Dissertação de Mestrado

Inatel
Instituto Nacional de Telecomunicações

AVALIAÇÃO DOS TRECHOS SONOROS E NÃO-SONOROS DO SINAL DE FALA PARA IDENTIFICAÇÃO DO LOCUTOR INDEPENDENTE DE TEXTO

ALEXANDER COELHO

NOVEMBRO / 2005

Avaliação dos Trechos Sonoros e Não Sonoros do Sinal de Fala para Identificação de Locutor Independente de Texto

Alexander Coelho

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

Orientador: Prof. Dr. Carlos Alberto Ynoguti

Santa Rita do Sapucaí
2005

Dissertação defendida e aprovada em 18/11/2005, pela comissão julgadora:

Prof. Dr. Carlos Alberto Ynoguti - DTE / INATEL

Prof. Dr. Abraham Alcaim - CETUC / PUC / Rio

Prof. Dr. Francisco José Fraga da Silva - USP / SP

Coordenador do Curso de Mestrado

Prof. Dr. Adonias Costa da Silveira

À minha família, Juarez Coelho, Maria Efigênia Ribeiro Coelho, meus irmãos Carlos Henrique Coelho e Virgilio Coelho, que me motivam a buscar novos desafios e oportunidades para o crescimento como ser humano.

Agradecimentos

Todo agradecimento, por mais que se diga ou se escreva palavras maravilhosas, será incapaz de expressar o que este sentimento denota; acredito que o agradecimento não apenas se faça por um muito obrigado ou por elogios, mas através da amizade.

Quero deixar aqui escrito meu muito obrigado ao meu **amigo** Carlos Alberto Ynoguti, além de um excelente professor e orientador, é também um exemplo de profissional a ser seguido.

Aos amigos e colegas José Antônio, Lidiane, Sandro, Daniela, e aos colegas do LabPg.

Às pessoas que emprestaram suas vozes na confecção das bases de dados utilizadas neste trabalho.

Ao Inatel, por proporcionar a oportunidade de realizar este mestrado através da concessão de bolsa e do estágio docente.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ASV/I *Automatic Speaker Verification / identification* - Verificação / Identificação Automática de Locutor

HMM *Hidden Markov Model* - Modelo Oculto de Markov

BW_c *Critical Band Width* - Largura de Faixa de Banda Crítica

FFT *Fast Fourier Transform* - Transformada Rápida de Fourier

MFCC *Mel Frequency Cepstrum Coefficients* - Coeficientes Mel Cepstrais

LPC *Linear Predictive Coding* - Codificação por Predição Linear

LBG *Linde-Buzo-Gray* - Algoritmo para Quantização Vetorial proposto por Linde, Buzo e Gray.

MLP *Multilayer Perceptron*

ANN *Artificial Neural Network*

QV *Quantizador Vetorial*

SVM *Support Vector Machines* - Máquina de Vetor de Suporte

VDT *Verificação dependente de texto*

VIT *Verificação independente de texto*

IDT *Identificação dependente de texto*

IIT *Identificação independente de texto*

CL close-set

OP open-set

FA *falsa aceitação*

FR *falsa rejeição*

GMM *Gaussian Mixture Model*

Lista de Símbolos

P	Probabilidade
S_n	Sinal digital de voz
a_{pre}	Coeficiente de pré-ênfase
N_s	Número de amostras de cada janela
n	Amostra em avaliação
fc	Frequência de corte
k	Número de janelas para derivação
$\tilde{S}(n)$	Sinal pré-enfatizado
y	Saída do Neurônio
$f(.)$	Função não linear
x_i	i-ésima entrada do neurônio
Θ	Limiar de ativação
O_t	Vetor de entrada

Resumo

Este trabalho tem três objetivos principais: a) verificar qual ou quais parâmetros acústicos são os mais adequados para a tarefa de reconhecimento de locutor; para este tópico foram investigados os parâmetros mel-cepstrais, pitch e perfil-energia, bem como suas derivadas primeira e segunda, b) identificar qual o tipo de classificador leva a um melhor desempenho em termos de taxa de acerto; foram analisados aqui os modelos ocultos de Markov, redes neurais artificiais, quantizadores vetoriais e os modelos de mistura de Gaussianas, c) investigar a influência das partes sonoras e não sonoras da fala no desempenho nos sistemas de classificação para reconhecimento de locutor.

Palavras chave: processamento de voz, reconhecimento de locutor,GMM.

Abstract

This work has three main goals: a) to verify which acoustic parameters lead to better discrimination among speakers; for this topic the investigated parameters were mel-cepstrum, pitch and energy-profile, as well as their first and second derivatives; b) to identify which classifier performs better in terms of recognition rate; a benchmark among Hidden Markov Models, Neural Networks, Vectorial Quantization and Gaussianas Mixture Models was performed; c) to investigate the influence of the voiced and unvoiced parts of speech in the performance of speaker recognition systems.

Keywords: speech processing, speaker recognition, GMM.

Capítulo 1

Introdução

1.1 Biometria

A Biometria refere-se à identificação de uma pessoa através de características físicas e fisiológicas tais como a forma e tamanho da face, o padrão da impressão digital, o tamanho e geometria das mãos, a forma e detalhes da íris, o tipo de voz, a assinatura, entre outras.

A carteira de identidade, documento padrão do brasileiro, permite a identificação de uma pessoa através de várias medidas biométricas: a forma do rosto, a assinatura e a impressão digital. Estas, quando analisadas em conjunto, são características únicas e exclusivas de um único indivíduo. Entretanto este sistema apresenta alguns problemas: obviamente existe o risco de falsificação dos documentos; na maioria das vezes não é checada a impressão digital da pessoa; as fotos costumam ser uma outra fonte de problemas, pois as pessoas mudam com o tempo, e desta forma, a confiabilidade destas características diminui.

O velho RG poderia ser substituído pelo armazenamento em um banco de dados digital de características como a impressão digital, fotos da íris, registros de voz, e outros. Na hora da identificação, aparelhos poderiam ser utilizados para coletar estas informações e enviá-las a um sistema que acessaria o banco de dados central, e faria a averiguação da identidade da pessoa. Desta forma, não seria necessário guardar mais o número no RG ou CPF. Tudo estaria ligado às características biométricas da pessoa.

Outras aplicações além do RG podem ser citadas, tais como:

Sistemas de atendimento e controle: O usuário realiza um *login* automático e obtém informações pertencentes a um banco de dados, como contas de consumo (telefone, água, luz), saldos de bancos e também acesso a serviços;

Identificação judicial: Pode ser realizada uma identificação do réu ou vítima

através da comparação de provas (locuções tiradas de um telefone grampeado, por exemplo) e gravações da voz dos mesmos.

Outras vantagens do método de identificação biométrica em relação aos métodos tradicionais são:

- a pessoa a ser identificada deve estar fisicamente presente no local de identificação,
- não precisa se lembrar de números ou senhas, carregar uma chave ou um cartão,
- o método é mais rápido que qualquer outro dispositivo,
- não precisaria mais portar documentos e cartões que podem ser esquecidos, roubados ou perdidos.

Em resumo, *a pessoa passa a ser o próprio documento.*

Com o avanço das técnicas automáticas para identificação biométrica, estas vêm sendo cada vez mais utilizadas. De fato, como mostram os dados a seguir, os sistemas mais modernos apresentam um desempenho melhor que o humano, para identificação por voz, tema abordado por este trabalho:

- As pessoas podem identificar vozes familiares facilmente, embora a taxa de erro para expressões breves, freqüentemente exceda 20%. Aproximadamente 2 a 3 segundos de fala são suficiente para identificar uma voz, embora o desempenho diminua para vozes menos conhecidas.
- Para locutores pouco conhecidos, o tempo de treinamento para humanos é muito maior que para as máquinas. O ser humano consegue reter em sua memória a curto prazo, aproximadamente 5 a 10 locutores. Um teste de reconhecimento de locutor por humanos, utilizando 8 a 10 locutores, obteve uma taxa de acerto de 96% para uma frase, desempenho que cai drasticamente, para 54% quando a duração da fala é pequena (menos de 1s). Passando a voz através de um filtro, o desempenho cai para 31% [9],
- Por outro lado, sistemas artificiais trabalham com taxas de acerto de 90 % a 99%, com uma quantidade muito maior de locutores e com material para reconhecimento muito menor [13][9][27].

1.2 Análise biométrica da voz

A fala contém informação não só sobre a identidade do locutor, mas também sobre a língua, características físicas, estado emocional e a informação propriamente dita (o que se quer dizer)[3].

Para determinar a identidade de um certo locutor são usados 2 níveis de informação: no nível alto, são levados em conta fatores como dialeto, estilo de fala e forma de se expor no contexto. Estas características são reconhecidas e analisadas por humanos. No nível baixo, são observados o período de pitch, o ritmo, o tom, a magnitude espectral, etc. Estas informações são usadas por máquinas. No entanto existem problemas:

- a voz se altera drasticamente quando as pessoas adoecem (com resfriados principalmente);
- de manhã bem cedo, logo ao acordar, as pessoas tendem a falar com voz mais grave;
- assim como para as assinaturas, é possível imitar a voz de uma pessoa e enganar o sistema;
- a qualidade dos microfones e o nível do ruído de fundo são fatores que influenciam diretamente na confiabilidade de tais sistemas.

Por conta disso, existem sistemas híbridos, onde se faz a identificação através de várias medidas simultaneamente: voz, íris, face, impressão digital, etc., para aumentar a sua confiabilidade. Alguns destes sistemas são adaptativos com as condições do ambiente: se há pouca luz, a identificação pela foto da face recebe um peso menor, por outro lado, o mesmo acontece com a voz se o ambiente é muito ruidoso.

1.3 O Problema do Reconhecimento de Locutor

O problema de reconhecimento de locutor pode ser encaixado dentro da área de processamento de voz, como ilustrado na Figura 1.1.

Existem duas formas de reconhecimento de locutor [9]:

- *identificação de locutor*: Quem é você?
- *verificação de locutor*: Você é quem você diz que é?

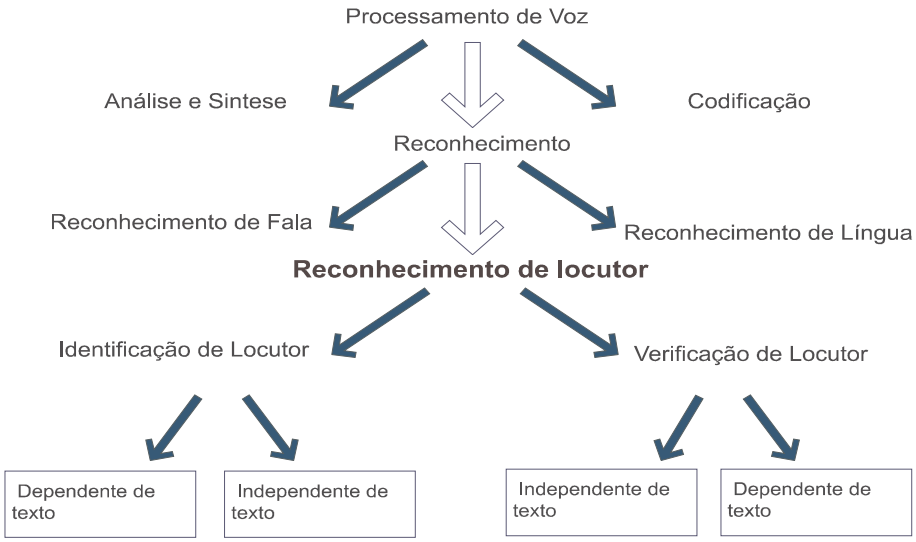


Figura 1.1: *Processamento de Fala*

No primeiro caso, a tarefa do sistema é identificar a qual pessoa pertence a locução sob teste. Desta forma, o sistema acerta se associa a locução sob teste à pessoa que a falou. O desempenho destes sistemas é medido em termos de taxas de acertos, ou seja, o quociente entre o número de associações corretas e o número de testes realizados.

Já no segundo caso, uma locução, supostamente associada a uma determinada pessoa é submetida a teste. A tarefa consiste em aceitar ou não tal locução como pertencente a um determinado locutor. Geralmente isto é feito medindo-se a similaridade da locução sob teste com as locuções usadas para treinar o sistema. Neste caso dois tipos de erro podem acontecer: a *falsa aceitação*, que consiste em aceitar um impostor como sendo um usuário válido e a *falsa rejeição*, que consiste em considerar um usuário válido como um impostor. Na literatura, os impostores são referenciados como "lobos", e os usuários válidos como "ovelhas".

O reconhecimento de locutor pode ser *dependente de texto*, que requer a pronúncia de sentenças específicas, geralmente as mesmas usadas para treinar o sistema; ou *independente de texto*, em que o locutor sob teste pode falar qualquer coisa. Obviamente, o primeiro caso é muito mais simples, e leva também a resultados melhores.

Para o reconhecimento independente de texto é desejável selecionar um material de treinamento que contenha pelo menos um exemplar de cada fone, de modo que o sistema tenha "conhecimento" da forma de pronúncia do locutor a ser modelado.

Quando todos os locutores que irão utilizar o sistema pertencem a um conjunto restrito de pessoas, tem-se uma abordagem conhecida como *closed-set*. No caso

de o sistema operar para além destes locutor, outros que não são conhecidos, tem-se a abordagem *open-set*.

Avaliação de sistemas de verificação de locutor: curvas DET x ROC

Na tarefa de verificação de locutor, mede-se a similaridade entre a locução sob teste e um modelo gerado para o locutor a quem supostamente esta pertence. Se esta ultrapassar um determinado limiar, aceita-se a locução como sendo válida.

Uma das medidas utilizadas para avaliar o desempenho de tais sistemas é a curva ROC (Receive Operating Characteristic), onde a taxa de falsa aceitação é apresentada no eixo horizontal e a taxa de aceitação correta é apresentada no eixo vertical, como mostrado na Figura 1.2

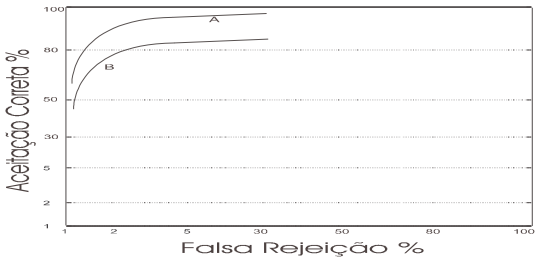


Figura 1.2: *Curva ROC, "A" para alto desempenho e "B" para baixo desempenho do sistema.*

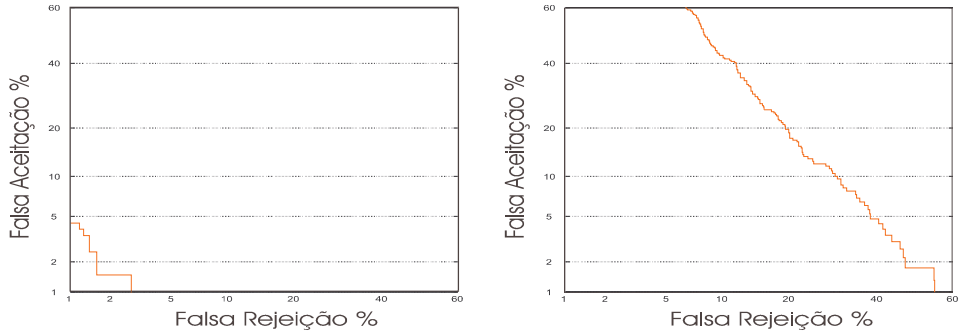
Entretanto, é fácil perceber que um limiar muito baixo para aceitação minimiza os erros de falsa rejeição, mas aumenta os números de falsa aceitação. Por outro lado, a escolha de um limiar muito alto determina uma alta taxa de falsa rejeição enquanto minimiza os erros de falsa aceitação.

Percebe-se então que existe um limiar ideal, o qual mantém um compromisso entre a falsa aceitação e a falsa rejeição. Esta situação é mostrada através da curva DET (Detection Error Tradeoff), onde se apresentam ambas as taxas de erro em seus eixos, [28]. Na Figura 1.3 tem-se exemplos de curvas DET para sistemas de alto desempenho e baixo desempenho.

Neste trabalho serão apresentadas as taxas de erro através das curvas DET.

Variação Intralocutor x Variação Interlocutor

A variação das características entre os diferentes locutores é chamada de variação interlocutor. Esta é causada por diferenças na forma e tamanho do trato vocal, na forma de pronúncia das palavras devido a regionalismos, ritmo da fala, etc.



(a) Sistema com alto desempenho (b) Sistema com baixo desempenho

Figura 1.3: *Desempenhos apresentados através da curva DET.*

A variação intralocutor refere-se à diferença de pronúncia de uma mesma palavra ou sentença pela mesma pessoa. Dentre os fatores que influenciam a variação intralocutor podem-se citar: o estado emocional (quando se está nervoso, fala-se mais rápido e mais agudo), o meio ambiente (força-se a voz para se fazer entender em ambientes muito ruidosos), o horário do dia (de manhã fala-se mais grosso).

No caso ideal teria-se uma baixa variação intralocutor e uma alta variação interlocutor.

1.4 Sobre a dissertação

Neste trabalho foi inicialmente realizado um estudo comparativo de diversas técnicas para o problema de reconhecimento de locutor. Especificamente foram feitas investigações sobre qual parâmetro acústico e qual classificador levaria a um melhor desempenho.

Depois foi feita uma investigação acerca da contribuição das partes sonora e não sonora da fala para o desempenho final do sistema. Os testes realizados mostraram uma melhora significativa no desempenho do sistema ao se utilizar apenas os trechos sonoros da fala.

Esta dissertação está organizada da seguinte forma: no Capítulo 2 são apresentados os parâmetros acústicos utilizados, bem como uma discussão sobre as características favoráveis e desfavoráveis ao problema de reconhecimento de locutor. No Capitulo 3, são mostrados os classificadores utilizados, bem como a forma de implementar os reconhecedores a partir dos mesmos. Os resultados experimentais são apresentados no Capítulo 4 e as conclusões do trabalho, no Capítulo 5.

Capítulo 2

Parâmetros acústicos

Tradicionalmente, os paradigmas de reconhecimento de padrões são divididos em três componentes: extração e seleção de características, casamento de padrões, e classificação. Neste capítulo serão tratados os problemas referentes à primeira etapa.

A *extração de parâmetros* consiste na estimação de variáveis, chamadas de vetores de observação (ou vetores acústicos no caso de processamento de voz), a partir de um outro conjunto de variáveis (por exemplo, uma sequência de amostras de um sinal de voz). O objetivo aqui é conseguir uma representação do sinal original em um espaço de características que seja útil para tarefa de reconhecimento. As técnicas mais utilizadas para a extração de parâmetros para a tarefa de reconhecimento de locutor são as que envolvem uma transformação para o domínio da frequência, tais como os parâmetros LPC e os mel-cepstrais.

A *seleção de parâmetros*, por sua vez, consiste na transformação destes vetores de observação em vetores de características. O objetivo da seleção de parâmetros é encontrar uma transformação para um espaço de dimensão relativamente pequeno que preserve a informação pertinente à aplicação. O método estatístico mais utilizado para realizar a seleção de parâmetros é a Análise de Componente Principal, embora outros tenham sido propostos [9].

Neste trabalho foi implementada apenas a extração de parâmetros, deixando a seleção de parâmetros para uma etapa futura nas pesquisas.

Os parâmetros acústicos podem ser vistos como a forma com que os sistemas de reconhecimento "ouvem" as locuções. Idealmente devem ter uma grande variabilidade interlocutor e pequena variabilidade intralocutor. Neste capítulo são apresentados os parâmetros utilizados neste trabalho, tentando analisá-los sob este ponto de vista.

Foram analisados os parâmetros mel cepstrais, LPC, Pitch e Perfil Energia, bem como suas derivadas primeira e segunda. Os parâmetros mel cepstrais e

LPC são bastante utilizados na tarefa de reconhecimento de fala, e também para o reconhecimento de locutor [3]. Os parâmetros pitch e perfil de energia parecem ter potencial para a tarefa de reconhecimento de locutor, e por essa razão foram inseridos no presente estudo.

Antes de entrar nos detalhes relativos a cada parâmetro em particular, será dada uma breve descrição dos pré-processamentos realizados sobre o sinal de fala.

2.1 Pré-Processamento

Para este trabalho, os sinais de voz foram gravados no formato WAV, a uma taxa de amostragem de 11025Hz, e resolução de 16 bits. Antes do cálculo dos parâmetros acústicos, os sinais passam por uma filtragem passa-altas através de um filtro de pré-ênfase de primeira ordem $H(z) = 1 - 0,95\,z^{-1}$, de modo a equilibrar o seu conteúdo espectral.

Depois da pré-ênfase, o sinal é parametrizado utilizando-se quadros de curta duração. Devido à natureza de alguns dos classificadores utilizados, todas as locuções foram parametrizadas utilizando-se o mesmo número de quadros. Isto foi possível variando-se o tamanho das janelas e ao grau de sobreposição entre elas.

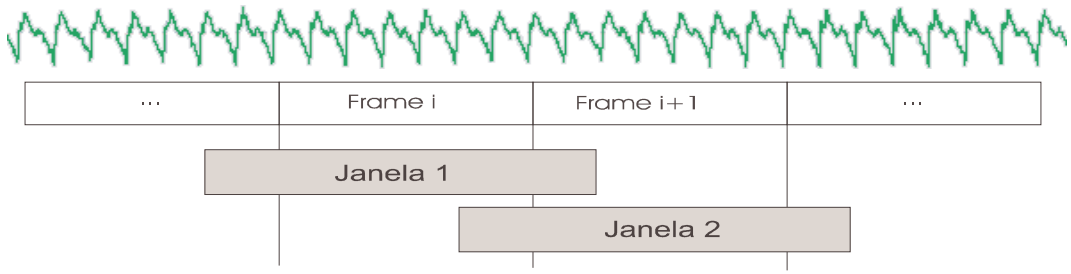


Figura 2.1: *Separação de Quadros*

Para a seleção do intervalo foram utilizadas janelas de Hamming:

$$W[n] = 0,54 - 0,46\cos[(2\pi n)/(Ns-1)] \tag{2.1}$$

$0 < n < Ns - 1$, onde Ns é o número de amostras de cada janela.

2.2 Parâmetros Mel Cepstrais

Os parâmetros mel cepstrais têm obtido grande sucesso tanto em sistemas de reconhecimento de fala como em reconhecimento de locutor [3]. Como são bas-

tante conhecidos, os detalhes de sua implementação não serão colocados nesta dissertação. Para mais detalhes, o leitor interessado pode consultar o artigo original de Davis e Melmertstein [11].

Estes parâmetros foram desenvolvidos tendo em mente a forma como o ouvido humano processa os sons [20]. Duas características são importantes nesta análise: o processamento homomórfico, que modela a forma como o ouvido humano processa a intensidade dos sons, e o banco de filtros não lineares, ligado à forma como a membrana basilar responde às diversas freqüências de um estímulo sonoro.

Estas características são bastante interessantes, pois fazem com que os sistemas "ouçam" da mesma forma que as pessoas. De fato, os resultados experimentais, apresentados adiante, mostram que os parâmetros mel cepstrais apresentam bons resultados, quando comparados com os demais parâmetros testados.

2.3 Parâmetros LPC

Os parâmetros LPC baseiam-se no modelo fonte-filtro para sistemas de produção de voz. Desta forma, diferentemente dos mel cepstrais, os LPC estão relacionados às características de quem está *falando*, e não de quem está *ouvindo*.

Esta técnica baseia-se no princípio de que a fala pode ser modelada como a saída de um sistema linear variante no tempo, excitada por dois tipos de sinais: ruído aleatório para sons não vocálicos e pulsos periódicos para sons vocálicos, como mostrado na Figura 2.2.

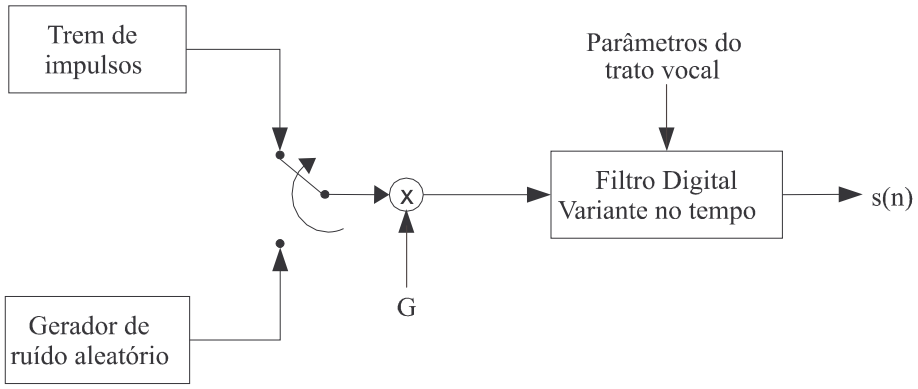


Figura 2.2: *Modelo LPC.*

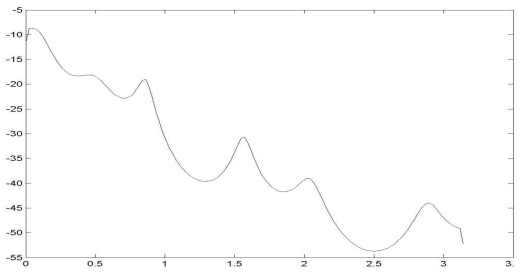
Pode-se mostrar que a função de transferência do filtro variante no tempo é dada por:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2.2}$$

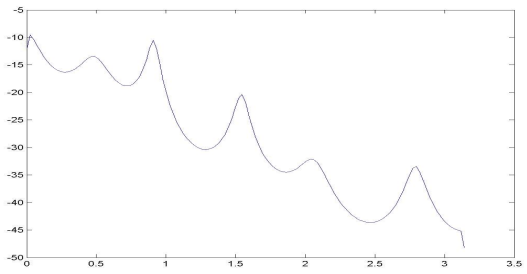
e os coeficientes a_i podem ser calculados eficientemente de forma recursiva através do algoritmo de Levinson-Durbin [3].

A escolha da ordem p do filtro é bastante importante, pois um filtro de ordem muito baixa não consegue modelar todas as características do trato vocal, enquanto um filtro de ordem muito alta começa a modelar também o sinal de excitação, e não apenas o filtro. Como regra de dedo, costuma-se utilizar um par de pólos para cada kHz de faixa do sinal. Como os sinais de voz foram gravados a uma frequência de amostragem de 11025 Hz, escolheu-se $p = 10$.

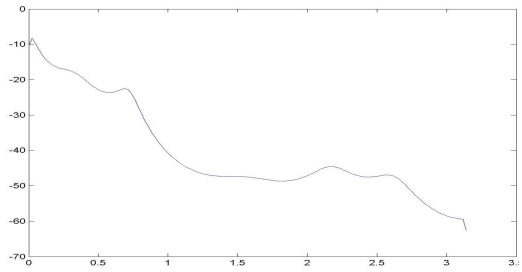
Uma investigação prévia foi realizada com o intuito de verificar a variância intralocutor e interlocutor: foram selecionados dois locutores, e cada um deles gravou duas vezes a vogal /a/ de forma sustentada. Depois disso, foi calculado o espectro LPC de um segmento de 20ms na região de cada locução, e os resultados são mostrados na Figura 2.3.



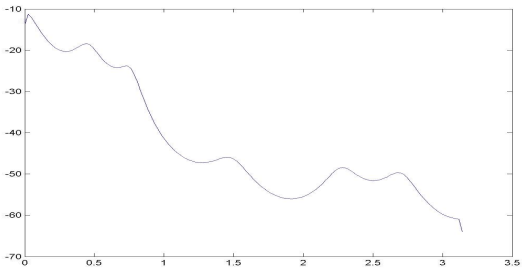
(a) Primeira repetição, locutor 1



(b) Segunda repetição, locutor 1



(c) Primeira repetição, locutor 2



(d) Segunda repetição, locutor 2

Figura 2.3: *Espectros LPC de duas locuções distintas de dois locutores masculinos para a vogal /a/.*

Comparando os gráficos (a) e (b), correspondentes ao locutor 1, e os gráficos (c) e (d), correspondentes ao locutor 2, observa-se uma baixa variação intralocutor, enquanto que comparando os espectros do locutor 1 e os espectros do locutor 2, observa-se uma variação interlocutor muito boa. Desta forma, os

parâmetros LPC também apresentam um bom potencial para a tarefa de reconhecimento de locutor, fato que foi confirmado nos testes realizados.

2.4 Parâmetro Perfil de Energia

Em um trabalho de doutorado desenvolvido para o auxílio no aprendizado à fala para crianças com deficiência auditiva [25], foi utilizado um parâmetro denominado perfil de energia, e que se mostrou bastante discriminativo na separação de sons fricativos e vocálicos sustentados. Com estes resultados, especulou-se sobre a sua possível utilização em um sistema de reconhecimento de locutor, para verificar se estas qualidades se manteriam neste tipo de aplicação.

Os perfis de energia têm por objetivo mostrar a região de concentração da energia do sinal dentro do espectro de freqüências. O algoritmo para o cálculo destes parâmetros é o seguinte:

1. para cada quadro a ser analisado, calcula-se a FFT do sinal, e depois o quadrado do seu módulo (energia em cada banda);

2. calcula-se a energia total do sinal naquele quadro (soma das energias de todas as bandas);

3. escolhe-se o número N de perfis desejados e os limiares par a cada um deles. Por exemplo, escolheu-se $N = 9$, com limiares de 10%, 20%, ..., 90%, respectivamente;

4. partindo da primeira banda do espectro, soma-se a energia de cada banda até que esta atinja o limiar para o primeiro perfil 10%, então o primeiro perfil é a frequência que atingiu 10% da energia total do quadro e assim por diante. Na maioria das vezes não existe um número inteiro de bandas que satisfaça o critério acima, e neste caso existem duas soluções possíveis:

   - escolhe-se a freqüência de uma das bandas (inferior ou superior) como sendo o perfil desejado;
   - faz-se uma interpolação (possivelmente linear) que calcule um valor intermediário entre as frequências das duas bandas.

   No presente estudo foi adotada a segunda alternativa, com interpolação linear.

A seguir são apresentados algumas figuras com exemplos dos perfis de energia para sons vocálicos (Figura 2.4), (/a/, /i/ e /u/) e fricativos (/f/, /s/ e /x/), (Figura 2.5), de um mesmo locutor. Nestes gráficos foram plotados 3 perfis, com

limiares de 25%, 50 % e 75 %. São sinais sustentados, formados por silêncio, locução e silêncio.

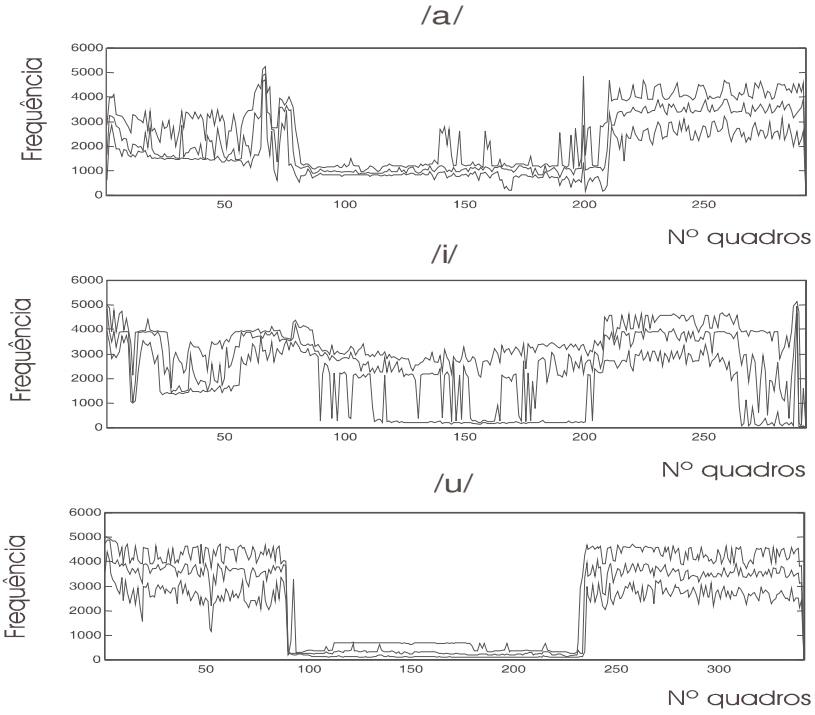


Figura 2.4: *Perfis de energia para as vogais /a/, /i/, /u/, utilizando FFT.*

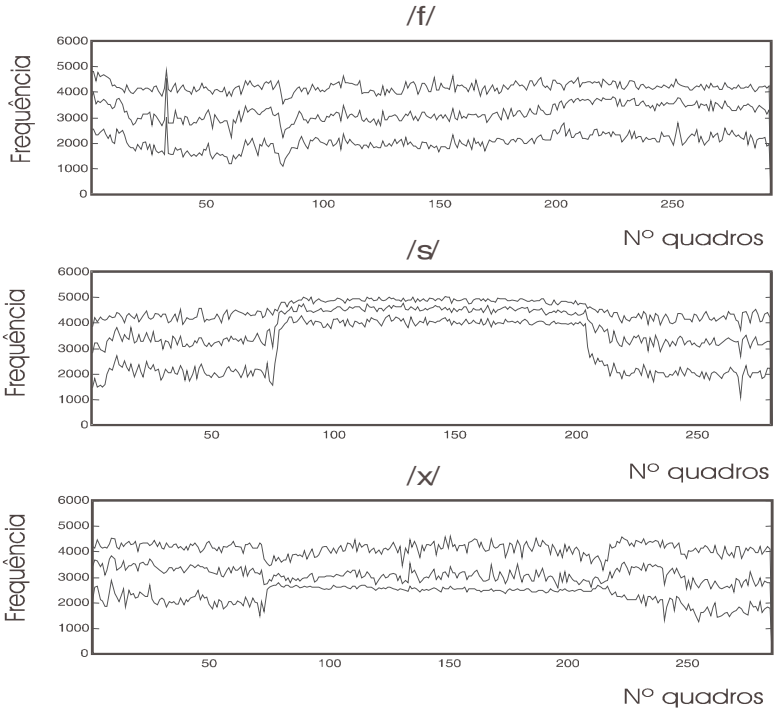


Figura 2.5: *Perfis de energia para as fricativas /f/, /s/, /x/, utilizando FFT.*

Em [25] é sugerido que ao invés de calcular o espectro do sinal diretamente utilizando o algoritmo de FFT, seja utilizado o espectro calculado a partir dos coeficientes LPC. Nas Figuras 2.6 e 2.7 são mostrados os mesmos perfis utilizando o espectro LPC para comparação.

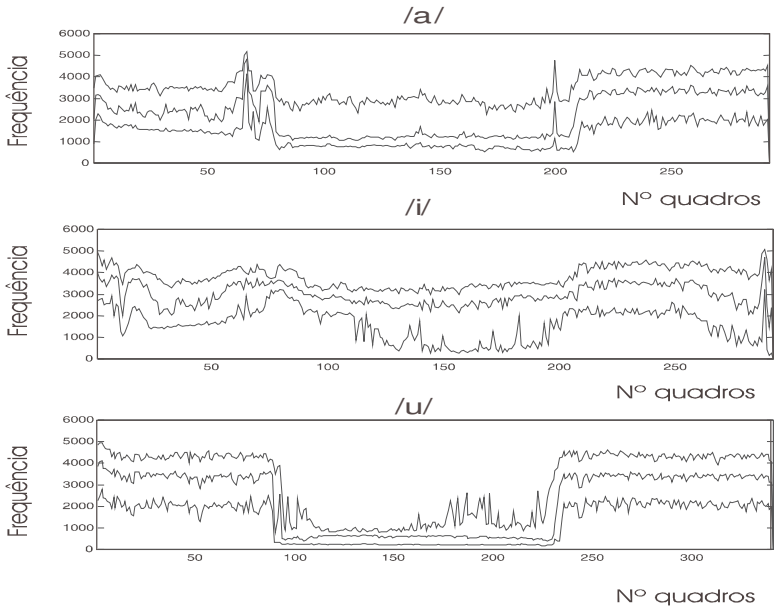


Figura 2.6: *Perfis de energia para as vogais /a/, /i/, /u/, utilizando LPC.*

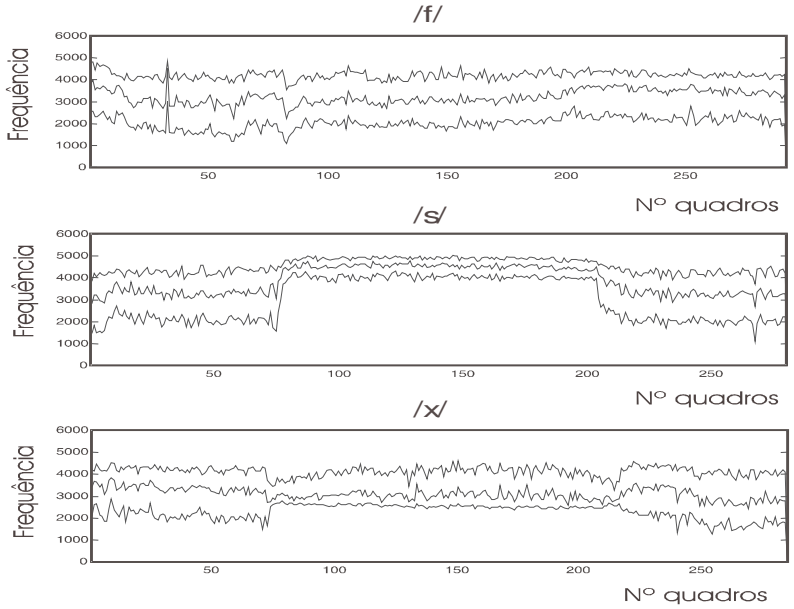
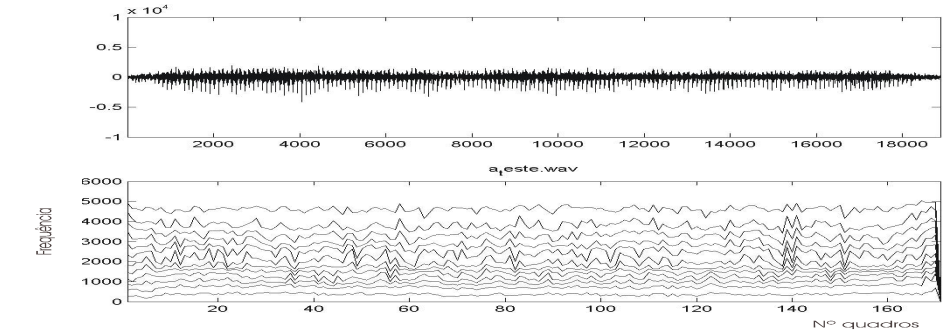


Figura 2.7: *Perfis de energia para as fricativas /f/, /s/, /x/, utilizando LPC.*

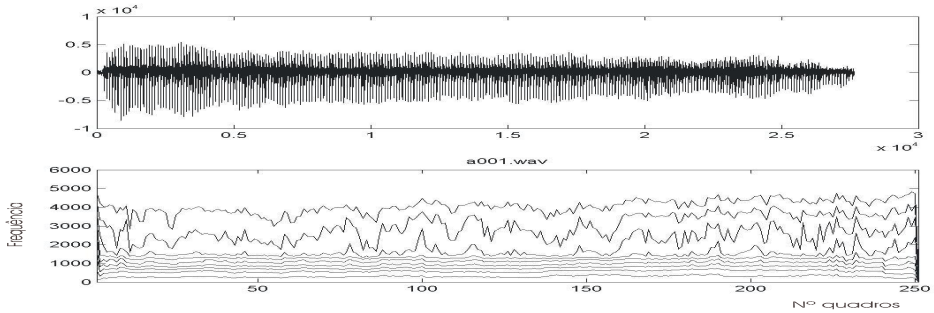
Uma análise destes gráficos, mostra que o espectro LPC é mais suavizado que aquele obtido através da FFT.

Embora pareça ter um bom potencial para a tarefa de reconhecimento de locutor, este parâmetro apresenta alguns problemas: a) é bastante sensível à relação sinal/ruído, b) apresenta uma baixa variabilidade interlocutor. A seguir, estes pontos são analisados com maiores detalhes.

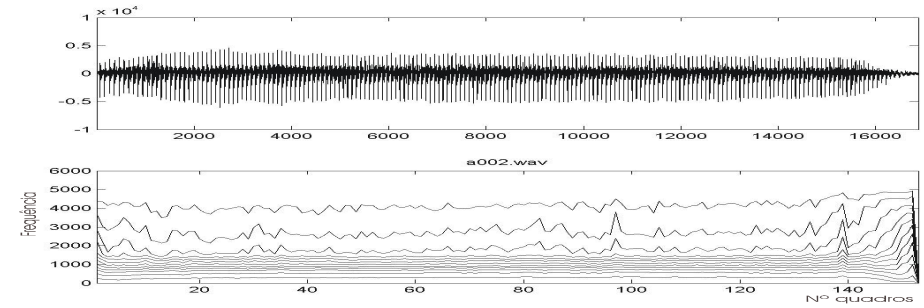
Nos gráficos das Figuras 2.8(a), 2.8(b) e 2.8(c), tem-se três locuções de uma vogal 'a' sustentada, pronunciadas pelo mesmo locutor, com diferentes SNRs.



(a)



(b)



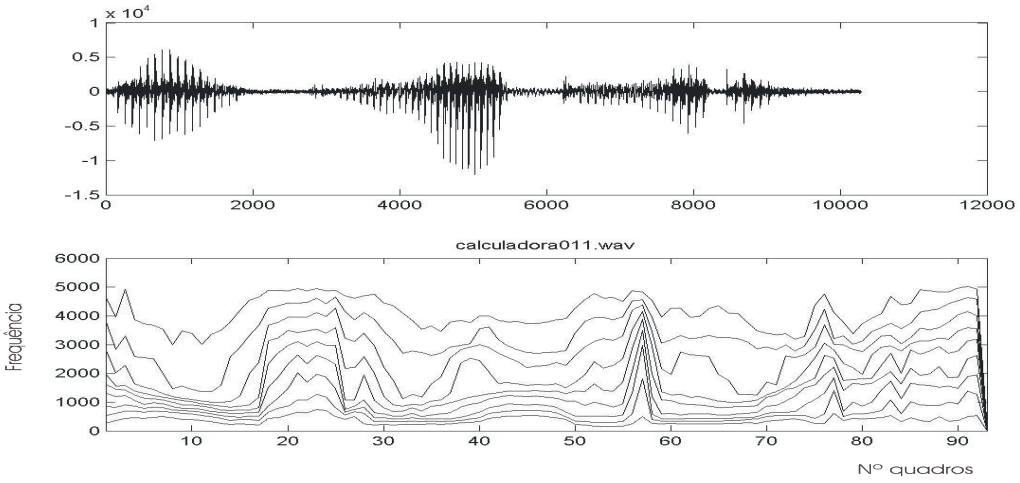
(c)

Figura 2.8: *Variação do perfil de energia com a relação sinal ruído: a) baixa, b)média, e c) alta.*

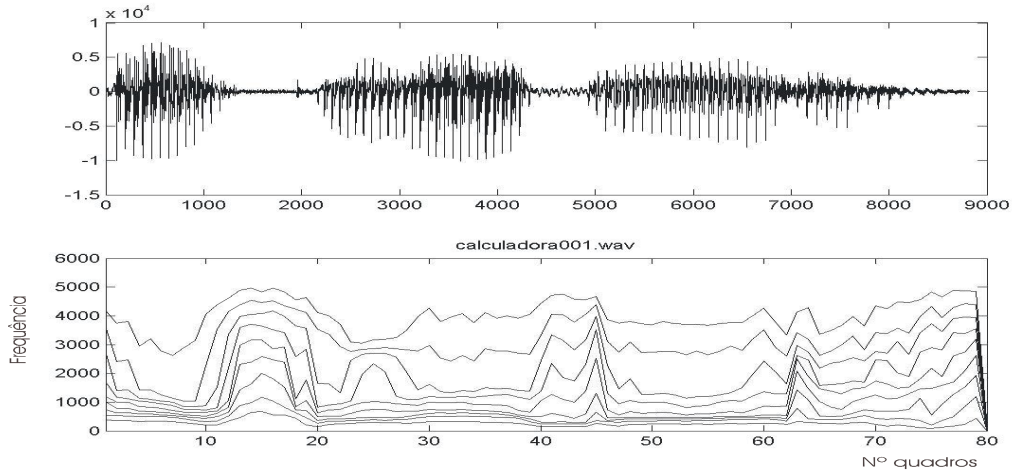
Observa-se que, à medida que a relação sinal/ruído piora, os perfis tendem a se espalhar. Desta forma, estes parâmetros refletem o que está acontecendo com

o sinal: quanto menor o nível de sinal, mais os perfis se aproximam do espectro do ruído de fundo. Esta característica torna estes parâmetros bastante úteis por exemplo na tarefa de determinar a relação sinal/ruído de uma dada locução, mas não servem à tarefa de reconhecimento de locutor, uma vez que esta característica tem impacto direto na variabilidade intralocutor.

Para verificar a variabilidade interlocutor foi realizado o seguinte teste: pediu-se a dois locutores diferentes que pronunciassem a palavra "calculadora", e extraíram-se os perfis de energia para estas duas locuções. Os resultados são mostrados na Figura 2.9.



(a) Locutor 1



(b) Locutor 2

Figura 2.9: *Perfis de energia para a palavra 'calculadora', pronunciadas por dois locutores diferentes.*

Percebe-se que há muita semelhança entre os perfis das duas locuções (baixa variabilidade interlocutor), o que é bastante ruim para diferenciar um locutor de outro. Esta conclusão foi também confirmada nos resultados experimentais, mostrados adiante.

2.5 Pitch

O *pitch* é dado pelo inverso da duração entre os pulsos da excitação do trato vocal, [3]. Desta forma, como cada pessoa tem um trato vocal diferente, o *pitch* é um potencial candidato a parâmetro para a tarefa de reconhecimento de locutor. Entretanto, existem alguns problemas:

- não pode ser medido nas regiões de sons não sonoros (unvoiced);
- varia com a idade;
- varia com o estado emocional: pessoas calmas falam com pitch mais baixo do que pessoas nervosas;
- varia com a potência: quando falamos alto ou gritamos, o pitch sobe.

Estas observações levam à conclusão de que o pitch apresenta alta variabilidade intralocutor, o que é ruim para a tarefa de reconhecimento de locutor. Apenas a título de ilustração, fez-se um teste com uma vogal "e" sustentada, pronunciada duas vezes por um mesmo locutor, uma com volume normal, e outra forçando a glote, para pronunciar mais alto. Verificou-se que falando de forma normal, o pitch médio ficou em torno de 103 Hz, e falando mais alto, o pitch médio pulou para 126Hz, além de flutuar mais; diferença mais do que suficiente para confundir um locutor com outro. Nas Figuras 2.10 e 2.11 tem-se os gráficos do pitch para as duas locuções analisadas.

Uma observação importante a ser feita para a implementação deste parâmetro, é que não se deve utilizar o filtro de pré-ênfase neste caso, pois este aumenta o conteúdo de altas freqüências do sinal, dificultando a determinação precisa do pitch.

Existem vários algoritmos para medição do pitch, e neste trabalho foi utilizado um procedimento clássico baseado na função de autocorrelação, proposto por Rabiner [10].

2.6 Parâmetros Delta

Os parâmetros delta fornecem informação contextual que não é dada pelos parâmetros estáticos, e desta forma estão ligados à dinâmica da fala. Estes parâmetros são

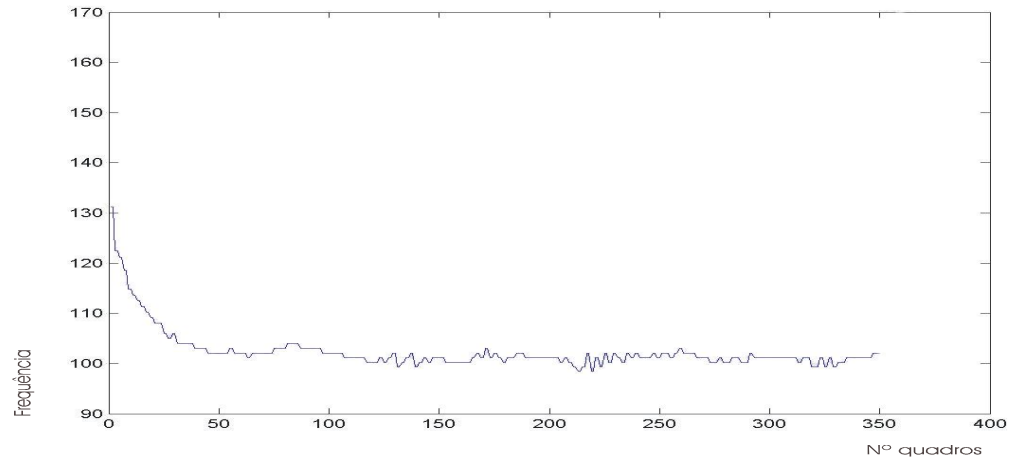


Figura 2.10: *Volume normal, pitch médio de 103Hz.*

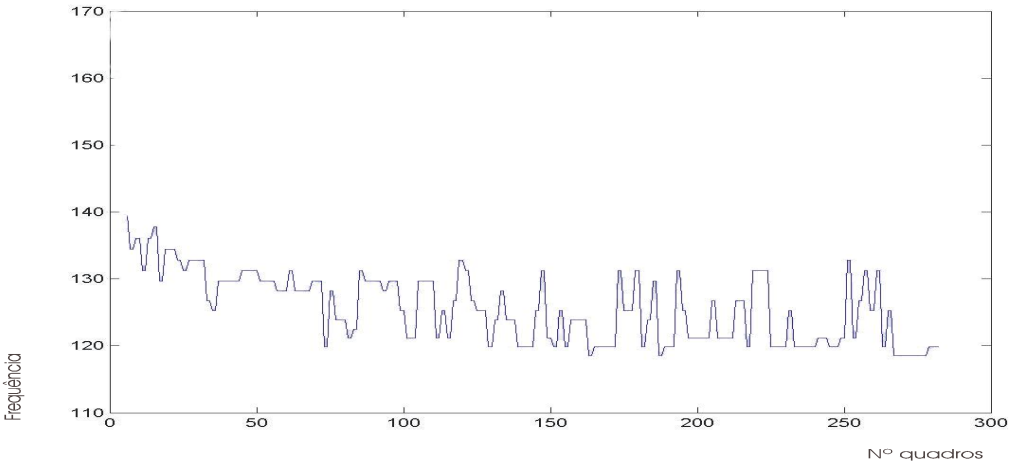


Figura 2.11: *Volume alto, pitch médio de 126Hz.*

calculados a partir da expressão:

$$\Delta(n) = \sum_{k=-k}^{k} kS_i - k(n)/(2k+1) \tag{2.3}$$

onde k é o número de janelas adjacentes a serem consideradas e os S_i são os parâmetros estáticos.

O valor adotado para k foi 5, conforme ensaios descritos em [1].

Capítulo 3

Sistemas de Classificação

Os parâmetros acústicos podem ser vistos como os "ouvidos" de um sistema de reconhecimento de locutor; por esta analogia, os classificadores seriam o "cérebro" do sistema, responsáveis por decidir se uma dada locução corresponde ou não a um determinado locutor.

Neste capítulo são apresentados os sistemas de classificação utilizados neste trabalho: o quantizador vetorial, os modelos ocultos de Markov, as redes neurais e os modelos de misturas de Gaussianas, bem como outras técnicas alternativas encontradas na literatura, mas que não foram implementados neste trabalho.

3.1 Quantização Vetorial

O problema da quantização vetorial pode ser definido da seguinte maneira: dados um conjunto de vetores, uma medida de distorção e um certo número de vetores código, encontrar um conjunto de vetores código e uma partição que resultem na mínima distorção média. Uma vez construído o quantizador vetorial, todos os vetores acústicos de entrada passam a ser representados pelos vetores código [26].

Na literatura, os vetores código são chamados de *codevectors*, e o conjunto destes, *codebook*. A dimensão dos vetores define a *dimensão* do quantizador vetorial, e o número de vetores código, a sua *ordem*. Na Figura 3.1 tem-se um esboço de um quantizador vetorial de dimensão 2 e ordem 16.

Para este trabalho foi utilizado o algoritmo LBG [26], com inicialização via "splitting" para a construção dos codebooks, e o método de busca exaustiva para a quantização vetorial dos vetores acústicos. Como métrica, foi utilizada a distância euclidiana.

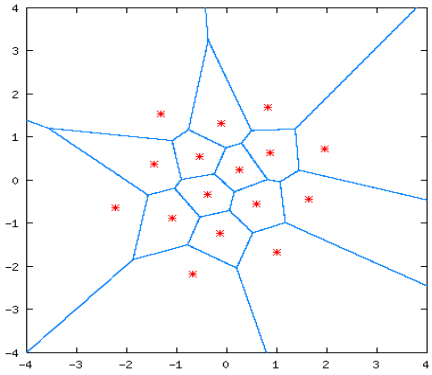


Figura 3.1: *Centróides (asteriscos) e partições (linhas) de um quantizador vetorial de dimensão 2 e ordem 16 [29].*

Quantização vetorial e reconhecimento de locutor

Para utilizar um quantizador vetorial como um sistema de reconhecimento de locutor, inicialmente todas as locuções devem ser parametrizadas com o mesmo número de quadros. Depois, os vetores acústicos de cada quadro da locução são agrupados em um único vetor, cuja dimensão é a dimensão do quantizador vetorial. Desta forma, cada locução passa a ser um ponto em um espaço cuja dimensão é o número de quadros vezes a dimensão de cada vetor acústico.

Espera-se que as locuções de um determinado locutor ocupem uma região mais ou menos definida neste espaço multidimensional. Desta forma, pode-se associar a cada locutor um vetor código no centro de sua "nuvem de pontos". Entretanto, esta estratégia leva a superfícies de separação que são hiperplanos no espaço multidimensional (veja Figura 3.1). Desta forma, é melhor associar a cada locutor não um, mas vários centróides de modo a permitir superfícies de separação mais complexas, como ilustrado na Figura 3.2:

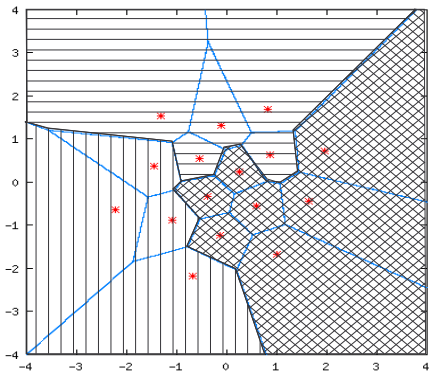


Figura 3.2: *Superfícies de separação para um quantizador vetorial de ordem 16 associado a um sistema de 3 locutores.*

O problema passa a ser então associar os vetores código a cada um dos locutores, pois sempre haverá casos de sobreposição das regiões de dois (ou mais) locutores. Para resolver este problema, o seguinte algoritmo foi utilizado:

1. os centróides aos quais estão associadas locuções de uma única pessoa são associados a ela;

2. os centróides associados a locuções de mais de uma pessoa são associados à pessoa que tem maior número de locuções na partição deste centróide.

Depois deste processo, cada locutor tem associado a ele (ela) um determinado número de centróides. O processo de identificação consiste então em adquirir uma locução do locutor que se quer testar, parametrizá-la e apresentá-la ao quantizador vetorial. Este vai associar um vetor código a este locutor, e uma consulta à tabela de vetores código versus locutor termina a identificação.

O problema de reconhecimento dependente e independente de texto é facilmente gerenciável com esta técnica: para o reconhecimento dependente de texto, são utilizadas para a confecção do quantizador vetorial somente locuções de uma mesma palavra (ou sentença) para cada locutor. Observe que as palavras/sentenças podem ser diferentes para cada locutor, mas devem ser as mesmas para um mesmo locutor. Já para o reconhecimento independente de texto usam-se locuções diversas para um mesmo locutor.

É importante observar que nesta técnica, a locução de *qualquer* pessoa irá ser necessariamente mapeada em um dos vetores código.

No caso de utilização em sistemas *open-set*, se acontecer de uma pessoa querer se passar pela ovelha correspondente (o que geralmente é o caso: os impostores tentam imitar a voz da pessoa a quem querem se fazer passar), o sistema irá aceitá-la, pois a voz do impostor é mais parecida com a desta ovelha do que das outras. Desta forma, os "lobos *open-set*" devem ser explicitamente modelados nesta técnica. Isto pode ser feito adicionando-se mais um locutor (o lobo open-set), e a este associam-se locuções de diversas pessoas que não fazem parte do conjunto close-set.

Obviamente esta solução depende da quantidade de pessoas disponíveis para modelar o "lobo *open-set*" e das características destas: por exemplo, se for utilizado um locutor com características muito próximas às de um locutor do conjunto close-set, a taxa de erro deve aumentar significativamente. Desta forma, é importante selecionar cuidadosamente as pessoas para compor o modelo dos lobos open-set, o que não é uma tarefa fácil. Apesar destes problemas, a inclusão deste modelo adicional melhora bastante o desempenho do sistema, ou seja, é melhor usá-lo do que não usá-lo.

3.2 Modelos Ocultos de Markov

Os modelos ocultos de Markov vêm sendo utilizados com sucesso no problema de reconhecimento de fala, e podem ser facilmente adaptados para a tarefa de reconhecimento de locutor.

No reconhecimento de fala, cada modelo representa uma palavra ou subunidade fonética; já no reconhecimento de locutor, cada modelo representa um locutor diferente. Desta forma, para a tarefa de reconhecimento de locutor, apresenta-se a cada modelo locuções de um único locutor.

Para o reconhecimento dependente de texto, treina-se cada modelo com locuções de uma mesma palavra/sentença, e para o reconhecimento independente de texto, apresentam-se locuções de diversas palavras/sentenças, mas sempre de um mesmo locutor.

Na etapa de reconhecimento, teria-se dois procedimentos a seguir, dependendo da tarefa:

Verificação: A locução é aplicada ao modelo referente à ovelha que o locutor afirma ser; se o valor da verossimilhança estiver abaixo de um limiar pré-determinado, o locutor é considerado um lobo e rejeitado; caso contrário, é aceito.

Identificação: aplica-se a locução a todos os modelos, e aquele que apresentar a maior verossimilhança é identificado como o sendo o locutor.

Para uso em sistemas *open-set* é necessário criar um modelo explícito para os lobos, assim como foi feito para o quantizador vetorial.

Neste trabalho foram utilizados HMMs contínuos, com 3 gaussianas por mistura, e 6 estados por modelo. Utilizou-se o modelo de Bakis, permitindo um salto de dois estados, como mostrado na Figura 3.3.

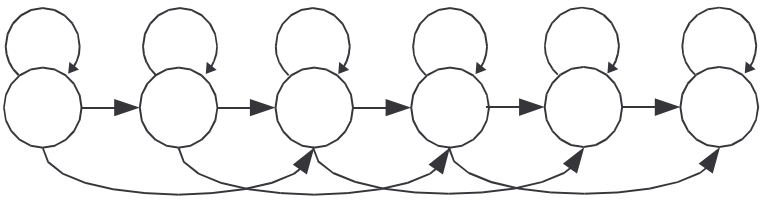


Figura 3.3: *Estrutura do HMM utilizado.*

O treinamento foi realizado através do algoritmo Baum-Welch, e o reconhecimento, através do algoritmo de Viterbi.

3.3 Redes Neurais Artificiais

O termo Rede Neural originalmente refere-se a uma rede de neurônios interconectados. Hoje esse termo é usado para designar qualquer arquitetura computacional que consista de interconexões paralelas de elementos simples de processamento, os neurônios. Estas estruturas têm sido utilizadas com grande sucesso nas tarefas de previsão de séries temporais e reconhecimento de padrões, [21].

Existem várias topologias possíveis para uma rede neural (que dependem de como os neurônios são conectados), bem como diferentes tipos de algoritmos de treinamento (supervisionado, não supervisionado). Neste trabalho foi escolhida a rede Multilayer Perceptron, treinada através do algoritmo Backpropagation [21].

As Multilayer Perceptrons (MLP) são redes em camadas, onde todos os neurônios de uma camada se conectam com todos os neurônios da camada seguinte. Uma rede deste tipo tem pelo menos duas camadas: a camada de saída e a camada escondida, como mostra a Figura 3.4:

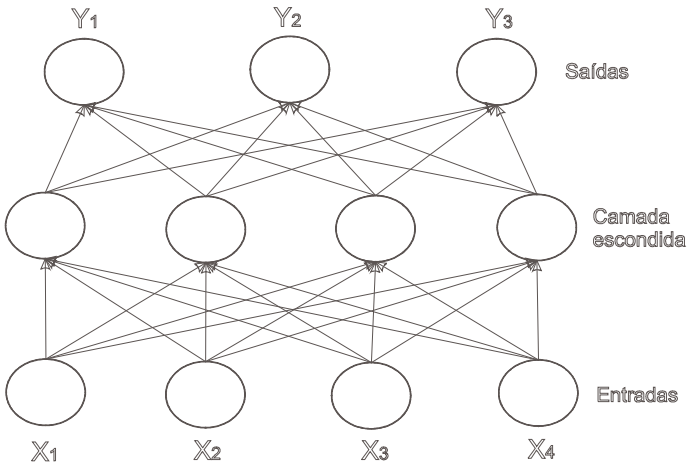


Figura 3.4: *Multilayer Perceptron com uma camada escondida*

O processo de treinamento consiste em apresentar vários pares estímulo-resposta, e alterar os parâmetros da rede para que esta passe a ter o comportamento desejado.

Para utilizar esta estrutura na tarefa de reconhecimento de locutor, associa-se a cada neurônio de saída um locutor diferente. Na etapa de treinamento, para cada locução de treinamento de um dado locutor, orienta-se a rede para aumentar o valor da saída do neurônio correspondente ao locutor que pronunciou a locução, ao mesmo tempo em que minimiza a saída dos outros neurônios. O processo se repete a cada locução de treinamento, até que todas as locuções tenham sido apresentadas à rede. Todo este procedimento compõe uma *época* de treinamento. O sistema é treinado durante várias épocas, até que seja atingida

uma convergência.

O procedimento de reconhecimento é bem parecido com aquele adotado para os HMMs e o quantizador vetorial.

Identificação: o locutor é identificado como sendo aquele correspondente à saída mais alta da rede.

Pode-se eventualmente utilizar a diferença entre a maior saída e a segunda maior saída como um fator de segurança. Se for menor que um determinado limiar, significa que o sistema está em dúvida, e pode por exemplo, pedir ao locutor que repita a locução.

Novamente, a exemplo do que foi feito com o quantizador vetorial e os HMMs, para o reconhecimento independente de texto treina-se a rede com locuções de diversas palavras/sentenças de um mesmo locutor.

Para problemas *open-set* associa-se um neurônio de saída aos "lobos open-set", treinado com vários locutores que não pertençam ao conjunto close-set. Neste caso, a esperança é de que, no caso de um impostor, a saída correspondente aos lobos open-set seja maior que a saída correspondente à ovelha em questão.

Embora o número de entradas e de neurônios na camada de saída sejam facilmente determinados, o mesmo não se pode dizer do número de neurônios da camada escondida: um número muito baixo resulta em uma rede sub-dimensionada, que não consegue resolver o problema, e um número muito alto resulta em uma rede super-dimensionada, que tem um bom desempenho para o material de treinamento, mas um desempenho pobre para dados novos (baixa capacidade de generalização), efeito conhecido na literatura como *overfitting*. Como não há procedimentos claros na literatura para se determinar o número ótimo de neurônios na camada escondida, optou-se por fazer uma varredura, e chegou-se ao número de 300 neurônios como um bom compromisso para o problema em questão.

3.4 Modelos de Misturas de Gaussianas

Os modelos de misturas de gaussianas (GMM), muitas vezes referenciados como modelos ocultos de Markov de um único estado, são considerados o estado da arte para reconhecimento de locutores [27].

Os modelos de mistura são um tipo de modelo de densidade de probabilidade multimodal que compreende um certo número de funções, geralmente gaussianas, que são combinadas para fornecer uma densidade multimodal, para modelar o conteúdo acústico do sinal de voz, como mostrado na Figura 3.5.

O GMM pode ser descrito da seguinte forma:

Uma densidade de misturas de gaussianas é uma soma de pesos de M densidades componentes.

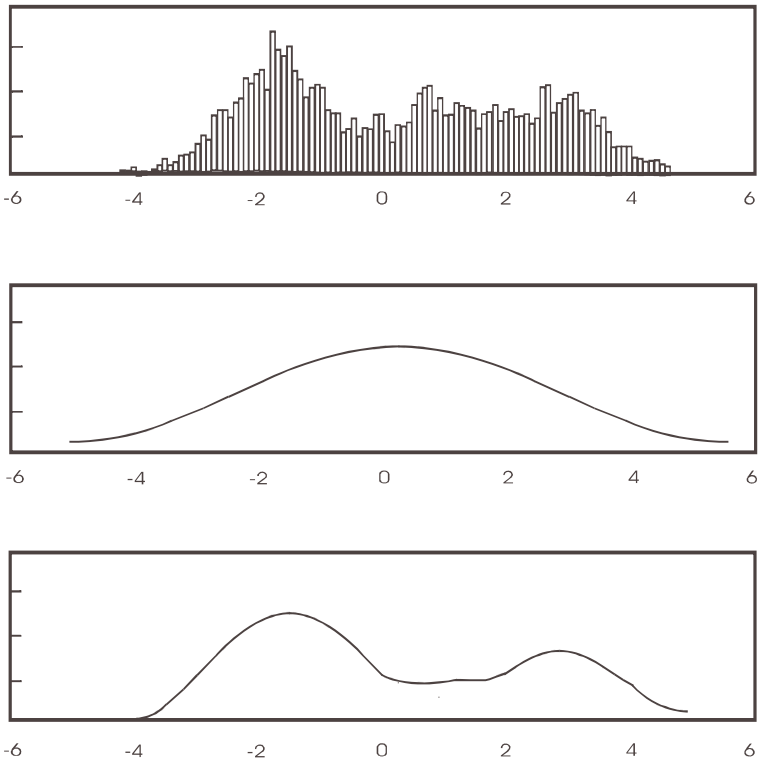


Figura 3.5: *Sistema GMM com 1 e 10 gaussianas [27]*

$$p(\overrightarrow{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\overrightarrow{x}) \tag{3.1}$$

onde:

- $\overrightarrow{x}$ é um vetor aleatório D-dimensional,
- b_i $(\overrightarrow{x})$, i=1,...,M são as densidades componentes,
- p_i i=1,...,M: são os pesos da mistura.

Cada densidade componente é uma função Gaussiana D-dimensional da forma:

$$b_i(\overrightarrow{x}) = (1/(2\Pi)^{D/2}|\Sigma_i|^{0,5})) \exp(-0,5(\overrightarrow{x} - \overrightarrow{\mu_i})' \sum_{i}^{-1}(\overrightarrow{x} - \overrightarrow{\mu_i})) \tag{3.2}$$

onde:

- $\overrightarrow{\mu_i}$: é o vetor média
- $\sum_i$: é a matriz de covariância

Os coeficientes p_i devem satisfazer à condição:

$$\sum_{i=1}^{M} p_i = 1; \tag{3.3}$$

O modelo λ é então dado por:

$$\lambda = \{p_i, \overrightarrow{\mu_i}, \Sigma_i\}, i = 1, ..., M \tag{3.4}$$

Para identificação de locutor, cada locutor é representado por um GMM e referenciado pelo seu modelo λ.

Os GMM's podem ter as seguintes formas:

- Covariância modal: uma matriz de covariância por componente Gaussiana,
- Grand Covariância: uma única matriz de covariância para todos as componentes gaussianas de um dado modelo,
- Covariância global: uma única matriz de covariância compartilhada pelos modelos de todos os locutores,

A matriz de covariância pode ainda ser cheia ou diagonal. Neste trabalho foram usadas matrizes de covariância modais e diagonais.

Estimação dos parâmetros por máxima verossimilhança

A partir da amostras de fala de um determinado locutor, estima-se os parâmetros λ do GMM os quais melhor representem a distribuição dos vetores acústicos de treinamento. Embora existam muitas técnicas para se fazer isto, a mais popular é a da estimação através de máxima verossimilhança (Maximun Likelihood - ML) [27].

O objetivo da estimação por máxima verossimilhança é encontrar os parâmetros do modelo que maximizem a verossimilhança do GMM, dado os vetores de treinamento. Para uma sequência de T vetores de treinamento $\mathbf{x} = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_T}\}$, a verossimilhança do GMM pode ser escrita como:

$$P[\mathbf{x}|\lambda] = \prod_{t=1}^{T} P[\overrightarrow{x_t}|\lambda] \tag{3.5}$$

A idéia do algoritmo EM é começar com um modelo inicial λ, estimar um novo modelo $\overline{\lambda}$, tal que $P[\mathbf{x}|\overline{\lambda}] \geq P[\mathbf{x}|\lambda]$. O novo modelo torna-se então o modelo inicial para a próxima iteração e o processo repete-se até que um limiar de convergência seja atingido.

Em cada iteração ao algoritmo EM as seguintes fórmulas são usadas, as quais garatem um aumento monotônico no valor da verossimilhança do modelo:

Pesos das misturas:

$$\overline{p_i} = \frac{1}{T\sum_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)} \tag{3.6}$$

Médias:

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)\overrightarrow{x_t}}{\sum_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)} \tag{3.7}$$

Variâncias:

$$\overline{\sigma_i}^2 = \frac{\sum_{t=1}^{T} p(i|\overrightarrow{xt}, \lambda)xt^2}{\sum_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda) - \overline{\mu_i}^2} \tag{3.8}$$

onde σ_i^2, x_t e μ_i referem-se a elementos arbitrários dos vetores $\overrightarrow{\sigma_i}^2$, $\overrightarrow{x_t}$ e $\overrightarrow{\mu_i}$ respectivamente.

A probabilidade a posteriori para a classe acústica é dada por:

$$p(i|\overrightarrow{x_t}, \lambda) = \frac{p_i b_i(\overrightarrow{x_t})}{\sum_{k=1}^{M} p_k b_k(\overrightarrow{x_t})} \tag{3.9}$$

Identificação do Locutor

Para a identificação de locutor, um grupo de locutores $l_1, l_2, ..., l_S$ é representado pelos GMMs, λ_1,λ_2,...,λ_S. O objetivo é encontrar o modelo de locutor que tenha a máxima probabilidade a posteriori para uma dada seqüência de observação.

Formalmente:

$$\widehat{S} = \underset{1\leq k\leq S}{\operatorname{argmax}} P[\lambda_k|\mathbf{x}] = \underset{1\leq k\leq S}{\operatorname{argmax}} \frac{P[\mathbf{x}|\lambda_k]P[\lambda_k]}{P[\mathbf{x}]} \tag{3.10}$$

onde a segunda equação é devida à regra de Bayes. Assumindo que os locutores são equiprováveis ($P[\lambda_k] = 1/S$) e notando que $P[\mathbf{x}]$ é a mesma para todos os modelos, a regra de classificação pode ser simplificada para:

$$\widehat{S} = \underset{1 \leq k \leq S}{\operatorname{argmax}} \sum_{t=1}^{T} \log P[\overrightarrow{x_t}|\lambda_k] \tag{3.11}$$

onde $P[\overrightarrow{x_t}|\lambda_k]$ é dada por 3.1

Pode-se afirmar que o processo de verificação e identificação de locutor são similares aos do HMM.

Verificação: para esta tarefa deve ser criado um modelo conhecido na literatura como UBM (*Universal Background Model*), que consiste em um modelo treinado com todos os locutores do conjunto de lobos. A locução é aplicada tanto ao modelo referente à ovelha que o locutor afirma ser quanto ao modelo UBM; se o valor da diferença entre as verossimilhanças estiver abaixo de um limiar pré-determinado, o locutor é considerado um lobo e rejeitado; caso contrário, é aceito.

Identificação: aplica-se a locução a todos os modelos, e aquele que apresentar a maior verossimilhança é identificado como o sendo o locutor.

3.5 Outras Técnicas Utilizadas em Reconhecimento de Locutor

Modelo AR-Vector

O AR-vetor é uma extensão do LPC, no entanto a predição não é através das amostras, mas sim da diferença entre elas, [17].

A ordem p do modelo AR-Vector para uma sentença de N vetores de dimensão m, no domínio do tempo, é dada por:

$$X_n = \sum_{k=1}^{p} A_k X_{n-k} + E_n \tag{3.12}$$

onde X_n e E_n são vetores m dimensionais, o E representa o erro de predição e A_k é uma matriz de predição de ordem $m \times m$. Pode-se representar a matriz de predição por uma de ordem $m \times (p+1)$, ficando $A = [A_0 A_1 A_2 \dots A_p]$, sendo A_0 uma matriz identidade.

Dos vetores X_n, pode-se definir uma estimativa de autocorrelação:

$$R_k = \sum_{n=0}^{N-k} X_n X_{n+k}^T \tag{3.13}$$

onde N é o número de vetores X. O resultado de R_k resulta em uma matriz m x m.

Os A_k são obtidos resolvendo a equação:

$$\begin{pmatrix} R_0 & R_{1^T} & ... & R_{p-1}^T \\ R_1 & R_0 & ... & R_{p-2} \\ ... & ... & ... & ... \\ R_{p-1} & R_{p-2} & ... & R_0 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ ... \\ A_p \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ ... \\ R_p \end{pmatrix} \tag{3.14}$$

Da equação acima, se for definido a matriz de autocorrelação Toeplitz como $\mathbf{R}$ a matriz de coeficientes como A e a matriz de autocorrelação do lado direito como R, tem-se:

$$\mathbf{RA} = R \Longrightarrow \mathbf{A} = \mathbf{R}^{-1}R \tag{3.15}$$

onde $\mathbf{R}$ é a matriz de Toeplitz, e como se sabe, há o algoritmo de recursão Levinson-Durbin para sua solução.

A utilização do AR-vetor em reconhecimento de locutor necessita de uma medida para avaliar a semelhança entre dois modelos autoregressivos. Uma distância extensamente usada é a de Itakura, [17], que provê a distância entre dois pólos LPC, baseado em coeficientes de predição linear e matriz de autocorrelação.

A medida se faz da seguinte forma:

Dado um modelo de locutor A e um modelo de locutor B, sendo um deles o modelo de referencia e o outro que se deseja testar, calcula-se 3 distancias:

- Distância de B para A:

$$d(B, A) = \log(\mathrm{tr}[\frac{AR_BA^T}{BR_BB^T}]) \tag{3.16}$$

- Distância de A para B:

$$d(A, B) = \log(\mathrm{tr}[\frac{BR_AB^T}{AR_AA^T}]) \tag{3.17}$$

- Distância Simétrica:

$$d_{sim} = \frac{1}{2}(d(B, A) + d(A, B)) \tag{3.18}$$

Em função destas distâncias serão realizados os testes de "Verificação de Locutor", onde se o locutor estiver acima de um determinado limiar, será aceito e caso contrário recusado.

O modelo AR-Vetor produz um modelo suavizado das características do locutor, capturando informações de sua dinâmica.

Distância Bhattacharyya

A distância Bhattacharyya compara o vetor média e a matriz de covariância estimados a partir do segmento sob teste com aqueles do locutor que deseja-se testar. Esta é dada por:

$$d_B^2 = \frac{1}{2}\ln\frac{|\ \frac{C_i+C_j}{2}\ |}{|\ |C_i\ |^{0.5}|\ C_j\ |^{0.5}} + \frac{1}{8}(\mu_i-\mu_j)^T(\frac{C_i+C_j}{2})^{-1}(\mu_i-\mu_j) \tag{3.19}$$

onde: C_i são as matrizes de covariância, e μ_i são as médias.

O segundo termo da expressão (3.19) é a distância de Mahalanobis, a qual considera apenas distância entre os vetores média (supõe que as matrizes de covariância são as mesmas para todas as distribuições).

Capítulo 4

Resultados Experimentais

Nete capítulo são apresentados os resultados das investigações realizadas. Estas foram realizadas em três etapas:

- Em uma etapa preliminar foram feitos testes com vários parâmetros acústicos e classificadores para determinar qual ou quais levariam a um melhor desempenho. Com este objetivo, quatro tipos de parâmetros (mel-cepstrais, LPC, perfil energia, pitch) e quatro classificadores (quantizadores vetoriais, HMMs, redes neurais e GMMs), foram testados na tarefa de identificação independente de texto.
- Como os GMMs são apontados pela literatura como o estado da arte, foram realizados testes para verificar o seu desempenho, tanto nas tarefas de verificação como de identificação de locutor. Os resultados dos testes de identificação irão também servir como base para comparação dos testes finais, descritos a seguir.
- Na etapa final é avaliada a contribuição dos trechos sonoros e não sonoros do sinal de fala para a tarefa de identificação de locutor, sendo esta uma contribuição original deste trabalho.

4.1 Testes Preliminares

O objetivo destes testes é investigar os seguintes pontos:

- qual ou quais parâmetros acústicos são melhores para a tarefa de reconhecimento de locutor?
- qual o quais classificadores têm um melhor desempenho na tarefa de reconhecimento de locutor?

Antes de apresentar os resultados obtidos será feita uma breve descrição da base de dados utilizada nestes testes preliminares.

Base de Dados

A base de dados utilizada nos ensaios preliminares é formada por 30 locutores, organizados em 2 grupos:

Close-set: 10 locutores (6 homens e 4 mulheres), onde cada locutor pronunciou 10 palavras distintas, repetidas 10 vezes, totalizando 1000 locuções;

Open-set: 20 locutores (10 homens e 10 mulheres), onde cada locutor pronunciou 5 palavras apenas 1 vez, totalizando 100 locuções.

As palavras que compõem o vocabulário desta base de dados são mostradas na Tabela 4.1.

Tabela 4.1: *Vocabulário da base de dados*

Acrobat	Calculadora
Corel Draw	Documentos
Internet	Lixeira
Matlab	Outlook
Power Point	Windows

As locuções foram gravadas no formato WAV, sem compressão, a uma taxa de amostragem de 11025Hz e 16 bits de resolução, em ambiente de escritório.

Todas as locuções foram editadas para eliminar os silêncios inicial e final, ficando cada locução com duração entre 0,7s e 1s.

Metade das locuções de cada locutor do conjunto *close-set* foi usada para treinar o sistema, e a outra metade para os testes. As locuções do conjunto *open-set* foram utilizadas para os testes *open-set*.

Testes para avaliação dos parâmetros acústicos

Estes testes foram realizados para avaliar qual ou quais parâmetros acústicos levam a um melhor desempenho na tarefa de reconhecimento de locutor. Como citado anteriormente, idealmente devem ter uma baixa variabilidade intralocutor e uma alta variabilidade interlocutor. Da análise prévia realizada no Capítulo 2, foram identificados os parâmetros LPC e os mel-cepstrais como os melhores, o que realmente foi verificado.

Para estes testes foram utilizadas somente as locuções *close-set*, com 50% delas para o treinamento do quantizador e 50% para os testes. Como classificador Foi utilizado um quantizador vetorial com codebook de 128 vetores.

As combinações de parâmetros testadas, bem como os resultados obtidos são mostrados na Tabela 4.2:

Tabela 4.2: *Combinações de parâmetros utilizados nos testes de identificação de locutor independente de texto para o quantizador vetorial.*

Parâmetros	Taxa de erros
LPC	19,6%
Mel	15,8%
Perfil Energia	34,6%
Mel + Pitch	20,6%
Mel + Perfil Energia + LPC	31,9%
Mel + Delta Mel	40,0 %
Mel + Delta Mel + Delta-Delta-Mel	39,0%

Conclusões

Pode-se observar que, dentre os parâmetros testados, os melhores foram os LPC e os mel-cepstrais, e que os parâmetros pitch e perfil energia não tiveram um bom desempenho, confirmando as hipóteses feitas no Capítulo 2. Ainda notou-se que os parâmetros delta tiveram uma influência negativa no desempenho do sistema.

Com estes resultados, optou-se por utilizar apenas os parâmetros mel-cepstrais para os testes a seguir.

Testes para avaliação dos classificadores

Uma vez determinado qual o parâmetro acústico é o mais adequado, passou-se à investigação sobre os classificadores. Foram testados os seguintes classificadores: quantizadores vetoriais, HMMs, redes neurais e GMMs.

Para cada classificador foram realizados vários testes com diferentes configurações para determinar, em cada caso, qual a melhor, sendo apresentados nesta dissertação apenas aquelas que apresentaram os melhores resultados.

Inicialmente pensou-se em fazer apenas testes de identificação de locutor *closed-set*. Entretanto, a alta taxa de acertos nestes testes (100% na maioria dos casos) sugeriu que fossem feitos também testes de identificação de locutor *open-set*. A seguir são apresentados os resultados destes dois testes para cada classificador.

Quantizador vetorial

O tamanho do codebook determina a complexidade das superfícies de separação das partições referentes a cada locutor. Por outro lado, o material disponível para cada locutor limita este tamanho máximo. Desta forma, o projeto do quantizador é um compromisso entre estes dois aspectos. Nos testes realizados com este classificador, observou-se que os melhores resultados foram obtidos com codebooks de 128 vetores.

Neste caso, a taxa de erros para os testes *close-set* foi de 15,80% e 21,45% para os testes *open-set*.

Modelos Ocultos de Markov

Para este classificador, os melhores resultados foram alcançados com 6 estados por modelo, e misturas de 3 gaussianas por estado.

Para identificação independente de texto, obteve-se uma taxa de erro de 0 % para os testes *close-set* e 0,9 % nos testes *open-set*.

Redes Neurais

Como a aquitetura da rede é fixa, torna-se necessário também parametrizar todas as locuções (tanto de treinamento como de teste) com o mesmo número de quadros. Ainda, a rede neural deverá ter um número de entradas igual ao número de quadros vezes a dimensão de cada vetor acústico. O número de saídas deve ser igual ao número de locutores que se quer reconhecer, e o número de neurônios da camada escondida deve ser suficiente para resolver o problema.

Neste trabalho, como as locuções foram dividdas em 43 quadros, e cada quadro foi parametrizados com parâmetros mel-cepstrais de dimensão 12, a rede tem $43 \times 12 = 516$ entradas. Para os testes *close-set* a rede tem 10 saídas, e para os testes *open-set*, 11 saídas. Quanto ao número de neurônios na camada escondida, após vários testes, escolheu-se 300.

Para identificação independente de texto, obteve-se uma taxa de erro de 0% no caso *close-set*, e 6% no caso *open-set*.

Modelos de Mistura de Gaussianas (GMM)

Para este classificador, o parâmetro a ser ajustado é o número de gaussianas na mistura. Nos testes realizados, o melhor desempenho foi verificado com 32 gaussianas. Com esta configuração conseguiu-se uma taxa de erros de 0% no caso *close-set*, e 6% no caso *open-set*.

Conclusões

Um resumo dos resultados da comparação entre os classificadores pode ser visto na Tabela 4.3:

Tabela 4.3: *Desempenho dos diversos classificadores (taxa de erros).*

Classificador	Close-set	Open-set
Quantizador Vetorial	15,80%	21,45%
HMM	0 %	0,9 %
Rede Neural	0 %	6 %
GMM	0 %	6 %

Dentre os classificadores testados, os Modelos Ocultos de Markov apresentaram os melhores resultados, seguidos do GMM e das redes neurais. Ainda, os quantizadores vetoriais apresentaram um desempenho muito inferior aos demais classificadores.

Estes resultados são um pouco diferentes daqueles observados na literatura, que aponta os GMMs como o estado da arte para a tarefa de reconhecimento de locutor. O tamanho reduzido da base de dados utilizada talvez explique esta diferença.

4.2 Testes com GMM

Apesar dos testes preliminares terem apontado os HMMs como os classificadores de melhor desempenho, optou-se por seguir as recomendações da literatura e adotar os GMMs para uma análise mais detalhada. Para isto foram feitos testes de verificação e identificação de locutor, que são descritos a seguir.

A exemplo do que foi feito nos testes preliminares, será feita uma breve descrição da base de dados utilizada para estes testes finais antes da apresentação dos resultados dos experimentos.

Base de Dados

Como a base de dados utilizada até o momento não possui material de treinamento suficiente para cada locutor, houve a necessidade de se utilizar uma outra base de dados para efeito de comparação com os resultados obtidos na literatura.

Utilizou-se então um subconjunto de locutores de uma base de dados desenvolvida através de uma parceria entre o Grupo de Processamento Digital de Sinais do Inatel, o Laboratório de Processamento Digital da Fala da FEEC-Unicamp e

o Instituto de Estudos da Linguagem da Unicamp [30], que é uma base multi-locutor para o português brasileiro, desenvolvida especificamente para a tarefa de reconhecimento de fala, mas que serve perfeitamente para os objetivos deste trabalho.

Foram selecionados 12 locutores, 6 masculinos e 6 femininos para compor o universo dos locutores destes testes. Para cada um deles foi selecionado material correspondente a aproximadamente 180 segundos de locução, que foram divididos da seguinte maneira:

- um subcojunto de *treinamento* de 90 segundos
- um subconjunto de *teste* de 90 segundos

Como a base de dados é formada basicamente por locuções de uma frase, para se conseguir este material foi preciso concatenar várias delas até obter o comprimento desejado.

As locuções foram gravadas em ambiente de escritório, no formato WAV, sem compressão, a uma taxa de amostragem de 22050Hz e 16 bits de resolução. Para este trabalho, os sinais foram reamostrados a uma taxa de 11025kHz.

Organização do material de treinamento e teste

Foram realizados vários testes variando-se o material disponível para treinamento e para teste. Com isto, foram feitos testes seguindo as seguintes configurações:

Material de treinamento: 30s, 60s e 90s

Material de teste: 1s, 5s e 10s

Todo o material foi parametrizado através de coeficientes mel-cepstrais de ordem 12, com janelas de 20ms, e deslocamento de 10ms. Como foram selecionados 90s de material para treinamento e 90s de material para teste, tem-se equivalentemente, 900 quadros de treinamento e 900 quadros de teste.

A geração do material de treinamento é bastante simples: para os testes com 30s, são selecionados os 30s iniciais do material (300 quadros); para os testes de 60s, os 60s iniciais (600 quadros); e para os testes de 90s, todo o material é utilizado (900 quadros).

Para a seleção do material de teste foi adotado um procedimento sugerido em [27] para aproveitar melhor o material disponível. Supondo que são desejadas locuções de T quadros, procede-se da seguinte maneira (veja Figura 4.1):

- a primeira locução é formada pelos T quadros iniciais (X_1 a X_T);

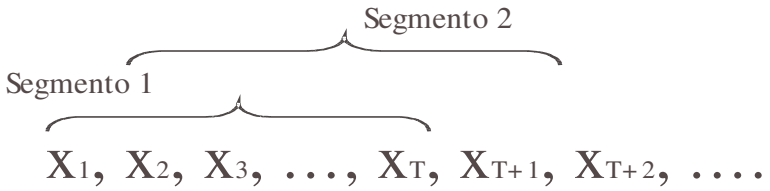


Figura 4.1: *Processo de construção dos segmentos de teste.*

- a segunda locução é formada pelos quadros X_2 a X_{T+1};

- este procedimento é repetido até chegar ao final dos 90s disponíveis.

Pode-se calcular o número de segmentos de teste, visto que, se há uma defasagem de 10 quadros entre eles, tem-se $9000/10 = 900$ segmentos no total, para um segundo de teste, 850 para 5s e 800 para 10s.

Verificação Independente de Texto

São apresentados resultados de testes com 4 locutores, 2 masculinos e 2 femininos, escolhidos aleatoriamente. Para estes testes, não foi utilizado o UBM.

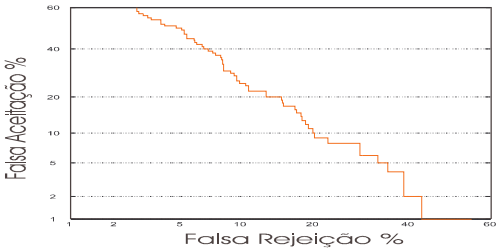
Estes testes tiveram o parâmetro Mel como parâmetro acústico e variou-se os tempos de treinamento, de segmentos de teste e número de gaussianas.

Para cada um dos 4 locutores, foi ensaiada a verificação com 1s teste e 8 e 16 gaussianas e 5s de teste e 8 gaussianas.

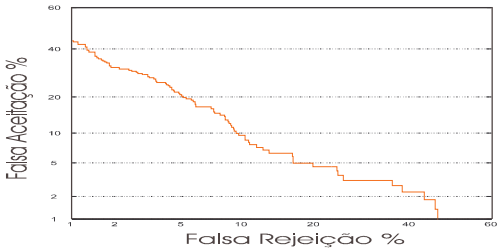
Para as demais combinações, 32 gaussianas e 5 ou 10 segundos, e 10 segundos 8 ou 16 ou 32, obteve-se 100% de acerto.

Desta forma, nas figuras 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13 e 4.14, são apresentados os resultados da verificação independente de texto no sistema GMM.

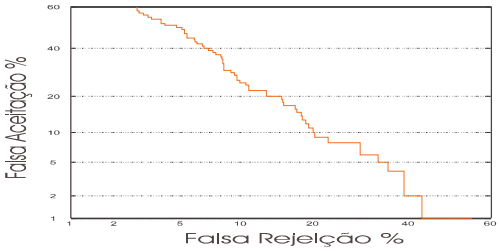
Locutor 1 (Feminino):



(a) 30 segundos de treinamento

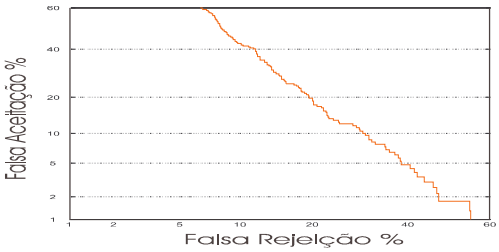


(b) 60 segundos de treinamento

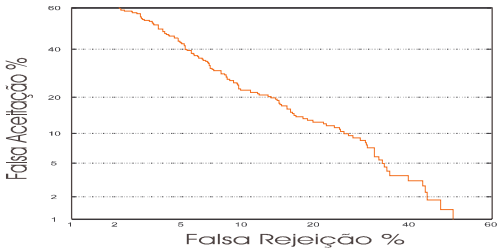


(c) 90 segundos de treinamento

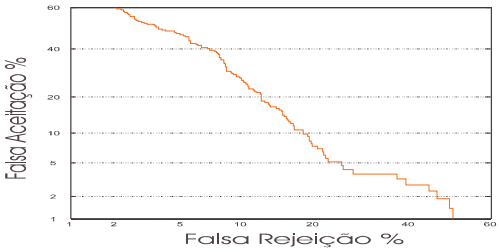
Figura 4.2: *Verificação para 1s de teste e 16 gaussianas, locutor 1.*



(a) 30 segundos de treinamento

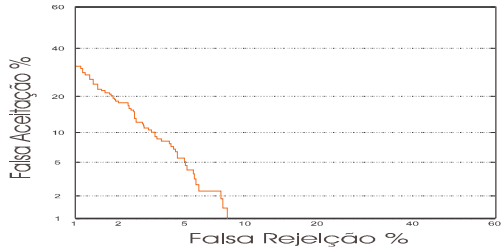


(b) 60 segundos de treinamento

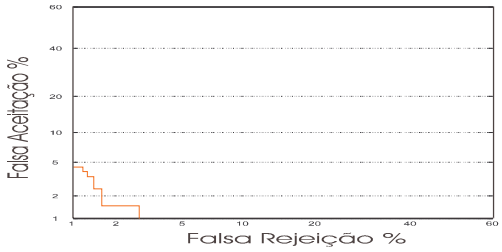


(c) 90 segundos de treinamento

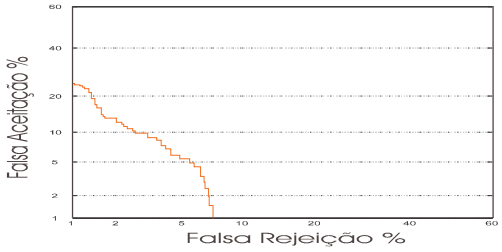
Figura 4.3: *Verificação para 1s de teste e 8 gaussianas, locutor 1.*



(a) 30 segundos de treinamento



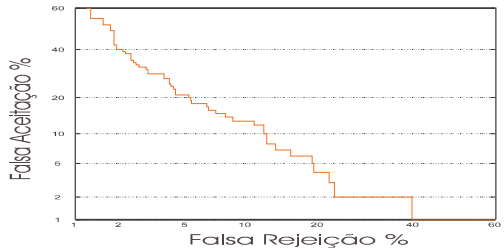
(b) 60 segundos de treinamento



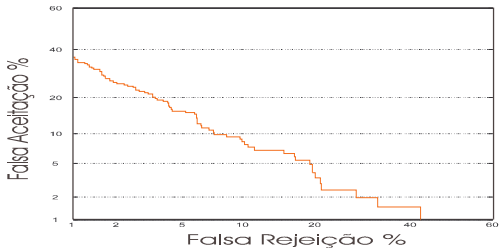
(c) 90 segundos de treinamento

Figura 4.4: *Verificação para 5s de teste e 8 gaussianas, locutor 1.*

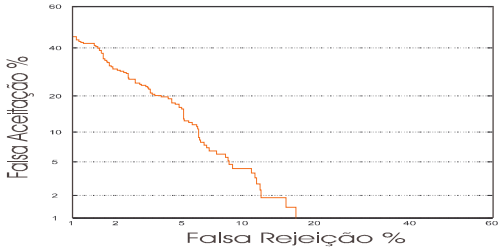
Locutor 2 (Feminino):



(a) 30 segundos de treinamento

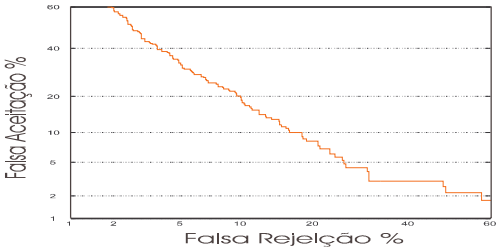


(b) 60 segundos de treinamento

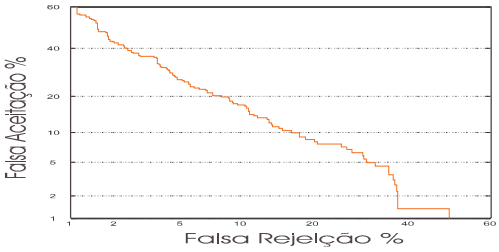


(c) 90 segundos de treinamento

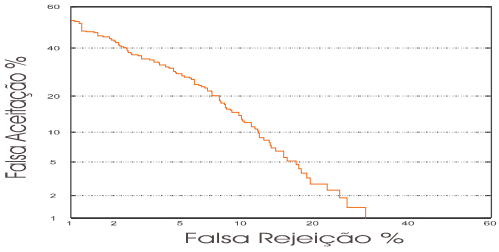
Figura 4.5: *Verificação para 1s de teste e 16 gaussianas, locutor 2.*



(a) 30 segundos de treinamento

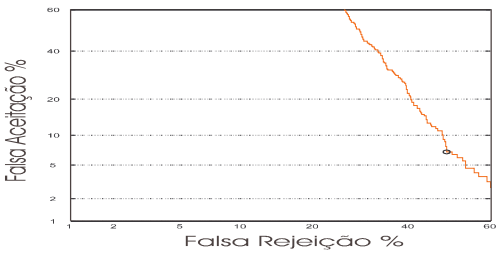


(b) 60 segundos de treinamento

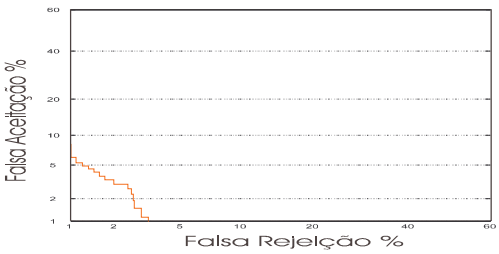


(c) 90 segundos de treinamento

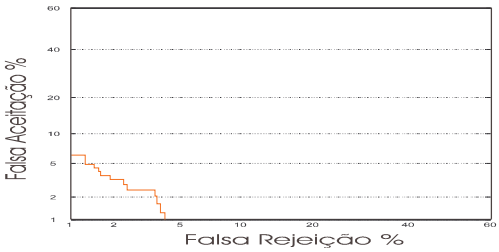
Figura 4.6: *Verificação para 1s de teste e 8 gaussianas, locutor 2.*



(a) 30 segundos de treinamento



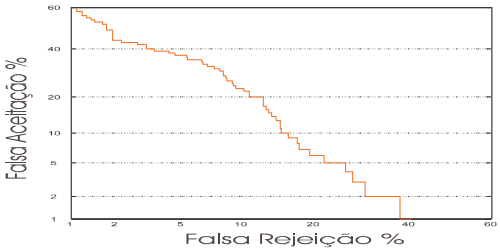
(b) 60 segundos de treinamento



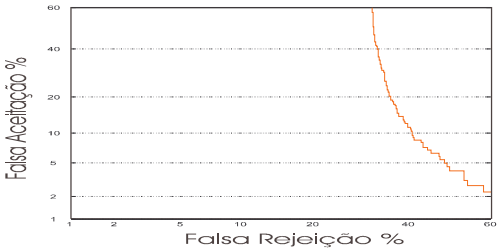
(c) 90 segundos de treinamento

Figura 4.7: *Verificação para 5s de teste e 8 gaussianas, locutor 2.*

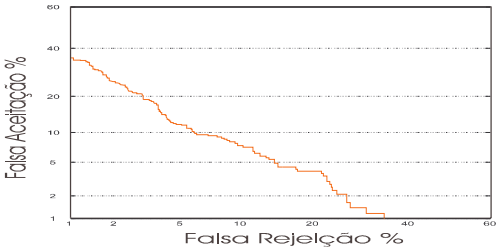
Locutor 3 (Masculino):



(a) 30 segundos de treinamento

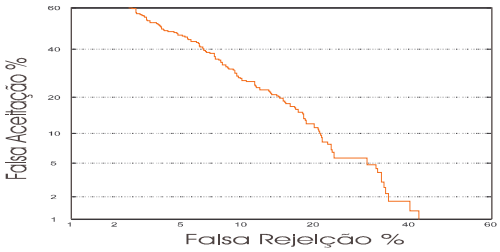


(b) 60 segundos de treinamento

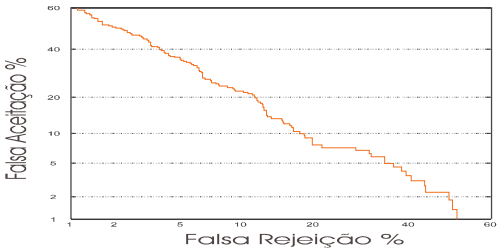


(c) 90 segundos de treinamento

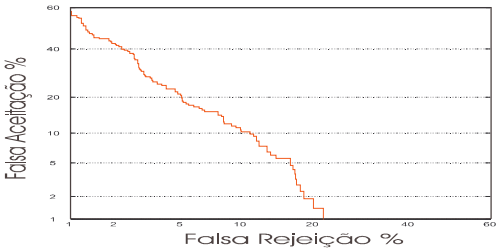
Figura 4.8: *Verificação para 1s de teste e 16 gaussianas, locutor 3.*



(a) 30 segundos de treinamento

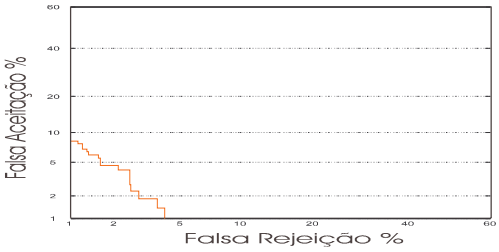


(b) 60 segundos de treinamento

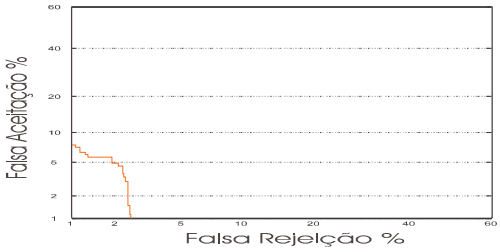


(c) 90 segundos de treinamento

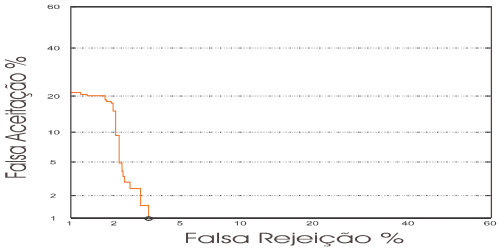
Figura 4.9: *Verificação para 1s de teste e 8 gaussianas, locutor 3.*



(a) 30 segundos de treinamento



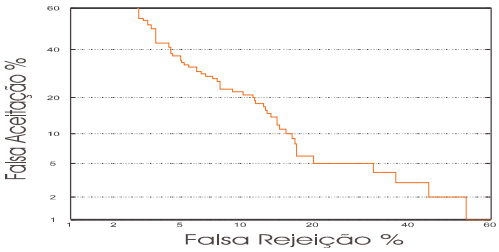
(b) 60 segundos de treinamento



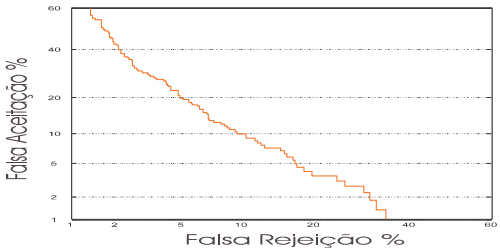
(c) 90 segundos de treinamento

Figura 4.10: *Verificação para 5s de teste e 8 gaussianas, locutor 3.*

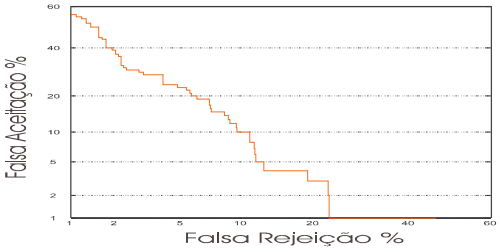
Locutor 4 (Masculino):



(a) 30 segundos de treinamento

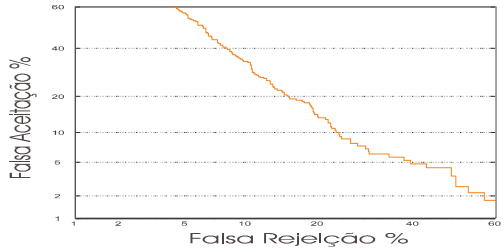


(b) 60 segundos de treinamento

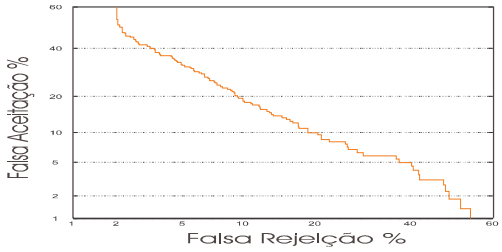


(c) 90 segundos de treinamento

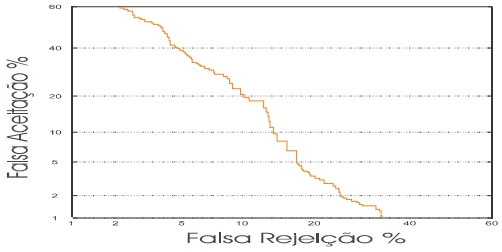
Figura 4.11: *Verificação para 1s de teste e 16 gaussianas, locutor 4.*



(a) 30 segundos de treinamento

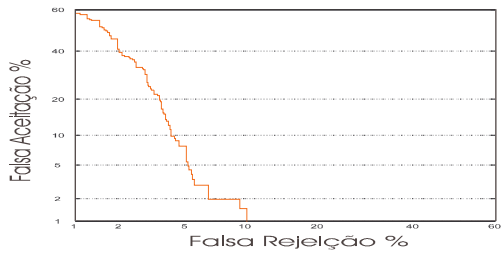


(b) 60 segundos de treinamento

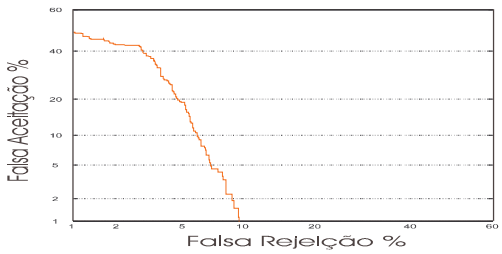


(c) 90 segundos de treinamento

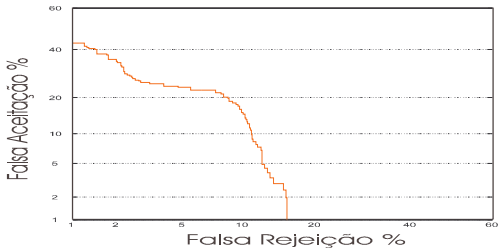
Figura 4.12: *Verificação para 1s de teste e 8 gaussianas, locutor 4.*



(a) 30 segundos de treinamento



(b) 60 segundos de treinamento



(c) 90 segundos de treinamento

Figura 4.13: *Verificação para 5s de teste e 8 gaussianas, locutor 4.*

Demais combinações, obteve-se 100% de acerto, bem como em todos as combinações de testes realizadas utilizando-se o UBM.

Figura 4.14: *Demais testes, 100% acerto.*

Conclusões

Pode-se concluir que, para a tarefa de verificação de locutor, de 5 a 10 segundos de fala são necessários para um funcionamento correto. Ainda, 32 gaussianas parece ser um número adequado para o problema em questão. Se houver disponibilidade de pelo menos 10s de material de voz, pode-se reduzir o número de gaussianas para 8, sem perda de desempenho.

Identificação Independente de Texto

Para estes testes foram treinados GMMs com 8, 16 e 32 gaussianas. Os resultados destes ensaios são mostrados na Tabela 4.4:

Tabela 4.4: *Taxa de acerto (%) dos testes de identificação de locutor (sistema básico).*

Tempo de treinamento	Gaussianas	1s	5s	10s
30s	8	39,87	72,68	88,05
-	16	51,15	88,05	93,76
-	32	60,52	95,50	98,64
60s	8	51,12	83,95	92,62
-	16	63,20	94,05	98,78
-	32	75,80	98,16	99,97
90s	8	57,19	85,53	92,79
-	16	72,08	92,50	96,67
-	32	81,17	97,92	99,99

4.3 Análise com os trechos sonoros e não sonoros da fala

O objetivo destes testes é verificar se há ganho de desempenho quando se utiliza apenas os trechos sonoros ou não sonoros da fala, ao invés de se utilizar todo o material disponível. Parte-se da hipótese de que os segmentos sonoros representam a parte nobre da fala, com grande quantidade de informação sobre a identidade do locutor, enquanto que os trechos não sonoros, por incluírem as regiões de plosivas, fricativas e partes de silêncio entre palavras, representariam um trecho do sinal com pouca informação útil.

Evidentemente, a remoção dos trechos não sonoros da fala reduz o material disponível para o treinamento/teste, mas espera-se que a (possível) melhor qualidade da informação contida nestes segmentos compense a quantidade. Os testes realizados mostraram que isto é apenas parcialmente verdadeiro. Estes testes finais foram realizados apenas na forma de identificação de locutor.

Separação dos trechos sonoros e não sonoros

A seleção dos trechos sonoros e não sonoros foi realizada com o extrator de pitch, descrito na Seção 2.5. Embora não seja muito preciso, serve ao propósito do trabalho, que é apenas o de separar os trechos sonoros dos não sonoros.

O procedimento adotado foi o seguinte: para cada quadro da locução, verificou-se o pitch associado: se fosse finito, o quadro seria considerado sonoro, caso contrário, seria considerado surdo.

Aplicando este procedimento, verificou-se que entre 50% e 70% do material disponível foi considerado sonoro. Desta forma, para ter-se a mesma quantidade de material útil, seria necessário que o locutor pronunciasse uma locução proporcionalmente maior. Isto pode ou não ser viável em aplicações práticas, e desta forma optou-se, neste trabalho, por testar as duas hipóteses. Desta forma, os testes foram realizados utilizando-se os seguintes conjuntos de dados:

- Todo material: todo o sinal de fala.
- Trechos sonoros: o locutor pronunciou locuções mais longas, de modo a compensar a perda dos trechos não sonoros.
- Trechos não-sonoros: apenas os trechos não sonoros do sinal (aqui também o locutor pronunciou locuções mais longas, de modo a compensar a perda dos trechos não sonoros).

- Trechos sonoros filtrados: para o caso do teste de 5s por exemplo, o material disponível é de aproximadamente 3s (pois os trechos não sonoros foram descartados).

Na Tabela 4.5 podem ser vistos os resultados dos testes realizados com 30 s de material de treinamento e GMMs com 8 gaussianas.

Tabela 4.5: *Resultados dos testes de identificação de locutor, taxa de acerto em %. Testes com seleção dos trechos sonoros e não sonoros. 30 s de material de treinamento. GMMs com 8 gaussianas. Na primeira linha é apresentado o desempenho do sistema básico, treinado e testado com todo o material, para comparação.*

Material para treinamento	Material para teste	1s	5s	10s
Todo material	Todo material	39,87	72,68	88,05
Trechos não sonoros	Trechos não sonoros	25,57	55,86	71,04
Trechos não sonoros	Todo material	38,85	48,44	35,98
Trechos sonoros	Todo material	55,61	79,02	89,93
Trechos sonoros filtrados	Todo material	52,41	69,57	73,97
Todo material	Trechos sonoros filtrados	64,59	86,72	92,98
Trechos sonoros	Trechos sonoros filtrados	49,45	69,54	77,02
Trechos sonoros filtrados	Trechos sonoros filtrados	49,62	68,23	73,97
Todo material	Trechos sonoros	72,65	94,21	99,32
Trechos sonoros	Trechos sonoros	53,83	76,56	85,63
Trechos sonoros filtrados	Trechos sonoros	51,75	68,93	73,46

Conclusões

Observando os resultados acima, podem ser tiradas as seguintes conclusões:

- a utilização apenas dos trechos não sonoros degrada bastante o desempenho do sistema;
- os trechos sonoros do sinal de fala provêm realmente informação nobre sobre a identidade do locutor
- como era de se esperar, se pudermos fazer com que o locutor pronuncie locuções mais longas, de modo que seja possível compensar os trechos descartados (não sonoros), o desempenho do sistema melhora sensivelmente.
- o melhor resultado foi obtido com o sistema treinado com todo o material, usando-se para os testes apenas os trechos sonoros. Este resultado é inesperado, e mais testes devem ser realizados para verificar este ponto em detalhes.

Capítulo 5

Conclusões e trabalhos futuros

Neste trabalho foram implementadas e testadas as principais técnicas utilizadas para o reconhecimento automático de pessoas pela voz. Os estudos basearam-se em dois tópicos principais: a) parâmetros acústicos e b) classificadores.

Em relação aos parâmetros acústicos, foram estudados os mel-cepstrais, LPC, perfil energia e pitch, bem como as suas derivadas primeira e segunda. Os testes realizados mostraram que os mel-cepstrais alcançaram o melhor desempenho, seguidos dos LPC, resultado coerente com a literatura e com a análise feita no Capítulo 2.

Os classificadores avaliados neste trabalho foram o quantizador vetorial, os HMMs, as redes neurais e o modelo de mistura de Gaussianas (GMM). De forma geral, pode-se dizer que os HMMs, as redes neurais e GMM tiveram um bom desempenho para o reconhecimento de locutor independente de texto, embora os HMMs tenham apresentado um desempenho levemente superior. Este resultado, aparentemente diferente do reportado pela literatura, que aponta os GMMs como o estado da arte, pode ser justificado pelo tamanho reduzido da base de dados utilizada nesta comparação.

Como contribuição original deste trabalho destaca-se o estudo da contribuição dos trechos sonoros e não sonoros da fala no desempenho destes sistemas. Verificou-se que a utilização apenas dos trechos não sonoros degrada bastante o desempenho do sistema. Já a utilização das partes sonoras promoveu uma substancial melhora na taxa de acertos do sistema, o que indica que realmente estas regiões possuem muita informação sobre a identidade do locutor. Entretanto, o melhor desempenho foi obtido quando o sistema foi treinado com todo o sinal de voz (partes sonoras e não sonoras), e testado apenas com os trechos sonoros. este é um resultado inesperado que merece um estudo mais aprofundado.

O reconhecimento de locutor é uma área específica que ainda tem muito a desenvolver. O desenvolvimento de parâmetros com menor variabilidade intralo-

cutor e maior variabilidade interlocutor poderá levar à criação de sistemas mais confiáveis e robustos.

Outra idéia seria utilizar informações de alto nível, tais como o sotaque, ritmo da fala, relação dos tempos de duração relativa entre vogais e consoantes, entre outros.

Para os sitemas dependentes de texto, a prosódia poderia ser utilizada como uma informação adicional: o locutor teria que pronunciar a frase correta, com a mesma prosódia utilizada para treinar o sistema. Isto aumentaria o grau de segurança do sistema.

Quanto aos sistemas de classificação, poderiam ser acrescentadas as máquinas de vetor suporte (Suport Vector Machines), que se preocupam em minimizar não apenas o risco empírico (erro no material de treinamento), mas também o risco estrutural (capacidade de generalização).

Os sistemas biométricos modernos fazem uso de mais de uma fonte de informação, e desta forma, a associação de reconhecedores de impressões digitais, face, íris e outros poderiam levar a um melhor desempenho.

Referências Bibliográficas

[1] Martins, J. A. Avaliação de Diferentes Técnicas Para Reconhecimento de Fala, Tese de Doutorado, Universidade de Campinas, 1997.

[2] Ynoguti, C. A., Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov, Tese de Doutorado, Universidade Estadual de Campinas,1999.

[3] Rabiner, L. and Juang, B. H., Fundamentals of Speech Recognition, Prentice Hall Signal Processing Series, 1993.

[4] Douglas A. Reynolds and Richard C. Rose, Robust Text-Independent Speaker identification Using Gaussian Mixture Speaker Models, in Proc. IEE, vol 3, No.1, pp. 72-83, 1995.

[5] Justinian Rosca and Andri Kofmehl, Cepstrum-Like ICA Representation for Text Independent speaker Recognition, Siemens Corporate Research, in ICA 2003, Japão, abril 2003.

[6] Andre G. Adami, Radu Mihaescu, Douglas A. Reynolds, Modeling Prosodic Dynamics for Speaker recognition, in ICASSP 2003, Japão, 2003.

[7] Hassan Ezzaidi, Jean Rouat and Douglas O'ShaughnessyErmetis, Combinning pitch and MFCC for Speaker Recognition Systems, in A speaker Odyssey, the Speaker Recognition Workshop, an ISCA Tutorial and Research Workshop (ITRW) on Speaker Recognition, junho, 18-22 2001. Paper nb: 1036

[8] M. Kemal Sonmez, Larry Heck, Mitchel Weintraub, Elizabeth Shriberg, A Lognormal tied Mixture Model of Pitch for Prosody-Based Speaker Recognition, 7th International Conference on Spoken Language Processing September 16-20, USA, 2002.

[9] Joseph P. Campbell, Jr., Speaker recognition: A Tutorial, proceedings of the IEE, vol.85, No.9, pp. 1437-1462, setembro 1997 .

[10] L. R. Rabiner/ R. W. Schafer - Digital Processing of Speech Signals, Prentice Hall, 1978 .

[11] DAVIS, S. & MELMERTSTEIN, P. Comparison of parametric representations for monossylabic word recognition in continously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, ASP-28(4):357-366. August, 1980.

[12] Herbert Gish e Michel Schmidt, Text-Independent Speaker Identification: A Tutorial, in IEE Signal Processing Magazine, 1053-588/94/S4.00, outubro 1994.

[13] Tomoko Matsui e Sadaoki Furui, Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion ans Discrete/Continuous HMM's, in Proc. IEE, vol.2, No.3, pp. 456-458, julho 1994.

[14] Minh Do e Michael Wagner, Speaker Recognition With Small Training Requirements Using a Combination of VQ and DHMM, Proc. of Workshop on Speaker Recognition and Its Commercial and Forensic Applications, pp. 169-172, França, abril 1998.

[15] Adriano Petry, Adriano Zanuz e Dante Augusto Couto, Utilização de Técnicas de Processamento Digital de Sinais para a Identificação Automática de Pessoas pela voz, in SSI'99 - Simpósio de Segurança em Informática, São José do Rio Preto, 1999.

[16] Northide Kitaoka, Daisuke Yamanda, Seiichi Nakagawa, Speaker Independent Speech Recognition Using Features Based on Glottal Sound Source, 7th International Conference on Spoken Language Processing September 16-20, USA, 2002.

[17] Charles B. de Lima, Abraham Alcaim e José Apolinário Jr., GMM Versus AR-Vector Models for Text Independent Speaker Verification, IME e CETUC-RJ, In: ITS'02, Natal - Brasil, setembro 2002 .

[18] Steven B. Davis, Paul Mermelstein, Comparision of Parametric Representations for Monosylabic Word Recognition in Continuosly Spoken Snetences, in proc. IEE, vol. ASSp28, No. 4, pp 357-366, agosto 1980.

[19] Frank K. Soog e Aaron E. Rosenberg, On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition, in proc. IEE, vol.36, No. 6, pp. 871-879, junho 1988.

[20] Picone, J.W., Signal Modeling Techiques in Speech Recognition, in proc. IEE, vol.81, No.81, pp.1215-1247, swetembro 1993.

[21] HAYKIN, Simon., Neural Networks: a comprehensive foundation. Prentice Hall, 1999.

[22] Braga, Antônio de Pádua, Carvalho, André Carlos P. L. F., Ludemir, Teresa Bernarda, Redes Neurais Artificiais: Teoria e Aplicações, LTC Livros Técnicos Científicos Editora, 1999.

[23] Todor Ganchev, Anastasios Tsopanoglou, Nikos Fakotakis, George Kokkinakis Probabilistic Neural Networks combined with GMMS for Speaker recognition over telephone channels, 14th International Conference on Digital Signal Processing (DSP2002), Volume II, pp.1081-1084 Julho 2002, Santorini, Greece

[24] Sawut Kasuriya, Chai Wutiwiwatchai, Varin Achariyakulporn, Chularat Tanprasert, Comparative Study of Continuous Hidden Markov Models (CHMM)and Artificial Neural Network (ANN) on Speaker Identification System, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 9, No. 6 (2001) 673-683.

[25] ARAÚJO, Antônio Marcos de Lima. Jogos Computacionais Fonoarticulatórios para Crianças com Deficiência Auditiva. Tese de Doutorado. UNICAMP. Campinas. 2000.

[26] Y. Linde, A. Buzo and R.M. Gray, "An algorithm for vector quantizer design," IEEE Transactions on Communications, vol. 28, pp. 84-95, January 1980.

[27] Douglas A. Reynolds, Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEE Transactions on Speech and Audio Processing, vol. 3, no 1, janeiro 1995.

[28] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The Det Curve in Assessment of Detection Task Performance, EuroSpeech 1997, Proceedings Volume 4, Pages 1895-1898, 1997.

[29] http://www.data-compression.com/vq.shtml. Acessado em 21/11/2004.

[30] YNOGUTI, C. A. ; BARBOSA, P. A. ; VIOLARO, F. A Large Speech Database for Brazilian Portuguese Spoken Language Research. In: VI Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, 2003, Faro - Algarve - Portugal. Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken. Berlin : Springer-Verlag, 2003. p. 193-196.