UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

ALGORITMOS ROBUSTOS DE RECONHECIMENTO DE VOZ APLICADOS A VERIFICAÇÃO DE LOCUTOR

Tarciano Facco Pegoraro

DISSERTAÇÃO DE MESTRADO

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

ALGORITMOS ROBUSTOS DE RECONHECIMENTO DE VOZ APLICADOS A VERIFICAÇÃO DE LOCUTOR

Tarciano Facco Pegoraro

Orientador: Prof. Dr. Néstor Jorge Becerra Yoma (Universidad de Chile)
Co-Orientador: Prof. Dr. João Marcos Travassos Romano

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como requisito parcial para a obtenção do Título de Mestre em Engenharia Elétrica.

Banca Examinadora:

Prof. Dr. João Marcos T. Romano (FEEC/UNICAMP)
Prof. Dr. Lee Luan Ling (FEEC/UNICAMP)
Prof. Dr. Jaime Portugheis (FEEC/UNICAMP)
Prof[a]. Dr[a]. Maria Miranda (Instituto Mackenzie)

Campinas, junho de 2000

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

P349a

Pegoraro, Tarciano Facco
Algoritmos robustos de reconhecimento de voz aplicados a verificação de locutor / Tarciano Facco Pegoraro.--Campinas, SP: [s.n.], 2000.

Orientadores: Néstor Jorge Becerra Yoma e João Marcos Travassos Romano.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Reconhecimento automático da voz. 2. Sistemas de reconhecimento de padrões. 3. Processamento digital de sinais. 4. Processos estocásticos. I. Becerra Yoma, Néstor Jorge. II. Romano, João Marcos Travassos. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

RESUMO

A voz é uma característica biométrica e, além das informações fonético-lingüísticas detectadas e classificadas pelos sistemas de reconhecimento de voz, também carrega informações que podem ser empregadas em tarefas de reconhecimento de locutor. Entretanto, sistemas de reconhecimento de voz e locutor sofrem uma sensível queda em seu desempenho na presença de ruído, tanto aditivo quanto convolucional. Esta dissertação mostra os estágios da implementação de um Sistema de Verificação de Locutor (SVL) e testes com algoritmos de robustez a ruído geralmente empregados em Sistemas de Reconhecimento de Voz. É realizado um breve estudo sobre a influência do ruído sobre a tecnologia de verificação de locutor e sobre as técnicas de robustez. Para os experimentos com o SVL são utilizadas três técnicas convencionais (subtração espectral (SS), a normalização da média cepstral (CMN), *Log*-RASTA) e um método de modelamento de duração de estados (MDE) com restrições temporais, recentemente proposto. Como verificado em reconhecimento de voz, todas estas técnicas também forneceram um bom desempenho para o SVL em questão. O ruído convolucional é quase que totalmente cancelado por CMN ou *Log*-RASTA, e o ruído aditivo tem sua influência bastante reduzida principalmente com o emprego conjunto de SS e MDE com truncamento simples. Com a presença de ambos os ruídos, SS, *Log*-RASTA e MDE com truncamento simples conjuntamente reduziram em até 87% a taxa de erros iguais. Verifica-se também que a presença de CMN e principalmente *Log*-RASTA reduz significativamente a variabilidade do limiar de decisão. A pesar dos resultados aqui apresentados mostrarem um importante avanço, a robustez de sistemas de reconhecimento de voz e de locutor a ruídos interferentes ainda são um problema complexo, e é o principal empecilho enfrentado em aplicações práticas reais.

ABSTRACT

The speech carries linguistic information that can be classified by speech recognition systems, and also information related to the speaker's characteristics, which is employed by speaker recognition methods. However, speech and speaker recognition tasks have the performance strongly degraded by noise environments, and this dissertation presents the results of experiments with a speaker verification system combined with noise robust algorithms usually used in speech recognition. Three conventional techniques were tested (spectral subtraction (SS), cepstral mean subtraction (CMN) and RASTA filtering) and a method for state duration modeling with temporal restrictions (MDE) that has recently been proposed. Firstly, an introduction to acoustic pattern matching algorithms is presented, and the speaker verification system employed in this dissertation is briefly described. Secondly, noise robust techniques are analyzed and discussed. Finally, these techniques are tested in the speaker verification system to cancel both additive and convolucional noise, and the combinations of the noise robust methods are evaluated and compared. This dissertation shows that the techniques here addressed can give a high improvement in a speaker verification system, although the noise robustness of speech and speaker recognition systems is still a complex topic and the main problem to be addressed to make successful real applications of this technology.

Este trabalho foi financiado pelo CNPq.

AGRADECIMENTOS

Ao final deste período de trabalho, gostaria de agradecer às pessoas as quais me ajudaram a vencer as barreiras que surgiram no decorrer destes anos. Agradeço:

- Aos meus pais, Valdir e Tereza, pelo seu apoio e amor infinito;
- Aos meus irmãos Rodinei e Ronaldo e a vó Adelina;
- Ao Néstor e ao Prof. João Marcos por sua orientação e amizade;
- À Laize por seu amor e sua companhia;
- Aos colegas do LRPRC e ao Prof. Lee pela presteza e, acima de tudo, amizade que se firmou no decorrer destes anos;
- Aos amigos, em especial ao Rodrigo, pelo companheirismo e compreensão;
- A Deus pela sua infinita bondade;
- E a todos aqueles que de uma forma ou de outra contribuíram para minha formação pessoal e profissional.

A eles, meus sinceros agradecimentos

Tarciano Facco Pegoraro

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

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

DTW – *Dynamic Time Warping*

EER – Taxa de erros iguais (*Equal Error Rate*)

EER_{SS} – Taxa de erros iguais média considerando-se limiares de decisão distintos para cada locutor

EER_{SI} – Taxa de erros iguais média considerando-se um limiar de decisão único para o sistema.

FA – Falsa Aceitação

FFT – Transformada rápida de Fourier (*Fast Fourier Transform*)

FR – Falsa Rejeição

HMM – Modelos ocultos de Markov (*Hidden Markov Models*)

IL – Identificação de Locutor

LPC – Coeficientes de predição linear (*Linear Predictive Coeficientes*)

MDE – Modelamento de Duração de Estados

PMC – *Parallel Model Combination*

RASTA – Espectro relativo (RelAtive SpecTrA)

RV – Reconhecimento de Voz

SNR – Relação sinal/ruído (*Signal-to-noise ratio*)

SRL – Sistema de Reconhecimento de Locutor

SS – Subtração espectral (*Spectral Subtraction*)

SVL – Sistema de Verificação de Locutor

$\mathbf{T_{EER}}$ – Limiar de decisão considerado *a posteriori* (*Threshold Equal Error Rate*)

VL – Verificação de Locutor

INTRODUÇÃO

Hoje em dia muitos dos sistemas de controle de acesso os quais restringem o acesso a serviços ou lugares protegidos utilizam senhas alfanuméricas ou números de identificação pessoais para a verificação de identidade. No entanto, a utilização de características físicas do cliente, chamadas características biométricas, pode melhorar o nível de segurança e a praticidade destes sistemas. Este trabalho visa contribuir para o aperfeiçoamento de sistemas de segurança que usam a voz para verificação de identidade. Estes sistemas são denominados Sistemas de Verificação de locutor (SVL) e consistem na utilização das características do sinal de voz para decidir se o locutor é quem afirma ser. Aplicações de verificação de locutor estão diretamente relacionadas com sistemas de segurança para vedar acesso de pessoas não autorizadas a lugares ou serviços protegidos.

Em SVL é preciso extrair parâmetros que representem as características de cada locutor ao contrário de reconhecimento de voz cuja finalidade é identificar a identidade fonética do sinal, de preferência independentemente do locutor. Contudo, a semelhança entre os dois sistemas são muitas e o sinal de voz do locutor, que alega uma dada identidade, é processado da mesma maneira que em reconhecimento de voz. No entanto, em verificação de locutor a verossimilhança (ou distância global) obtida no reconhecedor de padrões é utilizada para aceitar ou rejeitar o locutor, conforme um limiar de decisão previamente estabelecido.

Os problemas enfrentados pelos SVL em aplicações reais estão relacionados com o fato de que o limiar de aceitação/rejeição precisaria ser um valor fixo e predeterminado, mas é de fato função das condições de ruído (aditivo e convolucional) e do próprio locutor cuja voz pode sofrer alterações de acordo com o estado de saúde ou com a idade. Sendo

assim, robustez frente a condições adversas e adaptação às características do locutor são os principais empecilhos enfrentados pelos sistemas de verificação de locutor em aplicações reais. O tema central deste trabalho diz respeito à robustez frente a ruído aditivo e/ou ruído convolucional de Sistemas de Verificação de Locutor baseados em topologias convencionais de HMM.

Os ruídos aditivo e convolucional são de natureza diferentes: o primeiro corresponde a uma componente aditiva no domínio do tempo e no domínio linear da frequência, e pode variar acentuadamente ao longo do tempo. Por outro lado, o ruído convolucional pode ser modelado como um termo aditivo no domínio logarítmico da frequência e é geralmente invariante ao longo do tempo. Se as elocuções ou modelos (HMM's) de referência são obtidos utilizando sinal de voz limpo, a interferência no sinal de teste resulta num ruído no domínio da verossimilhança (HMM), e o desempenho dos sistemas de reconhecimento de voz e verificação de locutor é fortemente degradado. Várias técnicas têm sido propostas para melhorar a robustez dos sistemas de reconhecimento de voz, sendo que muitas delas também podem ser utilizadas com eficácia em verificação de locutor. Para ruído aditivo podemos mencionar subtração espectral (SS), normalização da SNR, *Parallel Model Combination* (PMC), J-RASTA e modelamento de duração de estados (MDE). O problema de ruído convolucional (distorção do canal transmissor e/ou microfone) é geralmente abordado utilizando filtragem *Log*-RASTA e normalização da média cepstral (CMN).

O procedimento experimental consistiu-se basicamente em gerar um HMM para cada locutor da base de dados, utilizar o HMM para calcular a verossimilhança das elocuções, e decidir se as elocuções pertencem ou não ao mesmo locutor. Isto foi feito com sinais corrompidos por ruído aditivo e/ou convolucional verificando e comparando-se os efeitos das técnicas de cancelamento de ruído na redução da taxa de erro nos sistemas de verificação de locutor. As técnicas empregadas foram CMN, *Log*-RASTA, SS e modelamento de duração de estados (MDE).

No capítulo 1 será feita uma breve revisão sobre verificação automática de locutor, considerando aspectos como medidas de performance, fatores que influenciam no desempenho da tarefa de verificação de locutor e técnicas de reconhecimento de padrões

geralmente empregadas em verificação de locutor. Ao final deste capítulo, será realizado um apanhado geral do estado da arte. As etapas que compõem um sistema de verificação de locutor serão descritas no capítulo 2. Todas as técnicas de processamento do sinal de voz serão detalhadas desde a transdução até a extração de parâmetros. Será feita uma discussão sobre as características da voz que melhor representam o locutor e sobre os métodos de escolha de parâmetros que melhor representam o sinal de voz. Também serão descritas as técnicas de modelamento estatístico do sinal de voz empregando reconhecedores de padrões com Modelos Ocultos de Markov (HMM). Estes serão descritos desde o seu treinamento, utilizando o algoritmo de Baum-Welch ou Viterbi, até o cômputo da verossimilhança final realizado pelo algoritmo de Viterbi. Ao final serão descritas as técnicas de normalização da verossimilhança.

No capítulo 3 será descrito o sistema de verificação de locutor construído neste trabalho e seu desempenho utilizando elocuções sem ruído da base da dados YOHO. No quarto capítulo terão destaques as técnicas de robustez a ruído e uma discussão sobre o efeito da presença do ruído aditivo e/ou convolucional em sistemas de verificação de locutor. As técnicas descritas serão CMN, *Log*-RASTA e J-RASTA, SS, a normalização da SNR, *Parallel Model Combination* (PMC) e modelamento de duração de estados com restrições temporais (MDE). No quinto capítulo serão realizados testes envolvendo sinal de voz corrompido por ruído aditivo e/ou convolucional no sistema de verificação de locutor com e sem técnicas de robustez. As técnicas de robustez testadas serão a CMN, o *Log*-RASTA, a SS e modelamento de duração de estados com restrições temporais. Por último, serão apresentas as conclusões do trabalho e propostas para a continuidade da pesquisa.

Capítulo 1

Uma Breve Revisão sobre Verificação Automática de Locutor

Atualmente, muitos dos sistemas de controle de acesso à transações bancárias, redes de computadores ou lugares protegidos, utilizam senhas alfanuméricas ou números de identificação pessoais para verificação de identidade pessoal. Entretanto, pessoas não autorizadas podem obter estas senhas ou números e usá-los sem o consentimento do cliente cadastrado. Além disso, senhas ou números podem ser esquecidos ou perdidos.

Sistemas biométricos de identificação, isto é, sistemas que munem-se de características físicas para verificação de identidade pessoal, são sistemas mais seguros que sistemas com senhas ou números (Frischholz & Dieckmann, 2000). Características biométricas são naturais e não podem ser emprestadas, roubadas ou perdidas. Dentre as mais utilizadas em sistemas biométricos podemos citar: íris, face, impressão digital, movimentos labiais, voz e assinatura (estática ou dinâmica). Uma característica simples, entretanto, falha algumas vezes por não ser precisa o suficiente. Assim, alguns sistemas utilizam-se de mais de uma característica e/ou senhas pessoais de forma a fornecer uma precisão muito maior do que com uma única característica biométrica. Como exemplo,

podemos citar o BioID, desenvolvido pela *Dialog Communication Systems*, que utiliza imagem facial, voz e movimentos labiais para identificar pessoas (Frischholz & Dieckmann, 2000).

Uma ou mais características biométricas podem ser armazenadas em um *Smart Card* (como de um cartão de crédito), por exemplo, e toda vez que uma pessoa não autorizada for utilizá-lo as características biométricas do impostor não serão compatíveis com as do cartão e o sistema rejeitará a transação.

Uma das características biométricas mais usuais para transações através da linha telefônica é a voz. Ao contrário de outras características biométricas, voz é fisicamente intrusiva (não é visualizada) e pode ser integrada com outros *softwares*, como de reconhecimento de fala. Na identificação vocal características acústicas da voz e estilo do locutor são utilizados para associar uma identidade ao locutor desconhecido e/ou verificar se a pessoa é quem afirma ser. Ao contrário de reconhecimento de voz, nenhuma das tecnologias está preocupada em verificar o que a pessoa disse (conteúdo fonético-lingüístico) mas apenas com quem está falando (conteúdo intra-locutor).

O desempenho de sistemas de identificação vocal é fortemente afetado pela presença de ruído, tanto do canal telefônico quanto ambiental. Esta tese tem por objetivo o estudo de técnicas direcionadas a melhorar a robustez dos sistemas de verificação de locutor ao descasamento entre as condições de treinamento no que diz respeito ao ruído.

1.1 DEFINIÇÕES

Reconhecimento de locutor é o processo de reconhecer automaticamente quem está falando com base nas informações individuais incluídas no sinal de voz (Doddington, 1985; Furui, 1986, 1989, 1991a,b, 1994, 1997; O'Shaugnessy, 1986; Rosenberg & Soong, 1991). Esta técnica torna possível a verificação de identidade para controle de acesso a serviços, informações e lugares protegidos.

Reconhecimento de locutor pode ser dividido em *Identificação de Locutor (IL)* e *Verificação de Locutor (VL)*. Identificação de locutor é o processo de determinar de quem

é, dentre os clientes registrados, uma dada elocução. Verificação de locutor é o processo de aceitar ou rejeitar a identidade alegada por uma pessoa a qual pronunciou uma ou várias elocuções. Em identificação, o número de alternativas de decisão é igual ao tamanho da população. Já na verificação existem duas decisões possíveis, aceitar ou rejeitar, independentemente do tamanho da população.

A performance de sistemas de IL diminui com o aumento no tamanho da população, enquanto que a performance de sistemas de VL se mantém mais estável e, consequentemente, é menos dependente do tamanho da população, exceto quando a distribuição das características físicas do locutor é extremamente tendenciosa (Furui, 1997).

Sistemas de reconhecimento de locutor (SRL) podem ser ainda divididos quanto ao texto em *dependente de texto*, *independente de texto* e *texto induzido*.

- *Dependente de texto:* Requer que o locutor forneça elocuções de sentenças ou palavras chave com o mesmo texto para ambos treinamento e reconhecimento.
- *Independente de texto:* Não requer que um texto específico seja falado.
- *Texto induzido:* O sistema requisita ao usuário, durante o reconhecimento, que este fale um texto específico diferente daquele do treinamento. Tanto características intra-locutor quanto o conteúdo lingüístico da voz são analisados.

Tarefas de SRL independentes de texto aplicam-se em condições onde não há cooperação do cliente; onde o locutor é reconhecido no curso de uma conversação normal; e onde o cliente pode cooperar mas é indesejado forçar o locutor a ser claro ou consistente, ou colocar alguma restrição ao seu vocabulário.

Tarefas de SRL dependente de texto geralmente envolvem alguma palavra pré-determinada ou senha induzida de forma a obter-se o texto requerido. É esperado do cliente cooperação, clareza e consistência na elocução para ser corretamente aceito pelo sistema.

Ambos os métodos, dependente e independente de texto, podem ser burlados por alguém que reproduza o sinal de voz de um locutor registrado, gravado durante a elocução da palavra chave ou sentença. Para enfrentar este problema existem métodos nos quais um pequeno número de palavras pré-determinadas, como dígitos numéricos, são usadas como palavras chaves e cada usuário é induzido a falar uma dada seqüência de palavras chave

que são aleatoriamente escolhidas cada vez que o sistema é usado. Não sendo este método confiável o suficiente, visto que ele poderia ser derrotado com equipamentos de gravação eletrônica avançados que podem reproduzir palavras chave na ordem desejada, métodos de reconhecimento de locutor com *texto induzido* tem sido propostos recentemente.

No método de reconhecimento de locutor empregando-se texto induzido, sentenças chaves são mudadas todas as vezes. O sistema aceita a elocução de entrada somente quando é determinado que o locutor registrado falou a elocução induzida. Pelo fato de ter um vocabulário ilimitado, impostores não podem saber com antecedência a sentença a ser induzida. Este método não pode somente reconhecer locutores precisamente, mas também rejeitar uma elocução enquanto o texto difere do texto induzido se isto é dito por um locutor registrado. Assim, uma reprodução de uma voz pré-gravada pode ser corretamente rejeitada.

1.2 VERIFICAÇÃO DE LOCUTOR

Em VL, um locutor desconhecido que alega uma determinada identidade tem sua elocução comparada com o modelo do locutor verdadeiro da afirmada identidade. Se a distância é o suficiente, acima de um limiar de decisão, o locutor é aceito tendo a identidade confirmada. Um limiar de decisão alto dificulta a aceitação de impostores mas aumenta o risco de falsa rejeição de usuários válidos. Por outro lado, um limiar baixo possibilita a usuários válidos serem facilmente aceitos, mas aumenta-se o ricos de aceitação de impostores. Este limiar de decisão é determinado de acordo com as exigências de sua aplicação.

Existem dois tipos de classificações corretas: a aceitação de clientes e a rejeição de impostores. Por outro lado, temos dois tipos de erros: a *falsa rejeição (FR)* (rejeição de locutores genuínos) e *falsa aceitação (FA)* (aceitação de impostores). A decisão é tomada comparando-se um limiar de decisão com uma medida de *semelhança*, que pode ser uma distância global (DTW) ou uma verossimilhança (HMM), entre o modelo do cliente e a elocução de teste.

1.2.1 Medidas de Performance

Várias medidas de performance são usadas:

- Taxa de falsa rejeição nula: é a taxa de erro de FA considerando-se o maior limiar de decisão possível para o qual nenhum locutor genuíno é rejeitado;
- Taxa de falsa aceitação nula: é a taxa de erro de FR considerando-se o menor limiar de decisão possível para o qual nenhum impostor é aceito;
- Taxa de erros iguais (EER – *Equal Error Rate*): é a taxa de erro onde os percentuais de FR e FA são iguais, no qual aplica-se um limiar de decisão *a posteriori* T_{EER}. Este limiar é dito *a posteriori* por ser o limiar correspondente ao ponto de interseção das curvas de FA e FR. escolhido após a realização dos testes. Em implementações reais, o limiar de decisão deve ser escolhido *a priori*, tendo em vista que necessita-se de seu valor para a tomada de decisão.

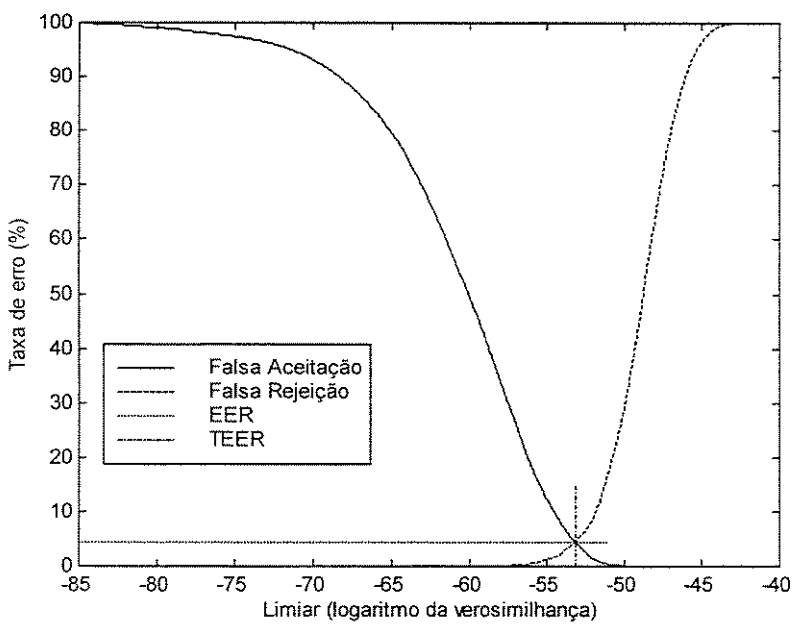

Figura 1.1 - Determinação da taxa de erros iguais EER e do limiar T_{EER}

EER é a medida de performance mais comumente utilizada e pode ser classificada como sendo de *locutor específico* (EER_{SS}) (calculada para cada locutor separadamente, com T_{EER} individuais) e *independente de locutor* (EER_{SI}) (o mesmo limiar T_{EER} é usado para

todos os locutores). A Fig. 1.1 mostra os valores de EER e T_{EER} para um determinado locutor.

Outro gráfico de performance freqüentemente usado em VL é a característica de operação do receptor (ROC – *Receiver Operating Characteristic*) (veja Fig. 1.2). Consiste em um gráfico da taxa de aceitação correta em relação a taxa de falsa rejeição. O ROC é um recurso importante para a escolha de um limiar a posteriori adequado a uma determinada aplicação, assim como as taxas de falsa aceitação e rejeição nulas.

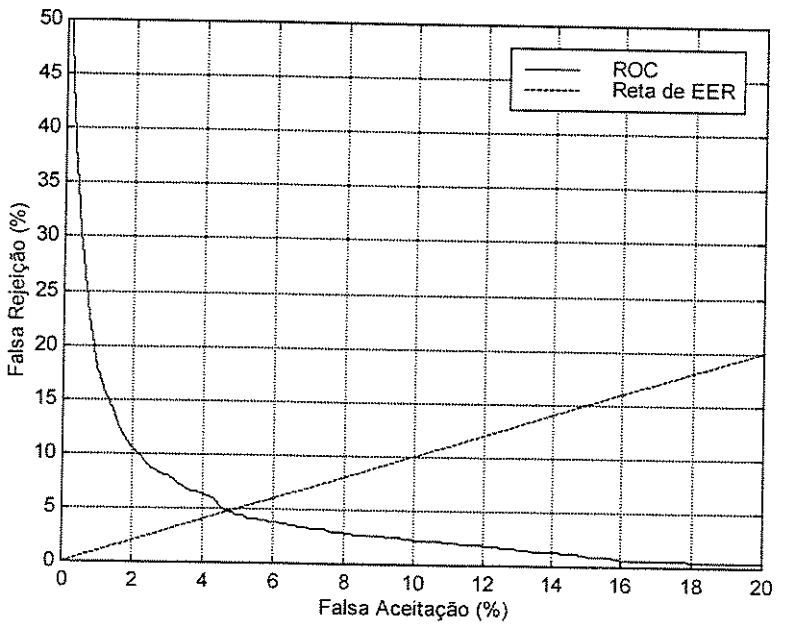

Figura 1.2 - Característica de operação do receptor (ROC)

1.2.2 Fatores que influenciam na tarefa de verificação de locutor

1.2.2.1 Variações intra e inter-locutor

A voz é uma característica biométrica que possui grande variação inter-locutor. Locutores de diferentes localidades costumam apresentar peculiaridades como sotaque e entoações distintas, o que influencia favoravelmente na precisão da verificação de identidade. Já a secreção provocada por uma gripe modifica o volume do trato vocal e pode modificar as características vocais do locutor. Fatores como estado de saúde, idade e o humor variam indesejavelmente as características intra-locutor empobrecendo o desempenho do sistema.

A presença de clientes de ambos os sexos no sistema melhora o desempenho. Homens e mulheres têm uma grande variação inter-locutor o que dificulta a ocorrência de erros de cruzamento de sexo. Soong & Rosenberg (1988) obtiveram entre 9,4% e 26,4% de erros entre cruzamento de sexos dependendo da configuração do sistema.

De forma a melhor modelar as variações temporais das características de cada locutor, muitas sessões espaçadas acima de um período de semanas ou meses deveriam ser usadas. É provável, entretanto, que mais do que uma sessão de treinamento poderia ser esperada em uma aplicação de banco por telefone mas certamente não muito mais do que duas ou três. Clientes estarão registrando-se porque eles querem usar o sistema – eles não querem esperar vários dias para isto ficar disponível.

1.2.2.2 Dados de treinamento

Como cada locutor é representado por um modelo, quanto mais completo for este modelo melhor se dará a verificação de locutor. Quanto maior for a quantidade de dados de treinamento disponível melhor será o modelo do locutor. Entretanto, as seções de treinamento costumam ser extensas e demoradas o que as pode tornar pouco viáveis em aplicações práticas. Em outras palavras, o montante de dados disponível para o treinamento do modelo do cliente é fortemente limitado ao que o cliente acha aceitável.

Por outro lado, a base de dados de treinamento disponível também depende de cada tarefa. Em aplicações de controle de acesso a áreas físicas, um cliente provavelmente estaria disposto a investir tempo significante e esforços para assegurar facilidade de uso e alto nível de segurança. Para aplicações como banco por telefone onde um cliente está substituindo um serviço existente por outro mais conveniente, o conjunto das dificuldades associados à obtenção do novo serviço serão cruciais ao seu sucesso.

1.2.2.3 Descasamento entre condições de treinamento e de teste

A diferença entre as condições de treinamento e de teste das elocuções de um determinado locutor é um dos fatores que mais prejudica a precisão da verificação de locutor (Openshaw *et al.*, 1993). Esta diferença pode ser causada por fatores como a taxa de amostragem, a qual impõe sua própria largura de banda, o microfone utilizado, o ruído de fundo e o canal transmissor. Em primeiro lugar, a taxa de amostragem é em geral próxima

da largura de banda do sinal que se quer representar. Entretanto, quando há a inserção de um canal transmissor, a largura de banda do sinal fica restrita à largura de banda do canal.

O tipo de microfone utilizado é muito importante já que cada qual colore o espectro do sinal de voz de forma distinta (Wang *et al.*, 1993). Entretanto, em muitos caso não há a possibilidade de utilizar sempre o mesmo microfone o que exige robustez do Sistema de Verificação de Locutor (SVL) ao descasamento entre condições de treinamento e de teste.

Da mesma forma que o microfone, o canal transmissor também distorce o sinal de voz transmitido de forma distinta a cada conexão. Tendo como exemplo o canal telefônico, duas ligações para o mesmo número de telefone podem fornecer respostas em freqüência do canal diferentes assim como conexões a diferentes distâncias. Esta distorção em freqüência do sinal (multiplicação da resposta em freqüência do microfone e/ou do canal transmissor pelo espectro do sinal) é chamado de ruído convolucional, já que é representado por uma convolução no tempo.

Para obter-se um bom modelamento do locutor, o ruído de fundo do ambiente de gravação no treinamento é mantido baixo. Já em condições de teste é difícil restringir o uso do sistema a ambientes de baixo ruído. Portanto, espera-se que o SVL possa operar confiavelmente com o locutor dentro de seu escritório, carro, próximo a pessoas em conversação ou quaisquer outros ambientes. A este tipo de ruído, em que o sinal em questão é corrompido pela soma temporal com outro sinal ruidoso, dá-se o nome de ruído aditivo.

1.2.2.4 Verificação de palavras isoladas e fala contínua

Tanto reconhecimento de voz quanto de locutor podem se dar com o processamento de elocuções com palavras isoladas ou fala contínua. Na primeira forma o locutor fala uma ou mais palavras forçando um intervalo entre elas. Já em fala contínua o locutor não tem nenhuma restrição quanto ao modo de falar, como em uma conversação. A grande diferença entre estes métodos está no efeito de coarticulação que as palavras em fala contínua sofrem, ou seja, a realização acústica das palavras será influenciada pelas palavras imediatamente anterior e posterior. Em conseqüência, é necessário que as elocuções sejam pronunciadas de maneira semelhante no treinamento e no teste. Caso se utilize fala contínua

na verificação, para um desempenho satisfatório é necessário que também se tenha treinado com elocuções em fala contínua. O mesmo também é válido para palavras isoladas.

Devido a maior variabilidade das palavras em fala contínua, a verificação de locutor com este método geralmente é mais difícil que verificação utilizando-se elocuções com palavras isoladas.

1.2.2.5 Tamanho da elocução de teste

É natural pensar que quanto maior for a informação obtida do locutor mais fácil será a tarefa de verificação de identidade. Até mesmo para humanos é muito mais fácil identificar uma pessoa conhecida pelo telefone depois de vários segundos de conversação do que apenas com um "alô". Assim, quanto maior for o número de palavras da elocução de verificação, maior será a precisão do resultado. Soong & Rosenberg (1988) em seus experimentos com identificação de locutor utilizando parâmetros espectrais estáticos verificaram no caso ótimo (comprimento da janela de 7 quadros) uma taxa de erro de 15% com elocuções de verificação de 1 dígito e de 1,5% para elocuções de 10 dígitos isolados.

Pode-se aumentar o número de palavras necessárias para a verificação o quanto se queira, sem necessidade de aumentar o vocabulário excessivamente, até encontrar uma taxa de erro aceitável. O limite é o tempo máximo que o cliente poderia ou desejaria gastar durante a verificação.

1.2.2.6 Armazenagem de dados e complexidade computacional

A performance em tempo real não é preocupante visto que para a tarefa de verificação de locutor não há a necessidade de associação a requisitos gramaticais como em reconhecimento de voz. Requisitos para armazenagem de dados são preocupantes quando:

- Os modelos de clientes devem ser portáveis, como um sistema usando cartões magnéticos ou *smart cards*, onde o modelo do cliente é codificado no cartão;
- A população de clientes é muito grande.

1.2.3 Conceitos Importantes

Na literatura de verificação de locutor utilizam-se termos para se caracterizar as pessoas envolvidas no processo. A seguir, são citados alguns deles:

- **Locutor:** Pessoa que requisita liberação de acesso utilizando-se da análise do seu sinal de voz.
- **Cliente:** Locutor registrado no sistema de verificação de locutor.
- **Impostor:** Locutor que tenta burlar o sistema tentando se passar por outra pessoa (cliente).
- **Bode (*goat*):** Locutor que é freqüentemente falsamente rejeitado pelo sistema (alta taxa de falsa rejeição).
- **Carneiro (*sheep*):** Locutor que não tem dificuldade para ser corretamente aceito pelo sistema (baixa taxa de falsa rejeição).
- **Cordeiro (*lamb*) (um carneiro jovem):** Locutor que é fácil de ser imitado (sua identidade tem uma alta taxa de falsa aceitação). Tem muita variabilidade na voz e/ou características de voz muito comuns.
- **Carneiro macho (*ram*):** Locutor dificilmente imitado por impostores (sua identidade tem uma pequena taxa de falsa aceitação). Possuem vozes distintas e/ou consistentes.
- **Lobo (*wolf*):** Locutor que tem facilidade em burlar o sistema e se passar por outro locutor (alta taxa de falsa aceitação de locutores verdadeiros).
- **Texugo (*badger*):** Locutor que tem dificuldade em se passar por outro locutor (pequena taxa de falsa aceitação de locutores verdadeiros).
- **Impostores casuais:** Impostores que estão tentando enganar o sistema e não têm conhecimento dos clientes que afirmam ser e podem ser assumidos que eles falem em sua voz usual.
- **Impostores dedicados:** O impostor tem conhecimento da voz do locutor da afirmada identidade e usa este conhecimento para tentar burlar o sistema.

1.3 RECONHECIMENTO DE PADRÕES ACÚSTICOS

Na tarefa de verificação de locutor, o locutor que deseja ter o acesso liberado pronuncia uma elocução de teste. Conforme o diagrama de bloco da Fig. 1.3, características que melhor representam o locutor são extraídas desta elocução de teste. Assim, a elocução fica representada por uma seqüência no tempo de n vetores de parâmetros $\tau = \{t_1, t_2, \ldots, t_n\}$. Para que uma decisão lógica seja tomada (aceitação ou rejeição do locutor) esta seqüência de vetores de parâmetros deve ser comparada com o modelo (padrão de referência) do locutor através de um classificador de padrões. O modelo do locutor é obtido previamente em seções de treinamento empregando m elocuções distintas das de teste $\{\tau_1, \tau_2, \ldots, \tau_m\}$. Ele pode ser tanto uma seqüência de vetores de parâmetros ótima (determinística) quanto um modelo estocástico. No classificador de padrões acústicos é obtida uma medida de similaridade entre o modelo do locutor hipotético e a seqüência τ do locutor de teste. Esta medida pode ser tanto uma distância global quanto um valor de verossimilhança. Três técnicas básicas de classificação de padrões são comumente utilizadas: o *Dynamic Time Warping* (DTW) (Sakoe & Chiba, 1978), os Modelos Ocultos de Markov (HMM – *Hidden Markov Models*) (Baker, 1975a, 1975b; Jelinek *et al.*, 1975) e, mais recentemente, as Redes Neurais (Haykin, 1994). Todos os SVL existentes trabalham com uma ou mais destas técnicas sendo que a maioria utiliza os HMM.

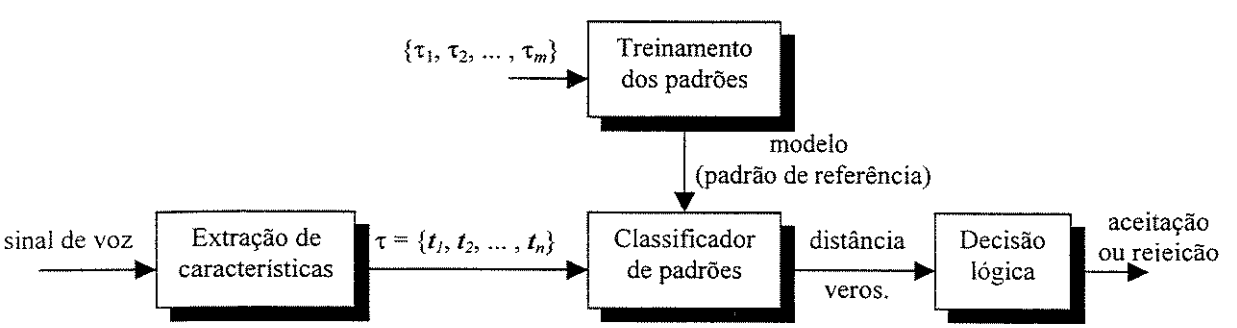

Figura 1.3- Diagrama de bloco do reconhecedor de padrões acústicos.

O DTW foi introduzido para contornar o problema causado pelas variações na velocidade de pronunciação das palavras de um forma determinística. Ou seja, a realização acústica de uma mesma elocução por um mesmo locutor (ainda mais em locutores diferentes) pode variar significativamente modificando a velocidade de articulação do

aparelho fonador. Assim, uma palavra pode ser pronunciada mais ou menos rapidamente, encurtando ou alongando os períodos estacionários do sinal de voz, enquanto que os períodos não estacionários se mantêm quase sempre constantes. O DTW tem a capacidade de alinhar seqüências de vetores de parâmetros com diferentes durações e assim obter um valor de distância global (Furui, 1981).

Os HMM constituem o modelamento estatístico da elocução através de uma seqüência de estados formando uma cadeia Markoviana. Este modelamento estatístico das seqüências de vetores de parâmetros é o grande mérito dos HMM pois fornece uma boa representação da evolução temporal da fala (Rabiner & Juang, 1993). Por outro lado, as redes neurais têm facilidade em assimilar superfícies de decisão formadas por estatísticas complexas mas têm dificuldades em lidar com processo variantes no tempo (Bennani & Gallinari, 1995; Morgan & Franco, 1997). Sistemas híbridos com HMM e redes neurais (Olsen, 1997; Albesano *et al*, 2000) têm fornecidos bons resultados, permitindo, em alguns casos, a combinação das melhores características de ambas arquiteturas.

1.4 ESTADO DA ARTE EM VERIFICAÇÃO DE LOCUTOR

Atualmente, a tarefa de verificação de locutor dependente de texto fornece um melhor desempenho que tarefas independentes de texto e de texto induzido. Contudo, vários trabalhos vêm sendo publicados com novas técnicas que melhoram gradativamente o desempenho destas últimas (Matsui & Furui, 1994a e 1994b). Tarefas de texto induzido são as mais adequadas para verificação de identidade vocal já que são praticamente invulneráveis à fraude eletrônica e constituem uma importante e promissora área de pesquisa.

Para a classificação de padrões, as técnicas que fornecem melhor desempenho são as que empregam HMM individualmente (por exemplo: Bimbot *et al.*, 1997; e Markov & Nakagawa, 1998). e as que formam modelos híbridos com redes neurais (como Olsen, 1997). Embora os modelos híbridos não tenham se mostrado significativamente mais eficientes que os HMM individuais, eles reúnem as vantagens individuais de ambas as técnicas e são potencialmente mais promissores.

O modelamento dos HMM é feito tanto para palavras quanto subpalavras (difones, trifones, polifones), sendo estas últimas as mais utilizadas em vocabulários de médio e grande porte. A topologia em esquerda-direita é utilizada quase que exclusivamente. Cada estado de cada modelo geralmente é constituído por uma ou múltiplas gaussianas dependendo da aplicação, sendo que somas de gaussianas fornecem um melhor desempenho quando há dados de treinamento suficientes.

Os parâmetros utilizados são quase que exclusivamente a energia e os coeficientes cepstrais e suas derivadas (delta-cepstrais). A seleção dos parâmetros que melhor representam o locutor e/ou a inclusão de pesos para o cálculo da verossimilhança são formas simples de melhorar o desempenho dos Sistemas de Verificação de Locutor (SVL) sem aumentar excessivamente a complexidade computacional.

Verificação de locutor pode ser utilizada em conjunto com outras técnicas de verificação de identidade por biometria, como movimentos labiais (Jourlin *et al.*, 1997; Wark *et al.*, 1998), de forma a aumentar a precisão na verificação. Um exemplo comercial deste tipo de sistema de verificação de identidade é o já citado *BioID* que utiliza imagem facial, voz e movimentos labiais para identificar pessoas (Frischholz & Dieckmann, 2000). Contudo, há casos como em transações pela rede telefônica onde se pode utilizar apenas o sinal de áudio para a verificação de identidade. Um exemplo comercial deste tipo de SVL é o *Nuance Verifier*, desenvolvido pela *Nuance Communications*.

O ruído convolucional tem seus efeitos bastante atenuados com a aplicação de técnicas como a normalização da média cepstral (CMN) e RASTA. Entretanto, o maior empecilho, e o que demanda maior esforço de pesquisa, é o problema do ruído aditivo em relações sinal/ruído (SNR – *Signal-to-noise ratio*) moderadas e baixas que ainda deteriora sensivelmente o desempenho de sistemas de verificação de locutor em muitas aplicações reais.

A taxa de erros iguais (EER) para SVL no estado da arte chega a valores inferiores a 1%, como em (Bimbot *et al.*, 1997; Markov & Nakagawa, 1998). Atualmente, boa parte dos SVL são dependentes de texto devido a sua maior confiabilidade. No entanto, a crescente melhora na taxa de acerto em pesquisas com texto induzido provavelmente leve a um maior emprego desta técnica.

No capítulo 2 serão descritas as etapas que compõem um Sistema de Verificação de Locutor, assim como uma discussão sobre as características que melhor representam o locutor. Também serão apresentados os algoritmos para modelamento estatístico do locutor (HMM) e a normalização da verossimilhança. Os detalhes da implementação e os resultados obtidos com o SVL implementado serão descritos no capítulo 3.

Capítulo 2

Sistema Automático de Verificação de Locutor (SVL)

2.1 INTRODUÇÃO

A Fig. 2.1 indica o fluxograma de um SVL típico. Existem duas entradas no sistema: a elocução de teste e a afirmada identidade. O fornecimento desta é necessário para o acesso aos dados de referência do locutor verdadeiro que incluem modelo e o limiar de decisão; esta identidade pode ser fornecida por um código, ou até mesmo falada para ser registrada por um atendente ou um sistema de reconhecimento de voz. A elocução de teste é requisitada assim que o locutor manifestar seu desejo de liberação de acesso. Após a aquisição do sinal, a elocução sofre um pré-processamento e é dividida em intervalos iguais chamados quadros ou *frames* dos quais se faz uma análise espectral para uma posterior extração de parâmetros. Estes parâmetros são comparados com o modelo do locutor verdadeiro através do cálculo de uma medida de distância global ou verossimilhança. Esta medida de distância ou verossimilhança é comparada com um limiar de aceitação e

representa a semelhança entre a elocução de teste e o modelo de referência da elocução pronunciada pelo locutor verdadeiro (para o caso dependente de texto). Se a distância for menor (ou a verossimilhança for maior) do que o limiar o locutor é aceito como tendo a alegada identidade e tem liberação de acesso. Caso contrário, o locutor é rejeitado.

O modelamento pode ser feito por várias técnicas, como HMM (Baker, 1975a, 1975b; Jelinek *et al.*, 1975), DTW (Sakoe & Chiba, 1978) e Redes Neurais (Haykin, 1994) e combinações destes. No caso específico deste trabalho utilizou-se HMM, sendo que a verossimilhança foi computada por um algoritmo recursivo chamado Algoritmo de Viterbi.

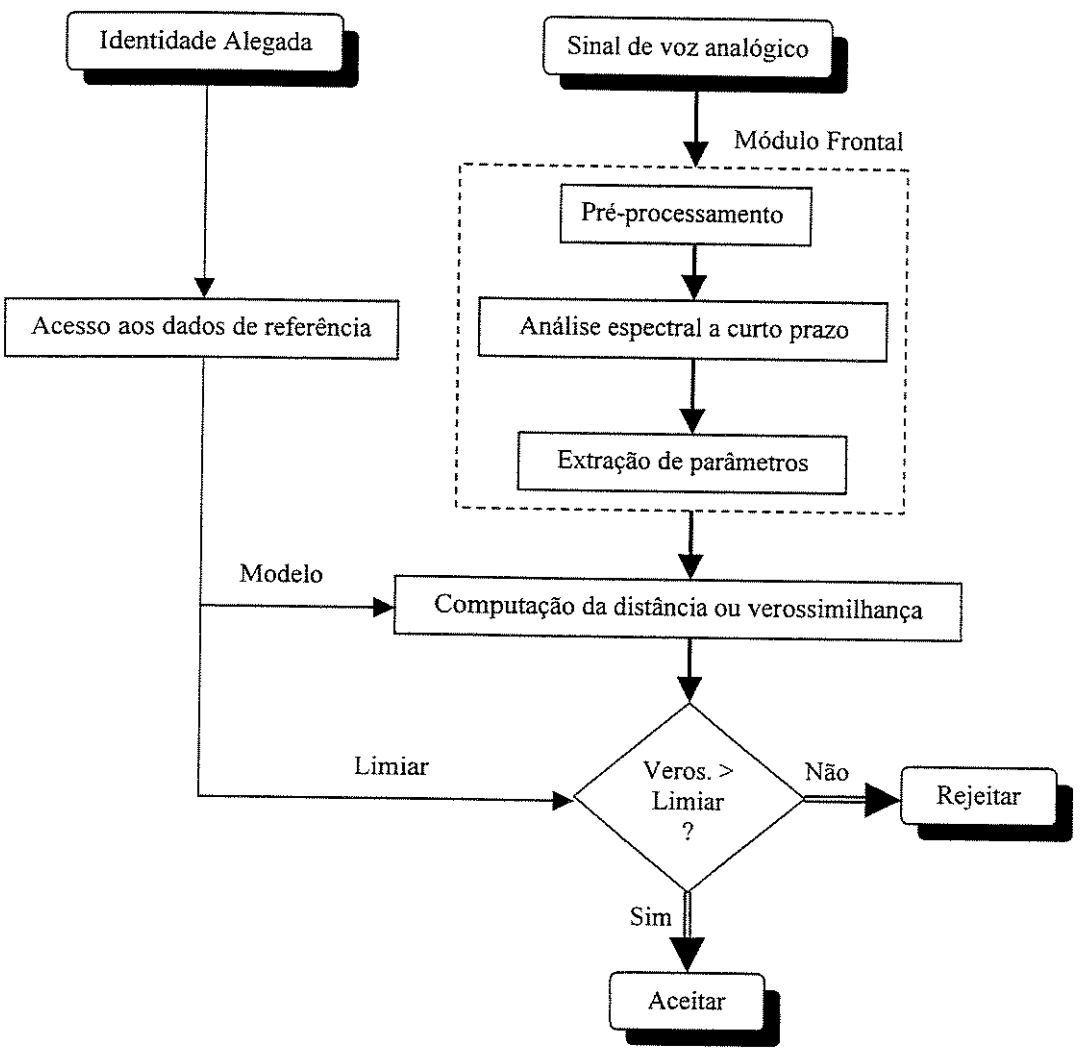

Figura 2.1 - Fluxograma de um SVL típico

2.2 PRÉ-PROCESSAMENTO

Na etapa de pré-processamento, o sinal de voz analógico é digitalizado, a elocução tem seu início e final detectados e seu sinal é pré-enfatizado, conforme Fig. 2.2.

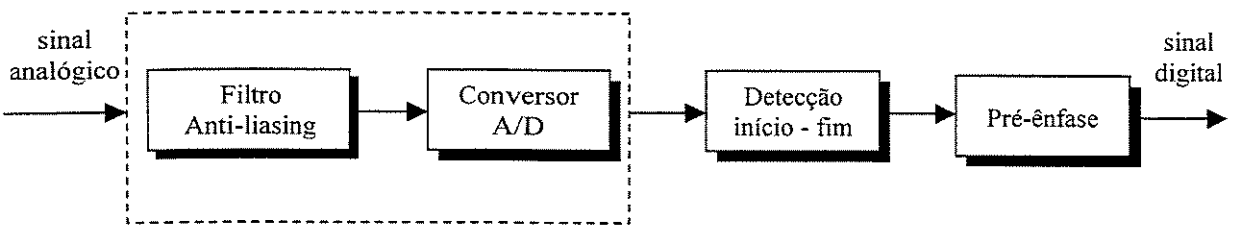

Figura 2.2 - Etapas do pré-processamento

2.2.1 Conversão A/D

O sinal analógico proveniente de um transdutor (microfone) precisa ser filtrado por um filtro passa-baixa de freqüência de corte igual à metade da freqüência de amostragem utilizada de forma a evitar o fenômeno conhecido como *aliasing* no sinal digital resultante. Em seguida o sinal passa por um conversor A/D o qual amostra e quantiza o sinal analógico transformando-o em um sinal digital.

2.2.2 Detecção de início e final

As partes de silêncio inicial e final não trazem nenhuma informação relevante para o modelamento do locutor. Além disso, sua incorporação aos modelos de palavras pode ser prejudicial, sendo necessário então a detecção e eliminação destes intervalos de silêncio.

Geralmente os detectores de início e final empregam os seguintes parâmetros: energia, estimação espectral e restrições temporais como em (Lamel *et al.*, 1981); cruzamento por zero ou por nível (Savoji, 1989); ou freqüência fundamental (Mak *et al.*, 1992). Nesta dissertação a detecção de início-fim utilizou apenas energia e restrições temporais o que mostrou bons resultados em situações com uma relação SNR alta.

Cada quadro da elocução digitalizada tem sua energia computada e comparada com limiares de energia. Para detectar-se o início, a energia deve ser maior que o limiar l_1 por

pelo menos dois quadros consecutivos. Analogamente, para detectar-se o final da elocução a energia deve ser menor que o limiar l_2 por pelo menos dois quadros consecutivos. Como a velocidade de coarticulação do trato vocal impede a produção de pulsos de voz com duração menor que 75 a 100ms, os pulsos de voz iniciais e finais são validados caso tenham a duração de pelo menos três quadros e pelo menos um deles tenha energia maior que o limiar l_{max}. Isto impede que pequenos ruídos indesejados, tais como "cliques", provenientes tanto do ambiente de gravação quanto do microfone, passem por um pulso de voz. Cabe destacar que a detecção do quadro final é feita para todos os pulsos de voz e o final da elocução será o quadro detectado no último pulso válido. Já o início é detectado apenas uma vez sendo igual ao quadro inicial do primeiro pulso de voz válido

Os limiares l_1, l_2 e l_{max} foram estimados a partir da máxima e mínima energias de quadro de cada elocução. Isto permite que o detector de início e final funcione adequadamente até uma SNR de mais ou menos 12dB sem qualquer ajuste. As Eq. 2.1 descrevem o cálculo dos limiares.

$$\begin{aligned} l_1 &= 10\log[E_{\min} + \alpha(E_{\max} - E_{\min})] \\ l_2 &= 10\log[E_{\min} + \beta(E_{\max} - E_{\min})] \\ l_{\max} &= 10\log[E_{\min} + \gamma(E_{\max} - E_{\min})] \end{aligned} \tag{2.1}$$

onde os valores α, β e γ são constantes obtidas empiricamente.

2.2.3 Pré-ênfase

O sinal de voz naturalmente tem um decaimento espectral de aproximadamente 20 dB por década devido ao efeito combinado dos pulsos glotais e da irradiação pelos lábios do locutor (Deller *et al.*, 1993). De forma a tentar compensar esta queda, o sinal é filtrado por um filtro de resposta ao impulso finita (FIR) de primeira ordem conhecido como filtro de pré-ênfase. A função de transferência $H(z)$ do filtro é dada pela Eq. 2.2.

$$H(z) = 1 - \mu Z^{-1} \tag{2.2}$$

A filtragem de pré-ênfase também reduz a instabilidade computacional associada a aritmética de precisão finita (Markel & Gray, 1980). O valor μ é o coeficiente de pré-ênfase

e tem valores ente 1 a 0,9. Para os experimentos desta dissertação manteve-se o valor de μ constante em 0,97.

2.3 ANÁLISE ESPECTRAL A CURTO PRAZO

A Fig. 2.3 representa o diagrama de blocos da análise espectral a curto termo. A elocução é dividida em quadros que posteriormente são janelados para a realização da análise espectral propriamente dita.

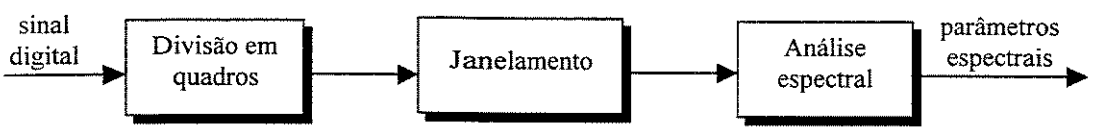

Figura 2.3 - Etapas da análise espectral a curto prazo

2.3.1 Divisão em quadros

Para a análise espectral a curto prazo divide-se a elocução pré-processada em pequenos intervalos iguais chamados quadros ou *frames*. A duração de cada quadro é escolhida para estar entre 10 a 30ms, intervalo no qual o sinal pode ser considerado estacionário. Outro aspecto que influencia na escolha da duração do quadro é o compromisso entre a resolução no tempo e a resolução na freqüência: quanto maior for a duração do quadro maior será a resolução em freqüência e menor será a resolução no tempo, e vice versa. Após a divisão, os quadros são superpostos de forma a permitir uma transição mais suave dos parâmetros extraídos. Valores de superposição variam de 0 a 70%.

2.3.2 Janelamento

Deve-se janelar o sinal de voz de forma a minimizar as transições abruptas nos extremos do sinal segmentado e concentrar a análise do sinal no centro do quadro. O janelamento consiste em multiplicar no tempo o sinal do quadro $x(n)$ por uma função chamada janela $w(n)$, a qual geralmente possui bordas suaves, conforme a seguinte equação:

(2.3)

$$\tilde{x}(n) = x(n).w(n)$$

onde n é o índice da amostra. Uma janela muito utilizada em processamento de voz é a de Hamming cuja função analítica é dada por:

$$w(n) = \begin{cases} 0{,}54 - 0{,}46.\cos\left(\dfrac{2\pi n}{N-1}\right) & , 0 \le n \le N-1 \\ 0 & , \text{caso contrário} \end{cases} \qquad (2.4)$$

onde N é o o tamanho da janela em amostras.

2.3.3 Análise espectral

Dois métodos de análise espectral do sinal de voz são comumente utilizados: o método de análise espectral por banco de filtros utilizando a transformada rápida de Fourier (FFT – *Fast Fourier Transform*); e o método de análise espectral por predição linear (LPC - *Linear Predictive Coding*) (Rabiner & Juang, 1993).

Yoma (1993) verificou em um sistema de reconhecimento de voz que a utilização de parâmetros mel-cepstrais obtidos através de banco de filtros é mais robusto a ruído branco gaussiano que coeficientes LPC-cepstrais obtidos por predição linear. Segundo Yoma, a maior robustez daqueles parâmetros, no que diz respeito a ruído aditivo, deve-se a utilização de filtros passa-banda enquanto que a análise LPC-cepstral utiliza polinômios preditores cujos pólos são muito susceptíveis ao ruído. Assim, quando o SVL é utilizado em ambiente ruidoso, como no caso dos experimentos desta dissertação, é preferível a utilização da análise espectral por banco de filtros.

2.3.3.1 Análise por banco de filtros

Experimentos sobre a percepção humana mostram que freqüências de um som complexo, como a voz, dentro de uma certa faixa, não podem ser percebidas individualmente (Picone, 1991). Baseado nisto, a análise do sinal de voz é feita para faixas de freqüência do seu espectro, ao invés de utilizar-se freqüências individuais. Cada faixa compõe um filtro e o conjunto de filtros é denominado *banco de filtros*.

A "teoria da posição" (Picone, 1991) afirma que a capacidade de resolução frequencial do ouvido humano não varia de forma linear com a freqüência mas de forma logarítmica. Desta maneira, escalas de freqüência baseadas na percepção auditiva humana, como a *mel* e a *Bark*, foram criadas. A *mel* é baseada na habilidade do ouvido em distinguir uma freqüência de outra adjacente, tendo em vista que para freqüências elevadas a capacidade de resolução do ouvido é reduzida, e vice-versa. A escala mel pode ser descrita pela Eq. 2.5 e ilustrada na Fig. 2.4, onde f representa a freqüência acústica.

$$mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{2.5}$$

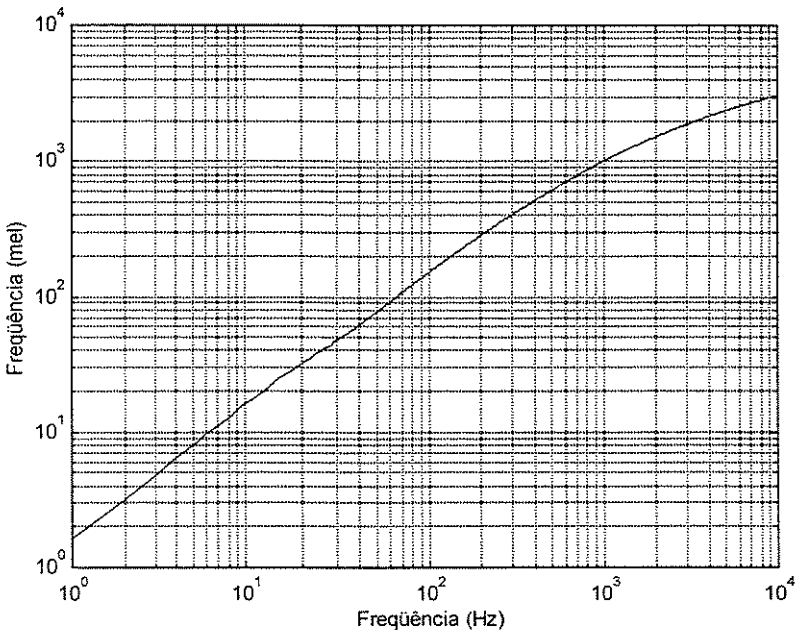

Figura 2.4 - Escala mel como função da freqüência acústica (representação logarítmica)

No entanto, a escala mel é freqüentemente aproximada como linear de 0 a 1000Hz e logarítmica acima de 1000Hz. Como a escala mel enfatiza aspectos importantes da percepção do sinal de fala, ela é largamente utilizada na aplicação do banco de filtros. Ao invés de utilizar a escala linear, os filtros são distribuídos uniformemente ao longo da mel.

A Fig. 2.5 mostra uma das formas de implementação do banco de filtros em escala mel. Os filtros são triangulares, de igual largura de banda (em mels), com ganho unitário na freqüência central, com sobreposição de 50% e dispostos sobre a faixa BW em mels. Sendo

K o número de filtros desejados do banco, a largura de banda bw (em mels) de cada filtro do banco pode ser obtida por

$$bw = \frac{2BW}{K+1} \tag{2.6}$$

com BW e bw em mels.

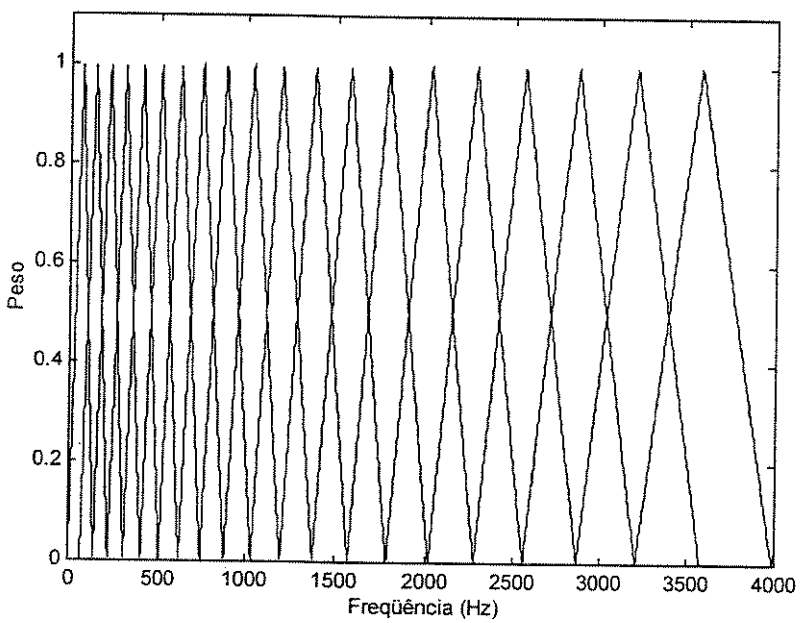

Figura 2.5 - Banco com 20 filtros dispostos em escala mel

Para uma maior eficiência computacional, geralmente utiliza-se a FFT para o cálculo do escpectro discreto. Para isso, sendo T_w a duração da janela em amostras, cada quadro janelado de duração T_w é acrescido de zeros até atingir o tamanho 2^n. Como resultado da aplicação de um banco de filtros a um determinado quadro janelado de análise, tem-se a energia de cada filtro i. O cômputo da energia E_i do filtro i é feito por

$$E_i = \sum_{j=1}^{2^n} w_{ij} e_j \quad , 1 \le i \le K \tag{2.7}$$

onde w_{ij} é o peso do ponto j do espectro de energia no filtro triangular i e e_j é a energia no ponto j da FFT.

2.4 EXTRAÇÃO DE PARÂMETROS

2.4.1 Parâmetros que melhor representam o locutor

Os humanos utilizam-se de características de "alto nível" (Naik, 1990) - tais como sotaque, estilo do locutor, entoação, estado emocional, etc - para reconhecer uma pessoa através de sua voz. Como este tipo de característica é difícil de ser adquirida e mensurada, parâmetros de "baixo nível" (Naik, 1990) derivados de medidas acústicas do sinal de voz – como freqüência fundamental, envoltória espectral, freqüência de formantes, energia, etc- (Atal, 1976; Doddington, 1985) são utilizados em SVL reais.

Para os sistemas de VL a informação que é relevante é aquela relacionada com o tamanho e a forma do aparato vocal, e a informação comportamental como sotaque e velocidade de fala do locutor. Idealmente, os parâmetros extraídos do sinal de voz devem ter pequena variação intra-locutor e grande variação inter-locutor, ou seja, ter pequena variabilidade para elocuções de mesmo conteúdo lingüistico do mesmo locutor e ser o mais distinto possível de um locutor em relação a outro. Estes parâmetros ainda devem ser de fácil extração, não devem mudar com o tempo, a saúde ou o estado emocional do locutor, não devem ser concientemente modificáveis, devem ser robustas ao canal transmissor e a ruído ambiental (Forsyth, 1995).

2.4.1.1 Freqüência Fundamental (F_0)

F_0 é a freqüência fundamental e é um aspecto da voz perceptualmente bastante saliente, o que o faz muito vulnerável a tentativas de disfarce. Foi inicialmente testado em alguns sistemas (Southerland & Jack, 1988) mas tem grandes variações intra-locutor de acordo com a saúde e o estado emocional do locutor (Rosenberg, 1976). Além disto, sua extração em ambientes com alto nível de ruído é pouco confiável . Assim, sua utilização em SVL reais fica bastante restrita.

2.4.1.2 Freqüência de Formantes

Cada trato vocal é caracterizado por uma série de freqüências de ressonância. Estas freqüências são chamadas de formantes e representam as freqüências de maior energia

acústica da fonte de voz (Rabiner & Juang, 1993). O primeiro formante é representado por F_1 e os seguintes por F_2, F_3, F_4, ...

Vários estudos têm verificado que as freqüências dos formantes de vogais produzidas por diferentes locutores apresentam uma considerável variabilidade inter-locutor, mesmo em contextos fonéticos fixos. Dentre os formantes, F_3 e F_4 são os que mais são afetados pela forma e volume do trato vocal, sendo assim características com boa distintividade (Figueiredo, 1994). Contudo, o principal empecilho para a utilização dos formantes em SVL reais é a dificuldade de sua estimação, principalmente em ambientes ruidosos.

2.4.1.3 Energia

É natural de se imaginar que a medida de energia de uma elocução contém informações importantes sobre a identidade fonética dos sons e sobre o locutor. Sons fricativos como o [*s*] têm menos energia que sons de vogais como o [*a*] e cada som é, ainda, de energia distinta para cada locutor. Contudo, para que a energia seja útil na tarefa de verificação de locutor, ela deve ser apropriadamente normalizada para cada elocução (Rabiner & Juang, 1993).

A energia é bastante empregada como parâmetro para VL, como em (Bimbot *et al.*, 1997). Em VL geralmente é computada a energia total por quadro janelado das elocuções. Sendo s_n a amostra n de um quadro janelado do sinal de voz de N_a amostras, a energia E deste quadro pode ser calculada por

$$E = \sum_{n=1}^{N_a} s_n^2 \tag{2.8}$$

2.4.1.4 Coeficientes Cepstrais

O cepstrum de um sinal é a transformada de Fourier inversa do logaritmo do espectro. A habilidade do cepstrum em capturar a estrutura formântica e a inclinação espectral do segmento de voz janelado o tem tornado a característica mais comum para aplicações em tecnologia de fala (Forsyth, 1995). Furui (1981) verificou a eficácia dos coeficientes cepstrais aplicados a verificação de locutor observando que coeficientes

cepstrais obtidos por análise LPC fornecem um resultado semelhante aos obtidos com a FFT, sendo que os primeiros demandam um menor esforço computacional embora sejam menos robustos em relação a ruído.

Através das energias obtidas na análise espectral por banco de filtros é possível calcular-se os coeficientes cepstrais por meio de

$$c_i = \sum_{j=1}^{K} \log E_j \cos\left[\frac{\pi i}{K}(j-0{,}5)\right] \tag{2.9}$$

onde K é o número de filtros do banco e E_j é a energia do filtro j.

O valor do coeficiente c_0 é função da energia total do quadro e algumas vezes é empregado como parâmetro de energia .

2.4.1.5 Coeficientes Delta-Cepstrais

De forma a capturar a informação dinâmica (de transição) do sinal de voz, são utilizadas diferenças entre coeficientes cepstrais de quadros vizinhos. Em verificação de locutor, geralmente são utilizadas as diferenças de primeira e segunda ordem, também denominadas respectivamente de *coeficientes delta-cepstrais* e *delta-delta-cepstrais*. Ao verificar a correlação entre coeficientes cepstrais estáticos e dinâmicos, Soong & Rosenberg (1988) observaram que estes contém informações bastante distintas daqueles e que as características instantâneas como o cepstrum têm melhor performance do que as características de transição, embrora a combinação de ambas melhore a performance do sistema. Em (Soong & Rosenberg, 1988) também foi observado que, devido à natureza diferencial no domínio do logaritmo do espectro, características espectrais de transição são mais resistentes às variações do canal transmissor (ruído convolucional) do que as características instantâneas.

Os coeficientes delta-cepstrais podem ser calculados pela seguinte regressão

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.10}$$

onde d_t é o coeficiente delta do quadro t calculado em termos dos coeficientes estáticos $c_{t+\theta}$ e $c_{t-\theta}$. θ é o intervalo no qual é computada a diferença cepstral (em quadros) e deve representar um intervalo de até 100ms. Os coeficientes dinâmicos superiores podem ser calculados pela mesma fórmula descrita na equação anterior, substituindo-se os coeficientes estáticos pelos dinâmicos de ordem imediatamente inferior. No que diz respeito a esta tese, calculou-se os coeficientes dinâmicos com a equação:

$$d_t = \frac{(c_{t+\theta} - c_{t-\theta})}{2\theta} \tag{2.11}$$

Os problemas referentes ao cômputo do coeficientes dinâmicos no início e no final da elocução podem ser resolvidos por

$$d_t = \begin{cases} c_{t+1} - c_t & , t < 1 \\ c_t - c_{t-1} & , t \geq T - 1 \end{cases} \tag{2.12}$$

onde T é o número de vetores de observação da seqüência.

2.4.2 Seleção de parâmetros para SVL

A presença de características como energia, delta da energia, coeficientes cepstrais estáticos e coeficientes delta-cepstrais melhora o desempenho de SVL de forma diferenciada. Características espectrais instantâneas como o cepstrum possuem mais informações relevantes sobre o locutor do que características espectrais de transição, como o delta-cepstrum (Soong & Rosenberg, 1988). Da mesma forma, Charlet & Jouvet (1997) verificaram em seus experimentos de verificação de locutor que coeficientes cepstrais estáticos de ordem elevada oferecem uma boa distinção entre locutores.

A otimização dos parâmetros que melhor individualiza as características de cada locutor melhora o desempenho do SVL sem exigir um aumento na complexidade computacional. Existem vários procedimentos de seleção de parâmetros entre os quais podemos destacar (Charlet & Jouvet, 1997):

- **Método dos melhores de *N*:** considerando-se os N parâmetros, são selecionados os n melhores parâmetros considerados individualmente. Este procedimento requer N+1 experimentos;
- **Seleção ascendente:** o melhor conjunto de $n + 1$ parâmetros é composto do melhor conjunto de n parâmetros e o parâmetro, entre os $N - n$ restantes, que fornece o melhor conjunto de $n + 1$ parâmetros. Este procedimento requer $N(N+1)/2$ experimentos.
- **Procedimento de eliminação:** é semelhante ao método de seleção ascendente. O parâmetro o qual não fornece os melhores resultados é descartado (Sambur, 1975). Este procedimento passo a passo também requer $N(N+1)/2$ experimentos.
- **Seleção por programação dinâmica:** ao contrário dos anteriores, este procedimento não implica em que o melhor conjunto de n parâmetros é incluído no melhor conjunto de $n + 1$ parâmetros. A seleção é feita por programação dinâmica, como descrito em (Cheung & Eisenstein, 1978). Este procedimento requer $N^2(N-1)/2$ experimentos.

Charlet & Jouvet (1997) compararam os procedimentos de seleção. Observou-se que o procedimento dos melhores de N é pior que os outros. Isto provou que o melhor conjunto de n parâmetros não é necessariamente composto pelos n parâmetros de melhor desempenho individuais, pelo menos quando o critério aplicado é a taxa de erro. Os três outros procedimentos forneceram desempenhos equivalentes.

Charlet & Jouvet (1997) escolheram 27 parâmetros (energia, delta-energia, delta-delta-energia, e os oito primeiros coeficientes com seus respectivos delta e delta-delta coeficientes) dos quais 14 foram selecionados de forma ascendente. A EER foi reduzida de 5,9% para 4,6%. Também foram feitos experimentos introduzindo um peso para a probabilidade de emissão para cada parâmetro, de acordo com a capacidade de cada parâmetro em discriminar locutores. Para cada um dos 27 parâmetros encontrou-se um peso ótimo que resultou em uma EER de 4,1%.

2.5 MODELOS OCULTOS DE MARCOV (HMM)

2.5.1 Definição do problema de classificação de padrões

Cada elocução é representada por uma seqüência de vetores de parâmetros ou observações $\boldsymbol{O}$, definida por

$$\boldsymbol{O} = \boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T \tag{2.13}$$

onde $\boldsymbol{o}_t$ é o vetor de observações no tempo t. Para a classificação de padrões acústicos em verificação de locutor dependente de texto é medida a distância entre o modelo da palavra do locutor afirmado j e a seqüência de vetores de observação $\boldsymbol{O}$ do locutor de teste i. Esta distância é comparada com um limiar de decisão de forma a aceitar ou não o locutor. Esta medida de distância pode ser uma probabilidade e o problema pode ser considerado como estimar:

$$P(S_i = S_j / \boldsymbol{O}, \lambda_j) \tag{2.14}$$

onde S_i e S_j denotam os locutores i e j, e λ_j corresponde ao modelo do locutor j. Esta probabilidade não é obtida diretamente mas usando-se a regra de Bayes:

$$P(S_i = S_j / \boldsymbol{O}, \lambda_j) = \frac{P(\boldsymbol{O} / S_i = S_j, \lambda_j).P(S_i = S_j)}{P(\boldsymbol{O})} \tag{2.15}$$

Como $P(S_i = S_j)$ e $P(\boldsymbol{O})$ são constantes, o cômputo $P(S_i = S_j / \boldsymbol{O}, \lambda_j)$ depende apenas do valor da verossimilhança $P(\boldsymbol{O} / S_i = S_j, \lambda_j)$ que é calculada considerando o processo de produção da voz como sendo Markoviano. Esta é a hipótese assumida pela técnica HMM que é descrita a seguir.

2.5.2 Classificação de Padrões com HMM

Em verificação de locutor baseado em HMM assume-se que cada seqüência de vetores de observação correspondente a cada palavra de cada locutor é gerada por um modelo de Markov, como ilustrado na Fig. 2.6. Um modelo de Markov é uma máquina de estados finita a qual tem uma transição de estado a cada unidade temporal, considerada

como o quadro no caso de voz. A cada quadro t em que há transição para o estado j, um vetor de observações $\boldsymbol{o}_t$ é gerado com a função densidade de probabilidade $b_j(\boldsymbol{o}_t)$. Além disso, a transição do estado i para o estado j acontece com a probabilidade a_{ij}.

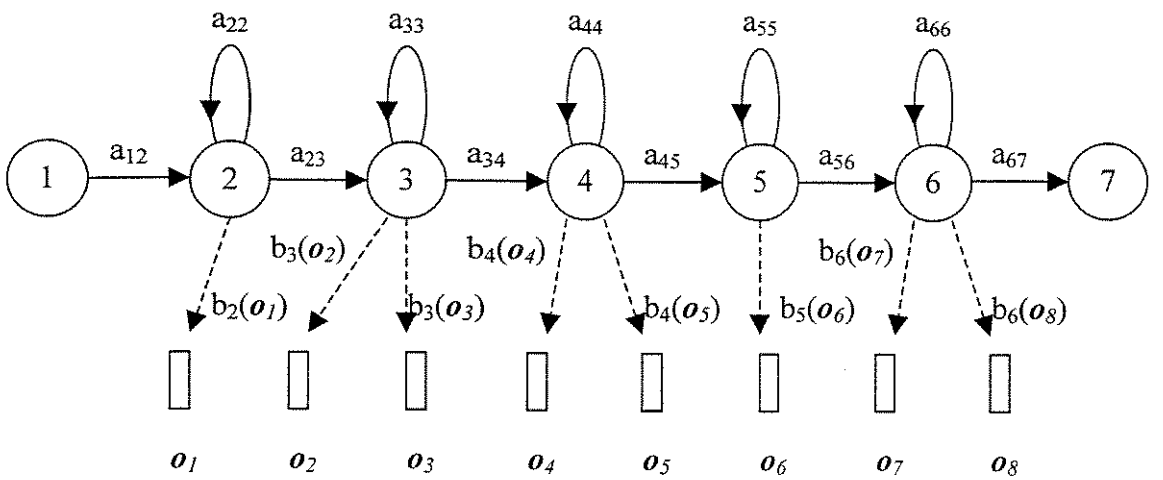

Figura 2.6 - Estrutura e um HMM com topologia de esquerda-direita sem pulo de estado

Tanto em reconhecimento de voz quanto de locutor geralmente se emprega uma topologia de HMM chamada *esquerda-direita* sem pulo de estado. Ela se caracteriza pela possibilidade de auto-transições de estado (transição para o mesmo estado) e de transição para o estado imediatamente posterior. A Fig. 2.6 ilustra um exemplo de topologia HMM esquerda-direita sem pulo de estado com cinco estados emissores (2 a 6). Os estados inicial e final (1 e 7, respectivamente) são não-emissores e têm a função de concatenar modelos no treinamento e no teste. Estes estados fornecem a probabilidade de transição do último estado emissor de um modelo para o primeiro estado emissor do próximo modelo.

A probabilidade conjunta de que $\boldsymbol{O}$ seja gerado pelo modelo λ_c da palavra do afirmado cliente e que a seqüência de estados seja X=1, 2, 3, 3, 4, 4, 5, 6, 6, 7, conforme a Fig. 2.6, é dada por:

$$P(\boldsymbol{O}, X / \lambda_c) = a_{12} b_2(\boldsymbol{o}_1) a_{23} b_3(\boldsymbol{o}_2) a_{33} b_3(\boldsymbol{o}_3) \ldots \tag{2.16}$$

Entretanto, na prática, somente a seqüência de observações $\boldsymbol{O}$ é conhecida e a seqüência de estados X correspondente é oculta. Por este motivo o modelo é chamado oculto (*Hidden*).

Dado que X é desconhecido, a verossimilhança requerida pode ser computada pelo somatório de todas as possíveis seqüências de estados $X = x(1), x(2), ..., x(T)$, que é

$$P(\boldsymbol{O}/\lambda_c) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(\boldsymbol{o}_t) a_{x(t)x(t+1)} \tag{2.17}$$

onde $x(0)$ é o estado inicial do modelo e $x(T+1)$ é o estado final. Como uma alternativa para a Eq. 2.17, a verossimilhança pode ser aproximada considerando-se apenas a seqüência de estados mais provável:

$$P(\mathbf{O}/\lambda_c) \cong \max_{X} \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(\mathbf{o_t}) a_{x(t)x(t+1)} \right\} \tag{2.18}$$

Os parâmetros a_{ij} e $b_j(\boldsymbol{o}_t)$ são conhecidos para cada modelo de cada locutor e previamente obtidos através de algoritmos recursivos, como o algoritmo de re-estimação de Baum-Welch (ver item 2.5.5.1), utilizando elocuções de treinamento. Por outro lado, a Eq. 2.18 é resolvida pelo algoritmo de Viterbi que fornece a seqüência de estados ótima (alinhamento ótimo) e a respectiva verossimilhança.

2.5.3 Probabilidade de Observação

O sinal de voz e, consequentemente, os vetores de observação assumem valores reais (contínuos). Sendo assim, a função densidade de probabilidade (fdp) de emissão de observações $b_j(\boldsymbol{o}_t)$ também deveria ser contínua e frequentemente é representada por uma gaussiana simples ou por uma mistura de M gaussianas (Eq. 2.19).

$$b_j(\boldsymbol{o}_t) = \sum_{k=1}^{M} c_{jk} N(\boldsymbol{o}_t; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \quad , 1 \le j \le \text{número de estados} \tag{2.19}$$

onde c_{jk} é o peso da k-ésima componente e $N(.;\mu, \Sigma)$ é uma gaussiana multivariável com vetor de médias μ e matriz de covariância Σ representada por:

$$N(\boldsymbol{o}_t; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n \mid \boldsymbol{\Sigma} \mid}} e^{-\frac{1}{2}(\boldsymbol{o}_t - \mu)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{o}_t - \mu)} \tag{2.20}$$

onde n é a dimensão de $\boldsymbol{o}_t$. Os pesos c_{jk} devem satisfazer as condições

$$\begin{cases} \sum_{k=1}^{N} c_{jk} = 1 \quad ,1 \le j \le N \\ c_{jk} \ge 0 \quad ,1 \le j \le N, 1 \le k \le M \end{cases} \tag{2.21}$$

Vale a pena mencionar que, embora não empregada nesta tese, uma alternativa a fdp´s contínuas é a utilização de funções de probabilidade discretas e quantização vetorial que caracterizam os HMM discretos (Rabiner & Juang, 1993).

2.5.4 Algoritmo de Decodificação de Viterbi

O algoritmo de Viterbi é responsável pela estimação da verossimilhança da elocução de teste (seqüência de vetores de observação) dado o modelo do locutor afirmado (padrão de referência). A Eq. 2.18 é calculada obtendo-se a seqüência de estados X mais provável, conforme indica a treliça da Fig. 2.7. A dedução completa do algoritmo de Viterbi pode ser encontrado em (Rabiner & Juang, 1993).

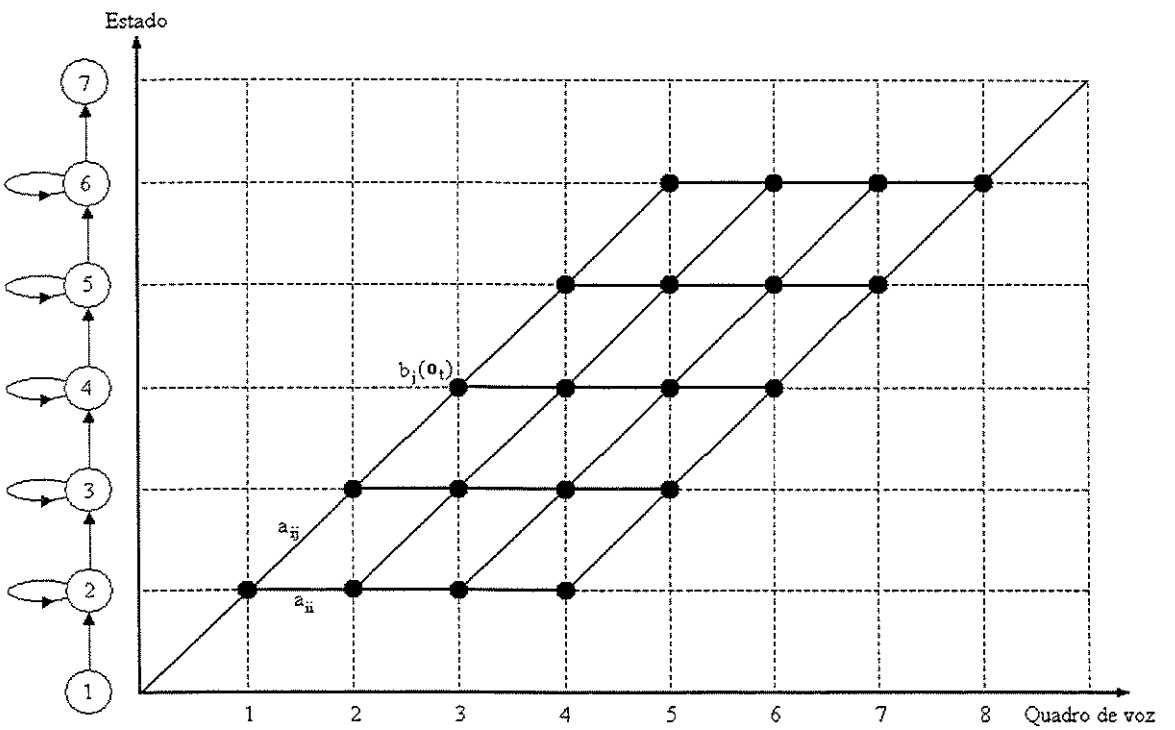

Figura 2.7 - Treliça do alinhamento de Viterbi

Considere $\boldsymbol{X} = x_1 x_2 ... x_T$ a melhor seqüência de estados para uma dada seqüência de vetores de observação $\boldsymbol{O} = \boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T$. Considere também $\delta_i(t)$ como sendo a

probabilidade máxima ao longo de um caminho, no tempo t, a qual é computada para as primeiras t observações e finalizando no estado i. Temos

$$\delta_i(t) = \max_{x_1, x_2, \ldots, x_{t-1}} \{P(x_1 x_2 \ldots x_{t-1}, x_t = i, \boldsymbol{o}_1 \boldsymbol{o}_2 \ldots \boldsymbol{o}_t / \lambda_c)\} \tag{2.22}$$

Por indução tem-se:

$$\delta_j(t+1) = [\max_i \delta_i(t) a_{ij}] \cdot b_j(\boldsymbol{o}_{t+1}) \tag{2.23}$$

Para recuperar a seqüência de estados, precisa-se guardar o caminho o qual maximiza a equação anterior, para cada t e j. Faz-se isto através da matriz $\psi_j(t)$. O algoritmo completo para obter-se a máxima verossimilhança e a melhor sequência de estados é mostrado a seguir:

Inicialização:

$$\delta_j(1) = \begin{cases} 1 & , i = 1 \\ a_{1j} b_j(\boldsymbol{o}_1) & , 2 \le j \le N \end{cases} \tag{2.24}$$

Recursão:

$$\delta_j(t) = \max_i [\delta_i(t-1) a_{ij}] \cdot b_j(\boldsymbol{o}_t) \quad ; 2 \le t \le T; 1 \le i, j \le N \tag{2.25}$$

$$\psi_j(t) = \arg\max_i [\delta_i(t-1) a_{ij}] \quad ; 2 \le t \le T; 1 \le i, j \le N \tag{2.26}$$

Finalização:

$$P(\boldsymbol{O} / \lambda_c) = \max_i [\delta_i(T)] \quad , 1 \le i \le N \tag{2.27}$$

$$x_T = \arg\max_i [\delta_i(T)] \quad , 1 \le i \le N \tag{2.28}$$

Recursão para a obtenção da melhor seqüência de estados

$$x_t = \psi_{x_{t+1}}(t+1) \quad , t = T-1, T-2, \ldots, 1 \tag{2.29}$$

Para evitar *underflow* e substituir somas por multiplicações no cômputo da verossimilhança, calculou-se o logaritmo da verossimilhança ao invés do valor direto da probabilidade.

Como mostrado na Fig. 2.7, este algoritmo pode ser visualizado como a obtensão do melhor caminho através de uma matriz onde a dimensão vertical representa os estados do HMM e a dimensão horizontal representa os quadros do sinal de voz. Cada nó na figura representa o logaritmo da probabilidade de observação daquele quadro naquele tempo. Cada conexão entre nós corresponde ao logaritmo da probabilidade de transição. O logaritmo da probabilidade de cada caminho é computado pelo somatório do logaritmo destas probabilidades ao longo do caminho.

2.5.5 Treinamento dos parâmetros de um HMM

Em verificação de locutor, é necessário comparar a sequência de vetores de observação (que representam a elocução) do locutor de teste com um padrão ou modelo (HMM) de referência para o cômputo da distância ou verossimilhança, e consequentemente a tomada de decisão. Cada palavra ou subunidade fonética de cada locutor (no caso dependente de texto) deve ter seu padrão ou modelo (HMM) de referência. Este modelo (no caso dos HMM) consiste nas probabilidades de transição (a_{ij}) e nas fdp de emissão de observações [$b_j(\boldsymbol{o}_t)$]. No caso contínuo, como esta última pode ser representada por uma distribuição gaussiana, é necessário apenas a obtenção da média e da variância da distribuição.

Para obter-se o padrão de referência, o locutor participa de seções de treinamento em que são coletadas várias elocuções com várias repetições de cada palavra ou subunidade fonética de cada locutor (no caso dependente de texto) de forma a modelar as várias fontes de variabilidade inerentes a fala (variabilidade intra-locutor). Assim, uma série de seqüências de vetores de observação de treinamento $\boldsymbol{O}^r$, $1 \leq r \leq R$, são usadas para estimar os parâmetros de um HMM. Neste trabalho as probabilidades de observação foram modeladas utilizando gaussianas simples, e considerando que os coeficientes não são correlacionados entre si o que conduz a uma matriz de covâriancia diagonal.

Para determinar-se os parâmetros de um HMM, primeiro é necessário fazer uma estimação grosseira de seus valores. As médias e as covariâncias de $b_j(\boldsymbol{o}_t)$ podem ser obtidas pelas Eq. 2.30 e 2.31, dividindo eqüitativamente cada uma das elocuções de treinamento pelo número de estados e associando cada parte da seqüência de vetores de observação ao estado correspondente:

$$\mu_{j,n} = \frac{1}{T}\sum_{t\in j} \mathbf{o}_{t,n} \tag{2.30}$$

$$\sigma^2{}_{j,n} = \frac{1}{T}\sum_{t\in j} (\boldsymbol{o}_{t,n} - \mu_{j,n})(\boldsymbol{o}_{t,n} - \mu_{j,n})' \tag{2.31}$$

onde $\mu_{j,n}$ e $\sigma^2{}_{j,n}$ .são a média e variância do coeficiente n no estado j. Uma vez feito isto, parâmetros mais precisos (no sentido da máxima verossimilhança) podem ser encontrados aplicando-se os algoritmo de reestimação de Baum-Welch ou de Viterbi ou ambos. Freqüentemente o algoritmo de Viterbi é utilizado inicialmente para obtêr-se uma melhor aproximação dos valores iniciais de um HMM para a reestimação de Baum-Welch. Este algoritmo de reestimação de parâmetros será descrito a seguir em maiores detalhes, conforme Young *et al.* (1997).

2.5.5.1 Algoritmo de Reestimação de Baum-Welch

Ao contrário da aproximação inicial, cada observação é associada a cada estado proporcionalmente à probabilidade do modelo estar naquele estado quando o vetor for observado. Assim, se $L_j(t)$ representa a probabilidade de estar no estado j no tempo t, têm-se as seguintes equações

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t)\boldsymbol{o}_t}{\sum_{t=1}^{T} L_j(t)} \tag{2.32}$$

$$\hat{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)(\boldsymbol{o}_t - \mu_j)(\boldsymbol{o}_t - \mu_j)'}{\sum_{t=1}^{T} L_j(t)} \tag{2.33}$$

As Eq. 2.32 e 2.33 são as fórmulas de reestimação de Baum-Welch das médias e covariâncias de um HMM para uma única $\boldsymbol{O}$. O cálculo da probabilidade de ocupação de estado $L_j(t)$ pode ser feito utilizando-se o algoritmo *Forward-Backward*. A probabilidade *forward* $\alpha_j(t)$ para um modelo M com N estados é definida como a probabilidade conjunta de observação dos primeiros t vetores de voz estando no estado j no tempo t, ou seja

$$\alpha_j(t) = P(\boldsymbol{o}_1,...,\boldsymbol{o}_t, x(t) = j / \lambda_c) \tag{2.34}$$

O algoritmo *forward* pode ser dado por

Condições iniciais (para $1<j<N$)

$$\alpha_1(1) = 1 \tag{2.35}$$

$$\alpha_j(1) = a_{1j} b_j(\boldsymbol{o}_1) \tag{2.36}$$

Recursão

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1) a_{ij} \right] b_j(\boldsymbol{o}_t) \tag{2.37}$$

Finalização

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN} \tag{2.38}$$

Pela definição de $\alpha_j(t)$, seu cálculo também produz a verossimilhança total

$$P(\boldsymbol{O} / \lambda_c) = \alpha_N(T) \tag{2.39}$$

A probablididade *backward* $\beta_j(t)$ é definida como

$$\beta_j(t) = P(\boldsymbol{o}_{t+1},...,\boldsymbol{o}_T / x(t) = j, \lambda_c) \tag{2.40}$$

O algoritmo *backward* é o seguinte:

Condições iniciais (para 1<*i*<*N*)

$$\beta_i(T) = a_{iN} \tag{2.41}$$

Recursão

$$\beta_i(t) = \sum_{i=2}^{N-1} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_j(t+1) \tag{2.42}$$

Finalização

$$\beta_1(1) = \sum_{i=2}^{N-1} a_{1j} b_j(\boldsymbol{o}_1) \beta_j(1) \tag{2.43}$$

A probabilidade de ocupação de estado pode ser determinada pelo produto das probabilidades *forward* e *backward*. Pela definição, temos

$$\alpha_j(t)\beta_j(t) = P(\boldsymbol{O}, x(t) = j / \lambda_c) \tag{2.44}$$

então,

$$L_j(t) = P(x(t) = j / \boldsymbol{O}, \lambda_c) = \frac{P(\boldsymbol{O}, x(t) = j / \lambda_c)}{P(\boldsymbol{O} / \lambda_c)} = \frac{\alpha_j(t)\beta_j(t)}{P(\boldsymbol{O} / \lambda_c)} \tag{2.45}$$

Resumindo, o algoritmo de Baum-Welch pode ser decomposto no seguintes passos:

1. Armazenar o somatório dos valores dos numeradores e denominadores das Eq. 2.32 e 2.33 para cada vetor de parâmetros em acumuladores.
2. Calcular as probabilidades *forward* e *backward* para todos os estados j e tempos t.
3. Para cada estado j e tempo t, usa-se a probabilidade $L_j(t)$ e o vetor de observações $\boldsymbol{o}_t$ corrente para atualizar os acumuladores para aquele estado.
4. Utiliza-se o valor final dos acumuladores para calcular um novo valor para os parâmetros.

5. Se o valor $P(\boldsymbol{O}/\lambda_c)$ para esta iteração não for maior que do que o valor da iteração anterior então pare, caso contrário repita os passos anteriores usando os valores reestimados dos parâmetros.

Os passos anteriores dizem respeito a reestimação de parâmetros de um HMM para uma única seqüência de vetores de observação. Na prática, muitos exemplos são necessários para conseguir-se uma boa estimação dos parâmetros. Para seqüências de observações múltiplas simplesmente repete-se os passos 2 e 3 para cada seqüência de treinamento distinta. Uma descrição mais detalhada do algoritmo de Baum-Welch pode ser obtida em Young *et al.* (1997).

2.6 NORMALIZAÇÃO DA VEROSSIMILHANÇA

Em um enfoque clássico, a decisão lógica é tomada comparando-se a verossimilhança obtida entre o modelo da afirmada identidade e a elocução de teste, com o limiar previamente estabelecido no treinamento. Neste caso, o valor da verossimilhança final leva em consideração tanto a similaridade entre locutores (de teste e da afirmada identidade) quanto a similaridade lingüística (entre a elocução de teste e as palavras do modelo). Ou seja, ao valor da distância desejada é adicionado outro valor que depende da variabilidade natural da fala intra-locutor. Sendo assim, um limiar de decisão estável é difícil de ser fixado.

Uma solução para o problema da variabilidade do limiar é aplicar a técnica de normalização da verossimilhança, a qual melhora significativamente a performance da verificação (Higgins *et al.*,1991; Rosenberg *et al.*, 1992; Carey & Parris, 1992; Matsui & Furui, 1993, 1994a, 1994b; Nakagawa & Markov, 1997; Markov & Nakagawa, 1998). Um dos métodos é baseado na relação de verossimilhanças (Higgins *et al.*,1991), onde a verossimilhança normalizada $L(\boldsymbol{O})$ é obtida pela razão

$$L(\boldsymbol{O}) = \frac{P(\boldsymbol{O}/S_c)}{P(\boldsymbol{O}/S_{\bar{c}})} \tag{2.46}$$

onde S_c é o afirmado locutor e $S_{\bar{c}}$ é o modelo representativo de todos os outros locutores.

O denominador da equação anterior é chamado de termo de normalização. Geralmente valores positivos de $log\ L(\boldsymbol{O})$ indicam que o locutor de teste é realmente o locutor afirmado e vice-versa.

A verossimilhança $P(\boldsymbol{O}/S_c)$, representada pela Eq. 2.18, é computada pelo algoritmo de decodificação de Viterbi. Idealmente, a probabilidade condicional do termo de normalização deveria ser calculada pela média das verossimilhanças da seqüência de observação $\boldsymbol{O}$ em relação ao modelo de cada impostor. Isto é computacionalmente inviável. Entretanto, a verossimilhança $P(\boldsymbol{O}/S_{\bar{c}})$ pode ser calculada por aproximação utilizando-se um subconjunto de B locutores representativos dos impostores. Este subconjunto de locutores é chamado de "*cohort*". Higgins *et al.* (1991) propuseram o uso de locutores que fossem representativos da população com modelos próximos ao do afirmado locutor. Assim, o temo de normalização pode ser calculado para os locutores "*cohort*" por

$$P(\boldsymbol{O}/S_{\bar{c}}) = \underset{b}{stat}\{P(\boldsymbol{O}/S_b)\} \quad , b = 1,\ldots,B \tag{2.47}$$

onde "*stat*" refere-se alguma medida estatística como mínimo, máximo ou a média das B probabilidades.

Rosenberg *et al.*(1992) verificaram também que a normazilação da verossimilhança reduz significativamente o erro provocado pela presença de ruído convolucional. Utilizando normalização com "*cohort speakers*", eles compararam elocuções de teste, gravadas com microfone de eletreto, com modelos construídos com elocuções de treinamento, gravadas com microfone de carbono.

Matsui & Furui (1993, 1994a, 1994b) propuseram um método de normalização baseado em uma probabilidade a posteriori. Este método consiste em obter o termo de normalização com os locutores com distribuição mais próxima do afirmado locutor, incluindo o próprio afirmado locutor. Experimentos indicaram que o método da relação de verossimilhanças e o da probabilidade a posteriori aumentam de forma semelhante a separabilidade entre locutores e reduzem a necessidade de limiares dependentes de texto e locutor. O método da probabilidade a posteriori é dado por

$$L(\boldsymbol{O}) = \frac{P(\boldsymbol{O}/S_c)}{\sum_{b=1}^{B} P(\boldsymbol{O}/S_b)} \tag{2.48}$$

Verificou-se que o método com *"cohort speakers"* torna o sistema vulnerável a ataques de impostores do sexo oposto. Assim, geralmente o *"cohort"* contém somente locutores do mesmo sexo.

Carey & Parris (1992) propuseram um método utilizando com modelo global no qual o modelo do termo de normalização representa a população em geral. O modelo global é obtido treinando-se conjuntamente todas as elocuções de treinamento de todos os locutores, de forma a obter-se o modelo linguístico da palavra ou sub-unidade, como ocorre no treinamento para sistemas de reconhecimento de voz independentes de locutor. Este método é computacionalmente melhor já que exige apenas um modelo global para toda a população e não necessita da soma de valores de verossimilhança como no método em que empregam-se *"cohort speakers"*.

Matsui & Furui (1994a, 1994b) também propuseram um método baseado na ponderação das misturas dos HMM no qual o modelo global é construido como um modelo de misturas reunidas representando a distribuição de parâmetros para todos os locutores registrados. Este modelo é criado fazendo-se as médias dos pesos das misturas do modelo de cada locutor de referência. Sempre que outro locutor cadastra-se no sistema, o modelo de misturas reunidas precisa ser novamente calculado. Este método mostrou bons resultados, melhores até que os utilizados anteriomente.

Outro método que mostrou-se bastante eficiente é o proposto por Nakagawa & Markov (1997) (Markov & Nakagawa, 1998) chamado *Weighting Models Rank* (WMR) que consiste em uma normalização ao nível de quadro.

Neste capítulo descreveram-se as etapas que compõem um Sistema de Verificação de locutor. No capítulo seguinte será apresentado um SVL dependente de texto implementado neste trabalho. Também serão mostrados os resultados dos testes obtidos empregando-se a base de dados YOHO, com elocuções livres de ruído.

Capítulo 3

Implementação e Resultados de um SVL

3.1 INTRODUÇÃO

Neste capítulo descrever-se-ão todas as etapas envolvidas na implementação de um sistema automático de verificação de locutor. O objetivo é demostrar experimentalmente toda a teoria descrita nos capítulos iniciais. Também serão apresentados os resultados obtidos nos testes do SVL com a base de dados YOHO. Os testes foram realizados sem a inserção de ruído e utilizando uma largura de banda equivalente à da linha telefônica. Ao final apresentar-se-ão sugestões para melhoria no desempenho do sistema.

3.2 BASE DE DADOS

Para realizar experimentos com um SVL é necessário que se grave tanto elocuções de treinamento quanto de teste para vários locutores. Este conjunto de elocuções forma uma base de dados que deveria ser representativa da população. Como discutido no capítulo 1, a

voz naturalmente tem muitas fontes de variabilidade, tanto de locutor para locutor (inter-locutor) como para o mesmo locutor ao longo do tempo (intra-locutor). Assim, uma base de dados de teste para verificação de locutor deve conter o maior número possível de locutores (para melhor representar a variabilidade inter-locutor) e de elocuções de cada locutor gravadas em várias seções (para melhor representar o locutor e as variações temporais de sua voz).

Outro aspecto importante na realização de experimentos tanto em SVL quanto em reconhecimento de voz em geral é a utilização de bases de dados padrões. Torna-se difícil comparar-se técnicas as quais tenham sido divulgadas com resultados obtidos com bases de dados diferentes. A utilização de uma base de dados de teste padrão impõe que todos os experimentos, com técnicas distintas, sejam realizados nas mesmas condições oferecendo uma referência de comparação. Assim sendo, os experimentos desta dissertação utilizaram a base de dados padrão YOHO de LDC (*Linguistic Data Consortium*) que é descrita a seguir.

3.2.1 Base de dados de Verificação de Locutor YOHO

A YOHO foi gravada pela ITT *Defense Communications Division* num contrato com o Departamento de Defesa norte-americano. Ela consiste de elocuções gravadas com vários locutores em um escritório com baixo nível de ruído onde cada locutor era induzido a falar frases em inglês contendo três dezenas, como "73 – 24 – 59" que pronunciada correponderia a "*seventy three, twenty four, fifty nine*".

As gravações foram feitas em várias seções com um microfone handset de eletreto unidirecional sem cancelamento de ruído com uma taxa de amostragem de 8 kHz. Cada locutor participou de 14 seções com 3 dias de separação entre elas (separação nominal) durante um intervalo de 3 meses. As primeiras quatro seções foram de treinamento, as quais requeriam cerca de 3 minutos cada; as 10 seções seguintes foram de teste, as quais levaram em torno de 20 segundos cada. As dezenas de cada frase foram escolhidas de 21 a 99 com as seguintes exceções: décadas exatas (30, 40, etc.), dígitos duplos (22, 33, etc.) e números que finalizem com "8" (28, 38, etc.). A pausa entre dezenas é opcional mas não encorajada. Cada seção de treinamento e de teste consiste, respectivamente, na gravação de 24 e 4

frases distintas. A base de dados contém 138 locutores distintos (108 masculinos e 30 femininos). A maioria deles residia na área de Nova Yorque, embora haja muitas exceções e inclusive locutores de inglês não-nativo.

Neste trabalho, utilizaram-se 41 locutores (32 masculinos e 9 femininos) para a construção de um modelo global para a normalização da verossimilhança e os outros 97 foram empregados nos testes do SVL.

3.3 IMPLEMENTAÇÃO DE UM SVL

3.3.1 Extração de Parâmetros

Cada sinal digitalizado correspondente a uma elocução deve ser representado por uma seqüência de vetores de parâmetros. Assim, inicialmente dividiu-se o sinal de voz em quadros de 25 ms (200 amostras) e se calculou as respectivas energias. De forma a tornar o parâmetro de energia confiável, as amostras do sinal foram normalizadas em amplitude em relação a máxima energia de quadro. A seguir, realizou-se as etapas finais de pré-processamento, detectando o inicio e o fim da elocução e realizando uma filtragem de pré-ênfase com um fator de 0,97. O sinal pré-processado é, então, processado com uma janela de Hamming de 25ms (200 pontos) a cada 10 ms (sobreposição de 60%). Cada quadro janelado foi representado por um vetor de 26 parâmetros obtidos através da analise espectral com uma FFT de 256 pontos e com um banco de 20 filtros. Os filtros foram dispostos linearmente em escala mel de 300 a 3400 Hz (banda passante do canal telefônico). A partir da energia dos filtros, foram calculados 12 coeficientes cepstrais (c_1 a c_{12}) e o logaritmo da energia do quadro (somatório das energias dos filtros) ($\log E$), e suas respectivas diferenças de primeira ordem (δc_1 a δc_{12}, e $\delta \log E$) compondo os 26 parâmetros do vetor.

3.3.2 Modelamento do Locutor

Utilizou-se Modelos Ocultos de Markov (HMM) como método para a classificação de padrões. Modelou-se cada palavra (década ou unidade) de cada locutor por um HMM do

tipo esquerda-direita, conforme a Fig. 2.6. Cada locutor é representado, então, por 16 modelos que representam a locução das palavras *one*, *two*, *three*, *four*, *five*, *six*, *seven*, *nine*, *twenty*, *thirty*, *forty*, *fifty*, *sixty*, *seventy*, *eighty* e *ninety*.

Não há uma regra geral para a determinação do número de estados ótimo de cada modelo de um HMM. Isto pode ser feito empiricamente, testando-se o SVL com modelos com diferentes números de estados. No caso desta dissertação, optou-se por adotar um modelo de 8 estados como os utilizados em alguns trabalhos de reconhecimento de voz com modelamento por palavras (um HMM por palavra). São utilizadas modelos contínuos com gaussiana simples e matrizes de covariância diagonais, com um limite inferior para a variância de 0,17.

3.3.3 Normalização da Verossimilhança com Modelo Global

Fez-se a normalização da verossimilhança utilizando-se um modelo global treinado com os 41 locutores restantes (41 masculinos e 10 femininos), diferentes dos locutores de teste. A idéia é extrair ou modelar toda a informação lingüística contida na fala de modo a isolar as características de cada locutor utilizando a normalização. Treinou-se conjuntamente todas as 96 elocuções de cada um dos 41 locutores para formar um único HMM para cada palavra.

3.3.4 Treinamento dos HMM

Para o treinamento dos modelos de palavras, individuais e global, utilizaram-se todas as elocuções de treinamento da YOHO, totalizando 13248 elocuções (96 elocuções por locutor). Neste trabalho não implementou-se os algoritmos de treinamento. Para a realização desta tarefa, treinou-se os modelos de forma concatenada através do *software HTK* (*Hidden Markov Models Tool Kit*) da *Entropic*.

Para obter-se boas condições iniciais para os parâmetros de cada HMM do modelo global, inicialmente algumas repetições das palavras de um único locutor foram segmentadas manualmente. Para cada palavra isoladamente, calculou-se a média e a variância global de seus vetores de observação sendo, então, assumidas como as novas médias e variâncias de cada estado. A seguir, empregou-se uma função para a reestimação

dos parâmetros dos HMM concatenados, utilizando o algoritmo de reestimação de Baum-Welch para todas as 96 elocuções de treinamento de cada um dos 41 locutores que formam o modelo global. Assim, obtêve-se 16 modelos de palavra globais, treinados com 41 locutores e 3936 elocuções (aproximadamente 1476 repetições para cada modelo).

Treinou-se os modelos de palavra individuais de cada locutor através do mesmo procedimento descrito anteriormente para o modelo global excetuando-se no fato de que o modelo inicial para o treinamento foi o próprio modelo global. Ao treinar os modelos individuais de cada locutor foram empregadas todas as 96 elocuções de treinamento do mesmo, consistindo em aproximadamente 36 repetições para cada modelo.

3.3.5 Verificação de Identidade

Tendo a seqüência de observação de um locutor que afirma ser um determinado cliente e conhecendo a priori a seqüência de palavras, os modelos (HMM) do cliente são concatenados formando um modelo único de 48 estados que representa o modelo da frase induzida. Então, a seqüência de observação é processada via algoritmo de Viterbi com o modelo da frase induzida, sendo que o valor do logaritmo da verossimilhança é obtido no último estado do último quadro. Obteve-se o termo de normalização da mesma maneira diferenciando-se apenas nos modelos ocultos de Markov, que devem ser os do modelo global. Assim, tem-se um valor do logaritmo da verossimilhança normalizada $\log L(\boldsymbol{O})$, resultante da subtração entre os logaritmos das verossimilhanças dos HMM do cliente e globais, para a comparação com o limiar

Para os testes com o SVL implementado, obteve-se as curvas de falsa aceitação e falsa rejeição para cada um dos 97 locutores. A curva de falsa rejeição de um determinado locutor, foi obtida calculando-se a verossimilhança para cada uma das suas 40 elocuções de teste. Já na curva de falsa aceitação, o locutor em questão teve seu modelo comparado com uma elocução de cada um dos outros 96 locutores, resultando em 96 valores de $\log L(\boldsymbol{O})$. Comparando-se as curvas de FA e FR obteve-se o valor da taxa de erros iguais (EER) para cada locutor. Neste caso, o limiar de decisão é considerado a posteriori.

3.4 RESULTADOS DE VERIFICAÇÃO DE LOCUTOR

Os valores de EER encontrados para o sistema são mostrados na Tab. 3.I.

Tabela 3.I - Taxa de Erros Iguais para o SVL

NORMALIZAÇÃO	EER_{SS}	EER_{SI}
Com	0,36	0,96
Sem	2,46	4,32

onde EER_{SS} é a taxa de erros iguais média (um limiar de decisão distinto para cada locutor) e EER_{SI} é a taxa de erros iguais independente de locutor (limiar de decisão único). Pode-se observar que há uma melhora de 85% no desempenho do sistema com a presença da normalização da verossimilhança, o que é significativo em função da base de dados utilizada.

De acordo com a aplicação do SVL e, consequentemente, com o grau de segurança exigido, um limiar de decisão poderia ser escolhido a posteriori conforme as tolerâncias máximas na aceitação de impostores e rejeição de clientes. Como exemplo, poderia ser fixado um limiar de decisão único de forma a ser condizente com um limite de aceitação de impostores de 0,2%, o que corresponderia a uma taxa de rejeição de clientes de aproximadamente 3%, conforme a Fig. 3.1.

Também pode ser observado na Tab. 3.I que a utilização de um limiar de decisão único fornece um desempenho pior do que quando da utilização de um limiar individual por locutor. No caso do SVL utilizado neste trabalho, a EER para um limiar único é de 0,96% enquanto que para limiares individuais é de 0,36%.

Uma pequena parcela dos locutores, os *goats*, têm uma taxa de falsa rejeição elevada. Adotando-se 1% como sendo uma boa EER individual, poder-se-á considerar os locutores com taxas de erros iguais maiores como sendo *goats* e menores como sendo *sheeps*. Neste caso, de acordo com a Fig. 3.2, têm-se cerca de 10% de *goats* no sistema implementado. Em sistemas comerciais, estes *goats* deveriam ser tratados de forma individual tendo seus modelos retreinados de forma a serem melhor modelados. Além

disso, parâmetros que melhor representam o locutor individualmente poderiam ser otimizados automaticamente.

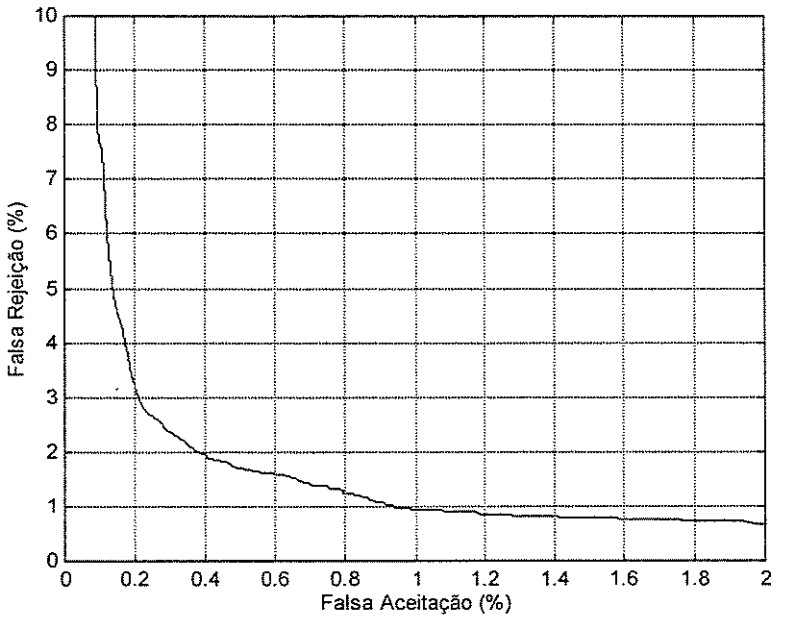

Figura 3.1 - Curva de Operação do Receptor (ROC) independente de locutor (geral) do SVL

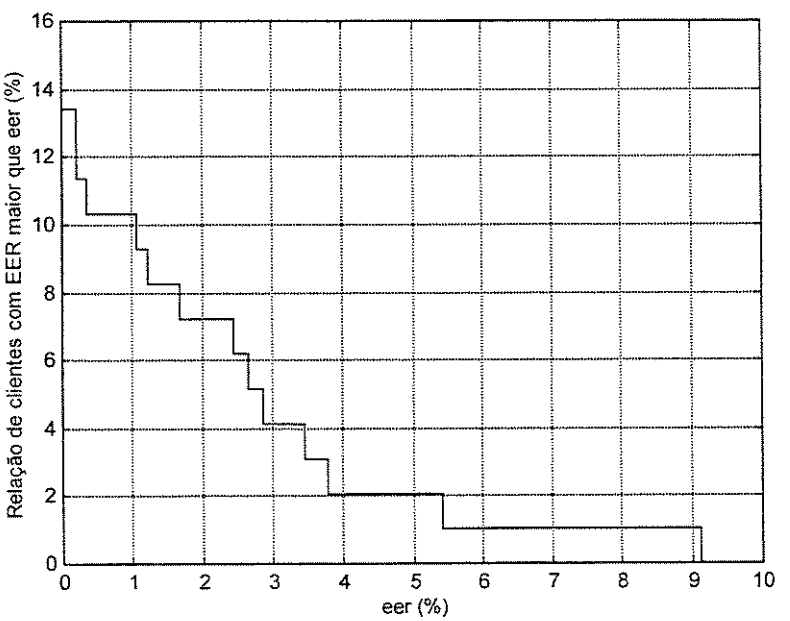

Figura 3.2 - Distribuição da taxa de erros iguais individuais

Os resultados aqui apresentados foram obtidos considerando-se aspectos de um sistema real com a utilização de uma banda de freqüência semelhante à da linha telefônica sem levar em consideração os efeitos do ruído. Vale a pena mencionar que a utilização da banda telefônica inutiliza totalmente as informações espectrais localizadas acima de 4kHz, tais como formantes de mais alta ordem (F4, F5, ...), que são importantes para a distinção entre locutores (Figueiredo, 1994). Em aplicações onde não há uma restrição de banda de freqüência do sinal de voz, como em controle de acesso físico, é esperado uma melhor desempenho do SVL. Neste caso, seria possível a utilização de todo o espectro da voz até os 8kHz.

3.5 SUGESTÕES PARA MELHORIA NO DESEMPENHO DO SVL

No caso de implementação de um SVL comercial, várias alterações e testes podem ser realizados de forma a melhorar o desempenho do sistema. Podemos citar:

- Modelamento por sub-unidade (fone, difone, polifone), o que aumentaria o vocabulário de teste sem aumentar significativamente o vocabulário de treinamento;
- Utilização de modelos com múltiplas gaussianas;
- Obter os valores de sobreposição, tamanho de quadro, número filtros do banco e número de estados por modelos mais adequados para o sistema em questão;
- Selecionar os parâmetros que melhor representam o locutor e/ou utilizar um peso para cada um no cálculo da verossimilhança, como descrito no item 2.4.2.
- Tamanho da elocução de teste pode ser aumentado ou várias elocuções podem ser utilizadas para a tomada de decisão.

Os relatos realizados até então representam uma descrição do SVL em si, mostrando sua viabilidade e desempenho com sinal de voz sem ruído. Entretanto, em condições ruidosas o SVL sofre uma abrupta queda no seu desempenho e precisa de técnicas de robustez a ruído, tanto convolucional quanto aditivo. O capítulo 4 descreve algumas das

técnicas de robustez a ruído mais empregadas em tecnologia de voz. São elas a normalização da média cepstral, a filtragem RASTA (*Log*-RASTA e J-RASTA), a subtração espectral, a normalização da SNR, *Parallel Model Combination* e modelamento de duração de estados. No capítulo 5 serão realizados testes com o SVL empregando-se sinal de voz corrompido por ruído aditivo e/ou convolucional e algumas das técnicas anteriormente citadas.

Capítulo 4

Técnicas de Robustez a Ruído

4.1 DESCRIÇÃO DO PROBLEMA

Atualmente as tecnologias de reconhecimento de voz e de locutor apresentam um desempenho razoável em condições de baixo nível de ruído. Este desempenho tende a melhorar ainda mais a medida que novos trabalhos agregam novas técnicas ao estado da arte. Entretanto, quando se fala em utilização de sistemas de reconhecimento de voz e/ou de locutor em aplicações reais, a presença de ruído (aditivo ou convolucional) degrada fortemente o desempenho destes sistemas conforme a diminuição da SNR. Assim, sistemas reais precisam ser suficientemente robustos para não sofrerem a influência tanto da distorção em freqüência provocada pela resposta impulsiva do microfone e/ou canal de transmissão, quanto do ruído ambiental como o de carro ou o de uma conexão por telefone celular. A distorção causada pelo canal de transmissão é denominada de ruído convolucional e aquela oriunda de um sinal que é adicionado ao sinal de voz recebe o nome de ruído aditivo.

4.1.1 Ruído Convolucional

O ruído convolucional consiste numa distorção no domínio frequencial, o que resulta numa convolução no domínio do tempo. Considere $h(t)$ como sendo a resposta ao impulso, que é considerado invariante no tempo e independente da energia do sinal, do canal transmissor e/ou do microfone utilizado nas gravações da elocução. Tem-se então:

$$y(t) = s(t) * h(t) \tag{4.1}$$

onde $y(t)$ é o sinal de voz corrompido pelo ruído e $s(t)$ o sinal de voz limpo. A Eq. 4.1 pode ser representado no domínio frequencial como sendo

$$Y(\omega) = S(\omega)H(\omega) \tag{4.2}$$

Para um quadro qualquer, através da análise espectral com banco de filtros com energias no domínio logarítmico e considerando a energia dentro de cada filtro como sendo constante, tem-se:

$$\log E_{Y_i} = \log E_{S_i} + \log H_i \quad , 1 \le i \le K \tag{4.3}$$

sendo E_{Y_i} a energia do i-ésimo filtro do sinal distorcido Y, E_{S_i} a energia do filtro correspondente referente ao sinal de voz limpo, H_i é o ganho do canal de transmissão no filtro i, e K é o número de filtros do banco.

Como $h(t)$ é considerada invariante no tempo e independente do sinal de voz, o ruído convolucional pode ser modelado como a soma de uma constante no domínio logarítmico e no domínio cepstral, que por sua vez causa um erro na medida da verossimilhança final. Técnicas como a normalização da média cepstral e a filtragem RASTA tentam subtrair ou diminuir esta constante no domínio cepstral de forma a retirar a influência da distorção do canal de transmissão e/ou microfone.

4.1.2 Ruído Aditivo

O ruído aditivo $n(t)$ representa um processo que é adicionado ao sinal de voz original $s(t)$ no domínio linear gerando o sinal ruidoso $y(t)$, tanto no domínio do tempo:

$$y(t) = s(t) + n(t) \tag{4.4}$$

como no frequencial:

$$Y(\omega) = S(\omega) + N(\omega) \tag{4.5}$$

Ao contrário do ruído convolucional, o ruído aditivo pode sofrer uma variação temporal significativa e corrompe diferenciadamente o sinal de voz de acordo com a SNR segmental ou local, o que dificulta ainda mais a sua supressão. De maneira geral, as técnicas de cancelamento de ruído aditivo dependem da precisão e capacidade em acompanhar a dinâmica temporal da estimação do ruído. Entretanto, por simplicidade, muitas técnicas de robustez consideram o ruído aditivo estacionário no intervalo de tempo onde elas estão atuando. Neste sentido a subtração espectral (SS - *Spectral Subtraction*) é freqüentemente empregada por sua simplicidade e sua capacidade em acompanhar razoavelmente bem a dinâmica do ruído aditivo.

4.1.3 Ruído Convolucional e Aditivo

Em muitas situações ambos ruídos, aditivo e convolucional, estão presentes, como por exemplo numa ligação telefônica. Uma das situações reais poderia ser representada pelo modelo ilustrado nas Eq. 4.6 e 4.7, e na Fig. 4.1. A interferência dos ruídos convolucional e aditivo pode ser expressa no domínio do tempo por:

$$y(t) = h(t) * [s(t) + n(t)] \tag{4.6}$$

e no frequencial

$$Y(\omega) = H(\omega)[S(\omega) + N(\omega)] \tag{4.7}$$

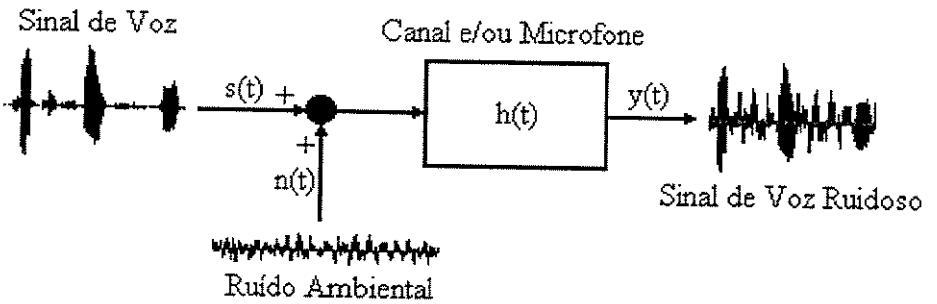

Figura 4.1 - Sinal de voz corrompido por ruído aditivo e convolucional

A geração dos sinais ruidosos de teste foi feita segundo as Eq. 4.6 e 4.7 e a Fig. 4.1. No entanto, para efeito de cancelamento de ruído, um modelo freqüentemente utilizados (e adotado nesta dissertação) considera que primeiro é introduzido o efeito do canal e depois é acrescido o ruído aditivo. Assim, o modelamento para a presença conjunta de ruído aditivo e convolucional, do ponto de vista das técnicas de cancelamento, pode ser feito pelas Eq. 4.8 e 4.9.

$$y(t) = h(t) * s(t) + n(t) \tag{4.8}$$

$$Y(\omega) = H(\omega)S(\omega) + N(\omega) \tag{4.9}$$

Isto facilita a consideração de técnicas conjuntas para o cancelamento de ambos os ruídos, como SS e filtragem RASTA. Como no caso deste exemplo, primeiro aplica-se SS de forma a cancelar o termo aditivo do ruído em freqüência ($N(\omega)$) e, depois, o filtro *Log*-RASTA para eliminar a componente de ruído multiplicativa em freqüência ($H(\omega)$). A natureza de ambos tipos de ruído também é coerente com este modelo: o ruído aditivo pode variar mais rapidamente do que o ruído convolucional e faz sentido que seja cancelado primeiro.

Para o cancelamento de ambos os ruídos podem tanto ser empregadas técnicas isoladas como PMC (Gales & Young, 1993a, 1993b, 1996; Gales, 1997), estimação por máxima verossimilhança (Acero & Stern, 1990; Raj *et al.*, 1996), e normalização da SNR (Van Compernolle, 1989; Claes & Van Compernolle, 1996; Claes *et al.*, 1996), quanto a combinação de técnicas geralmente empregadas quando há apenas um tipo de ruído. A seguir serão discutidas algumas das técnicas que podem ser empregadas para o

cancelamento destes dois tipos de ruídos, aditivo e convolucional, individual ou conjuntamente.

4.2 TÉCNICAS DE ROBUSTEZ A RUÍDO

Várias técnicas de robustez a ruído de sistemas de reconhecimento de voz vem sendo propostas. Muitas delas podem ser empregadas com sucesso também em reconhecimento de locutor. Dentre estas técnicas convencionais, podemos citar: normalização da média cepstral (CMN) (Atal, 1974; Furui, 1981; Rosenberg *et al.*, 1994), espectro relativo (RASTA) (Hermansky *et al.*, 1991; Hermansky *et al.*, 1993; Hermansky & Morgan, 1994), subtração espectral (SS) (Berouti *et al.*, 1979; Van Compernolle, 1989; Vasegui & Milner, 1997) normalização da SNR (Van Compernolle, 1989; Claes & Van Compernolle, 1996; Claes *et al.*, 1996), *Parallel Model Combination* (PMC) (Gales & Young, 1993a, 1993b, 1996; Gales, 1997) e modelamento de duração de estados com restrições temporais, esta recentemente proposta para robustez em SRV (Yoma *et al.*, 2000).

4.2.1 CMN

Também conhecida como *subtração da média cepstral* (CMS – *Cepstral Mean Subtraction*,), a *normalização da média cepstral* (CMN – *Cepstral Mean Normalization*) (Atal, 1974; Furui, 1981; Rosenberg *et al.*, 1994) tem se mostrado muito eficiente em compensar a distorção provocada pelo microfone e/ou pelo canal transmissor. Neste método, a média temporal de cada coeficiente cepstral, em uma dada elocução, é subtraída do correspondente coeficiente em todos os quadros. Considera-se que a média temporal de cada coeficiente cepstral não contém nenhuma informação relevante do ponto de vista acústico-fonético e do locutor, o que pode não ser verdade em algumas circunstâncias. Como o ruído convolucional pode ser modelado como uma constante no domínio logarítmico e cepstral, ao aplicar-se a CMN a média de todos os coeficientes cepstrais num dado intervalo de tempo é feita igual a zero eliminando assim aquela componente relativa ao ruído convolucional.

Entretanto, o desempenho da verificação em condições de treinamento e teste semelhantes pode ser prejudicado com CMN devido à consideração de que a média de cada coeficiente cepstral para o sinal limpo seria zero para todas as elocuções. Esta consideração só é válida integralmente quando a elocução é foneticamente balanceada, no qual incluem-se aproximadamente a mesma quantidade de sons sonoros, fricativos e plosivos (Mammone *et al.*, 1996). Assim, como a aplicação da CMN remove algumas características dependentes do texto e do locutor, ela não seria muito apropriada para verificação de locutor com elocuções de teste muito curtas (Furui, 1997).

4.2.2 RASTA

O *espectro relativo*, conhecido como RASTA (*RelAtive SpecTrA*) (Hermansky *et al.*, 1991; Hermansky *et al.*, 1993; Hermansky & Morgan, 1994), é uma técnica que tenta suprimir as componentes que variam mais lenta ou rapidamente do que a dinâmica natural de variação da voz. Existem duas técnicas oriundas deste conceito, o *Log*-RASTA e o J-RASTA. A primeira é aplicada para supressão de ruído convolucional, embora também possa ajudar no cancelamento dos efeitos do ruído aditivo. Já o J-RASTA é mais eficiente para a supressão do ruído aditivo. O *Log*-RASTA (Hermansky *et al.*, 1991) consiste em um filtro no domínio cepstral e classicamente tem a seguinte função de transferência, para aplicação de janelas a cada 10ms:

$$H(z) = 0{,}1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0{,}94z^{-1}} \tag{4.10}$$

A resposta em frequência do filtro *Log*-RASTA pode ser observada na Fig. 4.2. A freqüência de corte inferior é de 0,26 Hz. O declive do filtro a partir de 12,8 Hz é de 6 dB/oitava com um zero agudo em 18,9 Hz e 50Hz.

De forma a tornar a análise da voz menos sensível às variações lentas ou a componente contínua do logaritmo do espectro da voz, na filtragem *Log*-RASTA cada coeficiente cepstral é filtrado com um filtro com um zero espectral agudo na freqüência zero. Este efeito de filtragem passa alta elimina as baixas freqüências e alivia os efeitos do ruído convolucional cujas características variam lentamente, ou podem mesmo ser consideradas invariantes no tempo. Entretanto, ao salientar a faixa de variação cepstral que

representa a voz, a filtragem RASTA também suprime parte do ruído aditivo. Em relação à CMN, a filtragem RASTA tem a vantagem de poder ser aplicada a intervalos de sinais voz com duração variável embora exija uma carga computacional maior.

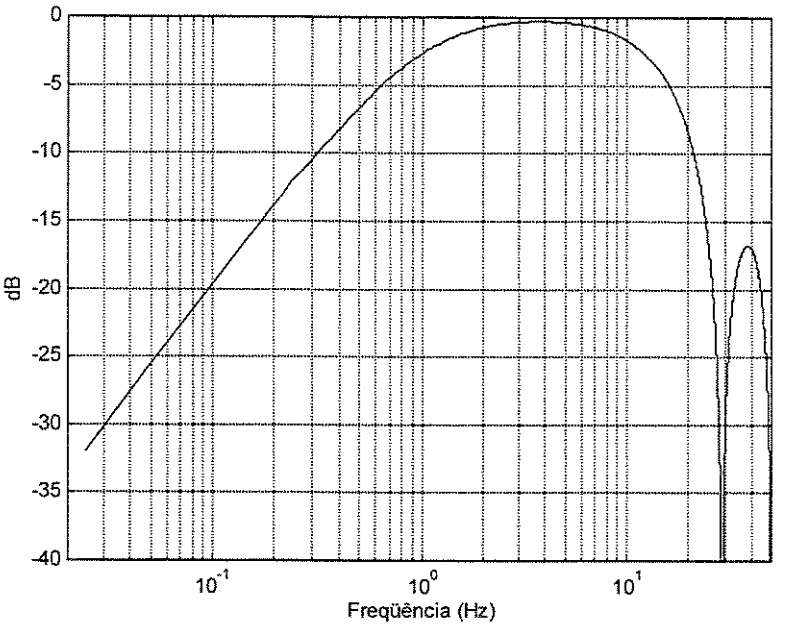

Figura 4.2 - Módulo da resposta em freqüência do filtro Log-RASTA

A filtragem *Log*-RASTA tem uma constante de tempo longa (160 ms para a Eq. 4.10) e, portanto, deve ser iniciada bem antes do início da elocução de forma a não distorcer o sinal de voz devido à resposta transitória. Nos experimentos que foram realizados neste trabalho o filtro *Log*-RASTA foi aplicado em toda a elocução a partir do silêncio que precede o sinal de voz (aproximadamente 500 ms em média).

O J-RASTA (Hermansky *et al.*, 1993) foi desenvolvido para cancelamento do ruído aditivo. Ele consiste num filtro no domínio espectral, que foi chamado de domínio espectral alternativo (Hermansky & Morgan, 1994). Este domínio espectral alternativo é linear para pequenos valores espectrais e logarítmico para os grandes. A equação que implementa o J-RASTA é a seguinte

$$y = \ln(1 + Jx) \tag{4.11}$$

onde J é uma constante positiva dependente do sinal e é ela que dá o nome a técnica. A variação na amplitude do espectro de energia x é linear para $J<<1$ e logarítmica para $J>>1$.

Existe um valor ótimo de J para cada SNR. A estimação de J pode ser feita da seguinte forma

$$J = \frac{1}{CE_{noise}} \tag{4.12}$$

onde E_{noise} é a energia média do ruído que pode ser medida nos intervalos de silêncio. C é uma constante definida experimentmente. Em (Hermansky & Morgan, 1994) utilizou-se quatro valores de C para o treinamento, 3000, 300, 30 e 3. Para os testes utilizou-se um C = 3. A grande desvantagem do J-RASTA é a dependência em relação ao parâmetro J para o qual não existe uma estimação analítica.

Naturalmente, RASTA deve ser aplicado tanto às elocuções de treinamento quanto as de teste, apenas aos parâmetros estáticos já que os dinâmicos são derivados a partir destes. Assim, como na CMN, RASTA também remove algumas características dependentes do texto e do locutor.

4.2.3 Subtração Espectral

A *subtração espectral* (SS – *Spectral Subtraction*) (Berouti *et al.*, 1979; Van Compernolle, 1989; Vasegui & Milner, 1997) é aplicada no cancelamento de ruído aditivo e consiste na estimação do espectro de energia do ruído $\widetilde{N}(\omega)$ e sua correspondente subtração do espectro de energia obtido em cada quadro do sinal corrompido por ruído $Y(\omega)$. Então, aplicando-se a SS teremos (Berouti *et al.*, 1979):

$$Y_{SS}(\omega) = \max\left\{Y(\omega) - \alpha\widetilde{N}(\omega), \beta Y(\omega)\right\} \tag{4.13}$$

onde α é o fator de sobre-estimação e β é o fator para o limite inferior. Há de se ressaltar que, na equação anterior, $\widetilde{N}(\omega)$, $Y(\omega)$ e $Y_{SS}(\omega)$ podem ser tanto a magnitude do espectro quanto o espectro de energia. O parâmetro α depende da SNR de acordo com a Fig. 4.3 e, conforme (Berouti *et. al*, 1979), pode ser obtido por

$$\alpha = \alpha_0 - \mu SNR \tag{4.14}$$

onde α_0 é o valor desejado de α para SNR = 0 dB e μ é a taxa de variação de α em função da SNR.

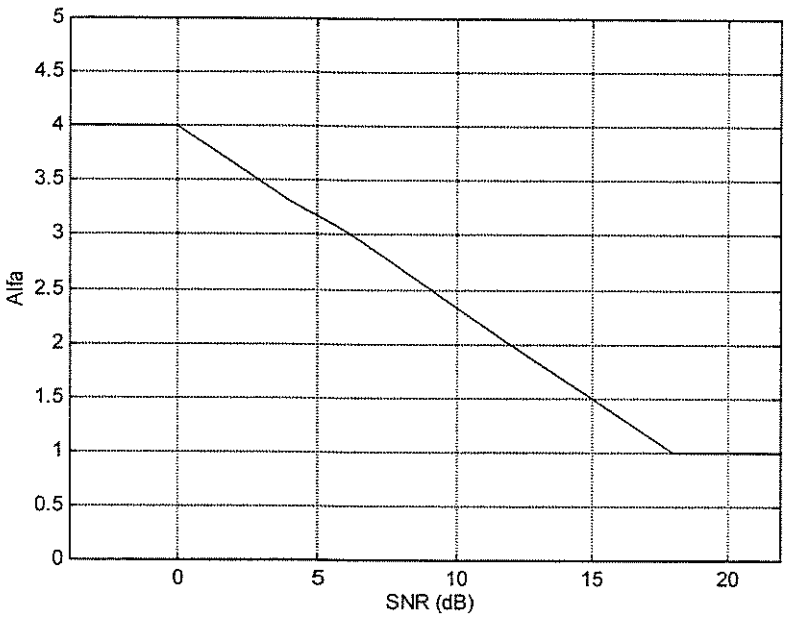

Figura 4.3 - Valor de α conforme a SNR

Uma estimação grosseira do espectro de energia do ruído $\widetilde{N}(\omega)$ pode ser feita pela média espectral dos quadros correspondentes ao silêncio tanto iniciais quanto no decorrer da elocução utilizando os quadros com menor energia (maior probabilidade de não haver sinal de voz). Há de se destacar que o ruído aditivo tem uma grande variabilidade temporal e uma melhor estimação do ruído pode ser significativa.

Para os experimentos deste trabalho β foi feita igual 0,25, α_0 igual a 4.0 e μ igual a 3/18. $\widetilde{N}(\omega)$ foi obtido pela média do espectro nos primeiros 10 quadros do sinal onde só havia ruído. O valor de SNR para cada quadro e raia espectral foi obtido pela relação entre $\widetilde{N}(\omega)$ e $Y(\omega)$.

4.2.4 Normalização da SNR

A normalização da relação sinal/ruído (Van Compernolle, 1989; Claes & Van Compernolle, 1996; Claes *et al.*, 1996) pode ser usada tanto para ruído aditivo quanto para o cancelamento da distorção convolucional. Permite a utilização em conjunto com técnicas

para cancelamento da distorção convolucional, como a filtragem RASTA (Claes & Compernolle, 1996), e cancelamento do ruído aditivo, como a SS (Van Compernolle, 1989). Neste método uma constante é adicionada ao valor final das energias dos bancos de filtro triangulares em escala mel no treinamento e no teste. A idéia é normalizar a SNR em cada banda de freqüência adaptando uma constante de acordo com a SNR medida ou a faixa dinâmica de cada banda. Isto faz com que os parâmetros extraídos sejam menos sensíveis ao nível de ruído, mas a influência da distorção do canal também é suprimida. A normalização da SNR é de fácil implementação embora precise de vários parâmetros para os quais não existe uma estimação analítica e são dependentes de cada caso.

4.2.5 Parallel Model Combination (PMC)

PMC (Gales & Young, 1993a, 1993b, 1996; Gales, 1997) baseia-se no fato de que a performance de sistemas de reconhecimento de voz é ótima quando não há descasamento entre condições de treinamento e de teste. Assim, é criado um HMM que representa o sinal de voz corrompido por ruído (veja Fig. 4.4). A técnica utiliza o HMM obtido no treinamento com sinal de voz limpo e um HMM que representa o ruído. Este HMM do ruído é semelhante ao modelo do sinal limpo, com a mesma distribuição de probabilidade de emissão, mas em geral com um número menor de estados. Para facilitar a combinação dos modelos do sinal limpo e do ruído, esta é usualmente feita no domínio espectral logarítmico utilizando uma função de descasamento para a combinação dos modelos. Após a adaptação no domínio logarítmico, o modelo estimado do sinal de voz corrompido por ruído é transformado novamente ao domínio cepstral. No PMC tanto os coeficientes estáticos quanto os dinâmicos (de primeira e segunda ordem) podem ser combinados com o modelo de ruído.

PMC fornece um bom desempenho e reduz significativamente a taxa de erro em sistemas de reconhecimento de voz (Gales & Young, 1993a, 1993b, 1996). Contudo, a estimação do modelo de ruído é crítica e envolve uma elevada carga computacional. A estimação do ruído aditivo não é difícil de ser feita, embora se suponha estacionariedade, o que não corresponde ao caso real. Já o ruído convolucional não é facilmente estimado e é geralmente desconhecido. Por outro lado, técnicas convencionais para cancelamento de

ruído convolucional, como CMN e RASTA, não podem ser aplicadas em combinação com PMC.

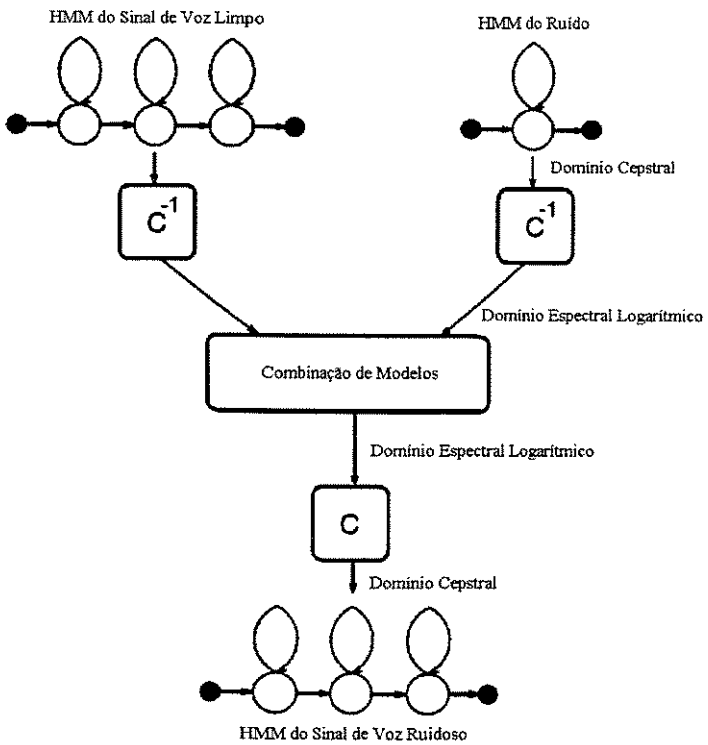

Figura 4.4 - Processo básico do PMC

4.2.6 Modelamento de Duração de Estados com Restrições Temporais

A probabilidade de transição de estados no HMM convencional é representada por uma constante que representa uma função densidade de probabilidade geométrica quanto a duração de estados. Na prática, esta probabilidade de transição tem pouca capacidade discriminatória quando comparada com as probabilidades de observação, e a distribuição geométrica não é um modelo apropriado para a probabilidade de duração de estados. Logo, a utilização de uma distribuição mais apurada quanto a duração de estados levaria a um melhor modelamento da palavra e/ou locutor e uma conseqüente melhora no desempenho do sistema, principalmente em condições ruidosas. Várias técnicas de robustez a ruído utilizando modelamento de duração de estados foram propostas para sistemas de reconhecimento de voz (Burshtein, 1996; Laurila, 1997; Yoma *et al.*, 2000).

Na técnica sugerida em (Yoma *et al.*, 2000), é utilizada a função densidade de probabilidade gamma e um truncamento mínimo e máximo para o modelamento de duração

de estados. Este modelamento é incluído no algoritmo de Viterbi por meio da seguinte probabilidade de transição condicional:

$$a_{i,j}^{(\tau)} = Prob(s_{t+1} = j \mid s_t = s_{t-1} = \ldots = s_{t-\tau+1} = i) \tag{4.15}$$

onde τ é a duração do estado i no quadro t. Define-se $d_i(\tau)$ como sendo o valor de probabilidade da duração do estado i ser igual a τ, e $D_i(\tau)$ a probabilidade do estado i estar ativo para $t \geq \tau$, ou seja,

$$D_i(\tau) = \sum_{t=\tau}^{\infty} d_i(t) \tag{4.16}$$

Usando-se a definição de probabilidade de transição da Eq. 4.15, e impondo-se uma duração mínima e uma duração máxima para cada estado, teremos as seguintes probabilidades de transição de estado

$$a_{i,i}^{\tau} = \begin{cases} 1 & \text{se } \tau < t_{\min_i} \\ 0 & \text{se } \tau \geq t_{\max_i} \\ 1 - \frac{d_i(\tau)}{D_i(\tau)} & \text{caso contrário} \end{cases} \tag{4.17}$$

$$a_{i,i+1}^{\tau} = \begin{cases} 0 & \text{se } \tau < t_{\min_i} \\ 1 & \text{se } \tau \geq t_{\max_i} \\ \frac{d_i(\tau)}{D_i(\tau)} & \text{caso contrário} \end{cases} \tag{4.18}$$

onde

$$\begin{cases} t_{min_i} = tol_{min} min_i(\tau) \\ t_{max_i} = tol_{max} max_i(\tau) \end{cases}$$

sendo que as constantes tol_{min} e tol_{max} introduzem uma tolerância na estimação dos valores de duração mínima e máxima para cada estado. A distribuição de probabilidade $d_i(\tau)$ é aproximada por uma distribuição gamma discreta dada por

$$d_i(\tau) = Ke^{-\alpha\tau}\tau^{p-1} \tag{4.19}$$

onde τ é um valor inteiro positivo, $\alpha > 0, p > 0$ e K é um termo de normalização. A média ($E_i(\tau)$), a variância ($Var_i(\tau)$), e os valores mínimo e máximo de duração de estados são estimados computando-se o algoritmo de Viterbi após o treinamento dos HMM, e obtendo a seqüência de estados ótima para cada elocução de treinamento. Os parâmetros α e p são estimados utilizando-se a média e variância por meio de:

$$\alpha_i = \frac{E_i(\tau)}{Var_i(\tau)} \tag{4.20}$$

$$p_i = \frac{E_i^2(\tau)}{Var_i(\tau)} \tag{4.21}$$

Em (Yoma *et al.*, 2000) é sugerida uma forma alternativa em que se conservam as probabilidades de transição clássicas $a_{i,i}$ e $a_{i,i+1}$ e se utiliza uma função de densidade de probabilidade truncada para modelar a duração de estados. Assim, as probabilidades de transição são modificadas para

$$a_{i,i}^{\tau} = \begin{cases} 1 & \text{se } \tau < t_{\min_i} \\ 0 & \text{se } \tau \geq t_{\max_i} \\ a_{i,i} & \text{caso contrário} \end{cases} \tag{4.22}$$

$$a_{i,i+1}^{\tau} = \begin{cases} 0 & \text{se } \tau < t_{\min_i} \\ 1 & \text{se } \tau \geq t_{\max_i} \\ a_{i,i+1} & \text{caso contrário} \end{cases} \tag{4.23}$$

Em (Yoma *et al.*, 2000) ainda é sugerido um modelamento de duração de estados dependente do contexto. Foi verificado uma sensível melhora no desempenho de um sistema de reconhecimento de três dígitos concatenados o qual empregou modelos de duração de estados para cada posição de cada dígito (primeiro, segundo ou terceiro dígito). Isto sugere que o efeito de coarticulação influencia bastante na duração de cada estado; para o primeiro dígito não há o efeito de coarticulação inicial e ele sofre influência apenas do segundo; o segundo sofre influência tanto do primeiro quanto do terceiro; já o terceiro é influenciado pela elocução do anterior.

Os detalhes da implementação de alguns destes algoritmos de robustez (CMN, *log*-RASTA, SS e modelamento de duração de estados com restrições temporais) para o SVL deste trabalho estarão no capítulo 5. Também serão feitos testes empregando-se ruído convolucional e/ou aditivo através da base de dados de ruído NOISEX-92.

Capítulo 5

Implementação das Técnicas de Robustez e Resultados com Ruído

5.1 INTRODUÇÃO

Neste capítulo descrever-se-á a implementação no SVL de algumas das técnicas de robustez comentadas no capítulo anterior. São elas: a CMN, *log*-RASTA, SS e modelamento de duração de estados com restrições temporais. As condições de ruído serão simuladas artificialmente empregando-se para isso um filtro e ruídos da base de dados NOISEX-92, representando respectivamente o ruído convolucional e o aditivo. Poder-se-ão verificar a queda no desempenho do sistema em condições ruidosas e o aumento na robustez provocada pelo emprego das técnicas mencionadas anteriormente, e avaliar o resultado da interação da combinação de mais de um destes métodos.

5.2 SIMULAÇÃO E RESULTADOS COM RUÍDO

Diante da dificuldade em se obter gravações com ruído e sinal limpo gravados simultaneamente, geralmente o ruído é acrescentado artificialmente ao sinal livre de ruído. Isto simula com boa precisão o caso real, com exceção do efeito Lombard causado pelo estresse na voz do locutor em situações com ruído muito alto. Assim, neste trabalho tanto o ruído convolucional quanto o aditivo foram introduzidos no sinal de voz limpo sendo que os resultados obtidos aqui apresentados podem ser considerados uma boa avaliação das técnicas estudadas. Contudo, há de se destacar que o acréscimo do ruído se deu após a detecção de início e fim da elocução, sendo esta realizada com o sinal limpo.

5.2.1 Ruído Convolucional

Para simular um ruído convolucional, as 40 elocuções de teste de cada locutor passaram por um filtro com a resposta em freqüência da Fig. 5.1.

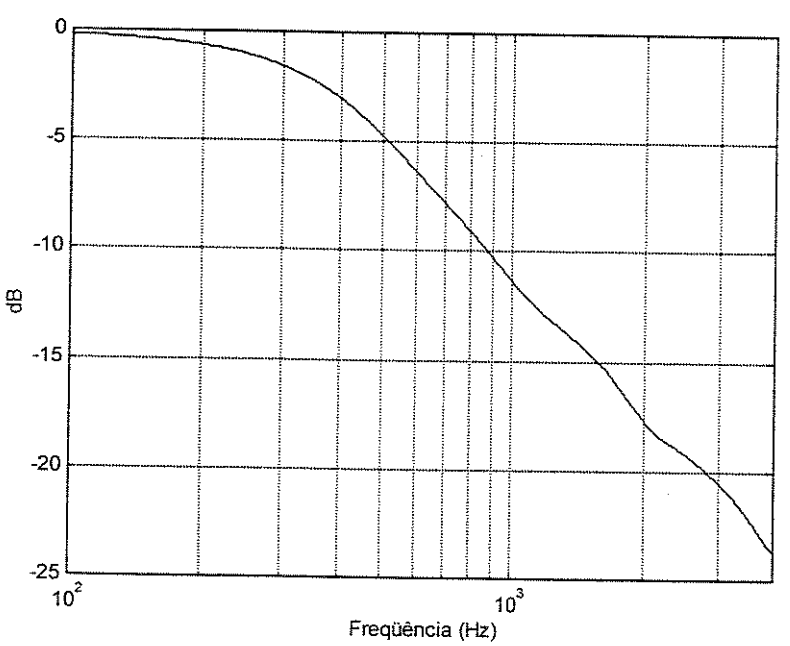

Figura 5.1 - Resposta em freqüência do filtro que simula o ruído convolucional

Com as elocuções de teste com ruído convolucional e não se utilizando nenhuma técnica de robustez, o SVL apresentou uma EER_{SS} de 1,54% e uma EER_{SI} de 8,83%. Isto representa um aumento na EER de cerca de 300% quando considerados limiares de decisão

individuais por locutor e de cerca de 800% quando considerado um limiar único. A Fig. 5.2 mostra o desempenho do SVL com sinal limpo e com ruído convolucional. Nela verifica-se a grande queda de desempenho do sistema quando há presença do ruído convolucional, considerando-se um limiar de decisão único para todos os locutores.

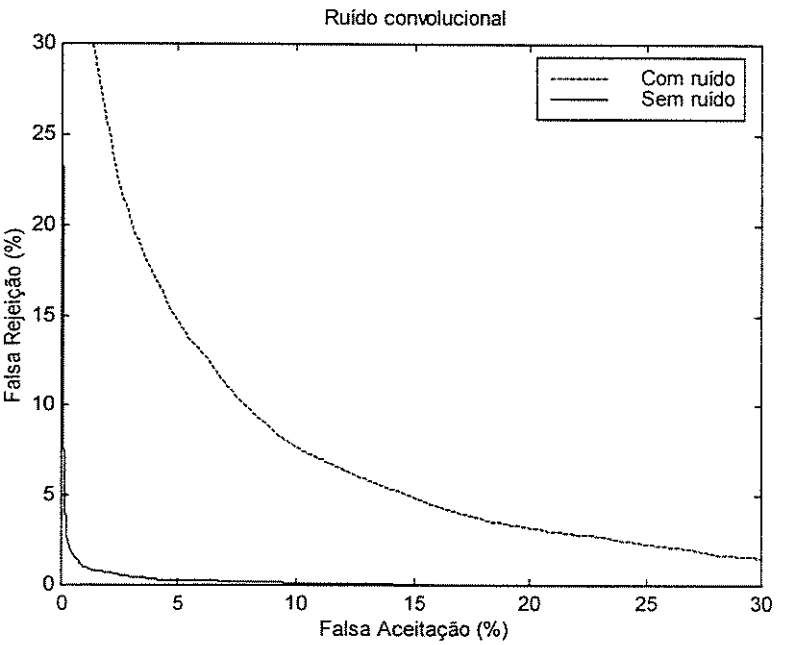

Figura 5.2 - ROC comparativa do SVL para sinal de voz limpo e com ruído convolucional considerando-se um limiar de decisão único

5.2.2 Ruído Aditivo

Para os testes com ruído aditivo, foram utilizados os ruídos de carro e de fala da NOISEX-92 em 4 relações sinal ruído diferentes, 0, 6, 12 e 18dB. O sinal de ruído de carro da NOISEX-92 foi gravado dentro de um Volvo em movimento e consiste em cerca de 4 minutos de gravação. O ruído de fala (cerca de 4 minutos de duração) consiste de ruído branco processado com um filtro passa baixa que modela o espectro médio da soma dos sinais provenientes das vozes de várias pessoas nas proximidades do microfone. O sinal de ruído foi filtrado para eliminar as freqüências inferiores a 300 Hz de forma a limitar a energia na banda de interesse entre 300 e 3400 Hz. O espectro médio dos sinais de ruído pode ser observado na Fig. 5.3.

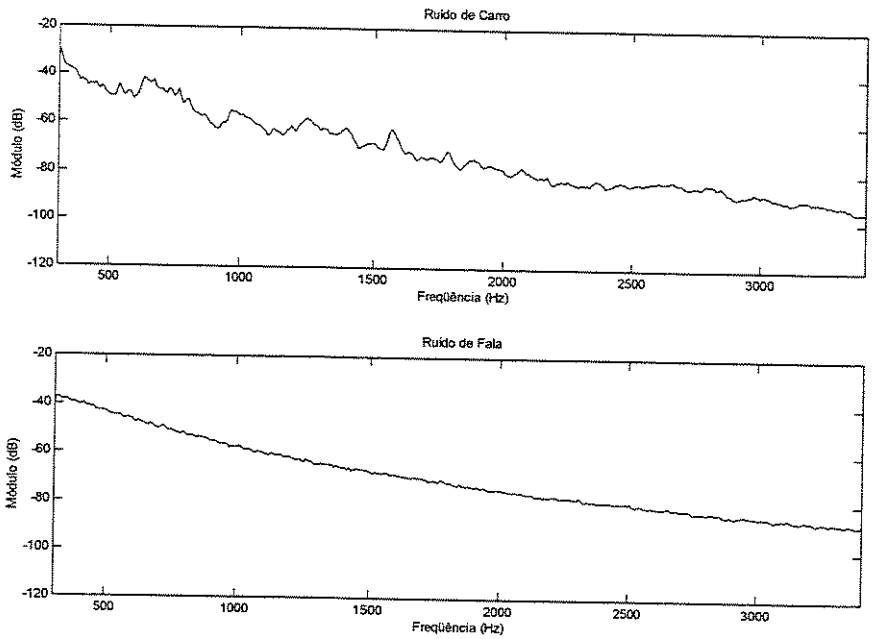

Figura 5.3 - Módulo médio do espectro dos sinais de ruído de carro e fala da NOISEX-92

A adição do ruído às elocuções foi feita em seqüência, de forma que a amostra inicial de ruído a ser adicionada a uma determinada elocução é imediatamente posterior à última amostra adicionada à elocução anterior. Para a adição do ruído às elocuções de teste, calculou-se a energia total de cada elocução (após a segmentação de início e fim) para estimar-se a energia do ruído em função da SNR desejada. Com as elocuções de teste corrompidos por ruído em várias SNR, verificou-se o desempenho do SVL. Os resultados dos testes são mostrados na Tab. 5.I e nas Fig. 5.4 e 5.5. Como pode ser observado, o ruído aditivo deteriora bastante o desempenho do SVL e representa um problema mais complexo até mesmo que o do ruído convolucional. Enquanto o efeito do ruído convolucional leva a um erro de 1,54%, o ruído aditivo a 0dB faz o sistema ter um erro de até 23,58%.

Tabela 5.I - Taxa de erros iguais com ruído aditivo

SNR (dB)	EER_{SS} (%)	
	Ruído de Carro	Ruído de Fala
0	22,9	23,58
6	6,34	6,77
12	1,80	1,88
18	0,68	0,68

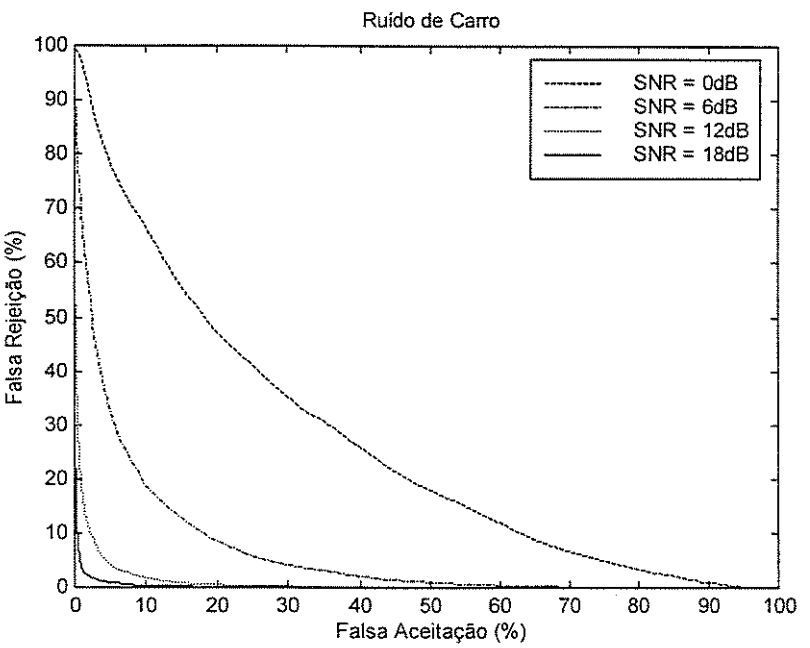

Figura 5.4 - ROC para o SVL com sinal de voz corrompido por ruído de carro em 0, 6, 12 e 18dB

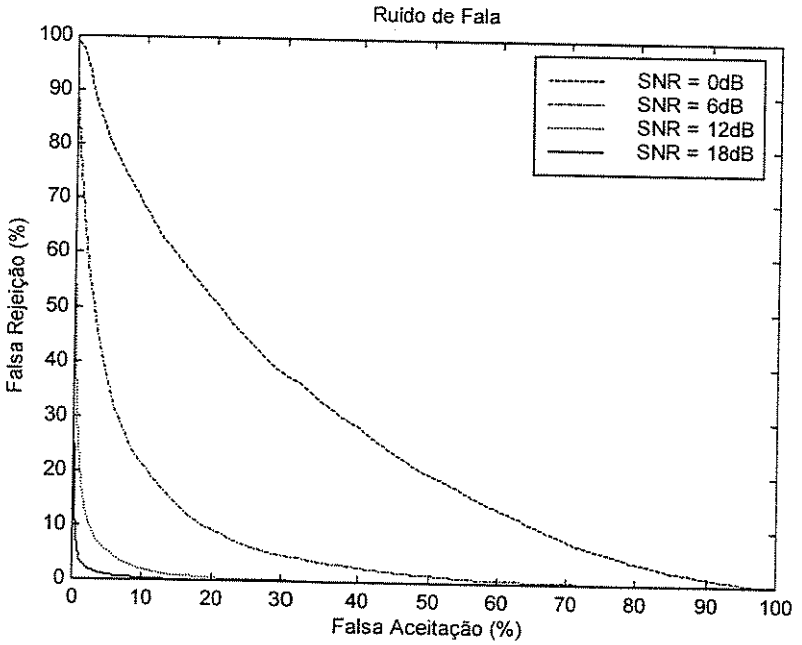

Figura 5.5 - ROC para o SVL com sinal de voz corrompido por ruído de fala em 0, 6, 12 e 18dB

5.2.3 Ruído aditivo e convolucional

Na maioria das situações reais há presença tanto do ruído aditivo quanto do convolucional. Para a simulação conjunta destes, primeiro foi feita a adição do ruído aditivo a cada elocução e a seguir todas elas foram filtradas pelo filtro que simula o ruído convolucional. Isto é mais ou menos o que acontece com o sinal de voz proveniente de um telefone, onde o locutor está em um ambiente ruidoso e o sinal sofre distorção tanto do microfone responsável pela transdução quanto do canal telefônico. Os resultados com sinal corrompido por ruído aditivo e convolucional obtidos com o SVL sem nenhuma técnica de robustez são mostrados na Tab. 5.II e nas Fig. 5.6 e 5.7.

Tabela 5.II - Taxa de erros iguais com sinal corrompido por ruído aditivo e convolucional

SNR (ruído aditivo) (dB)	EER_{SS} (%)	
	Ruído de Carro e convolucional	Ruído de Fala e convolucional
0	30,00	29,81
6	13,79	13,46
12	5,94	6,16
18	3,42	3,41

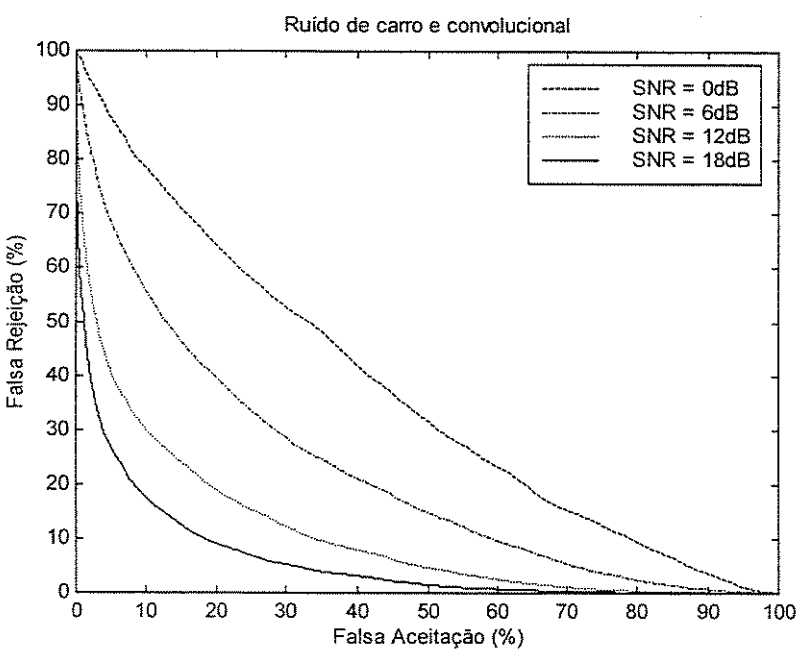

Figura 5.6 - ROC para o SVL com sinal de voz corrompido por ruído de carro a uma SNR de 0, 6, 12 e 18dB e ruído convolucional

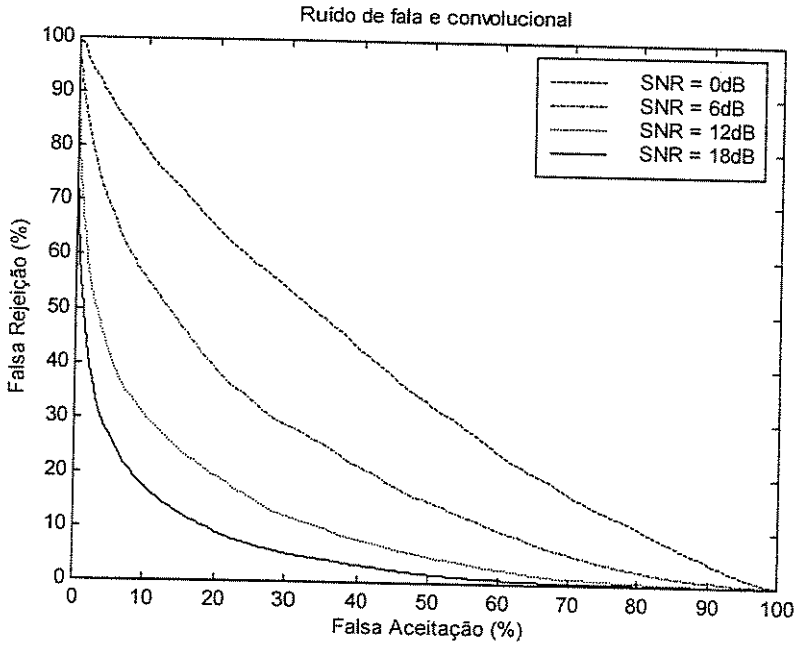

Figura 5.7 - ROC para o SVL com sinal de voz corrompido por ruído de fala a uma SNR de 0, 6, 12 e 18dB e ruído convolucional

Conjuntamente, o ruído aditivo e o convolucional representam um problema ainda maior do que cada um isoladamente. Nos testes realizados com ambos os ruídos, conforme a Tab. 5.II, o SVL chegou a apresentar 30% de EER_{SS}. Conforme verificado nos testes com os ruídos individuais, o ruído convolucional não é um grande problema quando aparece isoladamente; no entanto, prejudica em muito o sistema que já sofre danos devido ao ruído aditivo. Quando considerado somente o ruído aditivo, a pior situação levou a uma EER_{SS} de 23,58% enquanto que acrescentando-se o ruído convolucional houve um acréscimo de 6,42% na EER_{SS}.

5.3 IMPLEMENTAÇÃO E RESULTADOS DAS TÉCNICAS DE CANCELAMENTO DE RUÍDO

5.3.1 CMN

Para aplicar-se a CMN em uma dada elocução, calculou-se a média temporal de cada coeficiente cepstral inclusive utilizando os quadros de silêncio iniciais e finais. A

seguir, cada valor temporal de coeficiente cepstral foi subtraído da sua correspondente média. A CMN foi realizada tanto nas elocuções de teste quanto nas de treinamento. Em experimentos com sinal limpo, verificou-se uma EER_{SS} de 0,41%; o que representa um aumento de 14% na EER_{SS} do SVL. Isto mostra que a CMN pode também deteriorar o desempenho do SVL com sinal limpo, embora seja um aumento na taxa de erros pouco significativo.

Aplicando-se o filtro que simula o ruído convolucional às elocuções de teste, a CMN levou a uma EER_{SS} de 0,42% e uma EER_{SI} de 1,23%. Isto representa uma melhora de 95% no desempenho do SVL com ruído convolucional, quando considerado limiares de decisão distintos para cada locutor. A CMN elimina quase que totalmente o ruído convolucional, restando apenas a distorção provocada pela própria CMN.

5.3.2 *Log*-RASTA

Para cada elocução, a seqüência temporal de cada coeficiente temporal foi filtrada pelo filtro RASTA da Eq. 4.8. De forma a minimizar os efeitos da resposta impulsiva do filtro, a filtragem foi realizada desde os quadros iniciais de silêncio que antecedem a elocução (que varia de 300 a 500 ms). Tanto os coeficientes das elocuções de teste quanto os coeficientes das elocuções de treinamento foram filtrados. Em condições livre de ruído, o SVL com *Log*-RASTA apresentou uma EER_{SS} de 0,47%; cerca de 31% de aumento em relação ao sistema sem técnica de robustez. Assim, o decréscimo no desempenho verificado em SRV (Hermansky *et. al.*, 1991; Hermansky *et. al.*, 1993; Hermansky & Morgan, 1994) acaba se confirmando também para SVL mas, assim como o do CMN, é uma piora pouco significativa.

Com ruído convolucional e *Log*-RASTA, o SVL teve uma EER_{SS} de 0,50% e uma EER_{SI} de 1,12%. Considerando-se um limiar de decisão individual por locutor, isto representa uma melhora de 88% no desempenho do sistema. Assim como o CMN, o *Log*-RASTA elimina quase que totalmente o ruído convolucional e o decréscimo no desempenho em relação ao obtido em condições ideais deve-se mais à distorção provocada pelo filtro *Log*-RASTA.

A Fig. 5.8 ilustra a eficiência da CMN e do *Log*-RASTA no cancelamento do ruído convolucional. Considerando-se um limiar de decisão único para o SVL, através das características de operação do receptor verifica-se a grande distorção provocada pelo ruído convolucional e a melhora significativa proporcionada pelas técnicas de robustez empregadas. Tanto CMN quanto *Log*-RASTA levam o SVL a um desempenho muito próximo ao obtido com sinal de voz limpo.

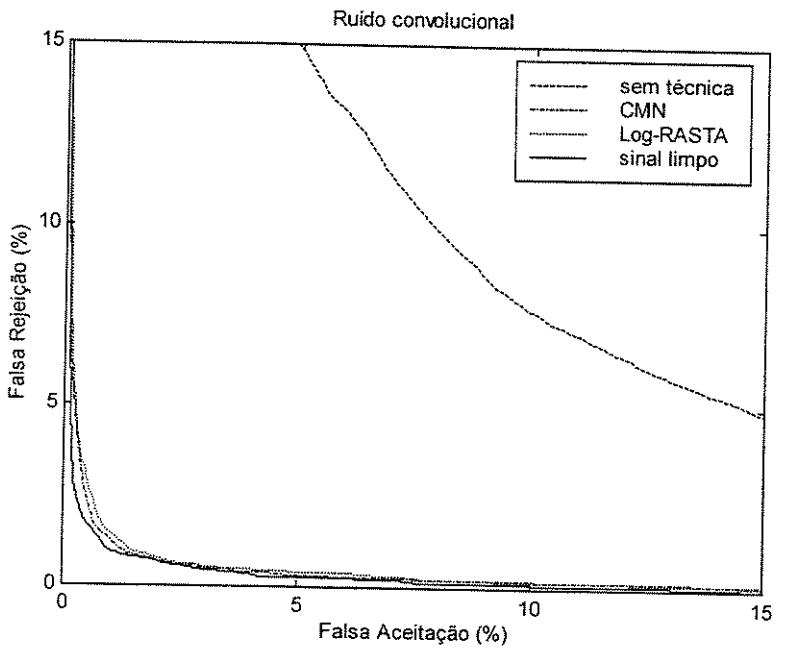

Figura 5.8 - Gráfico comparativo entre CMN e Log-RASTA para SVL com sinal de voz corrompido por ruído convolucional

5.3.3 SS

A SS foi realizada em cada um dos 128 pontos do espectro de energia de cada quadro do sinal de voz. Sejam $Y(\omega)$ e $\widetilde{N}(\omega)$ o espectro de energia num dado quadro do sinal ruidoso e o espectro de energia médio do ruído, respectivamente. $\widetilde{N}(\omega)$ foi obtido pela média do espectro nos primeiros 10 quadros do sinal os quais antecedem a elocução e são constituídos só por ruído. O valor de SNR para cada quadro foi obtido pela relação entre $\widetilde{N}(\omega)$ e $Y(\omega)$. Obtêve-se experimentalmente o melhor valor de α, conforme a SNR, e construiu-se o gráfico da Fig. 4.3. A partir dele considerou-se $\alpha_0 = 4$ e $\mu = 3/18$. β

também foi obtido experimentalmente e seu valor foi considerado igual a 0,25 para os experimentos deste trabalho.

Os resultados dos testes com ruído aditivo e SS podem ser avaliados através da Tab. 5.III. A SS melhorou a robustez do sistema ao ruído aditivo de 28 à 44%. Embora seja uma melhora significativa mostrando que a SS pode ser empregada com sucesso também em SVL, a subtração espectral por si só não resolve o problema do ruído aditivo. No entanto, pode ser utilizada com outras técnicas de robustez, como será descrito no item 5.4.

Tabela 5.III - Desempenho do SVL com SS e sinal corrompido por ruído aditivo

SNR (dB)	EER_{SS} com SS (%)	
	Ruído de Carro	Ruído de Fala
0	13,12	13,87
6	3,70	3,93
12	2,24	1,23
18	0,59	0,55

O ruído aditivo não foi totalmente cancelado pela SS principalmente pelo fato de que ele não é estacionário. Assim sendo, o espectro do ruído estimado nos quadros iniciais não é completamente igual ao espectro o ruído que corrompe o sinal de voz nos outros quadros. Para que a SS tenha um melhor efeito seria necessária um melhor estimação temporal do espectro do ruído aditivo.

A subtração do espectro do ruído apenas em módulo (desconsiderando a fase) é outro fator que faz com que a SS não cancele totalmente o ruído aditivo.

5.3.4 Modelamento de duração de estados com restrições temporais

Foram feitos testes com a distribuição gamma com e sem truncamento e com a distribuição geométrica clássica com truncamento simples. Para cada um dos testes com a distribuição gamma sem truncamento foram utilizadas as seguintes tolerância na duração de

estados: $tol_{min} = 0$ e $tol_{max} = 4$ com um mínimo de duração de 1 quadro por estado (isto é, praticamente sem restrição de duração máxima e duração mínima). A intenção ao se estender o intervalo de atuação da função gamma é o de modelar a duração de estados para todos os caminhos do algoritmo de Viterbi os quais têm uma mínima possibilidade de serem o ótimo, evitando o truncamento.

Para o modelamento com truncamento simples, considerou-se $tol_{min} = 0,8$ e $tol_{max} = 1,5$ e um mínimo para o valor do logaritmo da probabilidade de transição de estado. Este valor é igual a -10 e corresponde a uma probabilidade de cerca de 4.10^{-5}, tendo sido obtido experimentalmente. Este mínimo de probabilidade foi considerado de forma a não punir exageradamente o caminho, no alinhamento de Viterbi, o qual possui valores de duração de estados menor que a mínima ou maior que a máxima. Embora a variabilidade na duração de estados de uma determinada palavra de um locutor específico seja menor que a variabilidade da população, aquela também é bastante variável especialmente nos primeiros e últimos estados em função dos intervalos de silêncio intra-elocução. Consequentemente, é preciso evitar que a ocorrência de durações extremas levem à rejeição do locutor verdadeiro, já que sendo a probabilidade de emissão igual a zero o logaritmo da verossimilhança seria $-\infty$.

Da mesma forma que no truncamento simples, foram feitos testes com a função gamma truncada com $tol_{min} = 0,8$ e $tol_{max} = 1,5$ e um mínimo para a probabilidade de transição de estado de 4.10^{-5}.

5.3.4.1 Modelamento de duração de estados para o termo de normalização

Foram considerados quatro tipos de modelamento de duração de estados para o modelo global do termo de normalização. São os seguintes:

1. *Modelo global de duração de estados* (MGDE): O modelamento de duração de estados é obtido para cada palavra independentemente do locutor. Através do algoritmo de Viterbi e do HMM global, são obtidos a média, a variância, duração mínima e máxima de cada estado de cada palavra empregando-se todas as elocuções de treinamento de todos os locutores;

2. *Modelo individual de duração de estados* (MIDE): O modelo de duração de estados para o termo de normalização é obtido para cada locutor de forma distinta. Utilizando-se do HMM global, são obtidos os parâmetros de duração para cada locutor empregando-se todas as elocuções de treinamento do locutor em questão;
3. *Modelo de duração de estados individual em relação ao HMM individual* (modelo único): O modelo de duração de estados é o mesmo que o empregado para o cálculo da verossimilhança da elocução em relação ao HMM individual do locutor. Através do HMM individual, são obtidos os parâmetros de duração para cada locutor empregando-se todas as elocuções de treinamento do locutor em questão;
4. *Sem modelo de duração de estados*.

Para testar qual dos tipos de modelamento seria mais adequado, foram feitos testes utilizando-se sinal limpo e sinal de voz corrompido por ruído de carro da NOISEX-92. Para os testes utilizou-se uma variância mínima igual a 3 para a distribuição gamma sem truncamento, e modelos de duração dependentes de contexto. O resultado dos testes pode ser verificado na Fig. 5.9 para sinal limpo e na 5.10 para ruído de carro a 0, 6 12 e 18dB, conforme numeração do texto.

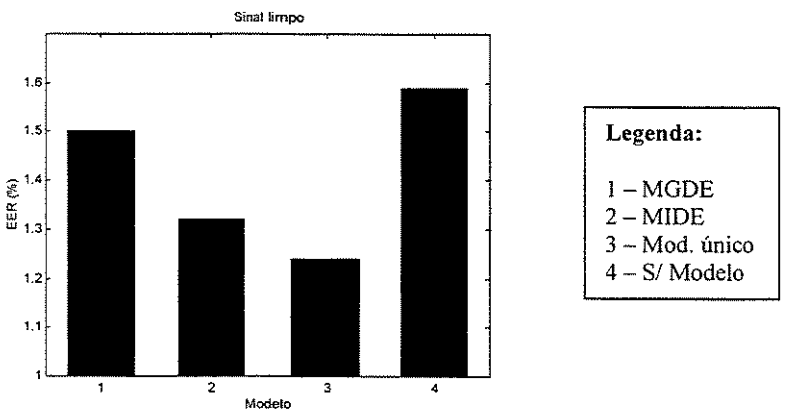

Figura 5.9 - Teste de modelamento de RT para o termo de normalização com sinal limpo, considerando-se um limiar de decisão único para toda a população

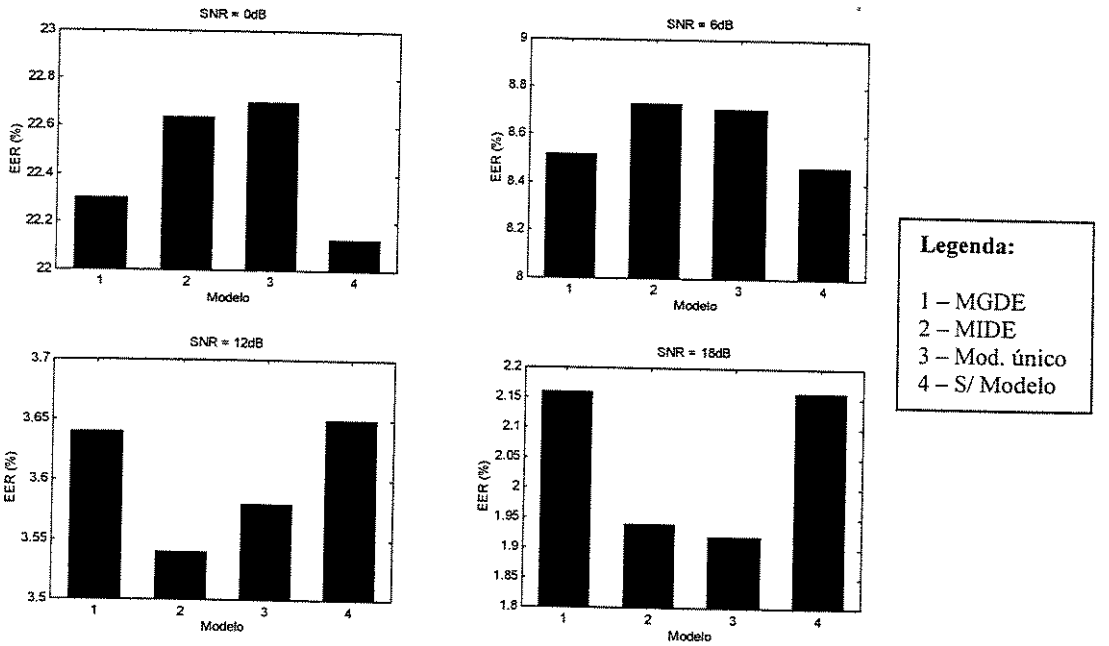

Figura 5.10 - Teste de modelamento de RT para o termo de norm. com ruído de carro, considerando-se um limiar de decisão único para toda a população

Quanto maior a SNR, melhor é o desempenho do modelo de duração de estados individual em relação ao HMM individual (número 3) para o termo de normalização. Quanto menor a SNR, melhor é o desempenho do SVL sem modelamento de duração de estados no termo de normalização. No entanto, a diferença de desempenho do sistema com os quatro modelos é muito pouco significativa, mostrando que o modelamento de duração de estados para o termo de normalização não é muito discriminativo. Isto pode ser justificado pela idéia de que o locutor individualmente fala de uma forma característica e tem uma duração de estados particular mas a população em geral fornece uma variabilidade muito grande no que se refere a duração de estados. Em (Yoma *et al.*, 2000), verificou-se que o modelamento de duração de estados para SRV dependente de locutor dava bons resultados em condições de ruído; no entanto, em SRV independente de locutor, o modelamento de duração de estados fornecia uma menor redução na taxa de erro.

Mesmo assim, por oferecer uma menor distorção a uma alta SNR, escolheu-se o mesmo modelo de duração de estados do cálculo da verossimilhança individual para o termo de normalização (modelo número 3). Todos os experimentos descritos nos itens seguintes serão realizados com este modelo.

5.3.4.2 Variância mínima

Para uma boa estimação da variância da duração de cada estado é necessário uma grande quantidade de elocuções de treinamento. Como não é possível ter-se infinitas elocuções de treinamento, é necessário limitar-se a variância a um valor mínimo de forma a não se ter uma variabilidade subestimada. Assim, utilizando-se o modelo número 3, fez-se testes com duração de estados utilizando-se uma modelamento com a função gamma sem truncamento com vários valores mínimos de variância. Os valores obtidos para a EER, considerando-se um limiar de decisão único e modelos de duração dependentes de contexto são ilustrados na Fig. 5.11.

Pela Fig. 5.11, verifica-se que o melhor desempenho com sinal limpo é conseguido com um limite inferior para a variância em torno de 3. Assim, todos os testes nos itens seguintes que utilizam modelamento de duração de estados com a distribuição gamma serão realizados com este limite inferior para a variância .

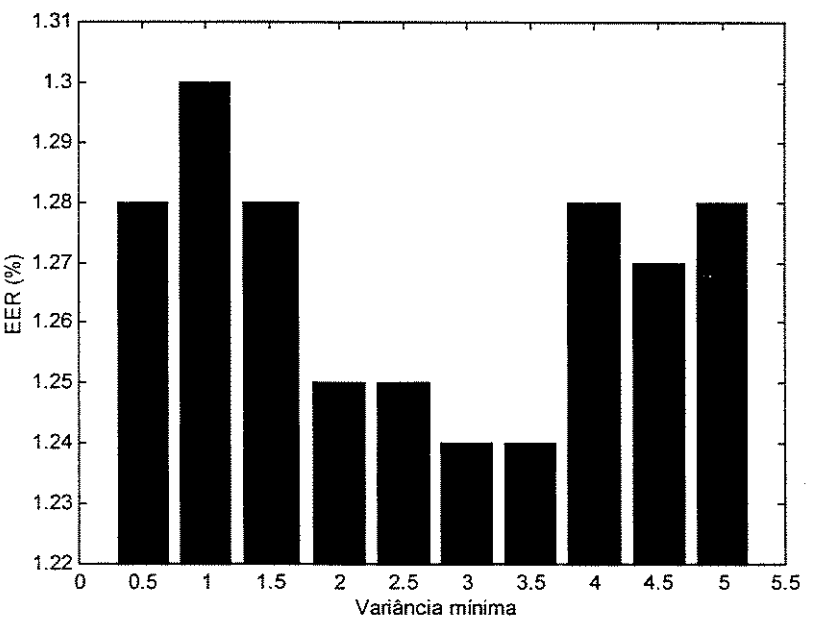

Figura 5.11 - Testes com variância mínima

5.3.4.3 Modelamento de duração de estados dependente de contexto e independente de contexto

Como comentado no capítulo 4, em (Yoma *et al.*, 2000), verificou-se uma dependência do desempenho do SRV em condições ruidosas quanto ao contexto. Assim, também foram realizados experimentos para o SVL em questão para condições dependentes e independentes de contexto. Como os experimentos são realizados com a base de dados YOHO, na qual as elocuções gravadas são compostas por três dezenas distintas, optou-se por dois modelos de posição por palavra no caso dependente de contexto. As décadas podem encontrar-se no início ou no meio da elocução e, assim, criou-se um modelo de duração para cada década inicial e outro para cada década do meio. Da mesma forma, existem dois modelos de duração para cada unidade, o modelo para a unidade que se encontra no meio da elocução e outro para a que se encontra no final. Assim, no caso dependente de contexto há 32 modelos de duração de estados (16 modelos de palavras por posição) e no independente de contexto os 16 modelos de duração, já que não se considera a posição.

Tabela 5.IV - Testes com do modelo de duração de estados dependente e independente de contexto

SNR (dB)	EER_{SI} (%)	
	Dependente de contexto	Independente de contexto
0	22,70	22,99
6	8,71	8,72
12	3,65	3,44
18	1,92	2,09
Sinal limpo	1,24	1,23

Foram realizados experimentos com a distribuição gamma sem truncamento e SS para modelos de duração dependentes e independentes de contexto. Utilizou-se para tanto,

ruído de carro e sinal limpo. Os valores de EER_{SS} para ambos os casos é mostrado na Tab. 5.IV.

Em geral, o modelo dependente de contexto ofereceu um desempenho melhor que o independente de contexto, embora a diferença seja pouco significativa. Isto pode se dever ao fato de que os locutores geralmente falam pausadamente entre a elocução de cada dezena, introduzindo um intervalo de silêncio e diminuindo o efeito de coarticulação de uma dezena sobre a outra. Mesmo assim, optou-se por utilizar um modelo dependente de contexto para os experimentos com modelamento de duração de estados, descritos nos itens seguintes.

5.3.4.4 Resultados com sinal limpo

A Tab. 5.V mostra comparativamente o desempenho do SVL com sinal limpo para as diversas técnicas de modelamento de duração de estados. Somente a técnica com a distribuição gamma sem truncamento forneceu um desempenho inferior ao sem técnica de robustez alguma, o que pode levar a dizer que a distribuição gamma sem truncamento não é um bom modelo para a duração de estados.

Tabela 5.V - Desempenho do SVL com modelamento de duração de estados com sinal limpo

Técnica	EER_{SS} (%)	EER_{SI} (%)
Gamma s/ truncamento	0,63	1,24
Gamma c/ truncamento	0,33	1,03
Truncamento simples	0,35	1,03
Nenhuma	0,36	0,96

5.3.4.5 Resultados individuais com ruído

Na Tab. 5.VI podemos verificar o desempenho fornecido pelo SVL com elocuções de teste corrompidas por ruído aditivo empregando-se modelamento de duração de estados.

Tabela 5.VI - Desempenho do SVL com modelamento de duração de estados e sinal corrompido por ruído aditivo

Tipo	SNR (dB)	EER_{SS} (%)		
		Gamma s/ truncamento	Gamma c/ truncamento	Truncamento simples
Carro	0	18,76	16,18	16,18
	6	6,10	5,44	5,45
	12	2,05	1,86	1,87
	18	0,91	0,68	0,66
Fala	0	20,41	17,85	18,05
	6	6,68	6,11	6,08
	12	2,30	2,16	2,21
	18	1,00	0,78	0,76

Comparando-se os valores da Tab. 5.VI com os obtidos com o sistema sem técnica de robustez (Tab. 5.I), verifica-se uma melhora de até 30% com ruído de carro a 0dB e com truncamento (distribuição gamma ou geométrica). Para uma SNR mais alta (12 e 18 dB), a melhora no desempenho é pouco significativa com truncamento (gamma e geométrica), sendo que a distribuição gamma sem truncamento parece introduzir uma certa distorção. Como dito anteriormente, a gamma sem truncamento pode não modelar adequadamente a duração de estados. O modelamento de duração de estados com distribuição gamma truncada e com distribuição geométrica truncada forneceram um desempenho equivalente. Assim, pode-se afirmar que, para o caso de verificação de locutor dependente de texto, a utilização de truncamento simples é um melhor modelo de duração de estados até mesmo que a própria distribuição gamma. O modelamento de duração de estados não é apropriado para ser empregado individualmente em SNR elevadas pois não leva a uma redução importante na taxa de erro em SNR´s altas. Esta redução não muito significativa na taxa de erro com o emprego de MDE pode ser justificada pelos fatos da base de dados de teste estar

muito bem casada com a de treinamento (depedendente do locutor e do texto), e desta técnica não ser propriamente um método de cancelamento de ruído, como SS. Consequentemente, os efeitos do ruído sobre o sistema podem ser minimizados mas não eliminados totalmente. Sua utilização é mais apropriada em conjunto com outras técnicas de robustez, como SS, como será experimentalmente mostrado no item 5.4.

5.4 RESULTADOS COM COMBINAÇÃO DE TÉCNICAS

5.4.1 Ruído Convolucional

De forma a verificar o desempenho do SVL com modelamento de duração de estados, utilizou-se modelamento com a distribuição gamma e geométrica com truncamento em combinação com o filtro *Log*-RASTA. Como este modifica o caminho ótimo do HMM, obteve-se um novo modelo de duração de estados para cada locutor utilizando todas as elocuções de treinamento após a filtragem *Log*-RASTA. Os resultados dos experimentos podem ser observados na Tab. 5.VII.

Tabela 5.VII - Taxa de erros iguais com ruído convolucional

Técnica	EER_{SS} (%)	EER_{SI} (%)
nenhuma	1,54	8,83
Log-RASTA	0,50	1,12
Log-RASTA e MDE c/ gamma truncada	0,52	1,19
Log-RASTA e MDE c/ truncamento simples	0,52	1,19

Os modelos de duração de estados (MDE) em combinação com *Log*-RASTA não forneceram uma melhora no desempenho do SVL. Isto é devido principalmente ao fato de que o *Log*-RASTA elimina quase que totalmente o ruído convolucional e o MDE não fornece uma melhora muito significativa a uma alta SNR. Assim, este seria mais adequado para o ruído aditivo ou na existência de ambos os ruídos.

5.4.2 Ruído Aditivo

Utilizando-se elocuções de teste corrompidas por ruído de carro e de fala da base de dados NOISEX-92 foram realizados testes do SVL com as técnicas de robustez já mencionadas. Os resultados de EER_{SS} fornecidas pelo sistema são mostrados na Tab. 5.VIII.

Tabela 5.VIII - Taxa de erros iguais com ruído aditivo

Ruído	SNR (dB)	EER_{SS} (%)					
		sem técnica	SS	SS e MDE c/ gamma	SS e MDE c/ gamma trunc.	SS e MDE c/ trunc. simples	SS, *Log*-RASTA e MDE c/ trunc. simples
carro	0	22,9	13,12	11,34	10,20	10,24	7,37
	6	6,34	3,70	3,76	3,15	3,11	2,70
	12	1,80	1,24	1,57	1,12	1,11	1,18
	18	0,68	0,59	0,74	0,49	0,49	0,63
fala	0	23,58	13,87	12,30	11,36	11,17	8.72
	6	6,77	3,93	3,76	3,76	3,72	3,00
	12	1,88	1,23	1,41	1,29	1,28	1,21
	18	0,68	0,55	0,84	0,59	0,59	0,63

O MDE em conjunto com SS forneceram um bom resultado principalmente em SNR baixa, inferior a 12dB. A utilização de SS e MDE com a distribuição gamma sem truncamento não oferece um desempenho muito significativo, tendo um resultado inferior do que apenas com SS para SNR igual ou maior que 6dB para ruído de carro e igual ou maior a 12dB para ruído de fala. Já SS e MDE com gamma truncada ou truncamento simples ofereceram um bom desempenho, chegando a uma melhora de até 56% com ruído de carro a 0dB. Comparada com a SS isolada, a combinação de SS e MDE não levou a uma

redução na taxa de erro com SNR igual a 12 e 18dB. SS e MDE com gamma truncada ou truncamento simples forneceram um desempenho semelhante, sendo que dá-se preferência ao último devido a sua baixa complexidade computacional.

Embora o *Log*-RASTA seja mais apropriado para ruído convolucional, ele também pode ser usado com eficácia para ruído aditivo, como mostra os resultados da Tab. 5.VIII. Isto se deve a própria idéia inicial do RASTA em que se atenuam freqüências que não podem ser naturalmente geradas pelo aparato vocal humano. *Log*-RASTA em conjunto com SS e MDE com truncamento simples forneceu uma melhora no desempenho do SVL de até 69% com ruído de carro a 0dB. Mesmo assim, a utilização em conjunto destas técnicas não foi melhor que SS individualmente com ruído de carro ou fala para SNR igual a 18dB. Isto deve-se a eliminação pelo *Log*-RASTA de características próprias do locutor, melhor verificada com a presença de um menor nível de ruído aditivo.

Teoricamente, MDE poderia ser empregado com qualquer outra técnica de robustez principalmente por ser aplicada sobre o modelo do locutor. Assim, não há influência sobre uma outra técnica empregada conjuntamente, como verificado com SS e *Log*-RASTA fazendo com que todas contribuam para um melhor desempenho final.

5.4.3 Ruído Aditivo e Convolucional

Os resultados dos testes com ruído aditivo e convolucional utilizando-se técnicas de robustez conjuntamente são ilustrados na Tab. 5.IX.

Inicialmente aplica-se uma técnica para cancelamento de ruído aditivo e outra para o convolucional. Verifica-se, então, que a SS utilizada em conjunto com a CMN melhora significativamente a robustez na presença de ruído convolucional. Mesmo assim, SS e *Log*-RASTA fornecem um desempenho muito melhor, já que este também elimina o ruído aditivo como verificado nos experimentos da Tab. 5.VIII. Com a combinação destes últimos houve uma melhora na EER_{SS} de aproximadamente 70% com ruído aditivo de SNR igual a 0dB; para um menor nível de ruído esta melhora foi ainda mais significativa, como no caso com ruído de fala a 18dB em que se conseguiu uma melhora de até 89%.

SS, *Log*-RASTA e MDE conjuntamente fornecem um bom desempenho para baixas SNR mas levemente inferior a SS e *Log*-RASTA em 12 e 18dB. Com ruído de carro a 6dB,

SS, *Log*-RASTA e MDE com gamma truncada chega-se a uma melhora de 87% no desempenho do SVL. A combinação de técnicas com MDE com gamma truncada e truncamento simples continua fornecendo em média um desempenho melhor que na ausência destas técnicas, sendo este desempenho ainda semelhante para ambas as combinações. A combinação com MDE com gamma sem truncamento continua fornecendo um acréscimo ao desempenho apenas em baixas SNR, não sendo este muito significativo.

Tabela 5.IX - Taxa de erros iguais com ruído de carro e convolucional

Ruído	SNR (dB)	EER_{SS} (%)						
		s/ técnica	SS e CMS	SS e *Log*-RASTA	SS, *Log*-RASTA e MDE gamma	SS, *Log*-RASTA e MDE gamma trunc.	SS, *Log*-RASTA e MDE trunc. simples	SS e MDE trunc. simples
carro	0	30,00	14,70	8,53	8,05	7,57	7,66	17,37
	6	17,79	5,08	2,78	2,85	2,65	2,67	7,90
	12	5,94	2,13	1,26	1,55	1,35	1,36	4,62
	18	3,42	0,88	0,72	0,98	0,78	0,77	2,70
fala	0	29,81	15,90	9,55	8,89	8,48	8,49	17,25
	6	13,46	5,48	3,09	3,24	2,91	2,94	8,15
	12	6,16	2,17	1,22	1,53	1,24	1,25	4,49
	18	3,41	0,90	0,71	1,17	0,73	0,74	2,80

5.5 VARIABILIDADE DO LIMIAR DE DECISÃO

Embora a normalização da verossimilhança faça o limiar de decisão considerado a posteriori se aproximar de zero e diminuir sua variabilidade, ela não consegue evitar

totalmente a variação do limiar com a presença de ruído. Assim, em condições reais, um limiar de decisão deve ser tomado a priori distintamente para cada locutor e para SNR diferentes. A variabilidade do limiar de decisão, conforme a SNR, depende também da técnica de robustez empregada. Nas Fig. de 5.12 a 5.15 observa-se a variabilidade do valor dos limiares de decisão (considerando-se um único limiar para o SVL) conforme a técnica empregada e a relação sinal/ruído.

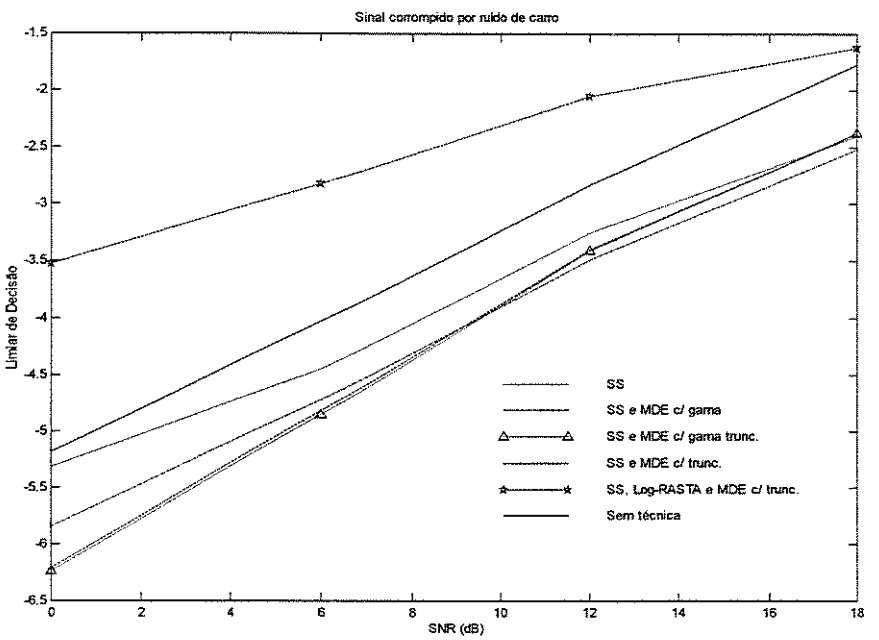

Figura 5.12 - Variabilidade do limiar de decisão para diversas técnicas de robustez a ruído com sinal corrompido por ruído de carro

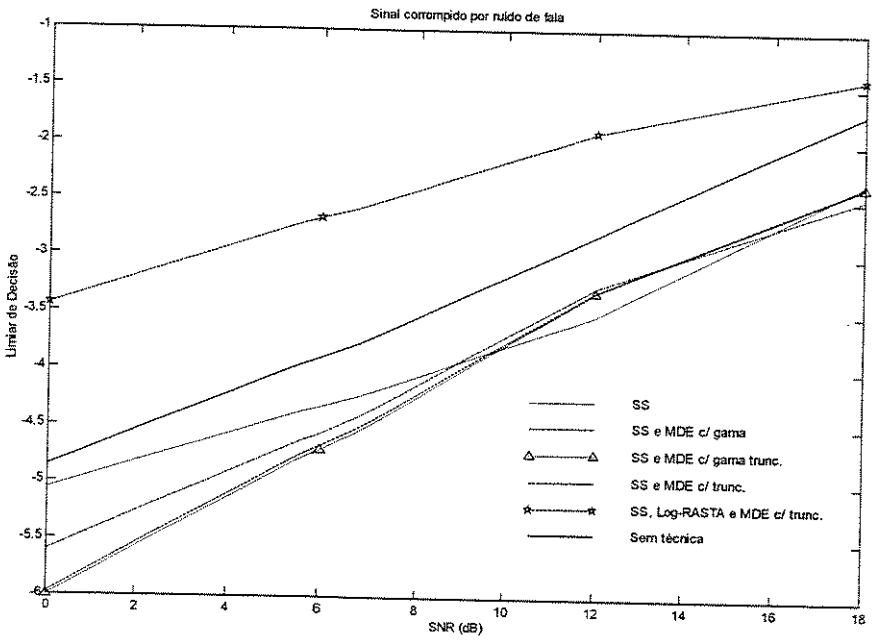

Figura 5.13 - Variabilidade do limiar de decisão para diversas técnicas de robustez a ruído com sinal corrompido por ruído de fala

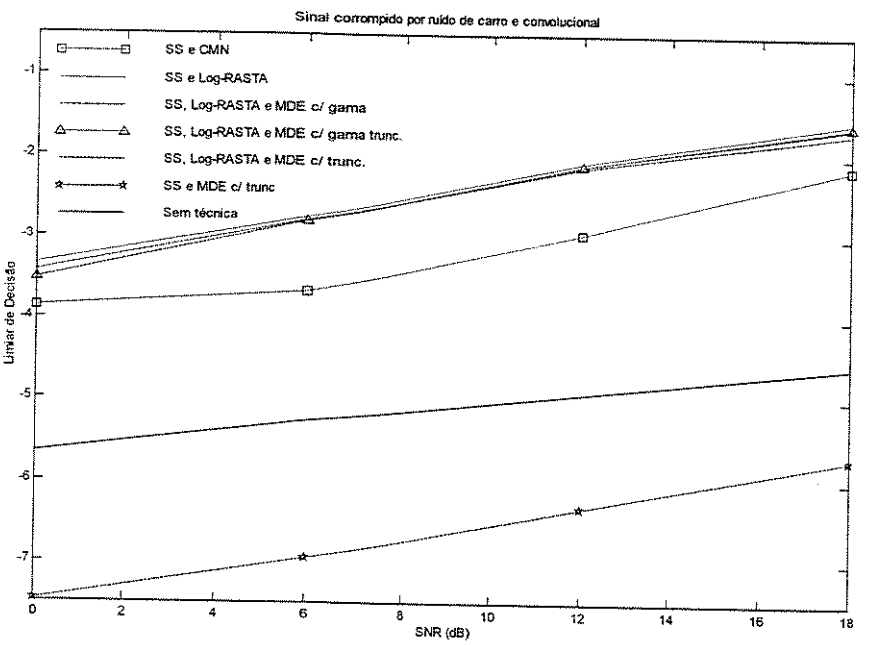

Figura 5.14 - Variabilidade do limiar de decisão para diversas técnicas de robustez a ruído com sinal corrompido por ruído de carro e convolucional

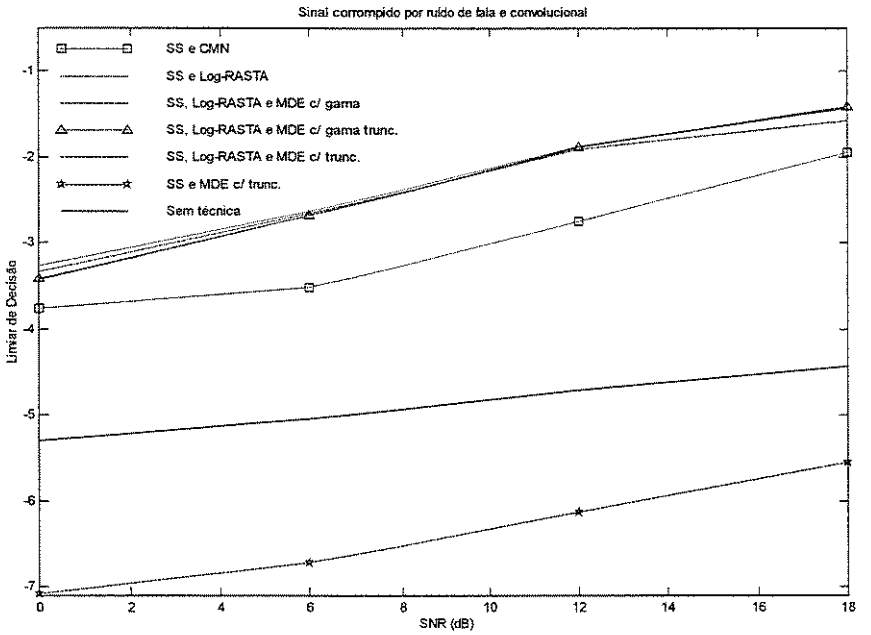

Figura 5.15 - Variabilidade do limiar de decisão para diversas técnicas de robustez a ruído com sinal corrompido por ruído de fala e convolucional

Nas Fig. 5.12 e 5.13, com testes com sinal corrompido apenas por ruído aditivo, verifica-se que empregando-se SS isoladamente e com MDE há um aumento na variabilidade do limiar de decisão em relação ao limiar encontrado onde nenhuma técnica de robustez é empregada. O mesmo pode ser verificado nas Fig. 5.14 e 5.15, no caso de ruído aditivo e convolucional, quando emprega-se SS e MDE com truncamento simples. Já quando emprega-se CMN em conjunto com SS ou *Log*-RASTA em conjunto com quaisquer outras técnicas mencionadas esta variabilidade é sensivelmente reduzida, principalmente com o emprego desta. Assim, mesmo que *Log*-RASTA ou até mesmo CMN possam não acrescer o desempenho do SVL eles são especialmente úteis no que se refere a tornar o limiar de decisão mais estável quando na presença de ruído e facilitar uma estimação a priori do mesmo.

Esta diminuição na variabilidade do limiar de decisão com CMN ou *Log*-RASTA pode ser justificada por estas técnicas tornarem nula (ou aproximadamente nula) a média temporal de cada parâmetro da elocução, bastante modificada pela presença de ruído. Assim, o acréscimo ao valor do logaritmo da verossimilhança, devido a comparação entre as médias dos parâmetros da elocução e as do modelo do locutor, é eliminado.

5.6 SUMÁRIO E CONCLUSÕES

Com o intuito de verificar o comportamento do SVL aqui desenvolvido na presença de ruído foram feitos testes inserindo-se ruído convolucional e/ou aditivo nas elocuções de teste de cada locutor. A performance do sistema com ruído foi muito degradada com ruído convolucional, aditivo ou ambos. Contudo, aplicando-se ao SVL técnicas convencionais de robustez de sistemas de reconhecimento de voz, como CMN, *Log*-RASTA, SS e MDE, houve uma sensível melhora em seu desempenho.

Para ruído convolucional, tanto CMS quanto *Log*-RASTA tiveram um desempenho satisfatório, com uma melhora na EER_{SS} de até 95%. A filtragem RASTA tem a vantagem de não depender do comprimento da elocução embora exija uma carga computacional maior do que CMN.

Para ruído aditivo a SS individualmente é a melhor técnica entre as utilizadas neste trabalho, com uma redução de até 44% na EER_{SS}. Mesmo assim, seu desempenho pode ser melhorado se usada em conjunto com outras técnicas. Ainda para ruído aditivo, a combinação de técnicas que dá melhor resultado é a com SS e MDE com gamma truncada ou truncamento simples chegando a uma melhora de até 56%. Apesar do J-RASTA ser mais apropriado para ruído aditivo, o *Log*-RASTA também diminui a taxa de erros do SVL em combinação com SS e MDE, chegando a uma redução de 69% na EER_{SS}.

Quando individualmente presente o ruído convolucional não chega a ser um problema muito significativo, diante das técnicas de robustez existentes. No entanto, quando há a presença tanto do aditivo quando do convolucional, o desempenho do SVL cai drasticamente. Com ambos os ruídos, a utilização em conjunto de SS, RASTA e MDE fornece uma melhora na taxa de erro de até 87% em SNR de 18dB o que, em muitos casos, poderia ser considerado satisfatório.

O MDE com a distribuição gamma não mostrou ser eficiente no cancelamento dos efeitos do ruído em SNR´s elevadas. Já o MDE com gamma truncada e com truncamento simples podem ser usados com sucesso para este fim, principalmente em baixa SNR. No que se refere a estas duas técnicas, dá-se preferência para a utilização do MDE com truncamento simples por requerer uma carga computacional menor do que a distribuição

gamma truncada, sendo que ambas mostraram um desempenho equivalente. Isto é coerente com o observado em (Yoma *et al.*, 2000) onde concluiu-se que em RV dependente de locutor, no contexto da duração de estados, o modelamento estatísitco parece ser superfluo quando comparado com a utilização de uma duração máxima e de uma mínima.

O limiar de decisão ótimo considerado a posteriori é muito dependente da relação sinal-ruído e de difícil estimação a priori para condições ruidosas. A presença de técnicas como CMN e principalmente *Log*-RASTA diminui significativamente a variabilidade do limiar de decisão, facilitando a sua determinação a priori.

5.7 SUGESTÕES PARA MELHORIA NA ROBUSTEZ DE SVL

Algumas modificações nas técnicas empregadas ou a utilização em conjunto com outras técnicas poderiam melhorar o desempenho do SVL em condições ruidosas. Entre elas, podem ser citadas:

- Uma estimação mais precisa do espectro do ruído utilizado na SS, levando em consideração as variações temporais do mesmo;
- A utilização de J-RASTA em conjunto com SS e MDE tanto para ruído aditivo quanto aditivo e convolucional;
- A inclusão do modelamento de duração de estados na fase de treinamento;
- Otimização da aplicação do MDE no algoritmo de Viterbi. As técnicas de modelamento de duração de estados utilizadas neste trabalho levam em consideração a probabilidade de transição de estado condicionada ao fato do estado já ter durado um determinado número de quadros. Mas com o algoritmo de Viterbi só conseguimos saber a duração de estados mais provável *a posteriori*.

CONCLUSÕES

O SVL implementado neste trabalho foi inicialmente testado sem a presença de ruído e mostrou uma EER média de 0,36% considerando um limiar de decisão a posteriori. Vale a pena mencionar que esta taxa de erro é muito próxima da obtida em artigos internacionais recentes sobre o estado da arte em VL (Bimbot *et al.*, 1997; Markov & Nakagawa, 1998). No entanto, alguns locutores (*goats*) apresentam uma EER inaceitável. Estes *goats* constituem a maior parte das elocuções rejeitadas e , em condições reais, precisam ser tratados individualmente.

Com a presença de ruído aditivo e convolucional o Sistema de Verificação de Locutor teve seu desempenho comprometido, principalmente pela presença de ambos os ruídos. A utilização de técnicas de robustez como CMN ou *Log*-RASTA reduziu em até 95% a taxa de erros iguais média do sistema com elocuções corrompidas por ruído convolucional. Contudo, a filtragem RASTA tem a vantagem de não exigir elocuções longas apesar de requerer uma maior carga computacional de que CMN. Para ruído aditivo, a melhor combinação de técnicas foi a de SS com MDE com truncamento simples mostrando uma redução de até 56% na taxa EER. A utilização destas com *Log*-RASTA reduz ainda mais os efeitos do ruído aditivo, sendo que com ruído convolucional SS, *Log*-RASTA e MDE com truncamento simples forneceram um bom desempenho apresentando uma redução na taxa de erros iguais de até 87%. Verificou-se também que CMN e principalmente *Log*-RASTA diminuem significativamente a variabilidade do limiar de decisão devido a presença de ruído, e poderiam ser usadas somente para este fim. É interessante destacar que, apesar dos resultados aqui apresentados mostrarem um importante avanço, a robustez de sistemas de reconhecimento de voz e de locutor a ruídos

interferentes ainda é um problema complexo, e é o principal empecilho enfrentado em aplicações práticas reais.

Como contribuições desta dissertação podemos citar:

1. A validação da técnica de MDE com restrições temporais aplicada para robustez a ruído em Sistemas de Verificação de Locutor;
2. O estudo de verificação automática de locutor, um assunto novo na FEEC da UNICAMP, sendo que o autor espera que este trabalho possa ser utilizado como referências para outras dissertações na área;

Para pesquisas futuras sugere-se:

1. Implementação de um SVL com texto induzido, com emprego de múltiplas gaussianas;
2. Desenvolvimento de um SVL texto independente;
3. Um estudo mais detalhado do limite de aplicabilidade de MDE em combinação com as técnicas aqui abordadas para cancelamento de ruído aditivo e convolucional.

REFERÊNCIAS BIBLIOGRÁFICAS

Acero, A.; Stern, R. "Environmental robustness in automatic speech recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 254-272, 1990.

Albesano, D.; Gemello, R.; Mana, F. "Hybrid HMM-NN modeling of stationary-transitional units for continuous speech recognition". *Information Sciences*, 123 (2): 3-11, 2000.

Atal, B. "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification". *Journal of the Acoustics Society of America*, 55 (6): 1304-1312, 1974.

Atal, B. S. "Automatic Recognition of Speakers from their Voices". *Proceedings of IEEE*, 64 (4): 460-475, 1976.

Baker, J. K. "Stochastic modeling for automatic speech understanding". In: D. R. Reddy, ed., *Speech Recognition*. New York: Academic Press, pp. 521-542, 1975a.

Baker, J. K. "The DRAGON system – an overview". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23 (2): 24-29, 1975b.

Bennani, Y.; Gallinari, P. "Neural networks for discrimination and modelization of speakers". *Speech Communication*, 17: 159-175, 1995.

Berouti, M.; Schwartz, R.; Makhoul, J. "Enhancement of speech corrupted by acoustic noisy". *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, (4): 208-211, 1979.

Bimbot, F.; Hutter, H-P.; Jaboulet, C.; Koolwaaij, J.; Lindberg, J.; Pierrot, J-B. "Speaker verification in the telephone network: research activities in the CAVE project". *Proceedings of the EUROSPEECH'97*, Rhodes, Grécia, 2: 971-974, 1997.

Burshtein, D. "Robust parametric modeling of durations in Hidden Markov Models". *IEEE Transactions on Speech and Audio Processing*, 4 (3): 240-242, 1996.

Carey, M.; Parris, E. "Speaker verification using connected words". *Proceedings on Institute of Acoustics*, 14 (6): 95-100, 1992.

Charlet, D.; Jouvet, D. "Optimizing feature set for speaker verification". *Pattern Recognition Letters*, 18: 873-879, 1997.

Cheung, R. S.; Eisenstein, B. A. "Feature selection via dynamic programming for text-independent speaker verification". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26 (5): 397-403, 1978.

Claes, T.; Van Compernolle, D. "SNR-normalization for robust speech recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 331-334, 1996.

Claes, T.; Xie, F.; Van Compernolle, D. "Spectral estimation and normalization for robust speech recognition". *Proceedings of the IEEE International Conference on Signal Processing*, 1997-2000, 1996.

Deller, J. R.; Proakis, J. G.; Hansen, J. H. L. *Discrete Time Processing of Speech Signal*. New York: MacMillan, 1993.

Doddington, G. R. "Speaker recognition – identifying people by their voices". *Proceedings of IEEE*, 73 (11): 1651-1664, 1985.

Figueiredo, R. M. *Identificação de falantes: aspectos teóricos e metodológicos*. Tese de Doutorado, Universidade Estadual de Campinas, 1994.

Forsyth, M. *Semi-continuous Hidden Markov Models for Automatic Speaker Verification*. PhD Thesis, The University of Edinburgh, United Kingdom, 1995.

Frischholz, R. W.; Dieckmann, U. "BioID: a multimodal biometric identification system". *IEEE Computer Magazine*, 33 (2): 64-68, 2000.

Furui, S. "Cepstral analysis technique for automatic speaker verification". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29 (2): 254-272, 1981.

Furui, S. "Research on individuality features in speech waves and automatic speaker recognition techniques". *Speech Communication*, 5 (2): 183-197, 1986

Furui, S. *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker, New York, 1989

Furui, S. "Speaker-independent and speaker-adaptive recognition techniques". In: Furui, S., Sondhi, M. M. (Eds.), *Advances in Speech Signal Processing.* Marcel Dekker, New York, pp. 597-622, 1991a.

Furui, S. "Speaker-dependent-feature extraction, recognition and processing techniques". *Speech Communication*, 10 (6): 505-520, 1991b.

Furui, S. "An overview of speaker recognition technology". In: *ESCA Workshop na Automatic Speaker Recognition, Identification and Verification*, pp. 1-9, 1994.

Furui, S. "Recent advances in speaker recognition". *Pattern Recognition Letters*, 18: 859-872, 1997.

Gales, M. J. F."'Nice" model-based compensation approach to robust speech recognition". *ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for unknown communication channel*, pp. 55-64, 1997.

Gales, M. J. F.; Young, S. J. *Parallel Model Combination for speech recognition in noise.* Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR135, Junho, 1993a.

Gales, M. J. F.; Young, S. J. *PMC for speech recognition in additive and convolucional noise*. Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR154, Dezembro, 1993b.

Gales, M. J. F.; Young, S. J. "Robust continuous speech recognition using Parallel Model Combination". *IEEE Transactions on Speech and Audio Processing*, 4 (5): 352-359, 1996.

Haykin, S. *Neural Networks: a comprehensive foundation*. Macmillan College Publishing Company, 1994.

Hermansky, H.; Morgan, N.; Bayya, A.; Kohn, P. "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)". *Proceedings of the Eurospeech*, pp. 1367-1370, 1991.

Hermansky, H.; Morgan, N.; Hirsch, H. "Recognition of speech in additive and convolutional noise based on RASTA spectral processing". *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2: 83-86, 1993.

Hermansky, H.; Morgan, N. "RASTA Processing of Speech". *IEEE Transactions on Speech and Audio Processing*, 2 (4): 578-589, 1994.

Higgins, A.; Bahler, L.; Porter, J. "Speaker verification using randomized phrase prompting". *Digital Signal Processing*, 1: 89-106, 1991.

Jelinek, F.; Bahl, L. R.; Mercer, R. L. "Design of a linguistic statistical decoder for the recognition of continuous speech". *IEEE Transactions on Information Theory*, 21(5): 250-256, 1975.

Jourlin, P.; Luettin, J.; Genoud, D.; Wassner, H. "Acoustic-labial speaker verification". *Pattern Recognition Letters*, 18: 853-858, 1997.

Lamel, L. *et al.* "An improved endpoint detector for isolated word recognition". *IEEE Transactions on Acoustic, Speech and Signal Processing*, 29 (4): 777-785, 1981.

Laurila, K. "Noise robust speech recognition with state duration constrains". *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 871-874, 1997.

Mak, B., Junqua, J.C., Reaves,B. "A robust speech/non-speech detection algorithm using time and frequency-based features". *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 1: 269-272, 1992.

Mammone, R. J.; Zhang, X.; Ramachandran, R. P. "Robust Speaker Recognition". *IEEE Signal Processing Magazine*, (9): 58-71, 1996

Markel, J. D. & Gray, A. H. Jr. *Linear Prediction of Speech.* New York: Springer-Verlag, 1980.

Markow, K. P.; Nakagawa, S. "Text-independent speaker recognition using non-linear frame likelihood transformation". *Speech Communication*, 24: 193-209, 1998.

Matsui, T.; Furui, S. "Concatenated phoneme models for text-variable speaker recognition". *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing, Minneapolis*, 2: 391-394, 1993.

Matsui, T.; Furui, S. "Similarity normalization method for speaker verification based on a posteriori probability". *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp.59-62, 1994a.

Matsui, T.; Furui, S. "Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker verification", *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, Adelaide, Australia, 1: 125-128, 1994b.

Morgan, N.; Franco, H. "Applications of neural networks to speech recognition". *IEEE Signal Processing Magazine*, 14 (6): 46-47, 1997.

Naik, J. M. "Speaker Verification: A Tutorial". *IEEE Communications Magazine*, (1): 42-48, 1990.

Nakagawa, S.; Markov, K. P. "Speaker verification using frame and utterance level likelihood normalization". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.1087-1090, 1997.

Olsen, J. "A two-stage procedure for phone based speaker verification". *Pattern Recognition Letters*, 18: 889-897, 1997.

Openshaw, J. P., Sun, S. P., Mason, J. S. "A comparison of Composite Features Under Degraded Speech in Speaker Recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2: 371-374, 1993.

O'Shaugnessy, D. "Speaker recognition". IEEE Transactions on Acoustics, Speech, and Signal Processing, 3 (4): 4-17, 1986

Picone, J. W. "Signal modeling techniques in speech recognition". *Proceedings of the IEEE*, 79 (4): 1214-1247, 1991.

Rabiner, L. R. & Juang, B., *Fundamentals of Speech Recognition*, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1993.

Raj, B.; Gouvea, E.; Moreno, P.; Stern, R. "Cepstral compensation by polynomial approximation for environment-independent speech recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2340-2344, 1996.

Rosenberg, A. E. "Automatic speaker verification: a review". *Proceedings of IEEE*, 64 (4): 475-487, 1976.

Rosenberg, A. E.; DeLong, J.; Lee, C.; Juang, B.; Soong, F. "The use of cohort normalized scores for speaker verification". *Proceedings of the International Conference on Spoken Language Processing*, Banff, pp. 599-602, 1992.

Rosenberg, A. E.; Lee, C.; Soong, F. K. "Cepstral channel normalization techniques for HMM-based Speaker Verification". *IEEE Proceedings of the International Conference on Signal Processing*, pp. 1835-1838, 1994.

Rosenberg, A.; Soong, F. "Recent research in automatic speaker recognition". In: Furui, S.; Sondhi, M. (Eds.), *Advances in Speech Signal Processing*. Marcel Dekker, New York, pp. 701-737, 1991.

Sakoe, H.; Chiba, S. "Dynamic programming algorithm optimization for spoken word recognition". *IEEE Transactions on Acoustic, Speech, and Signal Processing*, 26 (2): 43-49, 1978.

Sambur, M. R. "Selection of acoustic features for speaker identification". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23 (2): 176-182, 1975.

Savoji, M. H. "A robust algorithm for accurate endpointing of speech signals". *Speech Communication*, 8: 45-60, 1989.

Soong, F. K.; Rosenberg, A. E. "On the use of instantaneous and transitional spectral information in speaker recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(6): 871-879, 1988.

Southerland, A. & Jack, M. *Aspects of speech technology*. Edingurgh University Press, pp 197-200, 1988.

Van Compernolle, D. "Noise Adaptation in a Hidden Markov Model speech recognition system". *Computer Speech and Language*, 3 (2): 151-168, 1989.

Vasegui, S. V.; Milner, B. P. "Noise compensation methods for Hidden Markov Model speech recognition in adverse environments". *IEEE Transactions on Speech and Audio Processing*, 5 (1): 11-21, 1997.

Wang, H. C., Chen, M. S., Young, T. "A novel approach to the speaker identification over the telephone networks". *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2: 161-164, 1993

Wark, T. J.; Sridharan, S.; Chandran, V. "The use of speech and lip modalities for robust speaker verification under adverse conditions". *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, vol.1, 1998.

Yoma, N. B. *Reconhecimento automático de palavras isoladas: estudo e aplicação dos métodos determinístico e estocástico*. Tese de mestrado, FEEC-UNICAMP, Campinas, 1986.

Yoma, N. B. *et al.* "On including temporal constraints in Viterbi alignment for speech recognition in noise". Aceito para publicação em *IEEE Transactions on Speech and Audio Processing*, 2000.

Young, S.; Odell, J.; Valtchev, V.; Woodland, P. *The HTK Book (for HTK Version 2.1)*. Cambridge University, 1997.