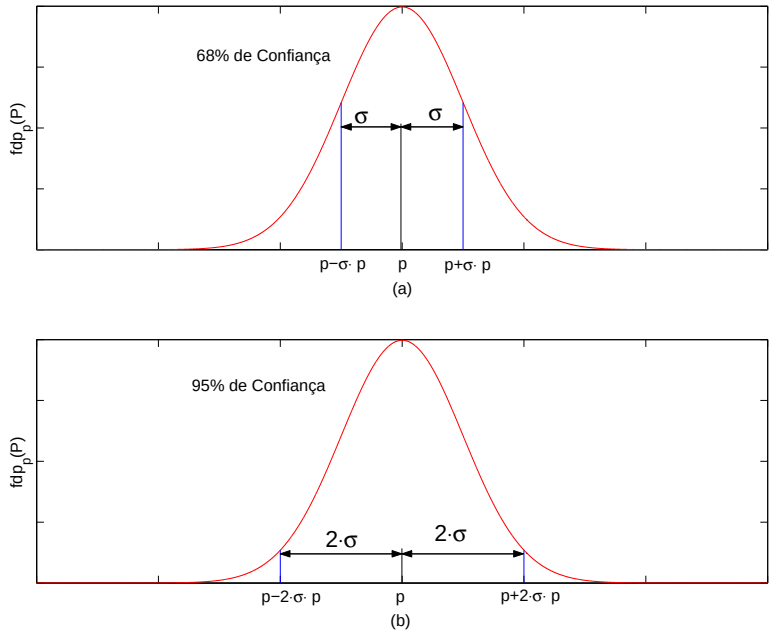

FIG. 5.2: Intervalo de confiança do estimador $\hat{p}$; (a) Para $k = 1$ (b) Para $k = 2$.

Para as condições de treinamento e teste do sistema, temos que $N_V = 600$ e $N_F = 23400$ para o cômputo de P_{FR} e P_{FA}, respectivamente. Nas FIG. 5.3 e FIG. 5.4, pode-se ver o intervalo de confiança, em porcentagem da probabilidade, para a probabilidades de falsa rejeição e falsa aceitação de um erro qualquer observado.

Da FIG. 5.5 podemos calcular o intervalo de confiança do EER, como:

$$\sigma_{EER} = \sqrt{\sigma^2_{P_{FR}} + \sigma^2_{P_{FA}}} \tag{5.7}$$

sendo que $\sigma_{P_{FA}}$ é, em geral, pequeno, $\sigma_{EER} \approx \sigma_{P_{FR}}$.

Pode-se observar que o intervalo de confiança do erro obtido no sistema é pequeno ($< 30\%$ do erro) para probabilidades de erro maiores a 5%, o que garante estatisticamente a validade dos resultados obtidos.

5.5 AVALIAÇÃO DO SISTEMA

O primeiro teste realizado foi usando sinais de treinamento limpos e sem uso de realce de voz. Os sinais de teste foram corrompidos com diferentes tipos de ruído e diferentes

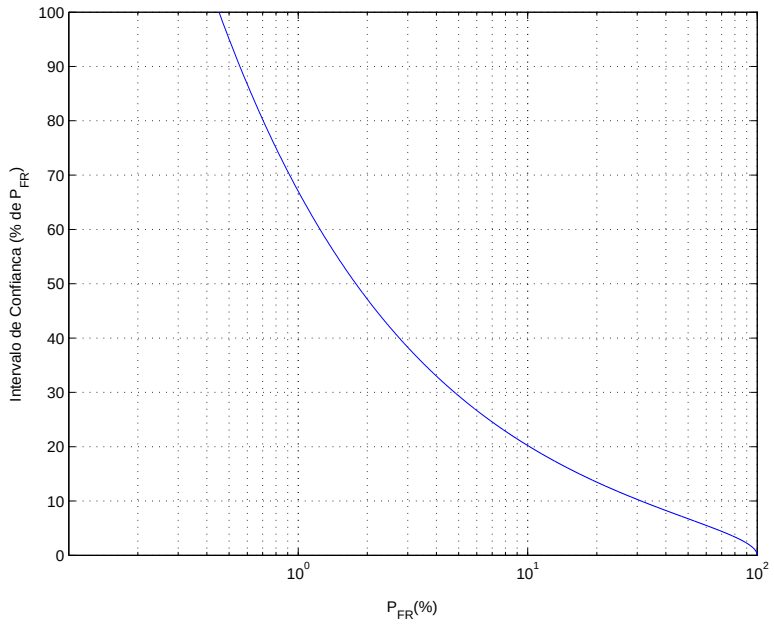

FIG. 5.3: Intervalo de confiança da P_{FR}.

relações sinal–ruído. Os resultados desta simulação estão listados na TAB. 5.2, para as seguintes situações: sem realce de voz no sinal de teste (WOSE), realce de voz no sinal de teste com o método de subtração espectral (SS), com o filtro de Eprahim-Malah (EMF), com o método de Virag, o método baseado em wavelets dos tipos SureShrink, VisuShrink, e o método proposto usando redes neurais (RN).

Observa-se, em geral, que o método de Virag apresenta um melhor desempenho que os demais métodos. O erro obtido em cada caso pode ser considerado muito alto. Isto torna difícil uma escolha adequada do limiar de decisão (θ) na verificação de locutor e estimula a busca de soluções alternativas que diminuam a taxa de erro obtida.

O segundo teste foi realizado corrompendo o sinal de treinamento, ou seja, foi somado ruído branco e depois aplicou-se o mesmo algoritmo de realce de voz do sinal de teste. A relação sinal–ruído do sinal de treinamento foi mantida constante e igual a SNR=5 dB. Os resultados estão listados na TAB. 5.3.

Dos resultados apresentados na TAB. 5.3, que complementam os obtidos em (**MEDINA**, 2003a), pode-se observar que:

1) Para ruído branco e exceto nos dois primeiros algoritmos (SS e EMF) a taxa de erro, para cada algoritmo, atinge o menor valor quando a SNR do sinal de teste é de $5dB$,

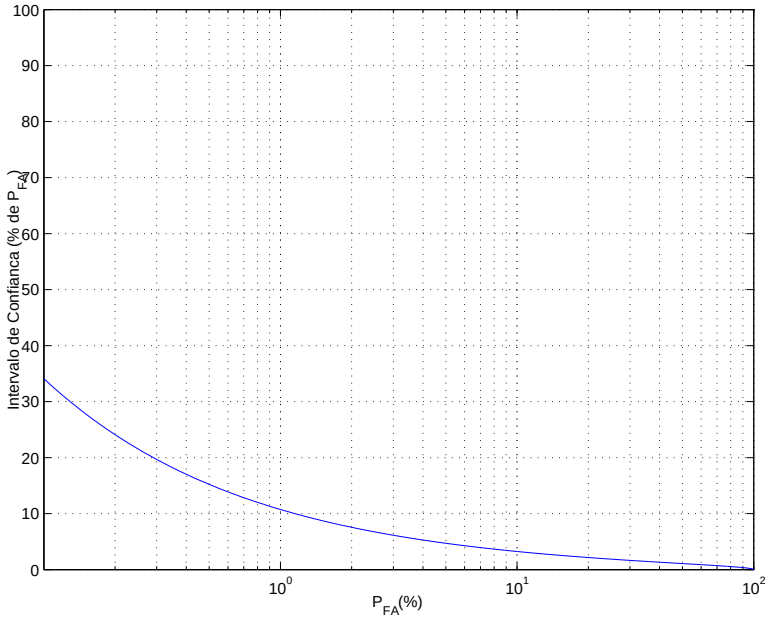

FIG. 5.4: Intervalo de confiança da P_{FA}.

isto é, igual ao sinal de treinamento. Repetiu-se as simulações variando a relação sinal–ruído do sinal de treinamento para $0\,dB$ e $-5\,dB$ e os resultados apresentaram o mesmo comportamento, ou seja, o menor erro obtido nas simulações foi quando a relação sinal–ruído do sinal de teste é a mesma que a relação sinal–ruído do sinal de treinamento, isto para o caso de ruído branco.

2) A taxa de erro, no caso de ruído branco, diminuiu em relação aos valores apresentados na TAB. 5.2. Para os outros tipos de ruído, a taxa de erro varia (desce ou sobe) dependendo do algoritmo.

   Para verificar o efeito do ruído empregado no sinal de treinamento, este experimento foi repetido, mas usando o ruído chamado sala de operações da base de dados Noisex-92. No caso de ruído branco no sinal de teste, a taxa de erro aumentou consideravelmente. Para o caso dos outros três tipos de ruído, as taxas de erros obtidas foram basicamente da mesma ordem do que as obtidas no caso da presença de ruído branco no sinal de teste e treinamento.

Estes resultados sugerem um cenário de treinamento diferente para cada sinal de teste, dependendo do ruído presente nele. Em (**MEDINA**, 2003b) é proposta uma abordagem

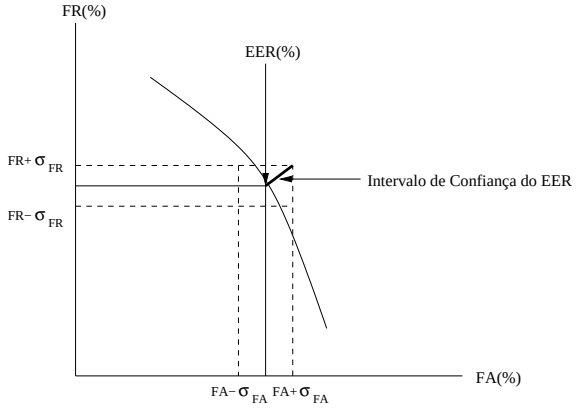

FIG. 5.5: Intervalo de confiança do EER.

para este cenário como indicado a seguir.

5.6 VERIFICAÇÃO DE LOCUTOR EM AMBIENTES DE RUÍDO COLORIDO

A FIG. 5.6 apresenta o módulo da transformada de Fourier[2] das 8192 primeiras amostras dos ruídos usados nas simulações. Observamos nesta figura que o espectro de amplitude do ruído branco é consideravelmente diferente dos espectros dos outros tipos de ruído da base Noisex-92.

Se usarmos como uma medida de similaridade o coeficiente de Pearson definido por:

$$S_p = \frac{cov(E_0,E_1)}{\sigma_0\sigma_1}; \quad -1 \leq S_p \leq 1 \tag{5.8}$$

onde E_0 e E_1 são os espectros de amplitude dos ruídos comparados, σ_0 e σ_1 são o desvio padrão destes espectros de media μ_0 e μ_1 e de tamanho N, e

$$cov\left(E_0,E_1\right) = \frac{1}{N}\sum_{k=0}^{N-1}\left(E_0(k)-\mu_0\right)\left(E_1(k)-\mu_1\right), \tag{5.9}$$

Observaremos que a similaridade entre o ruído sala de operações e os outros ruídos são: $0,02$ para ruído branco, $0,61$ para ruído *speech like*, $0,35$ para ruído de cabine de avião e $0,40$ para ruído de fábrica. Para ruído branco a similaridade entre este e os outros ruídos é: $0,05$ para ruído *speech like*, $0,04$ para ruído de cabine de avião e $0,06$ para ruído de fábrica.

[2]Também denominado de "Espectro de Amplitude"

TAB. 5.2: EER (%) para sinais de treinamento limpos e sinais de teste corrompidos com diferentes tipos de ruído e relação sinal–ruído.

SNR	WOSE	SS	EMF	Virag	SureShrink	VisuShrink	RN
Ruído Branco							
−5	47,58	47,76	45,76	**44,26**	47,92	50,58	49,75
0	44,76	47,59	45,92	**37,60**	47,59	45,92	48,42
5	41,10	49,09	44,43	**31,28**	44,93	43,93	45,76
10	28,79	47,92	41,60	**24,00**	43,10	39,60	39,60
Ruído *Speech Like*							
−5	29,95	49,92	43,10	**23,96**	31,95	33,94	33,44
0	20,30	45,59	38,94	**15,97**	21,46	23,63	21,96
5	**8,99**	46,08	35,44	12,31	11,98	11,98	12,15
10	4,66	43,26	30,78	9,65	**3,66**	3,83	3,99
Ruído de Cabine de Avião							
−5	49,25	48,09	**46,59**	47,25	47,75	50,08	48,91
0	46,92	47,75	41,93	**41,43**	45,92	46,08	46,92
5	43,10	46,92	39,60	**38,44**	40,76	41,43	41,26
10	33,28	46,26	33,28	**28,29**	30,28	30,94	31,11
Ruído de Fábrica							
−5	44,59	48,42	44,09	**36,44**	44,09	42,10	42,76
0	37,60	48,59	40,43	**24,63**	33,28	30,95	31,61
5	25,29	47,25	38,10	**14,48**	16,14	15,87	19,30
10	11,15	44,93	36,11	10,98	**7,99**	8,15	8,82

Isto justifica a diferença de desempenho do sistema de verificação de locutor quando é usado ruído branco ou ruído sala de operações no sinal de treinamento. Quanto maior a similaridade entre o ruído presente no sinal de teste e o ruído presente no sinal de treinamento, menor será a taxa de erro do sistema.

Como nem sempre se dispõe do sinal de treinamento gravado nas mesmas condições de ruído que o sinal de teste, faz-se necessário algum método para extrair do sinal de teste uma informação sobre o ruído, e com isso utiliza-lo para corromper um sinal de treinamento limpo e, daí, retreinar o sistema de verificação.

O método proposto, ilustrado na FIG. 5.7, é explicado a seguir:

Utilizando um classificador do sinal em voz/silêncio como o apresentado no Capítulo 2, extraímos as janelas de silêncio do sinal, aquelas onde se encontra presente somente o ruído. Com estas janelas estimamos o espectro de amplitude do ruído e a sua potência.

A seguir, o espectro de amplitude do ruído é modelado por um processo AR, MA ou ARMA. Foi escolhido para este trabalho o uso de um modelo de síntese de predição linear (LP). A análise de predição linear é baseada num preditor progressivo cuja minimização do

TAB. 5.3: EER (%) para sinais de treinamento corrompidos com ruído branco (SNR=5 dB) e sinais de teste corrompidos com diferentes tipos de ruído e relação sinal–ruído.

SNR	SS	EMF	Virag	SureShrink	VisuShrink	RN
Ruído Branco						
−5	36, 27	**33, 61**	35, 94	34, 61	33, 78	**33, 61**
0	24, 46	18, 30	17, 80	**16, 31**	16, 47	**16, 31**
5	15, 97	8, 65	6, 49	**6, 32**	**6, 32**	**6, 32**
10	13, 81	8, 15	8, 32	8, 15	**7, 82**	7, 99
Ruído *Speech Like*						
−5	**36, 77**	45, 92	47, 09	50, 08	48, 75	49, 08
0	**29, 78**	38, 10	44, 93	50, 25	48, 25	45, 92
5	**21, 80**	22, 96	45, 09	48, 09	45, 92	47, 92
10	**16, 64**	19, 80	43, 76	45, 09	45, 76	44, 43
Ruído de Cabine de Avião						
−5	**42, 10**	44, 59	48, 42	49, 58	48, 42	49, 75
0	34, 11	**31, 78**	47, 42	46, 76	48, 75	46, 76
5	22, 46	**20, 30**	40, 27	43, 76	45, 59	42, 76
10	**16, 14**	17, 14	40, 60	44, 43	44, 59	45, 09
Ruído de Fábrica						
−5	**42, 76**	46, 09	48, 42	48, 92	48, 59	48, 25
0	35, 44	**35, 11**	44, 09	46, 59	48, 42	49, 42
5	**23, 96**	25, 62	36, 27	46, 42	47, 75	42, 43
10	19, 47	**17, 14**	35, 44	42, 26	46, 42	45, 92

erro quadrático médio resulta em um modelo de síntese definido como (PICONE, 1991):

$$H_{LP} = \frac{1}{1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) z^{-i}} \tag{5.10}$$

onde a_{LP} são os coeficientes de predição linear (LPC). Para o cômputo eficiente destes coeficientes pode ser usado um algoritmo tal como o algoritmo recursivo de Levinson–Durbin (PICONE, 1991).

Nesta dissertação foram usados 15 coeficientes de predição linear para modelar o filtro H_{LP} que gera, a partir de uma entrada de ruído branco, uma saída de ruído colorido com aproximadamente o mesmo espectro do ruído original.

Após aplicar o procedimento de realce da voz do sinal de teste, calculamos a sua potência. Assim, juntamente com a potência do ruído, podemos estimar a relação sinal–ruído do sinal de teste original. Essa estimativa é usada no ruído gerado, o qual é multiplicado por uma constante S que faz com que a relação sinal–ruído do sinal de treinamento seja aproximadamente igual à relação sinal–ruído estimada do sinal de teste.

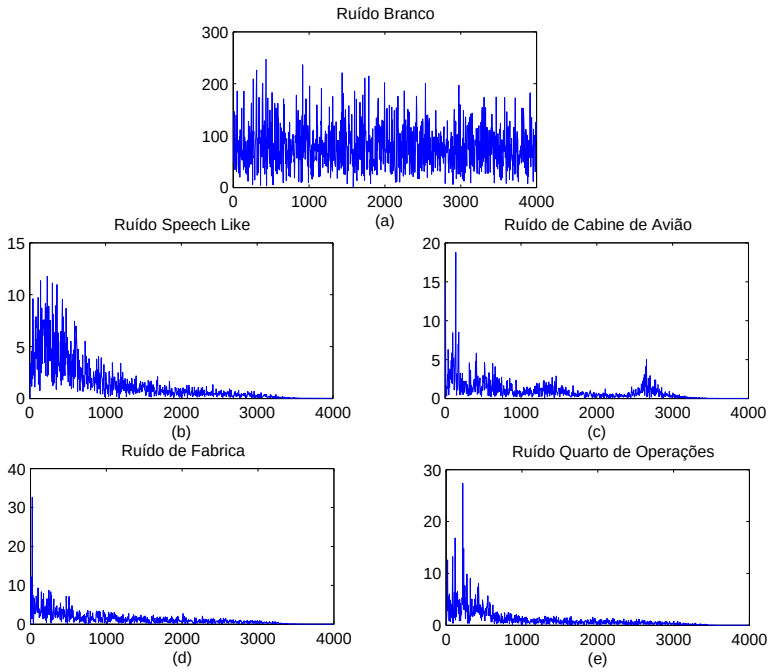

FIG. 5.6: Espectro de amplitude dos ruídos usados nas simulações.

O passo seguinte consiste em adicionar o ruído gerado ao sinal de treinamento. O sistema de verificação de locutor deve ser novamente treinado, mas desta vez, com este sinal corrompido com o ruído gerado. Deverá usar-se, na etapa de pré–processamento do sistema, um algoritmo de realce de voz, tanto para o treinamento como para o teste.

Neste trabalho, devido à grande quantidade de testes necessários para levantar a EER, o ruído não foi estimado do sinal de teste, mas diretamente do ruído armazenado na base de dados Noisex-92. Justifica-se esta simplificação pelos resultados obtidos da medida de similaridade entre o espectro estimado da base de dados e o espectro estimado do sinal de teste. Como exemplo, são apresentados na FIG. 5.8 três diferentes espectros de amplitude do ruído cabine de avião: o primeiro, foi extraído diretamente do sinal de ruído gravado na base de dados Noisex-92, o segundo é o estimado mediante o modelo de predição linear do ruído da base de dados e o terceiro é o estimado das janelas de silêncio presentes num sinal de teste.

Observa-se que a diferença entre os espectros de amplitude do ruído estimado diretamente do sinal da base de dados e o estimado das janelas de silêncio não é grande. Usando

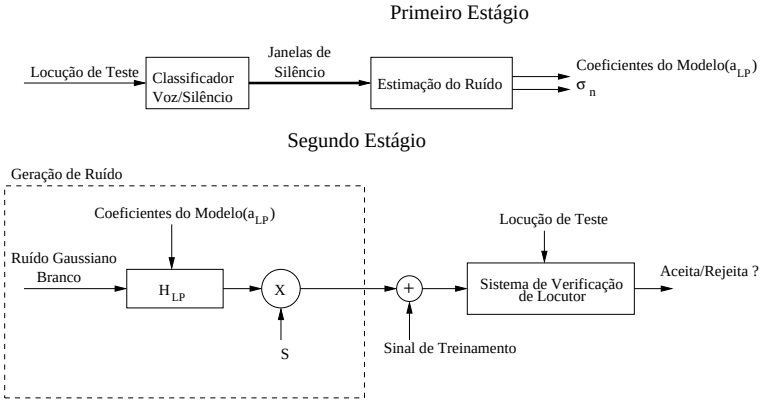

FIG. 5.7: Esquema de Verificação de Locutor com modelagem de ruído.

o coeficiente de Pearson (EQ. 5.8), a similaridade destes dois espectros é de $0,662$. A medida de similaridade entre o espectro real do ruído (obtido diretamente da base de dados Noisex-92) e os espectros estimados da base de dados e das janelas de silêncio é $0,593$ e $0,596$, respectivamente. Por isto (e pela facilidade de simulação) foi usado o primeiro.

Os resultados obtidos com este método estão listados na TAB. 5.4.

Observa-se que os esquemas de subtração espectral e Eprahim–Malah apresentam melhorias no desempenho, mas a taxa de erro é sempre maior que as apresentadas pelos outros esquemas. O método de Virag e os baseados em wavelets apresentam taxas de erro que podem ser consideradas relativamente baixas (principalmente se comparamos com os resultados inicias da TAB. 5.2), tornando o sistema robusto a diferentes tipos ruído e dando uma maior segurança ao escolher o limiar de decisão.

5.7 RESUMO

Neste capítulo foram apresentados os resultados de verificação de locutor nas simulações realizadas sob diferentes ambientes de ruído. Investigando as causas do desempenho relativamente baixo atingido pelo sistema, foi proposto um método para realizar a verificação de locutor baseado num sinal de treinamento somado a um ruído gerado artificialmente modelado a partir do ruído presente no sinal de teste. Dada a sua natureza, este método apresenta a limitação de ser útil só em aplicações *off-line* como aplicações forenses; isto se deve à necessidade de um novo treinamento em caso de mudar o ambiente de gravação

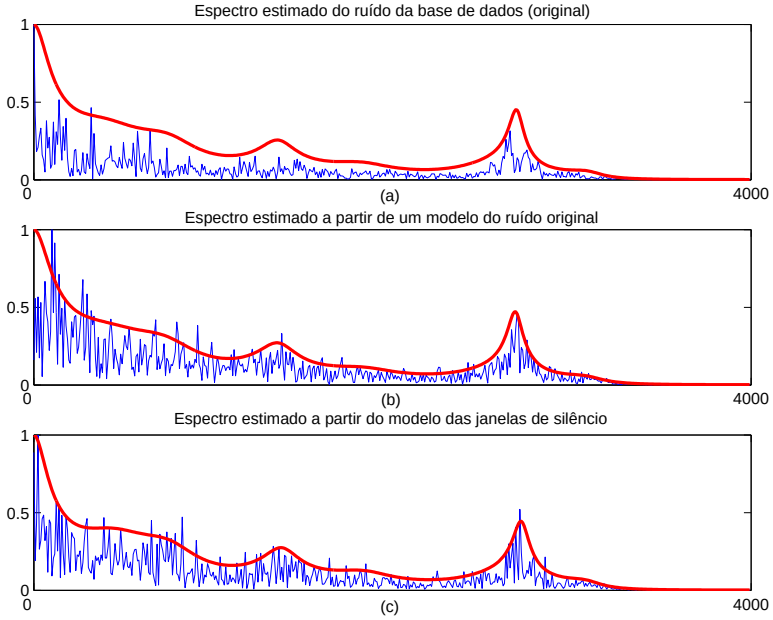

FIG. 5.8: Espectro do ruído cabine de avião.

(nível e tipo de ruído) do sinal de teste.

TAB. 5.4: EER (%) para sinais de treinamento corrompidos somando ruído modelado.

SNR	SS	EMF	Virag	SureShrink	VisuShrink	RN
Ruído Branco						
-5	$21,46$	$10,98$	$9,65$	$8,65$	$\mathbf{8,32}$	$8,65$
0	$19,47$	$10,15$	$7,15$	$7,15$	$7,15$	$\mathbf{6,66}$
5	$16,64$	$7,82$	$7,15$	$\mathbf{5,99}$	$6,16$	$6,16$
10	$9,98$	$6,99$	$\mathbf{4,66}$	$5,16$	$4,83$	$5,32$
Ruído *Speech Like*						
-5	$20,80$	$5,66$	$\mathbf{4,16}$	$5,16$	$5,16$	$4,33$
0	$8,82$	$3,33$	$\mathbf{2,50}$	$2,83$	$2,83$	$2,66$
5	$6,16$	$2,33$	$\mathbf{1,16}$	$2,33$	$2,33$	$2,16$
10	$4,16$	$2,00$	$\mathbf{1,16}$	$2,00$	$2,00$	$2,00$
Ruído de Cabine de Avião						
-5	$25,79$	$10,98$	$\mathbf{7,99}$	$11,65$	$11,48$	$10,82$
0	$12,98$	$6,32$	$\mathbf{4,83}$	$5,32$	$5,16$	$\mathbf{4,83}$
5	$7,49$	$4,00$	$3,33$	$3,33$	$3,33$	$\mathbf{3,16}$
10	$6,16$	$3,00$	$\mathbf{2,50}$	$2,66$	$2,66$	$2,83$
Ruído de Fábrica						
-5	$40,60$	$36,61$	$16,97$	$16,31$	$16,31$	$\mathbf{15,97}$
0	$20,63$	$16,97$	$6,65$	$6,16$	$\mathbf{5,66}$	$\mathbf{5,66}$
5	$9,15$	$7,49$	$4,16$	$\mathbf{3,00}$	$\mathbf{3,00}$	$3,16$
10	$7,32$	$4,16$	$3,16$	$\mathbf{2,83}$	$\mathbf{2,83}$	$\mathbf{2,83}$

6 CONCLUSÕES E COMENTÁRIOS FINAIS

6.1 CONCLUSÕES

Nesta dissertação, foi abordado o tema realce de voz e sua eficiência na verificação automática de locutor (VAL) independente do texto, quando os sinais de voz estão corrompidos com ruído aditivo.

As técnicas de realce de voz usadas neste trabalho são as derivadas de subtração espectral e as baseadas em wavelets. Uma investigação destas últimas, levada a cabo no Capítulo 4, conduziu à proposta de um novo método de cálculo do limiar baseado em redes neurais.

A avaliação objetiva (Ganho SRN_{seg}) dos algoritmos de realce de voz, levou-nos às seguintes conclusões:

- Os algoritmos derivados da subtração espectral apresentam bons resultados quando aplicados a sinais com ruído branco ou colorido;
- Para ruído branco e baixo SNR ($\leq 0\,dB$) os algoritmos baseados em wavelets apresentam resultados semelhantes aos resultados obtidos dos algoritmos derivados de subtração espectral;
- Para ruído branco e alto SNR ($> 0\,dB$) o desempenho dos algoritmos baseados em wavelets diminui bastante (com exceção do caso onde calculamos o limiar com o método proposto);
- Para o caso de ruídos coloridos, os algoritmos baseados em wavelets apresentam um desempenho menor daquele que obtiveram em presença de ruído branco;
- Os algoritmos baseados em wavelets propiciam a obtenção de sinais realçados com pouco ruído, mas com uma distorção similar a quando ocorre saturação em amplificadores de áudio. Por outro lado, os algoritmos derivados da subtração espectral não apresentam esta distorção de modo acentuado, mas introduzem ruídos residuais, principalmente o ruído musical, tornando-os menos confortáveis para o ouvinte.

Depois de realizada a avaliação objetiva (em termos de G_{SNR}) dos algoritmos de realce de voz, foram efetuadas simulações para avaliar o desempenho destes algoritmos na tarefa de verificação automática de locutor. Em todos os testes realizados, usou-se sinais de teste corrompidos com 4 tipos de ruído: ruído gaussiano branco, ruído *speech like*, ruído de cabine de avião e ruído de fábrica com diferentes relações sinal–ruído: $-5\,dB$, $0\,dB$, $5\,dB$ e $10\,dB$.

O primeiro teste realizado foi usando sinais de treinamento limpos. As taxas mais baixas de EER foram em sua maior parte obtidas usando o método de Virag, mas estas taxas ainda não são suficientemente baixas para a implementação prática de um sistema VAL.

O segundo teste implementado foi usando sinais de treinamento corrompidos com ruído gaussiano branco e aplicando os algoritmos de realce de voz na etapa de pré–processamento do sistema, usaram-se diferentes relações sinal–ruído: SNR = $5\,dB$, SNR = $0\,dB$ e SNR = $-5\,dB$. Os resultados sugerem que a menor taxa de erro é obtida quando existe um casamento entre a relação sinal–ruído presente no sinal de treinamento e a presente no sinal de teste. Além disso, os resultados só melhoraram (baixaram os valores de EER) para o caso de ruído gaussiano branco no sinal de teste.

A seguir, implementou-se o mesmo experimento com um outro tipo de ruído somado ao sinal de treinamento, o ruído sala de operações da base Noisex-92. Uma análise da taxa de erro obtida levou à conclusão que, se existe um alto grau de similaridade entre os espectros de amplitude do ruído presente no sinal de treinamento e do ruído presente no sinal de teste, a taxa de erro cai a valores até 20 vezes menores dos apresentados nos testes anteriores.

Estes últimos resultados levaram à proposição de um novo método. O método foi sugerido em (**MEDINA**, 2003b) e é baseado na modelagem do ruído presente no sinal de teste para corromper, com o ruído modelado, o sinal de treinamento. Este sinal de treinamento serve para retreinar o sistema de verificação de locutor, aplicando na etapa de pré–processamento um algoritmo de realce de voz. Com este método obteve-se uma taxa de erro menor que a apresentada nas simulações anteriores.

De uma análise geral das taxas de erro, conclui-se que:

- não é conveniente realizar o treinamento do sistema de verificação de locutor com sinais de treinamento limpos de ruído, sempre que o sinal de teste tenha ruído aditivo;

- no caso de usar o esquema de modelagem do ruído presente no sinal de treinamento, o método de Virag é sempre superior aos outros dois métodos derivados da subtração espectral;
- fazendo uso dos dois esquemas propostos -o de realce de voz baseado em wavelets e redes neurais e o esquema de modelagem do ruído presente no sinal de teste- os resultados obtidos são, em 70% dos casos, superiores aos resultados obtidos pelos outros dois métodos baseados em wavelets;
- em 40% dos casos, o uso dos dois esquemas propostos resulta superior ao uso do método de Virag. Entretanto, levando-se em consideração o intervalo de confiança, podemos afirmar que, estatisticamente, os dois métodos são equivalentes;
- os erros apresentados pelo esquema de modelagem de ruído presente no sinal de teste, possibilitam a implementação prática da verificação automática de locutor em ambientes de ruído aditivo colorido;
- o uso deste esquema de verificação automática de locutor limita seu uso em aplicações *off-line*, porém de muito interesse, como o caso de verificação de locutor para fins forenses;
- se for preciso usar um sistema de verificação de locutor *on-line*, sugere-se fazer uma análise previa das condições do ruído presente no ambiente comum do local de operação do sistema e realizar o treinamento do mesmo corrompendo o sinal de treinamento com ruído de características similares ao encontrado na análise realizada.

6.2 TRABALHOS FUTUROS

No processamento com wavelets, existem implementações que poderiam ser estudadas com vista a obter melhores resultados que os obtidos nesta dissertação. Podemos citar por exemplo:

- o uso de pacotes de wavelets no *denoising*;
- o uso de esquemas de decomposição wavelet que usam bases ortonormais escolhidas de um conjunto de bases previamente definidas. A escolha das bases usadas é

especifica para o sinal que está sendo processado, para isto segue-se um esquema como o proposto em (DONOHO, 1994a);

- nesta dissertação sempre foram processados todos os níveis da transformada wavelet do sinal com ruído, mas, sugere-se um estudo de quais níveis da transformada devem, efetivamente, ser processados.

6.3 COMENTÁRIOS FINAIS

A demanda por sistemas de verificação automática de locutor tem sido crescente, o que provoca o interesse por uma grande quantidade de pesquisa nesta área. Estas pesquisas visam a solução dos principais problemas encontrados pelos sistemas VAL tais como a degradação do sinal devido à distorção introduzida por um canal e pelo ruído aditivo.

Como vimos, pelos resultados das simulações, o ruído aditivo deteriora seriamente o desempenho de um sistema VAL. Os esquemas propostos conseguiram melhorar este desempenho para níveis bem mais aceitáveis.

As pesquisas realizadas com a língua portuguesa são escassas se comparadas com a língua inglesa; além disso, devido à falta de uma base de dados gravada em condições reais, os testes são limitados pela simulação de tais condições. Em um futuro próximo, espera-se poder contar com uma base de dados gravada em condições reais para uma melhor avaliação dos sistemas. Uma base piloto gravada a partir de telefones fixos e móveis está sendo produzida pelo IME.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AGBINYA, J. I. Discrete wavelet transform techniques in speech processing. *IEEE TENCON - Digital Signal Processing Applications*, págs. 514–519, 1996.

ALCAIM, A., SOLEWICZ, J. A. e MORAES, J. A. Freqüência de ocorrência dos fonemas e listas de frases fonéticamente balanceadas no português falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, 7(1), dezembro 1992.

APOLINÁRIO, J. A. J., WERNER, S. e DINIZ, P. S. R. Conjugate gradient algorithm with data selective updating. Em *Simpósio Brasileiro de Telecomunicações–SBT*, 2001 2001.

APOLINÁRIO JR., J. A., **MEDINA**, **C. A**. e DINIZ, P. S. R. Infinite precision analysis of the Fast QR Algorithm Based on Backward Prediction Errors. *Revista da Sociedade Brasileira de Telecomunicações*, 18:123–133, dezembro 2002.

APOLINÁRIO JR., J. A., WERNER, S., DINIZ, P. S. R. e LAAKSO, T. I. Constrained normalized adaptive filters for CDMA mobile communications. Em *Proceedings EUSIPCO - European Signal Processing Conference*, Rodes, Grécia, setembro 1998.

ATAL, B. S. Automatic recognition of speaker from their voices. Em *Proceedings of the IEEE*, volume 64, págs. 460–475, abril 1976.

BAHOURA, M. e ROUAT, J. Wavelets speech enhancement based on the teager energy operator. *IEEE Signal Processing Letters*, 8(1):10–12, janeiro 2001.

BEROUTI, M., SCHWARTZ, R. e MAKHOUL, J. Enhancement of speech corrupted by acoustic noise. Em *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, págs. 208–211, Paris, França, abril 1979.

BEZERRA, M. D. R. Reconhecimento automático de locutor para fins forenses, utilizando técnicas de redes neurais. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, 1994.

BOLL, S. F. A spectral subtraction algorithm for suppression of acoustic noise in speech. Em *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, págs. 200–203, Paris, França, abril 1979a.

BOLL, S. F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–27, abril 1979b.

BUCKLEY, K. M. Spatial/spectral filtering with linearly constrained minimum variance beamformers. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–35(3):249–266, março 1987.

BURRUS, C. S., GOPINATH, R. A. e GUO, H. *Introduction to Wavelets and Wavelets Transforms, A Primer*. Prentice Hall, 1998.

CAMPBELL JR., J. P. Speaker recognition: A tutorial. Em *Proceedings of the IEEE*, volume 85, págs. 1437–1462, setembro 1997.

CAMPOS, L. R. M. e APOLINÁRIO JR., J. A. The constrained affine projection algorithm–development and convergence issues. Em *First IEEE Balkan Conference on Signal Processing, Communications, Circuits and Systems*, Istanbul, Turquia, junho 2000.

CAMPOS, L. R. M., APOLINÁRIO JR., J. A. e LAAKSO, T. I. Constrained Quasi–Newton algorithm for CDMA mobile communications. Em *Proceedings SBT/IEEE International Telecommunications Symposium*, págs. 371–376, São Paulo, Brasil, agosto 1998.

DAVIS, S. B. e MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28(4), agosto 1980.

DELLER JR., J. R., HANSEN, J. H. L. e PROAKIS, J. G. *Discrete–Time Processing of Speech Signals*. IEEE Press, New York, 2000.

DINIZ, P. S. R. *Adaptive Filtering: Algorithms and Practical Implementation*. The Kluwer international series in engineering and computer science. Boston:Kluwer Academic Publishers, segunda edição, julho 1997.

DODDINGTON, G. R. Speaker recognition – identifying people by their voices. Em *Proceedings of the IEEE*, volume 73, págs. 1641–1664, novembro 1985.

DONOHO, D. L. De–noising by soft–thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, maio 1995.

DONOHO, D. L. e JOHNSTONE, I. M. Ideal spatial adaptation via wavelet shrinkage. Technical report, Department of Statistics, Stanford University, 1992.

DONOHO, D. L. e JOHNSTONE, I. M. Ideal denoising in an orthonormal basis chosen from a library of bases. Technical report, C. R. Acad. Sci. Paris, Ser. I,319; Stanford Statistics Dep. Report 461, setembro 1994a.

DONOHO, D. L. e JOHNSTONE, I. M. Threshold selection for wavelet shrinkage of noisy data. Em *Proceedings of the 16th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, págs. 24a–25a, Maryland, USA, novembro 1994b.

DONOHO, D. L. e JOHNSTONE, I. M. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

EPHRAIM, Y. e MALAH, D. Speech enhancement using a minimum mean–square error short–time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(6):1109–1121, dezembro 1984.

FLIELLER, A., LARZABAL, P. e CLERGEOT, H. Geometrical determination of ambiguities in bearing estimation for sparse linear arrays. Em *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Seattle, Washington, USA, maio 1998.

FROST, O. L. An algorithm for linearly constrained adaptive array processing. Em *Proceedings of the IEEE*, volume 60, agosto 1972.

FURUI, S. *Automatic Speech and Speaker Recognition, Advanced Topics*. Boston:Kluwer Academic Publishers, 1996.

GABREA, M. e TADJ, C. Speech enhancement for speaker identification. Em *International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, setembro 2001.

GRIFFITHS, L. J. e JIM, C. W. An adaptive approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, AP-30(1), janeiro 1982.

HAYKIN, S. *Adaptive Filter Theory*. Prentice-Hall Signal Processing Series, Englewood Cliffs, 1986.

HERMANSKY, H., MORGAN, N., BAYYA, A. e KOHN, P. RASTA-PLP speech analysis technique. Em *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, págs. I.121–I.124, San Francisco, USA, março 1992.

HÄRDLE, W., KERKYACHRIAN, G., PICARD, D. e TSYBAKOV, A. Wavelets approximation and statistical applications. Uma versão web do livro: Wavelets, Approximation and Statistical Applications, Lecture Notes in Statistics (Springer-Verlag), Vol. 129. Springer-Verlag, New York., setembro 1997. http://www.quantlet.de/scripts/wav/wavpdf.html [capturado em 1 de Maio de 2003].

JOHNSTON, J. D. Transform coding of audio signals using perceptual noise criteria. *Journal on Selected Areas in Communications*, 6(2):314–323, fevereiro 1988.

LIM, J. S. e OPPENHEIM, A. V. Enhanced and bandwidth compression of noisy speech. Em *Proceedings of the IEEE*, volume 67, dezembro 1979.

LIMA, C. B. Sistemas de verificação de locutor independente do texto baseados em GMM e AR–vetorial utilizando PCA. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, 2001.

MARTIN, A., DODDINGTON, G., KAMM, T., ORDOWSKI, M. e PRZYBOCKI, M. The DET curve in assessment of detection task performance. *Proceedings of the European Conference on Speech Technology*, págs. 1895–1898, 1997.

MCAULAY, R. J. e MALPASS, M. L. Speech enhancement using a soft–decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28, abril 1980.

MCCOWAN, I., PELECANOS, J. e SRIDHARAN, S. Robust speaker recognition using microphone arrays. Em *Proceedings of 2001: A speaker odyssey*, junho 2001.

MEDINA, **C. A**., APOLINÁRIO JR., J. A. e ALCAIM, A. Modern speech enhancement techniques in text–independent speaker verification. Aceito para publicação nos Anais do XX Simpósio Brasileiro de Telecomunicações–SBT 2003, outubro 2003a.

MEDINA, **C. A**., APOLINÁRIO JR., J. A., ALCAIM, A. e ALVES, R. G. Robust speaker verification in colored noise environment. Submetido a: Thirty-Seventh Annual Asilomar Conference on Signals, Systems, and Computers, novembro 2003b.

MEDINA, **C. A**., APOLINÁRIO JR., J. A. e DINIZ, P. S. R. Infinite precision analysis of the Fast QR Algorithm Based on a Posteriori Backward Prediction Errors. Em *Proceedings SBT/IEEE International Telecommunications Symposium*, Natal/RN , Brasil, setembro 2002a.

MEDINA, **C. A**., APOLINÁRIO JR., J. A. e SIQUEIRA, M. S. A unified framework for Multichannel Fast QRD-LS Adaptive Filters Based on Backward Prediction Errors. Em *45th IEEE International Midwest Symposium on Circuits and Systems*, volume 3, págs. 668–671, Oklahoma, USA, agosto 2002b.

NBTC, 2000. National biometric test center, collected works 1997–2000, agosto 2000. www.engr.sjsu.edu/biometrics/nbtccw.pdf [capturado em 22 de Maio de 2003].

OPPENHEIM, A. e SCHAFER, R. *Discrete–Time Signal Processing*. Prentice-Hall Signal Processing Series, 1989.

PAINTER, T. e SPANIAS, A. Perceptual coding of digital audio. Em *Proceedings of the IEEE*, volume 88, págs. 452–513. IEEE, abril 2000.

PICK JR., H. L., SIEGEL, G. M., FOX, P. W., GARBER, S. R. e KEARNEY, J. K. Inhibiting the Lombard effect. *The Journal of the Acoustical Society of America*, 85 (2):894–900, fevereiro 1989.

PICONE, J. W. Signal modeling techniques in speech recognition. Em *Proceedings of the IEEE*, volume 81, págs. 1215–1247, setembro 1991.

RABINER, L. e SAMUR, M. Technical report, The Bell System Technical Journal, fevereiro 1974.

RABINER, L. e SCHAFER, R. *Digital Processing of Speech Signals*. Prentice-Hall Signal Processing Series, 1978.

RESENDE, L. S., ROMANO, J. M. T. e BELLANGER, M. G. A fast least–squares algorithm for linearly constrained adaptive filtering. *IEEE Transactions on Signal Processing*, 44(5):1168–1174, maio 1996.

REYNOLDS, D. A. *A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification*. Tese de Doutorado, Georgia Institute of Technology, 1992.

REYNOLDS, D. A. Robust text–independent speaker identification using gaussian mixture speaker model. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, janeiro 1995.

REYNOLDS, D. A. e HECK, L. P. Automatic speaker recognition–recent progress, current applications and future trends. Em *AAAS 2000 Metting Humans, Computers, and Speech Symposium*, fevereiro 2000a.

REYNOLDS, D. A., QUATIERI, T. F. e DUNN, R. B. Speaker verification using adapted gaussian mixture models. Em *Digital Signal Processing*, volume 10, págs. 19–41. 2000b.

RODRÍGUEZ, J. G. Panorámica de los esquemas de mejora de voz en presencia de ruido. Em TELECOMUNICACIÓN, E., editor, *SEAF–2000, I Congreso de la Sociedad Española de Acústica Forense*, págs. 25–41, Universidad Politécnica de Madrid, outubro 2000.

ROSE, R. C. ANS HOFSTETTER, E. M. e REYNOLDS, D. A. Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, 2(2):245–257, abril 1994.

ROSEMBERG, A. E. Automatic speaker verification: A review. Em *Proceedings of the IEEE*, volume 64, págs. 475–487, abril 1976.

SHU, Y., LI, X. e ZHANG, R. Adaptive speech enhancement based on wavelet in high noise envinroment. Em *Proceedings of First International Conference on Machine Learning and Cybernetics*, págs. 885–889, Beijing, China, novembro 2002.

SILVA, D. G. Estudo de compensação de canal e análise fractal aplicada ao reconhecimento automático de locutor. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, 2002.

SINHA, D. e TEWFIK, A. H. Low bit rate transparent audio compression using adapted wavelets. *IEEE Transactions on Signal Processing*, 41(12):3463–3479, dezembro 1993.

SOLEWICZ, Y. A. Noise robustness in forensic speaker verification. Em *Proceedings of 2001: A speaker odyssey*, junho 2001.

STEVENS, S. S. e VOLKMAN, J. The relation of pitch to frequency. Technical report, Journal of Psychology, 1940.

VAN VENN, B. D. e BUCKLEY, K. M. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, págs. 4–24, abril 1988.

VARGA, A. P., STEENEKEN, H. J. M.AND TOMLINSON, M. e JONES, D. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, Defense Research Agency, Malvern, U.K., 1992.

VIRAG, N. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, 7(2):126–137, março 1999.

VUUREN, V. S. *Speaker Verification in a Time–Feature Space*. Tese de Doutorado, Oregon Graduate Institute of Science and Technology, 1999.

WAKAO, A., TAKEDA, K. e ITAKURA, F. Variability of Lombard effects under different noise conditions. Em *Proccedings ICSLP '96*, volume 4, págs. 2009–2012, Philadelphia, PA, 1996.

WOODLAND, P. *Speech Recognition*. The Institute of Electrical Engineers, Savoy Place, London, WC2R OBL, UK, 1998.