INSTITUTO MILITAR DE ENGENHARIA

CHARLES BORGES DE LIMA

SISTEMAS DE VERIFICAÇÃO DE LOCUTOR INDEPENDENTE DO TEXTO BASEADOS EM GMM E AR-VETORIAL UTILIZANDO PCA

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

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

Rio de Janeiro
2001

Ao Mestre.

AGRADECIMENTOS

Ao Exército Brasileiro, em especial ao Departamento de Engenharia Elétrica do Instituto Militar de Engenharia, pela oportunidade de crescimento pessoal e profissional durante o curso de Mestrado.

Ao professor Abraham Alcaim pela orientação, sugestões, apoio e dedicação transmitidos durante todo o trabalho.

Ao professor José Antônio Apolinário Jr. pela co-orientação, incentivo e ajuda.

Ao professor Roberto Miscow Filho pelos ensinamentos transmitidos e colaboração nas correções deste trabalho.

À professora Rosângela Fernandes Coelho pelas valiosas sugestões.

Ao amigo e colega Dirceu Gonzaga da Silva, por suas inumeras sugestões e ajudas no desenvolvimento de programas.

Aos amigos do IME em especial a Senhora Maria de Lourdes Santarem Rodrigues.

À CAPES pelo apoio financeiro que permitiu a elaboração desta dissertação.

A minha família e em especial a minha noiva Simone pelo apoio e compreensão.

A todos que direta ou indiretamente colaboraram na concretização deste trabalho.

"Sonhar mais um sonho impossível, lutar quando é fácil ceder, vencer o inimigo invencível, negar quando a regra é vender."

MIGUEL DE CERVANTES

SUMÁRIO

LISTA DE ILUSTRAÇÕES 10

LISTA DE TABELAS 12

LISTA DE SIGLAS 15

1 INTRODUÇÃO 19
1.1 Objetivo da Dissertação 20
1.2 Organização da Dissertação 20

2 RECONHECIMENTO AUTOMÁTICO DE LOCUTOR 22
2.1 Introdução 22
2.2 Conceitos Básicos 22
2.3 Pré-Processamento 26
2.4 Extração de Características 27
2.5 Sistemas de Classificação 32
2.6 Estado da Arte 36
2.7 Resumo e Conclusão 41

3 CARACTERÍSTICAS E TRANSFORMAÇÃO UTILIZADA 42
3.1 Introdução 42
3.2 Eliminação do Silêncio 42
3.3 Os Coeficientes Mel-cepestrais 44
3.4 A Análise de Componentes Principais 47
3.5 Aplicação do PCA aos Coeficientes Mel-cepestrais 49
3.6 Resumo e Conclusão 51

4 CLASSIFICADORES 52
4.1 Introdução 52
4.2 Modelo de Mistura de Gaussianas 52
4.3 O GMM no Reconhecimento de Locutor 53
4.3.1 Interpretações do Modelo no Reconhecimento de Locutor 54
4.3.2 Estimação dos Parâmetros de Máxima Verossimilhança 56
4.3.3 Sistema de Identificação com o GMM 57
4.3.4 Sistema de Verificação com o GMM 58

4.3.5 Background 60
4.4 Modelo Autorregressivo Vetorial 64
4.4.1 Relação entre o LPC e o AR-Vetorial 65
4.4.2 Medidas Usadas no AR-Vetorial para o Reconhecimento de Locutor 68
4.4.3 Sistema de Identificação com o AR-Vetorial 70
4.4.4 Sistema de Verificação com o AR-Vetorial 70
4.5 Resumo e Conclusão 71

5 AVALIAÇÃO DE DESEMPENHO E ANÁLISE COMPARATIVA 72
5.1 Introdução 72
5.2 Descrição das Simulações 72
5.3 Medida de Erro 74
5.4 Avaliação do GMM 75
5.4.1 Definição do Limiar de Decisão 75
5.4.2 Definição do Número de Coeficentes Mel-cepestrais 77
5.4.3 Avaliação do Número de Gaussianas, Tempo de Treinamento e Teste 78
5.4.4 Análise do Background 84
5.5 Avaliação do AR-Vetorial 86
5.5.1 Definição do Limiar de Decisão 87
5.5.2 Definição da Distância Utilizada no Modelo 88
5.5.3 Avaliação da Ordem do Modelo, Tempo de Treinamento e Teste 90
5.5.4 Avaliação entre MCC12 e MCC15 91
5.5.5 Desempenho do PCA no AR-Vetorial 92
5.6 Comparação entre os Resultados do GMM e AR-Vetorial 93
5.7 Comparação entre os Tempos de Processamento 95
5.8 A Curva DET 96
5.9 Resumo e Conclusão 98

6 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS 101
6.1 Conclusões 101
6.2 Sugestões para Trabalhos Futuros 103
6.3 Comentários Finais 104

7 BIBLIOGRAFIA 105

8 APÊNDICES 110
8.1 APÊNDICE 1: ESTIMAÇÃO DOS PARÂMETROS DO GMM 111
8.1.1 Estimação da Máxima Verossimilhança 111
8.1.2 Função Auxiliar 112
8.1.3 Equações para a Estimação dos Parâmetros 114
8.1.4 O algoritmo EM 117
8.1.5 Uso do Algoritmo EM 117
8.2 APÊNDICE 2: ALGORITMOS PARA OBTENÇÃO DO LPC E DO AR-VETORIAL 119
8.3 APÊNDICE 3: RESULTADOS COMPLEMENTARES DO AR-VETORIAL 121

LISTA DE ILUSTRAÇÕES

FIG.2.1 Algumas áreas do processamento de voz, com destaque para o reconhecimento de locutor. As palavras em negrito indicam a abrangência desta dissertação. 23
FIG.2.2 Sistema genérico para reconhecimento de locutor. 25
FIG.2.3 Resposta em freqüência de um filtro de pré-ênfase com $a_{pre} = -0,95$, para um sinal amostrado a uma taxa de $8KHz$. 27
FIG.2.4 Interpretação do janelamento no domínio do tempo. 29
FIG.2.5 Alinhamento temporal não linear de um sinal de teste $T(n)$ em relação a um modelo $R(m)$. 34
FIG.2.6 Estrutura HMM esquerda-direita de N estados. 35
FIG.2.7 Taxa de identificação de locutor em função do número de estados e misturas em HMMs ergódicos. Um único estado pode ser compreendido como um GMM (FURUI, 1996). 38

FIG.3.1 Ilustração do funcionamento do algoritmo proposto para eliminação do silêncio de gravações de voz. 43
FIG.3.2 Escala mel versus escala de freqüência normal. 45
FIG.3.3 Magnitude do espectro dos filtros de banda crítica utilizados na produção dos coeficientes mel-cepestrais (DAVIS, 1980). 46
FIG.3.4 Diagrama em blocos para a extração dos coeficientes mel-cepestrais (REYNOLDS, 1995). 46
FIG.3.5 Transformação PCA em uma variável bidimensional. 48
FIG.3.6 Transformação PCA sobre um vetor de características mel-cepestrais. . . 49
FIG.3.7 Gráfico tridimensional da matriz covariância de um conjunto de vetores com 15 MCCs. 50
FIG.3.8 Matriz covariância para 15 MCCPCAs. 50

FIG.4.1 M densidades de probabilidade formando um GMM (REYNOLDS, 1992). 54
FIG.4.2 (a) Histograma normalizado de um vetor de característica Log-energia, para uma locução com 30 segundos de duração de um locutor masculino, (b) modelagem da distribuição dos dados feita por um GMM composto por sete gaussianas, as quais estão abaixo da linha cheia. 55
FIG.4.3 Sistema de identificação de locutor, com S locutores. 58
FIG.4.4 Sistema para verificação de locutor. 59

FIG.4.5 Métodos mais comuns para criar o UBM. (a) Dados das subpopulações são agrupados antes do treinamento. (b) Modelos individuais de subpopulações são treinados e, então, combinados para criar um UBM final. 64
FIG.4.6 Modelo linear do trato vocal para a produção da voz. 66
FIG.4.7 Sistema de identificação de locutor com o AR-Vetorial. 70
FIG.4.8 Sistema de verificação de locutor com AR-Vetorial. 71

FIG.5.1 Resultado do GMM com o uso do *background* com 10 locutores em (a) e sem seu uso em (b). 76
FIG.5.2 Avaliação do número de MCCs e MCCPCAs no GMM com 32 gaussianas e 60s treinamento. 78
FIG.5.3 Desempenho do GMM (EER) para 30s de teste, em relação ao número de gaussianas e tempo de treino. 83
FIG.5.4 Desempenho do GMM (EER) para 10s de teste, em relação ao número de gaussianas e tempo de treino. 83
FIG.5.5 Desempenho do GMM (EER) para 3s de teste, em relação ao número de gaussianas e tempo de treino. 84
FIG.5.6 Limiar de decisão para uso no AR-Vetorial. 87
FIG.5.7 Resultados do AR-Vetorial usando o UBM. 88
FIG.5.8 Desempenho nas Distâncias utilizadas no AR-Vetorial, para $p = 2$ com 60s de treinamento. 89
FIG.5.9 Curva DET para o GMM com 32 gaussianas e AR-Vetorial de ordem 2 (distância simétrica) para um tempo de treinamento e teste de 10s. 97

LISTA DE TABELAS

TAB.2.1 Cronologia selecionada no reconhecimento de locutor (CAMPBEL, 1997). 39

TAB.5.1 Desempenho do GMM com a variação do número de coeficientes MCC e MCCPCA. 77
TAB.5.2 Desempenho do GMM com MCC e MCCPCA, para 32 gaussianas com 60s de treinamento. 79
TAB.5.3 Desempenho do GMM com MCC e MCCPCA, para 16 gaussianas com 60s de treinamento. 79
TAB.5.4 Desempenho do GMM com MCC e MCCPCA, para 8 gaussianas com 60s de treinamento. 80
TAB.5.5 Desempenho do GMM com MCC e MCCPCA, para 32 gaussianas com 30s de treinamento. 80
TAB.5.6 Desempenho do GMM com MCC e MCCPCA, para 16 gaussianas com 30s de treinamento. 80
TAB.5.7 Desempenho do GMM com MCC e MCCPCA, para 8 gaussianas com 30s de treinamento. 81
TAB.5.8 Desempenho do GMM com MCC e MCCPCA, para 32 gaussianas com 10s de treinamento. 81
TAB.5.9 Desempenho do GMM com MCC e MCCPCA, para 10s de treinamento, 16 gaussianas. 82
TAB.5.10 Desempenho do GMM com MCC e MCCPCA, para 10s de treinamento, 8 gaussianas. 82
TAB.5.11 Desempenho do GMM com 10 locutores de *background*. 85
TAB.5.12 Desempenho do GMM com 16 locutores de *background*. 85
TAB.5.13 Desempenho das distâncias usadas no AR-Vetorial para $p = 2$, 60s de treinamento. 89
TAB.5.14 Desempenho do AR-Vetorial para variações na ordem do modelo, utilizando a distância simétrica, com 60s de treinamento. 90
TAB.5.15 Desempenho do AR-Vetorial para variações na ordem do modelo, utilizando a distância simétrica, com 30s de treinamento. 90
TAB.5.16 Desempenho do AR-Vetorial para variações na ordem do modelo, utilizando a distância simétrica, com 10s de treinamento. 91
TAB.5.17 Desempenho do AR-Vetorial para MCC12 e MCC15, para ordem 2 e dis-

tância Simétrica. 92
TAB.5.18 Desempenho do AR-Vetorial para MCC e MCCPCA, para ordem 2 com distância simétrica, com 60s de treinamento. 92
TAB.5.19 Desempenho do AR-Vetorial para MCC e MCCPCA, para ordem 2 com distância simétrica, com 30s de treinamento. 93
TAB.5.20 Desempenho do AR-Vetorial para MCC e MCCPCA, para ordem 2 com distância simétrica, com 10s de treinamento. 93
TAB.5.21 Desempenho do GMM versus AR-Vetorial, para 60s de treinamento. . . . 94
TAB.5.22 Desempenho do GMM versus AR-Vetorial, para 30s de treinamento. . . . 94
TAB.5.23 Desempenho do GMM versus AR-Vetorial, para 10s de treinamento. . . . 95
TAB.5.24 Tempos de processamento para treinamento do GMM e AR-Vetorial. . . 96
TAB.5.25 Tempos de processamento para testes do GMM e AR-Vetorial. 96

TAB.8.1 Desempenho do AR-Vetorial para as diferentes distâncias usadas, $p = 2$ e 30s de treinamento. 121
TAB.8.2 Desempenho do AR-Vetorial para as diferentes distâncias usadas, $p = 2$ e 10s de treinamento. 121
TAB.8.3 Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 1 e 60s de treinamento). 122
TAB.8.4 Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 1 e 30s de treinamento. 122
TAB.8.5 Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 1 e 10s de treinamento. 122
TAB.8.6 Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 7 e 60s de treinamento. 123
TAB.8.7 Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 7 e 30s de treinamento. 123
TAB.8.8 Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 7 e 10s de treinamento. 123
TAB.8.9 Desempenho do AR-Vetorial com MCCPCA, para as ordens 1, 2 e 3, usando as distâncias 1, 5 e 7, com 60s de treinamento. 124
TAB.8.10 Desempenho do AR-Vetorial com MCCPCA, para as ordens 1, 2 e 3, usando as distâncias 1, 5 e 7, com 30s de treinamento. 124
TAB.8.11 Desempenho do AR-Vetorial com MCCPCA, para as ordens 1, 2 e 3, usando as distâncias 1, 5 e 7, com 10s de treinamento. 125

TAB.8.12 Desempenho do AR-Vetorial com MCC12, para as ordens 1, 2 e 3, com distâncias 1, 5 e 7 e 60s de treinamento. 125
TAB.8.13 Desempenho do AR-Vetorial com MCC12, para as ordens 1, 2 e 3, com distâncias 1, 5 e 7 e 30s de treinamento. 126
TAB.8.14 Desempenho do AR-Vetorial com MCC12, para as ordens 1, 2 e 3, com distâncias 1, 5 e 7 e 10s de treinamento. 126

LISTA DE SIGLAS

A/D	Conversor Analógico-Digital
AR-Vetorial	Modelo Autorregressivo Vetorial
DCT	*Discrete Cosine Transform*
DET	*Detection Tradeoff Error*
DFT	*Discrete Fourier Transform*
DTW	*Dynamic Time Warping*
EER	*Equal Error Rate*
EM	*Expectation-Maximization*
FA	Falsa Aceitação
FBI	*Federal Bureau of Investigations*
fdp	função densidade de probabilidade
FIR	*Finite Impulse Response*
FR	Falsa Rejeição
GMM	*Gaussian Mixture Model*
HMM	*Hidden Markov Model*
ICASSP	*International Conference on Acoustics, Speech, and Signal Processing*
ICSLP	*International Conference on Spoken Language Processing*
IDFT	*Inverse Discrete Fourier Transform*
KLT	*Karhuen-Loéve Transform*
LAR	*Log Area-Ratio*
LBG	*Linde Buzo Gray*
LDA	*Linear Discriminant Analysis*
LPC	*Linear Prediction Coefficients*
MCC	*Mel-Cepstrum Coefficients*
MCCPCA	*Mel-Cepstrum Coefficients with Principal Components Analysis*
ML	*Maximum Likelihood*
MSC	*Maximally Spread Closes*
MSF	*Maximally Spread Far*
NIST	*National Institute of Standards and Technology*
NLDA	*Not Linear Discriminant Analysis*
NPCA	Not Linear Principal Component Analysis
PCA	*Principal Component Analysis*
PLP	*Perceptual Linear Predictive*

RAL	Reconhecimento Automático de Locutor
RLA2C	*La Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques*
RN	Rede Neural
UBM	*Universal Background Model*
VQ	*Vector Quantization*

RESUMO

Esta dissertação apresenta uma avaliação detalhada do desempenho de dois sistemas de classificação usados na tarefa de verificação de locutor independente do texto, o GMM e o AR-Vetorial. Além disto, foi estudado o efeito do uso de análise de componentes principais (*Principal Component Analysis* ou PCA) sobre as características de voz. Os sistemas de classificação foram investigados em termos das variações de seus parâmetros (o número de Gaussianas no caso do GMM e a ordem do modelo e a medida de distância no caso do AR-Vetorial). As vantagens individuais e comparativas dos sistemas foram apontadas para diferentes durações das elocuções usadas para treinamento e teste. Os resultados obtidos mostraram a eficiência desses sistemas de classificação na verificação de locutor independente do texto.

ABSTRACT

This dissertation presents a detailed performance evaluation of two classification systems used in the task of text independent speaker verification, the GMM and the AR-vector. Moreover, the effect of using Principal Components Analysis (PCA) on the speech features was studied. The classification systems were investigated in terms of the variation of their parameters (the number of Gaussians for the case of GMM and the model order and the distance measure for the case of AR-vector). The individual and comparative system advantages were pointed out for different speech durations used for training and testing. The results obtained have shown the efficiency of these classification systems in text independent speaker verification.

1 INTRODUÇÃO

O fato de que as vozes das pessoas são diferentes é uma importante propriedade da fala. É o que possibilita o reconhecimento, através do som, de um ser humano dentre outros. A habilidade de reconhecer uma pessoa somente por sua voz é denominada reconhecimento de locutor. O reconhecimento de locutor é uma experiência comum e conhecida há muito tempo. Com o advento dos computadores, os pesquisadores começaram a utilizar a voz para a identificação do homem pela máquina, o que pode ser chamado de reconhecimento automático de locutor (RAL) (ATAL, 1976). Esta tarefa é dividida em identificação e verificação de locutor. A identificação é feita em um conjunto conhecido de locutores, enquanto a verificação avalia se um dado locutor é quem diz ser.

Existem outras características biométricas que podem ser utilizadas para a identificação de uma pessoa, tais como, impressão digital, geometria da mão, retina e atividades próprias de cada ser humano, como a forma de digitação. Em todos os casos, características são comparadas com outras armazenadas previamente e a identificação é realizada baseada em critérios de decisão. Dentre as várias características biométricas para identificação de uma pessoa, a voz é a que apresenta mais vantagens para aplicações práticas por ser um meio natural de comunicação, o que evita contatos com o sistema de identificação e garante a sua ampla aceitação pelos usuários do sistema.

Com o avanço do processamento digital de sinais, a utilização da fala tornou possível o projeto de sistemas de reconhecimento de locutor de baixo custo, rápidos e com desempenho razoável. Tais sistemas podem ser facilmente integrados às redes de telefonia e aos computadores equipados com microfones, justificando ainda mais as pesquisas nesta área.

As aplicações do reconhecimento de locutor têm, portanto, aumentado significativamente nos últimos anos (CAMPBELL, 1997). Dentre tais aplicações, podemos destacar:

- Controle de acesso: a dispositivos, redes de trabalho e dados.
- Autenticação de transações comerciais como ferramenta para prevenção de fraudes: compras por telefone com cartão de crédito, transações na internet e operações bancárias.
- Segurança pública: monitoração em penitenciárias, aplicações forenses[1].
- Auxílio a deficientes físicos.

[1]Relativo ao foro judicial (FERREIRA, 2000).

- Uso militar: informações que requeiram identificação de locutor.

A utilização da voz para segurança pode ser acompanhada por outros dispositivos que possam validar o locutor verdadeiro, tais como, cartões magnéticos e senhas. No futuro, aplicações incluirão interação entre homem e máquina, como ambientes *simpáticos* (salas e veículos que respondam apropriadamente aos comandos de um determinado ser humano); outras aplicações por voz, como correio eletrônico tomarão volume (NIST, 2000).

Na sociedade atual a identificação de cada indivíduo torna-se cada vez mais imprescindível. As pesquisas com voz têm sido intensificadas na comunidade científica, visando suprir esta crescente necessidade. Muitas lacunas estão sendo preenchidas mas ainda existem muitos problemas a serem solucionados e toda pesquisa e avaliação nesta área é importante. Espera-se que em poucos anos existam sistemas de RAL de alta confiabilidade e eficiência.

1.1 OBJETIVO DA DISSERTAÇÃO

O objetivo principal desta dissertação é analisar dois sistemas de classificação para verificação de locutor independente do texto, o GMM (*Gaussian Mixture Model* - modelo de mistura de gaussianas)e o AR-Vetorial (modelo autorregressivo vetorial), fazendo uso da transformada PCA (*Principal Components Analysis* - análise de componentes principais) nas características de voz utilizadas. Serão determinadas as principais vantagens e desvantagens de tais sistemas de classificação em função de seus respectivos parâmetros.

A análise visa obter um sistema que apresente maior taxa de acertos possível e conseqüentemente maior confiabilidade para o reconhecimento de locutor independente do texto. A utilização de um classificador puramente estatístico (GMM), e um que analisa a evolução espectral das características de voz (AR-Vetorial), mostrará a diferença de resultados para os diferentes sistemas de classificação. O uso de PCA, revelará a importância da utilização de características de voz com maior poder discriminativo, para o reconhecimento de locutor.

1.2 ORGANIZAÇÃO DA DISSERTAÇÃO

Serão apresentados no capítulo 2 os conceitos básicos e outros necessários à compreensão dos sistemas de reconhecimento automático de locutor, até o estado da arte. No capítulo 3 serão vistos a característica de voz utilizada e a análise de componentes principais com seu respectivo relacionamento no reconhecimento de locutor. O capítulo 4 possui

a explanação detalhada dos classificadores, o GMM e o AR-Vetorial e sua utilização no reconhecimento de locutor. Os resultados e simulações em conjunto com a análise comparativa dos sistemas de classificação, são vistos no capítulo 5 e, finalmente, as conclusões e sugestões para trabalhos futuros são apresentados no capítulo 6.

2 RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

2.1 INTRODUÇÃO

Neste capítulo serão apresentados os principais conceitos necessários para a compreensão dos sistemas de reconhecimento automático de locutor. Na seção 2.2 os conceitos básicos comuns na literatura são apresentados. A seção 2.3 trata do pré-processamento que deve ser efetuado no sinal de voz para sua posterior utilização. Na seção 2.4 são apresentadas as características de voz mais utilizadas com um comentário sucinto sobre as mesmas. Alguns sistemas de classificação muito utilizados no processamento de voz são vistos na seção 2.5. Na seção 2.6 é comentado o estado da arte no reconhecimento de locutor e, finalmente, na seção 2.7 é realizada a conclusão sobre os assuntos vistos neste capítulo.

2.2 CONCEITOS BÁSICOS

A voz é o resultado da soma de diferentes informações produzidas em uma complexa seqüência de transformações que ocorrem em diferentes níveis, tais como: semântico, lingüístico, articulatório e acústico (JAYANT, 1990). Primariamente o sinal de voz é composto por palavras que são unidas para formarem a mensagem que se quer transmitir; em um nível secundário, o sinal de voz contém informações sobre a identidade do locutor. Enquanto o reconhecimento de voz é baseado na lingüística do sinal, o reconhecimento de locutor é baseado na extração de informações sobre a identidade do locutor (REYNOLDS, 1992).

Variações na voz de uma pessoa são relacionadas em parte pelas diferenças anatômicas no trato vocal e em parte pelas diferenças no hábito de falar de diferentes locutores, relacionados à idiossincrasia[1].

As variações que podem ocorrer na voz pertinentes ao reconhecimento de locutor e que podem afetar seu desempenho são (WOODLAND, 1998):

- Intra-locutor (mesmo locutor): estado de saúde, estado emocional e ambiente.
- Inter-locutor (locutores diferentes): fisiológica, sotaque/dialeto.

[1]Maneira própria de ver, sentir, reagir, de cada indivíduo (FERREIRA, 2000).

- Estilo de falar: leitura/espontâneo, formal/casual.
- Distorções acústicas:
  - Meio de gravação das locuções (mídia).
  - Meio de transmissão (canal telefônico, canal radiofônico, etc.).
  - Ruídos aditivos (vozes de fundo, ruídos sonoros em geral).

O processamento de voz atua em diversas áreas, possuindo muitas aplicações, algumas já citadas para o reconhecimento de locutor (capítulo 1). Na FIG. 2.1 é apresentado algumas destas áreas, com maior detalhamento para o reconhecimento de locutor (CAMPBELL, 1997).

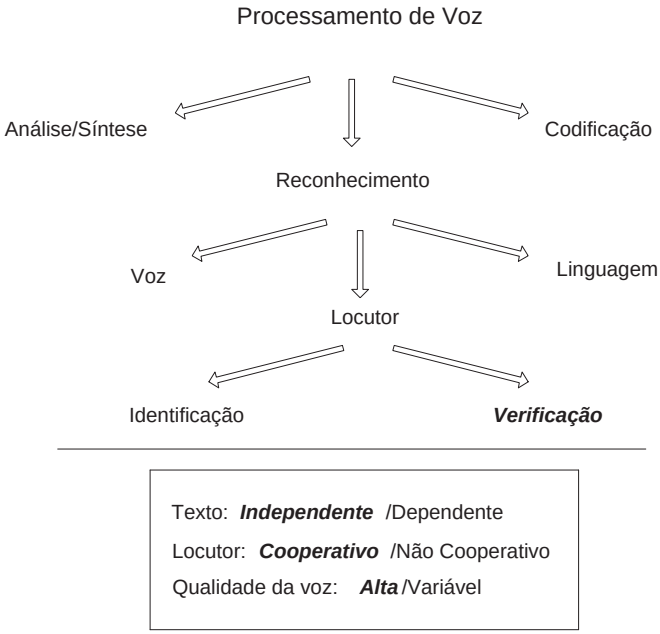

FIG. 2.1: Algumas áreas do processamento de voz, com destaque para o reconhecimento de locutor. As palavras em negrito indicam a abrangência desta dissertação.

Como já foi comentado no capítulo 1, o reconhecimento de locutor pode ser dividido em duas áreas clássicas: identificação e verificação.

- *Identificação*: É a tarefa de identificar uma pessoa através de sua voz em um conjunto conhecido de N locutores, denominado fechado quando envolver N decisões ou aberto quando houver $N + 1$ decisões (decide-se também, se a voz pertence a algum dos componentes do conjunto ou a nenhum deles) (REYNOLDS, 1992). O desempenho

de um sistema de identificação é degradado pelo aumento do número de locutores, pois aumenta-se o número possível de decisões (JAYANT, 1990).

- *Verificação*: É a tarefa de verificar se uma dada voz (locução) pertence ou não a uma determinada pessoa sendo, portanto, uma decisão binária. A decisão é feita no denominado conjunto aberto de locutores (REYNOLDS, 1995), porque o reconhecimento é efetuado em um grupo de locutores desconhecidos (possíveis impostores). Considerando-se um bom projeto do sistema, com estatística suficiente, o desempenho é independente do número de locutores (JAYANT, 1990).

Recentemente, o *National Institute of Standards and Technology* - NIST (MARTIN, 2000, NIST, 2000) apresentou uma nova tarefa no reconhecimento de locutor que é a segmentação e agrupamento por locutor. Esta tarefa visa determinar os trechos que foram falados por um determinado locutor durante uma conversação. O agrupamento consiste em juntar-se os segmentos de voz pertencentes a um mesmo locutor. Um dos objetivos desta tarefa são aplicações forenses.

Quanto à dependência do texto, o reconhecimento pode ser dependente ou independente. Os sistemas que exigem uma fala pré-determinada, são sistemas dependentes do texto. Estes sistemas podem fazer comparações precisas e confiáveis entre duas locuções com o mesmo texto, em ambientes foneticamente similares e exigindo de 2 a 3 segundos de voz para treinamento e teste. Em sistemas independentes do texto, tais comparações não são fáceis de serem obtidas, o que torna o desempenho de tais sistemas inferiores aos sistemas dependentes do texto. Para se obter uma razoável estatística do sinal, geralmente é exigido de 10 a 30 segundos de voz para treinamento e teste (JAYANT, 1990). Trabalhos recentes têm utilizado de 1 a 2 minutos de treinamento e de 3 a 30 segundos de teste, com qualidade variável do sinal de voz (MARTIN, 2000).

Quanto aos locutores (SOUSA, 1996):

- Cooperativos: Sabem que estão sendo reconhecidos, é o caso quando pronunciam palavras específicas ou não, mas falam de forma clara para ajudar o sistema de reconhecimento.

- Não-cooperativos: Não sabem que estão sendo reconhecidos, sua fala não é condicionada para ajudar o sistema de reconhecimento.

Quanto à qualidade da voz:

- Alta: Sinal de voz, gravado em ambiente com pouco ou nenhum ruído (limpo); voz pronunciada sem erros, também não afetada pelo estado de saúde de quem fala.

- Variável: Sinal de voz, gravado em ambiente com razão sinal/ruído variável, podendo ser corrompido por ruído aditivo e/ou convolucional (canal telefônico, meio de transmissão, cápsula do microfone, etc.), como também por problemas de saúde do locutor.

Uma aproximação geral para um sistema de reconhecimento de locutor, consiste basicamente de três etapas: aquisição do sinal de voz (conversão do sinal analógico para digital), extração das características do sinal de voz pertinentes ao reconhecimento e sistema classificador. O classificador atuará com base em modelos de referência estimados no treinamento, podendo utilizar modelos para simular impostores e condições de ruído, o denominado *background*, como mostrado na FIG. 2.2.

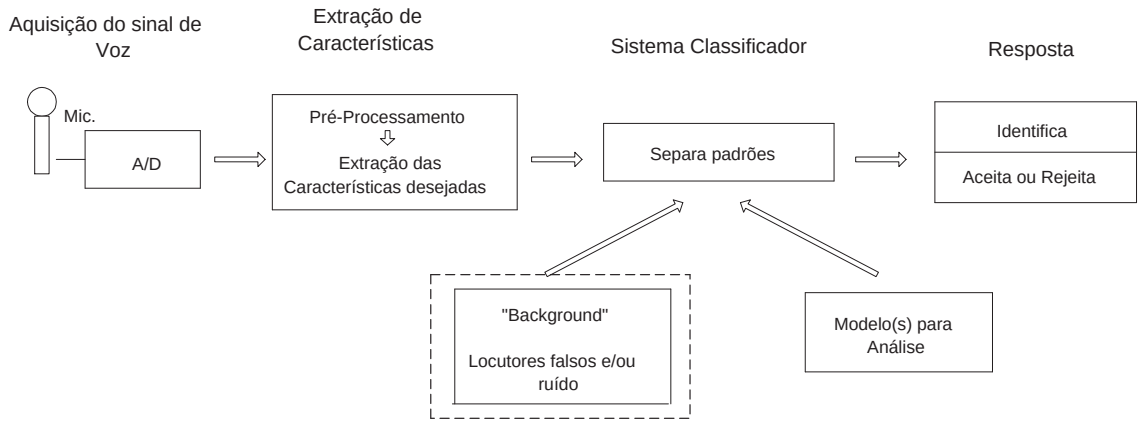

FIG. 2.2: Sistema genérico para reconhecimento de locutor.

Na aquisição do sinal, a voz (pressão acústica) é transformada em um sinal elétrico (analógico) através do microfone, sendo convertido para um sinal digital por um conversor analógico/digital. Esse sinal é então pré-processado, para supressão de informações desnecessárias e/ou ênfase nas importantes (RABINER, 1978). Finalmente, as características desejadas são extraídas gerando vetores de características, os quais darão os padrões para o classificador. Este, após treinamento, visará separar classes distintas. Na identificação de locutor, todos os modelos treinados são avaliados com uma locução de teste. O modelo que apresentar melhor resultado é aceito como *verdadeiro*, ou seja, a locução de teste pertence ao locutor cujo modelo venceu. Na verificação, o modelo de um pretenso locutor determinará se a locução de teste pertence ou não ao locutor. Tal modelo pode ser comparado com outro (vide caixa pontilhada da FIG. 2.2), que procura modelar o universo de locutores falsos e/ou ruído. A decisão será dada com base em limiares pré-estimados, de acordo com os valores apresentados pelo classificador. Maiores detalhes serão apresentados

no capítulo 4.

2.3 PRÉ-PROCESSAMENTO

O pré-processamento é efetuado sobre o sinal de voz antes da extração das características, servindo para adequar o sinal ao processamento que será efetuado; pode ser uma redução da taxa de amostragem, a eliminação de algum trecho inconveniente da gravação, uma filtragem ou uma normalização. A redução da taxa de amostragem pode ser conveniente quando o sinal adquirido possui uma taxa de amostragem que cobre uma faixa de freqüências maior que a desejada ou existente no sinal, diminuindo o número de amostras a serem processadas. A eliminação de trechos indesejáveis do sinal pode ser a eliminação dos trechos que não possuam voz, mas somente ruído de fundo, ou mesmo a eliminação de partes do som com as quais não se deseje trabalhar (LIU, 1999).

A filtragem é utilizada para ressaltar algumas freqüências do espectro e/ou diminuir outras, podendo, por exemplo, ser utilizada para a supressão de uma faixa de ruído. No processamento de voz costuma-se utilizar um filtro com resposta a impulso finita (FIR) (PICONE, 1991) dado por:

$$H_{pre} = \sum_{k=0}^{N_{pre}} a_{pre}(k) z^{-k} \tag{2.1}$$

Normalmente este filtro possui um único coeficiente, sendo conhecido como filtro de *pré-ênfase*, resultando em:

$$H_{pre} = 1 + a_{pre} z^{-1} \tag{2.2}$$

Valores típicos para a_{pre} estão compreendidos entre $-1,0$ e $-0,4$. Valores próximos a -1 são mais utilizados. A resposta em freqüência de um filtro de pré-ênfase com $a_{pre} = -0,95$ é apresentada na FIG. 2.3.

O filtro de pré-ênfase é utilizado para amplificar as regiões do espectro de maior freqüência. As explicações para se utilizar tal filtro são:

- Para eliminar a combinação da inclinação espectral negativa ($\simeq$ - 12 dB/oitava) introduzida pelo modelo glótico com a inclinação positiva ($\simeq$ + 6 dB/oitava) introduzida pelo modelo de radiação pelos lábios (PICONE, 1991).

- O ouvido humano é mais sensível para freqüências acima de $1KHz$. Como o filtro de pré-ênfase amplifica esta região do espectro, ele ressalta partes perceptualmente

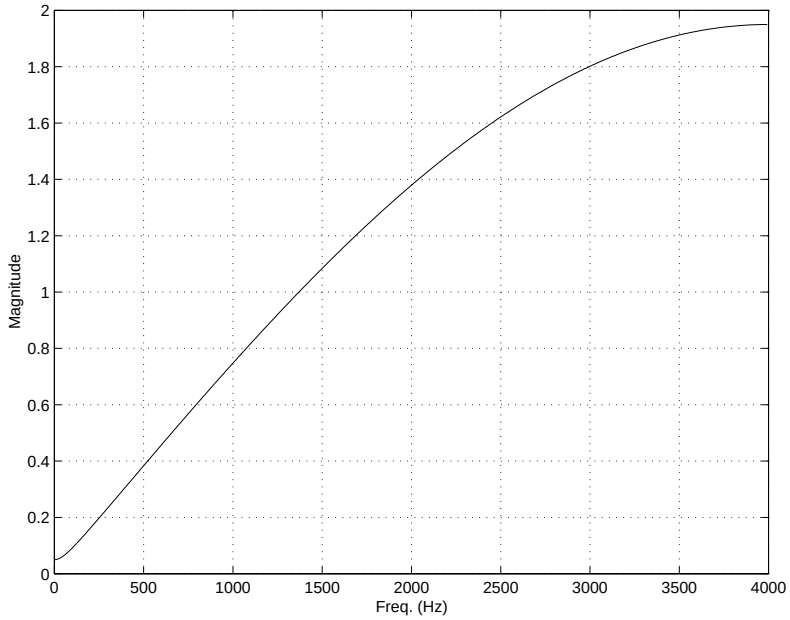

FIG. 2.3: Resposta em freqüência de um filtro de pré-ênfase com $a_{pre} = -0,95$, para um sinal amostrado a uma taxa de $8KHz$.

importantes do espectro (PICONE, 1991).

- Previnir instabilidade númerica, devido aos baixos valores nas regiões de maiores freqüências do espectro (DELLER, 1993).

Após o sinal ter sofrido ou não tal filtragem, algumas vezes se torna conveniente fazer a normalização do sinal de voz para que assuma valores dentro de uma faixa específica, tal como $[-0,5\ 0,5]$ e $[-1\ 1]$. Tal normalização visa corrigir a intensidade de sons gravados em seções diferentes ou por pessoas diferentes. A normalização é muito útil para reconhecimento de voz (SANTOS, 1997).

2.4 EXTRAÇÃO DE CARACTERÍSTICAS

A análise do sinal de voz consiste na extração de suas características relevantes, realizando-se uma compressão do sinal para posterior transmissão, armazenamento, síntese ou reconhecimento automático.

Estas características deverão atender, na medida do possível, às seguintes condições (ATAL, 1976):

- Eficiente na representação da informação do locutor ou do texto.
- Fácil de determinar.

- Estável ao longo do tempo.
- Ocorrer naturalmente e freqüentemente na voz.
- Mudar pouco de um ambiente de gravação para outro.
- Não ser vulnerável à mímica[2] (para o reconhecimento de locutor).

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

Para o uso das técnicas convencionais de análise aplicadas ao sinal de voz, é necessário trabalhar com pequenos intervalos do sinal, supostos estacionários. Sendo o sinal de voz um processo estocástico, em geral não estacionário, e sabendo-se que o trato vocal muda de forma muito lentamente com o passar do tempo na voz contínua, muitas partes da onda acústica podem ser supostas estacionárias num curto intervalo de tempo. Este intervalo caracteriza o tamanho da janela de análise a ser utilizada, em cuja duração, de 10 a 40 ms, o sinal de voz pode ser considerado como um processo estacionário (RABINER, 1978). Uma janela de comprimento longo tende a produzir uma melhor representação espectral do sinal, desde que este esteja na região de estacionariedade. Uma janela de comprimento pequeno tende a ser melhor em análises no domínio do tempo. Com o objetivo de atenuar o efeito do fenômeno de *Gibbs* (*ripple* em amplitude na resposta em freqüência da janela retangular) devido ao truncamento do sinal de análise no domínio do tempo, deve-se utilizar janelas que possuam, no domínio da freqüência, um lóbulo principal o mais estreito possível e uma grande diferença de amplitude entre o lóbulo principal e o primeiro lóbulo lateral. As janelas mais utilizadas na prática e que procuram atender a estas condições são as seguintes: Hamming, Hanning, Retangular, Bartlett (triangular), Blackman e Kaiser (OPPENHEIM, 1998). Para compensar o efeito provocado pelo amaciamento da janela temporal, uma sobreposição entre janelas é efetuada aumentando a correlação entre janelas adjacentes. Na FIG. 2.4 é apresentada a interpretação do janelamento com sobreposição. Cada janela representa uma parte do sinal do qual serão extraídas as características desejadas. Cada parte do sinal, após o janelamento e/ou processamento, pode ser chamada de quadro de voz (vetor).

A porcentagem de sobreposição, é dada por:

[2]Imitação da voz de uma pessoa por outra.

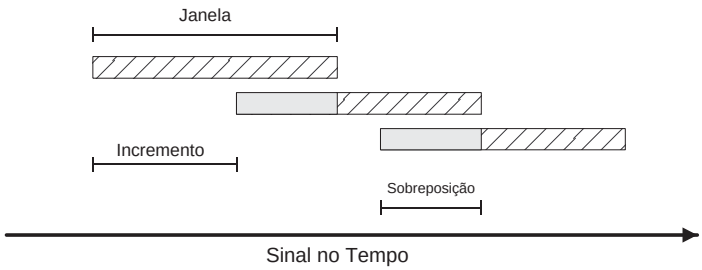

FIG. 2.4: Interpretação do janelamento no domínio do tempo.

$$Sobr = \frac{T_{jan} - T_{Inc}}{T_{jan}} \times 100\% \tag{2.3}$$

onde T_{jan} e T_{inc} são o tamanho da janela e o tamanho do incremento para uma nova janela, respectivamente, ambos em unidade de tempo ou em número de amostras. Sobreposições típicas estão na ordem de 50% ou mais.

Dentre as várias características de voz que podem ser utilizadas no processamento de voz, pode-se destacar:

- **Taxa de cruzamento por zero**: é um método simples de análise no domínio do tempo que é baseado na medida de cruzamento por zero de um sinal digital. No contexto da implementação digital, um cruzamento por zero ocorre entre instantes de amostragem n e $n-1$, se:

$$sinal[x(n)] \neq sinal[x(n-1)] \tag{2.4}$$

  ou seja, um cruzamento por zero ocorre sempre que duas amostras sucessivas trocam de sinal algébrico.

- **Energia**: É uma das informações mais simples extraídas de um sinal, calculada a cada N amostras, como:

$$E = \sum_{t=1}^{N} x^2(n) \tag{2.5}$$

  Costuma-se utilizar o logaritmo da energia, para suavizar grandes variações de magnitude e ressaltar as pequenas. Esta caraterística é conhecida como Log-energia.

- **Freqüência fundamental (*pitch*)**: É definida como a freqüência na qual as cordas vocais vibram durante um som sonoro. A freqüência fundamental é uma característica encontrada apenas em sons sonoros; tal fato é justificado fisicamente porque na produção de sons sonoros as cordas vocais vibram em tal freqüência, que os sons produzidos são quase periódicos.

- **Formantes**: A cavidade bucal, junto com a faringe, pode ser considerada como um tubo de formato irregular, quase fechado em uma extremidade e aberto na outra. Um tubo assim tem um número de freqüências ressonantes, que correspondem a picos, conhecidas como formantes. No contexto da produção da voz, as freqüências de ressonância do trato vocal são definidas como freqüências formantes ou simplesmente formantes, sendo portanto dependentes do locutor. As freqüências formantes dependem do formato e das dimensões do trato vocal, sendo cada formato caracterizado por um conjunto dessas freqüências. Diferentes sons são formados na variação do trato vocal, o que varia as propriedades espectrais do sinal de voz com o tempo (DELLER, 1993).

- **Coeficientes de predição linear (*Linear Prediction Coefficientes* - LPC)**: A idéia principal da predição linear é a de que uma amostra de voz pode ser aproximada por uma combinação linear de amostras passadas. O modelo é estimado por um filtro digital (pólos e/ou zeros) que simula o trato vocal e nasal. A ordem do filtro é escolhida de tal modo que se ele for excitado por impulsos, sua saída apresente um espectro muito próximo do espectro do sinal que está sendo modulado (RABINER, 1978). O LPC será visto em detalhe no capítulo 4.

- **Cepestro**: A obtenção do cepestro de um sinal é uma transformação homomórfica, ou seja, transforma no caso da modelagem de voz, uma convolução numa soma. O cepestro pode ser calculado como:

$$c(n) = \mathcal{F}^{-1}\Big[\log\Big[\mid \mathcal{F}[x(n)] \mid\Big]\Big] \tag{2.6}$$

  ou seja, é a transforma inversa de Fourier do logaritmo do módulo da transformada de Fourier de um sinal $x(n)$. Se o sinal $x(n)$ resultar de um processo de convolução, então a transformada de Fourier deste sinal representará uma multiplicação no domínio da freqüência, o logaritmo transformará este produto em uma soma[3], comprimindo de certa forma o sinal. A transformada inversa de Fourier retorna o cepestro ao domínio do tempo. Na voz a resposta ao impulso do trato vocal convoluida pela excitação glotal (RABINER, 1978) pode ser separada pelo uso do cepestro. O cepestro também pode ser utilizado para eliminar o ruído convolucional produzido por um filtro qualquer, além do que, pode ser utilizado para cálculo dos coeficientes LPC,

[3]Domínio de freqüência do cepestro: Qüefrência.

bem como, para o cálculo da *pitch*. Por tais motivos, o cepestro é uma característica muito utilizada no processamento de voz (DELLER, 1993).

- **Mel-cepestro**: É uma das características mais utilizadas no reconhecimento de locutor (REYNOLDS, 1994). Os coeficientes mel-cepestrais (*mel-cepstrum coefficients* - MCC), são obtidos de um sistema que aproxima a resposta em freqüência do ouvido humano, do qual são extraídos os coeficientes cepestrais (PICONE, 1991). Presume-se que o sucesso no reconhecimento de locutor é devido à capacidade que o MCC tem em capturar as diferenças interlocutor, pois carrega informações sobre o trato vocal em conjunto com características da percepção auditiva (SARMA, 1999). No capítulo 3 o MCC será visto em detalhe.

- **Predição linear perceptual (*Perceptual Linear Predictive* - PLP)**: Os coeficientes LPC fornecem uma envoltória espectral suavizada, se a ordem do modelo é bem escolhida. Um modelo só de pólos faz uma boa aproximação das áreas de alta concentração de energia do trato vocal (formantes), enquanto despreza harmônicos de baixa energia e outros detalhes espectrais menos relevantes. Entretanto, essa propriedade não é consistente com as peculiaridades do ouvido humano. Com o objetivo de levar em conta as características do sistema auditivo, HERMANSKY (1990) estudou uma classe alternativa de transformações espectrais a partir da análise LPC. O espectro do sinal de voz é modificado de acordo com características acústicas antes da aproximação pelo método autorregressivo. A idéia é semelhante à utilizada no cálculo dos coeficientes mel-cepestro. Entretanto, Hermansky escolheu filtros assimétricos com banda maior que a dos filtros triangulares (vide capítulo 3) para simular as bandas críticas e a escala *Bark* (PICONE, 1991) para espaçamento desses filtros. Além disso, incorporou também pré-ênfase e compressão com o objetivo de simular determinadas áreas do ouvido humano. Da mesma forma que se pode obter os coeficientes cepestrais a partir do LPC, pode-se obter os coeficientes PLP-cepestro do PLP, os quais são muito utilizados para reconhecimento de voz (SANTOS, 1997).

- **Coeficientes delta e delta-delta**: Estes parâmetros são obtidos através das derivadas de primeira e segunda ordem das características de voz, respectivamente. São utilizados para representar as mudanças dinâmicas no espectro de voz e, desse modo, detectar variações bruscas dentro do espectro, como também analisar a variação dinâmica das características. Em processamento digital de sinais existem várias maneiras para aproximação do cálculo de derivadas de primeira ordem, uma aproximação muito popular é dada por (PICONE, 1991):

$$\dot{s}(n) \equiv \frac{d}{dt}s(n) \approx s(n) - s(n-1) \tag{2.7}$$

onde $s(n)$ é a característica sob análise na janela n. A derivada de segunda ordem pode ser obtida aplicando a equação acima sobre $\dot{s}(n)$. A primeira derivada é referida como uma característica de velocidade e a segunda derivada como uma característica de aceleração. Os coeficientes delta são geralmente calculados sobre a log-energia e os coeficientes cepestrais.

2.5 SISTEMAS DE CLASSIFICAÇÃO

Os modelos dos locutores são construídos a partir das características extraídas do sinal de voz. Quando um locutor novo entra no sistema, um modelo de sua voz, baseado nas características extraídas, é gerado e armazenado. Para a verificação de locutor, um sistema de classificação comparará valores do sinal de entrada com o modelo armazenado do pretenso locutor, aceitando-o ou rejeitando-o. Na identificação de locutor o modelo que apresentar maior similaridade indica a quem pertence o sinal de entrada.

Existem dois tipos clássicos de modelos: o modelo estatístico ou estocástico e o modelo baseado em casamento de padrões característicos (CAMPBEL, 1997). No modelo estatístico, o sistema de classificação é probabilístico e resulta numa medida de verossimilhança, ou probabilidade condicional, dada pela observação do modelo. Para modelos baseados em casamento de padrões característicos, o sistema de classificação faz comparações. É assumido que a observação é uma réplica imperfeita do modelo armazenado e geralmente o alinhamento dos quadros do modelo para os quadros observados é feito para minimizar uma medida de distância d. Um outro modelo é fornecido pelas redes neurais, que são consideradas aproximadoras universais de funções, não se enquadrando adequadamente nos modelos acima. A seguir os modelos referidos serão descritos.

Modelos Baseados em Casamento de Padrões Característicos:

- **Média de Longo Termo** (CAMPBELL, 1997): consiste de um simples modelo $\overline{\mathbf{x}}$, o qual representa o modelo do sinal de referência. A distância de valores entre o modelo $\overline{\mathbf{x}}$ armazenado e o vetor de entrada $\mathbf{x}_i$ de um locutor desconhecido é dada por $d(\mathbf{x}_i, \overline{\mathbf{x}})$. O modelo de um locutor pode ser um centróide (média) de um conjunto de N vetores de treinamento:

$$\overline{\mathbf{x}} = \mu = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j \tag{2.8}$$

Muitas medidas de distâncias diferentes entre os vetores $\mathbf{x}_i$ e $\overline{\mathbf{x}}$ podem ser expressas como:

$$d(\mathbf{x}_i, \overline{\mathbf{x}}) = (\mathbf{x}_i - \overline{\mathbf{x}})^T \mathbf{W} (\mathbf{x}_i - \overline{\mathbf{x}}) \tag{2.9}$$

onde $\mathbf{W}$ é uma matriz de ponderação. Se $\mathbf{W}$ é uma matriz identidade a distância é a *Euclidiana*; se $\mathbf{W}$ é a matriz covariância[4] inversa correspondente aos vetores de referência $\mathbf{x}$, então esta é a distância de *Mahalanobis* (RABINER, 1993). A distância de *Mahalanobis* dá uma menor ponderação aos componentes com maior covariância e é equivalente à distância *Euclidiana* sobre os componentes principais, os quais são os autovetores da matriz covariância do espaço original (FUKUNAGA, 1990).

- **Alinhamento Temporal Dinâmico (*Dynamic Time Warping* - DTW)**: é um método muito popular para compensar a variabilidade entre dois sinais no domínio do tempo. Um modelo dependente do texto é uma seqüência de amostras $(\overline{\mathbf{x}}_\mathbf{1}, ..., \overline{\mathbf{x}}_\mathbf{N})$ a qual deve ser casada com uma seqüência de entrada $(\mathbf{x_1}, ..., \mathbf{x_M})$. Em geral, N não é igual a M devido à inconsistência temporal na voz humana. O valor assimétrico de casamento z é dado por:

$$z = \sum_{i=1}^{M} d(\mathbf{x}_i, \overline{\mathbf{x}}_{j(i)}) \tag{2.10}$$

onde o índice de casamento $j(i)$ é dado por um algoritmo DTW. Dado o sinal de referência e o de entrada, o DTW faz um limite, mapeamento por partes de um (ou ambos) eixo(s) temporal(ais) para alinhamento não linear dos dois sinais, enquanto minimiza z. Ao final do DTW, a distância acumulada é a base do valor de casamento entre os sinais. Este método mede a variação no tempo (trajetória) dos parâmetros correspondentes a variações dinâmicas na produção da voz (CAMPBEL, 1997). A FIG. 2.5 mostra como o DTW trabalha para fazer o alinhamento temporal entre dois sinais.

- **Quantização Vetorial (*Vector Quantization* - VQ)**: outra forma para a modelagem de sinais de voz é a quantização vetorial. Esta faz o agrupamento dos dados de treinamento de um locutor qualquer, em regiões distintas do espaço, gerando um dicionário de códigos (*codebook*) com N regiões distintas (RABINER, 1993). Para avaliar a similaridade entre um conjunto de dados de entrada e o modelo, é calculada a menor distância entre os vetores de entrada num dicionário C com relação a seus N grupos. O valor de similaridade para L vetores de entrada (características de voz)

[4] $K = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$.

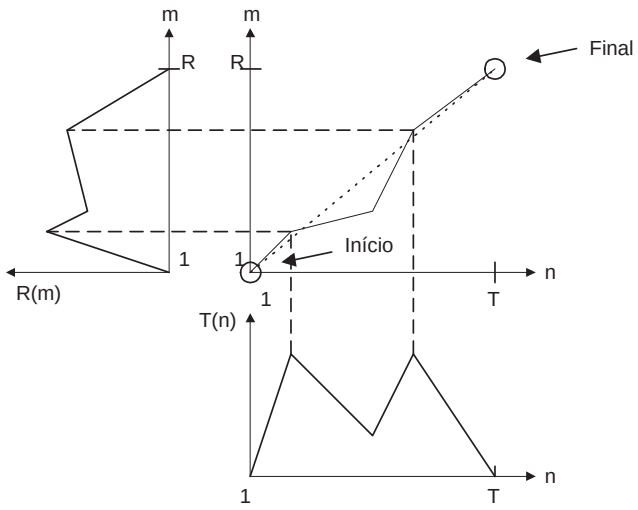

FIG. 2.5: Alinhamento temporal não linear de um sinal de teste $T(n)$ em relação a um modelo $R(m)$.

é portanto, dada por:

$$z = \sum_{i=1}^{L} \min_{\overline{\mathbf{x}} \in C} \{d(\mathbf{x}_i, \overline{\mathbf{x}})\} \tag{2.11}$$

O processo de agrupamento usado para formar o dicionário de códigos não leva em conta a informação temporal dos dados de treinamento. Assim não há a necessidade de fazer um alinhamento temporal entre dados, o que simplifica bastante o classificador. Portanto, a informação temporal, mesmo importante, é negligenciada.

- **Modelo Autorregressivo Vetorial - AR-Vetorial**: é um modelo baseado nos coeficientes de predição linear, que avaliam a evolução espectral das características de voz. A distância de Itakura (ITAKURA, 1975) é utilizada na comparação entre modelos. O AR-Vetorial será visto em detalhe no capítulo 4.

Modelos estatísticos:

- **Modelo Escondido de Markov (*Hidden Markov Model* - HMM)**: é um modelo probabilístico do sinal de voz que descreve as características variantes no tempo. Este modelo é um processo duplamente estocástico, no qual um processo estocástico não é observável (daqui o termo escondido), mas que pode ser observado através de outro processo estocástico que produz uma seqüência de observações (RABINER, 1989). A técnica do HMM representa o modelo de produção da voz como um sistema modelado por um número finito de estados. Em intervalos discretos de tempo, o sistema passa de um estado a outro, cada estado produzindo uma saída. A transição entre estados é aleatória, como é a saída associada a cada estado. Devido à

aleatoriedade das transições entre estados e saídas, o modelo adapta a variação temporal do sinal de voz (JAYANT, 1990). Uma estrutura comum de HMM é o modelo esquerda-direta, mostrado na FIG. 2.6.

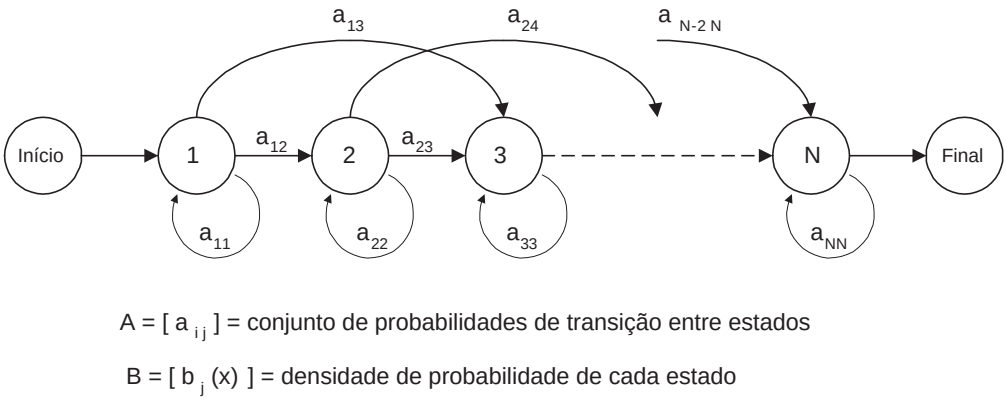

FIG. 2.6: Estrutura HMM esquerda-direita de N estados.

Este modelo apresenta um simples estado de partida e um final, e os demais estados têm probabilidades de transições para si próprios e para os dois próximos. Cada estado é formado por um conjunto de densidades de probabilidade, geralmente misturas de fdp's gaussianas, que modelam algum evento acústico, como um sílaba ou palavra. A montagem de um sistema HMM consiste das seguintes operações:

- Otimização dos parâmetros do modelo, para descrever da melhor forma a seqüência observada (treinamento).
- Dada uma seqüência de observações, escolher uma seqüência de estados que é ótima de acordo com algum critério pré-definido.
- Dada uma seqüência de observações e um modelo, calcular a probabilidade da seqüência de observações.

Um sistema de verificação de locutor com HMM pode usar modelos de locutor baseados em sentenças com algumas palavras, uma simples palavra ou fonemas. Tipicamente, várias palavras em uma frase (uma cadeia de sete a dez dígitos, por exemplo) são usadas, e modelos para cada palavra individual são combinados de acordo com a sentença.

O HMM que possui probabilidade de transição entre todos os estados é chamado de *ergódico*, sendo mais complexo que o modelo esquerda-direita e capaz de fornecer uma modelagem mais ampla de classes fonéticas, pode ser utilizado para reconhecimento

de locutor independente do texto (FURUI, 1996). Para maiores detalhes sobre HMM consultar (RABINER, 1989, DELLER, 1993).

- **Modelo de Mistura de Gaussianas - GMM**: Este modelo é composto por um conjunto de funções densidade de probabilidade gaussianas, que podem modelar várias classes fonéticas. É um modelo que não considera a evolução temporal do sinal, sendo próprio para sistemas de reconhecimento de locutor independente do texto. Detalhes deste modelo serão vistos no capítulo 4.

Redes Neurais (RNs):

- RNs são sistemas paralelos distribuídos e compostos por unidades de processamento simples, os chamados neurônios, que computam determinadas funções matemáticas (normalmente não lineares). Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais. Na maioria dos modelos estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede (BRAGA, 1998). O funcionamento destas redes é inspirado em uma estrutura concebida na natureza: o cérebro humano.

  A solução de problemas através de RNs é bastante atrativa, já que a forma como estes são representados internamente pela rede e o paralelismo natural inerente à arquitetura das RNs criam a possibilidade de um desempenho superior ao dos modelos convencionais. Nas RNs, o procedimento usual na solução de problemas passa inicialmente por uma fase de aprendizado, onde um conjunto de exemplos é apresentado para a rede, a qual extrai automaticamente as características necessárias para representar a informação recebida. Essas características são utilizadas posteriormente para gerar respostas para o problema.

  RNs conhecidas como *feed-forward* têm sido utilizadas no reconhecimento de locutor (BEZERRA, 1994), onde cada locutor tem uma rede neural treinada com suas características. Assume-se que a inclusão de muitas pessoas nos dados de treinamento habilita a RN a modelar diretamente as diferenças da voz de cada locutor em relação a impostores (FURUI, 1996).

2.6 ESTADO DA ARTE

No reconhecimento automático de locutor, o que se vem tentando é imitar a capacidade que o ser humano tem em conseguir reconhecer pessoas através de suas vozes. Antigamente

a máquina não conseguia superar o homem, hoje ela já é capaz de tal feito, mas está aquém da capacidade que o ser humano tem de reconhecer a voz quando corrompida por ruído. Em NIELSEN (2000) é feito um estudo comparativo entre o desempenho da máquina versus o do ser humano na verificação de locutor. A experiência foi desenvolvida dentro do rigor científico possível, demonstrando que para sinais de voz com qualidade (gravados em ambiente pouco ruidoso) o ser humano e a máquina empatam, com uma taxa de erro médio de 8%. Em gravações corrompidas por ruído, o ser humano e a máquina tiveram taxas de erro superiores aos anteriores, mas o homem superou a máquina em aproximadamente 40%. Nesta experiência foram utilizados 60 segundos de dados para treinamento e um tempo de teste de 3 segundos. Hoje, portanto, um dos fatores mais pesquisados é a diminuição do efeito do ruído no sinal de voz, para tentar superar o efeito degradativo que o mesmo insere no reconhecimento de locutor.

Uma das referências mundiais, aceita pela comunidade científica, para conhecer a tecnologia utilizada e pesquisada atualmente no reconhecimento de locutor, é dada pelo NIST. Este faz um concurso anual de âmbito mundial, para avaliar a melhor tecnologia no reconhecimento de locutor, ditando as regras para o concurso e fornecendo o banco de dados para desenvolvimento dos sistemas (MARTIN, 2000). Atualmente o modelo de mistura de gaussianas (GMM) tem sido uma ferramenta muito utilizada na verificação de locutor independente do texto (REYNOLDS, 1995, 2000). Comparações do GMM com outros sistemas de reconhecimento de locutor podem ser vistos em REYNOLDS (1992) e FURUI (1996). Algumas tendências podem ser vistas em trabalhos recentes as quais apontam novas direções a serem pesquisadas, tais como, a pesquisa de novas características (MALAYATH, 2000), o tratamento dos dados extraídos das locuções (AVENDAÑO, 1997) visando principalmente conseguir maior separabilidade entre classes e diminuir o efeito degradativo inserido pelo ruído nos sistemas de decisão, e o processamento em sub-bandas que tenta diminuir o efeito do ruído que atinge algumas faixas do espectro (SHARMA, 1999, AVENDAÑO, 2000).

Pesquisas sobre modelos de comparação (*background*) mais robustos a ruído no sistema de decisão, também estão continuamente em estudo (REYNOLDS, 2000, CHAGNOLLEU, 2000, GRAVIER, 2000, AUCKENTHALER, 2000). Outras tendências estão focadas na verificação de locutor usando segmentos específicos da voz (análise fonética), as quais exigem sistemas de segmentação relativamente complexos (FURUI, 1996, SARMA, 1999, FREDOUILLE, 2000, DELACRÉTAZ, 2000, SAVIC, 1992). Muitas destas pesquisas têm sido feitas primeiramente para reconhecimento de voz e estendidas posteriormente para o reconhecimento de locutor. Técnicas muito utilizadas, como a quantização vetorial e

o DTW, foram abandonadas pelo baixo desempenho em sistemas independes do texto e corrompidos por ruído. A TAB. 2.1 extraída de CAMPBEL (1997) apresenta a cronologia do progresso no reconhecimento de locutor, onde a última coluna indica a taxa de erro e o tempo utilizado para testar o sistema de reconhecimento. Dessa tabela, percebe-se que o número de locutores para testes aumentou nos últimos anos e que as condições das gravações de voz tornaram-se mais reais, com predomínio do HMM e GMM.

Outras técnicas utilizadas no reconhecimento de locutor são o HMM (RABINER, 1989) e as RNs (FURUI, 1996, MARTIN, 2000), por apresentarem bons resultados. O HMM é utilizado em modelagens que possuam dependência temporal para reconhecimentos dependentes do texto. Em modelagens (sem dependência temporal) independentes ao texto, o uso do HMM não tem apresentado vantagens sobre o GMM (REYNOLDS, 1992). A FIG. 2.7 extraída de FURUI (1996) apresenta uma comparação entre o HMM ergódico e o GMM para a identificação de locutor. Observa-se que um HMM de um único estado, sem probabilidades de transição entre estados (GMM), supera outro com mais estados, onde existam probabilidades de transição, demonstrando que em casos sem dependência temporal as probabilidades de transição entre estados não possuem relevância.

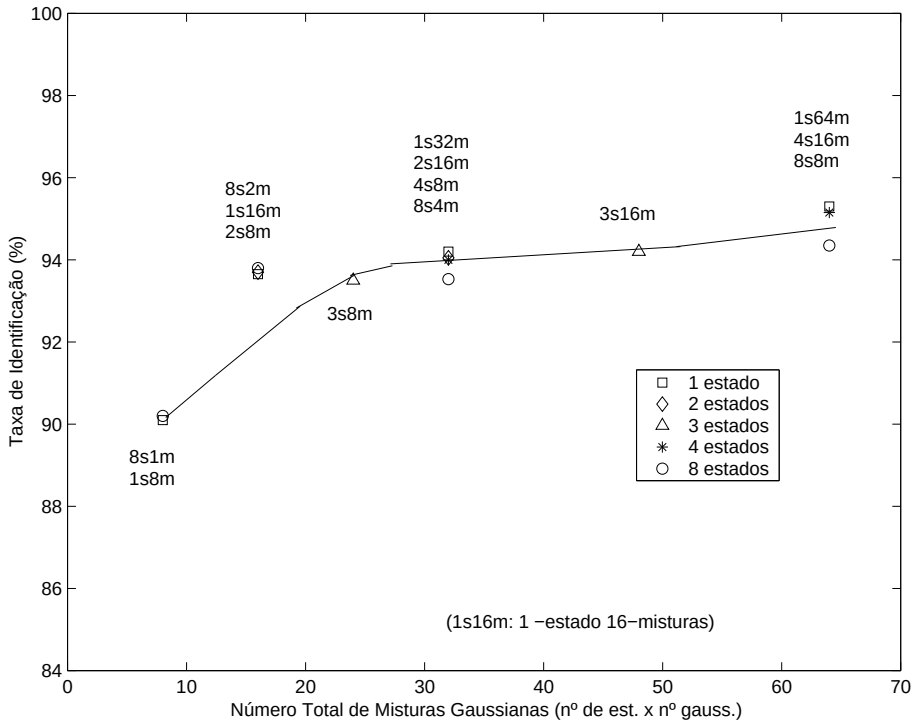

FIG. 2.7: Taxa de identificação de locutor em função do número de estados e misturas em HMMs ergódicos. Um único estado pode ser compreendido como um GMM (FURUI, 1996).

As RNs devem apresentar bons resultados no reconhecimento de locutor independente

TAB. 2.1: Cronologia selecionada no reconhecimento de locutor (CAMPBEL, 1997).

Fonte	Organiz.	Característ.	Método	Qualid. da voz	Texto	Nº Loc.	Erro(%) temp. teste(s) i:identif. v:verific.
Atal, 1974	AT& T	cepestro	Associação de padrões	Laborat.	Depend.	10	i:2% 0,5s v:2% 1s
Markel and Davis, 1979	STI	LPC	Estatist. de longo termo	Laborat.	Independ.	17	i:2% 39s
Furui, 1981	AT& T	cepestro normalizado	Associação de padrões	Telefone	Depend.	10	v:0,2% 3s
Schwartz, et al. 1982	BBN	LAR (RABINER, 1978)	fdp não paramétrica	Telefone	Independ.	21	i:2,5% 2s
Li and Wrench, 1983	ITT	LPC cepestro	Associação de padrões	Laborat.	Independ.	11	i:21% 3s v:4%, 10s
Doddington, 1985	TI	Banco de Filtros	DTW	Laborat.	Depend.	200	v:0,8% 6s
Soong, et al. 1985	AT& T	LPC	VQ (64)	Telefone	10 dígitos isolados	100	i:5% 1,5s i:1,5% 3,5s
Higgins and Wohlford, 1986	ITT	cepestro	DTW	Laborat.	Independ.	11	v:10% 2,5s v:4,5% 10s
Attili, et al. 1988	RPI	cepestro LPC Autocorr.	Projeção estatística de longo termo	Laborat.	Depend.	90	v:1%, 3s
Higgins, et al. 1991	ITT	LAR LPC-cepestro	DTW	Escritório	Depend.	186	v:1,7% 10s
Tishby, 1991	AT& T	LPC	HMM	Telefone	10 dígitos isolados	100	v:2,8% 1,5s v:0,8% 3,5s
Reynolds and Carlson, 1995	MIT-LL	Mel-cepestro	GMM	Escritório	Depend.	138	i: 0,8%, 10s v: 0,12% 10s
Che and Li, 1995	Rutgers	cepestro	HMM	Escritório	Depend.	138	i:0,56% 2,5s i:0,14% 10s v:0,62% 2,5s
Colombi, et al. 1996	AFIT	cepestro energia Dcepestro DDcepestro	HMM Monofone	Escritório	Depend.	138	i:0,22% 10s v:0,28% 10s
Reynolds, 1996	MIT-LL	Mel-cepestro DMel-cepest.	GMM	Telefone	Independ.	416	v:11%/16% 3s v:6%/8% 10s v:3%/5% 30s mesmo telef./ outro telef.

ao texto. Seus inconvenientes devem-se à difícil adaptação a novos locutores e condições de ruído, necessitando retreinamentos, além da exigência de grande quantidade de dados para treinamento (REYNOLDS, 1992). Segundo FURUI (1996), as RNs para reconhecimento de locutor comparam-se em desempenho a um quantizador vetorial. O GMM possui menor complexidade e treinamento mais rápido comparado aos sistemas com RNs e/ou HMMs, sendo facilmente adaptável a novas condições de ruído e a novos locutores. Maiores detalhes serão vistos no capítulo 4. Em NAKASONE (2001) é apresentado um trabalho de reconhecimento de locutor utilizando GMM, desenvolvido pelo FBI para uso forense. Esse trabalho demonstra a confiança no uso do GMM, principalmente em situações que exijam margens de erro muito pequenas.

Existem outros trabalhos que utilizam o AR-Vetorial no reconhecimento de locutor (BIMBOT, 1992, MONTACIÉ, 1992, CHAGNOLLEU, 2000). Como o AR-Vetorial apresenta um bom desempenho na identificação de locutor, surgiram trabalhos dele em união com o GMM (FLOCH, 1996) e em união com redes neurais (HADJITODOROV, 1994). Entretanto, uma análise mais completa do AR-Vetorial na verificação de locutor independente do texto, bem como suas vantagens e desvantagens em função dos diversos parâmetros envolvidos, não esta disponível na literatura.

Com base na literatura técnica, a característica que tem predominado no reconhecimento de locutor é o mel-cepestro, que incorpora características da percepção auditiva. Quase todas as referências bibliográficas citadas, fazem uso do mel-cepestro, devido ao seu bom desempenho. Uma comparação entre características robustas para identificação de locutor, é feita em REYNOLDS (1994), confirmando a preferência dada ao mel-cepestro. Incorpora-se geralmente o delta mel-cepestro, uma característica de velocidade, carregando mais informações sobre a identidade dos locutores, melhorando a resposta do sistema de reconhecimento.

No estado da arte de reconhecimento de locutor, uma grande preocupação é dada aos efeitos degradativos produzidos pelo ruído principalmente devido a condições diferentes de gravação para treinamento e teste. O sistema de classificação mais utilizado tem sido o GMM independente do texto. Outro ramo de pesquisa é o reconhecimento baseado em segmentos específicos de fala e poucas pesquisas aparecem na busca de novas características de voz, que possuam maior poder discriminativo para o reconhecimento de locutor.

2.7 RESUMO E CONCLUSÃO

Neste capítulo foram apresentados os conceitos básicos envolvidos no reconhecimento de locutor, como:

- Diferença entre identificação e verificação, a dependência do texto e os tipos de locutores.
- O pré-processamento para adequar o sinal de voz para a extração de características. O filtro de pré-ênfase para ressaltar as altas freqüências, a normalização do sinal e a eliminação de trechos irrelevantes ao reconhecimento.
- Características mais comuns no processamento de voz: taxa de cruzamento por zero, energia, *pitch*, formantes, coeficientes de predição linear (LPC), cepestro, mel-cepestro, coeficientes PLP e coeficientes delta e delta-delta.
- Os sistemas de classificação, especificamente: (1) baseados em casamento de padrões característicos, como a média de longo termo, DTW, VQ, AR-Vetorial. (2) estatísticos, como HMM e GMM e (3) redes neurais, que se adaptam dinamicamente para solucionar o problema de classificação, sendo consideradas aproximadores universais de funções.
- Estado da arte: observou-se que no reconhecimento de locutor independente do texto, o GMM com o uso do mel-cepestro e seus deltas tem apresentado melhores resultados. Estudos se intensificaram no sentido de evitar o efeito degradativo do ruído nos sistemas de RAL, principalmente nos sistemas que usam o GMM.

O reconhecimento automático de locutor é uma tarefa difícil de ser efetuada com precisão devido à dinâmica encontrada no sinal de voz e aos diferentes ruídos existentes em ambientes reais. A pesquisa está voltada para a superação desses problemas; entretanto, a máquina ainda está muito aquém do desempenho humano. O estudo de novas características, mais discriminantes e robustas ao ruído, faz-se necessário, bem como, de sistemas de decisão mais eficientes.

3 CARACTERÍSTICAS E TRANSFORMAÇÃO UTILIZADA

3.1 INTRODUÇÃO

Neste capítulo será abordado o conjunto de características da voz utilizado nesta dissertação para o reconhecimento de locutor: os coeficientes mel-cepestrais, bem como a transformação utilizada com a intenção de aumentar o seu poder discriminativo (análise de componentes principais).

Na seção 3.2 será descrito o algoritmo utilizado para a eliminação do silêncio entre palavras e a importância de tal fato; na seção seguinte, 3.3, a teoria que levou à utilização da característica de percepção auditiva humana na extração de coeficientes cepestrais é explanada. A seção 3.4 apresenta o fundamento teórico da utilização da análise de componentes principais ao processamento de dados. A ligação entre os coeficientes mel-cepestrais e o PCA, para o reconhecimento de locutor, e as vantagens dessa ligação é vista na seção 3.5. Finalmente, na seção 3.6, é feito um breve resumo e a conclusão sobre os temas vistos neste capítulo.

3.2 ELIMINAÇÃO DO SILÊNCIO

Na extração das características da voz, obviamente o que se deseja é processar somente trechos de voz. Portanto, deve-se eliminar o silêncio no início e no final de uma locução[1], bem como o silêncio entre palavras. Estes trechos de silêncio compreendem o ruído de fundo das gravações, constituindo informação inútil na extração de características e proporcionando um acréscimo computacional desnecessário.

Existem vários algoritmos para a eliminação destes trechos de ruído de fundo. Um dos algoritmos mais simples, baseado na taxa de cruzamento por zero e energia do sinal e muito utilizado para determinar os *endpoints*, é dado por RABINER (1974). Nesta dissertação foi proposto um algoritmo simples, adaptado de RABINER (1974) e baseado somente na magnitude média do sinal, para a eliminação do silêncio encontrado nas gravações.

O algoritmo proposto é baseado na estimação da magnitude média do sinal, considerando janelas de 10 ms sem sobreposição. Os 100 ms iniciais (RABINER, 1974) e 30 ms finais da locução são considerados como ruído de fundo. Desta estimação inicial e final, é calculada a média e o desvio padrão. Um limiar do valor da magnitude média é esti-

[1] Denominado *endpoint*.

pulado e todo sinal abaixo deste limiar é considerado como ruído de fundo. Considera-se sempre 3 janelas adjacentes para evitar ruídos espúrios. O tempo de 30 ms para o final da locução foi utilizado porque as gravações possuem um tempo de silêncio final menor que o inicial, uma vez que a parada na gravação é mais rápida que o início da fala no começo da gravação. Para melhor compreensão observar a FIG. 3.1.

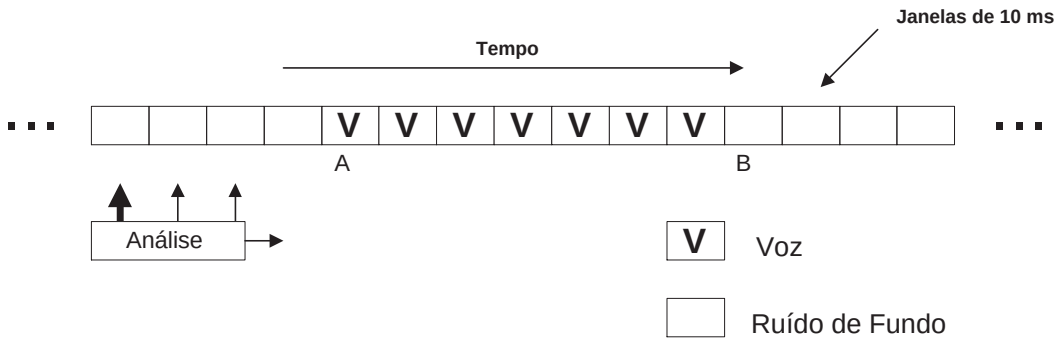

FIG. 3.1: Ilustração do funcionamento do algoritmo proposto para eliminação do silêncio de gravações de voz.

Na FIG. 3.1, a "bloco de análise" anda uma janela por vez até o final do sinal, analisando consecutivamente 3 janelas. Quando as 3 janelas estiverem acima do limiar estimado, determina-se o ponto de início da voz, dado pela flecha mais escura, o ponto A da figura. Quando a *caixa de análise* encontra 3 pontos abaixo do limiar, a flecha mais escura indicará o final do trecho de voz, letra B. Assim, serão conhecidos os trechos contendo voz no sinal e, posteriormente, efetua-se o recorte. Este sistema evita o ruído espúrio em janelas anteriores de algum trecho de voz, entretanto, não eliminando o mesmo se estiver contido nas 3 janelas posteriores ao mesmo trecho.

Apesar da simplicidade, o algoritmo proposto mostrou-se eficiente na eliminação do silêncio nas locuções gravadas. Seu inconveniente é necessitar de sinais com alta razão sinal/ruído.

Os passos do algoritmo são:

1. Cálculo da magnitude média do sinal (de todas as janelas de 10ms).
2. Cálculo da média e desvio padrão das 13 janelas (100 ms iniciais e 30 ms finais).
3. Com a média e desvio padrão do passo 2, se define o limiar: limiar = média + 0.5x(desvio padrão).

4. Comparação do valor da magnitude média de 3 janelas consecutivas com o limiar estimado: se estiverem acima, ou abaixo deste, o ponto de início ou final do trecho de voz é determinado, respectivamente.

5. Sabendo-se quais janelas são pertencentes ao início e final do sinal de voz, é efetuada a eliminação dos trechos sem voz.

3.3 OS COEFICIENTES MEL-CEPESTRAIS

Para a explicação do mel-cepestro (DAVIS, 1980) é necessário um breve estudo no campo físico-acústico em que se estuda a percepção auditiva humana.

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

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

$$mel = 2595 \log_{10}(1 + \frac{f}{700}) \tag{3.1}$$

onde f é a freqüência linear em Hz.

A escala mel, portanto, resulta do mapeamento da freqüência percebida de um tom sobre uma escala linear. Este mapeamento é mostrado na FIG. 3.2.

A percepção de uma freqüência particular para o sistema auditivo humano é influenciada pela energia dentro de uma banda crítica centrada em torno da freqüência em questão. Por esse motivo usam-se filtros de banda crítica - filtros passa faixas - para calcular o mel-cepestro. Alguns pesquisadores sugerem usar a log-energia total encontrada dentro das bandas críticas em torno de cada freqüência, em vez de usar a log-magnitude (DELLER, 1993). Além disso, a largura da banda dos filtros varia com a freqüência, começando por volta de $100Hz$ para freqüências abaixo de $1KHz$, e aumentando logaritmicamente acima

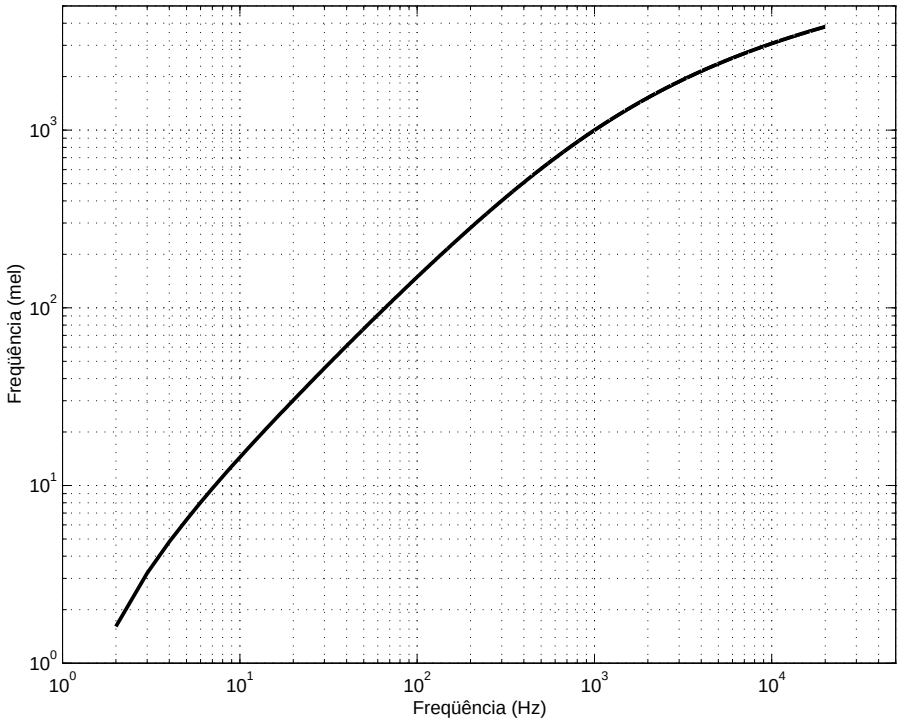

FIG. 3.2: Escala mel versus escala de freqüência normal.

de $1KHz$. Para o cálculo dos coeficientes mel-cepestrais, costuma-se utilizar 20 filtros passa-banda triangulares (DELLER, 1993). A FIG. 3.3 mostra os filtros de banda crítica para o cálculo do mel-cepestro de um sinal amostrado a $8KHz$. Cada filtro é centrado em uma freqüência *mel*, que determina a largura de banda crítica do filtro.

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

$$LB_{crit} = 25 + 75[1 + 1,4(f/1000)^2]^{0,69} \tag{3.2}$$

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

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

$$MCC_i = \sum_{k=1}^{N} X_k \cos\Big[i\Big(k - \frac{1}{2}\Big)\frac{\pi}{N}\Big] \qquad i = 1, 2, \ldots, M \tag{3.3}$$

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

O digrama em blocos que ilustra a extração dos coeficientes mel-cepestrais, é apresentado na FIG 3.4.

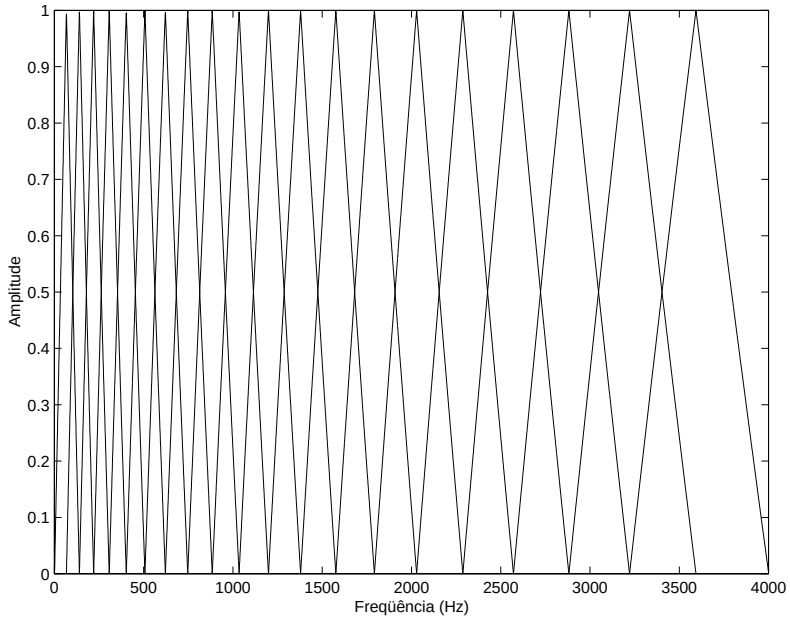

FIG. 3.3: Magnitude do espectro dos filtros de banda crítica utilizados na produção dos coeficientes mel-cepestrais (DAVIS, 1980).

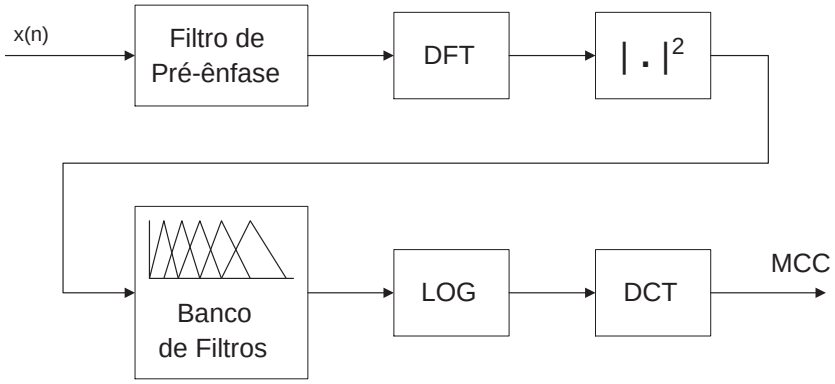

FIG. 3.4: Diagrama em blocos para a extração dos coeficientes mel-cepestrais (REYNOLDS, 1995).

Inicialmente o sinal de voz digitalizado passa por um filtro de pré-ênfase (ver capítulo 2), depois extrai-se o espectro do sinal por meio da transformada discreta de Fourier (*Discrete Fourier Transform* - DFT) ou dos coeficientes de predição linear. Nesta dissertação o espectro foi obtido através da DFT. Em seguida, é calculada a potência espectral, que é filtrada por sua multiplicação por uma série de filtros triangulares espaçados segundo a escala *mel* - escala projetada para simular a resposta de freqüências do ouvido humano. A energia resultante da filtragem é aplicada a uma função logarítmica, e finalmente, é utilizada a transformada cosseno discreta (*Discrete Cosine Transform* - DCT) para se obter os coeficientes no domínio cepestro de freqüência (qüefrência). Essa transformação possui a propriedade de comprimir a informação espectral nos coeficientes de baixa ordem

e também produz uma descorrelação adicional. Tal descorrelação permite a utilização, com menor perda de precisão, de matrizes covariâncias diagonais para o modelamento estatístico da voz.

3.4 A ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais (PCA) é uma técnica de mapeamento linear muito usada em reconhecimento de padrões (MALAYATH, 2000). Esta técnica extrai características de um vetor aleatório a partir de sua projeção sobre um conjunto de vetores base. Consideremos um vetor aleatório X de dimensão N. O conjunto de características Y pode ser extraído de X pela projeção de X sobre o conjunto de vetores bases, da seguinte forma:

$$Y = \Phi^T X \tag{3.4}$$

Na equação acima, Φ é uma matriz, composta dos vetores base $(\phi_1, \phi_2, \ldots, \phi_N)$, que extraem características de X pela combinação linear de seus componentes. Para que cada característica de Y carregue informação única sobre X, os vetores base ϕ_i, $i = 1, \ldots, N$, devem ser linearmente independentes. Pode-se notar que a DCT é um caso especial de mapeamento, onde os vetores base são funções cosseno.

Se os vetores base ϕ_i forem os autovetores da matriz covariância de X, então a resultante extração de características é chamada de *análise de componentes principais* (FUKUNAGA, 1990). Esta é também a versão discreta da transformada de Karhunen-Loéve (*Karhunen-Loéve Transform* - KLT) (GARCIA, 1994).

Os vetores base ϕ_i podem ser usados para representar X em uma dimensão M, menor que a original ($M < N$). Isto é feito utilizando-se os M principais autovetores ϕ_i e desprezando os autovetores com menores autovalores.

As bases derivadas do PCA garantem em uma dimensão M menor, que a representação resultará no menor erro de reconstrução. Desde que os ϕ_i sejam ortogonais entre si[2], X pode ser reconstruído a partir de Y usando a seguinte equação:

$$\hat{X} = \Phi Y \tag{3.5}$$

No PCA os autovetores ϕ_i são organizados de acordo com a grandeza dos autovalores, ou seja, o índice dos autovetores são dados pelos correspondentes autovalores em ordem decrescente ($\upsilon_1 \geq \upsilon_2 \geq \ldots \geq \upsilon_N$) (FUKUNAGA, 1990). O conjunto dos autovetores

[2]Se a matriz covariância é simétrica, os autovetores são sempre ortogonais entre si (STRANG, 1988).

organizados formará a matriz de transformação Φ. Se os autovetores da matriz covariância de X são usados como ϕ_i, então o erro médio quadrático de reconstrução, dado por:

$$\varepsilon = E[||X - \hat{X}||^2] \tag{3.6}$$

é mínimo para qualquer M (NANDAKISHORE, 1996). A transformação efetuada pela matriz Φ, provocará a descorrelação das variáveis aleatórias do vetor transformado Y, o que resultará em uma matriz covariância diagonal.

O PCA projeta os dados transformados na direção de maior variância destes. Isto pode ser visto na FIG. 3.5, que exemplifica uma transformação sobre uma variável bidimensional. Percebe-se que os eixos são rotacionados na direção de maior variância dos dados ((x, y) -> (x', y')). Isto significa que o PCA preserva a direção de máxima variabilidade dos dados (MALAYATH, 2000).

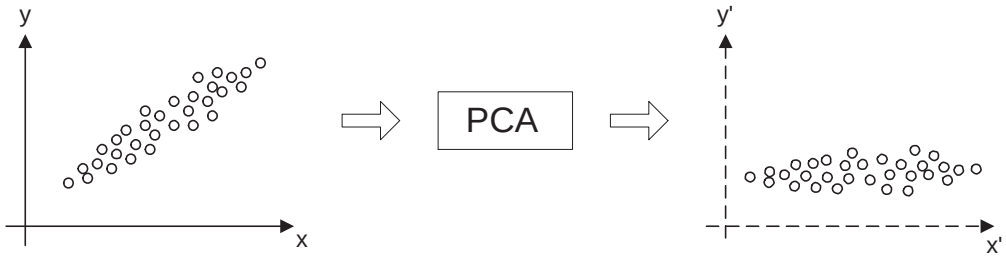

FIG. 3.5: Transformação PCA em uma variável bidimensional.

O uso do PCA possui duas vantagens principais (NANDAKISHORE, 1996):

- A redução de dimensionalidade garante o mínimo erro de reconstrução.
- A matriz covariância das características extraídas será diagonal, resultando em modelos estatísticos mais simples.

Nesta dissertação o objetivo de usar o PCA não é a extração de características, nem a redução de dimensionalidade, mas a transformação linear das características utilizadas no reconhecimento, de modo a gerar matrizes covariância diagonais. Cada locutor tem uma matriz de transformação própria, extraída das suas características (MCC). Portanto, o PCA de cada locutor é dependente da estrutura (autovetores) de sua matriz covariância associada, o que é bem interessante ao sistema de reconhecimento, porque as características serão transformadas de acordo com a informação estrutural da matriz transformação Φ. Portanto, para cada locutor L existirá uma matriz de transformação Φ_L, que atuará sobre

cada um de seus vetores de características. O vetor aleatório X aqui empregado é um vetor de características, mel-cepestrais ($C = \{c_1, c_2, ..., c_N\}$). O processo de transformação é ilustrado na FIG. 3.6.

FIG. 3.6: Transformação PCA sobre um vetor de características mel-cepestrais.

3.5 APLICAÇÃO DO PCA AOS COEFICIENTES MEL-CEPESTRAIS

A aplicação da DCT sobre um conjunto de dados produz uma modificação nesses dados, fazendo para alguns sinais, uma descorrelação parcial dos mesmos (MALAYATH, 2000). A energia resultante do banco de filtros (X_k na FIG. 3.4) sofre tal transformação, resultando nos coeficientes mel-cepestrais[3]. A transformada cosseno é composta por um conjunto fixo de vetores de transformação, que são independentes dos dados. A matriz covariância de um conjunto de vetores com 15 MCCs é apresentada de forma tridimensional na FIG. 3.7.

No referido gráfico os eixos horizontais representam o índice das linhas e colunas da matriz covariância, enquanto o eixo vertical apresenta os valores existentes na mesma. A diagonal principal apresenta os maiores valores, porém existem valores menores em sua volta. Como descorrelação dos dados implica matriz covariância diagonal, o gráfico da FIG. 3.7 demonstra que os dados não foram completamente descorrelacionados pela transformada cosseno.

Neste trabalho o PCA será aplicado sobre os coeficientes mel-cepestrais de cada locutor, idéia semelhante utilizada por LIU (1999), que utilizou somente sons sonoros no reconhecimento de locutor. Cada locutor terá sua matriz de transformação própria, conforme visto anteriormente, resultando em coeficientes mel-cepestrais modificados (MCCPCA). O PCA estimado nos dados a serem transformados provocará a descorrelação dos mesmos.

[3]Na extração do cepestro comum, se utiliza a transformada discreta inversa de Fourier, como X_k é uma seqüência simétrica e par, a IDFT pode ser substituída pela DCT (DELLER, 1993).

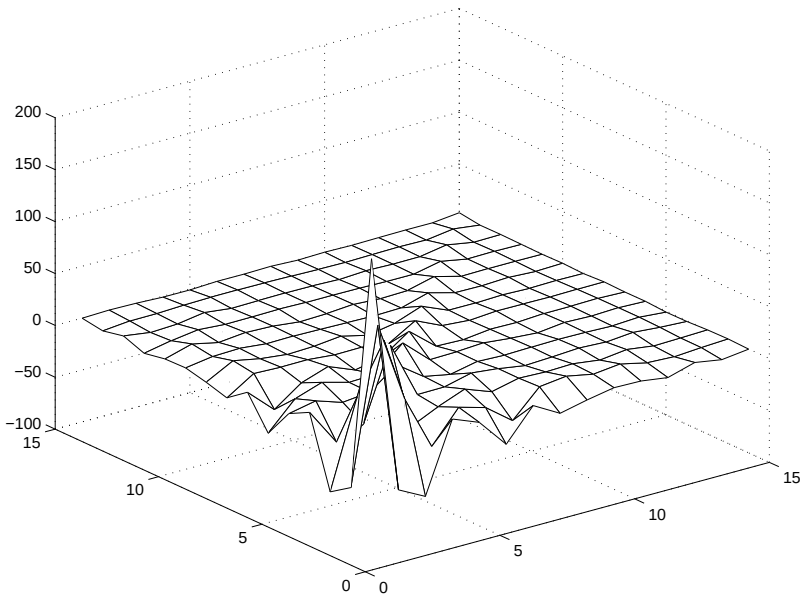

FIG. 3.7: Gráfico tridimensional da matriz covariância de um conjunto de vetores com 15 MCCs.

Isto pode ser visto na FIG. 3.8, que apresenta de forma tridimensional a matriz covariância do MCCPCA para vetores com 15 coeficientes.

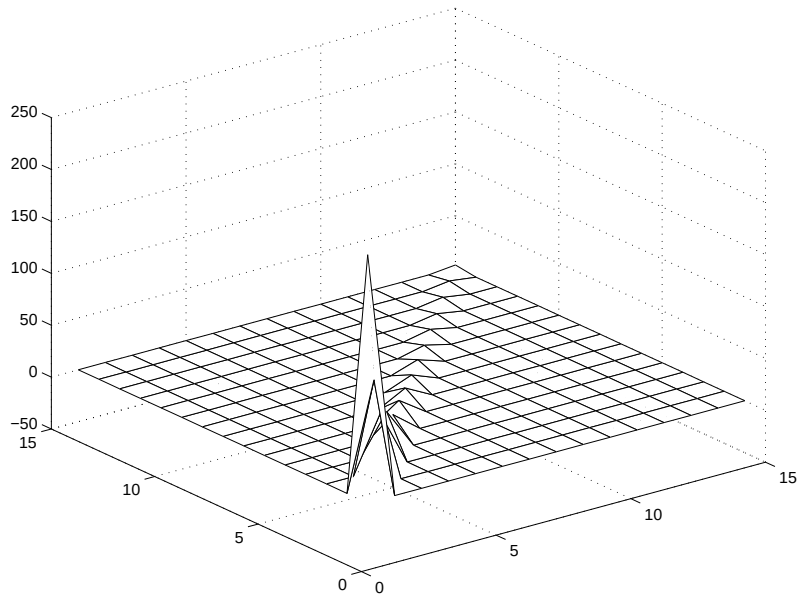

FIG. 3.8: Matriz covariância para 15 MCCPCAs.

Percebe-se claramente que só existem valores na diagonal principal, o que demonstra a descorrelação dos dados transformados. É importante ressaltar que os vetores base do PCA são dependentes dos dados a partir dos quais foram estimados.

A vantagem do PCA em relação à descorrelação dos dados se deve, porque a correla-

ção implica em redundância de informações nos dados. O número real de características requeridas para descrever a informação pode ser bem menor que o número de medidas efetivamente realizadas. Pode-se conseguir algum nível de compressão ou redução selecionando um subconjunto de características (ou combinação linear de características). Na classificação de sinais, a redundância nos dados pode acarretar problemas no processo de aprendizagem; isto ocorre porque o sistema tenta aprender esta redundância ao invés de modelar as características efetivas para a classificação (PICONE, 1991). Dados descorrelacionados tornam mais simples os cálculos estatísticos, como no caso de variáveis conjuntamente gaussianas.

3.6 RESUMO E CONCLUSÃO

Neste capítulo foram apresentadas as características do locutor utilizadas nesta dissertação e a transformação empregada, bem como a forma do processamento inicial realizado sobre as gravações antes da extração das características.

A eliminação do ruído de fundo entre palavras numa gravação é relevante no processo de reconhecimento de locutor aqui usado, pois evita-se a utilização de informação inútil, sem valor agregado à identidade do locutor. Essa informação só aumentaria a carga computacional, podendo mesmo prejudicar o desempenho dos sistemas de classificação. Foi apresentado um algoritmo simples para efetuar tal tarefa.

O uso do MCC (uma característica espectral que leva em conta a percepção auditiva), tem apresentado bons resultados na tarefa de reconhecimento de locutor (REYNOLDS, 1992, 1995, 2000), sendo uma característica muito utilizada para verificação de locutor independente do texto, o que justifica seu uso na verificação de locutor.

O PCA é utilizado principalmente por dois motivos: a descorrelação dos dados e a tentativa de aumentar o poder discriminativo dos MCCs. Este aumento de poder discriminativo estaria associado à dependência da transformada aos dados, sendo um parâmetro próprio de cada locutor. A descorrelação simplifica o uso de modelos de classificação estatísticos, como o modelo de mistura de gaussianas.

4 CLASSIFICADORES

4.1 INTRODUÇÃO

Este capítulo trata dos sistemas de classificação para a verificação de locutor independente do texto utilizados nesta dissertação: um deles é o modelo estatístico GMM, e o outro é um modelo baseado no casamento de padrões característicos, o AR-Vetorial. Na seção 4.2 são apresentados os fundamentos teóricos do modelo de mistura de gaussianas e o seu relacionamento para a verificação de locutor, bem como os requisitos exigidos para que se possa montar um sistema de reconhecimento de locutor. Na seção 4.3, é apresentada a aplicação do modelo autorregressivo vetorial para a mesma tarefa de reconhecimento, iniciando com sua relação ao LPC; também é apresentada uma forma de estimação dos modelos LPC e AR-Vetorial, usando a autocorrelação do sinal. Na seção 4.5, encontram-se o resumo e as conclusões deste capítulo.

4.2 MODELO DE MISTURA DE GAUSSIANAS

O modelo de mistura de gaussianas pode ser visto como um modelo híbrido de dois modelos efetivos para o reconhecimento de locutor: um classificador uni-modal gaussiano e um quantizador vetorial (VQ), combinando a robustez e o amaciamento do modelo gaussiano paramétrico com a modelagem arbitrária de um modelo VQ não-paramétrico. De certa forma, o GMM faz a separação espacial de classes acústicas. A diferença em relação ao VQ está no fato de que não são distâncias que separam as classes, mas sim probabilidades provindas de conjuntos de funções densidade de probabilidade (fdp's) gaussianas estimadas previamente.

O GMM também pode ser entendido como um HMM de um único estado, tendo como observações mistura de fdp's gaussianas. Estas componentes podem modelar um amplo conjunto de classes fonéticas, para caracterizar o som produzido por uma pessoa. Na aproximação VQ, cada locutor é representado por um dicionário de amostras espectrais representando grupos de classes fonéticas. Esta técnica tem demonstrado bom desempenho no reconhecimento de locutor com vocabulários pequenos, como dígitos (DELLER, 1993), sendo limitada em modelar possíveis variações encontradas na verificação de locutor independente do texto. Tem sido mostrado que modelos estatísticos fornecem uma melhor modelagem acústica da voz (RABINER, 1993, SANTOS, 1997). HMMs de várias

formas, têm sido usados para esta modelagem no reconhecimento de locutor dependente e independente do texto. O HMM não modela somente classes acústicas desconhecidas, mas também a seqüência temporal entre essas classes. Embora a modelagem de estruturas temporais seja vantajosa para a tarefa de reconhecimento de locutor dependente do texto, no caso de independência do texto, esta modelagem não apresenta relevância. Por esse motivo, o HMM apresenta limitações no desempenho de tarefas independentes do texto.

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

4.3 O GMM NO RECONHECIMENTO DE LOCUTOR

Uma mistura de densidades de probabilidade gaussianas é uma soma ponderada de M densidades, FIG. 4.1, dada pela equação:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{4.1}$$

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

$$b_i(\vec{x}) = \frac{e^{\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' K_i^{-1}(\vec{x}-\vec{\mu}_i)\right)}}{(2\pi)^{\frac{N}{2}}\sqrt{|K_i|}} \tag{4.2}$$

com vetor média $\vec{\mu}_i$ e matriz de covariância K_i, onde $|.|$ indica determinante. A ponderação das misturas satisfaz à condição $\Sigma_{i=1}^{M} p_i = 1$.

A densidade de mistura gaussiana completa é parametrizada por um vetor de médias, por uma matriz covâriancia e pelos coeficientes da mistura ponderada de todas as densidades componentes (modelo λ). Estes parâmetros são representados coletivamente pela notação:

$$\lambda = \{p_i, \vec{\mu}_i, K_i\} \qquad i = 1, ..., M. \tag{4.3}$$

O GMM pode ter diferentes fdp's dependendo da escolha da matriz covariância. Além disso, o GMM pode ter a matriz covariância distribuída a cada componente gaussiana como indicado na FIG. 4.1, uma matriz covariância para todas as componentes gaussianas

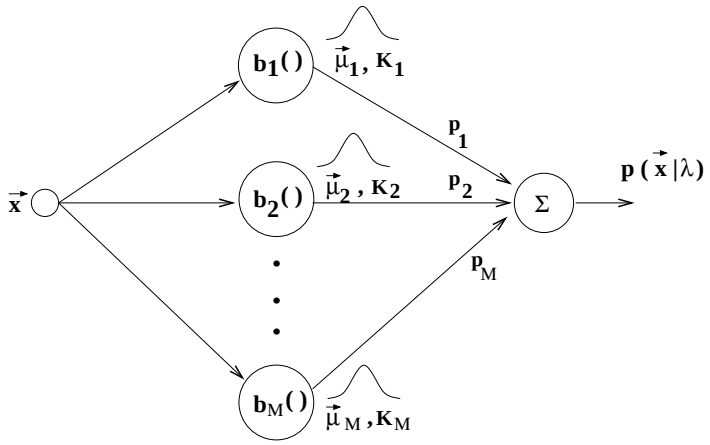

FIG. 4.1: M densidades de probabilidade formando um GMM (REYNOLDS, 1992).

para um dado modelo ou uma única matriz covariância para todos os modelos. A matriz covariância também pode ser completa ou diagonal (REYNOLDS, 1992).

Como as componentes gaussianas agem conjuntamente para modelar um função densidade de probabilidade, uma matriz covariância completa não é necessária, mesmo que os vetores de entrada não sejam estatisticamente independentes. A combinação linear das matrizes covariância diagonais no GMM é capaz de modelar a correlação entre os vetores de dados. O efeito de usar um conjunto de M matrizes covariância completa pode ser igualmente obtido usando-se um conjunto maior de matrizes covariância diagonais (REYNOLDS, 1995).

4.3.1 INTERPRETAÇÕES DO MODELO NO RECONHECIMENTO DE LOCUTOR

Existem dois motivos principais para o reconhecimento de locutor utilizando mistura de densidades gaussianas. O primeiro deles é dado pela noção intuitiva de que as componentes individuais de uma densidade multi-modal, como o GMM, podem modelar conjuntos não distingüíveis de classes acústicas. É razoável supor que o espaço acústico correspondente à voz de um locutor (características extraídas), possa ser caracterizado por um conjunto de classes acústicas representando eventos fonéticos, tais como, sons vogais, nasais, fricativos, etc. Estas classes acústicas refletem a dependência do locutor em relação ao seu trato vocal. A i -ésima classe acústica pode ser representada por sua função densidade de probabilidade, caracterizada pela média $\vec{\mu}_i$ da i -ésima densidade componente e pela matriz covariância K_i. Como as locuções de treinamento e teste não possuem dependência de texto, as classes acústicas ficam *escondidas* porque as observações são desconhecidas. Supondo vetores de características independentes, as densidades de probabilidades aproximadas por estas classes acústicas escondidas formam um conjunto de

gaussianas (mistura).

O segundo motivo para o uso da mistura de densidades gaussianas é a observação empírica de que uma combinação de funções de base gaussiana é capaz de representar uma ampla classe de distribuições de probabilidade (REYNOLDS, 1995). Um das características mais poderosas do GMM é sua capacidade de aproximação para formar densidades de probabilidades desconhecidas, o que pode ser observado pelo exemplo dado na FIG. 4.2. A FIG. 4.2 (a) representa um histograma de um vetor de característica Log-energia, para uma locução com 30 segundos de duração de um locutor masculino. Na FIG. 4.2 (b) é apresentado a modelagem da distribuição dos dados feita por um GMM composto por sete gaussianas, as quais estão abaixo da linha cheia, que é a soma das mesmas, conforme a EQ. 4.1, onde os pesos p_i, da mistura são iguais para todas as gaussianas.

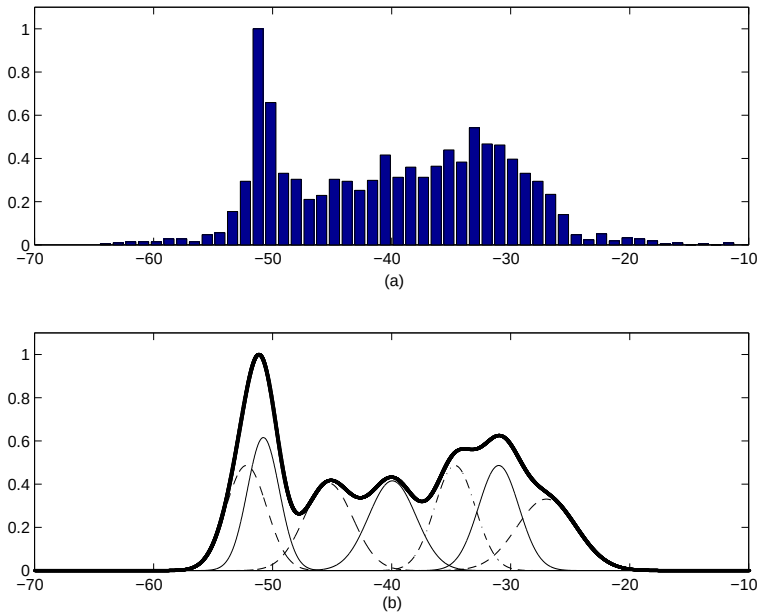

FIG. 4.2: (a) Histograma normalizado de um vetor de característica Log-energia, para uma locução com 30 segundos de duração de um locutor masculino, (b) modelagem da distribuição dos dados feita por um GMM composto por sete gaussianas, as quais estão abaixo da linha cheia.

Resumindo, o GMM representa de forma geral a dependência das características espectrais da voz associadas ao locutor, em conjunto com a capacidade de modelar densidades de probabilidades desconhecidas (REYNOLDS, 1995, VUUREN, 1999). Especificamente: a distribuição dos vetores de características extraídos de uma locução é modelada por uma mistura de densidades de probabilidade gaussianas.

4.3.2 ESTIMAÇÃO DOS PARÂMETROS DE MÁXIMA VEROSSIMILHANÇA

Em um sistema de reconhecimento de locutor, cada locutor é representado por um GMM com seu modelo λ. Na verificação, tal modelo é comparado com uma base formada por outros modelos (*background*). Existem vários métodos de estimação dos parâmetros do GMM (REYNOLDS, 1995). Um método bem difundido e que apresenta bons resultados é a estimação da máxima verossimilhança (*Maximum Likelihood* - ML) (THEODORIDIS, 1999), vide apêndice 1.

Para um conjunto de dados de treinamento, a estimação ML tenta encontrar os parâmetros do modelo que maximizem a verossimilhança do GMM. Para uma seqüência de vetores de características (supondo independência entre os mesmos), $X = \{\vec{x}_1, ..., \vec{x}_T\}$, a verossimilhança do GMM é dada por:

$$p(X|\lambda) = \prod_{i=1}^{T} p(\vec{x}_t|\lambda) \tag{4.4}$$

Normalizando pelo número total de vetores T e usando o logaritmo, chega-se a:

$$\log p(X|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda) \tag{4.5}$$

Infelizmente, esta expressão é uma função não-linear de parâmetros λ e cuja maximização direta não é possível. Entretanto, a estimação dos parâmetros obtidos pelo método ML poderá ser conseguida iterativamente, utilizando-se um caso especial do algoritmo de máxima expectativa (*Expectation Maximization* - EM) (REYNOLDS, 1995, VUUREN, 1999).

A idéia básica do algoritmo EM é a de iniciarmos com um modelo inicial λ para a estimação de um novo modelo $\overline{\lambda}$, tal que $p(X|\overline{\lambda}) \geq p(X|\lambda)$. O novo modelo torna-se, então, o modelo inicial para a próxima iteração, e o processo é repetido até que um limiar de convergência seja alcançado. Esta é a mesma idéia básica para estimação dos parâmetros do HMM através do algoritmo de reestimação de Baum-Welch (RABINER, 1989). Detalhes da estimação dos parâmetros do GMM são dados no apêndice 1.

Em cada iteração do EM, as seguintes fórmulas de reestimação são usadas para a modelagem da i -ésima gaussiana, as quais garantem um crescimento monotônico do modelo de verossimilhança:

Ponderação das misturas:

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} P(i|\vec{x}_t, \lambda) \tag{4.6}$$

Média:

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} P(i|\vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} P(i|\vec{x}_t, \lambda)} \tag{4.7}$$

Variâncias:

$$\overline{\sigma}_{ij}^2 = \frac{\sum_{t=1}^{T} P(i|\vec{x}_t, \lambda)x_{tj}^2}{\sum_{t=1}^{T} P(i|\vec{x}_t, \lambda)} - \overline{\mu}_{ij}^2 \tag{4.8}$$

onde $\overline{\sigma}_{ij}^2$, x_{tj} e $\overline{\mu}_{ij}$, $j = 1, ..., N$, $i = 1, ..., M$ (N é o comprimento do vetor $\vec{x}_t$ e M é o número de gaussianas do GMM) referem-se aos elementos dos vetores $\vec{\overline{\sigma}}_i^{\,2}$, $\vec{x}_t$ e $\vec{\overline{\mu}}_i$ respectivamente, e $P(i|\vec{x}_t, \lambda)$ é a probabilidade *a posteriori* para uma classe acústica i (ver apêndice 1), dada por:

$$P(i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{4.9}$$

Dois fatores críticos no treinamento do GMM para modelagem de um locutor são a seleção do número de gaussianas e a inicialização dos parâmetros *a priori* para o algoritmo EM. Não existem métodos teóricos para a determinação precisa destes parâmetros. A determinação empírica deve adaptar-se à tarefa desejada. No trabalho referido nesta dissertação utilizou-se o algoritmo LBG - *Linde Buzo Gray* (LINDE, 1980) de quantização vetorial, para fornecer o modelo inicial usado no EM, ver apêndice 1. Resultados satisfatórios foram conseguidos com esse algoritmo em (REYNOLDS, 1992).

4.3.3 SISTEMA DE IDENTIFICAÇÃO COM O GMM

O sistema de identificação é um classificador simples de máxima-verossimilhança. Para a identificação de locutor, um grupo de S locutores $S = \{1, 2, ..., S\}$ é representado pelos GMM's $\lambda_1, \lambda_2, ..., \lambda_S$. O objetivo é encontrar o modelo do locutor que tenha a máxima probabilidade *a posteriori* para uma dada seqüência de observações. Matematicamente, temos:

$$\hat{S} = \arg \max_{1 \leq k \leq S} P(\lambda_k|X) = \arg \max_{1 \leq k \leq S} \frac{p(X|\lambda_k)P(\lambda_k)}{p(X)} \tag{4.10}$$

onde a segunda parcela da equação corresponde à regra de Bayes. Supondo igual probabilidade para os locutores ($P(\lambda_k) = 1/S$) e que $p(X)$ é a mesma para todos os locutores, independente de k, a regra de classificação é simplificada para:

$$\hat{S} = \arg \max_{1 \leq k \leq S} p(X|\lambda_k) \tag{4.11}$$

Usando logaritmo e supondo independência entre observações, a regra de decisão para identificação de locutor torna-se:

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^{T} \log p(\vec{x}_t | \lambda_k) \tag{4.12}$$

onde $p(\vec{x}_t|\lambda_k)$ é dado pela EQ. 4.1. O diagrama em blocos de um sistema de identificação de locutor é apresentado na FIG. 4.3.

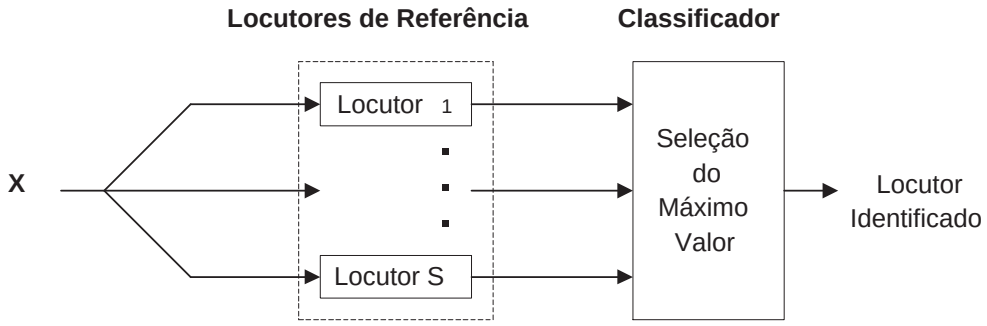

FIG. 4.3: Sistema de identificação de locutor, com S locutores.

4.3.4 SISTEMA DE VERIFICAÇÃO COM O GMM

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

H_0: X pertencer ao locutor.

H_1: X não pertencer ao locutor.

Para desenvolver uma razão de verossimilhança de teste, que decidida entre H_0 e H_1 é usualmente empregado algum modelo do universo de possibilidades falsas, o denominado *background*, que é composto por um conjunto de falsos locutores.

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

$$\frac{P(X \text{ pertence ao locutor})}{P(X \text{ não pertence ao locutor})} = \frac{P(\lambda_L|X)}{P(\lambda_B|X)} \tag{4.13}$$

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

$$\Lambda(X) = \log p(X|\lambda_L) - \log p(X|\lambda_B) \tag{4.14}$$

O termo $p(X|\lambda_L)$ é a verossimilhança da locução do pretenso locutor e $p(X|\lambda_B)$ é a verossimilhança dada por um modelo não pertencente ao mesmo locutor (*background*). A razão de verossimilhança é comparada com um limiar θ e o pretenso locutor é aceito se $\Lambda(X) > \theta$ e rejeitado se $\Lambda(X) \leq \theta$. Este limiar pode ser global, estimado com os dados de todos os locutores (independente ao locutor), usando o resultado de um grande número de testes disponíveis verdadeiros e falsos. O limiar também pode ser dependente do locutor, isto é, cada locutor possui um limiar próprio. Neste caso existe a exigência de uma quantidade maior de informação do locutor (mais tempo de voz), para fornecer um limiar com significado estatístico. Um diagrama em blocos para um sistema de verificação é mostrado na FIG 4.4.

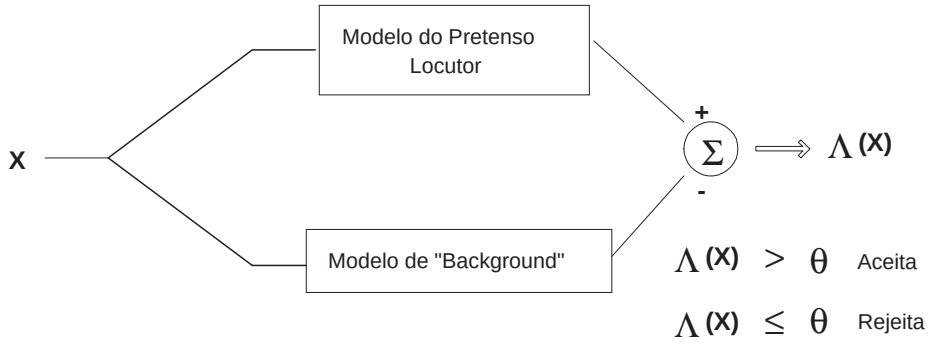

FIG. 4.4: Sistema para verificação de locutor.

A razão de verossimilhança mede essencialmente o quão melhor o locutor de teste se aproxima da modelagem do locutor verdadeiro, comparado com algum modelo falso. A partir disso, o limiar de decisão é ajustado a um limiar que obedeça um compromisso entre a rejeição de locutores verdadeiros (falsa rejeição) e aceitação de locutores falsos (falsa aceitação).

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

$$\log p(X|\lambda_L) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_L) \tag{4.15}$$

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

A verossimilhança das locuções não pertencentes ao locutor verdadeiro é formada usando-se uma coletânea de vozes de locutores falsos, gerando o *background*, que também simula as condições de gravação das locuções, agregando assim, informações do ruído

presente nas mesmas (REYNOLDS, 2000). Esse *background* pode ser montado de duas formas:

- Utilizando vários modelos individuais de locutores (GMMs), escolhidos de acordo com certas regras, para formar o universo de supostos impostores, com características de voz próximas a um determinado locutor verdadeiro (*cohorts*) (REYNOLDS, 1995). Supondo um conjunto com K locutores e seus respectivos modelos $\{\lambda_1, ..., \lambda_K\}$, a verossimilhança em log para os locutores de *background* é calculada da seguinte forma:

$$\log p(X|\lambda_B) = \frac{1}{K} \sum_{b=1}^{K} \log p(X|\lambda_b) \tag{4.16}$$

  onde $\log p(X|\lambda_b)$ é calculado como na EQ. 4.15. Esta é a densidade de probabilidade das locuções dos locutores utilizados para *background*, assumindo igual probabilidade entre os locutores. O fator $1/K$ é utilizado para normalizar a verossimilhança de acordo com o número de modelos utilizados no *background*.

- Utilizando vários locutores em um único modelo. São utilizadas características de vários locutores modelados por um único GMM. Esse é o denominado modelo Universal de *Background* (REYNOLDS, 2000).

A normalização da verossimilhança logarítmica $\Lambda(X)$ provida pelo *background* é importante na tarefa de verificação porque ajuda a diminuir as variações relacionadas com a decisão a ser tomada, permitindo limiares de decisão mais confiáveis.

O valor absoluto da verossimilhança de uma locução é influenciado por muitos fatores dependentes ao locutor, tais como: características do trato vocal, conteúdo lingüístico e qualidade da locução. Estes fatores tornam difícil a determinação de um limiar de decisão para diferentes testes de verificação. A normalização produzida pela razão de verossimilhança produz uma medida mais estável em relação às características do locutor e menos sensível a outras tipos de variações (REYNOLDS, 1995). O *background* ajuda a minimizar informações não relacionadas à identidade do locutor. Na tarefa de identificação não há necessidade desta normalização porque as decisões são tomadas diretamente dos valores de verossimilhança das locuções de teste.

4.3.5 BACKGROUND

Duas questões surgem na determinação do *background*: a determinação dos locutores e sua quantidade. Intuitivamente os locutores para o *background* deveriam ser selecionados

para representar a população de impostores esperados. Em alguns casos, pode-se supor que os impostores que tentarão corromper o sistema possuam voz similar ao locutor verdadeiro, ou ao menos sejam do mesmo sexo (impostores dedicados). Em outros casos, pode existir uma grande quantidade de supostos impostores e alguns deles podem ter voz muito dissimilares à verdadeira (impostores casuais). Neste caso, os impostores podem ser de ambos os sexos.

Muitos sistemas operam com locutores escolhidos de modo a ficarem mais próximos ao verdadeiro, formando um conjunto muito próximo *fechado*, conhecido na literatura por *cohorts* (REYNOLDS, 1995, SARMA, 1999). Isto pode ser apropriado para aplicações onde os impostores são dedicados; mas, como visto em (HIGGINS, 1991), isso deixa o sistema muito vulnerável a impostores que tenham voz com características muito dissimilares à verdadeira. Isto ocorre porque, com esta abordagem, vozes muito dissimilares não são bem modeladas pela razão de verossimilhança (REYNOLDS, 1995). Embora seja possível empregar métodos para rejeitar vozes muito dissimilares (HIGGINS, 1991), a escolha de um *background* adequado é fundamental para um bom desempenho do sistema.

Idealmente o número de locutores de *background*, deve ser o maior possível, para melhor modelar a população de impostores. Neste trabalho, o número de locutores de *background* foi igual a dez, tamanho este motivado pelas considerações de processamento em tempo real, a disponibilidade de um banco de dados pequeno e o desejo de ter um conjunto constante de *background* para os experimentos. Na tarefa de verificação sobre um dado banco de dados, cada locutor é usado como verdadeiro e os demais como impostores, repetindo o processo para todos, sendo os locutores de *background* excluídos dos testes. Isso significa que para um dado número de locutores disponíveis, o aumento do número de locutores de *background* diminui o número de locutores para testes falsos.

Background para o caso de impostores dedicados

Para sistemas suceptíveis a impostores dedicados, a seleção do *background* é realizada utilizando os dados de treinamento do GMM para modelar todos os locutores disponíveis. Para isto, é realizado o cálculo das distâncias duas-a-duas dos modelos. Para os locutores i e j com modelos (λ_i, λ_j) e seqüências de treinamento (X_i, X_j), a distância é definida como:

$$d(\lambda_i, \lambda_j) = \log \frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)} + \log \frac{p(X_j|\lambda_j)}{p(X_j|\lambda_i)} \tag{4.17}$$

A razão $p(X_i|\lambda_i)/p(X_i|\lambda_j)$ mede como o modelo do locutor j se relaciona com a locução do locutor i. Isto é feito em relação ao modelo do locutor i comparado com sua própria

locução. A razão torna-se menor, quanto mais similares forem os modelos. A distância medida é, então, uma combinação simétrica das razões de comparação dos modelos λ_i e λ_j.

Os locutores para análise, próximos ao locutor verdadeiro i, são selecionados para formar um conjunto fechado *cohort*, denominado $\mathcal{C}(i)$. Cada locutor i terá N locutores próximos ($N > B$, onde B é o número final de locutores para o *background*). Do conjunto $\mathcal{C}(i)$, o *background* final de B locutores, denominado $\mathcal{B}(i)$, é selecionado encontrando aqueles $B's$ que são os mais separados uns dos outros (espalhados). Especificamente, o processo é realizado da seguinte forma (REYNOLDS, 1995):

1. Iniciar movendo o locutor mais próximo à $\mathcal{C}(i)$ para $\mathcal{B}(i)$. $N = N - 1$, $B' = 1$ (B' é o número corrente de locutores em $\mathcal{B}(i)$).

2. Mover o locutor c de $\mathcal{C}(i)$ para $\mathcal{B}(i)$, onde c é encontrado por:

$$c = \arg \max_{c \in \mathcal{C}(i)} \left\{ \frac{1}{B'} \sum_{b \in \mathcal{B}(i)} \frac{d(\lambda_b, \lambda_c)}{d(\lambda_i, \lambda_c)} \right\}$$

   $N = N - 1$, $B' = B' + 1$

3. Repetir o passo (2) até que $B' = B$.

Cada locutor verdadeiro i terá seu próprio *background* $\mathcal{B}(i)$. Portanto, o *background* é calculado para cada locutor que se deseja verificar.

A condição de que os $B's$ estejam espalhados é usada para eliminar locutores de *backgrounds* muito similares e, assim, obter uma melhor cobertura, tendo em vista o número limitado de locutores selecionados. Os locutores selecionados para *background* desta forma são denominados: conjunto próximo maximamente espalhado (*Maximally Spread Close* - MSC) (REYNOLDS, 1995).

Background para o caso de impostores casuais

Quando a população de impostores é formada por locutores dissimilares ao verdadeiro (como de outro sexo, por exemplo), a seleção do *background* deve incluir modelos de locutores distantes do modelo verdadeiro, bem como modelos próximos. Portanto, o conjuto de locutores é dividido igualmente entre modelos próximos e distantes. Considerando B locutores para o background, os $B/2$ modelos próximos ao locutor verdadeiro i, são selecionados como no caso de impostores dedicados. Para os $B/2$ locutores que faltam, cada locutor i terá N locutores mais distantes, calculados com a EQ. 4.17, formando o chamado

cohort distante $\mathcal{F}(i)$. O *background* dos $B/2$ locutores maximamente separados é então, selecionado da seguinte forma (REYNOLDS, 1995):

1. Iniciar movendo o locutor mais distante de $\mathcal{F}(i)$ para $\mathcal{B}(i)$. $N = N - 1$, $B' = 1$.

2. Mover o locutor f de $\mathcal{F}(i)$ para $\mathcal{B}(i)$, onde f é encontrado por:

$$f = \arg \max_{f \in \mathcal{F}(i)} \left\{ \frac{1}{B'} \sum_{b \in \mathcal{B}(i)} d(\lambda_b, \lambda_f) \times d(\lambda_i, \lambda_f) \right\}$$

   $N = N - 1$, $B' = B' + 1$

3. Repetir passo (2) até que $B' = B/2$.

O processo é repetido para cada locutor verdadeiro a ser modelado.

Os locutores dissimilares de *background* são denominados conjunto distante maximamente espalhado (*Maximally Spread Far* - MSF) (REYNOLDS, 1995).

Modelo Universal de Background (*Universal Background Model* - UBM)

O modelo universal de *background* é formado por um único GMM que modela o universo independente ao pretenso locutor. O UBM é um grande GMM treinado com um conjunto de caracterísitcas de alguns supostos impostores. Especificamente deseja-se selecionar locutores que reflitam os possíveis impostores que sejam encontrados durante o reconhecimento. No *background*, aplicam-se vozes de vários tipos e qualidades, modelando de certa forma o ruído também. Se as pessoas que serão testadas pelo sistema de reconhecimento forem de um único sexo o *background* deve conter locutores com o mesmo sexo. Sem este conhecimento o UBM deve conter locutores de ambos os sexos, o que aproxima o sistema da realidade prática. O UBM pode ser montado de acordo com o tipo de ruído que será encontrado no sistema de verificação, sempre tentando simular as condições reais de teste.

Não existe nenhuma medida objetiva para determinar o número exato de locutores ou a quantidade de voz necessária para treinar o UBM (REYNOLDS, 2000).

A partir dos dados para treinamento do UBM, existem dois métodos gerais que podem ser utilizados para obtenção do modelo final. O mais simples é agrupar os dados do treinamento e treinar o UBM, conforme ilustra a FIG. 4.5(a). Isto significa agrupar os dados balanceados de sub-populações, como de homens e mulheres, para evitar que o modelo final fique melhor modelado com uma das sub-populações, e portanto, tendencioso. Outro método é treinar UBMs individuais para cada sub-população e, então, agrupar os

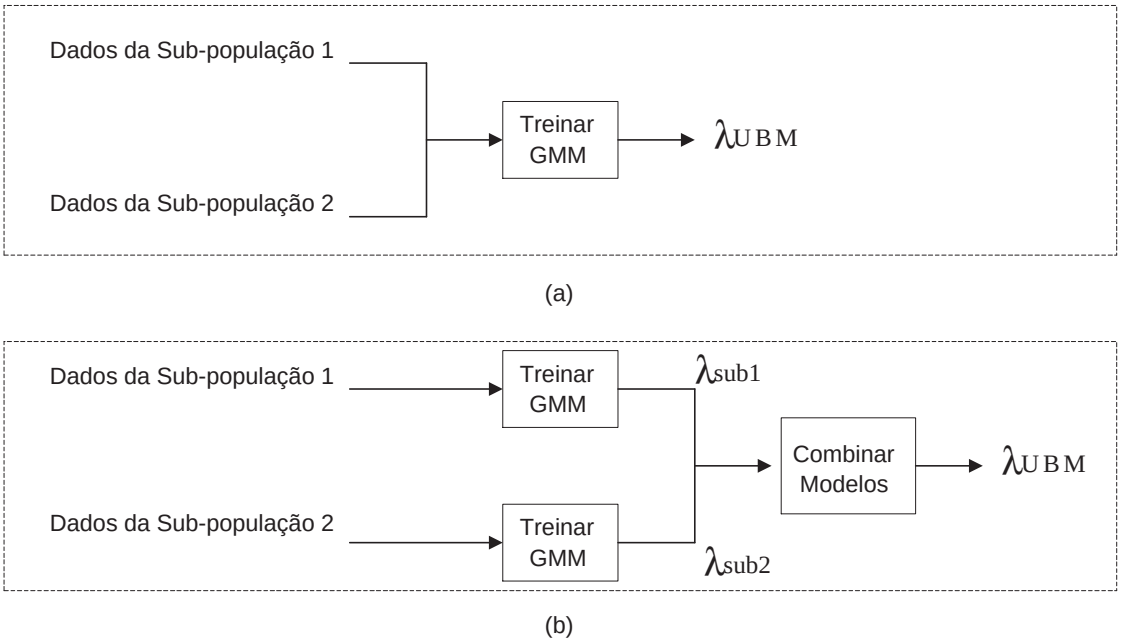

FIG. 4.5: Métodos mais comuns para criar o UBM. (a) Dados das subpopulações são agrupados antes do treinamento. (b) Modelos individuais de subpopulações são treinados e, então, combinados para criar um UBM final.

modelos para formar um único, como mostra a FIG. 4.5(b) para duas subpopulações. Este método tem a vantagem de poder usar dados desbalanceados e a composição final do UBM poder ser controlada (REYNOLDS, 2000), ou seja, pode-se utilizar subpopulações modeladas com números de gaussianas e tempo de treinamento diferentes.

O UBM é o modelo *background* mais utilizado atualmente pelos pesquisadores, devido aos bons resultados, sua proximidade com o mundo real e simplicidade. Em MARTIN (2000) tem-se empregado em torno de 1 hora de voz de cada sexo para treinar UBMs independentes de 1024 gaussianas, agrupados para formarem um único de 2048 gaussianas, utilizando em torno de 200 locutores.

No capítulo 5 a necessidade do *background* para o GMM é avaliada, bem como, é apresentada uma pequena análise do aumento do tempo de treinamento para o UBM.

4.4 MODELO AUTORREGRESSIVO VETORIAL

O modelo autorregressivo vetorial (AR-Vetorial) faz a modelagem da evolução espectral da voz, sendo uma generalização de um modelo muito comum na análise da voz, o modelo de predição linear (LPC). O AR-Vetorial é estimado a partir de uma sequência de vetores extraídos do sinal (geralmente vetores de características), enquanto que o LPC é estimado a partir de escalares.

No reconhecimento de locutor, o AR-Vetorial é utilizado para medir a similaridade entre modelos de locutores previamente estimados. Seu uso é motivado por extrair de forma aproximada, as características dinâmicas do locutor, ou seja, a forma como fala com o passar do tempo. Assim, o AR-Vetorial modela a velocidade média e aceleração da fala (MONTACIÉ, 1992), ao contrário das modelagens estáticas, baseadas na distribuição estatística dos dados como é o caso do GMM.

4.4.1 RELAÇÃO ENTRE O LPC E O AR-VETORIAL

O sons da voz podem ser classificados em duas grandes classes distintas: sonoros e não-sonoros ou surdos (LIMA, 1994). A voz é uma onda acústica de pressão que se origina do movimento fisiológico voluntário de estruturas anatômicas, tais como, as cordas vocais, trato vocal, cavidade nasal, língua e lábios (RABINER, 1978). O trato vocal pode ser modelado como uma concatenação de tubos não uniformes sem perdas, com diâmetros variáveis que começam nas cordas vocais e terminam nos lábios (DELLER, 1993). Sons sonoros tais como /i/ e /e/ são produzidos forçando o ar atráves da glote (abertura das cordas vocais) com a tensão das cordas vocais oscilando entre vibração e relaxamento, excitando o trato vocal com pulsos de ar quase-periódicos. Quanto maior a tensão das cordas vocais, maior a *pitch* ou freqüência fundamental da voz. Os sons surdos são gerados mantendo as cordas vocais abertas, formando um constrição usando os articuladores, e forçando o ar através da constrição a uma velocidade alta suficiente para produzir turbulência. O trato vocal é excitado por um ruído de banda larga (ruído branco) durante a produção dos sons surdos (RABINER, 1978).

Um modelo linear de produção da fala foi desenvolvido por Fant no final dos anos 50 (FANT, 1960), onde o pulso glotal, trato vocal, e radiação são individualmente modelados por um filtro linear. Um modelo completo da produção da voz é apresentado na FIG. 4.6.

A fonte é um sequência de impulsos quase periódicos para sons sonoros e um seqüência de ruído aleatório para os sons surdos com um fator de ganho G para controlar a intensidade da excitação. A função de transferência $V(z)$ para o trato vocal, relaciona o volume de ar e velocidade da fonte com o volume de ar e velocidade nos lábios (MAMMONE, 1996). $V(z)$ é modelado geralmente por um modelo só de pólos, para a maioria dos sons (MAKHOUL, 1975). Cada pólo de $V(z)$ corresponde a uma freqüência formante ou ressonante do som produzido. Para sons que requerem ambas freqüências, ressonantes e antiressonantes (pólos e zeros), um modelo só de pólos ainda pode ser utilizado, porque o efeito de um zero na função de transferência pode ser conseguido incluindo mais pólos

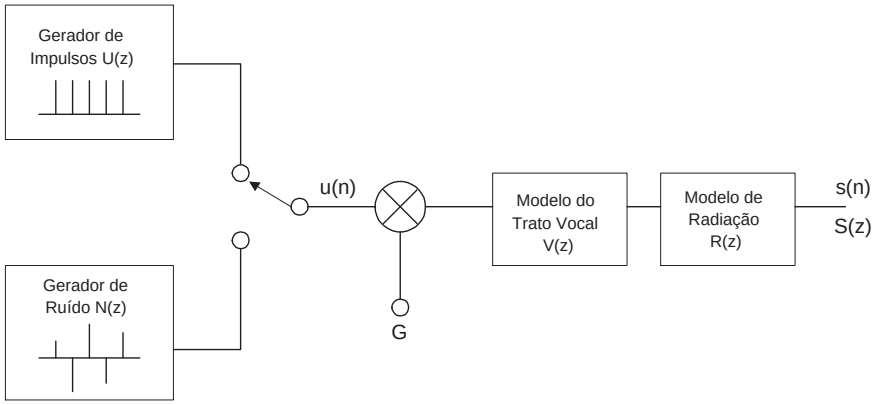

FIG. 4.6: Modelo linear do trato vocal para a produção da voz.

no modelo (DELLER, 1993). O modelo de radiação $R(z)$ descreve a pressão de ar nos lábios, podendo ser razoavelmente aproximado uma equação recursiva de primeira ordem (MAMMONE, 1996).

Combinando o pulso glotal, trato vocal, e radiação, resulta em uma simples função de transferência dada por (MAKHOUL, 1975):

$$H(z) = GV(z)R(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.18}$$

Com esta função de transferência, obtém-se a equação diferença para sintetizar as amostras de voz $s(n)$. De acordo com a notação empregada na FIG. 4.6, tem-se então:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{4.19}$$

A idéia da predição linear é a de que uma amostra voz pode ser aproximada por uma combinação linear de p amostras passadas. Por isso o termo modelo de predição linear ou modelo autorregressivo. Os coeficientes de predição linear (LPC) são os a_i das equações acima, de modo que a ordem do modelo p será dada pelo número de coeficientes utilizados.

O LPC é calculado com as amostras de uma seqüência numérica (trechos de voz). O AR-Vetorial é uma extensão do LPC no sentido de que realiza a predição entre vetores e não entre amostras, modelando a evolução dos vetores com o passar do tempo.

A notação utilizada nas equações seguintes visa facilitar a compreensão matemática existente entre o LPC e o AR-Vetorial. A seguir, será apresentado o método de autocorrelação para o cálculo dos coeficientes de ambos os modelos.

- LPC (MAMMONE, 1996):
  - Equação geral no domínio do tempo:

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} + e_n \tag{4.20}$$

onde x_n e e_n são valores numéricos da seqüência modelada, com e_n representando o erro de predição linear (excitação do modelo). O conjunto de coeficientes de predição linear é definido por: $\mathbf{a} = [a_0 \quad a_1 \quad a_2 ... a_p]$, com $a_0 = 1$.

  - Autocorrelação do sinal x_n:

$$r_k = \sum_{n=0}^{N-k} x_n x_{n+k} \tag{4.21}$$

onde N é o número de amostras da seqüência numérica modelada.

  - Cálcula-se os a_k resolvendo o seguinte conjunto de equações:

$$\begin{pmatrix} r_0 & r_1 & \dots & r_{p-1} \\ r_1 & r_0 & \dots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \dots & r_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{pmatrix} \tag{4.22}$$

Denominando a matriz de autocorrelação Toeplitz no lado esquerdo de R, o vetor de coeficientes por $\mathbf{a}$ e o vetor de autocorrelação da direita por $\mathbf{r}$, tem-se:

$$R\mathbf{a} = \mathbf{r} \tag{4.23}$$

Portanto:

$$\mathbf{a} = R^{-1}\mathbf{r} \tag{4.24}$$

- AR-Vetorial (BIMBOT, 1992):
  - De maneira similar ao LPC, um modelo AR-Vetorial de ordem p, para uma seqüência de N vetores de dimensão m, no domínio do tempo é dado por:

$$X_n = \sum_{k=1}^{p} A_k X_{n-k} + E_n \tag{4.25}$$

onde X_n e E_n são vetores de dimensão m, com E_n representando o erro de predição linear, e A_k é uma matriz de predição, de dimensão $(m \times m)$. O conjundo de matrizes de predição pode ser representado por uma matriz $\mathbf{A} = [A_0 \quad A_1 \quad A_2 ... A_p]$ de dimensão m x $(p+1)m$, com $A_0 = I$ (matriz identidade).

– Matriz autocorrelação dos vetores X_n:

$$R_k = \sum_{n=0}^{N-k} X_n X_{n+k}^T \tag{4.26}$$

onde N é o número de vetores X. R_k resulta numa matriz quadrada (m x m).

– Cálcula-se os A_k resolvendo o seguinte conjunto de equações:

$$\begin{pmatrix} R_0 & R_1^T & \dots & R_{p-1}^T \\ R_1 & R_0 & \dots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \dots & R_0 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \end{pmatrix} = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{pmatrix} \tag{4.27}$$

Denominando a matriz de autocorrelação Toeplitz no lado esquerdo de **R**, a matriz de coeficientes por **A** e a matriz de autocorrelação da direita por R, tem-se:

$$\mathbf{RA} = \mathrm{R} \tag{4.28}$$

Portanto:

$$\mathbf{A} = \mathbf{R}^{-1}\mathrm{R} \tag{4.29}$$

Como **R** é uma matriz Toeplitz, um algoritmo computacionalmente eficiente, conhecido como recursão de Levinson-Durbin pode ser utilizado para resolver os sistemas de equações (HAYKIN, 1996). No apêndice 2 são apresentados os algoritmos para os cômputos do LPC e do AR-Vetorial.

4.4.2 MEDIDAS USADAS NO AR-VETORIAL PARA O RECONHECIMENTO DE LOCUTOR

A utilização do AR-Vetorial no reconhecimento de locutor exige o emprego de alguma medida para avaliar a similaridade entre dois modelos autorregressivos. Medidas de distância sensíveis a variações entre modelos são, portanto, necessárias. Uma medida muito utilizada é a distância de Itakura (ITAKURA, 1975), a qual fornece a distância entre dois modelos de LPCs só de pólos, com base nos coeficientes de predição linear e matriz de autocorrelção.

Supondo dois modelos de LPCs, compostos dos coeficentes **a** e **b** respectivamente, a distância de **a** para **b** , com suas respectivas matrizes de autocorrelação R_a e R_b, é dada por[1]:

[1] tr = traço da matriz

$$d(\mathbf{a},\mathbf{b}) = \log \frac{\mathbf{b}R_a\mathbf{b}^T}{\mathbf{a}R_a\mathbf{a}^T} \tag{4.30}$$

A distância de Itakura não é simétrica portando a distância de **a** para **b** não é igual a **b** para **a**, assim:

$$\log \frac{\mathbf{b}R_a\mathbf{b}^T}{\mathbf{a}R_a\mathbf{a}^T} \neq \log \frac{\mathbf{a}R_b\mathbf{a}^T}{\mathbf{b}R_b\mathbf{b}^T} \tag{4.31}$$

A aplicação da distância de Itakura para o AR-Vetorial é apresentada em (BIMBOT, 1992). Supondo um modelo armazenado **A** (previamente estimado de um locutor), e um modelo **B** de um pretenso locutor, são definidas sete medidas de distância, entre os referidos modelos, com suas respectivas matrizes de autocorrelação $\mathbf{R_A}$ e $\mathbf{R_B}$. Estas medidas são:

1. Distância de **B** para **A**:

$$d(\mathbf{B},\mathbf{A}) = \log(\mathrm{tr}\Big[\frac{\mathbf{A}\mathbf{R_B}\mathbf{A}^T}{\mathbf{B}\mathbf{R_B}\mathbf{B}^T}\Big]) \tag{4.32}$$

2. Distância de **A** para **B**:

$$d(\mathbf{A},\mathbf{B}) = \log(\mathrm{tr}\Big[\frac{\mathbf{B}\mathbf{R_A}\mathbf{B}^T}{\mathbf{A}\mathbf{R_A}\mathbf{A}^T}\Big]) \tag{4.33}$$

3. Distância reversa (ver apêndice 2) de $\hat{\mathbf{B}}$ para $\hat{\mathbf{A}}$:

$$d(\hat{\mathbf{B}},\hat{\mathbf{A}}) = \log(\mathrm{tr}\Big[\frac{\hat{\mathbf{A}}\mathbf{R_B}\hat{\mathbf{A}}^T}{\hat{\mathbf{B}}\mathbf{R_B}\hat{\mathbf{B}}^T}\Big]) \tag{4.34}$$

4. Distância reversa de $\hat{\mathbf{A}}$ para $\hat{\mathbf{B}}$:

$$d(\hat{\mathbf{A}},\hat{\mathbf{B}}) = \log(\mathrm{tr}\Big[\frac{\hat{\mathbf{B}}\mathbf{R_A}\hat{\mathbf{B}}^T}{\hat{\mathbf{A}}\mathbf{R_A}\hat{\mathbf{A}}^T}\Big]) \tag{4.35}$$

5. Distância simétrica:

$$d_{\mathrm{sim}} = \frac{1}{2}(d(\mathbf{B},\mathbf{A}) + d(\mathbf{A},\mathbf{B})) \tag{4.36}$$

6. Distância reversa simétrica:

$$\hat{d}_{\mathrm{sim}} = \frac{1}{2}(d(\hat{\mathbf{B}},\hat{\mathbf{A}}) + d(\hat{\mathbf{A}},\hat{\mathbf{B}})) \tag{4.37}$$

7. Distância mista:

$$d_{\mathrm{mis}} = \frac{1}{2}(d_{\mathrm{sim}} + \hat{d}_{\mathrm{sim}}) \tag{4.38}$$

4.4.3 SISTEMA DE IDENTIFICAÇÃO COM O AR-VETORIAL

Para realizar a identificação de um locutor, o modelo para identificação é comparado com os modelos do conjunto de locutores de referência. A distância de Itakura é utilizada e o modelo que apresentar menor distância em relação ao modelo para identificação é reconhecido, como ilustra a FIG. 4.7. Os modelos $\mathbf{A}_S$ são armazenados; quando se deseja identificar um locutor desconhecido, seu modelo $\mathbf{B}$ é estimado e a distância de Itakura indicará a similaridade entre os modelos. Escolhe-se, então, o modelo mais próximo a $\mathbf{B}$, o qual é identificado.

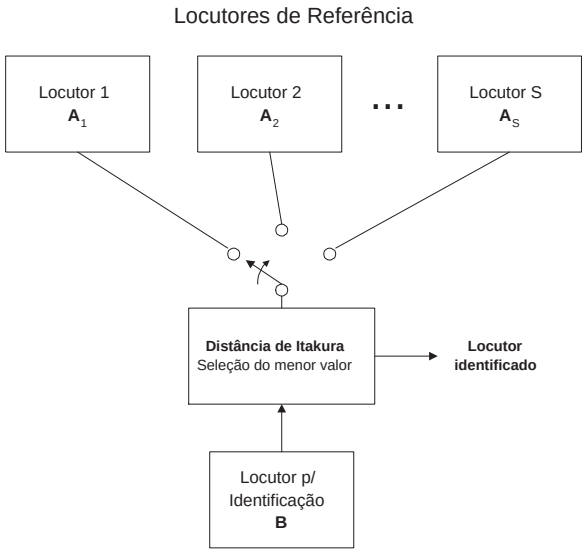

FIG. 4.7: Sistema de identificação de locutor com o AR-Vetorial.

4.4.4 SISTEMA DE VERIFICAÇÃO COM O AR-VETORIAL

O sistema de verificação fornece uma resposta binária, aceita ou rejeita um pretenso locutor. Isto é contrário ao sistema de identificação, no qual existe a necessidade de se estimar um limiar de aceitação θ com base em locuções verdadeiras e falsas. Este limiar é estimado com as distâncias verdadeiras, os dois modelos sob comparação são da mesma pessoa - e com as distâncias falsas - dadas pelo modelo do pretenso locutor comparado com outros modelos não pertencentes a ele. A partir destas distâncias, o limiar é estimado levando em conta os erros de falsa aceitação e falsa rejeição. Quando um locutor for analisado, ele será aceito se a distância resultante for menor que o limiar, e rejeitado caso contrário. A FIG. 4.8 apresenta o sistema de verificação utilizando o AR-Vetorial. No capítulo 5 será discutido o problema da utilização de *background* dentro desse contexto assim como a escolha do limiar de decisão.

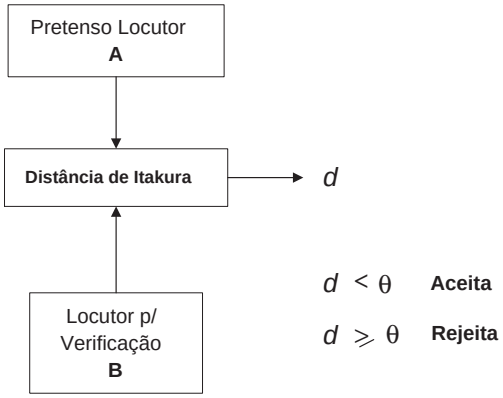

FIG. 4.8: Sistema de verificação de locutor com AR-Vetorial.

4.5 RESUMO E CONCLUSÃO

O modelo de mistura gaussiano, devido às suas propriedades de modelar conjuntos distintos de classes acústicas e à sua capacidade de representar de forma suave distribuições de probabilidade, é atualmente o sistema de classificação mais utilizado no reconhecimento de locutor independente do texto. Esse sistema fornece um desempenho muito bom quando o sinal de voz apresenta uma boa razão sinal/ruído. Quando o sinal de voz sofre influência dos mais diversos tipos de ruído, o desempenho do GMM cai consideravelmente. Existem pesquisas para adaptação dos modelos do locutor e *background*, de acordo com diferentes tipos de ruídos (REYNOLDS, 2000), resultando em sistemas com melhor desempenho nesses ambientes.

O modelo autorregressivo vetorial gera um modelo amaciado da evolução das características, capturando informações da dinâmica da fala do locutor. O AR-Vetorial é um modelo simples que utiliza uma medida de distância (Itakura) para avaliar a similaridade entre dois modelos. É um modelo que apresenta certa robustez ao ruído (CHAGNOLLEU, 1996) em sistemas de identificação de locutor. Entretanto, em sistemas de verificação essa característica ainda necessita ser investigada.

O GMM é um modelo estatístico estático, que leva em consideração a estatística do sinal e não a ordem como os dados estão distribuídos, ao contrário do AR-Vetorial, cujo modelo é estimado de acordo com a evolução dinâmica do sinal. Os dois modelos trabalham, portanto, de forma distinta.

Neste capítulo foram apresentados os sistemas de classificação. No próximo capítulo, seus desempenhos serão avaliados com o uso do MCC e do MCCPCA.

5 AVALIAÇÃO DE DESEMPENHO E ANÁLISE COMPARATIVA

5.1 INTRODUÇÃO

Neste capítulo é examinado o desempenho dos sistemas de classificação usados para a tarefa de verificação de locutor independente do texto. Na seção 5.2 é descrito como foram realizadas as gravações de voz, e como estas gravações foram processadas. A seção 5.3 traz a medida de erro utilizada e seu significado na avaliação do desempenho dos sistemas de verificação de locutor. O primeiro sistema a ser avaliado é o que faz uso do GMM. Seus resultados são apresentados na seção 5.4, onde se mostra como determinar o limiar de decisão, define-se o número de coeficientes mel-cepestrais, avalia-se a variação do número de gaussianas, tempo de treinamento e teste, com o uso do MCC e MCCPCA, e por final é realizada um análise da variação do número de locutores no *background*. Na seção 5.5 o desempenho do AR-Vetorial é avaliado. Nesta seção descreve-se como foi determinado o limiar de decisão. Em seguida, escolhe-se a melhor medida de distância, para posterior avaliação da ordem do modelo com variações do tempo de treinamento e teste. É feita uma comparação entre o uso do MCC12 e MCC15 no desempenho do AR-Vetorial para determinar a importância do número de coeficientes MCC no sistema, como realizado para o GMM e, por último, avalia-se o uso do MCCPCA no desempenho do sistema. Comparações entre os resultados do GMM e AR-Vetorial são fornecidas e discutidas na seção 5.6; comparações para seus tempos de processamento são vistos na seção 5.7. Na seção 5.8, é apresentada uma outra maneira de avaliação dos sistemas de verificação de locutor, a curva DET. Por último, na seção 5.9 faz-se o resumo e conclusão deste capítulo.

5.2 DESCRIÇÃO DAS SIMULAÇÕES

Para a obtenção dos resultados de desempenho dos sistemas de verificação de locutor independente do texto, foram utilizadas 200 frases (20 listas de 10 frases) fonéticamente balanceadas para o português falado na cidade do Rio de Janeiro, extraídas de ALCAIM (1992). Utilizou-se 36 locutores, 23 masculinos e 13 femininos, cada um deles falando as 200 frases. Estas locuções foram gravadas a uma taxa de $22{,}05KHz$ com 16 bits. Para a extração dos coeficientes mel-cepestrais, as locuções sofreram um rebaixamento da freqüência de amostragem (fs), com a filtragem correta para evitar *aliasing*, chegando-se a $8KHz$. Esse é o valor mais empregado pela comunidade científica para

pesquisas no reconhecimento de locutor, tendo em vista que esta é a freqüência de amostragem utilizada pelos sistemas telefônicos em geral.

Com os sinais de voz a $8KHz$, foi efetuado o recorte do silêncio existende nas gravações, e a filtragem de pré-ênfase (com $a_{pre} = -0,95$, de acordo com a EQ. 2.2). Após este pré-processamento, extraiu-se 15 coeficientes mel-cepestrais, fazendo uso de 20 filtros triangulares espaçados segundo a escala mel. O tamanho da janela foi de 20ms com sobreposição de 50%, como utilizado por REYNOLDS (1995). Com a massa de dados processada, foi feita a separação entre dados de treinamento e testes.

Os sistemas de classificação utilizados, o GMM e o AR-Vetorial, foram treinados com 60, 30 e 10s de voz processada, sendo testados com 30, 10 e 3s. Com as gravações disponíveis, conseguiu-se por locutor: 10 testes de 30s, 30 testes de 10s e 100 testes de 3s, correspondendo a um total de 260, 780 e 2600 testes, respectivamente.

Os dados de treinamento não participaram dos testes e os testes verdadeiros utilizaram locuções pertencentes ao pretenso locutor, ao contrário dos testes falsos. Quando o tempo de treinamento foi o menor (10s), os testes foram com 10 e 3s, porque não há sentido em testar um sistema com mais tempo do que foi usado no seu treinamento.

Como existe uma dissimilaridade razoável entre locuções masculinas e femininas, o uso de testes cruzados com locuções de ambos os sexos, deve melhorar o desempenho dos sistemas de classificação. Para evitar este mascaramento, testes cruzados com locutores de sexos diferentes não foram realizados.

Para o GMM foram utilizados 10 locutores, 5 masculinos e 5 femininos, escolhidos aleatoriamente no conjunto total de locutores, para formarem um modelo universal de *background*, cada um participando com 6s de voz processada, para a obtenção, portanto, de 60s de treinamento. Devido à modelagem destes locutores no *background*, eles foram excluídos dos testes. Restaram, assim, 26 locutores para testar o sistema. Utilizou-se 32, 16 e 8 gaussianas para avaliar o desempenho do GMM.

O AR-Vetorial foi avaliado para os modelos de ordem 1, 2, 3, 4 e 5. Para posterior comparação entre os sistemas de classificação, foram utilizados os mesmos números de locutores de teste do GMM, no AR-Vetorial. Isto é feito, porque o AR-Vetorial não faz uso do modelo de *background*.

Ambos sistemas de classificação foram testados com os conjuntos de características MCC e MCCPCA. Foi feita uma avaliação com o uso de 12 coeficeintes mel-cepestrais nos dois sistemas de classificação, e uma compressão de 3 coeficientes pelo PCA para o GMM (MCCPCA1512). Os resultados visam determinar a importância do número de coeficientes mel-cepestrais no sistema de reconhecimento.

A importância do *background* também é abordada e o resultado do seu uso é apresentado para ambos os sistemas de classificação. Foi realizado, também, um teste com a expansão do número de locutores de *background* para 16 locutores, apresentado na seção 5.4.4, deixando precedentes para pesquisas futuras.

5.3 MEDIDA DE ERRO

Como mencionado no capítulo 2, para um sistema de verificação de locutor podem ocorrer dois tipos de erro: o erro E_{FR} de falsa rejeição (FR), onde o sistema rejeita o locutor verdadeiro, e o erro E_{FA} de falsa aceitação (FA), no qual um locutor falso é aceito. Os limiares nos sistemas de verificação de locutor, foram estimados visando obter iguais erros de FR e FA, ou seja, foi dada a mesma importância a esses erros na avaliação dos sistemas.

Para facilitar a avaliação do desempenho dos sistemas de classificação foi utilizada uma medida simples, ponderando-se igualmente os dois tipos de erros, que geralmente não foram iguais devido aos procedimentos práticos. Esta medida é dada por:

$$E = (E_{FR} + E_{FA})/2 \tag{5.1}$$

ou seja, a média entre os erros de FR e FA.

Existe uma outra medida do erro conjunto de FR e FA, surgida com o NIST, que pondera diferentemente os erros, de acordo com a exigência do projeto. Supondo um bom projeto do sistema de verificação, teria-se então:

- $P(FR|H0)$: A probabilidade de rejeitar um locutor verdadeiro, dado que a locução de teste pertence ao pretenso locutor.
- $P(FA|H1)$: A probabilidade de aceitar um locutor falso, dado que a locução de teste não pertence ao pretenso locutor.

Utilizando valores de custo C_{FR} e C_{FA} para ambos os erros , define-se uma medidade de erro total, dada por (NIST, 2000):

$$E = E_{FR} + E_{FA} = C_{FR}\ P(FR|H0)P(H0) + C_{FA}\ P(FA|H1)P(H1) \tag{5.2}$$

onde $P(H0)$ e $P(H1)$ são as probabilidades de ocorrerem resultados verdadeiros e falsos, respectivamente.

Os custos utilizados na EQ. 5.2 servem para avaliar o desempenho do sistema de verificação no caso de um tipo de erro ser mais prejudicial que o outro. Por exemplo, em

um sistema de acesso privado por voz, o erro de FA é muito mais prejudicial que o de FR. Portanto, o valor de C_{FA} seria maior que o de C_{FR}, e a medida de erro seria mais rígida na aceitação de locutores falsos.

Ponderando igualmente a probabilidade de falsa rejeição $P(FR|H0)$ e falsa aceitação $P(FA|H1)$ com custo unitário para ambos $C_{FR} = C_{FA} = 1$, com probabilidades iguais de ocorrerem testes verdadeiros e falsos $P(H0) = P(H1) = 0,5$, chega-se a taxa igual de erro, conhecida por EER (VUUREN, 1999).

Nesta dissertação, o erro definido na EQ. 5.1 é uma aproximação do EER pois a probabilidade de ocorrerem testes verdadeiros e falsos é a mesma (foram utilizados as mesmas quantidades de testes verdadeiros e falsos). O limiar de decisão foi escolhido para ponderar igualmente os dois tipos de erro, ou seja, $C_{FR} = C_{FA} = 1$. Nesse sentido, os sistemas de classificação trabalharam na região do EER.

5.4 AVALIAÇÃO DO GMM

5.4.1 DEFINIÇÃO DO LIMIAR DE DECISÃO

Conforme ilustrado na FIG. 4.4, os resultados do GMM dependem diretamento do limiar de decisão escolhido (θ), o qual é baseado na subtração da verossimilhança do modelo de pretenso locutor e do modelo de *background*, EQ. 4.14. Testes realizados comprovaram a importância do *background* na escolha do limiar de decisão. A FIG. 5.1 apresenta os resultados para o GMM utilizando 32 gaussianas treinado com 60s e testado com 30s, com e sem o uso de *background*. Nos dois gráficos aparecem duas linhas, a linha cheia indica os resultados dos testes com locutores verdadeiros e a linha fina indica os resultados dos testes com locutores falsos. O eixo horizontal representa o número de testes efetuados, 260 testes falsos e verdadeiros. Os resultados destes testes são dados em verossimilhança logarítmica normalizada, conforme a EQ. 4.5. Quanto mais afastadas estiverem as duas linhas, ou seja, a verossimilhança dos testes verdadeiros em relação a verossimilhança dos testes falsos, melhor será o desempenho do GMM pois a escolha do limiar de decisão implicará em menos erros na decisão entre locutores verdadeiros e falsos.

Na prática espera-se que quando o sistema seja testado, os resultados estejam compreendidos na faixa dada pelas linhas de resultados dos testes verdadeiros e falsos e o limiar de decisão consiga validar corretamente o locutor. Observa-se no gráfico da FIG. 5.1 (a), em que se utiliza o *background*, que existe uma separação clara entre os resultados dos testes verdadeiros e falsos. Desta forma, um limiar entre as duas curvas as separaria de forma adequada. Já o gráfico da FIG. 5.1 (b), onde o *background* não é usado, mostra

uma mistura acentuada entre os resultados dos testes verdadeiros e falsos, nos quais um limiar de decisão não apresentaria resultados satisfatórios.

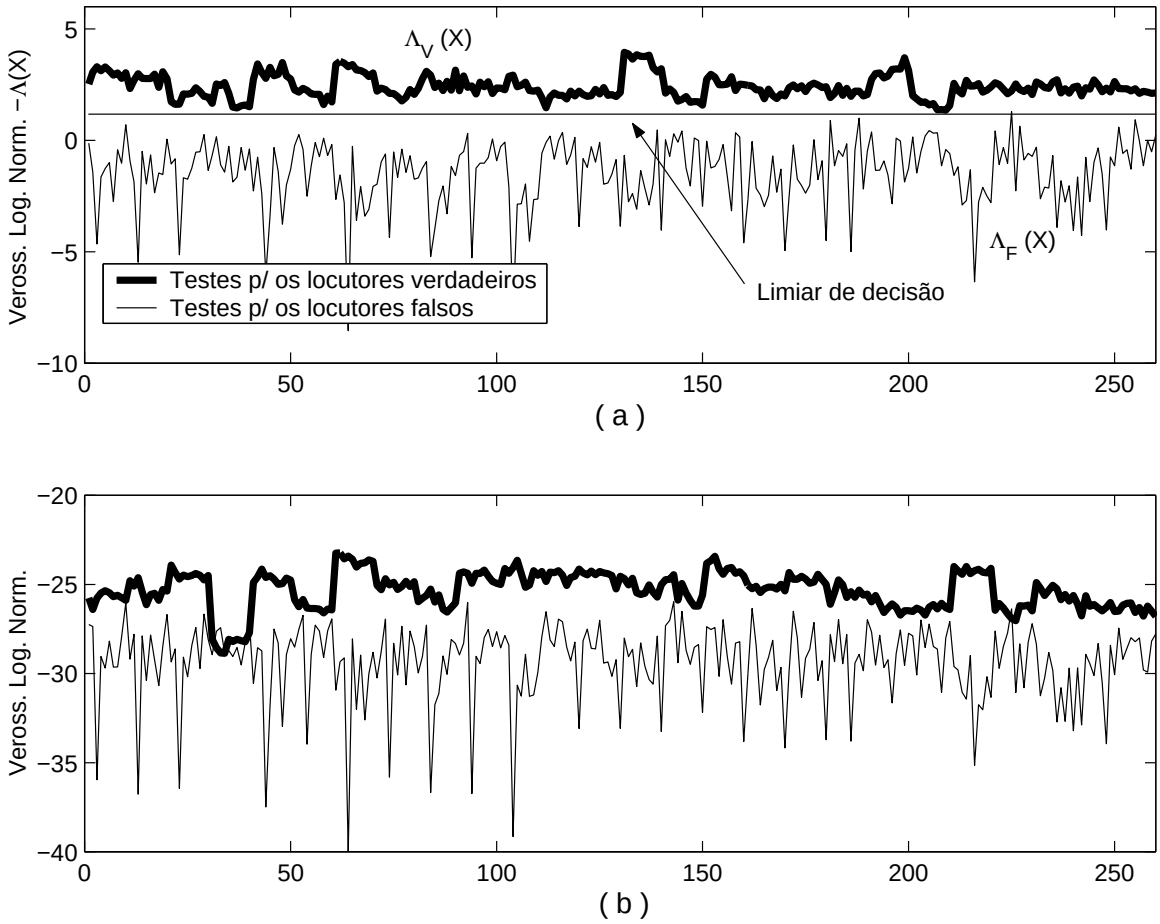

FIG. 5.1: Resultado do GMM com o uso do *background* com 10 locutores em (a) e sem seu uso em (b).

Este limiar de decisão é estimado visando igualar o erro entre FA e FR, sendo um limiar global, utilizado para todos os pretensos locutores, independente do locutor que esteja sendo avaliado. Um limiar para cada locutor deve resultar em melhores resultados. Entretanto, o inconveniente dessa estratégia seria a necessidade de uma massa de dados razoável para definição deste limiar, dificultando de forma acentuada o projeto. Na prática, o emprego de um limiar já estaria estimado e o sistema de verificação necessitaria somente treinar o modelo do novo locutor que fosse pertencer ao sistema, sem a necessidade de dados deste locutor para estimação do limiar (REYNOLDS, 1995).

Os resultados aqui apresentados, mostraram claramente a necessidade do *background*, visando eliminar as variações não pertencentes à identidade do locutor e fornecendo um limiar de decisão mais confiável.

5.4.2 DEFINIÇÃO DO NÚMERO DE COEFICENTES MEL-CEPESTRAIS

Foi realizada uma comparação entre o uso de 15 coeficentes mel-cepestrais e sua redução para 12, incluíndo a compressão de 3 coeficientes conseguida pelo uso do PCA. Fez-se uso do GMM treinado com 60s e 32 gaussianas, o modelo mais preciso utilizado nesta dissertação. Na comparação utilizou-se o MCC e o MCCPCA, notação dada por MCC15, MCC12, MCCPCA15 e MCCPCA12, para 15 e 12 coeficientes respectivamente. A compressão foi representada por MCCPCA1512. Os resultados podem ser vistos na TAB. 5.1. Com um tempo de teste de 10s o uso de MCC15 ou MCC12 é indiferente.

TAB. 5.1: Desempenho do GMM com a variação do número de coeficientes MCC e MCC-PCA.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	0	0	**0**	0,38	0,38	**0,38**	1,19	1,58	**1,38**
MCC12	0	0,38	**0,19**	0,26	0,51	**0,38**	2,19	1,42	**1,80**
MCCPCA15	0	0	**0**	0,38	0,38	**0,38**	1,27	1,19	**1,23**
MCCPCA12	0	0,38	**0,19**	0,26	0,51	**0,38**	1,65	1,65	**1,65**
MCCPCA1512	0	0,38	**0,19**	0,90	0,51	**0,70**	1,96	1,88	**1,92**

De um modo geral, observa-se que o MCC15 obteve melhores resultados que o MCC12, como esperado, visto a maior quantidade de informação contida no MCC15. Para 30s de teste o MCC15 não apresentou erros. Seu desempenho foi melhor em relação ao MCC12 para 3s de teste. Estes resultados demonstram que os 3 coeficientes desprezados carregam informação útil para o reconhecimento de locutor e, portanto, desprezá-los prejudica o desempenho do GMM.

Analisando a utilização do PCA, o MCCPCA15 teve o melhor desempenho que os demais para o tempo de teste de 3s. O MCCPCA12 obteve resultados iguais ao MCC12 para 30 e 10s de teste e superando o anterior para 3s, ou seja, quando o número de MCCs diminuiu, o PCA resultou em ganho somente no teste com menor estatística.

A compressão de 3 coeficientes – MCCPCA1512, resultou nos piores resultados para 10 e 3s de teste. Estes resultados indicam que com uma menor estatística nos dados de teste, a compressão dada pelo PCA acarreta perda de informação sobre a identidade do locutor. Apesar do PCA possuir menor erro de reconstrução, isto não indica que características comprimidas por ele, forneçam maior poder discriminativo para os sistemas de classificação (MALAYATH, 2000).

Uma visualização gráfica dos resultados pode ser vista na FIG. 5.2, na qual é apresentado um gráfico em barras do desempenho EER para 30, 10 e 3s de teste.

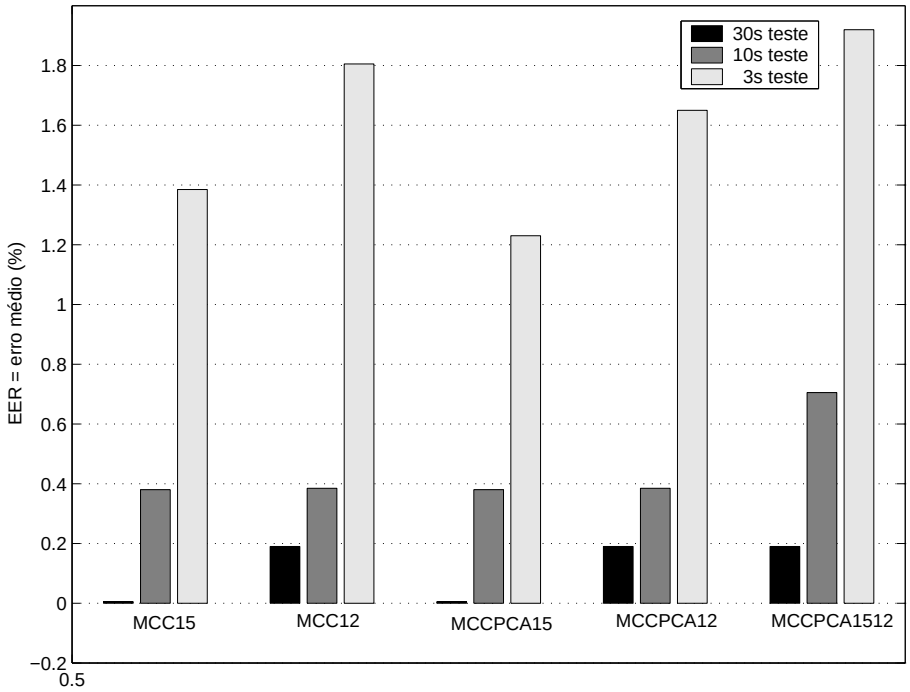

FIG. 5.2: Avaliação do número de MCCs e MCCPCAs no GMM com 32 gaussianas e 60s treinamento.

Os resultados apresentados nesta seção demonstram a importância dos 3 coeficientes mel-cepestrais, do 13º ao 15º. Como os resultados foram obtidos com a modelagem mais precisa utilizada nesta dissertação para o GMM: 32 gaussianas e 60s de treinamento, espera-se que com um GMM mais simples, o desempenho do MCC12 seja mais inferior ao do MCC15. A partir desta análise inicial, os demais resultados foram obtidos utilizando-se 15 coeficientes mel-cepestrais.

5.4.3 AVALIAÇÃO DO NÚMERO DE GAUSSIANAS, TEMPO DE TREINAMENTO E TESTE

Nesta seção são apresentados os desempenhos do GMM em função da variação do número de gaussianas e do tempo de treinamento e teste para os conjuntos de características MCC e MCCPCA. Podendo, assim determinar, o relacionamento entre o número de gaussianas, tempo de treinamento e teste, assim como o efeito da utilização do PCA para melhorar o desempenho do GMM. O número de gaussianas é avaliado respectivamente com 60, 30 e 10s de treinamento. Na TAB. 5.2 estão os resultados conseguidos pelo

TAB. 5.2: Desempenho do GMM com MCC e MCCPCA, para 32 gaussianas com 60s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	0	0	**0**	0,38	0,38	**0,38**	1,19	1,58	**1,38**
MCCPCA15	0	0	**0**	0,38	0,38	**0,38**	1,27	1,19	**1,23**

GMM, com 32 gaussianas e 60s de treinamento. Analisando os resultados da TAB. 5.2, observa-se que para 30s de teste o MCC e o MCCPCA não obtiveram erros com os dados testados. Quando o tempo de teste diminui para 3s, o MCCPCA dá um resultado um pouco melhor. Se o objetivo fosse obter o melhor sistema de verificação de locutor, pelos resultados obtidos, usaria-se um GMM com 32 gaussianas treinado com 60s e o teste teria de ser realizado com 30s, e tanto o MCC como o MCCPCA poderiam ser usados. Neste caso, a opção pelo MCC forneceria um sistema com menor complexidade. Contudo, se dispormos somente de 3s de voz, o MCCPCA apresentará melhor resultado.

Na TAB. 5.3 os resultados são fornecidos para um menor número de gaussianas (16) e o mesmo tempo de treinamento anterior.

TAB. 5.3: Desempenho do GMM com MCC e MCCPCA, para 16 gaussianas com 60s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	0,77	0,38	**0,57**	0,38	0,77	**0,57**	1,65	2,23	**1,94**
MCCPCA15	0	0,38	**0,19**	0,51	0,77	**0,64**	1,65	1,92	**1,78**

Percebe-se pelos resultados da TAB. 5.3, um menor desempenho do GMM comparado à TAB. 5.2. Para 30s de teste o MCCPCA não rejeitou nenhum locutor verdadeiro, superando o MCC. Para 10s de teste o MCC foi ligeiramente superior ao MCCPCA para o erro de FR. No caso de 3s de teste o MCCPCA apresentou um menor erro de FA.

Na TAB. 5.4 são apresentados os resultados para o menor número de gaussianas utilizado (8), com 60s de treinamento. Comparando os resultados desta tabela com os GMMs de 32 e 16 gaussianas com 60s de treinamento, percebe-se que o desempenho do sistema diminui consideravelmente. Com os tempos de testes de 30 e 10s os erros estão entre 1 e 2%, para o tempo de teste de 3s o erro se torna praticamente o dobro dos anteriores. Nesta condição crítica em relação ao número de gaussianas, o MCCPCA levou

TAB. 5.4: Desempenho do GMM com MCC e MCCPCA, para 8 gaussianas com 60s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	2,31	0,38	**1,34**	1,92	1,67	**1,79**	3,54	3,62	**3,58**
MCCPCA15	1,92	0,38	**1,15**	1,67	0,90	**1,28**	3,46	2,15	**2,80**

vantagem sobre o MCC , nos dois tipos de erro, FR e FA. Serão apresentados a seguir o desempenho do GMM variando-se o número de gaussianas para um menor tempo de treinamento.

Na TAB. 5.5 são apresentados os resultados para o GMM treinado com 30s, utilizando 32 gaussianas. O erro que estava próximo a zero na TAB. 5.2, passa com 30s de teste para

TAB. 5.5: Desempenho do GMM com MCC e MCCPCA, para 32 gaussianas com 30s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	0,38	1,92	**1,15**	1,41	1,54	**1,47**	2,50	3,50	**3,00**
MCCPCA15	0,77	1,15	**0,96**	1,28	0,90	**1,09**	2,77	2,69	**2,73**

a casa do 1%. O desempenho do sistema foi claramente prejudicado pela redução do tempo de treinamento, o mesmo é válido para o 10s de teste da TAB. 5.5 com um erro próximo ao caso descrito na TAB. 5.2 utilizando 3s. O desempenho do MCCPCA foi superior em todos os tempos de teste ao MCC, para o EER.

Os resultados utilizando 16 gaussianas e 30s de treinamento podem ser vistos na TAB. 5.6. Nesta tabela a superioridade do MCCPCA em relação ao MCC é mais signi-

TAB. 5.6: Desempenho do GMM com MCC e MCCPCA, para 16 gaussianas com 30s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	3,08	2,69	**2,88**	3,21	3,46	**3,33**	2,88	5,92	**4,40**
MCCPCA15	1,15	1,15	**1,15**	1,54	1,92	**1,73**	3,15	3,54	**3,34**

ficativa do que a obtida nos resultados anteriores. Para 30s de teste, o MCCPCA apresenta menos da metade do erro do MCC, para 10s de teste esta diferença diminui, mas ainda é

considerável. Com 3s de teste, o desempenho do MCCPCA cai pela metade comparado com 10s.

Com 8 gaussianas o desempenho do GMM para 30s de treinamento é mostrado na TAB. 5.7. Esta tabela apresenta os melhores resultados do MCCPCA em relação ao MCC

TAB. 5.7: Desempenho do GMM com MCC e MCCPCA, para 8 gaussianas com 30s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	5,38	4,62	**5,00**	6,15	4,49	**5,32**	6,42	6,88	**6,65**
MCCPCA15	1,92	2,31	**2,11**	1,54	2,82	**2,18**	4,00	3,85	**3,92**

obtidos até aqui. Percebe-se claramente a significativa superioridado do MCCPCA quando o número de gaussianas é 8 e o tempo de treinamento é de 30s. O caso mais crítico usado para treinar o GMM (estatística mais pobre), que utiliza locuções de 10s, será apresentado nas próximas tabelas.

Na TAB. 5.8 é apresentado o desempenho do GMM para 32 gaussianas com 10s de treinamento. O erro obtido na TAB. 5.8, é bastante considerável se levarmos em

TAB. 5.8: Desempenho do GMM com MCC e MCCPCA, para 32 gaussianas com 10s de treinamento.

Característica	Testes (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
MCC15	4,10	4,74	**4,42**	6,15	7,62	**6,68**
MCCPCA15	4,10	4,74	**4,42**	6,77	7,62	**7,19**

conta que estamos trabalhando com "sinais limpos", estando entre 4 e 8%. Para 10s de teste o MCC e MCCPCA forneceram resultados iguais. Para o tempo de teste de 3s o MCC superou ligeiramente o MCCPCA. Percebe-se que com uma pequena massa de dados para treinamento o desempenho do GMM, que é um classificador estatístico, cai consideravelmente.

Os resultados com 16 gaussianas para 10s de treinamento podem ser vistos na TAB. 5.9. A variação do erro praticamente não sofreu alteração com relação ao uso de 32 gaussianas, para 10s de treinamento. O MCCPCA conseguiu erros iguais de FR e FA, um caso particular que ocorreu na escolha do limiar de decisão. O MCC manteve os erros próximos

TAB. 5.9: Desempenho do GMM com MCC e MCCPCA, para 10s de treinamento, 16 gaussianas.

Característica	Testes (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
MCC15	4,49	5,26	**4,87**	6,00	7,73	**6,86**
MCCPCA15	4,23	4,23	**4,23**	6,69	6,69	**6,69**

aos do MCCPCA, superando este apenas no erro de FR para 3s de teste. O ganho do MCCPCA em relação ao MCC foi muito baixo.

O piores resultados são mostrados na TAB. 5.10, onde se utilizou 8 gaussianas e 10s de treinamento. Obteve-se um EER de 7,73%, com o MCCPCA para 3s de teste. O MCC

TAB. 5.10: Desempenho do GMM com MCC e MCCPCA, para 10s de treinamento, 8 gaussianas.

Característica	Testes (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
MCC15	3,59	5,90	**4,74**	7,15	7,19	**7,17**
MCCPCA15	4,62	5,64	**5,13**	7,81	7,65	**7,73**

superou os resultados do MCCPCA para 3s de teste e só perdeu para o erro de FA com 10s de teste.

Quando o tempo de treinamento é muito baixo (10s), o número de gaussianas é irrelevante para melhorar o desempenho do sistema. Em geral, em todas as configurações do GMM analisadas, o MCCPCA superou o MCC, a não ser com tempo de treinamento de 10s. Neste caso o erro de ambos foi grande e tal sistema não seria interessante na verificação de locutor. A utilização do PCA forneceu poder discriminativo para evitar o locutor falso e agregou mais informação ao locutor verdadeiro.

A FIG. 5.3 apresenta de forma gráfica o desempenho do MCCPCA e MCC para a variação do número de gaussianas com seus respectivos tempos de treinamento (tr), para o tempo de teste de 30s.

O ganho produzido pelo PCA na FIG. 5.3 é mais acentuado para 30s de treinamento. Neste caso, o GMM com 16 gaussianas utilizando MCCPCA obteve desempenho comparável ao GMM com 32 gaussianas e MCC.

Os resultados para o tempo de teste de 10s, podem ser vistos na FIG. 5.4.

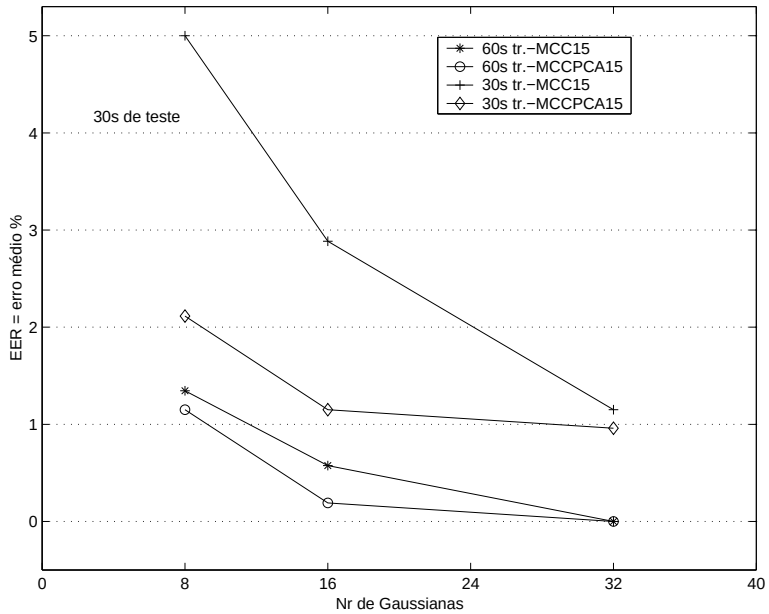

FIG. 5.3: Desempenho do GMM (EER) para 30s de teste, em relação ao número de gaussianas e tempo de treino.

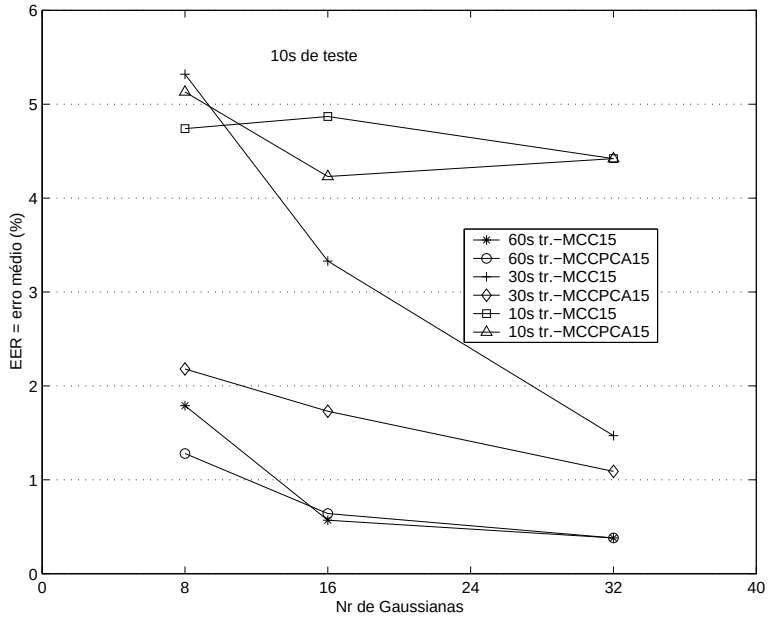

FIG. 5.4: Desempenho do GMM (EER) para 10s de teste, em relação ao número de gaussianas e tempo de treino.

Novamente, o ganho dado pelo PCA é mais perceptível para 30s de treinamento. O GMM com 16 gaussianas e MCCPCA teve resultado similar ao GMM com 32 gaussianas com MCC. Para 60s de treinamento o MCC e MCCPCA tiveram resultados próximos. Com 10s de treinamento o desempenho do GMM é ruim para ambos MCC e MCCPCA e a vantagem do MCCPCA deixa de ser aparente. O pior resultado neste gráfico é dado pelo GMM com 8 gaussianas treinado com 30s utilizando MCC.

Na FIG. 5.5 o desempenho do GMM para 3s de teste é apresentado com conclusões semelhantes às da FIG. 5.4.

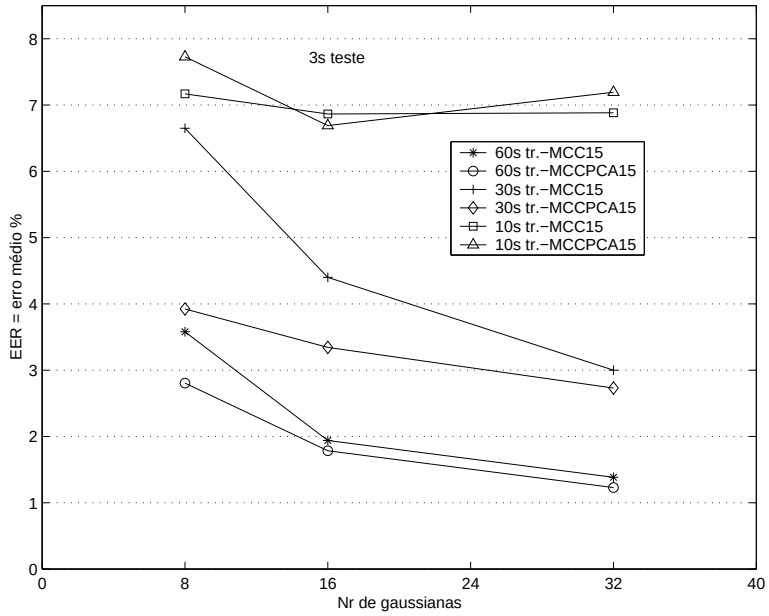

FIG. 5.5: Desempenho do GMM (EER) para 3s de teste, em relação ao número de gaussianas e tempo de treino.

5.4.4 ANÁLISE DO BACKGROUND

Não existe uma medida objetiva para determinar a quantidade de voz necessária para treinar o modelo universal de *background*, e nenhum experimento mais cuidadoso foi realizado para avaliar o número de locutores presentes no treinamento do modelo universal de *background* (REYNOLDS, 2000). No presente trabalho foi realizado um experimento com o MCC e MCCPCA utilizando um GMM com 32 gaussianas e 60s de treinamento para avaliar o desempenho do *background* com 10 e 16 locutores. Foram utilizados 6s de voz de cada locutor, gerando *backgrounds* com 60 e 96s, respectivamente, com um número igual de vozes masculinas e femininas. É importante ressaltar que, considerando a limitada base de dados para esse experimento, o aumento do número de locutores no *background*

acarretou um número menor de locutores para teste (20). Desta maneira, diminuiu-se a estatística dos resultados uma vez que os locutores de *background* não participaram dos testes. Ambos UBMs com 10 e 16 locutores utilizaram 20 locutores para testes. Não são apresentados os resultados para 30s de teste pois os sistemas não apresentaram erros com os dados de teste utilizados.

Na TAB. 5.11 é apresentado o desempenho do sistema para um UBM com 10 locutores. Verifica-se que existe uma pequena diferença em favor do MCCPCA para o erro de FA

TAB. 5.11: Desempenho do GMM com 10 locutores de *background*.

Característica	Testes (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
MCC15	0,50	0,50	**0,50**	1,35	1,65	**1,50**
MCCPCA15	0,50	0,33	**0,42**	1,25	1,30	**1,28**

com 10s de teste. Já para 3s de teste o MCCPCA superou todos os resultados do MCC.

O desempenho do GMM usando um UBM com 16 locutores, pode ser visto na TAB. 5.12 Para 10s de teste o MCC e MCCPCA não cometeram erros, o que é um ganho razoável

TAB. 5.12: Desempenho do GMM com 16 locutores de *background*.

Característica	Testes (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
MCC15	0	0	**0**	1,10	1,20	**1,15**
MCCPCA15	0	0	**0**	1,25	1,10	**1,18**

comparado com a TAB. 5.11. Para 3s de teste o MCC forneceu melhores resultados do que com um UBM de 10 locutores, o MCCPCA obteve resultado inferior no erro de FR e superior no erro de FA, ficando um pouco atrás do MCC para o EER. Aparentemente, os resultados mostram que um UBM com mais locutores e tempo de voz produz melhores resultados e que o PCA não resulta em um ganho sobre o MCC para um tempo de teste pequeno. Com base nestes testes, além do desempenho em termos da taxa de erro, chegou-se às seguintes conclusões para o aumento do *background*:

1. Passar as características de voz de um indivíduo por um modelo de *background* maior aumenta a verossimilhança dada por este modelo. Isso se deve ao fato do aumento do background retratar melhor a população a que o índividuo pertence.

2. De 1, a diferença entre a verossimilhança produzida pelo modelo do pretenso locutor e a verossimilhança produzida pelo *background* diminui ($\Lambda(X)$, conforme EQ. 4.14), tanto para um locutor verdadeiro como para um falso.

3. No sistema de verificação (FIG. 4.4) o resultado de $\Lambda(X)$ é comparado um determinado limiar de decisão baseado nas diferenças produzidas pelos locutores verdadeiros ($\Lambda_V(X)$) e falsos ($\Lambda_F(X)$), conforme ilustrado na FIG. 5.1 (a). Logo, quanto mais afastadas forem estas diferenças, melhor tenderá a ser o desempenho. Assim o desejável seria aumentar $\Lambda_V(X)$ e diminuir $\Lambda_F(X)$.

4. Com o aumento do locutores de *background*, as duas diferenças caem ($\Lambda_V(X)$ e $\Lambda_F(X)$), ou seja, ocorre algo indesejável que é a queda de $\Lambda_V(X)$ e algo desejável que é a queda de $\Lambda_F(X)$. À princípio não se pode afirmar nada a respeito da melhora de desempenho com o aumento do número de locutores do *background*. A não ser que $\Lambda_F(X)$ caia significativamente mais que $\Lambda_V(X)$, aumentando a distância entre esta diferenças, como desejável.

5. Dos resultados numéricos obtidos:

   a) Sem usar PCA $\Lambda_F(X)$ cai mais que $\Lambda_V(X)$.

   b) Usando PCA, essa queda de $\Lambda_F(X)$ em relação a $\Lambda_V(X)$ é um pouco menor.

   c) De *a* e *b*, usando ou não PCA, o aumento de desempenho ocorre com o aumento do *background*. Porém sem o PCA, o aumento de desempenho é maior.

Para validar realmente estas conclusões, outro experimentos teriam que ser conduzidos utilizando uma quantidade maior de dados de treinamento e teste, bem como várias configurações com diferentes números de locutores e tempos de voz disponíveis para treinar o UBM. Esta experiência deve ser considerada, portanto, apenas como um ponto de partida para futuras pesquisas.

5.5 AVALIAÇÃO DO AR-VETORIAL

Na avaliação do AR-Vetorial foram utilizados 15 coeficientes mel-cepestrais, como definido no uso do GMM. Utilizaram-se também o mesmo conjunto de tempos de treinamento e teste.

5.5.1 DEFINIÇÃO DO LIMIAR DE DECISÃO

Como o GMM, o AR-Vetorial necessita de um limiar de decisão para a verificação de locutor, estimado entre os testes verdadeiros e falsos. Este limiar é dado pela distância de Itakura do modelo armazenado de um locutor e suas locuções de teste (verdadeiras) e a distância do mesmo modelo para locuções de teste de outros locutores (falsos). Espera-se que se o teste for do locutor verdadeiro a distância de Itakura será pequena, e se o teste for de um locutor falso será grande. Na FIG. 5.6, isto pode ser visto usando a distância simétrica de Itakura (ver capítulo 4), para um modelo de ordem 2 ($p = 2$), estimado com 60s de voz e testado com 30s.

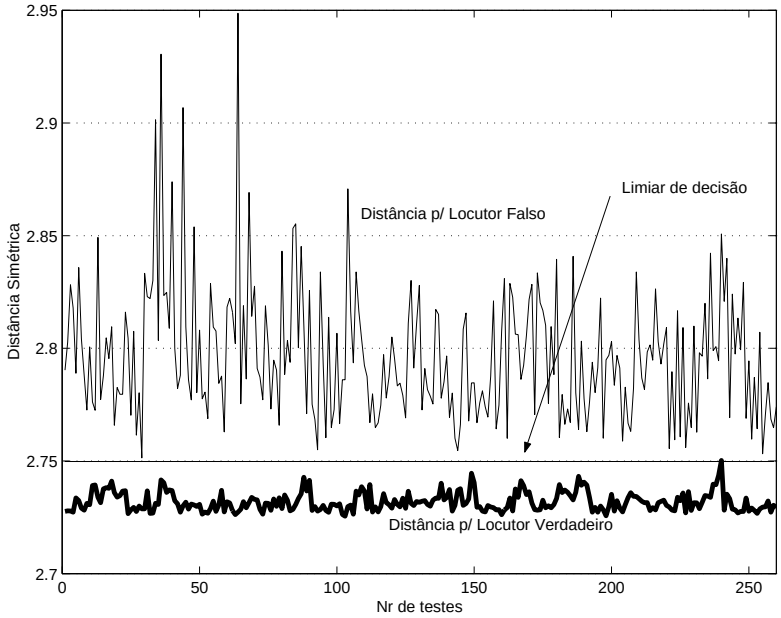

FIG. 5.6: Limiar de decisão para uso no AR-Vetorial.

Observa-se que, ao contrário do GMM, a linha pertencente aos testes relativos ao locutor verdadeiro (linha cheia) está abaixo da linha pertencente aos testes associados aos locutores falsos, porque a distância dos testes verdadeiros ao modelo a que pertencem é menor que a dos testes falsos para o mesmo modelo. No GMM, os testes verdadeiros resultarão num valor maior de verossimilhança (FIG. 5.1).

A idéia de utilizar o *background* não faz sentido no AR-Vetorial, pelo menos da forma como utilizado no GMM. A razão disso é que o AR-Vetorial utiliza uma distância para comparar modelos. A FIG. 5.7 mostra isto claramente (mesma modelagem da figura anterior): subtraindo-se a distância entre o modelo do locutor e a do UBM, obtêm-se duas linhas, a cheia representando os testes verdadeiros e a fina os falsos. Percebe-se que existe

uma mistura entre as duas linhas, indicando que a utilização do UBM produzirá resultados ruins quando um limiar for escolhido.

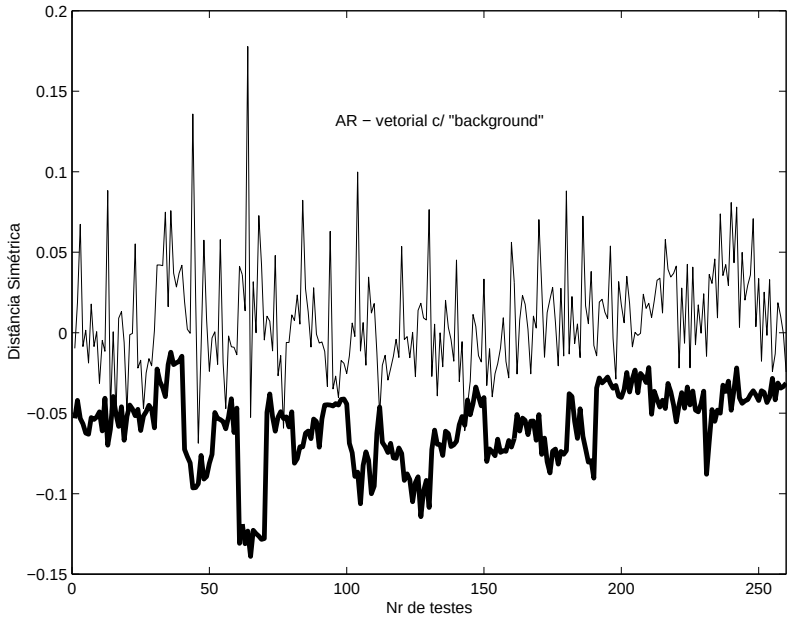

FIG. 5.7: Resultados do AR-Vetorial usando o UBM.

5.5.2 DEFINIÇÃO DA DISTÂNCIA UTILIZADA NO MODELO

O AR-Vetorial pode fazer uso de 7 medidas de distância, dadas pelas EQs 4.32 a 4.38, utilizando os modelos direto e reverso para efetuar a medida entre dois modelos (apêndice 2). Tais medidas serão referenciadas pelos números de 1 a 7, de acordo com a ordem que aparecem no capítulo 4, respectivamente.

Para avaliar a melhor distância, foi utilizado um AR-Vetorial de ordem 2[1], estimado com 60s de treinamento. Estes resultados podem ser vistos na TAB. 5.13. No apêndice 3 são dados os resultados para o desempenho das distâncias, com o modelo de ordem 2, para 30 e 10s de treinamento, resultados semelhantes ao de 60s de treinamento. Para 30s de teste a distância 5 obteve os melhores resultados, seguida pelas distâncias 1, 2 e 7 todas com os mesmos resultados, com erro apenas de FR. Após estas, vem a distância 6. As distâncias 3 e 4 não obtiveram bons resultados. Com 10s de teste o menor EER foi da distância 5 seguido pela 7 e 1. A distância 5 perdeu apenas no erro de FA para a distâncias 2 e 7. As distâncias 3 e 4 tiveram os piores resultados. Para 3s de teste o erro dado pelas medidas se torna muito grande. Nesse caso, os melhores resultados de EER foram para as distâncias 1, 5 e 7.

[1] vide capítulo 4

TAB. 5.13: Desempenho das distâncias usadas no AR-Vetorial para $p = 2$, 60s de treinamento.

Dist.	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
1	0,38	0	**0,19**	1,15	2,30	**1,72**	8,80	9,60	**9,20**
2	0,38	0	**0,19**	2,40	1,41	**1,90**	14,2	13,4	**13,8**
3	3,8	3,8	**3,8**	7,4	8,30	**7,85**	19,0	16,0	**17,5**
4	3,8	3,8	**3,8**	6,6	3,2	**4,9**	15,5	15	**15,25**
5	0	0	**0**	0,89	1,6	**1,24**	8,5	11,4	**9,95**
6	1,15	0,76	**0,95**	3,1	2,8	**2,95**	12,7	12,3	**12,5**
7	0,38	0	**0,19**	1,92	0,90	**1,41**	8,2	12,6	**10,4**

Uma visualização gráfica dos resultados da TAB. 5.13 é mostrada na FIG. 5.8 em termos do EER. Como se vê nesta figura, os melhores desempenhos são conseguidos pelas

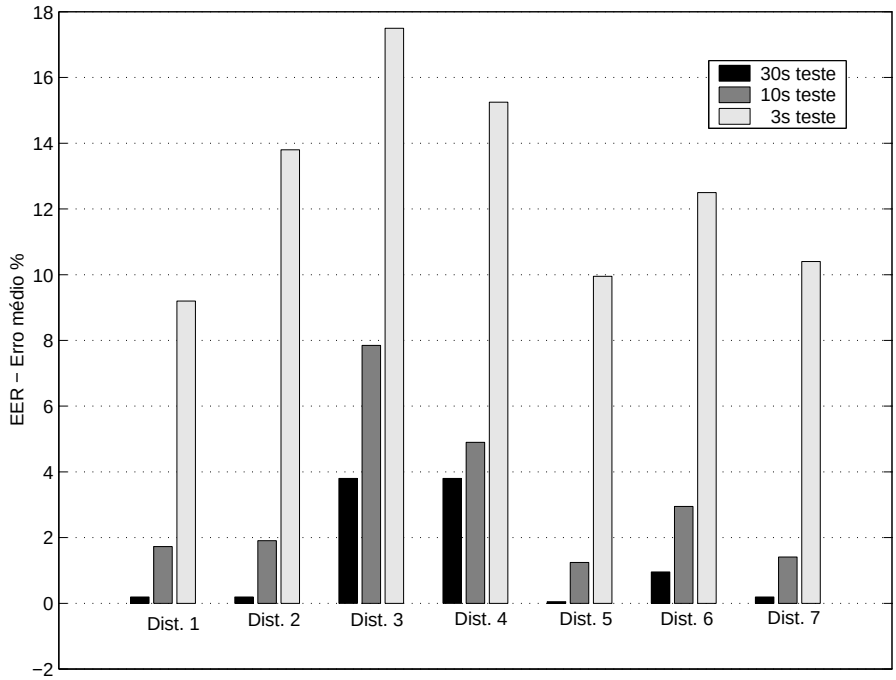

FIG. 5.8: Desempenho nas Distâncias utilizadas no AR-Vetorial, para $p = 2$ com 60s de treinamento.

distâncias 5, 7 e 1, respectivamente, para os tempos de teste de 30 e 10s. Já para o tempo de teste de 3s o desempenho do AR-Vetorial é ruim, e os melhores resultados foram obtidos pela distância 1, seguida da 5 e da 7. A partir destes resultados, optando-se por um melhor desempenho do AR-Vetorial, a distância 5, ou seja, a distância simétrica é utilizada.

5.5.3 AVALIAÇÃO DA ORDEM DO MODELO, TEMPO DE TREINAMENTO E TESTE

A ordem do AR-Vetorial é dado pelo número de coeficientes estimados no modelo, matrizes A_k da EQ. 4.25. Foram feitas variações na ordem do modelo de 1 até 5 e os resultados para 60s de treinamento, usando a distância simétrica (5), são apresentados na TAB. 5.14. Os resultados para as distâncias 1 e 7 são dados no apêndice 3. Da

TAB. 5.14: Desempenho do AR-Vetorial para variações na ordem do modelo, utilizando a distância simétrica, com 60s de treinamento.

Ordem do modelo	Teste (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
1	0,38	0,77	**0,57**	1,8	1,7	**1,75**	10,4	9,8	**10,1**
2	0	0	**0**	0,89	1,6	**1,24**	8,5	11,4	**9,95**
3	0	0	**0**	1,9	0,38	**1,14**	10,6	8,8	**9,7**
4	0	0	**0**	1,3	1,4	**1,35**	11,2	10,15	**10,67**
5	0	0	**0**	1,4	1,0	**1,2**	13,6	13,4	**13,5**

TAB 5.14 nota-se que com 30s de teste, somente o modelo de ordem 1 resultou em erros. Os modelos de ordens 2 a 5 produzem resultados semelhantes para 10s de teste, sendo os melhores desempenhos conseguidos pelos modelos de ordens 3, 5, 4, 2 e 1. Percebe-se que a ordem do modelo não influencia no desempenho do AR-Vetorial para o tempo de teste de 3s, com EER na faixa de 10%, com exceção para o modelo de ordem 5, com EER na faixa de 13%.

Na TAB. 5.15 é apresentado o desempenho do AR-Vetorial em função da ordem do modelo para 30s de treinamento. Novamente o erro do modelo de ordem 1 foi o maior

TAB. 5.15: Desempenho do AR-Vetorial para variações na ordem do modelo, utilizando a distância simétrica, com 30s de treinamento.

Ordem do modelo	Teste (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
1	0,77	0,77	**0,77**	2,2	2,2	**2,2**	10,4	10,5	**10,45**
2	0	0	**0**	1,8	1,4	**1,6**	10,2	10,3	**10,25**
3	0	0	**0**	1,6	1,6	**1,6**	10,8	10,2	**10,5**
4	0	0	**0**	1,4	1,3	**1,35**	11,5	10,7	**11,1**
5	0	0	**0**	1,7	1,7	**1,7**	13,7	13,3	**13,5**

para 30 e 10s de teste. O erro para 3s de teste esteve na faixa de 10%. Para 10s de teste há um certo balanceamento de desempenho dos modelos de ordem 2 a 5, com melhor

resultado para o modelo de ordem 4. Nota-se que o desempenho do AR-Vetorial para 30s de treinamento não sofre alterações significativas em relação ao treinamento com 60s.

Na TAB. 5.16 é apresentado o desempenho do AR-Vetorial em função da ordem do modelo , para o menor tempo de treinamento utilizado (10s). O erro obtido com um tempo

TAB. 5.16: Desempenho do AR-Vetorial para variações na ordem do modelo, utilizando a distância simétrica, com 10s de treinamento.

Ordem do modelo	Teste (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
1	3,5	3,2	**3,35**	11,0	11,3	**11,15**
2	2,4	3,8	**3,1**	11,6	11,2	**11,4**
3	3,3	2,9	**3,1**	12,9	11,5	**12,2**
4	3,6	3,3	**3,45**	12,3	12,4	**12,35**
5	3,8	3,5	**3,65**	14,4	14,9	**14,65**

de treinamento de 10s praticamente dobrou para 10s de teste e aumentou em 1% o erro para 3s. Os melhores desempenhos para 10s de teste foram conseguidos para os modelos de ordem 2 e 3 e os piores para as ordens 4 e 5. Com esse tempo de treinamento o modelo de ordem menor superou os de ordem 4 e 5 e produziu o menor EER para 3s de teste.

Quando o tempo disponível para estimação do AR-Vetorial é pequeno, o modelo de menor ordem fornece resultados superiores aos modelos de maior ordem porque o tempo para a estimação destes, não foi adequado para uma modelagem precisa.

No geral, os modelos de ordem 2 e 3 forneceram os melhores resultados. Desta forma, pode-se optar pelo modelo de ordem 2, pela sua maior simplicidade.

5.5.4 AVALIAÇÃO ENTRE MCC12 E MCC15

Assim como no GMM, uma avaliação do AR-Vetorial para os conjuntos de características MCC12 e MCC15 foi realizada.

Como era esperado, a diminuição do número de coeficientes mel-cepestrais piorou o desempenho do AR-Vetorial, devido à perda de informação sobre a identidade do locutor. Isto pode ser visto na TAB. 5.17 para um AR-Vetorial de ordem 2, utilizando a distância simétrica de Itakura e para os tempos de 60, 30 e 10s de treinamento. Nota-se que para todos os tempos de teste o AR-Vetorial estimado com MCC12 resultou num decréscimo de desempenho, quando comparado com o MCC15, a não ser para um tempo de treinamento de 60s e 30s de teste, quando não ocorreram erros em nenhum dos casos. Esses resultados comprovam a importância do número de coeficientes MCC na verificação de locutor

TAB. 5.17: Desempenho do AR-Vetorial para MCC12 e MCC15, para ordem 2 e distância Simétrica.

Caract. - tr.	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC12 - 60s tr.	0	0	**0**	1,54	1,67	**1,6**	11,88	11,61	**11,74**
MCC15 - 60s tr.	0	0	**0**	0,89	1,6	**1,24**	8,5	11,4	**9,95**
MCC12 - 30s tr.	0,38	0,38	**0,38**	1,92	2,05	**1,98**	11,85	11,42	**11,63**
MCC15 - 30s tr.	0	0	**0**	1,8	1,4	**1,6**	8,5	11,4	**9,95**
MCC12 - 10s tr.	x	x	x	3,2	3,97	**3,58**	13,73	12,46	**13,09**
MCC15 - 10s tr.	x	x	x	2,4	3,8	**3,1**	11,6	11,2	**11,4**

utilizando o AR-Vetorial. Os demais testes com o AR-Vetorial foram realizados com o MCC15, incluindo o MCCPCA15. Não foram realizados testes com o PCA nos MCC12, tendo em vista os resultados do MCC12. No apêndice 3 são fornecidos os resultados para o MCC12 e MCC15 usando os modelos de ordem 1,2 e 3, com as distâncias 1,5 e 7.

5.5.5 DESEMPENHO DO PCA NO AR-VETORIAL

Tendo em vista os resultados obtidos para o AR-Vetorial, os testes com o PCA foram realizados apenas nos modelos de ordem 1, 2 e 3 com as distâncias 1, 5 e 7. Os resultados presentados nesta seção são para o AR-Vetorial de ordem 2 usando distância simétrica, os demais resultados podem ser vistos no apêndice 3.

Os desempenhos obtidos com o uso do MCCPCA em comparação com o MCC, aplicado ao AR-Vetorial de ordem 2, utilizando a distância simétrica, são mostrados na TAB. 5.18 Dos resultados da TAB. 5.18 nota-se que ambos MCC e MCCPCA não obtiveram erros

TAB. 5.18: Desempenho do AR-Vetorial para MCC e MCCPCA, para ordem 2 com distância simétrica, com 60s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	0	0	**0**	0,89	1,6	**1,24**	8,5	11,4	**9,95**
MCCPCA15	0	0	**0**	1,02	1,28	**1,15**	10,04	9,92	**9,98**

nos testes com 30s. No teste com 10s o MCCPCA superou o MCC no erro de FA e no EER. Quando o tempo de teste diminuiu para 3s, o MCC e MCCPCA continuaram com erros elevados, onde o primeiro mostrou menor erro de FR, e o segundo menor erro de FA.

A TAB. 5.19 apresenta os resultados para um tempo de treinamento de 30s. Nota-se

TAB. 5.19: Desempenho do AR-Vetorial para MCC e MCCPCA, para ordem 2 com distância simétrica, com 30s de treinamento.

Característica	Testes (%)								
	30s			10s			3s		
	FR	FA	**EER**	FR	FA	**EER**	FR	FA	**EER**
MCC15	0	0	**0**	1,8	1,4	**1,6**	8,5	11,4	**9,95**
MCCPCA15	0	0	**0**	1,54	1,54	**1,54**	10,23	10,35	**10,29**

que o MCC e MCCPCA produziram valores de EER próximos para 10s de teste, onde o MCC obteve um menor erro de FA. Para 3s de teste o erro continua da ordem de 10%, para ambos, com desvantagem para o MCCPCA no erro de FR e no EER.

O caso mais crítico em relação ao tempo de treinamento pode ser visto na TAB. 5.20. Novamente o MCC e o MCCPCA obtiveram resultados próximos para 10s de teste, onde

TAB. 5.20: Desempenho do AR-Vetorial para MCC e MCCPCA, para ordem 2 com distância simétrica, com 10s de treinamento.

Característica	Testes (%)					
	10s			3s		
	FR	FA	**EER**	FR	FA	**EER**
MCC15	2,4	3,8	**3,1**	11,6	11,2	**11,4**
MCCPCA15	3,08	3,33	**3,2**	12,11	11,58	**11,84**

o MCCPCA ficou aquém do MCC para o erro de FR e EER. O erro para 3s de teste subiu em relação aos resultados das tabelas 5.18 e 5.19. O ganho do PCA no AR-Vetorial é pequeno e em alguns casos não existe. Em resumo, o uso do MCCPCA no AR-Vetorial acarreta um acréscimo computacional sem fornecer ganhos significativos.

5.6 COMPARAÇÃO ENTRE OS RESULTADOS DO GMM E AR-VETORIAL

Os resultados obtidos com o GMM, usando MCC e MCCPCA, com a variação do número de gaussianas, são comparados com os resultados obtidos pelo AR-Vetorial de ordem 2, utilizando a distância simétrica. A razão desta última escolha é que os modelos de ordem 2 à 5 forneceram resultados similares, e o modelo de ordem 2 apresenta uma menor complexidade computacional. Os resultados obtidos com 3s de teste não serão comentados, pois, o erro do AR-Vetorial é muito superior ao do GMM.

Os resultados com 60s de treinamento podem ser vistos na TAB. 5.21 Analisando o MCC para 30s de teste o AR-Vetorial foi superior ao GMM com 16 e 8 gaussianas. Para

TAB. 5.21: Desempenho do GMM versus AR-Vetorial, para 60s de treinamento.

Sistema Utilizado	Testes (%)								
	30s			10s			3s		
	FA	FR	**EER**	FA	FR	**EER**	FA	FR	**EER**
GMM-MCC-32G	0	0	**0**	0,38	0,51	**0,44**	1,19	1,58	**1,38**
GMM-MCC-16G	0,77	0,38	**0,57**	0,38	0,77	**0,57**	1,65	2,23	**1,94**
GMM-MCC-8G	2,31	0,38	**1,34**	1,92	1,67	**1,79**	3,54	3,62	**3,58**
ARV-MCC15	0	0	**0**	0,89	1,60	**1,24**	8,5	11,4	**9,95**
GMM-MCCPCA-32G	0	0	**0**	0,38	0,38	**0,38**	1,27	1,19	**1,23**
GMM-MCCPCA-16G	0	0,38	**0,19**	0,51	0,77	**0,64**	1,65	1,92	**1,78**
GMM-MCCPCA-8G	1,92	0,38	**1,15**	1,67	0,90	**1,28**	3,46	2,15	**1,74**
ARV MCCPCA15	0	0	**0**	1,02	1,28	**1,15**	10,04	9,92	**9,98**

10s de teste o AR-Vetorial resultou em melhores resultados que o GMM somente para 8 gaussianas, apresentando mais do que o dobro de erro que o GMM com 32 e 16 gaussianas. Com 3s de teste o AR-Vetorial teve um desempenho muito inferior ao GMM.

Com o MCCPCA e um tempo de teste de 30s o AR-Vetorial superou o GMM com 16 e 8 gaussianas. Para 10s de teste o erro do AR-Vetorial continua praticamente o dobro em relação do GMM com 32 e 16 gaussianas, ganhando do GMM com 8 gaussianas, com um pequena diferença no EER. Para 3s de teste o AR-Vetorial sempre produz erros muito mais elevados que o GMM.

A TAB. 5.22 apresenta os resultados para um tempo de treinamento de 30s. Nesta

TAB. 5.22: Desempenho do GMM versus AR-Vetorial, para 30s de treinamento.

Sistema Utilizado	Testes (%)								
	30s			10s			3s		
	FA	FR	**EER**	FA	FR	**EER**	FA	FR	**EER**
GMM-MCC-32G	0,38	1,92	**1,15**	1,41	1,54	**1,47**	2,5	3,5	**3,0**
GMM-MCC-16G	3,08	2,69	**2,88**	3,21	3,46	**3,33**	2,88	5,92	**4,4**
GMM-MCC-8G	5,38	4,62	**5,0**	6,15	4,49	**5,32**	6,42	6,88	**6,65**
ARV-MCC15	0	0	**0**	1,8	1,4	**1,6**	10,2	10,3	**10,25**
GMM-MCCPCA-32G	0,77	1,15	**0,96**	1,28	0,90	**1,09**	2,77	2,69	**2,73**
GMM-MCCPCA-16G	1,15	1,15	**1,15**	1,54	1,92	**1,73**	3,15	3,54	**3,34**
GMM-MCCPCA-8G	1,92	2,31	**2,11**	1,54	2,82	**2,18**	4,0	3,85	**3,92**
ARV MCCPCA	0	0	**0**	1,54	1,54	**1,54**	10,23	10,35	**10,29**

tabela o AR-Vetorial superou o GMM para 30s de teste, não apresentando erros. Usando o MCC com 10s de teste, o AR-Vetorial superou o GMM com 16 e 8 gaussianas, ficando com resultados próximos ao GMM com 32 gaussianas. Para este tempo de teste o desempenho

do AR-Vetorial usando o MCCPCA foi praticamente igual ao seu desempenho com o MCC, ficando aquém do GMM com 32 gaussianas. Com o MCCPCA, os ganhos nos 10s de teste do AR-Vetorial sobre o GMM com 16 e 8 gaussianas foram bem menores que os obtidos quando se usa o MCC.

Na TAB. 5.23 são fornecidos os resultados para o menor tempo de treinamento, o qual resultou nos maiores erros nos sistemas de classificação. Quando o tempo de treinamento

TAB. 5.23: Desempenho do GMM versus AR-Vetorial, para 10s de treinamento.

Sistema Utilizado	Testes (%)					
	10s			3s		
	FA	FR	**EER**	FA	FR	**EER**
GMM-MCC-32G	4,10	4,74	**4,42**	6,15	7,62	**6,88**
GMM-MCC-16G	4,49	5,26	**4,87**	6,0	7,73	**6,86**
GMM-MCC-8G	3,59	5,90	**4,74**	7,15	7,19	**7,17**
ARV-MCC15	2,4	3,8	**3,1**	11,6	12,0	**11,8**
GMM MCCPCA-32G	4,10	4,74	**4,42**	6,77	7,62	**7,19**
GMM-MCCPCA-16G	4,23	4,23	**4,23**	6,69	6,69	**6,69**
GMM-MCCPCA-8G	4,62	5,64	**5,13**	7,81	7,65	**7,73**
ARV MCCPCA15	3,08	3,33	**3,20**	12,11	11,58	**11,84**

é de 10s, o GMM com MCC e MCCPCA independente do número de gaussianas, produz resultados próximos. O AR-Vetorial para 10s de teste obteve resultados entre 1 e 2% superiores ao GMM.

5.7 COMPARAÇÃO ENTRE OS TEMPOS DE PROCESSAMENTO

O processamento da massa de dados para os sistemas de verificação usados, foi realizado em um computador Pentium-II MMX© 333MHz, 128Mb de memória RAM, usando o programa Matlab© 5.3. Os tempos encontrados nesse processamento fornecem uma maneira de inferir a carga computacional do GMM e do AR-Vetorial. Tais valores são, entretanto, dependentes do tipo de *hardware* e *software* empregado.

O tempo de processamento para cálculo do MCC não é apresentado, pois deseja-se somente avaliar o desempenho entre o GMM e AR-Vetorial. Uma vez calculados, os tempos de processamento (treinamento e teste) são os mesmos para o MCC e MCCPCA. O cálculo do PCA está em torno de 0,38 a 0,05s para 60, 30 e 10s de treinamento e sua utilização nos testes em torno de 0,06 a 0,01s.

A TAB. 5.24 apresenta os resultados dos tempos de processamento para os diferentes tempos de treinamento utilizados no GMM (32, 16 e 8 gaussianas) e AR-Vetorial (ordem 2

com distância 5). Percebe-se que o tempo de processamento do AR-Vetorial é muito menor

TAB. 5.24: Tempos de processamento para treinamento do GMM e AR-Vetorial.

Sistema Utilizado	Tempo de Processamento (s)		
	60s tr.	30s tr.	10s tr.
GMM-32G	302	164	28
GMM-16G	180	68	14
GMM-8G	67	45	7,5
ARV-P2	4,8	2,4	0,8

que do GMM, em torno de 60 vezes, comparado com o GMM de 32 gaussianas treinado com 60s. A redução do número de gaussianas no GMM em duas vezes, praticamente reduz o tempo de processamento na mesma quantia.

Na TAB. 5.25 são apresentados os resultados para os diferentes tempos de teste utilizados. O AR-Vetorial apresenta um tempo de processamento maior que o GMM nos

TAB. 5.25: Tempos de processamento para testes do GMM e AR-Vetorial.

Sistema Utilizado	Tempo de Processamento (s)		
	30s teste	10s teste	3s teste
GMM-32G	1,6	0,5	0,16
GMM-16G	0,9	0,3	0,11
GMM-8G	0,4	0,15	0,05
ARV-P2	2,45	0,9	0,27

diferentes tempos de teste. Isto é devido à necessidade da estimação do AR-Vetorial nos dados de teste para o uso da distância de Itakura, enquanto que no GMM calcula-se a verossimilhança dos dados de teste (processamento mais simples). O GMM com 32 gaussianas, 30s de teste, é aproximadamente 1,5 vezes mais rápido que o AR-Vetorial para a mesma configuração.

5.8 A CURVA DET

Quando se deseja avaliar o desempenho dos sistemas de verificação para diferentes taxas de erros de FR e FA, pode-se utilizar a curva introduzida pelo NIST e denominada DET (*Detection Error Tradeoff*) (MARTIN, 1997). Esta curva é produzida pelas coordenadas (FA,FR) de um gráfico com eixos horizontal e vertical em escala logarítmica, como ilustrado na FIG. 5.9.

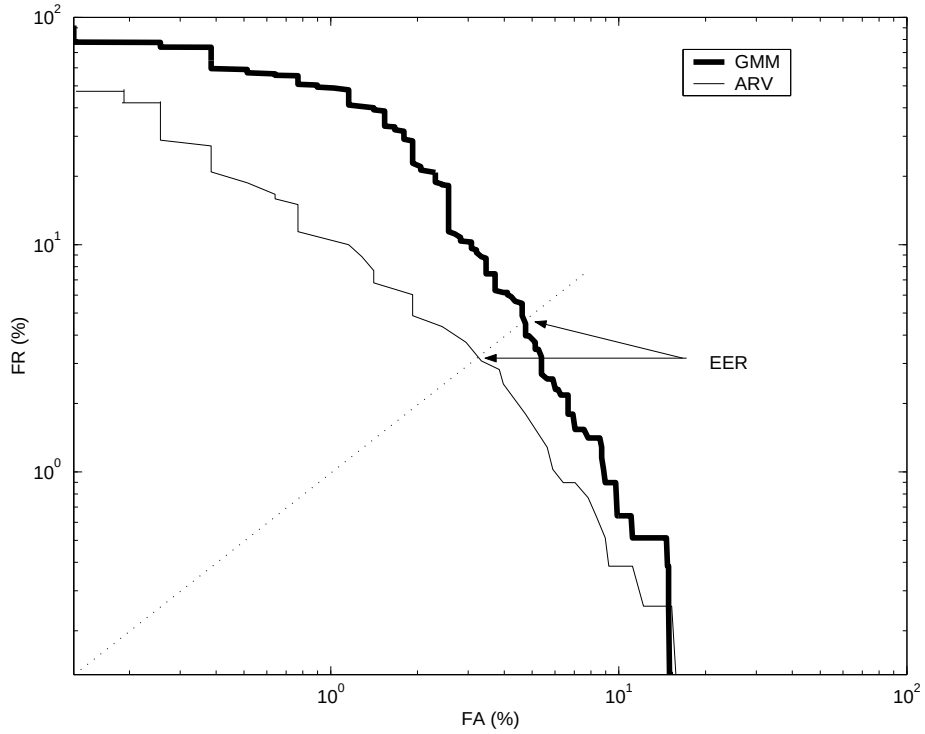

FIG. 5.9: Curva DET para o GMM com 32 gaussianas e AR-Vetorial de ordem 2 (distância simétrica) para um tempo de treinamento e teste de 10s.

Para traçar a curva DET, as coordenadas (FA,FR), são obtidas variando-se o limiar de decisão do sistema de verificação. O limiar é variado no sentido de obter-se o erro de FR igual a zero e depois em sentido contrário até obter-se o erro de FA igual a zero, ou vice-versa. Com base nos valores de erros encontrados nessa variação de limiar, traça-se a curva DET em um gráfico com eixos em escala logarítmica. Se a distribuição dos erros for gaussiana, a curva DET se aproximará de uma reta (MARTIN, 1997).

A distância entre as curvas representa de forma mais clara o desempenho entre dois ou mais sistemas de verificação de locutor. Na FIG. 5.9 pode-se ver o desempenho do GMM com 32 gaussianas versus o do AR-Vetorial de ordem 2, usando a distância simétrica. Nessa figura ambos sistemas usam MCC15 com 10s de treinamento e teste. Deste gráfico percebe-se a superioridade de desempenho do AR-Vetorial para o tempo de 10s de treinamento e teste. Quando os erros de FR e FA são iguais tem-se o EER, dado pelo encontro da linha tracejada com as curvas DET do GMM e AR-Vetorial. Notar que este foi o ponto (aproximado) analisado nas seções anteriores.

A curva DET não é adequada para análise de sistemas com desempenhos muito bons, pois existirão poucas coordenadas (FA, FR) para gerar as curvas. Esse é o caso do GMM (32G) e AR-Vetorial para 60s de treinamento e 30s de teste. A curva DET é muito utilizada quando os desempenhos dos sistemas de verificação de locutor são ruins, como é o caso

quando se analisa sistemas corrompidos por ruído.

Dependendo do sistema em questão, pode-se escolher o ponto de operação do sistema (limiar de decisão) de acordo com os erros desejados de FR e FA. A curva DET pode também ser obtida utilizando-se custos para os erros, como na EQ. 5.2, o que produzirá diferentes inclinações das curvas.

5.9 RESUMO E CONCLUSÃO

Neste capítulo foram apresentados os desempenhos do GMM e do AR-Vetorial, com a variação dos seus parâmetros, para a tarefa de verificação de locutor independente do texto.

Para o GMM, com um *background* de 10 locutores, chegou-se às seguintes conclusões:

- O uso de *background* é fundamental para o bom desempenho do sistema, e aparentemente o aumento do número de locutores e tempo de voz para background também melhoram o desempenho.
- Quanto maior o tempo de treinamento, melhor o desempenho do sistema, devido à maior quantidade de dados para treinamento, o que acarreta um modelagem mais precisa.
- Quanto maior o tempo de teste, melhor o desempenho do sistema, pois mais informação do locutor é disponibilizada para comparação com o modelo armazenado.
- Quanto maior o número de gaussianas, melhor é o desempenho, pois consegue-se um modelagem mais precisa das classes acústicas de cada locutor, melhorando assim a representação do seu espaço acústico.
- O uso do PCA melhora o desempenho do sistema, principalmente para um tempo de treinamento de 30s. Isto sugere que a matriz de transformação estimada nos dados de treinamento contém informação da estrutura espacial dos dados (autovetores da matriz covariância), que possibilita o aumentando da verossimilhança dos locutores verdadeiros e a diminuição dos falsos. O uso do PCA também gera uma matriz de covariância diagonal para o treinamento, que é utilizada no modelo.
- Quando o tempo de treinamento cai para 10s, o desempenho do GMM torna-se muito baixo e o aumento do número de gaussianas bem como do tempo de teste, com ou sem PCA, praticamente não afeta o desempenho do sistema.

- A compressão de 3 MCC pelo PCA produz resultados inferiores ao uso do mesmo número de MCC sem compressão. A combinação linear dos coeficientes mel-cepestrais na compressão acabou reduzindo a informação referente à identidade do locutor.

O GMM com um *background* de 16 locutores, com 32 gaussianas e um tempo de treinamento de 60s, apresentou melhores resultados que o GMM com as mesmas configurações e um *background* de 10 locutores. O uso de PCA para o GMM de 16 locutores de *background*, nessa configuração, não acarretou melhora de desempenho.

Para o AR-Vetorial, as conclusões são as seguintes:

- A ordem do modelo influi pouco no resultado, a menos da ordem 1. Portanto, o modelo de ordem 2 é o mais indicado, por ser mais simples e fornecer bons resultados.
- A distância utilizada pode ser a simétrica (5) ou a distância mais simples (1), ambas fornecendo bons resultados. A distância 7 também poderia ser utilizada, mas é a distância com maior custo computacional e seu desempenho fica aquém das distâncias 1 e 5.
- Um tempo de teste muito pequeno não deve ser usado, pois não se consegue uma modelagem precisa no AR-Vetorial.
- O uso do *background*, como no GMM, é insatisfatório e não faz sentido, pois é utilizada uma distância na avaliação, o que difere grandemente da subtração de verossimilhança do GMM.
- O PCA não mostrou-se uma técnica viável para ser usada junto com o AR-Vetorial, em geral, não resultou em acréscimo de desempenho significativo, aumentando a carga computacional.
- Como o AR-Vetorial é um modelo estimado a partir da autocorrelação de vetores de características, o comprimento destes afeta a modelagem.
- O modelo de ordem dois com MCC15 apresentou o mesmo desempenho para 60 e 30s de treinamento e testado com 30s, demonstrando que com 30s de voz o AR-Vetorial já consegue capturar informações suficientes do locutor.
- Para modelos de ordens maiores, o uso de maiores tempos de teste e treino produzem melhores resultados.

O AR-Vetorial se baseia na extração da informação espectral dinâmica do sinal, enquanto o GMM baseia-se na estatística presente no sinal, realizando uma modelagem estática do sinal. Comparando o GMM (10 locutores de *background*) com o AR-Vetorial, o desempenho do GMM sempre foi superior ao AR-Vetorial para 3s de teste, concluiu-se então, para os tempos de 30 e 10s de teste, que:

- O AR-Vetorial produz resultados iguais ou superiores ao GMM para o mesmos tempos de teste e treinamento.
- O AR-Vetorial captura a informação dinâmica do sinal, e o GMM a estática.
- O PCA produz resultados satisfatórios no GMM, ao contrário do AR-Vetorial.
- A ordem do modelo afeta grandemente o GMM, ao contrário do AR-Vetorial de ordens maiores que 1.
- O GMM necessita do modelo de *background* ao contrário do AR-Vetorial.
- O GMM é treinado, e os vetores de características de teste são passados pelo modelo treinado, gerando verossimilhanças. No AR-Vetorial os dados de treinamento e teste são modelados, e após se utiliza uma medida de distância (Itakura, por exemplo), para avaliar a similaridade dos modelos.
- O GMM tem um custo de processamento computacional muito maior que o AR-Vetorial no treinamento. No teste o AR-Vetorial perde, devido à necessidade da obtenção do modelo de teste.

Na escolha de um sistema de verificação de locutor independente do texto, dos resultados obtidos com o GMM e AR-Vetorial (ordem 2 e distância 5), pode-se optar por:

- Melhor desempenho com menor carga computacional (sem erros): AR-vetorial com MCC, para 30s de treinamento e teste.
- Melhor desempenho para os menores tempos de teste (10 e 3s): GMM 32 Gaussianas com MCCPCA, para 60 e 30s de treinamento, com erros na faixa de 0,5% a 3%. Ainda para 60s de treinamento, o GMM com 16 gaussianas com MCC ou MCCPCA produziu erros na faixa de 0,6% a 2%, superando o AR-Vetorial.
- Melhor desempenho com o menor tempo de treinamento e teste (FR=2,4% e FA=3,8%): AR-Vetorial com MCC para 10s de treinamento e teste.

6 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

6.1 CONCLUSÕES

Nesta dissertação foram avaliados dois sistemas de classificação para a verificação de locutor independente do texto: o GMM e o AR-Vetorial. Foram considerados os papéis representados por diferentes tempos de treinamento e teste sobre o desempenho dos referidos sistemas. O treinamento foi realizado com 60, 30 e 10s e os testes com 30, 10 e 3s de voz. Esses tempos foram avaliados para variações do número de gaussianas do GMM (32, 16 e 8) e da ordem do modelo para o AR-Vetorial (1, 2, 3, 4 e 5).

Na tarefa de verificação de locutor, foi empregada uma característica espectral que leva em conta a percepção auditiva: o mel-cepestral. Na tentativa de melhorar o poder discriminativo desta característica, foi utilizada uma transformação ortogonal sobre o conjunto de características espectrais: o PCA. O efeito do emprego dessa transformação sobre o desempenho dos diversos sistemas foi analisado de forma criteriosa.

Na avaliação do GMM verificou-se a dependência de desempenho associada ao número de gaussianas utilizadas. Quanto maior é este número (32), mais precisa é a modelagem do espaço acústico do locutor em questão, dado por suas características. A dependência de desempenho também é notada, quando se varia a quantidade de voz disponível para testar e treinar o sistema, ou seja, quanto mais dados existirem para treinamento, melhor será a modelagem do GMM, e quanto mais dados existirem para testes, mais informação que possa caracterizar a identidade do locutor será disponível. O GMM não apresentou erros com 32 gaussianas, 60s de treinamento e 30s de teste. Já com 3s de teste nesta configuração, os erros chegaram a aproximadamente 1,3 %. Com 30s de treinamento os erros variaram entre 1 % (32 gaussianas) e 6% (8 gaussianas). Esses resultados mostram a dependência entre o número de gaussianas e o tempo de treinamento e teste. Quanto maiores estes forem melhor será o desempenho do sistema. Estas conclusões não são válidas, entretanto, nos resultados encontrados para o tempo de treinamento de 10s, devido à falta de estatística para treinar o GMM. Os erros encontrados com 10s de treinamento estiveram na faixa de 6 % a 7 % independente do número de gaussianas empregado.

O teste aumentando o número de locutores de *background* de 10 para 16, com o GMM de 32 gaussianas e 60s de treinamento, demonstrou uma melhora de desempenho usando o MCC e o MCCPCA. O MCCPCA para 3s de teste com 16 locutores de background

apresentou EER 0,03 % inferior ao MCC. No entanto, para comprovação destes resultados novos experimentos com maior estatística precisariam ser efetuados (maiores tempo de treinamento, teste e número de locutores).

O uso do PCA nos coeficientes mel-cepestrais forneceu uma melhora de desempenho no GMM. Isso se deve aos seguintes fatos:

- Para simplificações no modelo utiliza-se a diagonal principal da matriz covariância dos dados de treinamento, o restante da informação contida nessa matriz é desprezada. O PCA provoca uma descorrelação dos dados transformados, fazendo com que a matriz covariância estimada nesses dados seja diagonal, o que não produz perda de informação no treinamento do GMM.

- Como o PCA é estimado sobre a matriz covariância dos dados de treinamento, ele carrega informação estrutural desses dados (autovetores da matriz covariância). Assim, para um locutor verdadeiro a matriz de transformação o projetará no espaço modelado, correspondente ao seu modelo. Quando o locutor for falso a transformação projetará os dados em um espaço diferente do modelado, reduzindo a verossimilhança do sistema de verificação.

O aumento do desempenho do GMM com o MCCPCA em relação ao MCC foi mais notável para 30s de treinamento, chegando a superar em aproximadamente 2 vezes o GMM (MCC) para 16 e 8 gaussianas com 30 e 10s de teste. Com 10s de treinamento o desempenho do GMM é bastante afetado e o uso de PCA não traz melhoras.

No AR-Vetorial avaliou-se a melhor distância de Itakura para comparação entre dois modelos. Dentre as sete distâncias analisadas os melhores resultados foram obtidos com as distâncias 1, 5 e 7, sendo a distância 5 (simétrica) usada para os demais experimentos apresentados, pois foi um pouco superior às demais. O desempenho do sistema também foi avaliado para variações da ordem do modelo de 1 até 5. Como era de se esperar, os piores resultados foram obtidos com a menor ordem. As demais ordens tiveram desempenho comparáveis, sendo que o modelo de ordem dois, foi usado devido à sua menor complexidade computacional. Os tempos de treinamento e testes foram variados, da mesma forma que no GMM. Verificou-se que os melhores desempenhos foram alcançados para os maiores tempos, pois mais precisa foi a estimação dos modelos. Para 60 e 30s de treinamento com 30s de teste o AR-Vetorial não apresentou erros. Com 3s de teste, independente do tempo de treinamento, as estimativas dos modelos foram imprecisas e os erros obtidos grandes, na faixa de 10 %.

O uso do PCA no sistema AR-Vetorial não acarretou acréscimo de desempenho significativo, com resultados até inferiores ao sistema sem PCA para o caso de 10s de treinamento. Uma possível explicação para isso é que o uso do PCA descorrelaciona os coeficientes dentro de cada vetor o que, à princípio, não altera a relação temporal entre os vetores, sobre a qual o AR-Vetorial se baseia.

Comparando o desempenho do GMM e do AR-Vetorial, verificou-se que ambos os sistemas não resultaram em erros para 60s de treinamento e 30s teste (GMM 32 gaussianas e AR-Vetorial(ordem 2 e distância 5)). O melhor desempenho com a menor carga computacional foi conseguido pelo AR-Vetorial, usando o MCC, para 30s de treinamento e teste (sem erros). O melhor desempenho com o menor tempo de teste foi conseguido pelo GMM com 32 gaussianas e MCCPCA (60s de treinamento e 3s de teste), erros próximos a 1,2 %. Para o menor tempo de treinamento (10s) com 10s de teste o AR-Vetorial de ordem 2 com distância de Itakura simétrica obteve erros de 3,1 % superando o GMM que forneceu erros próximos a 4,4 %. Para os tempos de treinamento de 60 e 30s, testados com 10 e 3s, o GMM com 32 gaussianas apresentou desempenho superior ao AR-Vetorial, com erros na faixa de 0,5 % à 3 %, independente de utilizar-se o MCC ou o MCCPCA. Essa superioridade se mantém para o GMM com 16 gaussianas treinado com 60s de voz (erros na faixa de 0,6% a 2%).

O uso de um número menor de coeficentes mel-cepestrais resultou em um decréscimo de desempenho de ambos os sistemas de classificação. No AR-Vetorial este decréscimo esteve em torno de 0,4 % para 10s de teste (60, 30 e 10s de treinamento) com o mesmo decréscimo para 30s de teste com 30s de treinamento. O GMM treinado com 60s, apresentou piora no desempenho em torno de 0,2 % para 30s de teste e 0,4 % para 3s de teste.

A complexidade computacional, avaliada em função do tempo de processamento, mostrou que o GMM consome um tempo muito maior de treinamento que o AR-Vetorial de ordem 2 (em torno de 60 vezes maior com 32 gaussianas, para 60s de treinamento). Por outro lado, no teste, o AR-Vetorial é mais lento que o GMM, aproximadamente 1,5 vezes mais lento que um GMM de 32 gaussianas para 30s de teste. Isto é devido à necessidade da obtenção dos coeficientes do AR-Vetorial nos dados de teste para o uso da distância de Itakura.

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

A partir dos resultados obtidos neste trabalho, fica óbvia a necessidade de avaliar um sistema híbrido entre o GMM e o AR-Vetorial. Tal sistema pode fazer uso de uma Rede

Neural para pesar a saída dos sistemas, fazendo desta forma uma ponderação que pode ou não ser linear. A união dos dois sistemas de classificação terá a vantagem de definir um modelo dinâmico, que carrega tanto informações da evolução temporal da fala, como informações estatísticas que modelam o espaço acústico produzido por um determinado locutor. Acredita-se que tal sistema híbrido possa fornecer melhores resultados.

Existe a necessidade de executar testes com sinal de voz corrompido por um canal e por ruído aditivo, para avaliar o desempenho dos diversos sistemas de verificação de locutor em condições mais práticas e reais, bem como avaliar o papel do PCA na tentativa de melhorar as caraterísticas de voz corrompidas.

Outras transformações, como o LDA (*Linear Discriminant Analysis*- análise discriminante linear) (MALAYATH, 2000) e transformações não lineares (MAO, 1995) como o NPCA (PCA não linear) (NANDAKISHORE, 1996) e NLDA (LDA não linear) (YOCHAI, 1998), podem constituir ferramentas úteis à tarefa de verificação de locutor.

Uma pesquisa mais profunda quanto ao aumento do número de locutores e o tempo de voz necessários para treinar o UBM ainda necessita ser conduzida. Os resultados preliminares apresentados indicam a possibilidade de melhoria de desempenho.

6.3 COMENTÁRIOS FINAIS

As pesquisas na área de reconhecimento de locutor tem aumentado consideravelmente nos últimos anos, tendo em vista a crescente demanda por sistemas dessa natureza. As pesquisas estão focadas para resolver os problemas produzidos pelo ruído no sinal de voz, os quais prejudicam muito o desempenho dos sistemas de classificação. A avaliação de tais sistemas tem sido feita com o uso da curva DET.

As pesquisas no Brasil nessa área são reduzidas quando comparadas com o restante do mundo. Existe ainda a necessidade de um banco de dados de voz brasileiro, como por exemplo o TIMIT disponível para a língua inglesa (CAMPBELL, 1999).

Nesta dissertação foram abordados os sistemas de reconhecimento de locutor mais empregados na atualidade. Este trabalho é a base para trabalhos futuros mais amplos, que possam utilizar um conjunto maior de locutores, com pesquisas voltadas para aplicações em ambientes encontrados na prática.

7 BIBLIOGRAFIA

ALCAIM, Abraham, José Alberto Solewicz e João Antônio de Moraes. **Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro**. Revista da Sociedade Brasileira de Telecomunicações, v. 7, n. 1, dez. 1992.

ATAL, Bishnu S. **Automatic Recognition of Speakers from Their Voices**. Proceedings of the IEEE, v. 64, n. 4, p. 460-475, Apr. 1976.

AVENDAÑO, Carlos. **Temporal Processing of Speech in a Time-Feature Space**. 1997. Tese (Doctor of Philosophy) - Oregon Graduate Institute, 1997.

AUCKENTHALER, Roland, Michael Carey, and Harvey Loyd-Thomas. **Score Normalization for Text-Independent Speaker Verification System**. Digital Signal Processing, v. 10, p. 42-45, 2000.

BEZERRA, Marconi dos Reis. **Reconhecimento Automático de Locutor para Fins Forenses, Utilizando Técnicas de Redes Neurais**. 1994. Dissertação (Mestrado em Ciências) - Instituto Militar de Engenharia, 1994.

BIMBOT, F., L. Mathan, A. de Lima, and G. Chollet. **Standard and Target Driven AR-vector Models for Speech Analysis and Speaker Recognition**. Proceeding of ICASSP, San Francisco, USA, v. 2, p. II5-II8, Mar. 1992.

BRAGA, Antônio de Pádua, André P. L. de Carvalho, e Teresa B. Lurdemir. **Fundamentos de Redes Neurais Artificiais**. Rio de Janeiro, 11º Escola de Computação, DCC/IM, COPPE/Sistemas, NCE/UFRJ, 1998.

CAMPBELL, Joseph P., Jr. **Speaker Recognition: A Tutorial**. Proceedings of IEEE, v. 85, n. 9, p. 1437-1462, Sept. 1997.

CAMPBELL, Joseph P., Jr., and Douglas A. Reynolds. **Corpora For The Evaluation of Speaker Recognition Systems**. Proceedings of ICASSP, Phoenix, USA, p. 829-832, May 1999.

CHAGNOLLEU, Ivan Magrin, and Geoffrey Durou. **Application of Time-Frequency Principal Componente Anaysis to Speaker Verification**. Digital Signal Processing, v. 10, p. 226-236, 2000.

CHAGNOLLEU, Ivan Magrin, Joachim Wilke, and Frédéric Bimbot. **A Further Investigation on AR-vector Models For Text-Independent Speaker Identification**. Proceeding of ICASSP, Atlanta, USA, v. 1, p. 101-104, May 1996.

DAVIS, Steven B., and Paul Mermelstein. **Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences**. IEEE Transactions on Acoustics, Speech, and Signal Processing, v. ASSP-28, n. 4, Aug. 1980.

DELACRÉTAZ, Dijana Petrovska, et al. **Segmental Approaches for Automatic Speaker Verification**. Digital Signal Processing, v. 10, p. 198-212, 2000.

DELLER, John R., Jr., John G. Proakis, John H. L. Hansen. **Discrete-Time Processing of Speech Signals**. New Jersey: Prentice Hall, 1993.

FANT, G. **Acoustic Theory of Speech Production**. Gravenhage, The Netherland: Mouton and Co., 1960.

FERREIRA, Aurélio Buarque de Olanda. **Miniaurélio Século XXI: O minidicionário da língua portuguesa**. 4. ed. Rio de Janeiro: Nova Fronteira, 2000.

FLOCH, J. L. Le, C. Montacié, and M. J. Caraty. **GMM and ARVM Cooperation and Competition for Text-independent Speaker Recognition on Telephone Speech**. Proceedings of the ICSLP, v. 4, p. 2411-2414, 1996.

FREDOUILLE, Corinne, Jean-François Bonastre, and Teva Merlin. **AMIRAL: A Block-Segmental Multirecognizer Architecture for Automatic Speaker Recognition**. Digital Signal Processing, v. 10, p. 172-197, 2000.

FUKUNAGA, Keinosuke. **Introduction to Statistical Pattern Recognition**. 2. ed. USA: Morgan Kaufmann, 1990.

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

GARCIA, Alberto Leon. **Probability and Random Processes for Electrical Engineering**. 2. ed. USA: Addison-Wesley Publishing Company, 1994.

GRAVIER, Guillaume, Jamal Kharroubi, and Gérard Chollet. **On the Use of Prior Knowledge in Normalization Schemes for Speaker Verification**. Digital Signal Processing, v. 10, p. 213-225, 2000.

HADJITODOROV, S., B. Boyanov, T. Ivanov, and N. Dalakchieva. **Text-independent Speaker Identification Using Neural Nets and AR-vector Models**. Electronics Letters, v. 30, n. 11, May 1994.

HAYKIN, Simon. **Adaptive Filter Theory**. 3. ed. New Jersey: Prentice Hall, 1996.

HERMANSKY, Hynek. **Perceptual Linear Predictive(PLP) Analysis of Speech**. Journal of Acoustic Society of America, 87(4), Apr. 1990.

HIGGINS, L. Bahler, J. Porter. **Speaker Verification using Randomized Phrase Prompting**. Digital Signal Processing, v. 1, p. 89-106, 1991.

HSU, Hwei. **Theory and Problems of Probability, Random variables, and Random Processes**. USA: McGraw-Hill, Schaum Outline, 1996.

ITAKURA, Fumitada. **Minimum Prediction Residual Principle Applied to Speech Recognition**. IEEE Transactions on Acoustics, Speech, and Signal Processing, v. ASSP-23, n. 1, Fev. 1975.

JAYANT, M. Naik. **Speaker Verification: A Tutorial**. IEEE Communications Magazine, p. 42-47, Jan. 1990.

YOCHAI, Honig, et al. **Nonlinear Discriminant Features Extraction for Robust Text-Independent Speaker Recognition**. Proceedings of RLA2C, Avignon, França, 1998.

LIMA, Carlos Henrique da Rocha. **Gramática Normativa da Língua Portuguesa**. 32 ed. Rio de Janeiro, RJ: José Olímpio, 1994.

LINDE, Yoseph, Andrés Buzo, and Robert M. Gray. **An Algorithm for Vector Quantizer Desing**. IEEE Transactions on Communications, v. COM-28, n. 1, Jan. 1980.

LIU, Li, and Jialong He. **On The Use of Orthogonal GMM in Speaker Recognition**. Proceedings of ICASSP, Phoenix, USA, May 1999.

MAKHOUL, John. **Linear Prediction: A Tutorial Review**. Proceedings of IEEE, v. 63, p. 561-580, Apr. 1975.

MALAYATH, Narendranath. **Data-driven Methods for Extracting Features From Speech**. 2000. Thesis (Doctor of Philosophy) - Oregon Graduate Institute, 2000.

MALAYATH, Narendranath, Hynek Hermansky, et al. **Data-driven Temporal Filters and Alternatives to GMM in Speaker Verification**. Digital Signal Processing, v. 10, p. 55-74, 2000.

MAO, Jianchang, and Anil K. Jain. **Artificial Neural Networks for Feature Extraction and Multivariate Data Projection**. IEEE Transactions on Neural Networks, v. 6, n. 2, p. 296-317, Mar. 1995.

MAMMONE, Richard J., Xiaoyu Zhang, and Ravi P. Ramachandran. **Robust Speaker Recognition**. IEEE Signal Processing Magazine, p. 58-71, 1996.

MARTIN, Alvin, et al. **The DET Curve in Assessment of Detection Task Performance**. In Proceedings of EuroSpeech 97, v. 4, p. 1895-1898, 1997.

MARTIN, Alvin, and Mark Przybocki. **The NIST 1999 Speaker Recognition Evaluation - An Overview**. Digital Signal Processing, v. 10, p. 1-18, 2000.

MONTACIÉ, Claude, and Jean-Luc Le Floch. **AR-vector Models for Free-text Speaker Recognition**. Proceedings of the ICSLP, Banff, Canada, p. 611-614, 1992.

NAKASONE, Hirotaka and Steven D. Beck. **Forensic Automatic Speaker Recognition**. 2001 A Speaker Odyssey: The Speaker Recognition workshop. Creta, Grécia, Jun. 18-22, 2001.

NANDAKISHORE, Kambhatla. **Local Model and Gaussian Mixture Models for Statistical Data Processing**. 1996. Thesis (Doctor of Philosophy) - Oregon Institute of Science and Tecnology, 1996.

NIELSEN, Astrid Schmidt and Thomas H. Crystal. **Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data**. Digital Signal Processing, v. 10, p. 249-266, 2000.

NIST **Introduction to the Issue**. Digital Signal Processing, v. 10, p. XI-XV, p. 2000.

OPPENHEIM, Alan V., Ronald W. Schafer com John R. Buck. **Discrete-Time Signal Processing**. 2. ed. New Jersey: Prentice-Hall, 1998.

PICONE, Joseph W. **Signal Modeling Techniques in Speech Recognition**. Proceedings of IEEE, v. 81, n. 9, p. 1215-1247, Sept. 1991.

RABINER, Lawrence, and M. Samur. **The Bell System Technical Journal** . Feb. 1974.

RABINER, Lawrence, and Ronald Shafer. **Digital Signal Processing of Speech Signals** . USA: Prentice Hall Inc., 1978.

RABINER, Lawrence. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**. Proceedings of The IEEE, v. 77, n. 2, Fev. 1989.

RABINER, Lawrence, and Biing-Hwang Juang. **Fundamentals of Speech Recognition**. 1. ed. New Jersey: Prentice-Hall, 1993.

REYNOLDS, Douglas A. **A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification**. 1992. Tese (Doctor of Philosophy) - Georgia Institute of Techonology, 1992.

REYNOLDS, Douglas A. **Robust Text-Idenpendent Speaker Identification Using Gaussian Mixture Speaker Model**. IEEE Transaction on Speech and Audio Processing, v. 3, n. 1, p. 72-83, Jan. 1995.

REYNOLDS, Douglas A. **Speaker Identification and Verification Using Gaussian Mixture Speaker Models**. Speech Communication, v. 17, p. 91-108, 1995.

REYNOLDS, Douglas A. Thomas F. Quatieri, and Robert B. Dunn. **Speaker Verification Using Adapted Gaussian Mixture Models**. Digital Signal Processing, v. 10, p. 19-41, 2000.

REYNOLDS, Douglas A. **Experimental Evaluation of Features for Robust Speaker Identification**. IEEE Transactions on Speech and Audio Processing, v. 2, n. 4, p. 639-643, Oct. 1994.

SANTOS, Sidney Cerqueira Bispo dos. **Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos**. 1997. Tese (Doutorado em Ciências) - Pontífica Universidade Católica do Rio de Janeiro, 1997.

SARMA, Sridevi Vedula. **A Segment-Based Speaker Verification System Using SUMMIT**. 1999. Dissertation (Master of Science) - Massachusetts Institute of Technology, 1999.

SAVIC, M., J. Sorensen. **Phoneme Based Speaker Verification**. Transactions of IEEE, p. II-165-II-168, 1992.

SHARMA, Sangita R. **Multi-Stream Aprroach to Robust Speech Recognition**. 1999. Tese (Doctor of Philosophy) - Oregon Graduate Institute of Science and Tecnology, 1999.

SOUSA, Ricardo Honório Guedes de. **Estudo de Características Relevantes do Sinal de Voz para o Reconhecimento Automático do Locutor Desprevenido, Independente ao Texto**. 1996. Dissertação (Mestrado em Ciências) - Instituto Militar de Engenharia, 1996.

STEVENS, S. S., and J. Volkman. **The Relation of Pitch to Frequency**. Journal of Psychology, v. 53, p. 329, 1940.

STRANG, Gilbert. **Linear Algebra and Its Applications**. 3. ed. USA: Harcourt College Publishers, 1988.

THEODORIDIS, Sergios, Konstantinos Koutroumbas. **Pattern Recognition**. San Diego: Academic Press, 1999.

VUUREN, Van Sarel. **Speaker Verification in a Time-Feature Space**. 1999. Tese (Doctor of Philosophy) - Oregon Graduate Institute of Science and Tecnology, 1999.

WOODLAND, Phil. **Speech Recognition**. The Institution Of Electrical Engineers, IEE, Savoy Place, London, WC2R 0BL, UK, 1998.

8 APÊNDICES

8.1 APÊNDICE 1: ESTIMAÇÃO DOS PARÂMETROS DO GMM

Neste apêndice é descrito um procedimento para estimação da máxima verossimilhança dos parâmetros das densidades de probabilidade gaussianas de um conjunto de misturas (GMM), para um conjunto de observações. Este procedimento foi baseado em REYNOLDS (1992).

A técnica de Máxima Expectativa (*Expectation-Maximization* - EM) aqui apresentada ajusta os parâmetros do conjunto de densidades de probabilidade, maximizando a função de verossimilhança. São também consideradas questões relacionadas à execução do algoritmo EM no treinamento do GMM para o reconhecimento de locutor.

8.1.1 ESTIMAÇÃO DA MÁXIMA VEROSSIMILHANÇA

A estimação dos parâmetros de máxima verossimilhança é um método geral e poderoso para estimar os parâmetros de um conjunto de observações de um processo estocástico. Esta estimação busca encontrar um modelo λ com maior probabilidade de ter produzido o conjunto de observações. Ela determina o conjunto de parâmetros do modelo que maximizam a função de verossimilhança de um conjunto de observações. Para um conjunto de observações referidos coletivamente por $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, a função de verossimilhança para o modelo λ é definida como a função densidade de probabilidade conjunta de X dado o modelo λ, referida como $p(X|\lambda)$. A condição necessária para a estimação dos parâmetros de máxima verossimilhança satisfaz a equação de verossimilhança[1]:

$$\frac{\partial p(X|\lambda)}{\partial \lambda} = 0 \tag{8.1}$$

A estimativa dos parâmetros de máxima expectativa tem algumas propriedades interessantes, tais como, consistência assintótica e eficiência. Isto significa que, dado um conjunto suficiente de vetores de treinamento, o modelo estimado convergirá para os parâmetros do modelo verdadeiro com probabilidade um. Infelizmente, resolver a EQ. 8.1 diretamente dos parâmetros do GMM, $\lambda = \{p_i, \vec{\mu}_i, K_i\}$, $i = 1, \ldots, M$, resulta em um conjunto não fechado de soluções.

A estimação dos parâmteros de máxima verossimilhança do GMM pode ser realizada através de um processo iterativo de estimação, o qual é um caso especial do algoritmo de Máxima Expectativa (EM). O amplo uso do algoritmo EM é devido ao fato de que

[1] A derivada na EQ. 8.1 é a notação do gradiente de $p(X|\lambda)$ em relação ao vetor de parâmetros que caracterizam λ. Esta equação representa um conjunto simultâneo de equações, uma para cada elemento de λ

ele garante um função de verossimilhança não decrescente, após cada iteração, provendo uma maneira poderosa de lidar com problemas de estimação complicados. A idéia básica do algoritmo EM é iniciar com um modelo λ, para estimar um novo modelo $\overline{\lambda}$, tal que $p(X|\overline{\lambda}) \geq p(X|\lambda)$. O modelo estimado torna-se, então, o novo modelo e o processo é repetido até que algum limiar de convergência seja alcançado.

8.1.2 FUNÇÃO AUXILIAR

Sejam M classes acústicas que podem ocorrer no tempo, i_t ($i_t \in [1, M]$), formando um conjunto de variáveis aleatórias discretas $I = \{i_1, \ldots, i_T\}$, a probabilidade de uma determinada classe é dada por:

$$P(i_t = i) = p_i \tag{8.2}$$

No GMM, p_i é a ponderação das gaussianas, ou seja, a probabilidade da ocorrência das mesmas.

Seja $\vec{x}_t$ um vetor de observação em t produzido por I, formando um conjunto de vetores aleatórios contínuos $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, a fdp condicional de $\vec{x}_t$ dado uma classe i é, então:

$$p(\vec{x}_t|i_t = i) = b_i(\vec{x}_t) \tag{8.3}$$

que representa a fdp da gaussiana correspondente a classe i no GMM. Assim, a fdp de $\vec{x}_t$ é dada por:

$$p(\vec{x}_t) = \sum_{i=1}^{M} p(\vec{x}_t|i_t = i)P(i_t = i) \tag{8.4}$$

que é equação que descreve o GMM dado um modelo λ.

A fdp de X dado um modelo λ, é então:

$$p(X|\lambda) = \sum_{I} p(X|I, \lambda)P(I|\lambda) \tag{8.5}$$

Considerando que os vetores $\vec{x}_t$ e as classes i são estatisticamente independentes e que a fdp de um vetor aleatório no instante t dadas todas as classes nos instantes $1, ..., T$ depende apenas da classe no instante t (i_t), resulta em

$$p(X|I, \lambda) = \prod_{t=1}^{T} p(\vec{x}_t|i_t, \lambda) \tag{8.6}$$

Do que foi visto, chega-se à:

$$p(X, I|\lambda) = p(X|I, \lambda)P(I|\lambda)$$

onde

$$p(X, I|\lambda) = \prod_{t=1}^{T} p_{i_t} b_i(\vec{x}_t) \tag{8.7}$$

O que REYNOLDS (1992) chama de $p(X, I|\lambda)$ não é uma fdp, pois:

$$\int_X p(X, I|\lambda) dX \neq 1 = P(I|\lambda)$$

Porém, nesta dissertação, usaremos a notação dada na EQ. 8.7.

O objetivo é encontrar um novo conjunto do modelo de parâmetros $\overline{\lambda}$, o qual aumenta sua função de verossimilhança para um dado modelo λ. Matematicamente, dado λ encontra-se $\overline{\lambda}$, tal que:

$$p(X|\overline{\lambda}) \geq p(X|\lambda) \tag{8.8}$$

A maximização é alcançada usando a função auxiliar a seguir. A motivação para utilizar esta função é que ela permitem um meio iterativo de aumentar a função de verossimilhança de dados observados, maximizando uma função dos dados completos[2]. A função:

$$Q(\lambda, \overline{\lambda}) = \sum_I p(X, I|\lambda) \log p(X, I|\overline{\lambda}) \tag{8.9}$$

exibe a propriedade de que maximizar $Q(\lambda, \overline{\lambda})$, resulta em $p(X|\overline{\lambda}) \geq p(X|\lambda)$ (THEODORIDIS, 1999).

O próximo passo é manipular a função auxiliar para uma forma que explicite os parâmetros do modelo de interesse. Substituindo a EQ. 8.7 (com $\lambda = \overline{\lambda}$) na EQ. 8.9 resulta em:

$$Q(\lambda, \overline{\lambda}) = \sum_I p(X, I|\lambda) \sum_{t=1}^{T} \log[\overline{p}_{i_t} \overline{b}_{i_t}(\vec{x}_t)] \tag{8.10}$$

onde $\overline{p}_{i_t}$ é a nova ponderação da mistura e $\overline{b}_{i_t}(\vec{x}_t)$ é a densidade componente usando novos parâmetros de média e covariância do modelo.

[2] O termo dados completos é freqüentemente usado para fazer referência a ambos dados observáveis e escondidos. Neste caso, os dados observáveis são X e os escondidos são I.

Para elimiar a dependência dos novos parâmetros do modelo sobre as variáveis de estado escondidas i_t, define-se uma função de contagem:

$$\eta_t(i, I) = \begin{cases} 1 & i_t = i \\ 0 & \text{caso contrário} \end{cases} \tag{8.11}$$

a qual age como uma função delta de Kronecker, agindo somente quando a seqüência de estados I contém o estado i no tempo t. Quando usado na EQ. 8.10, resulta:

$$Q(\lambda, \overline{\lambda}) = \sum_{t=1}^{T} \sum_{I} p(X, I|\lambda) \sum_{i=1}^{M} \log[\overline{p}_i \overline{b}_i(\vec{x}_t)] \eta_t(i, I) \tag{8.12}$$

Reagrupando os termos, a forma final da função auxiliar resulta em

$$Q(\lambda, \overline{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{M} \log[\overline{p}_i \overline{b}_i(\vec{x}_t)] \gamma_t(i) \tag{8.13}$$

onde

$$\gamma_t(i) = \sum_{I} \eta_t(i, I) p(X, I|\lambda) \tag{8.14}$$

Além disso, pode ser mostrado (REYNOLDS, 1992) que:

$$\gamma_t(i) = p(X|\lambda) P(i_t = i|\vec{x}_t, \lambda) \tag{8.15}$$

onde

$$P(i_t = i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{8.16}$$

é a probabilidade a *posteriori* do estado i.

8.1.3 EQUAÇÕES PARA A ESTIMAÇÃO DOS PARÂMETROS

As equações que aumentam a função de verossimilhança são obtidas agora maximizando a EQ. 8.13 com respeito a cada novo modelo de parâmetros $\overline{\lambda} = \{\overline{p}_i, \overline{\vec{\mu}}_i, \overline{K_i}\}$. Uma vez que, $Q(\lambda, \overline{\lambda})$ é uma função estritamente côncava dos parâmetros de interesse, a maximização pode ser feita encontrando funções de valores críticos, isto é, encontrando os parâmetros do modelo $\overline{\lambda}$, tais que, $\partial Q(\lambda, \overline{\lambda})/\partial \overline{\lambda} = 0$.

Ponderação da Mistura

A ponderação da mistura é obtida maximizando $Q(\lambda, \overline{\lambda})$ em relação a $\overline{p}_i$, com a imposição de que $\sum_{i=1}^{M} \overline{p}_i = 1$, para assegurar que $\overline{p}_i$ sejam probabilidades válidas. Pode-se mostrar que (REYNOLDS, 1992):

$$\overline{p}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(k)} \tag{8.17}$$

ou, substituindo na expressão para $\gamma_t(i)$ e cancelando termos,

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda) \tag{8.18}$$

onde $P(i_t = i|\vec{x}_t, \lambda)$ é dado pela EQ. 8.16.

Vetor Média das Densidades

Tomando o gradiente da EQ. 8.13 em relação a um vetor média $\vec{\overline{\mu}}_i$ de uma densidade, resulta (REYNOLDS, 1992):

$$\frac{\partial Q(\lambda, \overline{\lambda})}{\partial \vec{\overline{\mu}}_i} = \frac{\partial}{\partial \vec{\overline{\mu}}_i} \sum_{t=1}^{T} \gamma_t(i) \log \overline{b}_i(\vec{x}_t) = \sum_{t=1}^{T} \gamma_t(i) \frac{\partial}{\partial \vec{\overline{\mu}}_i} \left[-\frac{1}{2} (\vec{x} - \vec{\overline{\mu}}_i)' \ \overline{K}_i^{-1} (\vec{x} - \vec{\overline{\mu}}_i) \right] \tag{8.19}$$

O símbolo $'$ indica transposição. Neste ponto, uma regra útil de diferenciação para um vetor $\vec{a}$ $(D \times 1)$ e uma matriz C $(D \times D)$ é (REYNOLDS, 1992):

$$\frac{\partial}{\partial \vec{a}} \vec{a}\,' C \vec{a} = 2C\vec{a}$$

Aplicando esta regra, fazendo a EQ. 8.19 igual a zero e resolvendo para $\vec{\overline{\mu}}_i$, resulta (REYNOLDS, 1992):

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) \vec{x}_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{8.20}$$

A fórmula para estimação do vetor média final é obtida, substituindo $\gamma_t(i)$ na EQ. 8.15 e cancelando alguns termos, o que resulta:

$$\vec{\overline{\mu}}_i = \frac{\sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda) \vec{x}_t}{\sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda)} \tag{8.21}$$

Matriz Covariância das Densidades

Como a maximização do vetor de médias, $Q(\lambda, \overline{\lambda})$ pode ser maximizado em relação a todos os elementos da matriz covariância simultaneamente, encontrando os zeros do gradiente da EQ. 8.13 em relação a matriz $\overline{K}_i$. O gradiente da EQ. 8.13 é (REYNOLDS, 1992):

$$\frac{\partial Q(\lambda, \overline{\lambda})}{\partial \overline{K}_i} = \frac{\partial}{\partial \overline{K}_i} \sum_{t=1}^{T} \gamma_t(i) \log \overline{b}_i(\vec{x}_t)$$

$$= \sum_{t=1}^{T} \gamma_t(i) \frac{\partial}{\partial \overline{K}_i} \left[-\frac{1}{2}(\vec{x} - \vec{\overline{\mu}_i})' \, \overline{K}_i^{-1} (\vec{x} - \vec{\overline{\mu}_i}) - \frac{1}{2} \log |\overline{K}_i| \right] \quad (8.22)$$

Novamente, algumas regras de diferenciação de matrizes são necessárias. Para um vetor $\vec{a}$ $(D \times 1)$ e uma matriz não singular C $(D \times D)$, a seguinte regra é mantida (REYNOLDS, 1992):

$$\frac{\partial}{\partial \vec{a}} \vec{a}\,' C^{-1} \vec{a} = -\vec{a}\vec{a}\,' C^{-1} (C^{-1})'$$

e

$$\frac{\partial}{\partial C} \log |C| = C^{-1}$$

Aplicando estas regras, igualando a EQ. 8.22 a zero e resolvendo para $\overline{K}_i$ resulta (REYNOLDS, 1992):

$$\overline{K}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(\vec{x} - \vec{\overline{\mu}_i})(\vec{x} - \vec{\overline{\mu}_i})'}{\sum_{t=1}^{T} \gamma_t(i)} \quad (8.23)$$

Finalmente, usando a equação para $\gamma_t(i)$ produz-se a equação para a estimação da covariância:

$$\overline{K}_i = \frac{\sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda) \vec{x}_t \vec{x}\,'}{\sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda)} - \vec{\overline{\mu}}_i \vec{\overline{\mu}}_i\,' \quad (8.24)$$

Assumindo matrizes covariância diagonal, somente os elementos da diagonal ou as variâncias precisam ser estimadas e a seguinte equação é usada (REYNOLDS, 1992):

$$\overline{\sigma}_{ij}^2 = \frac{\sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda) x_{tj}^2}{\sum_{t=1}^{T} P(i_t = i | \vec{x}_t, \lambda)} - \overline{\mu}_{ij}^2 \quad (8.25)$$

onde $\overline{\sigma}_{ij}^2$, x_{tj} e $\overline{\mu}_{ij}$, $j = 1, ..., N$, $i = 1, ..., M$ (N é o comprimento do vetor $\vec{x}_t$ e M é o número de gaussianas do GMM) referem-se aos elementos dos vetores $\vec{\overline{\sigma}}_i^{\,2}$, $\vec{x}_t$ e $\vec{\overline{\mu}}_i$ respectivamente.

8.1.4 O ALGORITMO EM

Coletivamente, as EQs. 8.18, 8.21, e 8.24 ou 8.25, formam a base do algoritmo EM para iterativamente estimar os parâmetros de um GMM. O algoritmo consiste dos seguintes passos:

- **Inicialização** : inicializar os parâmetros do modelo $\lambda^{(0)}$.
- **Passo E**: estimar novos parâmetros do modelo $\overline{\lambda}$, usando os parâmetros do modelo $\lambda^{(m)}$ via equações de estimação.
- **Passo M**: substituir os parâmetros do modelo corrente com os parâmetros do novo modelo: $\lambda^{(m+1)} \leftarrow \overline{\lambda}$.
- **Iteração**: iteração entre os passos E e M até que a função de verossimilhança pare de crescer. Geralmente 15 iterações são suficientes (REYNOLDS, 1995).

Desde que os parâmetros obtidos do novo modelo, através das equações de estimação, garantem um crescimento monotônico da função de verossimilhança, o algoritmo garante a convergência para um ponto estacionário da função de verossimilhança (THEODORIDIS, 1999).

8.1.5 USO DO ALGORITMO EM

- Inicialização: O algoritmo para treinamento do GMM deve ser iniciado com algum modelo $\lambda^{(0)}$. Para isto, por exemplo, pode-se utilizar um algoritmo de VQ, como o LBG (LINDE, 1980), para gerar o modelo inicial. Dos vetores mais próximos a classe i, dada pelo VQ, cálcula-se μ_i e K_i, os pesos p_i são dados somando-se todos os vetores pertencentes a classe i e dividindo-se pelo número total de vetores de treinamento.
- Limite de variância: algumas vezes os valores de variância podem assumir valores muito pequenos, principalmente quando se utiliza modelos de ordem maior ou igual a 32 (REYNOLDS, 1995) ou quando se utiliza vetores de características com valores pequenos como o caso do mel-cepestro. Isto também pode acontecer quando existirem poucos dados para o treinamento. Este problema produz uma singularidade no modelo da função de verossimilhança, podendo degradar o desempenho do GMM. Para evitar isto, utiliza-se um limite de variância. Este limite é usado sempre que algum valor de variância num modelo λ atinge valor inferior ao determinado. Para

um conjunto de misturas com vetores de variância (matriz covariância diagonal), $\vec{\sigma}_i^2$, e um mínimo valor de variância, σ_{min}^2, o limite:

$$\overline{\sigma}_{ij}^2 = \begin{cases} \sigma_{ij}^2 & \text{se} \quad \sigma_{ij}^2 > \sigma_{min}^2 \\ \sigma_{min}^2 & \text{se} \quad \sigma_{ij}^2 \leq \sigma_{min}^2 \end{cases} \tag{8.26}$$

é aplicado a estimativa de variância após cada iteração do EM, para evitar singularidades no modelo final. Na equação acima, j é o índice do vetor $\vec{\sigma}_i^2$. Segundo REYNOLDS (1995), um limite entre $\sigma_{min}^2 = 0,01$ e $\sigma_{min}^2 = 0,1$ é uma boa opção para este objetivo. Nesta dissertação o GMM foi treinado com o limite $\sigma_{min}^2 = 0,01$. Os programas para o Matlab©, usados para o treinamento do GMM, foram obtidos do *Imperial College of Science, Techology & Medice*, no site *www.ee.ic.ac.uk*, que desponibiliza um conjunto de programas para processamento de voz no Matlab© (*voicebox*).

- Ordem do Modelo: determinar o necessário número de componentes gaussianas no GMM para modelar adequadamente um locutor é um problema importante. Porém, não existe um caminho teórico para estimar este número *a priori*. Para a modelagem de um locutor, o objetivo é escolher o menor número de componentes necessárias para modelá-lo adequadamente, visando um bom desempenho do sistema de reconhecimento. Escolher poucas componentes pode produzir um modelo impreciso. Entretanto, escolher muitas componentes pode reduzir o desempenho quando existirem muitos parâmetros para se estimar relativos aos dados para treinamento, além de resultar numa excessiva complexidade computacional para ambos treinamento e teste do sistema.

8.2 APÊNDICE 2: ALGORITMOS PARA OBTENÇÃO DO LPC E DO AR-VETORIAL

Na obtenção do LPC ou do AR-Vetorial é empregado o algoritmo de Levison-Durbin (HAYKIN, 1996), gerando modelos direto e reverso dos coeficientes de predição, a e $\hat{a}$ ou A e $\hat{A}$, respectivamente. Os modelos são calculados a partir das matrizes de autocorrelação. A notação utilizada na seção 4.4.1 é a mesma utilizada para os algoritmos que seguem.

- **Algoritmo para estimação do LPC**:

  Inicialização:

  $a_0 = 1$, $e_0 = \hat{e}_0 = r_0$, $a^0 = \hat{a}^0 = a_0$

  Para q, variando de 1 até p, faça:

  $$f_q = \sum_{k=0}^{q-1} a_k^{q-1} r_{q-k}$$

  $$k_q = -f_q/\hat{e}_{q-1} \qquad \hat{k}_q = -f_q/e_{q-1}$$

  $$a^q = (a^{q-1} \quad 0) + k_q(0 \quad \hat{a}^{q-1})$$

  $$\hat{a}^q = (0 \quad \hat{a}^{q-1}) + \hat{k}_q(a^{q-1} \quad 0)$$

  $$e_q = e_{q-1} + k_q f_q \qquad \hat{e}_q = \hat{e}_{q-1} + \hat{k}_q f_q$$

  Fim

  $a^q = (a_0 \quad a_1 \quad a_2 \ldots a_p)$

  $\hat{a}^q = (\hat{a}_0 \quad \hat{a}_1 \quad \hat{a}_2 \ldots \hat{a}_p)$

- **Algoritmo para estimação do AR-Vetorial**:

  Inicialização:

  $A_0 = I$, $E_0 = \hat{E}_0 = R_0$, $A^0 = \hat{A}^0 = A_0$, $Z = [zeros]_{m \times m}$

  Para q, variando de 1 até p, faça:

$$F_q = \sum_{k=0}^{q-1} A_k^{q-1} R_{q-k}^T$$

$$K_q = -F_q / \hat{E}_{q-1} \qquad \hat{K}_q = -F_q^T / E_{q-1}$$

$$A^q = (A^{q-1} \quad Z) + K_q (Z \quad \hat{A}^{q-1})$$

$$\hat{A}^q = (Z \quad \hat{A}^{q-1}) + \hat{K}_q (A^{q-1} \quad Z)$$

$$E_q = E_{q-1} + K_q F_q^T \qquad \hat{E}_q = \hat{E}_{q-1} + \hat{K}_q F_q$$

Fim

$A^q = [A_0 \quad A_1 \quad A_2 \ldots A_p]$

$\hat{A}^q = [\hat{A}_0 \quad \hat{A}_1 \quad \hat{A}_2 \ldots \hat{A}_p]$

8.3 APÊNDICE 3: RESULTADOS COMPLEMENTARES DO AR-VETORIAL

TAB. 8.1: Desempenho do AR-Vetorial para as diferentes distâncias usadas, $p = 2$ e 30s de treinamento.

Dist.	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1	0	0	**0**	2,3	1,8	**2,05**	10,6	9,3	**9,95**
2	0	1,15	**0,57**	2,56	2,95	**2,75**	15,6	13,6	**14,6**
3	7,3	5,4	**6,35**	8,2	8,7	**8,45**	14,8	19,1	**16,95**
4	5,0	3,8	**4,4**	5,4	4,7	**5,05**	16,7	15,5	**16,1**
5	0	0	**0**	1,8	1,4	**1,6**	10,2	10,3	**10,25**
6	1,9	1,5	**1,7**	3,8	3,2	**3,5**	13,7	12,4	**13,05**
7	0	0,77	**0,38**	2,2	1,4	**1,8**	10,9	10,2	**10,55**

TAB. 8.2: Desempenho do AR-Vetorial para as diferentes distâncias usadas, $p = 2$ e 10s de treinamento.

Dist.	Testes (%)					
	10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**
1	3,0	2,2	**2,6**	11,5	11,2	**11,35**
2	4,7	7,2	**5,95**	16,2	16,0	**16,1**
3	8,6	8,2	**8,4**	16,8	16,6	**16,7**
4	8,3	7,4	**7,85**	16,3	14,9	**15,6**
5	2,4	3,8	**3,1**	11,6	12,0	**11,8**
6	5,9	4,7	**5,3**	13,6	12,2	**12,9**
7	3,1	2,9	**3,0**	11,5	11,0	**11,25**

TAB. 8.3: Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 1 e 60s de treinamento).

Ordem do modelo	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1	0,38	0,38	**0,38**	1,8	1,5	**1,65**	8,1	7,7	**7,9**
2	0,38	0	**0,19**	1,15	2,3	**1,72**	8,8	9,6	**9,2**
3	0	0	**0**	1,6	1,5	**1,55**	9,2	8,8	**9,0**
4	0	0	**0**	1,4	1,4	**1,4**	9,1	8,8	**8,95**
5	0	0	**0**	1,7	1,5	**1,6**	14,9	14,3	**14,6**

TAB. 8.4: Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 1 e 30s de treinamento.

Ordem do modelo	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1	0,77	0,38	**0,57**	2,7	2,4	**2,55**	8,8	8,1	**8,45**
2	0	0	**0**	2,3	1,8	**2,05**	10,6	9,3	**9,95**
3	0	0	**0**	1,9	1,7	**1,8**	9,8	8,9	**9,35**
4	0	0	**0**	1,9	1,9	**1,9**	9,5	9,3	**9,4**
5	0	0	**0**	1,9	1,9	**1,9**	15,3	14,4	**14,85**

TAB. 8.5: Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 1 e 10s de treinamento.

Dist.	Testes (%)					
	10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**
1	4,2	3,8	**4,0**	10,9	10,0	**10,45**
2	3,0	2,2	**2,6**	11,5	11,2	**11,35**
3	2,9	2,8	**2,85**	11,7	10,8	**11,25**
4	2,9	3,1	**3,0**	11,8	10,6	**11,2**
5	3,7	3,7	**3,7**	15,8	15,0	**15,4**

TAB. 8.6: Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 7 e 60s de treinamento.

Ordem do modelo	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1	0,38	0,38	**0,38**	1,8	1,7	**1,75**	9,9	10,1	**10,0**
2	0,38	0	**0,19**	1,92	0,90	**1,41**	8,2	12,6	**10,4**
3	0,38	0	**0,19**	1,5	1,3	**1,4**	10,4	9,7	**10,05**
4	0	0	**0**	1,3	1,15	**1,22**	10,15	10,7	**10,42**
5	0	0	**0**	1,7	1,0	**1,35**	13,6	13,3	**13,45**

TAB. 8.7: Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 7 e 30s de treinamento.

Ordem do modelo	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1	0,77	0,77	**0,77**	2,2	2,0	**2,1**	10,5	10,3	**10,4**
2	0	0,77	**0,38**	2,2	1,4	**1,8**	10,9	10,2	**10,55**
3	0	0,38	**0,19**	2,0	1,7	**1,85**	10,9	9,9	**10,4**
4	0	0	**0**	1,8	1,5	**1,65**	11,6	10,7	**11,15**
5	0	0	**0**	2,3	1,9	**2,1**	13,8	13,5	**13,65**

TAB. 8.8: Desempenho do AR-Vetorial, com a variação da ordem do modelo, distância 7 e 10s de treinamento.

Dist.	Testes (%)					
	10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**
1	3,7	3,2	**3,45**	10,9	11,5	**11,2**
2	3,1	2,9	**3,0**	11,5	11,0	**11,25**
3	3,3	2,8	**3,05**	11,8	10,9	**11,35**
4	3,2	2,9	**3,05**	11,5	11,3	**11,4**
5	3,3	3,1	**3,2**	13,9	13,8	**13,85**

TAB. 8.9: Desempenho do AR-Vetorial com MCCPCA, para as ordens 1, 2 e 3, usando as distâncias 1, 5 e 7, com 60s de treinamento.

Ordem - Dist.	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1-D1	1,15	0	**0,57**	1,79	1,54	**1,66**	8,08	7,77	**7,92**
2-D1	0,38	0	**0,19**	1,15	1,53	**1,34**	9,15	9,31	**9,23**
3-D1	0	0	**0**	1,67	1,54	**1,60**	9,0	9,15	**9,07**
1-D5	0,38	0	**0,19**	2,18	1,67	**1,92**	10,0	10,23	**10,11**
2-D5	0	0	**0**	1,02	1,28	**1,15**	10,04	9,31	**9,67**
3-D5	0	0	**0**	1,28	1,02	**1,15**	10,15	9,77	**9,96**
1-D7	0,38	0,38	**0,38**	2,05	1,92	**1,98**	9,92	10,11	**10,01**
2-D7	0,38	0	**0,19**	1,28	1,41	**1,34**	9,77	10,31	**10,04**
3-D7	0	0	**0**	1,28	1,28	**1,28**	10,07	10,19	**10,13**

TAB. 8.10: Desempenho do AR-Vetorial com MCCPCA, para as ordens 1, 2 e 3, usando as distâncias 1, 5 e 7, com 30s de treinamento.

Ordem - Dist.	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1-D1	0,77	0,77	**0,77**	2,69	2,43	**2,56**	8,54	8,31	**8,42**
2-D1	0	0	**0**	1,79	2,18	**1,98**	9,81	10,35	**10,08**
3-D1	0	0	**0**	1,92	1,67	**1,79**	9,19	9,88	**9,53**
1-D5	0,77	0,77	**0,77**	2,30	1,67	**1,98**	10,38	10,46	**10,42**
2-D5	0	0	**0**	1,54	1,54	**1,54**	10,23	10,35	**10,29**
3-D5	0	0	**0**	1,67	1,67	**1,67**	10,23	10,69	**10,46**
1-D7	0,77	0,77	**0,77**	2,18	2,05	**2,11**	10,35	10,58	**10,46**
2-D7	0,38	0,77	**0,57**	1,79	2,05	**1,92**	10,69	10,46	**10,57**
3-D7	1,54	0	**0,77**	2,05	1,41	**1,73**	10,19	10,69	**10,44**

TAB. 8.11: Desempenho do AR-Vetorial com MCCPCA, para as ordens 1, 2 e 3, usando as distâncias 1, 5 e 7, com 10s de treinamento.

Ordem - Dist.	Testes (%)					
	10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**
1-D1	4,23	3,85	**4,04**	10,88	10,03	**10,44**
2-D1	2,69	2,82	**2,75**	11,0	11,81	**22,81**
3-D1	2,82	2,95	**2,88**	11,69	10,85	**11,27**
1-D5	3,46	3,20	**3,33**	11,46	11,04	**11,25**
2-D5	3,08	3,33	**3,20**	12,11	11,58	**11,84**
3-D5	2,95	3,08	**3,51**	12,31	12,0	**12,15**
1-D7	3,08	3,46	**3,27**	11,42	10,85	**11,13**
2-D7	3,08	2,95	**3,01**	11,46	11,00	**11,23**
3-D7	2,82	3,08	**2,95**	11,15	11,35	**11,25**

TAB. 8.12: Desempenho do AR-Vetorial com MCC12, para as ordens 1, 2 e 3, com distâncias 1, 5 e 7 e 60s de treinamento.

Ordem - Dist.	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1-D1	0	0,38	**0,19**	2,05	1,92	**1,98**	9,58	9,61	**9,59**
2-D1	0	0	**0**	1,67	1,92	**1,79**	10,27	10,61	**10,44**
3-D1	0	0	**0**	1,92	1,28	**1,6**	10,85	10,31	**10,58**
1-D5	0,38	0	**0,19**	1,67	1,92	**1,79**	12,0	11,88	**11,94**
2-D5	0	0	**0**	1,54	1,67	**1,60**	11,88	11,61	**11,74**
3-D5	0	0	**0**	1,54	1,67	**1,60**	11,27	11,96	**11,61**
1-D7	0	0,38	**0,19**	1,92	1,54	**1,73**	11,46	11,92	**11,69**
2-D7	0	0	**0**	1,41	1,54	**1,47**	11,85	11,69	**11,78**
3-D7	0	0	**0**	1,41	1,28	**1,34**	12,15	11,54	**11,84**

TAB. 8.13: Desempenho do AR-Vetorial com MCC12, para as ordens 1, 2 e 3, com distâncias 1, 5 e 7 e 30s de treinamento.

Ordem - Dist.	Testes (%)								
	30s			10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**	FR	FA	**ERR**
1-D1	0,77	0,77	**0,77**	2,82	2,95	**2,88**	9,69	10,38	**10,03**
2-D1	0,38	0	**0,19**	2,31	2,43	**2,37**	11,0	11,04	**11,02**
3-D1	0	0	**0**	1,92	1,92	**1,92**	11,04	11,58	**11,31**
1-D5	0,38	0,38	**0,38**	2,43	2,69	**2,56**	12,38	11,96	**12,17**
2-D5	0,38	0,38	**0,38**	1,92	2,05	**1,98**	11,85	11,42	**11,63**
3-D5	0,38	0	**0,19**	1,67	1,79	**1,73**	12,27	11,73	**12,0**
1-D7	0,38	0,38	**0,38**	2,31	2,82	**2,56**	12,31	12,50	**12,4**
2-D7	0,38	0,38	**0,38**	2,05	1,92	**1,98**	12,31	11,65	**11,98**
3-D7	0,38	0,38	**0,38**	1,79	1,79	**1,79**	12,46	11,81	**12,13**

TAB. 8.14: Desempenho do AR-Vetorial com MCC12, para as ordens 1, 2 e 3, com distâncias 1, 5 e 7 e 10s de treinamento.

Ordem - Dist.	Testes (%)					
	10s			3s		
	FR	FA	**ERR**	FR	FA	**ERR**
1-D1	4,36	4,74	**4,55**	13,0	12,46	**12,73**
2-D1	3,08	3,59	**3,33**	12,73	12,46	**12,59**
3-D1	3,08	3,72	**3,4**	12,5	12,77	**12,63**
1-D5	3,97	4,10	**4,03**	13,96	14,08	**14,02**
2-D5	3,20	3,97	**3,58**	13,73	13,81	**13,77**
3-D5	3,85	3,97	**3,91**	13,42	13,23	**13,32**
1-D7	3,97	4,10	**4,03**	13,78	13,88	**13,83**
2-D7	3,08	3,08	**3,08**	13,31	12,81	**13,06**
3-D7	3,85	3,59	**3,72**	13,0	13,15	**13,07**