INSTITUTO MILITAR DE ENGENHARIA

1º TEN JORGE FREDERICO VIEIRA CAMPOS FLORES

NOVAS CONTRIBUIÇÕES À VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR PARA FINS FORENSES

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. José Antonio Apolinário Jr. - D. Sc.
Co-orientador: Dirceu Gonzaga da Silva - M. C.

Rio de Janeiro
2008

c2008

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

Este exemplar é de propriedade do Instituto Militar de Engenharia, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es) e do(s) orientador(es).

006.454 Flores, J.F.V.C.
F634n Novas contribuições à Verificação Automática de Locutor para Fins Forenses / Jorge Frederico Vieira Campos Flores. - Rio de Janeiro: Instituto Militar de Engenharia, 2008.
126 p.: il., graf., tab.

Dissertação (mestrado) - Instituto Militar de Engenharia - Rio de Janeiro, 2008.

1. Processamento de Sinais. 2. Reconhecimento Automático de Locutor. 3. Fonética Forense. I. Título. II. Instituto Militar de Engenharia.

INSTITUTO MILITAR DE ENGENHARIA

1º TEN JORGE FREDERICO VIEIRA CAMPOS FLORES

NOVAS CONTRIBUIÇÕES À VERIFICAÇÃO AUTOMÁTICA DE LOCUTOR PARA FINS FORENSES

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. José Antonio Apolinário Jr. - D. Sc.

Co-orientador: Dirceu Gonzaga da Silva - M. C.

Aprovada em 11 de fevereiro de 2008 pela seguinte Banca Examinadora:

Prof. José Antonio Apolinário Jr. - D. Sc. do IME - Presidente

---

Dirceu Gonzaga da Silva - M. C. do IME

Prof. Ernesto Leite Pinto - D. C. do IME

Prof. Roberto Miscow Filho - M. C. do IME

Prof. Edson Cataldo - D. C. da UFF

Amaro Azevedo de Lima - Ph. D. da COPPE/UFRJ

Rio de Janeiro
2008

A Deus, acima de tudo.
A minha amada família, base de tudo que sou e fonte de ânimo nos momentos precisos.
A meus grandes amigos, que sempre me impulsionaram em todos os aspectos.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por permitir que todos os acontecimentos que permearam esses dois anos de Mestrado pudessem transcorrer sem dissabores maiores, e por ter dado Seu toque de providência no dia da Defesa.

Ao Instituto Militar de Engenharia / Seção de Engenharia Elétrica, pelos meios necessários para o andamento e a conclusão do meu trabalho. Ao meu orientador, José Antonio Apolinário Jr., pela dedicação diuturna; ao meu co-orientador Dirceu Gonzaga da Silva e aos docentes Ernesto Leite Pinto e Juraci Ferreira Galdino, pelas orientações técnicas valiosíssimas à conclusão deste trabalho e do artigo que fora submetido ao XXV SBrT em 2007; aos demais membros da Banca Examinadora, pelas palavras francas e sinceras, pelo apoio valioso na revisão ortográfica, sintática e semântica, pels conselhos e contribuições valiosas e pela atenção os mínimos detalhes. Aos amigos maravilhosos que fiz e revi nesta jornada nos bancos escolares da Pós-Graduação e Graduação, que propiciaram um maravilhoso convívio tanto no IME — Laboratórios de Processamento de Sinais de Voz, Comunicações Digitais, Eletromagnetismo e Controle — quanto na COPPE — nas cadeiras de Redes Neurais Feedforward, Processamento de Sinais da Fala, Otimização e Reconhecimento de Padrões.

A minha valorosa família, base sólida e inestimável do meu viver. Agradeço também a meus "amigos-irmãos" que estimo como se de minha família fossem, que muito contribuíram para esse trabalho com assessoria técnica, apoio moral e pitadas de otimismo. Ao magnífico mestre Paulo Roberto de Carvalho (em memória), um dos principais responsáveis por eu ter conseguido galgar os passos que galguei até hoje. Aos novos companheiros de jornada profissional no CTEx / Projeto Termal, pelo apoio moral e institucional na "reta final" dos trabalhos. Ao Núcleo Evangélico da Praia Vermelha, cujo apoio foi fundamental para que esse grande objetivo se concretizasse.

Aos profissionais do ramo forense, dedicados e atentos a este trabalho — Andréa Martiny, pelo constante contato por correio eletrônico; Andréa Porto-Carreiro, Nelly Soares Reis e, em especial, ao Instituto Carlos Éboli, pela oportunidade de mostrar a Perícia exercida na prática; César Braid e Alessandro Travassos, pelo curso ministrado no fim de outubro e início de novembro, que abriu novas oportunidades de desenvolvimento deste trabalho; em especial ao perito Alessandro, pelas correções fundamentais para este trabalho ter um toque final de perfeição no âmbito da Fonética Forense.

“Tudo vale a pena se a alma não é pequena”

(Fernando Pessoa)

SUMÁRIO

LISTA DE ILUSTRAÇÕES 10

LISTA DE TABELAS 15

LISTA DE SÍMBOLOS E ABREVIATURAS 17

1 INTRODUÇÃO 25
1.1 Objetivo da Dissertação 26
1.2 Contribuições da Dissertação 27
1.3 Organização da Dissertação 27

2 ASPECTOS CONCEITUAIS DA FALA 29
2.1 Mecanismo de Produção da Fala 29
2.2 Análise acústica da fala 31
2.3 Análise no domínio da freqüência em tempo curto 34
2.4 Modelo de tubos 36
2.5 Características Físicas dos sinais de voz 37
2.5.1 Freqüência fundamental (*pitch*) 38
2.5.2 Formantes 39
2.5.3 Coeficientes *mel-Cepstrais* (MFCC) 42
2.5.4 Tonalidade, SFM e SCF 45
2.5.5 Centróides Espectrais por Sub-banda (SSC) 46
2.6 Características com sincronismo de *pitch* 46
2.6.1 Critério de detecção de trechos sonoros dos sinais de voz 47
2.6.2 Detector de *pitch* e sincronismo 48
2.7 Conclusão 50

3 SISTEMAS DE PERÍCIA EM FONÉTICA FORENSE 52
3.1 Introdução 52
3.1.1 A voz no contexto da biometria 54
3.1.2 A Fonética Forense 55
3.1.3 Tipos de perícia em fonética forense 55
3.1.4 A verificação de locutor no contexto de perícia 56
3.2 Testes perceptuais e acústicos 57

3.3 O estado da arte da atividade de perícia no Brasil 60
3.3.1 Verificação das condições de gravação dos sinais de voz 60
3.3.2 Escolha do conjunto de peças-padrão para análise 63
3.3.3 Escolha da metodologia de comparação 64
3.3.4 Escolha dos parâmetros extraídos dos sinais de voz 65
3.3.5 Forma atual de apresentação do resultado da perícia 67
3.4 Análise de *pitch* para fins periciais 69
3.5 Análise de formantes para fins periciais 70
3.5.1 Emprego dos formantes 70
3.5.2 Análise da Distribuição Espectral dos Formantes 72
3.6 Conclusão 74

4 COMPUTAÇÃO EVOLUCIONÁRIA E SELEÇÃO DE CARACTERÍSTICAS APLICADAS À ESTIMAÇÃO DE PARÂMETROS DE MODELOS GMM 76
4.1 Introdução 76
4.2 Modelos de misturas de gaussianas (GMM) 76
4.2.1 O algoritmo EM 76
4.2.2 Forma clássica do algoritmo **EM** para modelos GMM 77
4.2.3 Emprego de projeções aleatórias 78
4.2.4 Emprego de algoritmos genéticos 80
4.2.5 Estudo de caso com os algoritmos EM-RP, EM-GA e a medida BIC 81
4.3 Seleção de características 87
4.4 Conclusão 93

5 AVALIAÇÃO DE RESULTADOS EM CONTRIBUIÇÃO À METODOLOGIA DE FONÉTICA FORENSE 94
5.1 Introdução 94
5.2 Esquema de verificação por GMM 94
5.2.1 Composição do *background* 97
5.3 Descrição sumária dos testes realizados 98
5.4 Modelagem dos erros 100
5.5 Testes de formantes por LTF 101
5.6 Testes das características pelo discriminante de Fisher 108

5.7 Testes de VAL englobando os algoritmos EM, EM-RP e EM-GA para estimação dos parâmetros dos modelos GMM das características 112
5.8 Resumo e Conclusão 117

6 CONCLUSÃO E SUGESTÕES PARA TRABALHOS FUTUROS 120
6.1 Limitações do trabalho 121
6.2 Sugestões para trabalhos futuros 121

7 REFERÊNCIAS BIBLIOGRÁFICAS 123

LISTA DE ILUSTRAÇÕES

FIG.1.1 Diagrama básico da atividade de perícia em Fonética Forense. 25

FIG.2.1 Os sistemas respiratório, laringeal e supralaringeal — figura extraída de (MORISSON, 2003). 30

FIG.2.2 Marcas de estimativa de período de *pitch*. 32

FIG.2.3 Processo de janelamento: (a) Sinal de voz $s(m)$; (b) Sinal de voz $s(m+n)$ e janela de Hamming $w(m)$. 35

FIG.2.4 Quadro $s_n(m)$ do sinal de voz após ser janelado por $w(m)$. 36

FIG.2.5 O modelo de tubo ressonante para três vogais diferentes (MORISSON, 2003). 37

FIG.2.6 Banco de filtros que representa o modelo auditivo humano. 43

FIG.2.7 Conversão das freqüências em hertz para a escala *mel*. 44

FIG.2.8 Diagrama em blocos que representa a extração dos coeficientes MFCC. 44

FIG.2.9 Diagrama básico da extração de características com sincronismo de *pitch*. 47

FIG.2.10 A grandeza P_{sonoro} (*voicing*) é dada pela razão do valor do pico principal pelo valor do primeiro pico secundário da autocorrelação de um quadro de sinal de voz. No gráfico acima, foi analisado um quadro do /a/ sustentado de um locutor do sexo feminino, com *voicing* de 0,7944. 48

FIG.2.11 Amplitude de trechos de sinais de voz em diferentes configurações de segmentação — (a),(c): duplas adjacentes de períodos consecutivos de *pitch*; (b),(d): idem (a) e (c) + 3,75ms; a resolução dos harmônicos (em verde) é melhor em (a) e (c) (casamento com a *pitch*) do que em (b) e (d) (em vermelho). 49

FIG.2.12 Comparação entre os espectros de trechos de sinais de voz em duas configurações diferentes — (a) configuração de janela fixa de 20 ms; (b) três períodos de pitch consecutivos. 50

FIG.3.1 1. O perito recebe o conteúdo das peças-motivo para análise; 2. O perito efetua a coleta de padrão; 3. Ocorre a comparação das peças analisadas. 57

FIG.3.2 Níveis de complexidade de extração das características dos sinais de voz. 59
FIG.3.3 Espectrogramas em banda larga (superior) e banda estreita (inferior) do conjunto de fones ['apə]. As elipses vermelhas indicam, em seu interior, a localização das barras de explosão, visíveis no espectrograma em banda larga. 66
FIG.3.4 Taxas de falsa aceitação e rejeição: (a) *Scores* dos locutores verdadeiros e falsos (impostores); (b) taxa de falsa aceitação (dada pela área em azul) e taxa de falsa rejeição (dada pela área em vermelho) para o limiar igual a 5,5; (c) Novos valores de falsa aceitação e falsa rejeição para o limiar em 2,0. Percebe-se a redução da taxa de falsa rejeição e o aumento da taxa de falsa aceitação com o novo limiar escolhido. 68
FIG.3.5 Gráfico de F_1 versus F_2 da vogal /i/ para os locutores U e K, com janelamento de 20 ms e sobreposição (*overlap*) de 15 ms. Foram extraídos três vetores de formantes [F_1 F_2]: um do centro do fone, um 5ms antes do centro e outro 5ms após o centro. Para o caso ilustrado, U $\neq$ K. 71
FIG.3.6 Gráfico de F_2 inicial versus F_2 final do ditongo /ei/, com janelamento de 20 ms e sobreposição (*overlap*) de 15 ms. Para este caso específico, U $\neq$ K. A linha diagonal preta indica o lugar geométrico de F_2 caso permanecesse constante ao decorrer do ditongo. 71
FIG.3.7 Diagrama explicativo da técnica LTF. 73
FIG.3.8 Comparação de distribuições LTF entre as gravações: (a) K1 e K2 (dois trechos diferentes de fala do mesmo locutor suspeito); (b) K e U (trechos da fala do suspeito e da fala do real autor das chamadas). Nota-se, em (b), que a dissimilaridade entre as curvas da distribuição LTF é bem maior do locutor K (em linhas cheias) para o locutor U (em linhas tracejadas). 74

FIG.4.1 Projeções aleatórias. O histograma indica no eixo horizontal os valores de verossimilhança alcançados para cada projeção. O eixo vertical mostra o número de valores de verossimilhança correspondente a cada faixa de valores do eixo horizontal. A verossi-

milhança do modelo GMM estimada pelo algoritmo EM simples está indicada pela seta rotulada com "EM". Os passos de "b" a "d" do algoritmo EM-RP podem também ser facilmente observados no gráfico. 80

FIG.4.2 Comparação entre as médias das verossimilhanças calculadas pelos algoritmos EM-RP e EM-GA. Foi computada a média de 5 e 30 realizações para as bases PIMA e LANDSAT, respectivamente. Pode ser percebida a maior eficiência do algoritmo EM-GA na média das realizações. 85

FIG.4.3 Comportamento do valor da medida BIC em função do coeficiente γ. 86

FIG.4.4 Comparação entre as Medidas BIC obtidas dos modelos GMM estimados pelo algoritmo EM e pelo algoritmo genético (EM-GA) até 25 gaussianas. Em (a) a medida BIC foi calculada sobre um modelo GMM sintético de 10 gaussianas (4000 vetores do $\mathbb{R}^{10}$ projetados ao $\mathbb{R}^2$). Notar os valores de medida BIC menores e de comportamento mais suave devido à implementação do algoritmo proposto. Além disso, o mínimo local (HALBE, 2005) (coincidente com o mínimo global) da medida BIC em (a) é de 10 gaussianas para ambos os algoritmos, exatamente a ordem arbitrada para o modelo sintético. 89

FIG.4.5 Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 2 e 3 da base de dados LANDSAT do $\mathbb{R}^{36}$ ao $\mathbb{R}^3$. 89

FIG.4.6 Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 4 e 5 da base de dados LANDSAT do $\mathbb{R}^{36}$ ao $\mathbb{R}^3$. 90

FIG.4.7 Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 1 e 2 da base de dados PIMA do $\mathbb{R}^8$ ao $\mathbb{R}^3$. 90

FIG.4.8 Comparação entre as Medidas BIC calculadas pelo algoritmo EM e

pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 1 e 2 da base de dados SEGMENT do $\mathbb{R}^{19}$ ao $\mathbb{R}^3$. 91

FIG.4.9 Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 3 e 4 da base de dados SEGMENT do $\mathbb{R}^{19}$ ao $\mathbb{R}^3$. 91

FIG.4.10 Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 5 e 6 da base de dados SEGMENT do $\mathbb{R}^{19}$ ao $\mathbb{R}^3$. 92

FIG.5.1 Sistema de verificação de locutor. Pode ser percebida a razão de verossimilhanças logarítmica $\Lambda(X)$ na entrada do bloco do teste de hipóteses 96

FIG.5.2 Plano de testes a ser seguido neste trabalho. 99

FIG.5.3 Gráfico do MME em função do tempo do teste de LTF para os casos indicados na legenda — resolução de 200 bandas — vide tabelas TAB. 5.3,TAB. 5.5 e TAB. 5.9. 106

FIG.5.4 Gráfico do MME em função da resolução do histograma para os casos indicados na legenda — teste de 30 s — vide tabelas TAB. 5.4, TAB. 5.6 e TAB. 5.10. 107

FIG.5.5 Gráfico do MME em função do tempo do teste de LTF para os casos indicados na legenda — resolução de 200 bandas — vide tabelas TAB. 5.3,TAB. 5.7 e TAB. 5.11. 108

FIG.5.6 Gráfico do MME em função da resolução do histograma para os casos indicados na legenda — teste de 30 s — vide tabelas TAB. 5.4,TAB. 5.8 e TAB. 5.12. Nota-se que, ao contrário da FIG. 5.4, o teste FSS não obtém os melhores resultados, e sim o teste FS1PS. 109

FIG.5.7 Ganho de verossimilhança dos modelos GMM estimados pelos algoritmos EM-GA ($\mathcal{L}_{i,EM-GA} - \mathcal{L}_{i,EM}$) e EM-RP ($\mathcal{L}_{i,EM-RP} - \mathcal{L}_{i,EM}$) indicados pela legenda (8 gaussianas) para cada locutor i, sendo $i = 1, \ldots, 40$. 114

FIG.5.8 Ganho de verossimilhança dos modelos GMM estimados pelos algo-

ritmos EM-GA ($\mathcal{L}_{i,EM-GA} - \mathcal{L}_{i,EM}$) e EM-RP ($\mathcal{L}_{i,EM-RP} - \mathcal{L}_{i,EM}$) indicados pela legenda (16 gaussianas) para cada locutor i, sendo $i = 1, \ldots, 40$. 114

FIG.5.9 Ganho de verossimilhança dos modelos GMM estimados pelos algoritmos EM-GA ($\mathcal{L}_{i,EM-GA} - \mathcal{L}_{i,EM}$) e EM-RP ($\mathcal{L}_{i,EM-RP} - \mathcal{L}_{i,EM}$) indicados pela legenda (32 gaussianas) para cada locutor i, sendo $i = 1, \ldots, 40$. 115

FIG.5.10 Impacto no MME dos modelos GMM estimados pelos algoritmos indicados pela legenda *versus* ganho médio de verossimilhança por locutor (testes de 3 s). 115

FIG.5.11 Teste de VAL para 3 s e 10 s de peça-motivo. 116

FIG.5.12 Teste de VAL para 30 s de peça-motivo. 117

FIG.5.13 Teste de VAL para 3 s, 10 s e 30 s de peça-motivo. O treinamento dos modelos GMM foi efetuado pela Medida BIC. 117

LISTA DE TABELAS

TAB.4.1 Bases de dados 82
TAB.4.2 Condição para economia de parâmetros 83
TAB.4.3 Verossimilhança da base PIMA usando um modelo GMM com 5 gaussianas, com projeção em $d = 2$ 83
TAB.4.4 Verossimilhança da base LANDSAT usando um modelo GMM com 5 gaussianas, com projeção em $d = 8$ 84
TAB.4.5 Verossimilhança da base SEGMENT usando um modelo GMM com 5 gaussianas, com projeção em $d = 4$ 84
TAB.4.6 Verossimilhança - medida BIC - PIMA - $d = 2$ / $d = 3$ 86
TAB.4.7 Verossimilhança - medida BIC - LANDSAT - $d = 2$ / $d = 3$ 87
TAB.4.8 Verossimilhança - medida BIC - SEGMENT - $d = 2$ / $d = 3$ 88

TAB.5.1 Avaliação do discriminante de Fisher para os testes de LTF (vide legenda na TAB. 5.2.) 103
TAB.5.2 Legenda auxiliar para a TAB. 5.1. 103
TAB.5.3 Teste FSS — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas. 103
TAB.5.4 Teste FSS — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s. 104
TAB.5.5 Teste FS1P — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas. 104
TAB.5.6 Teste FS1P — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s. 104
TAB.5.7 Teste FS1PS — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas. 105
TAB.5.8 Teste FS1PS — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s. 105
TAB.5.9 Teste FS2P — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas. 105
TAB.5.10 Teste FS2P — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s. 105
TAB.5.11 Teste FS2PS — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas. 106

TAB.5.12 Teste FS2PS — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s. 106
TAB.5.13 Discriminantes de Fisher para as características extraídas dos sinais de voz 109
TAB.5.14 Avaliação da VAL para os coeficientes MFCC de ordem 15, delta e delta-delta 110
TAB.5.15 Avaliação da VAL para os coeficientes SSC 110
TAB.5.16 Avaliação da VAL para os coeficientes de tonalidade 110
TAB.5.17 Avaliação da VAL para os coeficientes SFM 111
TAB.5.18 Avaliação da VAL para os coeficientes SCF 111
TAB.5.19 Avaliação da VAL para todos os 61 coeficientes 111
TAB.5.20 Avaliação comparativa de desempenho dos algoritmos EM, EM-RP e EM-GA para os coeficientes MFCC, delta e delta-delta 113

LISTA DE SÍMBOLOS E ABREVIATURAS

BIC	-	*Bayesian Information Criterion*
CD	-	*Compact Disc*
DCF	-	*Detection Cost Function*
DCT	-	*Discrete Cosine Transform*
DET	-	*Detection Error Tradeoff*
DFT	-	*Discrete Fourier Transform*
DVD	-	*Digital Video Disc*
EER	-	*Equal Error Rate*
EM	-	*Expectation Maximization*
FA	-	*False Acception*
FR	-	*False Rejection*
f.d.p	-	Função de Densidade de Probabilidade
FFT	-	*Fast Fourier Transform*
GA	-	*Genetic Algorithm*
GMM	-	*Gaussian Mixture Model*
HD	-	*Hard-Disk*
HMM	-	*Hidden Markov Models*
LDA	-	*Linear Discriminant Analysis*
LPC	-	*Linear Prediction Coefficients*
LTF	-	*Long-Time Formants*
LTS	-	*Long-Time Spectrum*
MFCC	-	*Mel-Frequency Cepstrum Coefficients*
MME	-	*Minimum Mean Error*
MMI	-	*Maximum Mutual Information*
PCA	-	*Principal Component Analysis*
RAL	-	Reconhecimento Automático de Locutor
ROC	-	*Receiver Operating Characteristics*
RP	-	*Random Projection*
SCF	-	*Spectral Crest Factor*
SFM	-	*Spectral Flatness Measure*

SFS	-	*Speech File System*
SNR	-	*Signal to Noise Ratio*
SSC	-	*Sub-band Spectrum Centroid*
UBM	-	*Universal Background Model*
VAL	-	Verificação Automática de Locutor
VOT	-	*Voice Onset Time*

SIGLAS

IAI	-	*International Association for Identification*
NIST	-	*National Institute of Standards and Technology*

SÍMBOLOS

$a(k)$	-	Amplitude da freqüência k de uma sub-banda do quadro de um sinal de voz
$\hat{\mathbf{a}}$	-	Vetor de coeficientes LPC
B_i	-	Sub-banda de ordem i do espectro do quadro de um sinal de voz
C_{FA}	-	Custo de falsa aceitação
C_{FR}	-	Custo de falsa rejeição
C_m	-	Coeficiente SSC de ordem m
d	-	Dimensão de projeção do vetor de características
D	-	Dimensão do vetor de características
E	-	Erro MME
$E[\cdot]$	-	Valor esperado
E_{FA}	-	Erro de falsa aceitação
E_{FR}	-	Erro de falsa rejeição
$f_L^{(trein)}$	-	Histograma de treinamento do L-ésimo locutor
$f_l^{(teste)}$	-	Histograma de teste do l-ésimo locutor
f_s	-	Freqüência de amostragem do sinal de voz
F_n	-	Formante de ordem n
$g(n)$	-	Pulso glotal
H_0	-	Hipótese verdadeira
H_1	-	Hipótese falsa
$\mathcal{H}(z)$	-	Modelo da glote no domínio da freqüência
k_c	-	Fator de recombinação (*crossover*) do algoritmo genético
K	-	Locutor conhecido
$\mathcal{L}$	-	Verossimilhança logarítmica
$m^{(k)}$	-	Média interlocutor da k-ésima característica
$m_i^{(k)}$	-	Média intralocutor da k-ésima característica relativa ao i-ésimo locutor
M	-	Número de componentes gaussianas do modelo GMM
n_g	-	Número de gerações do algoritmo genético
n_{RP}	-	Número de projeções aleatórias
$n_{\mathbf{W}}$	-	Número de matrizes de projeção (população) do algoritmo genético
N_w	-	Número de amostras do quadro de um sinal de voz
$N(\cdot)$	-	*f.d.p.* normal

$\mathbb{N}^*$	-	Conjunto dos números naturais não-nulos (o mesmo que $\mathbb{Z}_+^*$)
$P(f)$	-	Densidade espectral de potência
$\wp$	-	Vetor de pertinência do modelo GMM projetado
$r_s(\eta; m)$	-	Autocorrelação do sinal de voz em tempo curto
$\mathbf{r}_n$	-	Vetor de autocorrelação do método de Levinson-Durbin
$\mathbf{R}_n$	-	Matriz de autocorrelação do método de Levinson-Durbin
$\mathbb{R}$	-	Conjunto dos números reais
$s(m)$	-	Sinal de voz
$s(m+n)$	-	Sinal de voz deslocado de n amostras
$s_n(m)$	-	Quadro do sinal de voz, dado por $s(m+n)w(m)$
S	-	Parâmetro de controle da redução da variância dos indivíduos da população do algoritmo genético
$\mathbf{S}(\omega; m)$	-	Transformada de Fourier em tempo curto do quadro $s_n(m)$
$\mathrm{Tr}\{\cdot\}$	-	Traço da matriz
U	-	Locutor desconhecido
v	-	Velocidade do som no ar
$\mathbf{V}$	-	Cromossomo do algoritmo genético
$\mathcal{V}(z)$	-	Modelo do trato vocal no domínio da freqüência
$w(\cdot)$	-	Janela de amostragem do sinal de voz
$\mathbf{W}$	-	Matriz de projeção
$\mathbf{x}$	-	Vetor de características
$\mathbf{X}$	-	Conjunto de vetores de características
$\mathcal{X}(k; m)$	-	DFT de cada quadro do sinal de voz relativa aos coeficientes LPC
$\mathbf{y}$	-	Vetor projetado de características
$\mathbf{Y}$	-	Conjunto de vetores projetados de características
$\Delta_f^{(k)}$	-	Discriminante de Fisher da k-ésima característica
η	-	Número de amostras da autocorrelação
Λ	-	Razão logarítmica de verossimilhança
$\boldsymbol{\mu}$	-	Vetor médio interlocutor
$\boldsymbol{\mu}_i$	-	Vetor médio intralocutor do i-ésimo locutor
$\boldsymbol{\mu}_j$	-	Vetor médio da j-ésima componente gaussiana do modelo GMM
ν	-	Fator de penalidade da medida BIC
π_j	-	Peso da j-ésima componente gaussiana do modelo GMM
$\sigma_{inter}^{2(k)}$	-	Variância interlocutor da k-ésima característica

$\sigma_{intra}^{2(k)}$	-	Variância intralocutor da k-ésima característica
$\mathbf{\Sigma}_j$	-	Matriz de covariância da j-ésima componente gaussiana do modelo GMM
$\mathbf{\Sigma}_{inter}^{2(k)}$	-	Variância interlocutor da k-ésima característica
$\mathbf{\Sigma}_{intra}^{2(k)}$	-	Variância intralocutor da k-ésima característica

RESUMO

Esta dissertação aborda contribuições à Verificação Automática de Locutor (VAL) independente do texto para fins forenses. Foi realizado um estudo prévio do mecanismo de produção da fala, das principais características dos sinais de voz em âmbito forense (*pitch* e formantes), de características de reconhecimento de áudio e fala que foram introduzidas no estudo da VAL (coeficientes de tonalidade, SFM, SCF e SSC) e dos coeficientes mel-cepstrais (MFCC) e de suas duas primeiras derivadas (coeficientes delta e delta-delta, respectivamente). Foi estudado, ainda, o mecanismo de extração de características com sincronismo por períodos de *pitch*.

Para o contexto forense, foi estudado também o estado da arte dos protocolos de perícia fonética atualmente existentes no Brasil. Ao fim desse estudo, implementou-se uma técnica que emprega histogramas dos formantes extraídos dos sinais de voz, conhecida como Distribuição de Formantes a Tempo Longo (LTF).

Sob o ponto de vista matemático, foram estudadas as técnicas que dão suporte às avaliações de VAL: a estimação ML dos parâmetros dos modelos GMM (algoritmo EM), uma nova técnica de estimação ML do modelo GMM baseada na redução de dimensão utilizando matrizes de projeção selecionadas via algoritmos genéticos (EM-GA) e uma técnica anteriormente proposta que emprega projeção aleatória (EM-RP). Também foi realizado um estudo dos Discriminantes de Fisher sob o aspecto de avaliação das características mais discriminativas na tarefa de verificação.

Os resultados da avaliação da VAL, com suporte da análise do Discriminante de Fisher, são mostrados em três fases distintas: na primeira, é avaliada a VAL pela técnica LTF; na segunda, são avaliados os modelos GMM (via algoritmo EM simples) das características MFCC, SSC, SCF, tonalidade e SFM de forma isolada e em conjunto; finalmente, na terceira, é avaliado o ganho de desempenho do algoritmo EM-GA sobre os algoritmos EM e EM-RP e sua correspondência com o ganho de verossimilhança dos modelos GMM obtido em duas situações: EM-RP *versus* EM e EM-GA *versus* EM.

ABSTRACT

This dissertation addresses contributions to text independent automatic speaker verification (ASV) for forensic applications. Initially, a brief review of important topics is presented: the mechanism of speech production, the main features of speech signals used in a phonetic forensic environment (pitch and formants), the features used in audio recognition (tonality coefficients, SFM, SCF, and SSC) that were applied to ASV, and the mel-frequency cepstral coefficients (MFCC) with their first and second derivatives (delta and delta–delta coefficients, respectively). Also addressed was the feature extraction synchronized with the pitch period.

Aiming to collect subsidies for the context of this work, the state of the art of the forensic phonetics protocols currently used in Brazil was investigated. After this, a technique named Long-Term Formant Distribution (LTF), that employs histograms of formants extracted from speech signals, was employed.

Under the mathematical point of view, techniques used to give support to evaluation of ASV were studied: ML estimation of GMM model parameters (algorithm EM), a new technique for ML estimation of GMM model based on the dimension reduction employing projection matrices selected via genetic algorithms (EM-GA), and a previously proposed technique that employs random projection (EM-RP). A study was also carried out on the Fisher Discriminant aiming the evaluation of the most discriminative features for the task of speaker verification.

The evaluation results of ASV, with the support of the Fisher Discriminant, are shown in three distinct phases: in the first one, the ASV is evaluated via the LTF technique; in the second one, the GMM models (via the simple version of the algorithm EM) of the features MFCC, SSC, SCF, tonality, and SFM are evaluated in both isolated and combined forms; finally, in the third one, the performance gain of the algorithm EM-GA is evaluated over the algorithms EM and EM-RP as well as their correspondence with likelihood gain of GMM models obtained in two situations: EM-RP *versus* EM and EM-GA *versus* EM.

1 INTRODUÇÃO

O trabalho de perícia em Fonética Forense (BRAID, 2003; MORISSON, 2003), representado de forma sintética na FIG. 1.1, envolve técnicas de comparação de vozes de locutores diferentes, visando inferir se a voz coletada do locutor envolvido na cena do crime (chamada de peça-motivo) é a mesma voz de um determinado locutor cujo padrão de voz (também conhecido como peça-padrão) foi coletado posteriormente. As técnicas atualmente conhecidas no ambiente pericial são baseadas em comparação direta entre pares de espectrogramas de locutores (o locutor de treinamento — peça-padrão — e o locutor de teste — peça-motivo), buscando características visuais em comum, tais como barras de explosão, barras de vozeamento, perfis de formantes, perfis de contorno melódico de *pitch*, entre outros (ROSE, 2002; BRAID, 2003; MORISSON, 2003). A questão da visualização é fundamental no ambiente da Perícia, pois o espectrograma agrega a análise tempo-freqüência e informações de energia em duas dimensões. Além dos espectrogramas, a atividade de perícia em Fonética Forense também engloba a comparação perceptual (lingüística, sociolingüística e parâmetros supra-segmentares) e outras transformações, tais como o LTF (*Long-Term Formants*), comparando visualmente dois histogramas de formantes (NOLAN, 2005). Todas estas formas de comparação servem de apoio à decisão do perito e à posterior elaboração do laudo.

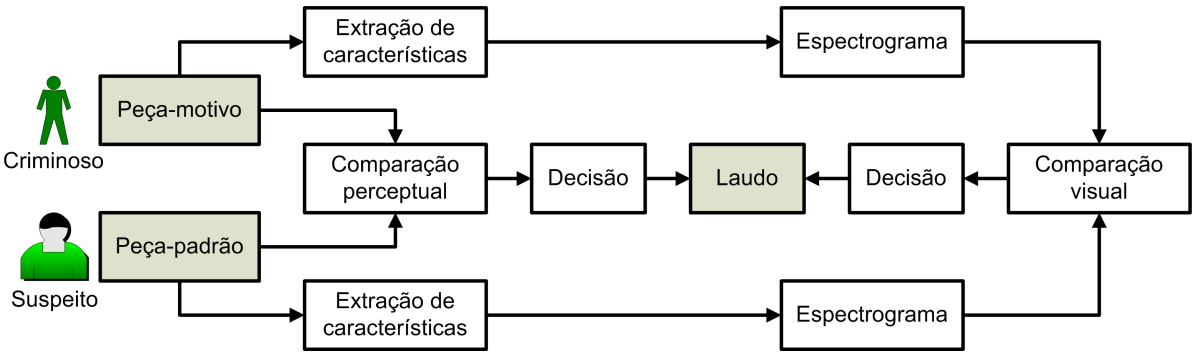


FIG.1.1: Diagrama básico da atividade de perícia em Fonética Forense.

O estado da arte em reconhecimento automático de locutor (RAL) utiliza técnicas probabilísticas em sua decisão, como o GMM — modelo de misturas de gaussianas, de *Gaussian Mixture Model* (REYNOLDS, 1995a,b, 2000) — que não utiliza a interferência

humana (medidas automáticas) e é mais utilizado em reconhecimento independente do texto, no qual as frases proferidas pelo locutor no instante do treinamento são diferentes das frases de teste. O RAL normalmente utiliza vetores N-dimensionais, $N >> 2$, que não permitem a visualização humana (dado ser impossível o ser humano enxergar um vetor em um número de dimensões maior do que três, fica impraticável para o perito visualizar as características que melhor discriminam os locutores). A característica principal atualmente utilizada são os MFCC (Coeficientes Mel-Cepstrais, de *Mel-Frequency Cepstral Coefficients*) (DAVIS, 1980; PEETERS), apresentando bons resultados quando utilizados (REYNOLDS, 1995a) com dados controlados (com alta relação sinal-ruído, microfones de boa qualidade e canais cujas distorções são conhecidas ou invariantes no tempo), justamente o contrário do encontrado no ambiente de perícia em Fonética Forense. Atualmente, outras características vêm sendo pesquisadas de modo a introduzir maior robustez a ruído aditivo e distorção de canal. Além disso, técnicas de decisão mais aprimoradas também vêm sendo pesquisadas.

Com o objetivo de estimar de forma mais eficiente modelos GMM, foram desenvolvidas técnicas de projeções aleatórias (DASGUPTA), que estimam os parâmetros das misturas de gaussianas na dimensão da projeção (efetuada por uma matriz de projeção aleatória) e posteriormente na dimensão original (utilizando uma pré-clusterização com base nos dados projetados), e de algoritmos genéticos (LIN), que estimam estes mesmos parâmetros de forma evolucionária. Além disso, foi desenvolvido um método evolucionário (FLORES) fundindo essas duas técnicas, buscando a matriz de projeção de forma genética, com ganhos em termos de verossimilhança do modelo obtido e menor tempo de otimização na média das realizações do algoritmo.

Sob o aspecto da extração das características, foram pesquisados métodos de extração síncronos à freqüência de vibração das cordas vocais[1], conhecida como freqüência fundamental (ou *pitch*). Esses métodos atingem uma melhor definição espectral, contribuindo para uma melhor visualização do espectrograma de sinais de voz, e alcançam menores taxas de erro na verificação automática de locutor (MORGAN; LEE; EZZAIDI; KIM; ZENG).

1.1 OBJETIVO DA DISSERTAÇÃO

Esta dissertação tem o objetivo principal de realizar um estudo comparativo e a implementação de variações da estimação dos modelos GMM — convencional, por projeções

[1]Também conhecidas como pregas vocais

aleatórias e por algoritmos genéticos — conjugadas às técnicas de Verificação Automática de Locutor (VAL) hoje existentes, visando a proposta de uma nova metodologia de perícia em Fonética Forense. Dessas técnicas, serão buscadas aquelas que atingirem melhor potencial de discriminação de dados de locutores distintos para auxiliar a tomada de decisão do perito em um ambiente forense.

1.2 CONTRIBUIÇÕES DA DISSERTAÇÃO

Esta dissertação, visando obter melhores resultados de VAL no contexto de perícia, introduz os conceitos da estimação de ordem[2] de modelo GMM (medida BIC — medida de critério de informação bayesiana, de *Bayesian Information Criterion.*) por projeções aleatórias e algoritmos genéticos aplicados à estimação do modelo GMM. Esses algoritmos listados possuem a vantagem de acelerar os algoritmos de verificação por serem implementados em uma dimensão muito menor do que a dimensão original dos vetores de características. Além disso, este trabalho também implementa, de forma inovadora no contexto pericial, características de reconhecimento de áudio (SCF — fator de crista espectral, de *Spectrum Crest Factor*, SFM — medida de planura espectral, de *Spectrum Flatness Measure* — e tonalidade) e de reconhecimento de voz (SSC — Centróides de espectro por sub-banda, de *Subband Spectrum Centroids*) em algoritmos de VAL, buscando atingir melhores taxas de acerto de verificação.

Sob o ponto de vista institucional, a médio prazo, este trabalho contribui para a definição formal de uma nova metodologia de perícia e a subseqüente implementação de um sistema de apoio à decisão pericial no Instituto Militar de Engenharia.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

Serão apresentados no Capítulo 2 os conceitos básicos que servem de subsídio às técnicas analisadas, tais como o mecanismo de produção da fala, as características extraídas dos sinais de voz e o sincronismo por *pitch*. O Capítulo 3 apontará os conceitos básicos decorrentes da atividade de perícia em Fonética Forense, que permitirão a contribuição à proposta de uma nova metodologia. O Capítulo 4 discorrerá sobre as projeções aleatórias e os algoritmos genéticos na estimação do modelo GMM para as características extraídas dos sinais de voz; nesse capítulo, também serão enunciadas as técnicas de seleção de características, em especial o Discriminante de Fisher, com o intuito de buscar as

[2]Número de componentes gaussianas.

características mais discriminativas de locutores. O Capítulo 5 abordará a verificação automática de locutor, dando ênfase aos resultados de técnicas abordadas nos Capítulos 3 e 4. O Capítulo 6 tratará das conclusões e fará propostas de trabalhos futuros.

2 ASPECTOS CONCEITUAIS DA FALA

2.1 MECANISMO DE PRODUÇÃO DA FALA

O sistema de produção da fala do ser humano (BRAID, 2003) não se compara a nenhum outro sistema motor existente; são milhares de movimentos por minuto, desencadeados por cerca de oitenta músculos diferentes, responsáveis pela articulação do aparelho fonador, e que geram os padrões sonoros conhecidos simplesmente por fala. O locutor produz o padrão sonoro através do controle sobre a alteração da posição dos músculos e órgãos responsáveis pela fala, aliado à passagem de ar pelas cavidades e tubos existentes.

Visíveis ao exterior, compõem o aparelho fonador humano os lábios, a língua e os dentes. Entretanto, não são apenas estes dispositivos os responsáveis pela produção da fala; muito pelo contrário, existe uma concomitância entre o aparelho respiratório — pulmões, traquéia, laringe e pregas (ou cordas) vocais, faringe (parte comum aos aparelhos respiratório e digestivo, constituída por faringe oral e faringe nasal) e cavidade nasal — e a cavidade bucal, limitada pela mandíbula, pelos lábios, pelos dentes, pela língua, pelos palatos duro e mole (conhecidos popularmente como "céu-da-boca") e pela faringe.

Fisiologicamente, estes tubos e cavidades compõem três subsistemas (BRAID, 2003; MORISSON, 2003) atuantes de modo sucessivo na produção da fala, representados pela FIG. 2.1:

- Respiratório — O subsistema respiratório é o responsável pela passagem da corrente de ar dos pulmões para a traquéia e a laringe;

- Laringeal (ou laríngeo) — O subsistema laringeal (ou laríngeo), ou simplesmente laringe, situado na parte superior da traquéia, é o mais importante subsistema do aparelho fonador. Nele estão localizados a glote, a epiglote (válvula elástica que obstrui a glote durante a deglutição) e as cordas vocais. A parte mais importante deste subsistema é a glote, que consiste numa pequena abertura de forma triangular situada próxima ao pomo-de-adão. Graças à chegada do fluxo de ar vindo dos pulmões, a glote pode abrir-se ou fechar-se, bastando que as bordas das pregas vocais se afastem ou se aproximem. Com a glote aberta, o ar passa livremente sem fazer vibrar as cordas vocais produzindo um fone surdo ou não vozeado. Com o movimento cíclico de abertura da glote, causado pelo aumento da pressão do ar sub-

glotal vindo dos pulmões, e fechamento, causado pela força de recuperação elástica e pelo efeito de Bernoulli [3], as cordas vocais vibram numa freqüência fundamental característica (a ser definida na Seção 2.2), e o fone produzido, então, é dito sonoro ou vozeado. A taxa na qual a glote abre e fecha é controlada pela pressão de ar imposta pelos pulmões, pela tensão e rigidez das cordas vocais e pela área da abertura glotal em condições de repouso. Resumindo, o subsistema laringeal é o responsável pela passagem da corrente de ar que pode provocar ou não a vibração das cordas vocais.

- Supralaringeal (ou supralaríngeo) — Passagem dos pulsos ou da corrente contínua de ar pela faringe, sujeitos a obstruções ou constrições em vários pontos de articulação (nas cavidades nasal e bucal). Este sistema completa o mecanismo da produção da fala.

O trato vocal compreende os subsistemas laringeal e supralaringeal, por ser a região situada desde as pregas vocais até as extremidades da cavidade nasal (as narinas) e da cavidade bucal (os lábios) (BRAID, 2003).

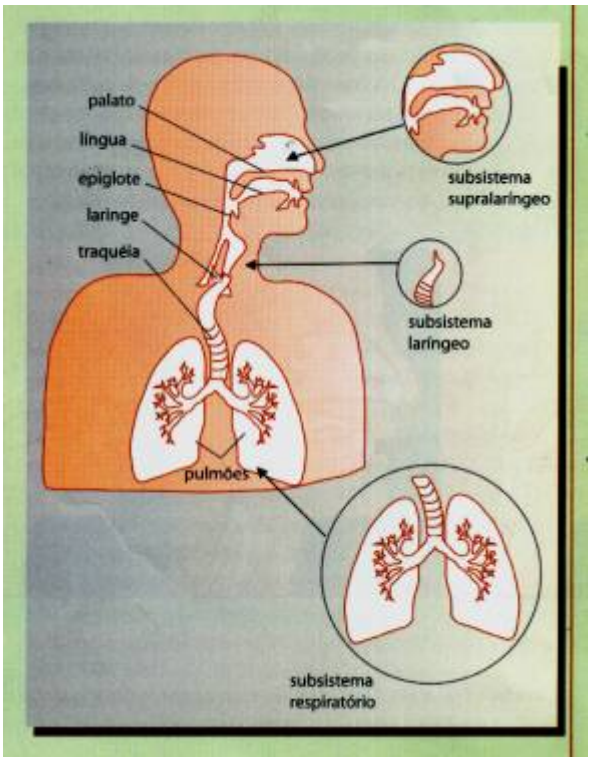


FIG.2.1: Os sistemas respiratório, laringeal e supralaringeal — figura extraída de (MORISSON, 2003).

[3]Tendência de um orifício se fechar devido à redução da pressão quando da passagem de um fluido.

A fonte sonora pode ser classificada como periódica ou aperiódica (BRAID, 2003). Quando o som produzido é vocálico, obrigatoriamente ocorre a vibração das pregas vocais, produzindo um fluxo de ar periódico. Quando o locutor profere uma consoante, o fluxo de ar pode ser puramente aperiódico (consoantes surdas — por exemplo: [p],[t],[k],[f],[s]) ou pode haver combinação de fluxo de ar periódico e aperiódico (consoantes sonoras — por exemplo: [b],[d],[g],[v],[z],[j]).

Quanto à articulação, as consoantes se classificam em função das particularidades envolvendo os pontos e modos de articulação (BRAID, 2003). O ponto de articulação é o local do aparelho fonador onde o som é efetivamente produzido (chamado de constrição) — lábios, dentes, língua, alvéolos (ossos onde se sustentam os dentes incisivos superiores), palato (parte superior da cavidade bucal), úvula, faringe e glote; o modo de articulação se refere às modificações do fluxo de ar durante a produção da consoante no interior da cavidade bucal.

As vogais também estão sujeitas à articulação (BRAID, 2003); ao contrário das consoantes, dependentes de pontos de articulação (constrição) para obstrução do som e sua conseqüente produção, as vogais dependem fortemente da posição da língua e dos lábios durante sua produção. Além disso, quando há encontros vocálicos — ditongos e tritongos — ocorre efeito de transição do posicionamento do trato vocal do início ao fim da combinação das vogais.

Aliado ao efeito de articulação, ocorre o fenômeno da coarticulação, definido como a interação de segmentos de fala, provocando um efeito transicional diferente do percebido na emissão dos fones de forma isolada, causando um certo ajuste temporal na fala.

2.2 ANÁLISE ACÚSTICA DA FALA

A análise acústica da fala consiste em expressar, através de métodos matemáticos, estatísticos e gráficos, o fenômeno sonoro desencadeado pela produção da voz causada pelo locutor, nos domínios do tempo, da freqüência, ou ambos. Esta análise é realizada com base na voz do locutor registrada em uma mídia qualquer. O modelo acústico trata da forma de onda em si, da característica da freqüência fundamental derivada do mecanismo ondulatório e do modelo fonte-filtro.

Fisicamente, a forma de onda de um sinal de voz (TITZE, 1994; BRAID, 2003; QUATIERI, 2002) representa o deslocamento do ar proveniente dos pulmões, desencadeando a transmissão de energia mecânica de uma molécula de ar para outra que, devido à elasticidade do ar, causa um movimento vibratório de moléculas. Esta vibração pode ser

transmitida para o ouvido de um interlocutor onde, convertida em estímulos nervosos, é recebida pelo cérebro e então percebida. De forma análoga, o som da fala pode ser captado por um microfone, convertido em sinais elétricos e posteriormente processado.

Uma das características mais importantes de um sinal de fala, sob os pontos de vista físico e perceptual, é a freqüência fundamental. Ela é oriunda do fluxo de ar que, após sofrer a excitação da vibração das pregas vocais, se transforma em pulsos de ar; destes pulsos, a freqüência fundamental é o valor do menor componente periódico medido em Hertz (Hz). Este valor, sob a ótica perceptual, recebe o nome de *pitch* (TITZE, 1994). Para um ouvinte, um *pitch* maior ou menor significa, perceptualmente falando, que o som escutado é mais agudo ou grave, respectivamente. Atribui-se ao *pitch* grande importância quanto à prosódia; a percepção acústica de uma frase como sendo afirmativa, exclamativa ou interrogativa se deve à variação do *pitch* no domínio do tempo, que implica perceptualmente na variação da melodia da voz. Da mesma forma, *pitch* mais aguda é uma condição necessária, porém não suficiente, para se detectar se a voz do locutor é masculina ou feminina. A FIG. 2.2 mostra a estimativa de marcas de períodos de *pitch* para o ditongo [iu] da palavra "rio" proferida por um locutor masculino.

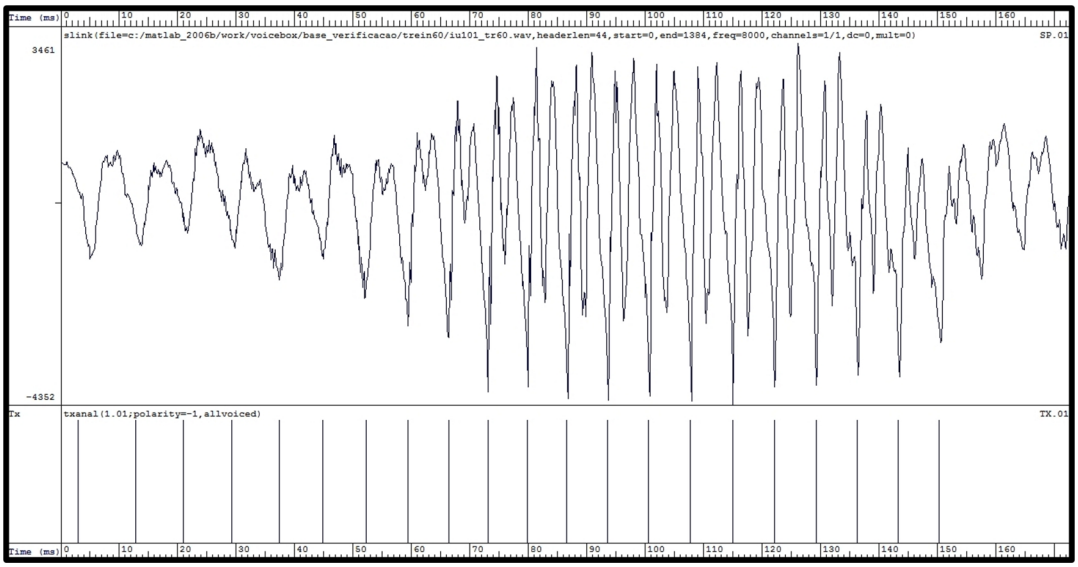


FIG.2.2: Marcas de estimativa de período de *pitch*.

Fisiológica e acusticamente, os três sistemas citados na Seção 2.1 (respiratório, laringeal, supralaringeal) participam diretamente da distinção de um som quanto à sonoridade, à nasalidade e às constrições. O sistema respiratório é responsável pela passagem do ar dos pulmões ao sistema laringeal; naquele, ainda não existe a distinção entre fone sonoro ou surdo. Esta distinção apenas se dá no sistema laringeal, onde a vibração das

pregas vocais transforma o fluxo de ar contínuo em pulsos de ar, impondo o quesito da sonoridade ao fone; caso contrário, o fluxo de ar permanecerá contínuo, à mercê de sofrer obstruções (constrições) ao longo do trato vocal. Assim como não existe a distinção entre "sonoro" e "surdo" no sistema respiratório, não existe a distinção de um fone para outro no sistema laringeal; esta distinção se dará apenas no sistema supralaringeal, dependendo da configuração geométrica do trato vocal (BRAID, 2003).

Por fim, cabe ao sistema supralaringeal, e somente a ele, o processamento do fluxo de ar, constituindo o modelo fonte-filtro para sinais de voz, com características peculiares para vogais, consoantes nasais e consoantes não-nasais:

- Para vogais, o sistema supralaringeal age como um filtro sobre o espectro em freqüência dos pulsos de ar oriundos da laringe (pulsos glotais), provocando maior concentração de energia em certas freqüências devido à configuração geométrica das cavidades por onde o ar passa (nasal ou oral). Essas freqüências são denominadas de **formantes**. Os formantes são causados pela ressonância dos pulsos glotais no trato vocal (vide EQ. 2.8), e são diretamente influenciados pela posição dos órgãos fonadores e pela coarticulação. Por exemplo, a altura da língua afeta o valor do primeiro formante; o avanço ou o recuo da língua, o segundo formante. Os formantes serão estudados, em detalhes, no Capítulo 3;

- Para consoantes fricativas — por exemplo [v], [f] e [j] — o fluxo de ar contínuo encontra no sistema supralaringeal um filtro; o ponto de constrição causa turbulência, sob forma de ruído, devida à obstrução parcial do trato vocal em algum ponto à frente da glote, independente de haver vibração das pregas ou não. A filtragem ocorre antes do ponto de constrição. Embora o comportamento do ar não seja periódico, há também a ocorrência de formantes.

- Para consoantes nasais, que recebem a soma de fluxos de ar periódicos e aperiódicos (sendo, portanto, sonoras), o modelo de fonte é análogo ao modelo das vogais. Neste caso, ao contrário das vogais, a função de filtro é desempenhada por ambas as cavidades, oral e nasal (TITZE, 1994; BRAID, 2003).

Com o objetivo do bom entendimento das características dos sinais de voz, serão detalhados, nas seções seguintes, alguns conceitos relativos a processamento digital de sinais de voz e modelos do trato vocal.

2.3 ANÁLISE NO DOMÍNIO DA FREQÜÊNCIA EM TEMPO CURTO

A grande motivação da análise em tempo curto é a necessidade de se adaptar conceitos matemáticos e estatísticos, até então concebíveis a tempo longo (como por exemplo, potência de um sinal), a uma forma limitada no tempo (DELLER, 2000). Uma prova disto é a detecção sonoro/surdo, baseada na energia de um trecho de um sinal de voz $s(n)$. Presume-se que trechos sonoros de sinais de voz possuam maior energia do que trechos surdos[4]. Desta forma, estima-se para cada trecho de sinal de voz (considerado estacionário) (DELLER, 2000) o valor esperado da energia, dado por $E[s^2(n)]$.

Por definição, um quadro (ou *frame*) de um sinal de voz, parte intrínseca da análise em tempo curto, é dado por uma seqüência de N_w amostras do sinal $s(m)$, tal que $m \in [n, n + N_w - 1]$, janelado no tempo. Este janelamento, de uma forma prática, pode ser inicializado através do deslocamento do quadro do sinal de voz em n amostras, tal que se obtenha o sinal $s(m + n)$ (QUATIERI, 2002); após o deslocamento, o quadro, representado por $s_n(m)$, é multiplicado pela janela $w(m)$, tal que $m \in [0, N_w - 1]$. O quadro de sinal de voz janelado é expresso matematicamente pela EQ. 2.1. A janela $w(m)$ pode ser de diversos tipos, como por exemplo — retangular, Hamming, Hanning, Kaiser e Bartlett. Além disso, as N_w amostras do quadro corresponderão a um intervalo de tempo entre 20 e 40 ms, para o qual se considera o trato vocal praticamente sem variações (DELLER, 2000; QUATIERI, 2002). Para um melhor entendimento, a FIG. 2.3 e a FIG. 2.4 ilustram um exemplo do processo de janelamento com janela de Hamming $w(m)$ atuando sobre um quadro do sinal de voz $s(m)$, de modo que

$$s_n(m) = s(m + n)w(m). \tag{2.1}$$

Da formulação da EQ. 2.1, além da necessidade de se representar o espectro do sinal de voz em trechos curtos (sob forma de espectrograma), surgiu o conceito da STFT (Transformada de Fourier em tempo curto - *Short Time Fourier Transform*), representada pela EQ. 2.2

$$\mathbf{S}(\omega; m) = \sum_{m=-\infty}^{\infty} s_n(m)e^{-j\omega m}$$

[4]Neste compêndio, *trecho surdo* é um intervalo de voz composto apenas por fones surdos, e *trecho sonoro* é um intervalo de voz composto apenas por fones sonoros

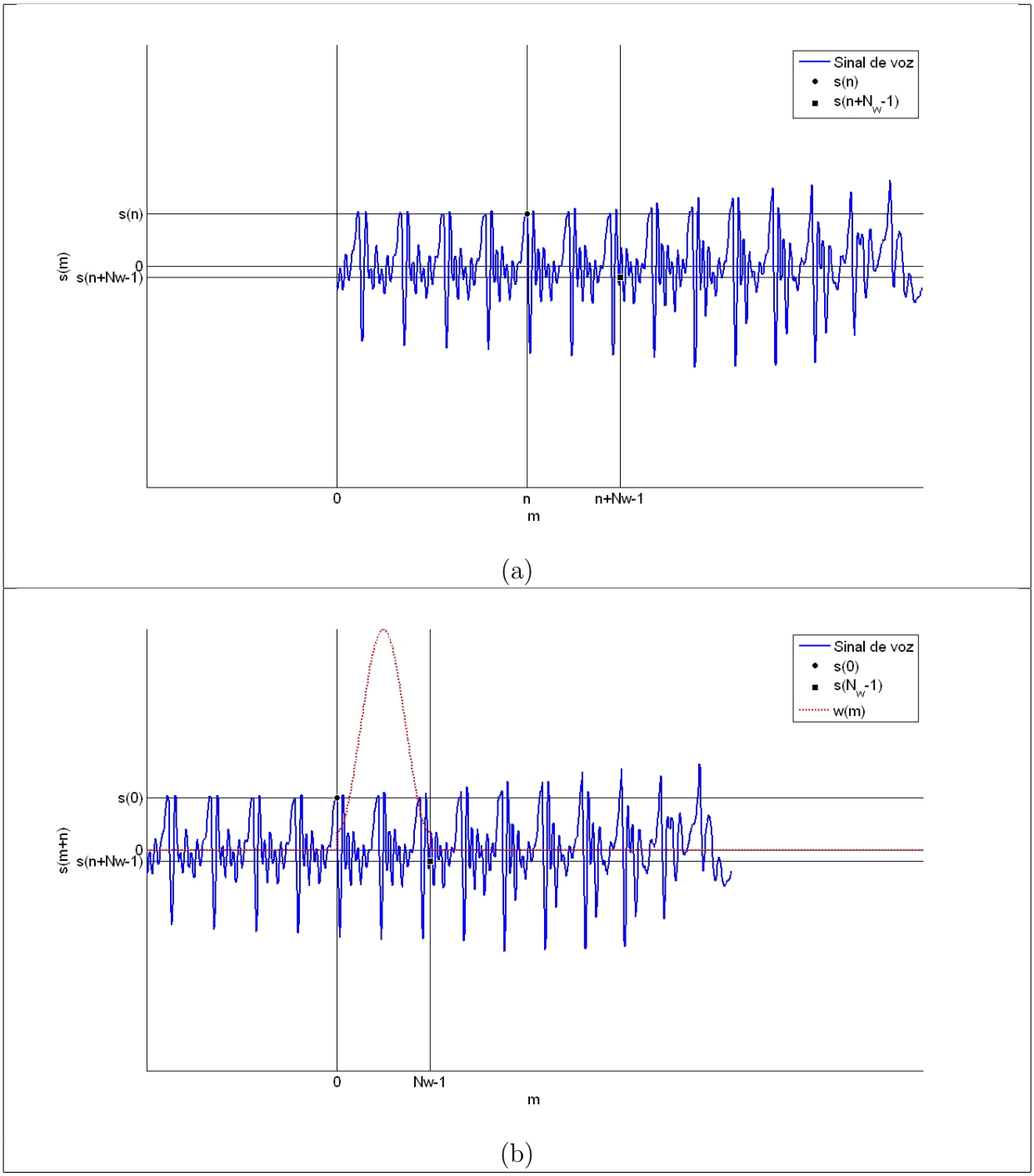


FIG.2.3: Processo de janelamento: (a) Sinal de voz $s(m)$; (b) Sinal de voz $s(m+n)$ e janela de Hamming $w(m)$.

$$= \sum_{m=0}^{N_w-1} s_n(m)e^{-j\omega m} \tag{2.2}$$

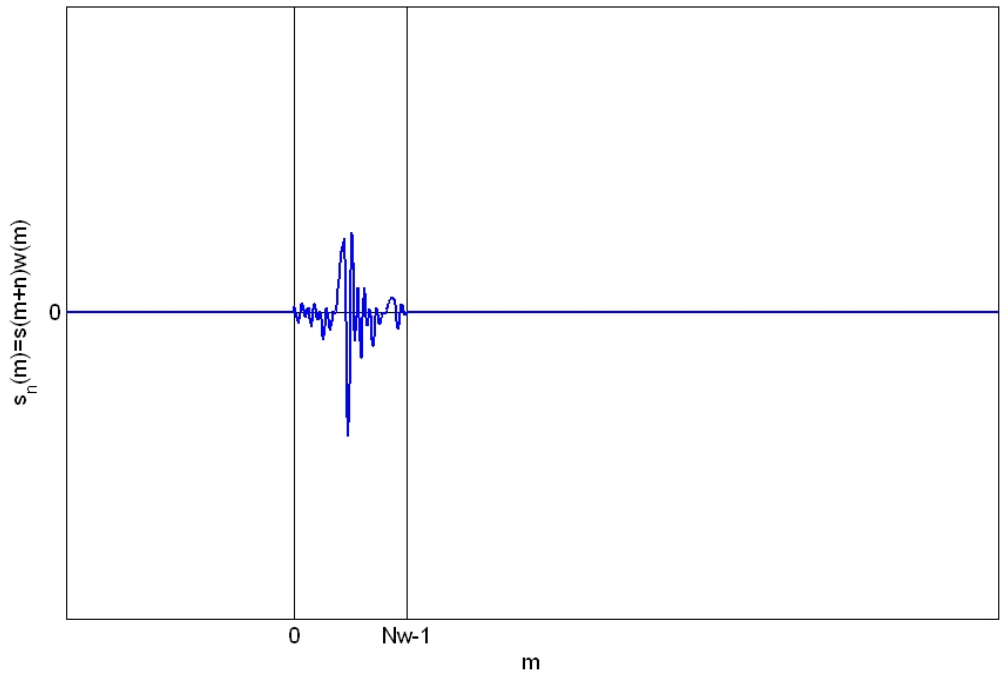


FIG.2.4: Quadro $s_n(m)$ do sinal de voz após ser janelado por $w(m)$.

2.4 MODELO DE TUBOS

O trato vocal pode ser modelado como um tubo reto e cilíndrico, cuja extremidade fechada representa a glote (fonte de energia) e cuja extremidade aberta representa os lábios (saída de ar e, conseqüentemente, da voz do locutor). Pode ser comprovada fisicamente, para uma vogal neutra /ə/ (PRATOR JR, 1972) — equivalente, na língua portuguesa, a uma vogal central meio-fechada como o /ə/ final da palavra "cas**a**" (SILVA, 2005) — e assumindo o trato vocal como um ressonador de ar aberto em uma das extremidades e fechado na outra, a validade da EQ. 2.3, em que $n \in \mathbb{N}^*$, v é a velocidade do som no ar e l é o comprimento do tubo (TITZE, 1994; QUATIERI, 2002). De acordo com o modelo da EQ. 2.3, à velocidade do ar de 340 m/s, para um locutor homem com l=17 cm, as freqüências de ressonância F_n, ou freqüências fundamentais, são dadas por 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz, e assim por diante.

$$F_n = \frac{(2n-1)\,v}{4l} \tag{2.3}$$

Assume-se esses valores de freqüências de ressonância como sendo os valores aproximados de formantes da vogal neutra. Contudo, esses valores são calculados com um modelo linear e invariante no tempo do trato vocal, que não corresponde à realidade

(QUATIERI, 2002). Além disso, para as outras vogais, este modelo de tubo se modifica de acordo com os pontos de articulação (FIG. 2.5), concebendo novos valores para as freqüências de ressonância, pois o modelo do tubo pode ser imaginado como trechos de tubos concatenados de dimensões distintas, sujeitos a modificações de configuração geométrica devidas ao movimento dos lábios e da língua.

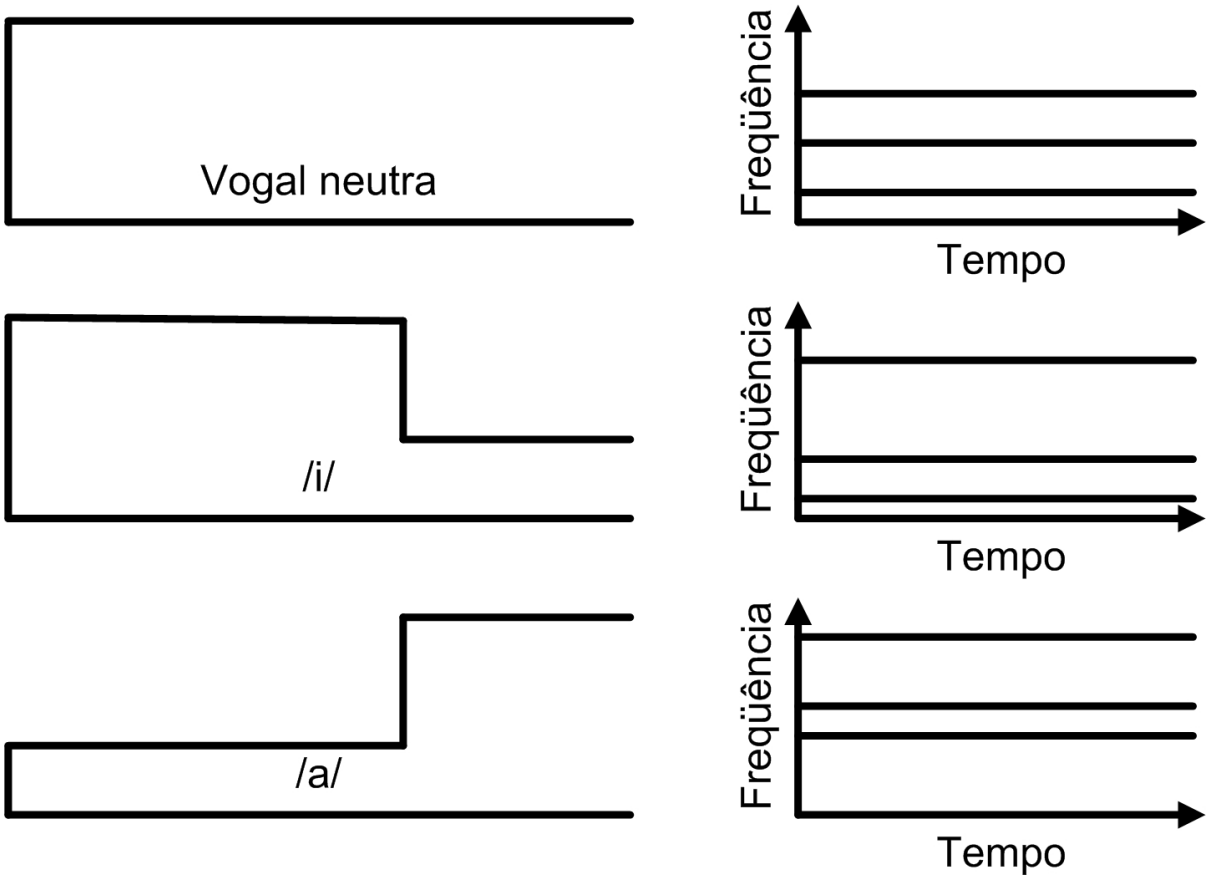


FIG.2.5: O modelo de tubo ressonante para três vogais diferentes (MORISSON, 2003).

2.5 CARACTERÍSTICAS FÍSICAS DOS SINAIS DE VOZ

Esta seção visa apresentar brevemente o conceito das características dos sinais de voz derivadas diretamente da modelagem física do trato vocal, comentando os aspectos peculiares a cada uma e seus métodos mais conhecidos de estimação.

2.5.1 FREQÜÊNCIA FUNDAMENTAL (*PITCH*)

O tempo de duração de um pulso glotal é conhecido como *período de pitch*. Obviamente, o inverso desta grandeza é conhecido como *pitch* ou freqüência fundamental. Para efeito de análise de sinais de voz, a *pitch* nada tem a ver com o conceito de percepção psicoacústica de sons musicais explorado em processamento de áudio; é uma característica física relacionada à vibração das pregas vocais.

O valor da freqüência fundamental pode variar devido à pressão do ar exercida sobre a glote (dependente das dimensões da glote), à tensão muscular exercida sobre as pregas vocais, ao peso e ao tamanho das pregas vocais (estas últimas são peculiares a cada locutor).

Aspectos emocionais (QUATIERI, 2002) também podem revelar características do locutor em termos de perícia (podem auxiliar ou mascarar a identidade do locutor). O nervosismo (*stress*) pode implicar em tensão parcial das pregas vocais, provocando uma voz com *pitch* alto e irregular (voz rangida - *creaky voice*). O efeito de entorpecentes, por exemplo, pode provocar uma *pitch* baixa e irregular (voz chiada - *fry voice*) pois as pregas vocais estarão mais pesadas e lentas.

Aspectos patológicos (QUATIERI, 2002) também fazem provocam padrões da *pitch* fora do normal. Um exemplo de patologia que pode caracterizar um locutor é a diplofonia, que consiste na geração de vibrações adicionais das pregas vocais, gerando pulsos glotais $\overline{g}(n)$ expressos abaixo como uma soma de um pulso glotal secundário modelado com atraso n_0 e atenuado em relação aos pulso glotal principal $g(n)$:

$$\overline{g}(n) = g(n) + \alpha g(n - n_0) \tag{2.4}$$

Um exemplo simples de método de extração da *pitch* é o método da autocorrelação. Para um sinal de voz $x(n)$, define-se o estimador de autocorrelação em tempo curto pela EQ. 2.5 (DELLER, 2000). Qualitativamente, a autocorrelação mede o grau de similaridade entre o mesmo sinal sem defasagem e defasado de um total de η amostras, dentro de uma janela terminando na amostra m para o caso de tempo curto. Para trechos sonoros de sinais de voz, fica nítido que a autocorrelação assumirá valores máximos quando houver máxima similaridade com respeito a periodicidade, construindo-se desta forma um detector e rastreador de *pitch* eficiente.

$$r_s(\eta; m) = \frac{1}{N} \sum_{n=m-N+|\eta|+1}^{m} x(n)x(n - |\eta|) \tag{2.5}$$

2.5.2 FORMANTES

O trato vocal, constituído pelas cavidades oral e nasal, pode ser modelado por um filtro LIT [5] (Linear Invariante no Tempo) *all-pole* (apenas com pólos) (QUATIERI, 2002) modelado pela função de transferência

$$\mathcal{V}(z) = \frac{A}{\prod_{i=1}^{N}(1 - c_i z^{-1})(1 - c_i^* z^{-1})}, \tag{2.6}$$

reduzindo-se, por frações parciais, a

$$\mathcal{V}(z) = \sum_{i=1}^{N} \frac{B_i}{(1 - c_i z^{-1})(1 - c_i^* z^{-1})}, \tag{2.7}$$

dependente de sua conformação geométrica. Quando ocorre qualquer mudança de disposição geométrica no trato vocal, conseqüentemente é gerado um modelo diferente de cavidade ressonante e, conseqüentemente, são alteradas as freqüências de ressonância, chamadas de formantes, designadas por $F_{\mathbf{1}}$ a $F_{\mathbf{n}}$, ordenadas do menor para o maior valor.

É interessante ressaltar que a maior parte dos fones é percebida pois sua resposta em freqüência $\mathcal{Y}(z)$ é denotada pela convolução da função de transferência do filtro $\mathcal{V}(z)$ que representa o trato vocal com a função de transferência do filtro $\mathcal{H}(z)$ que representa a glote:

$$\mathcal{Y}(z) = \mathcal{H}(z)\mathcal{V}(z). \tag{2.8}$$

Espectralmente, o movimento de lábios, dentes, língua e maxilares torna a fonte "colorida" (QUATIERI, 2002).

Sob o aspecto pericial, diferentes locutores apresentam diferentes configurações geométricas do trato vocal para produzir um mesmo fone. Os formantes de maior ordem - normalmente F_3 e F_4 - são mais dependentes da geometria do trato e, portanto, são de maior valia para reconhecimento de locutor (NOLAN, 2005).

Os formantes podem ser estimados, dentre outras técnicas, por:

- **Método da autocorrelação:**

  A estimativa (DELLER, 2000) se baseia nos (M) coeficientes LPC (*Linear Prediction Coefficients* — coeficientes de predição linear) extraídos pelo Método de

[5] Sejam pares de entrada e saída do filtro $[x_i(n), y_i(n)]$, $i = 1, \ldots, N$. Para filtros LIT, também serão pares de entrada e saída $[\sum_{i=1}^{N} C_i x_i(n), \sum_{i=1}^{N} C_i y_i(n)]$ e $[x_i(n - n_0), y_i(n - n_0)]$, sendo C_i constantes.

Levinson-Durbin, representado da EQ. 2.9 à EQ. 2.13. Cada termo da EQ. 2.9 representa a matriz de autocorrelação $\mathbf{R}_n$, o vetor de coeficientes LPC $\hat{\mathbf{a}}$ e o vetor de autocorrelações $\mathbf{r}_n$ representados pelas equações EQ. 2.10, 2.11 e 2.12, respectivamente. A EQ. 2.9 atende ao critério de mínimo erro médio quadrático de predição

$$\mathbf{R}_n\hat{\mathbf{a}} = \mathbf{r}_n, \tag{2.9}$$

no qual

$$\mathbf{R}_s = \begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \cdots & r_n(M-1) \\ r_n(1) & r_n(0) & r_n(1) & \cdots & r_n(M-2) \\ r_n(2) & r_n(0) & r_n(1) & \cdots & r_n(M-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(M-1) & r_n(M-2) & r_n(M-3) & \cdots & r_n(0) \end{bmatrix}, \tag{2.10}$$

$$\hat{\mathbf{a}} = \begin{bmatrix} \hat{a}(1) \\ \hat{a}(2) \\ \hat{a}(3) \\ \vdots \\ \hat{a}(M) \end{bmatrix}, \tag{2.11}$$

e

$$\mathbf{r}_n = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(M) \end{bmatrix}. \tag{2.12}$$

O valor teórico de $r_n(m)$ é dado por:

$$r_n(m) = E[x(n)x(n-m)]. \tag{2.13}$$

Contudo, cada termo da matriz da EQ. 2.10 é dado pela estimativa da autocorrelação (EQ. 2.13) do sinal de voz $x(n)$, calculado nesse método pela EQ. 2.14. Através desta equação, nota-se a igualdade da autocorrelação, em termos práticos,

à convolução do quadro do sinal de voz com ele mesmo, sendo $s_n(m)$ já definido na EQ. 2.1.

$$\begin{aligned} r_n(\tau) &= \sum_{m=0}^{N_w-1-\tau} s_n(m)s_n(m+\tau) \\ &= s_n(\tau) * s_n(-\tau) \end{aligned} \tag{2.14}$$

A EQ. 2.15 simboliza a Transformada Discreta de Fourier (DFT, de *Discrete Fourier Transform*) que representa o espectro do quadro $s_n(m)$ do sinal de voz. Esse espectro pode ser obtido do vetor $\hat{\mathbf{a}}$ de coeficientes LPC, cujos elementos são representados, com respeito ao quadro do sinal de voz, pelos termos $-\hat{a}(n;m)$, tal que $n = 1, 2, \ldots, M$. Tomando a DFT $\mathcal{X}(k;m)$, então

$$\mathcal{X}(k;m) = 1 - \sum_{n=1}^{M} \hat{a}(n;m)e^{-j(\frac{2\pi}{N})kn}, k = 0, 1, \ldots, N-1, \tag{2.15}$$

sendo a estimativa dos formantes os pontos, em freqüência, dos máximos consecutivos do inverso de $\mathcal{X}(k;m)$ em amplitude.

- **Método da covariância:**

  O método da covariância (QUATIERI, 2002) difere do método da autocorrelação por considerar as amostras do sinal atendendo ao critério do erro quadrático médio mínimo apenas no intervalo de predição, e não em todo o sinal ao contrário do método da autocorrelação. Assumindo o sinal $s(n)$ modelado somente com pólos (*all-pole*), o método da covariância estima os parâmetros exatos para um trecho do sinal de voz em tempo curto.

  A EQ. 2.9 assume a forma da EQ. 2.16. A matriz $\mathbf{S}_n$ descrita pela EQ. 2.17 é constituída por vetores de amostras do sinal de voz $s(m)$, formando uma base de um espaço vetorial (QUATIERI, 2002). O vetor de M coeficientes LPC $\hat{\mathbf{a}}$ é definido da mesma forma que na EQ. 2.11. O vetor $\mathbf{s}_n$ da EQ. 2.18 é constituído pelas $(2p+1)$ amostras consecutivas do sinal de voz $s(m)$ compreendidas no intervalo $m \in [n-p, n+p]$.

$$\mathbf{S}_n\hat{\mathbf{a}} = \mathbf{s}_n \tag{2.16}$$

$$\mathbf{S}_n = \begin{bmatrix} s(n-p+0-1) & s(n-p+0-2) & \cdots & s(n-p+0-M) \\ s(n-p+1-1) & s(n-p+1-2) & \cdots & s(n-p+0-M) \\ s(n-p+2-1) & s(n-p+2-2) & \cdots & s(n-p+0-M) \\ \vdots & \vdots & \ddots & \vdots \\ s(n+p-1) & s(n+p-2) & \cdots & s(n+p-M) \end{bmatrix} \tag{2.17}$$

$$\mathbf{s}_n = \begin{bmatrix} s(n-p) \\ s(n-p+1) \\ s(n-p+2) \\ \vdots \\ s(n+p) \end{bmatrix} \tag{2.18}$$

A EQ. 2.16 pode ser reescrita como

$$\begin{aligned} \left(\mathbf{S}_n^T \mathbf{S}_n\right) \hat{\mathbf{a}} &= \mathbf{S}_n^T \mathbf{s}_n \\ \hat{\mathbf{a}} &= \left(\mathbf{S}_n^T \mathbf{S}_n\right)^{-1} \mathbf{S}_n^T \mathbf{s}_n, \end{aligned} \tag{2.19}$$

onde se pode notar o termo $\left(\mathbf{S}_n^T \mathbf{S}_n\right)^{-1} \mathbf{S}_n^T$, conhecido como a matriz pseudoinversa (STRANG, 1988) da matriz $\mathbf{S}_n$.

2.5.3 COEFICIENTES *MEL-CEPSTRAIS* (MFCC)

Para explicar os coeficientes *mel-cepstrais*, ou MFCC (de *Mel-Frequency Cepstrum Coefficients*), é necessário introduzir o mecanismo de percepção dos sons pelo sistema auditivo humano. O ser humano possui um mecanismo de percepção da audição não linear (DELLER, 2000) modelado pela escala *mel*. Um *mel* é definido como uma unidade de medida da freqüência de um tom percebida pelo ouvido, ou melhor dizendo, da *pitch* percebida (neste caso, *pitch* se refere ao conceito psicoacústico de som percebido pelo ouvido) (DELLER, 2000). Desta forma, o valor da freqüência percebida pelo ouvido humano não corresponde linearmente ao valor de freqüência em Hz normalmente utilizado. Para este modelo, a percepção de uma determinada freqüência de valor Ω_0 é influenciada pela energia em uma banda crítica de freqüências em torno do valor de Ω_0 (ALLEN, 1985). A variação destas freqüências do ouvido humano, na realidade, funciona como um banco de filtros linearmente espaçados para baixas freqüências em Hz e logaritmicamente

espaçados para altas freqüências, representado pela FIG. 2.6. Fant, por exemplo (FANT, 1959), sugere a conversão da EQ. 2.20 de freqüências em hertz para freqüências em *mel*, representada também pela FIG. 2.7.

$$f(mel) = 1000 \log_2 \left(1 + \frac{f(\text{Hz})}{1000}\right) \tag{2.20}$$

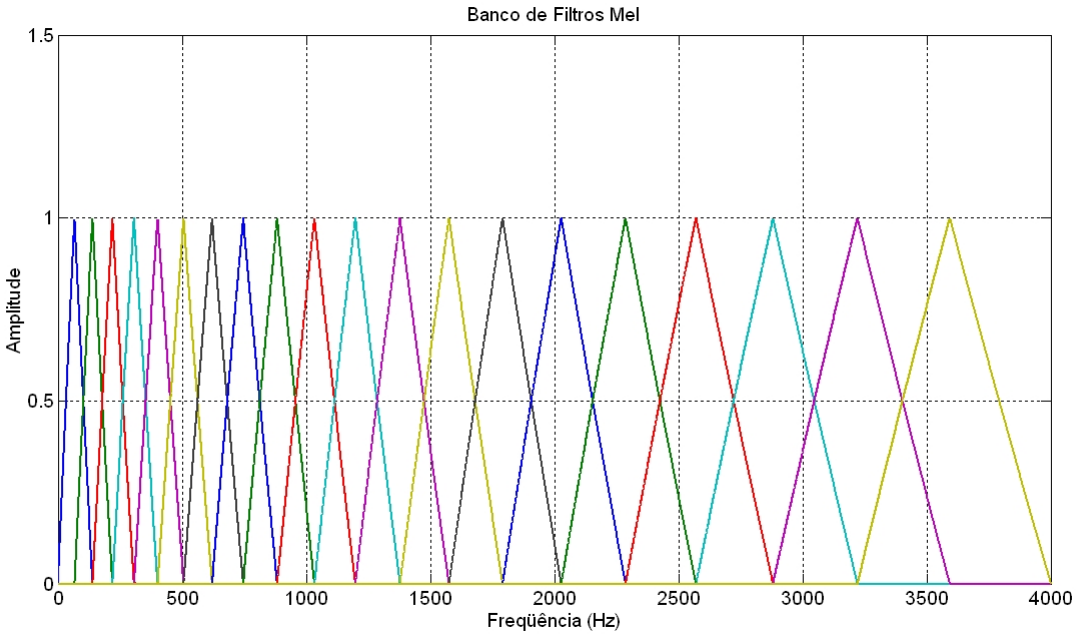


FIG.2.6: Banco de filtros que representa o modelo auditivo humano.

Quanto às informações fonéticas, os coeficientes MFCC surgiram da necessidade de comprimir, de alguma forma, os dados da fala de um locutor, de modo a eliminar o máximo de informação não-pertinente à análise fonética, realçando os atributos que efetivamente contribuissem para manter a identidade fonética a ser detectada (DAVIS, 1980). Uma vez que o espaçamento logarítmico implica na supressão de informação espectral insignificante nas bandas de alta freqüência (DAVIS, 1980), as características fonéticas são preservadas, acarretando boa discriminação inclusive de sons consonantais (além, é claro, dos sons vocálicos). A ênfase dada às altas freqüências também contribui sobremaneira para a boa discriminação das consoantes, além de compensar a fraca amplitude do sinal nesta faixa do espectro, contribuindo para melhorar a eficiência dos sistemas de reconhecimento de locutor (sobretudo de VAL).

Além de se beneficiar dos princípios da percepção auditiva, a análise *mel-cepstral* se mostra bastante robusta aos efeitos de distorção convolucional produzida pelos canais de comunicações (QUATIERI, 2002). Esta robustez é associada também ao baixo número

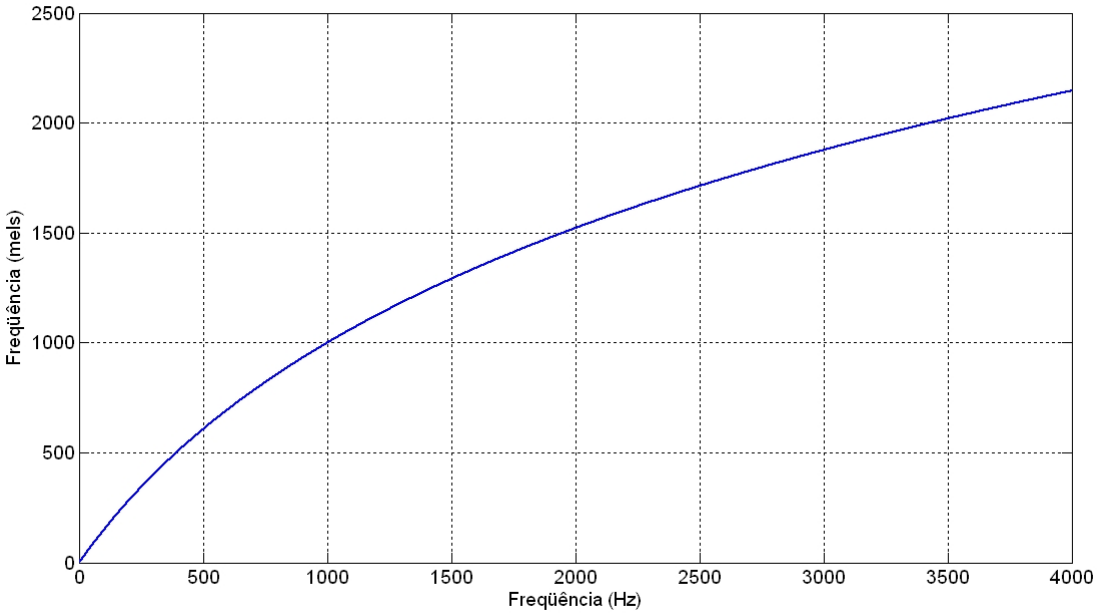


FIG.2.7: Conversão das freqüências em hertz para a escala *mel*.

de parâmetros considerados; tem sido demonstrado que 6 coeficientes já são suficientes para adquirir a maior parte das informações relevantes dos sinais de voz (DAVIS, 1980; ZENG).

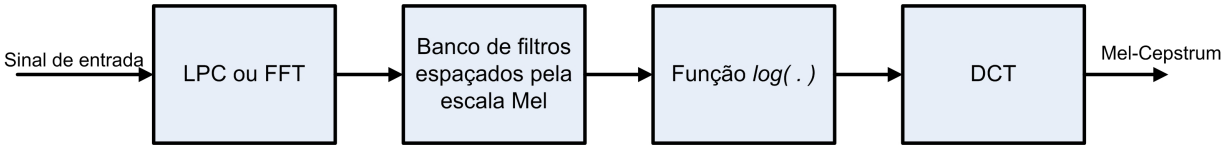


FIG.2.8: Diagrama em blocos que representa a extração dos coeficientes MFCC.

O diagrama em blocos da FIG. 2.8 mostra, de forma simplificada, a extração dos coeficientes MFCC. Inicialmente, deverá ser efetuado o cálculo da FFT ou do espectro suavizado do sinal de voz através dos coeficientes LPC. Em seguida, este espectro é filtrado através da sua multiplicação pela série de filtros composta pelo banco de filtros de freqüências críticas espaçados segundo a escala *mel*, de forma a simular a resposta em freqüência do ouvido humano. Após esta filtragem, é aplicada a função logarítmica aos componentes do espectro (conversão do domínio espectral para o domínio cepstral) e, em seguida, a transformada cosseno discreta (DCT, de *Discrete Cosine Transform*). A DCT permite a obtenção dos coeficientes no domínio da *qüefrência*[6]. Esta trans-

[6]Nome dado às freqüências no domíno cepstral (DELLER, 2000)

formação permite a compressão da informação espectral nos coeficientes de mais baixa ordem e promove uma descorrelação dos componentes espectrais na escala *mel*, uma vez que a DCT se aproxima da Transformada de Karhunen-Loeve (QUATIERI, 2002). Esta descorrelação se traduzirá em melhor modelagem GMM, objeto do Capítulo 5.

A **robustez ao canal**, comentada acima, também é uma qualidade das duas primeiras derivadas numéricas dos coeficientes MFCC, conhecidas como coeficientes *delta* (primeira derivada) e *delta-delta* (segunda derivada) (DELLER, 2000; QUATIERI, 2002). Os coeficientes delta são calculados pela diferença dos coeficientes MFCC extraídos de dois quadros de sinal de voz consecutivos). De forma semelhante, os coeficientes delta-delta são as diferenças dos coeficientes delta de dois quadros consecutivos. É comum, como será visto adiante no Capítulo 5, o emprego conjunto dos coeficientes delta e delta-delta e dos coeficientes MFCC na VAL visando ressaltar a **diferença dos aspectos dinâmicos de fala dos locutores envolvidos no processo pericial**.

2.5.4 TONALIDADE, SFM E SCF

A tonalidade, a SFM (de *Spectral Flatness Measure*) e a SCF (de *Spectral Crest Factor*) são medidas diretamente relacionadas com a planura (em inglês, *flatness*) do espectro de um sinal, originalmente empregadas em codificação (FLANAGAN, 1979), podendo ser adaptadas para sinais de voz (PEETERS). Para as atividades de perícia em Fonética Forense, é bem interessante discriminar vozes de locutores pela diversidade espectral, ressaltando a tendência de alguns locutores à maior concentração de energia em certas freqüências ou à maior presença de ruído aditivo no espectro. Além disso, por serem características perceptuais (HERRE), são importantes para a obtenção de resultados intuitivos e eficazmente discriminativos.

A característica SFM mede, qualitativamente, quanto um espectro se distancia de um espectro senoidal. Numericamente, é computada pela razão entre a média geométrica e a média aritmética do valor da energia das freqüências k em cada sub-banda B_i de K sub-bandas (250 a 500Hz, 500 a 1000Hz, 1000 a 2000 Hz e 2000 a 4000 Hz), dada pela EQ. 2.21

$$SFM(i) = \frac{\left[\prod_{k \in B_i} a(k)\right]^{1/K}}{\frac{1}{K}\sum_{k \in B_i} a(k)} \tag{2.21}$$

A característica SCF é computada pela razão entre o máximo valor dentro de cada banda e a média aritmética do valor de energia de cada banda, dada pela EQ. 2.22.

$$SCF(i) = \frac{\max\left[a(k \in B_i)\right]}{\frac{1}{K}\sum_{k \in B_i} a(k)} \tag{2.22}$$

A tonalidade (Ton) é derivada da característica SFM através da EQ. 2.23. Para sinais puramente senoidais (compostos de tons senoidais à mesma amplitude), a tonalidade é 1. Para sinais muito contaminados por ruído aditivo ou sinais com espectro pouco plano, a tonalidade é próxima de zero.

$$\begin{aligned} SFM_{dB} &= 10\log(SFM) \\ Ton &= \min\left(\frac{SFM_{dB}}{-60}, 1\right) \end{aligned} \tag{2.23}$$

2.5.5 CENTRÓIDES ESPECTRAIS POR SUB-BANDA (SSC)

Os centróides espectrais por sub-banda (SSC, de *Subband Spectrum Centroids*) surgiram (PALIWAL) com a necessidade da busca de maior poder de reconhecimento de voz, compensando os problemas de estimação de formantes (picos espectrais espúrios, picos próximos entre si) e de extração do *cepstrum* (sensível à distorção por ruído aditivo, podendo causar problema em casos reais de perícia). De acordo com (PALIWAL), os SSC são similares aos formantes, além de serem fácil e confiavelmente extraídos.

Para extrair os SSC, divide-se o espectro em tempo curto de cada janela do sinal de voz, entre as freqüências $f = 0$ e $f = f_s/2$ (sendo f_s a freqüência de amostragem do sinal) em um número fixo de sub-bandas e se computa o centróide através da média, para cada sub-banda, da densidade espectral de potência. Sejam o número de sub-bandas M e os intervalos de freqüência de cada sub-banda m ($m = 1, \ldots, M$) dados por $[l_m, h_m]$; calculam-se, então, os coeficientes SSC C_m pela EQ. 2.24, assumindo cada sub-banda filtrada por uma janela de formato $w_m(f)$, e a densidade espectral de potência de cada sub-banda dada por $P(f)$. A constante γ assume valor menor que 1, de modo a controlar a faixa dinâmica do espectro de potência. Normalmente se utiliza $\gamma = 0,5$.

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^{\gamma}(f) df}{\int_{l_m}^{h_m} w_m(f) P^{\gamma}(f) df} \tag{2.24}$$

2.6 CARACTERÍSTICAS COM SINCRONISMO DE *PITCH*

A extração de características dos sinais de voz com sincronismo de *pitch* deve obedecer ao diagrama de blocos da FIG. 2.9, que representa o funcionamento de um detector de

vozeamento (trechos sonoros) simples. Baseado na decisão deste detector, o sistema optará por qual metodologia de extração de características implementar - por janelas de 20 ms, para trechos surdos, ou janelas de sincronismo de *pitch*, para trechos sonoros.

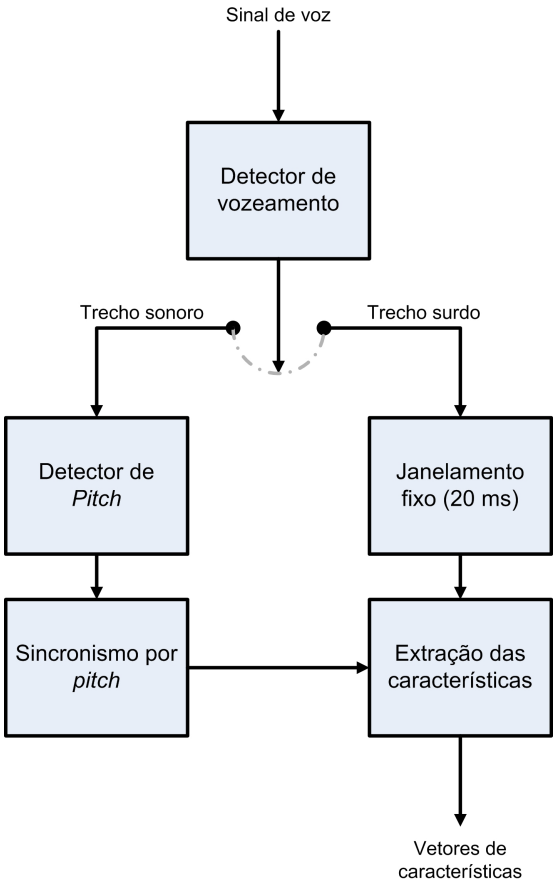


FIG.2.9: Diagrama básico da extração de características com sincronismo de *pitch*.

2.6.1 CRITÉRIO DE DETECÇÃO DE TRECHOS SONOROS DOS SINAIS DE VOZ

A necessidade de se discriminar os fones sonoros e surdos das gravações de voz (peças-motivo e peças-padrão) surge da necessidade do perito de buscar similaridades entre locutores através de características peculiares, isoladamente, a fones sonoros e surdos. Dentre estas características, pode-se destacar o contorno melódico de *pitch* e a análise de formantes em ditongos (caso sonoro).

Para a detecção dos trechos sonoros nas peças-motivo e nas peças-padrão, é necessário atribuir, dentro da gravação, quadro a quadro do sinal de voz, uma probabilidade P_{sonoro} (também conhecida como *voicing*; vide exemplo da FIG. 2.10) de aquele quadro corresponder a um intervalo de som sonoro. Pode-se inferir a correspondência do quadro analisado corresponde a um trecho sonoro se a probabilidade P_{sonoro} for maior que um

determinado limiar P_{limiar}; em outras palavras, caso $P_{sonoro} \geq P_{limiar}$, infere-se o rótulo de "sonoro" ao quadro do sinal de voz analisado. Caso contrário, infere-se o rótulo "surdo ou silêncio".

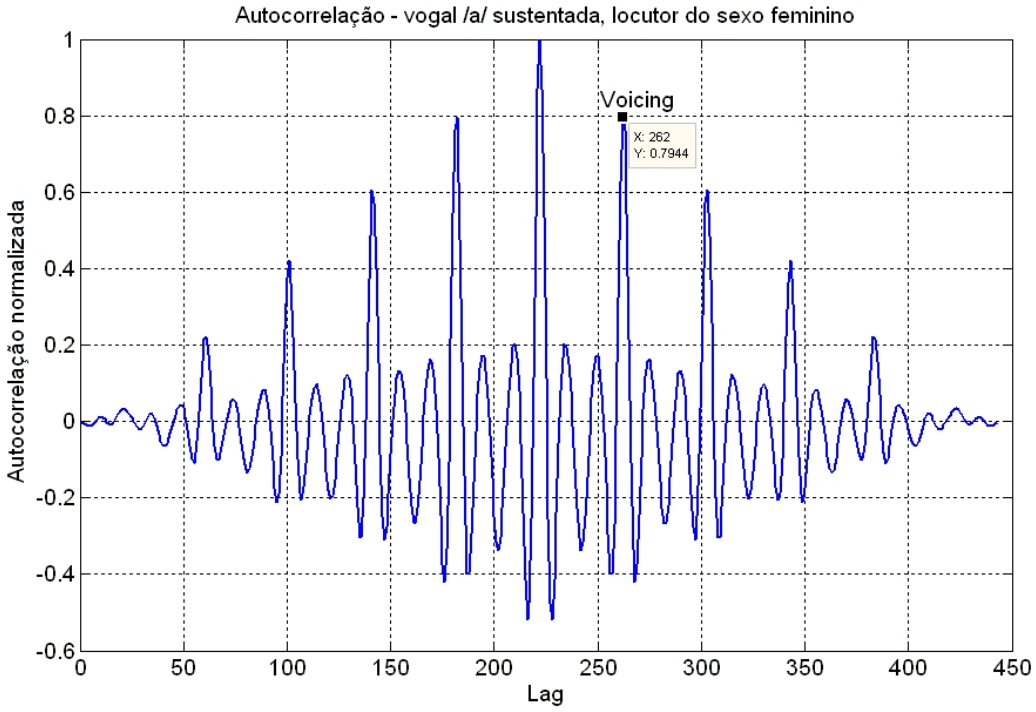


FIG.2.10: A grandeza P_{sonoro} (*voicing*) é dada pela razão do valor do pico principal pelo valor do primeiro pico secundário da autocorrelação de um quadro de sinal de voz. No gráfico acima, foi analisado um quadro do /a/ sustentado de um locutor do sexo feminino, com *voicing* de 0,7944.

Uma vez estimados os trechos sonoros do sinal de voz, efetua-se a etapa de sincronismo de *pitch*, lembrando que só faz sentido atribuir a *pitch* aos trechos sonoros, por serem estes decorrentes da vibração das pregas vocais.

2.6.2 DETECTOR DE *PITCH* E SINCRONISMO

Partindo dos trechos sonoros previamente estimados dos sinais de voz, efetuam-se duas tarefas subseqüentes associadas à *pitch* - a detecção da *pitch* e a extração das características com sincronismo da *pitch*.

A detecção da *pitch* consiste em estimar os períodos de *pitch* referentes aos trechos sonoros dos sinais de voz considerados no processo pericial. Esta estimação de processos é denominada *Point Process* (BOERSMA, 2001).

De posse dos intervalos de período, faz-se a extração das características de interesse sincronizados com os intervalos considerados. A estas características se dá o nome

de características *pitch*-síncronas. Sistemas de VAL baseados em características *pitch*-síncronas costumam possuir menores taxas de erro de verificação do que sistemas não-síncronos (EZZAIDI). Em detalhes, para cada período de *pitch* estimado, extrai-se um vetor de características.

O modelo quase-estacionário de sinais de voz, empregando quadros com duração fixa de 20 ou 30 ms com sobreposição de 10 ou 15 ms respectivamente, apresenta distorção causada pelo descasamento da posição do quadro com respeito à *pitch*. Um caso interessante em perícia é a diversidade de contornos melódicos de *pitch* (caracterizada na Seção 3.4) que pode ocorrer com as peças-padrão e a peça-motivo do mesmo locutor, causando distorção na extração das características (ZENG). De maneira a compensar as distorções surgidas quadro a quadro com o modelo de janelamento fixo, a extração de características pitch-síncronas é benéfica à atividade pericial, por conseguir uma definição espectral melhor do que pela extração não-síncrona (KIM; MORGAN; LEE; ZENG).

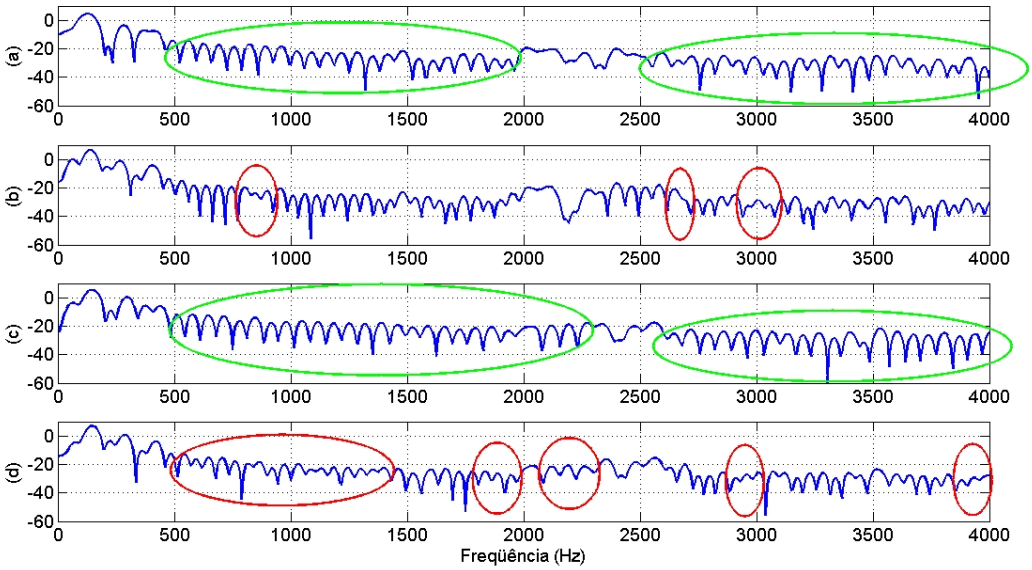


FIG.2.11: Amplitude de trechos de sinais de voz em diferentes configurações de segmentação — (a),(c): duplas adjacentes de períodos consecutivos de *pitch*; (b),(d): idem (a) e (c) + 3,75ms; a resolução dos harmônicos (em verde) é melhor em (a) e (c) (casamento com a *pitch*) do que em (b) e (d) (em vermelho).

A FIG. 2.11 ilustra o efeito do descasamento da extração do espectro em tempo curto em função do sincronismo de *pitch*. Os espectros representados em (a) e (c) são retirados de dois períodos de *pitch* consecutivos, com sobreposição de um período de *pitch*. Os espectros representados em (b) e (d) são extraídos de dois trechos do mesmo sinal de voz sincronizados pelos mesmos períodos de *pitch*, porém aumentados de 30 amostras a 8

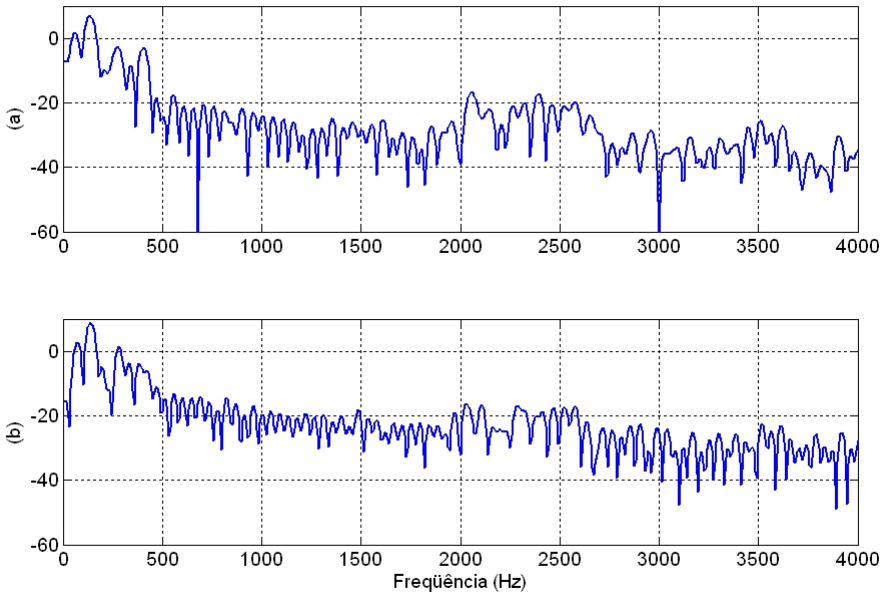


FIG.2.12: Comparação entre os espectros de trechos de sinais de voz em duas configurações diferentes — (a) configuração de janela fixa de 20 ms; (b) três períodos de pitch consecutivos.

KHz de amostragem — o equivalente a 3,75 ms. Nesta representação, é fácil perceber que os espectros em (a) e (c) são bem similares em praticamente toda a faixa de freqüências, ao passo que os pares de espectros (a)-(b) e (c)-(d) são bem similares apenas ao longo das freqüências mais baixas.

Um outro exemplo simples, ilustrado pela FIG. 2.12, compara os espectros de uma situação *pitch*-síncrona e em uma situação de janela fixa de 20 ms. O gráfico em (a) mostra o espectro do caso de janela fixa (amostras de 241 a 400) e, em (b), o espectro para uma janela de 3 períodos de *pitch* (amostras de 239 a 419). Pode-se perceber a maior riqueza de detalhes do espectro em uma situação de sincronismo (em máximos e mínimos locais) comparado com a abordagem por janela fixa.

2.7 CONCLUSÃO

Este capítulo serve de base para todo o estudo das atividades de perícia em Fonética Forense abordados neste trabalho, com respeito às características dos sinais de voz. As Seções 2.1 a 2.4 serviram de preâmbulo para a modelagem teórica e extração das características, mostrando a importância da discriminação dos locutores sob o ponto de vista de dimensões do trato vocal. A Seção 2.5 abordou, de forma simplificada, características de aspecto físico, tais como *pitch* e formantes, e características de aspecto perceptual,

tais como MFCC, SCF, SFM, tonalidade e SSC. A Seção 2.6 introduziu os conceitos de extração de características com sincronismo de *pitch*, mostrando possíveis problemas decorrentes da extração de forma assíncrona. Além disso, a discriminação de trechos sonoros comentada nesta mesma seção é importante para a delimitação dos trechos de interesse dos sinais de voz sob aspecto forense, como será visto no capítulo seguinte.

3 SISTEMAS DE PERÍCIA EM FONÉTICA FORENSE

A investigação pericial pressupõe uma captura de evidências colhidas da cena do crime. Quando essas evidências são coletadas, documentadas e registradas pela equipe de perícia, poderão consistir em provas irrefutáveis que incriminem ou inocentem um suspeito, conhecidas como **peças-motivo**. As peças-motivo são os objetos aos quais o perito fará referência quando realizar as comparações com os padrões coletados de fonte seguramente identificada. Esses padrões identificados são as **peças-padrão**. Estas provas podem ser impressões digitais, marcas de sangue, resíduos de pólvora ou compostos químicos, entre outros. No caso de perícia em fonética forense, essas provas serão gravações de voz oriundas do criminoso e padrões coletados de suspeitos, respectivamente as **peças-motivo** e **peças-padrão**, de forma a permitir ao perito descrever as peças, tentar reproduzir os eventos, buscar a verdade técnica e, por fim, materializar o crime, de sorte a apontar qual suspeito é o criminoso de fato; caso contrário, nenhum suspeito será incriminado (nada pode ser afirmado ou de fato nenhum suspeito é o criminoso). As peças-motivo podem ser oriundas de diversos meios - interceptações telefônicas, gravações de sinais por microfones escondidos, gravações de sinais em ambiente ruidoso, entre outros.

Este capítulo se encontra subdividido da seguinte forma — a Seção 3.1 introduz o conceito de Biometria e contextualiza a Fonética Forense com ênfase em Verificação de Locutor. A Seção 3.2 destaca os testes perceptuais e acústicos no ambiente pericial, comentando as peculiaridades e as diferenças entre eles; a Seção 3.3 aborda a seqüência de procedimentos que constitui o estado da arte da perícia fonética no Brasil; as Seções 3.4 e 3.5 comentam aspectos relativos às características *pitch* e formantes para fins periciais, abrangendo o conceito da distribuição LTF.

3.1 INTRODUÇÃO

Define-se **Biometria** (BESACIER, 2003) (MASON, 2005) (POH) como a ciência que busca o reconhecimento automático de indivíduos por meio de métodos estatísticos e processamento de sinais que mapeiam elementos fisiológicos dos seres humanos - por exemplo, íris, voz, face, movimento labial, batimentos cardíacos, impressões digitais ou palmares, geometria da mão, forma de deslocamento e DNA - em características numéricas. Posteriormente, ocorre o armazenamento destas características de cada indivíduo em

bases de dados numéricas, podendo ou não ser combinadas em um esforço de identificação de indivíduos de forma unívoca, atendendo aos seguintes requisitos:

- Todas as pessoas devem possuir um conjunto de características peculiar que permitem sua posterior identificação;
- As características não devem ser iguais em pessoas diferentes;
- As características não devem variar com o tempo (para isso, deve haver a coleta de padrão de amostras contemporâneas) nem com as circunstâncias emocionais e patológicas, entre outras. Detalhes sobre a coleta de padrão serão comentados na Seção 3.3;
- As características devem ser medidas quantitativamente (deve haver quantidade suficiente de amostras para compor padrões de modelos individuais);
- Devem ser contemplados graus de precisão e parâmetros de confiabilidade do método de identificação biométrico utilizado, bem como da inclusão e da exclusão de novos indivíduos ao sistema (adaptabilidade do sistema);
- O sistema de biometria deve ser robusto contra fraudes tais como mímica, supressão de textos e adulteração de conteúdo.

Dentre as aplicações beneficiadas pela Biometria, podem ser citadas a autenticação de acesso de usuários a sistemas privados (por exemplo, acesso a sistemas bancários ou a salas de acesso restrito) e a identificação de indivíduos para fins forenses.

Cabe ressaltar aqui que as aplicações em Biometria não fazem parte do cotidiano por questões relacionadas a custo e facilidade de implementação - equipamentos de identificação por imageamento de íris, por exemplo, são dispendiosos (MASON, 2005). Sistemas de autenticação por voz, pelo contrário, demandam menor custo e maior facilidade de implementação que os baseados em íris e impressões digitais, além de ser um método não-invasivo, não necessitando o contato físico com o suspeito (ao contrário de coleta de sangue ou fios de cabelo para identificação por DNA, por exemplo).

A Lei Federal nº 10.054, de 7 de dezembro de 2000, fornece amparo legal à identificação criminal de indivíduos presos em flagrante delito, indiciados em inquéritos policiais, infratores penais em menor gravidade e indivíduos contra os quais tenha sido expedido mandado de prisão judicial, desde que não sejam identificados civilmente. O suspeito é obrigatoriamente identificado sob ambas as formas, civil e criminalmente, em caso de:

- Acusação ou indiciamento por homicídio doloso, crimes contra o patrimônio praticados mediante violência ou grave ameaça, receptação qualificada, crimes contra a liberdade sexual e falsificação de documento público;
- Falsificação ou adulteração do documento de identidade;
- Impossibilidade de completa identificação dos caracteres essenciais por mau estado, extravio e demora na expedição do documento de identificação;
- Existência de múltiplos nomes ou diferentes qualificações nos registros policiais;
- Recusa, por parte do indiciado, em apresentar o documento de identificação.

A identificação do indiciado ou acusado, para fins forenses, deve permanecer em arquivo policial específico para compor os autos do processo. Além disso, não existe uma proibição do emprego de outras técnicas de identificação além da documental (por certidão de identidade) e a papiloscópica (impressões digitais), abrindo margem à coleta de outras formas de padrões, respeitando o princípio básico de Direito de que um cidadão não pode produzir prova contra si mesmo. (Constituição Federal, Art. 5º, inciso LXIII). A coleta de padrões de voz se encontra embutida nesse aspecto.

3.1.1 A VOZ NO CONTEXTO DA BIOMETRIA

Diante do exposto nesta seção, percebe-se o enquadramento perfeito da voz no contexto biométrico, devido aos seguintes aspectos:

- Individualização, pois duas pessoas não possuem a mesma voz. As ondas sonoras são sempre diferentes de falante para falante, e mesmo padrões de voz de um mesmo locutor podem não ser repetidos identicamente. Contudo, podem ser extraídas características que permitam ao perito inferir a dissimilaridade entre locutores diferentes;
- A voz apresenta características fisiológicas, comportamentais e culturais;
- Não é necessário ver o falante para inferir o grau de educação, idade, sexo.

Enquadrada, pois, no contexto biométrico, a voz constitui um ramo à parte da área forense, denominado de Fonética Forense.

3.1.2 A FONÉTICA FORENSE

Entende-se por Fonética: (SILVA, 2005)

> "Fonética é a ciência que apresenta os métodos para a descrição, classificação e transcrição dos sons da fala, principalmente aqueles sons utilizados na linguagem humana."

O trabalho da Fonética Forense, subdividido em sub-áreas de atuação, requer uma configuração apropriada de equipamentos (ROMITO, 2004; BRAID, 2003; MORISSON, 2003; TONACO, 2003), e exige um protocolo de coleta de padrão. Estas questões serão discutidas nas seções seguintes.

3.1.3 TIPOS DE PERÍCIA EM FONÉTICA FORENSE

A perícia em Fonética Forense abrange diversos braços de atuação, tais como:

- **Verificação de Locutor**: exames periciais que buscam determinar se as falas armazenadas numa mídia provêm ou não de um determinado indivíduo. Ocorre de um para um; exemplo: se uma medida de distância que representa a dissimilaridade entre o locutor da peça-motivo e o locutor da peça-padrão se situar abaixo de um valor (conhecido como *limiar*), ou um *score* que representa a similaridade entre os referidos locutores se situar acima de um valor limiar, é inferido que a peça-motivo e a peça-padrão são provenientes do mesmo locutor.

- **Verificação de edição**: exames objetivando verificar se os registros de áudio sofreram algum tipo de edição ou supressão;

- **Transcrição fonográfica ou degravação**: mudança, da forma oral para a forma escrita, do registro de áudio, tendo em vista que o sistema penal brasileiro não é oral e necessita nos autos, de maneira escrita, da fração do material de maior relevância ao fato julgado.

- **Tratamento de sinal de áudio degradado**: muitas vezes o som da mídia está degradado por ruído não-correlacionado (por exemplo: ar-condicionado, som de viatura, ruído ambiente) e precisa ser tratado numa tentativa de aumentar a relação sinal-ruído, melhorando com isto a inteligibilidade do sinal.

A Verificação de Locutor é, portanto, um dos diversos ramos de atuação da Fonética Forense.

3.1.4 A VERIFICAÇÃO DE LOCUTOR NO CONTEXTO DE PERÍCIA

No ramo da atividade pericial (BRAID, 2003; MORISSON, 2003), a verificação de locutor visa, por meio de locuções armazenadas em mídias de gravação, atestar se determinada locução é de um locutor específico ou não, através da comparação entre duas falas distintas. Cabe ao perito, através de análises técnico-comparativas, julgar a qualidade da mídia contendo a voz do locutor, colher as características da mídia e do canal de comunicações (quando for o caso), comparar padrões baseados em características técnicas extraídas dos sinais de voz, compor exames técnicos, utilizar equipamento adequado e, enfim, fazer tudo ao seu alcance para chegar a uma conclusão que forneça subsídios para incriminar ou inocentar um suspeito. Convém ressaltar que é eticamente preferível inocentar um possível culpado a incriminar um inocente. A voz, nesse caso, funciona como prova importante em casos de crime de suborno, extorsão, chantagem, coação, seqüestro, ou seja, casos em que o uso imperioso da voz pretere outras provas palpáveis. Também se aplica a casos em que não há a confissão direta do crime por parte do suspeito - nesse caso, aborda-se o suspeito por meio de entrevistas em ambiente descontraído com o objetivo de permitir a coleta de padrão de forma mais natural possível; caso contrário, efeitos psicológicos, tais como o nervosismo, podem promover alteração das características naturais da fala do locutor cotejado. A FIG. 3.1 mostra um fluxo simplificado das atividades da perícia em Fonética Forense.

A perícia fonética prevê uma seqüência metodológica de passos dentro de um elevado rigor técnico e reportada em laudo específico (MORISSON, 2003):

a) Uma introdução, contendo as condições de captura de voz (judicial - com a presença do locutor - ou investigativa - sem a presença de locutor), as características do canal de comunicações de proveniência da voz (telefone fixo, telefone celular, rádio, voz sobre IP etc), o tipo de compressão do canal de voz a ser efetuada, a presença ou não de ruído de canal, o grau de integridade da mídia (se houve ou não montagem). Este item abrange as atividades de coleta de padrão listadas na Seção 3.3. Além disso, deve haver uma descrição detalhada dos sinais de voz adquiridos pelo perito (cena do crime) oriundos de algum tipo de canal de comunicações aqui citado, e dos sinais de voz colhidos pela polícia no momento da entrevista dos suspeitos;

b) Uma listagem dos exames técnicos, dos equipamentos e das ferramentas matemáticas empregadas na verificação de locutor, plenamente justificados;

c) Uma listagem de procedimentos preliminares efetuados (pré-processamento do

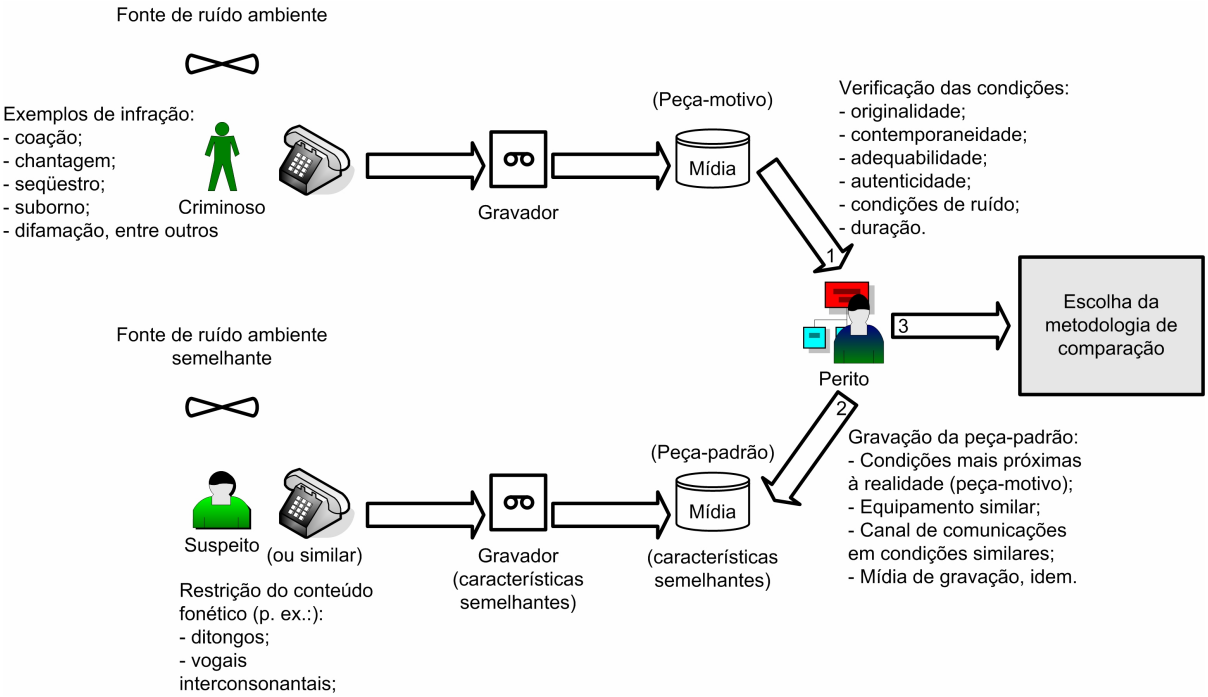


FIG.3.1: 1. O perito recebe o conteúdo das peças-motivo para análise; 2. O perito efetua a coleta de padrão; 3. Ocorre a comparação das peças analisadas.

sinal), tais como: redução de taxa de amostragem (também conhecido como *down-sampling*), filtragem, equalização, compensação de ruído;

d) Uma composição do resultado, com base na fundamentação teórica dos exames técnicos empregados, justificando o emprego (ou não) de determinadas características do sinal de voz em função dos aspectos citados no item “a”;

e) A conclusão, que consiste no parecer final do perito (locutor identificado, locutor não-identificado ou indefinição quanto à identificação) - explicitando os motivos que levaram à respectiva conclusão - e quanto às margens de aceitação dos resultados.

3.2 TESTES PERCEPTUAIS E ACÚSTICOS

A metodologia de perícia em fonética forense abrange duas grandes categorias de testes — os testes perceptuais e os testes acústicos.

Os **testes perceptuais** englobam medições subjetivas dependentes — como o próprio nome já sugere — da percepção, pelo ouvido humano, de detalhes predominantemente lingüísticos, exigindo o acompanhamento dos testes feito por pessoal credenciado e experiente, em um esforço multidisciplinar — por exemplo, lingüistas e fonoaudiólogos — de forma a analisar as gravações de voz com maior precisão perceptual. Por exemplo, em

uma situação na qual se procura comparar vozes reais de uma pessoa que efetuou ameaças por telefone com vozes de suspeitos, estes profissionais buscarão similaridades em diversas características, subdivididas abaixo em três níveis (DUNN), atribuindo-lhes um grau de similaridade que apoiará a decisão futura de um juiz. Esses níveis se encontram resumidos na FIG. 3.2.

- **Alto nível**: Neste nível estão englobadas características cuja dificuldade de extração a partir de gravações de sinais de voz é mais complexa. Dentre essas características, podem ser citadas a semântica, a dicção, a fonética (seqüência de fones para caracterizar a pronúncia e os padrões de fala dos supostos locutores) e as idiossincrasias, definidas como seqüências de palavras que revelam detalhes particulares dos locutores, dizendo respeito a suas origens (aspectos dialetais), ao seu estado sócio-econômico e nível educacional (aspectos socioletais) e a variações lingüísticas apresentadas por um mesmo indivíduo (aspectos idioletais).

- **Médio nível**: Este nível abrange características dos sinais de voz de dificuldade média de extração, menor do que a das características supracitadas. Como exemplos, podem ser destacadas a prosódia (evolução de padrão de *pitch* e de energia dos locutores ao longo de uma fala, denotando padrões de frases interrogativos, exclamativos, negativos entre outros), ritmo de fala (fala pausada, rápida ou com ritmo inconstante), entonação (ênfase de interrogações e exclamações) e modulação. São características que deixam transparecer em maior grau a personalidade do locutor.

- **Baixo nível**: Neste nível se encontram as características de mais fácil extração por métodos físicos e matemáticos. São as características fortemente dependentes dos aspectos acústicos da fala e da configuração do trato vocal dos locutores. Podem ser citadas a *pitch*, os formantes, o VOT[7], bem como indicadores de qualidade de fala tais como a nasalização, a aspereza, a incidência de ruído sobre trechos sonoros, a rouquidão e o excesso de aspiração, entre outras.

Os níveis supracitados ordenam as características dos sinais da fala, sob o ponto de vista pericial, de um grau de maior dificuldade para menor dificuldade de extração. Também pode ser percebido que as características se encontram graduadas do aspecto perceptual para o aspecto acústico, ou seja, quanto mais perceptuais - no sentido de

[7]Intervalo de tempo existente entre a soltura de uma oclusão (por exemplo, final da emissão de uma consoante [p]) e o início do vozeamento (ROSE, 2002).

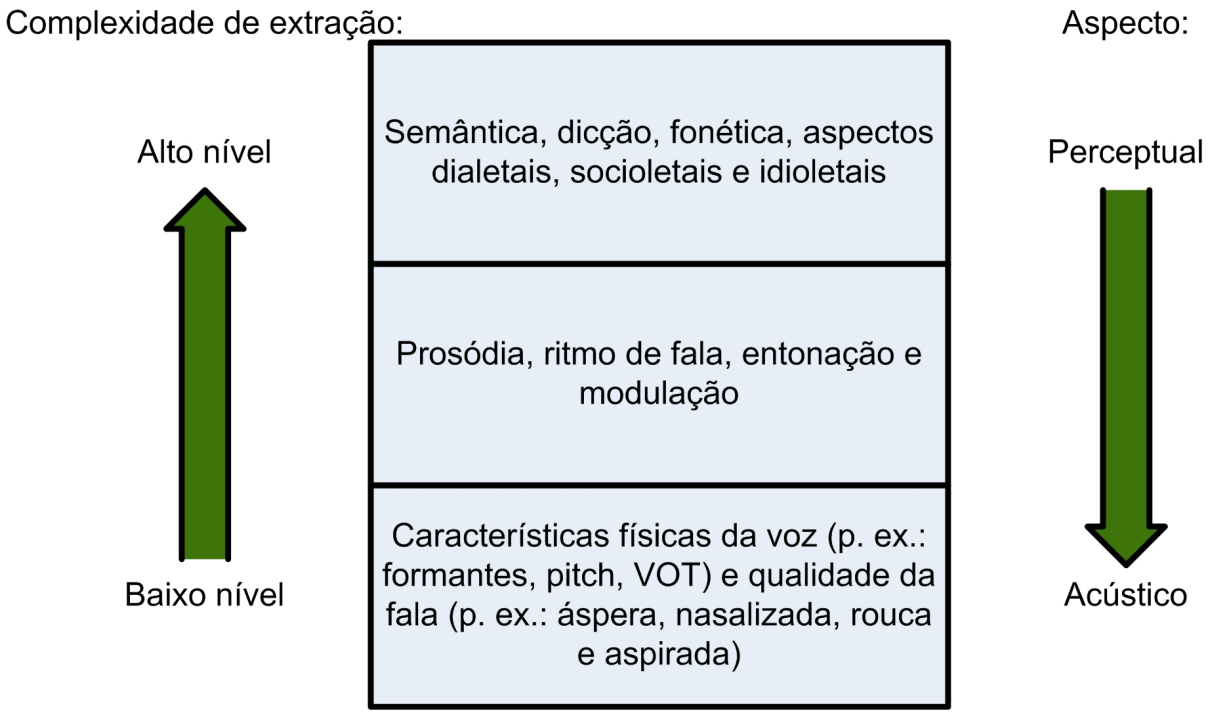


FIG.3.2: Níveis de complexidade de extração das características dos sinais de voz.

abranger uma análise cognitiva, e não matemática - são as características, maior será a dificuldade de extraí-las por métodos automáticos.

Os **testes acústicos** abrangem medidas matemáticas, físicas e estatísticas, correspondentes ao baixo nível de extração supracitado. Ao contrário dos testes perceptuais, qualitativos, a análise acústica determinará quantitativamente o grau de similaridade entre locutores. É importante, em âmbito pericial, ambas as abordagens — perceptual e acústica — de testes se complementarem, porque diversos detalhes lingüísticos podem não ser captados por medidas matemáticas, e a análise acústica pode descobrir detalhes imperceptíveis ao ouvido humano, tais como diferenças de formantes.

Devido a essa divisão ampla em testes perceptuais, dependentes da audição humana, e acústicos, dependentes das ciências exatas, o Reconhecimento Automático de Locutor para fins periciais se torna um desafio crescente. Em vários aspectos lingüísticos e qualitativos, é imprescindível a decisão do perito, ou seja, deve haver uma interatividade entre os procedimentos perceptuais e acústicos, dada a dificuldade de se estimar matematicamente as características ditas de alto nível. Pode-se afirmar, portanto, que os sistemas de reconhecimento de locutor para fins periciais são sistemas semi-automáticos, ou seja, **de apoio à decisão**. A título de exemplo, diferenças perceptuais refletem sempre diferenças acústicas; porém, falas muito parecidas perceptualmente não necessariamente refletem semelhanças acústicas. Da mesma forma, análises matemáticas aprofundadas

podem deixar passar despercebidos detalhes sociolingüísticos, demonstrando como uma área complementa a outra.

3.3 O ESTADO DA ARTE DA ATIVIDADE DE PERÍCIA NO BRASIL

O primeiro passo da metodologia de perícia consiste na verificação da legalidade, da procedência e da integridade do objeto a ser periciado, devendo ser registrado todo o material recebido para exames, inclusive por meio de fotografias, nos controles do protocolo geral da instituição executora da perícia, visando a garantir a custódia da peça-motivo (no caso da atividade de perícia em fonética forense, a peça-motivo é equivalente ao corpo de delito). Nesta etapa, deve ser descrita e caracterizada a mídia de suporte do áudio coletado da cena do crime. A mídia pode ser analógica ou digital. Dentre as mídias analógicas mais conhecidas, destacam-se as fitas cassete, as fitas microcassete e as fitas VHS. Dentre as digitais, pode-se citar os CDs, DVDs, HDs, cartões de memória, *pen-drives*, correio de voz e telefones celulares. O material de perícia sempre deverá estar em poder do perito responsável pela atividade pericial, guardado em local seguro, com acesso restrito e controlado.

3.3.1 VERIFICAÇÃO DAS CONDIÇÕES DE GRAVAÇÃO DOS SINAIS DE VOZ

Após o recebimento do material, a metodologia de perícia se incumbe de regular as condições satisfatórias de gravação de voz atendendo a aspectos tais como a originalidade, o tipo de mídia de gravação, as condições da mídia (integridade física), a autenticidade, a contemporaneidade, a adequabilidade (TONACO, 2003), as condições do canal de comunicações, as condições do ambiente de gravação, restrições quanto a ruído, tempo de fala, entre outros (ROSE, 2002).

Após o recebimento do material, a equipe de perícia deverá verificar a originalidade do material recebido. Neste aspecto, um registro de voz do locutor em questão pode ser classificado como original, cópia de registro de áudio ou cópia idêntica ao registro original de áudio. O registro original é aquele que não provém de nenhum outro registro (por exemplo, uma mídia magnética contendo a interceptação telefônica de uma conversação no exato momento de seu transcurso). A cópia de registro de áudio possui as mesmas informações do registro original, mas sem ser a gravação extraída no momento da fala do locutor. A cópia idêntica ao registro original possui o mesmo conteúdo do locutor gravado sem conversões ou mudanças de formato de áudio, ao contrário da cópia de registro de áudio.

A mídia recebida deverá ser rigorosamente descrita pela equipe de perícia nos seguintes quesitos — tipo, marca, modelo, numeração de lote, estado de conservação, acondicionamento, capacidade de armazenamento, existência de anotações manuscritas sobre a mídia ou sobre seu invólucro (também detalhando cor da escrita e tipo de caneta utilizada – esferográfica ou hidrográfica), data de recebimento e data de gravação do material perquirido (dado importante quanto ao quesito contemporaneidade).

Cabe à equipe de perícia realizar o exame físico (de integridade) minucioso da mídia visando constatar se há partes danificadas (por exemplo, em caso de CD, se a mídia está arranhada; em caso de fita magnética, se o estojo de proteção está violado ou se a fita está rompida). Caso necessário, a equipe de perícia poderá efetuar a substituição ou o reparo do material recebido, registrando no laudo a linha de ação a ser tomada. Além disso, deve ser verificado se a mídia recebida possui dispositivo de bloqueio de gravações posteriores, registrando no laudo pericial se a mídia possui o bloqueio contra gravação, se já se encontrava bloqueada ou se foi bloqueada posteriormente. Devem ser inspecionadas, também, as condições acústicas (presença de ruído e distorção) e as condições perceptuais (isto é, se há a inteligibilidade da voz do locutor).

Sob o aspecto autenticidade, é importante que as gravações das peças-motivo consistam em cópias idênticas ao padrão de voz adquirido, e que sejam realizadas por pessoal plenamente capacitado (TONACO, 2003), de preferência o próprio perito. Não é necessária a originalidade das gravações, pois é possível efetuar cópias da peça-motivo, se esta for digital, ou efetuar a conversão da peça-motivo em digital e posteriores cópias, no caso de mídia analógica. As cópias digitais futuras garantem a conservação da qualidade e a perpetuação das amostras.

Para efeitos de contemporaneidade, é necessário que não haja uma latência de tempo muito grande entre a gravação da peça-motivo e da peça-padrão, de forma a não haver grandes modificações nas características extraídas dos locutores envolvidos no processo pericial. Estas modificações surgem por motivos diversos, como por exemplo a aculturação (contato com a fala de locutores de localidades geográficas distintas), a evolução do padrão cultural, algumas enfermidades, entre outros (TONACO, 2003).

Quanto à adequabilidade das gravações de sinais de voz, é necessária a máxima naturalidade e clareza do padrão de fala coletado dos locutores entrevistados na atividade pericial, possibilitando o entendimento de todos os níveis de detalhes lingüísticos (frases, palavras, expressões típicas) em nível de análise perceptual, inclusive detalhes lingüísticos equivalentes — para esta análise, deve ser feito um levantamento prévio das caracterísi-

cas particulares do locutor da peça-motivo (principalmente as de alto nível e médio nível listadas na Seção 3.2), buscando repeti-las ao máximo na coleta de padrão — que possam permitir a identificação do acusado. No ato da coleta de padrão, a má pronúncia por parte de locutores, condições de nervosismo, raiva, desânimo, situações patológicas (alergia, calos nas cordas vocais, congestão nasal), cansaço, sede ou fadiga podem provocar alterações de características de voz dos locutores (XAFOPOULOS). É importante que o tempo de gravação do padrão coletado de cada locutor cotejado seja extenso o bastante para garantir maior confiabilidade no parecer final (TONACO, 2003). Contudo, destaca-se que as coletas de padrão devem ser realizadas de maneiras e em datas distintas ao se entrevistar um mesmo locutor. Além disso, é necessário o controle cuidadoso do nível de entrada de ganho do reprodutor da mídia, do equipamento e do aplicativo utilizado na captura do áudio do locutor entrevistado, para não haver a saturação do sinal de áudio adquirido e o posterior prejuízo do exame pericial. A taxa de amostragem, por se tratar de voz, deverá ser sempre igual ou superior a 8 KHz.

A coleta de padrão deverá ser realizada, preferencialmente, no mesmo tipo de mídia, com a utilização do mesmo equipamento de gravação, em local que apresente a maior similaridade possível com o ruído ambiental presente na peça-motivo; gravações em cabines acústicas não são ideais, por se distanciarem da situação real da análise forense. O espectro do som das peças-motivo raramente será sem distorções (aditivas ou convolucionais), devido às condições atípicas de gravação para fins forenses. Em muitos casos, a pessoa que está inquirindo o acusado não deseja que o entrevistado perceba estar sendo gravado. O uso de gravadores escondidos gera ruído e distorção proveniente de diversas fontes — por exemplo, abafamento do áudio, objetos no bolso e ruído ambiente de salas fechadas sem tratamento acústico. Além disso, o locutor acusado estará em seu ambiente de convivência, onde sempre existirá ruído ambiente (as gravações serão contaminadas por ruídos de televisores, rádios, ventiladores, aparelhos de ar-condicionado, veículos, falas de pessoas à volta, entre outros). A existência destes fatores, além do fator humano (manuseio dos gravadores pelo operador), prevê o tratamento do sinal — filtrado ou compensado em termos de ruído — de forma a garantir que todos os tipos de análise possam ser realizados pela equipe de perícia. Se a qualidade da mídia for muito precária nestes aspectos, de nada adiantarão os equipamentos e as técnicas de análise mais sofisticadas: os resultados fatalmente não serão satisfatórios. O canal de comunicações deve possuir banda de freqüência de pelo menos 3 kHz, adequada para voz. Desta forma, o sistema de perícia deve ser robusto a interferência, ruído, distorção, descasamento de uma forma

geral (para casos de gravações diferentes realizadas com microfones ou por canais de comunicações distintos) e ao tratamento acústico ruim (proposital, para manter o ambiente o mais similar possível ao ambiente de gravação da peça-motivo) da sala de coleta de padrão.

Como o trabalho possui aspectos forenses, é desejável que, acompanhando as mídias com as gravações de voz, acompanhe um relatório por escrito contendo as condições de gravação; por exemplo: a posição do microfone, a velocidade de gravação, o equipamento de gravação utilizado, o tipo de mídia empregado, as condições do ambiente da gravação (em estúdio, em campo aberto, em sala fechada sem facilidades de estúdio, por telefone fixo, por telefone celular; a operadora de telefonia fixa ou celular) e quaisquer outros dados que permitam a reprodução das gravações a qualquer momento.

Quanto ao tempo mínimo de fala para o processo de perícia, é importante salientar que outros fatores são determinantes. Por exemplo, pode ser arbitrado, para um dado número de locutores, e para um mesmo conjunto de condições de gravação de voz, que o tempo mínimo será aquele cuja taxa de falsa aceitação seja a mínima possível. Em termos práticos, quanto maior o tempo de gravação de voz, menor a probabilidade de se acusar injustamente um locutor inocente (ROSE, 2002).

Para efeitos deste trabalho, as mídias de gravação serão consideradas íntegras, contemporâneas, originais e autênticas. Também deverá ser levada em conta a mesma taxa de digitalização para todas as amostras de áudio consideradas.

3.3.2 ESCOLHA DO CONJUNTO DE PEÇAS-PADRÃO PARA ANÁLISE

Um segundo passo importante em uma metodologia de perícia é a restrição dos fones, sílabas ou trechos de fala para análise. Não necessariamente todas as falas de todos os locutores precisam ser analisadas. A análise acústica pode ser, por exemplo, restrita a (BRAID, 2003; ROMITO, 2004):

- Vogais - com ou sem acentuação; uma vogal em particular ou várias vogais; com efeito de arredondamento dos lábios ou não; vogais abertas, meio-abertas, médias, meio-fechadas ou fechadas;
- Ditongos, pela questão da variação dinâmica dos formantes;
- Vogais intercaladas com consoantes intervocálicas, importantes devido ao fenômeno da coarticulação;

- Fones sonoros, devido à variação dos formantes ao longo do tempo que lhes são destinados;

- Alguns fones surdos, tais como consoantes plosivas (por exemplo, a consoante [p]), devido à diferença de tempo entre a barra de explosão e o início do vozeamento (em outras palavras, o VOT).

Deve ficar bem claro que a escolha de um fone em particular requer diversos pormenores, tais como (ROMITO, 2004):

- A determinação de todos os fones a considerar;

- A determinação do estado da vogal. Por exemplo, a comparação de medições com vogais com e sem tonicidade, redução, centralização e coarticulação;

A restrição do conjunto de fones a considerar no confronto pericial da peça-padrão com a peça-motivo corrobora a noção de adequabilidade exposta na subseção anterior.

3.3.3 ESCOLHA DA METODOLOGIA DE COMPARAÇÃO

Após os procedimentos preliminares e a delimitação do conjunto de fones a considerar para cada locutor abrangido pelas peças-motivo e padrão, deverá ser escolhida a metodologia de comparação entre o padrão coletado dos suspeitos e a voz coletada na cena do crime.

Atualmente, a perícia em Fonética Forense emprega a técnica conhecida como *matching* (BROEDERS, 1999), que consiste em atribuir escalas de probabilidade ao resultado do exame pericial, baseado na comparação entre a peça-motivo e os padrões de voz coletados. São utilizadas escalas verbais de certeza, probabilidade e possibilidade, em vez de escalas numéricas percentuais, para expressar o grau de similaridade. O uso do *matching* gera uma falta de padrão, pois diversos peritos tendem a expressar as escalas de formas distintas e de forma subjetiva, sem seguir um padrão preconcebido. Podem ser escolhidas frases, palavras, unidades silábicas ou unidades sonoras semelhantes, desde que o contexto (entonação, tipo de emoção demonstrada, ritmo de fala, entre outros) seja semelhante.

Existem várias classes de *matching*, desde as mais extremistas, variando entre a identificação subdividida em negativa e positiva (BALDWIN, 1990), a identificação puramente comparativa (BROEDERS, 1999) e a identificação que suprime o quesito de possibilidade (BROEDERS, 1999). O IAI (*International Association for Identification*) (ROSE, 2002), por exemplo, convenciona um protocolo fixando, em ordem decrescente, sete níveis

de similaridade: identificação, provável identificação, possível identificação, resultado inconclusivo, possível eliminação, provável eliminação e eliminação. O grande problema é que não existe uma convenção regulando os percentuais de dissimilaridade ou de taxa de erro abrangidos por esta classificação de dissimilaridade (ROSE, 2002). Dentre tantos modelos de escala de comparação por *matching*, é preferível ao perito (BRAID, 2003) concluir categoricamente se o locutor cujo padrão foi coletado corresponde ao mesmo locutor da peça-motivo.

> "Os peritos concluem que (fulano de tal) é o interlocutor (n) da conversação intitulada (X)..."
>
> "Os peritos concluem que (fulano de tal) não é o interlocutor (n) da conversação intitulada (X)..."

Pode-se também (não sendo de rara ocorrência) mencionar que não houve meios de afirmar que o locutor tenha sido o mesmo locutor da peça-motivo.

O perito é soberano na emissão do laudo (BRAID, 2003); contudo, é necessária a argumentação do grau de conclusão do exame mencionando as técnicas de exame utilizadas, o instrumental operado — desde que seja adequado ao método empregado — no exame pericial, os elementos técnicos que fundamentam cada técnica, quais passos da atividade pericial foram desempenhados, muito embora a conclusão seja baseada em aspectos subjetivos.

Como a análise subjetiva depende muito da interpretação pessoal de ambas as partes (perito e juiz), é importante a organização do laudo contemplar uma metodologia visando incluir no trabalho do perito ferramentas automáticas que tornem a interpretação menos subjetiva e mais estatística (numérica), onde poderão ser incluídas as técnicas de Verificação Automática de Locutor, a serem abordadas no Capítulo 5.

3.3.4 ESCOLHA DOS PARÂMETROS EXTRAÍDOS DOS SINAIS DE VOZ

A última etapa da perícia fonética envolve a análise acústica, a extração das características, a análise estatística e a interpretação dos resultados. Podem ser utilizados aplicativos (*softwares*) livres, tais como o Praat[8], o Wavesurfer[9] ou o SFS[10]. Deve haver um padrão na escolha, que consiste em:

- Escolha de uma lista de características acústicas a analisar;

[8] http://www.fon.hum.uva.nl/praat/

[9] www.speech.kth.se/wavesurfer/download.html

[10] http://ftp.phon.ucl.ac.uk/resource/sfs/download.htm

- Uso do mesmo conjunto de aplicativos para todas as extrações efetuadas.

Cabe ressaltar aqui o comentário realizado na Seção 3.3.1 — as amostras das peças-padrão deverão ser analisadas nas mesmas condições de aquisição da peça-motivo, tais como freqüência de amostragem, largura de banda de freqüência, nível de energia sonora e SNR (relação sinal-ruído).

Os espectrogramas deverão ser analisados em banda estreita (boa definição em freqüência e má definição no tempo) e em banda larga (boa definição temporal e má definição em freqüência). Por exemplo, apenas em banda larga é possível analisar as barras de explosão, definidas como barras verticais ocupando quase todo o espectro do sinal de voz, causadas pela emissão de consoantes plosivas[11]; em banda estreita, pode-se analisar com maior definição a estrutura dos harmônicos. O exemplo da FIG. 3.3 evidencia tal fato – a barra de explosão (indicada pela elipse vermelha na FIG. 3.3) aparece com melhor definição no espectrograma superior, em banda larga, ao passo que não aparece em banda estreita. No espectrograma em banda estreita, percebe-se os harmônicos da freqüência fundamental como faixas contínuas ao longo do tempo. Além disso, somente os espectrogramas em banda larga evidenciam as características de articulação e coarticulação da fala.

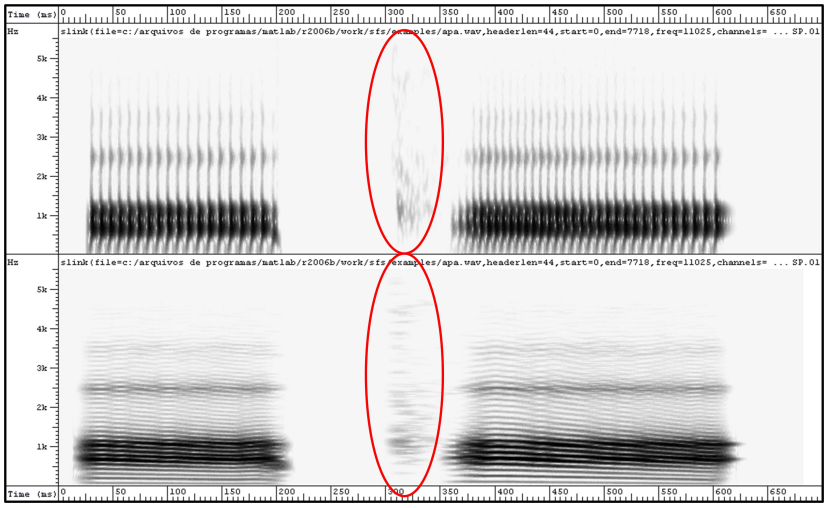


FIG.3.3: Espectrogramas em banda larga (superior) e banda estreita (inferior) do conjunto de fones ['apə]. As elipses vermelhas indicam, em seu interior, a localização das barras de explosão, visíveis no espectrograma em banda larga.

Através dos espectrogramas, é possível analisar o movimento dos formantes e suas transições, a estruturação dos harmônicos em uma unidade ou em um conjunto de

[11]Por exemplo, [p], [t], [k], [b], [d] e [g] (SILVA, 2005)

unidades sonoras e características articulatórias e coarticulatórias, como ficou sugerido no exemplo acima. Os espectrogramas consistem, então, em suporte importante para determinar as similaridades dos formantes dos núcleos vocálicos. Devem ser também analisados os perfis do primeiro e do segundo formantes em conjunto, transcrevendo-os do espectrograma para um registro escrito (laudo) se necessário.

3.3.5 FORMA ATUAL DE APRESENTAÇÃO DO RESULTADO DA PERÍCIA

A subseção anterior mostrou exemplos de características facilmente obtidas e transcritas a partir de um espectrograma — barras de explosão e formantes. Se a metodologia empregada for baseada no *matching*, o resultado deve ser expresso em um percentual associado a um grau de similaridade entre locutores. Um exemplo dado em (ROSE, 2002) fixa como critério de identificação 90% de similaridade perceptual e acústica de palavras possuindo os três primeiros formantes muito próximos entre si. A grande questão é: o que cada profissional que interpreta o processo (perito ou juiz) entende como "similar" ou "dissimilar"?

Para tornar o critério de similaridade mais objetivo e padronizado, sem suscitar este questionamento, sugere-se a adoção de resultados expressos em função das taxas de falsa aceitação, falsa rejeição e EER[12] do sistema como um todo (ROSE, 2002). Podem ser adotados outros critérios, como por exemplo a DCF (função-custo de detecção, de *Detection Cost Function*) e o MME[13] (erro médio mínimo, de *Minimum Mean Error*). A FIG. 3.4 exemplifica o efeito da escolha de um **limiar de decisão**, simbolizado pela linha verde vertical, sobre o erro de verificação. Dadas duas distribuições de probabilidade dos *scores* (verossimilhanças médias) dos locutores verdadeiros e dos locutores falsos, mostra-se que a escolha adequada do limiar de decisão influencia as taxas de falsa aceitação (igual à área representada em azul) e de falsa rejeição (igual à área representada em vermelho). No caso teórico exemplificado, a taxa de falsa aceitação aumenta e a taxa de falsa rejeição diminui em função do deslocamento do limiar exibido na FIG. 3.4(b) para o valor da FIG. 3.4(c). Além disso, poderão ser inseridas características no contexto pericial tais como as enunciadas no capítulo anterior — SFM, SCM, tonalidade, MFCC e SSC — permitindo um grau de liberdade amplo, muito embora ainda sejam importantes as características de maior apelo físico como a *pitch* e os formantes. É importante frisar que todas as etapas de todos os testes estatísticos realizados devem ser descritas em seus

[12]Situação na qual as taxas de falsa aceitação e de falsa rejeição são iguais.

[13]Situação de menor média aritmética das taxas de falsa aceitação e de falsa rejeição.

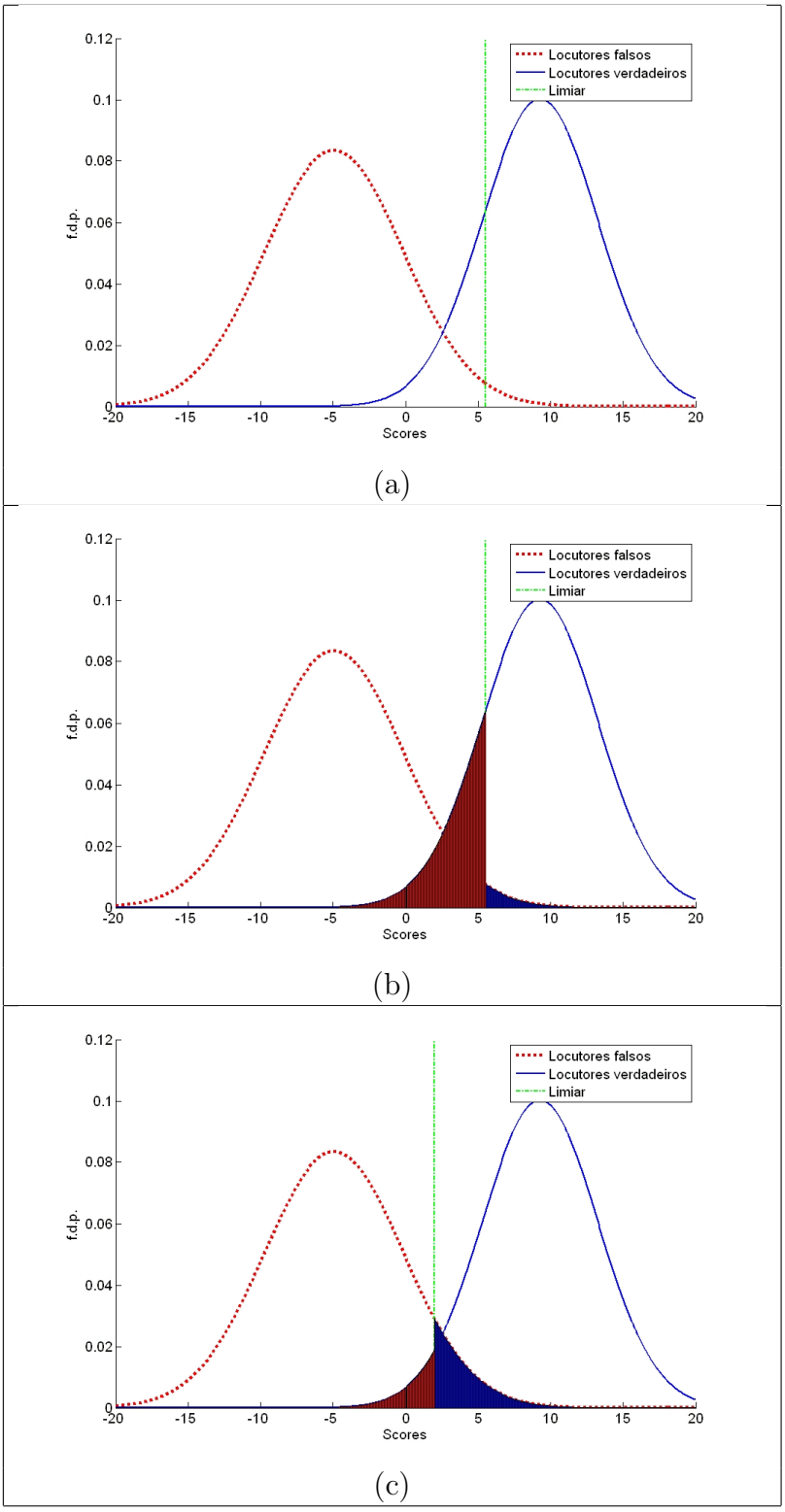


FIG.3.4: Taxas de falsa aceitação e rejeição: (a) *Scores* dos locutores verdadeiros e falsos (impostores); (b) taxa de falsa aceitação (dada pela área em azul) e taxa de falsa rejeição (dada pela área em vermelho) para o limiar igual a 5,5; (c) Novos valores de falsa aceitação e falsa rejeição para o limiar em 2,0. Percebe-se a redução da taxa de falsa rejeição e o aumento da taxa de falsa aceitação com o novo limiar escolhido.

mínimos detalhes. Da mesma forma, os resultados devem ser explicitados em todos os seus pormenores.

Por fim, todos os resultados obtidos devem ser encaminhados ao juiz, permitindo que ele chegue ao veredito de acusação ou inocência do suspeito baseado nos resultados, devidamente fundamentados pela interpretação dos peritos. Isto demonstra que, de fato, o sistema de verificação de locutor para fins periciais, do ponto de vista do juiz (decisor), sempre será um sistema de apoio à decisão.

3.4 ANÁLISE DE *PITCH* PARA FINS PERICIAIS

O contorno de *pitch*, também denominado contorno melódico, é considerado como sendo uma das características dos sinais de voz mais importantes para análise em Fonética Forense (MORISSON, 2003), por ser extremamente dependente do locutor em diversos aspectos (ROSE, 2002):

- **Prosódia**: locutores tendem a apresentar um padrão de contorno melódico em função da acentuação, do ritmo de fala, da oscilação da *pitch* dentro de uma mesma fala, do ataque de *pitch* (variação da *pitch* dentro de uma mesma palavra, fazendo a freqüência fundamental aumentar ou decrescer ao longo de uma palavra;

- **Estado emocional:** a *pitch* de um locutor depende de seu estado emocional - por exemplo, excitado, irritado, depressivo, contente, eufórico. Sob o aspecto Forense, numa situação de crime em que o agressor em questão intimide a vítima por telefone, ele poderá exibir um contorno melódico tal que a acentuação das palavras faça a *pitch* assumir valores mais altos na peça-motivo; entretanto, na coleta do padrão de voz, o suspeito poderá se sentir desconfortável, acuado ou intimidado com a situação da coleta, exibindo naturalmente discrepâncias do contorno melódico em comparação com a peça-motivo. Outro caso é a diferença de horário, que poderá inferir um grau de cansaço menor ou maior ao locutor, fazendo surgir novas discrepâncias. Outra questão é a alteração intencional do tom de voz pelo locutor na coleta do padrão de voz, de forma a burlar o sistema de perícia. Nestes casos, o locutor não poderá ser descartado (ROSE, 2002), tendo em vista as desigualdades de situação da peça-motivo para a peça-padrão. Contudo, um perito com anos de prática usualmente consegue reconhecer uma maneira de o locutor entrevistado burlar a perícia (alterando intencionalmente sua voz) e vencê-lo pelo cansaço.

3.5 ANÁLISE DE FORMANTES PARA FINS PERICIAIS

Os formantes apresentam uma vantagem muito interessante das quais os sistemas de perícia fonética podem usufruir - a **robustez a ruído aditivo** (não são robustos a variações de canal). São uma grandeza interessante devido a sua própria associação com os máximos espectrais do sinal de voz. Estes máximos superam o nível espectral de ruído, mesmo quando a relação sinal-ruído média é próxima de zero ou negativa (WET, 2004).

3.5.1 EMPREGO DOS FORMANTES

Para fins de análise acústica dos dois primeiros formantes, pode ser escolhido um fone vocálico muito freqüente nas gravações de voz dos locutores (NOLAN, 2005). Esta tarefa pode ser realizada de duas maneiras distintas:

a) Através da seleção não-automática dos fones, diretamente das gravações em formato digital, por um perito;

b) Através do reconhecimento automático de fala, que identificará o fone por alguma técnica matemática ou estatística.

Por simplicidade, será abordada a primeira situação acima. Por exemplo, Nolan e Grigoras (NOLAN, 2005) utilizam um fone vocálico (i) e três ditongos distintos. Nesse exemplo, um homem é acusado de realizar ligações obscenas a uma funcionária de um banco em Londres, sendo arrolado em um processo disciplinar, além de ter passado por investigações policiais. O real autor das ligações (locutor desconhecido) é designado por **U** (**U**nknown, desconhecido). Um suspeito é designado por **K** (**K**nown, conhecido). Este experimento foi reproduzido com sinais de voz em Português, apresentando-se em seguida os resultados. São levantados os dois primeiros formantes dos locutores U e K e confeccionados o gráfico F_1 versus F_2 para os formantes centrais da vogal, e os gráficos de F_2 no início versus F_2 no final do fone para o caso dos ditongos.

Pode ser percebida na FIG. 3.5, para os formantes extraídos das vogais [i] das vogais "analfabet**i**smo", "just**i**ça" e "T**i**to", uma boa separação e consistência dos valores medidos discriminando os locutores U e K (as vogais foram extraídas na mesma configuração fonética, posto que os fones [i] estão situados entre duas consoantes surdas). A FIG. 3.6 também apresenta uma boa discriminação entre locutores distintos, cujos formantes foram extraídos dos ditongos /ei/ das expressões "L**ei**la tem um lindo jardim","S**ei** que atingiremos o objetivo" e "Desculpe se mago**ei** o velho". No caso desta figura, U $\neq$

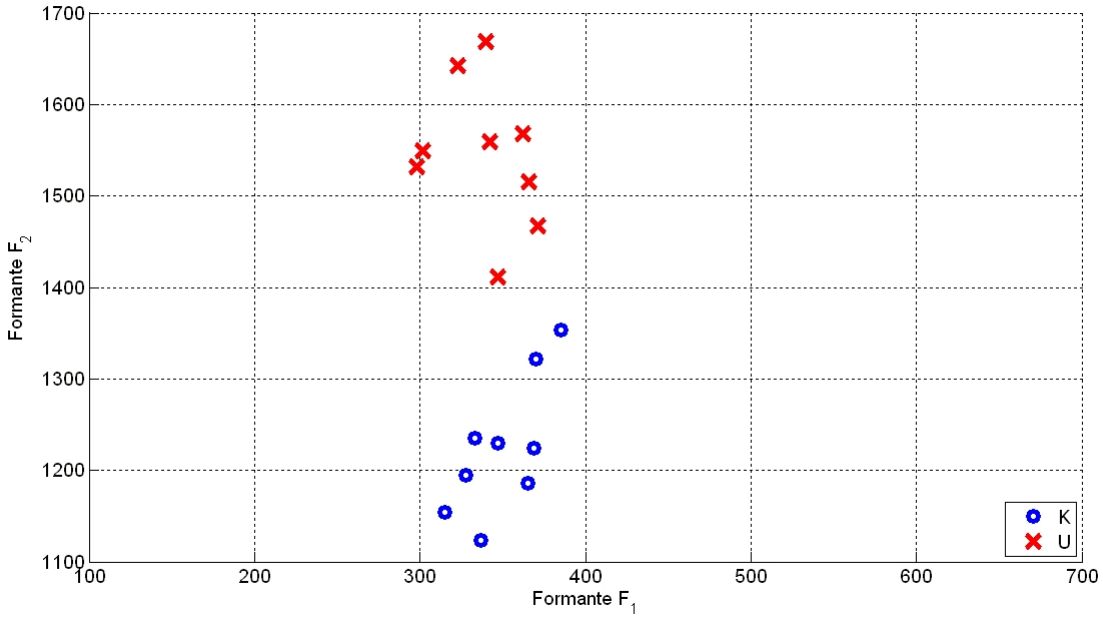


FIG.3.5: Gráfico de F_1 versus F_2 da vogal /i/ para os locutores U e K, com janelamento de 20 ms e sobreposição (*overlap*) de 15 ms. Foram extraídos três vetores de formantes $[F_1\ F_2]$: um do centro do fone, um 5ms antes do centro e outro 5ms após o centro. Para o caso ilustrado, U $\neq$ K.

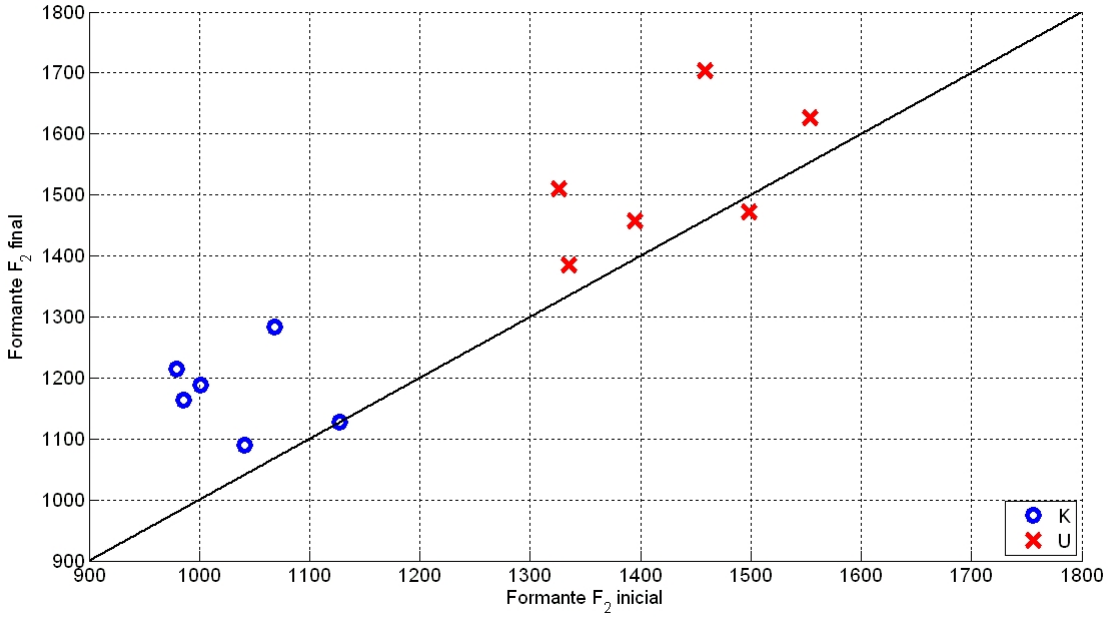


FIG.3.6: Gráfico de F_2 inicial versus F_2 final do ditongo /ei/, com janelamento de 20 ms e sobreposição (*overlap*) de 15 ms. Para este caso específico, U $\neq$ K. A linha diagonal preta indica o lugar geométrico de F_2 caso permanecesse constante ao decorrer do ditongo.

K. É importante ressaltar que o valor do primeiro formante foi evitado na análise dos ditongos por ser um dos formantes afetados pela banda inferior de freqüência do canal telefônico (KÜNZEL, 2001). Os dois valores de F_2 extraídos de cada ditongo são capturados 20 ms após o início e 20 ms antes do término do ditongo, respectivamente, para evitar efeitos coarticulatórios. A linha diagonal preta na FIG. 3.6 indica o lugar geométrico de F_2 caso permanecesse constante ao decorrer do ditongo, mostrando que há sempre a tendência do segundo formante final ser maior[14] que o inicial.

O resultado exibido reflete as diferenças anatômicas entre os locutores K e U. Tratos vocais diferentes estão associados a valores de formantes distintos. Além disso, os padrões articulatórios de locutores diferentes, via de regra, são distintos, o que motivou a boa discriminação nos ditongos na FIG. 3.6. Contudo, seguindo a lógica do trabalho de perícia, estes resultados podem ser conjugados ao exame perceptual, podendo ser, ou não, aceito por um juiz.

Apesar da boa discriminação aqui exibida, esses gráficos de formantes não são conclusivos para um resultado final. Embora o foco do resultado em (NOLAN, 2005) acima reproduzido seja focado em duas situações de *matching* (uma para ditongos e outra para vogais entre fones surdos), a técnica deveria ser testada para um grande número de locutores (*corpus*) para ser validada por meio de taxas de erros estatisticamente representativas. Uma outra técnica acústica, que pode ser implementada a título de contraprova na tentativa de aumentar a similaridade visual entre locutores, é conhecida como técnica LTF (de *Long-Term Formants*), objeto de estudo da subseção seguinte.

3.5.2 ANÁLISE DA DISTRIBUIÇÃO ESPECTRAL DOS FORMANTES

A distribuição LTF (de *Long-Term Formants*) consiste na contagem do número de vezes que cada freqüência é escolhida como estimativa do formante, através de um estimador de predição linear. Faz-se a ressalva que este estimador pode ser implementado livremente em linguagem C, programação em Matlab, ou pode ser utilizado via *softwares* livres como os citados neste capítulo.

Através da simulação da FIG. 3.8, percebe-se que não existe a possibilidade do suspeito K levantado no processo (em linhas cheias, cujas locuções foram indicadas por K_1 e K_2) ter sido o mesmo locutor identificado como U. Muito pelo contrário, a falta de aspectos em comum da distribuição LTF do suspeito em comparação ao real autor das

[14]Quanto maior o avanço da língua na emissão da vogal, maior será o valor do segundo formante (TITZE, 1994).

chamadas (mais nítida para os formantes de mais alta ordem — F_3 e F_4) descarta que o suspeito tenha de fato cometido o crime. A FIG. 3.7 explica sinteticamente a extração do LTF de diferentes sinais de voz.

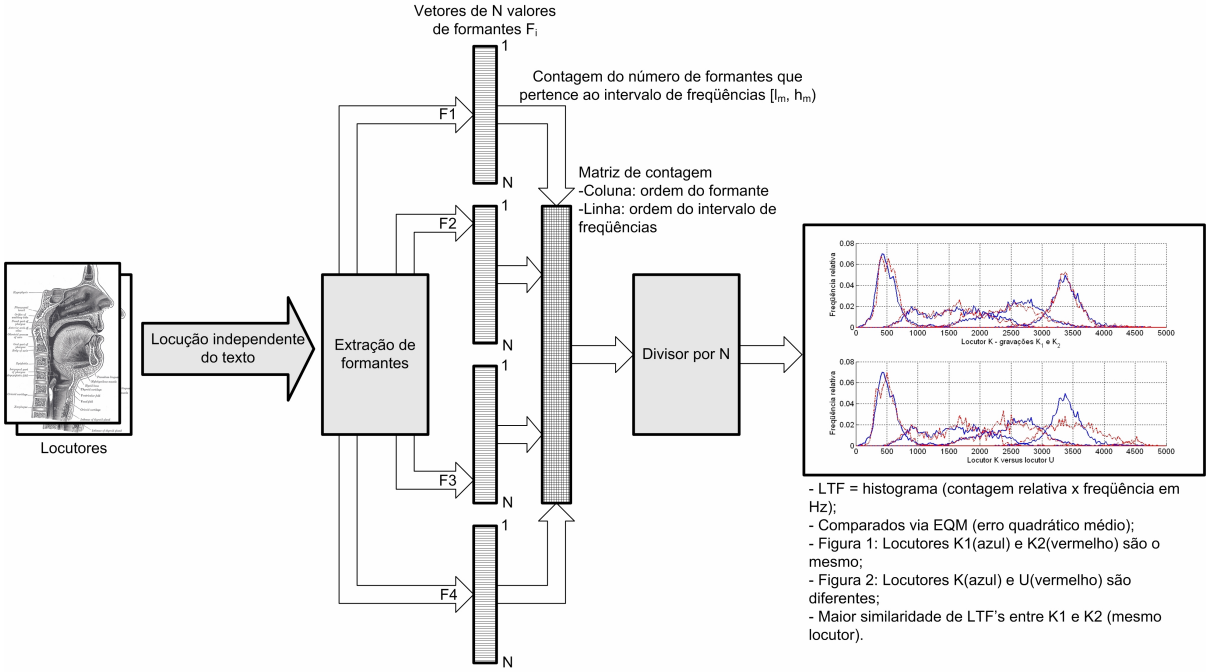


FIG.3.7: Diagrama explicativo da técnica LTF.

Além disso, a distribuição LTF é beneficiada por características intrínsecas a cada locutor, tais como:

- Tendência do locutor a arredondamento dos lábios (também chamado *lip rounding*);
- Diferenças anatômicas e articulatórias entre locutores;
- Diferenças de idioleto entre locutores, ou seja, cada locutor tende a utilizar naturalmente conjuntos de verbetes diferentes durante suas falas.

No atual estado da arte de processamento de sinais de voz, não se tem conhecimento de uma característica que possa ser extraída dos sinais de voz (matemática ou física) discriminando total e exclusivamente um locutor dos demais. Ou seja, não se pode inferir com 100% de certeza se a voz de uma gravação é ou não de um determinado locutor. Nesse sentido, as características estáticas de médias espectrais (por exemplo, a distribuição LTF) são insuficientes para a discriminação de locutores no contexto de identificação (no qual se aponta um locutor mais provável dentre N locutores — normalmente $N > 2$), devido à sobreposição de curvas percebida na FIG. 3.8.

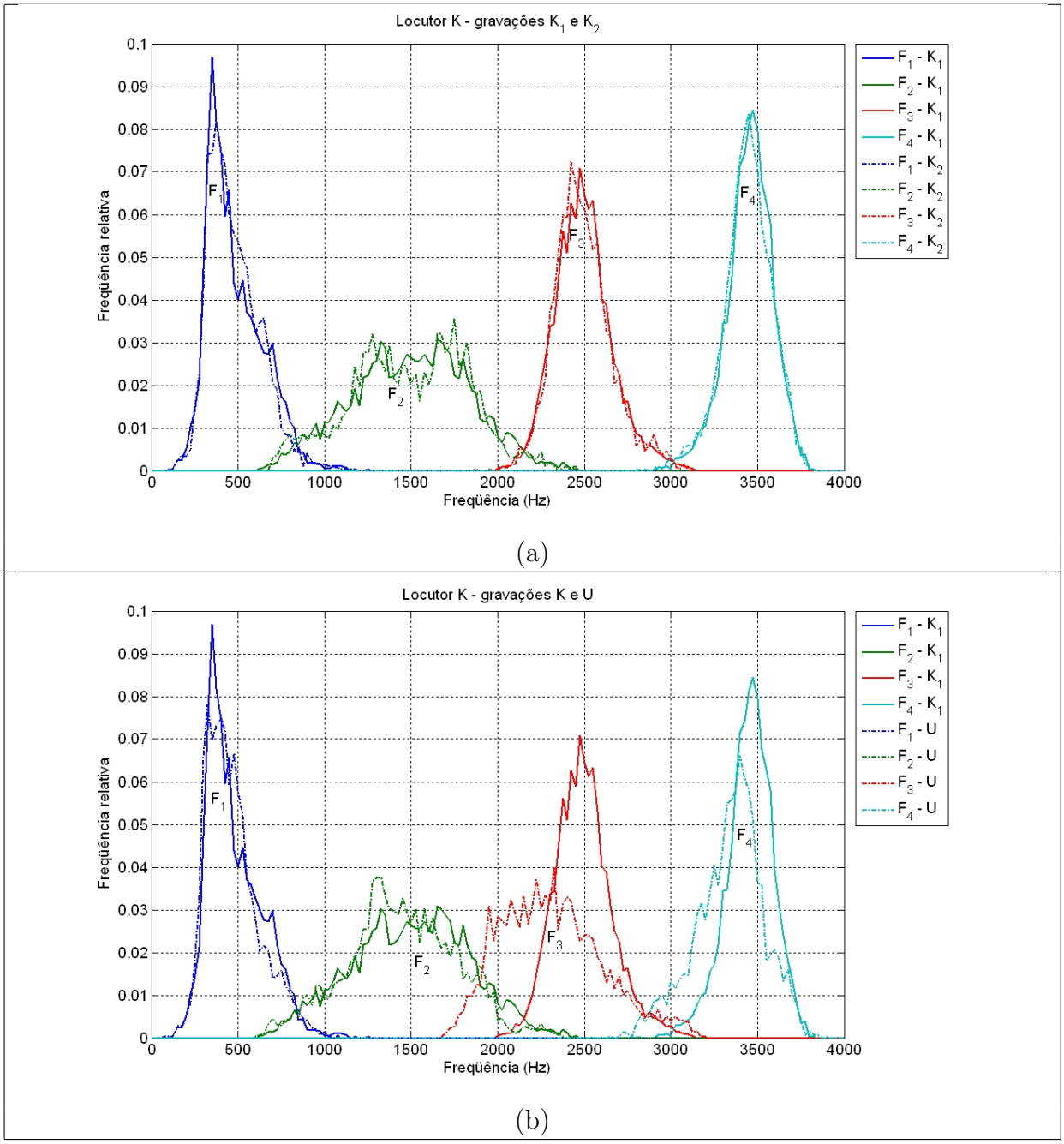


FIG.3.8: Comparação de distribuições LTF entre as gravações: (a) K1 e K2 (dois trechos diferentes de fala do mesmo locutor suspeito); (b) K e U (trechos da fala do suspeito e da fala do real autor das chamadas). Nota-se, em (b), que a dissimilaridade entre as curvas da distribuição LTF é bem maior do locutor K (em linhas cheias) para o locutor U (em linhas tracejadas).

3.6 CONCLUSÃO

Este capítulo apresentou diversos conceitos relativos a sistemas de perícia em fonética forense, descrevendo o trabalho do perito em termos de testes perceptuais e acústicos.

A Seção 3.2 definiu os testes perceptuais e acústicos, mostrando a necessidade de um esforço cooperativo entre profissionais de diversos ramos do conhecimento — foneticistas, lingüistas, estatísticos e fonoaudiólogos — visando complementar e interfacear ambos os tipos de testes, tornando os sistemas de perícia mais abrangentes. Houve também, na Seção 3.3, o detalhamento dos passos de uma possível metodologia de perícia em Fonética Forense, demonstrando ser necessário um novo protocolo de perícia para facilitar o trabalho do perito, restringir o universo de busca a um subconjunto de locutores, ou mesmo tornar os resultados dos sistemas de perícia menos subjetivos. Além disso, nas Seções 3.4 e 3.5 foram introduzidos os conceitos de *pitch* e formantes voltados para a atividade de perícia, mostrando ao final que a distribuição LTF tende a discriminar melhor, de forma visual, locutores diferentes, abrindo margem ao emprego das técnicas de verificação automática de locutor no ambiente forense, como será visto no Capítulo 5.

4 COMPUTAÇÃO EVOLUCIONÁRIA E SELEÇÃO DE CARACTERÍSTICAS APLICADAS À ESTIMAÇÃO DE PARÂMETROS DE MODELOS GMM

4.1 INTRODUÇÃO

Este capítulo trata basicamente da estimação dos parâmetros de modelos de misturas de gaussianas (GMM, *Gaussian Mixture Models*). A Seção 4.2 define e descreve o modelo GMM, comparando o desempenho de três variantes do algoritmo EM. A Seção 4.3 define e comenta a seleção de características para redução de dimensão de vetores de características extraídas das falas dos locutores em ambiente forense.

4.2 MODELOS DE MISTURAS DE GAUSSIANAS (GMM)

A expressão matemática da função de densidade de probabilidade (*f.d.p.*) $f(\cdot)$ de um modelo de misturas de gaussianas (GMM) em função de um vetor $\mathbf{x} \in \mathbb{R}^D$, em termos do número M de gaussianas, dos parâmetros peso π_j, vetor médio $\boldsymbol{\mu}_j$ e matriz de covariância $\boldsymbol{\Sigma}_j$, expressos conjuntamente pelo modelo $\lambda = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \mid j = 1, \ldots, M\}$, é dada pela EQ. 4.1. Esta equação representa, na verdade, uma combinação linear[15] de M f.d.p.'s gaussianas, dadas pela EQ. 4.2, com a restrição contida na EQ. 4.3 de que os pesos do modelo GMM devem ser compreendidos entre 0 e 1 e ter soma unitária.

$$f(\mathbf{x}; \lambda) = \sum_{j=1}^{M} \pi_j N(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{4.1}$$

$$N(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi^{D/2})\,|\boldsymbol{\Sigma}_j|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)} \tag{4.2}$$

$$\sum_{j=1}^{M} \pi_j = 1 \tag{4.3}$$

4.2.1 O ALGORITMO EM

O objetivo principal do algoritmo EM (*Expectation-Maximization*) é estimar os parâmetros de um modelo GMM de forma iterativa. Dado um conjunto de N vetores

[15] Convexa.

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, e assumindo esses vetores estatisticamente independentes, a verossimilhança logarítmica $\mathcal{L}$ associada ao modelo GMM da Seção 4.2 é definida pela EQ. 4.4

$$\mathcal{L}(\mathbf{X}; \lambda) = \sum_{i=1}^{N} \ln f(\mathbf{x}_i; \lambda) \tag{4.4}$$

4.2.2 FORMA CLÁSSICA DO ALGORITMO **EM** PARA MODELOS GMM

O algoritmo EM para o modelo GMM reestima os parâmetros do modelo GMM pelo critério da maximização da função-alvo verossimilhança pela estimação em Máxima Verossimilhança (Estimação ML, de *Maximum-Likelihood*) (REYNOLDS, 1995a; BISHOP, 2006). As condições para maximização da verossimilhança se encontram derivando a EQ. 4.4 em função dos parâmetros do GMM (BISHOP, 2006) e igualando a zero. Desta forma, chega-se às duas etapas principais do algoritmo EM, conhecidas como *passo-E* e *passo-M* (BISHOP, 2006).

- **Passo-E (expectância)**: O passo-E se desenrola com o cálculo das probabilidades *a posteriori* $\gamma_k(\mathbf{x}_n)$ de cada vetor, $n = 1, \ldots, N$, $k = 1, \ldots, M$ representadas na EQ. 4.5, com respeito a cada componente gaussiano de ordem k.

$$\gamma_k(\mathbf{x}_n) = \frac{\pi_k N(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{M} \pi_j N(\mathbf{x}_n; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{4.5}$$

- **Passo-M (maximização)**: O passo-M consiste na reestimação dos parâmetros do modelo GMM a partir do resultado da EQ. 4.5. Reestima-se, portanto, o valor dos pesos, dos vetores médios e das matrizes de covariância do modelo, de acordo com as equações EQ. 4.6, 4.7 e 4.8, respectivamente, para cada gaussiana $k = 1, \ldots, M$. Essas três equações são deduzidas, respectivamente, igualando a zero a derivada da EQ. 4.4 em função dos pesos, vetores médios e matrizes de covariância.

$$\begin{aligned} \pi_k^{novo} &= \frac{N_k}{N} \\ &= \frac{1}{N} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \end{aligned} \tag{4.6}$$

$$\boldsymbol{\mu}_k^{novo} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \mathbf{x}_n \tag{4.7}$$

$$\boldsymbol{\Sigma}_k^{novo} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) (\mathbf{x}_n - \boldsymbol{\mu}_k^{novo}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{novo})^T \tag{4.8}$$

O novo valor de verossimilhança (EQ. 4.9) é calculado em função dos parâmetros do modelo reestimados. São executadas várias iterações até a condição de convergência entre iterações consecutivas. Contudo, essa verossimilhança atinge um valor máximo local (REYNOLDS, 1995a; DASGUPTA).

$$\mathcal{L}(\mathbf{X};\lambda) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{M} \pi_k N\left(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\} \tag{4.9}$$

4.2.3 EMPREGO DE PROJEÇÕES ALEATÓRIAS

As projeções aleatórias (DASGUPTA) visam melhorar o ajuste dos parâmetros do modelo GMM, buscando se aproximar da estimação ML global através da redução de dimensionalidade do conjunto $\mathbf{X}$ formado por N vetores de características $\mathbf{x}_i \in \mathbb{R}^D, i = 1, \ldots, N$, definido na Seção 4.2.1. O conjunto $\mathbf{X}$ é projetado a um conjunto $\mathbf{Y}$ formado por vetores de características $\mathbf{y}_i \in \mathbb{R}^d, i = 1, \ldots, N$ tal que $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i$. O aprendizado dos parâmetros do modelo GMM envolve a técnica de redução de dimensionalidade através de uma matriz de redução $\mathbf{W}$ descrita na seqüência de passos abaixo:

a) Escolher os elementos de $\mathbf{W}$ aleatoriamente ($\mathbf{W} \in \mathbb{R}^{d\mathbf{x}D}$, $d < D$) a partir de uma função de densidade de probabilidade gaussiana de média zero e variância unitária. A dimensão d de projeção independe da dimensão original dos vetores de características (DASGUPTA);

b) Ortonormalizar as linhas da matriz $\mathbf{W}$ através do método de Gram-Schmidt (STRANG, 1988);

c) Calcular o conjunto de vetores $\mathbf{Y}$ tal que $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i$, $1 \leq i \leq N$;

d) Calcular, via algoritmo EM, os parâmetros do modelo GMM para o conjunto de vetores $\mathbf{Y}$;

e) Calcular o vetor de pertinência de $\mathbf{Y}$ em relação aos componentes gaussianos do modelo GMM acima, ou seja, construir um vetor $\wp = [k_1, k_2, \cdots, k_N]^T$ tal que $k_j = \arg\max_\theta \left(\mathcal{L}\left(\mathbf{y}_j; \lambda_\theta\right)\right)$; $j \in \mathbb{N}$; $\theta \in \mathbb{N}$; $1 \leq j \leq N$; $1 \leq \theta \leq M$;

f) Tomar o vetor de pertinência do modelo GMM de $\mathbf{Y}$ e repeti-lo na inicialização do algoritmo EM sobre $\mathbf{X}$, ou seja, estimar os parâmetros do modelo GMM empiricamente para cada subconjunto $\mathbf{X}^{(j)}$ dos dados originais, tal que $\mathbf{x}_i \in \mathbf{X}^{(j)} \Leftrightarrow \wp_i = j$;

g) Aplicar o algoritmo EM aos vetores de $\mathbf{X}$ para o cálculo dos parâmetros do modelo GMM;

h) Armazenar o valor final da verossimilhança.

Em comparação ao algoritmo EM, as vantagens do algoritmo EM com projeção aleatória se resumem ao menor tempo computacional, ao fato dos *clusters* dos dados projetados em geral serem menos excêntricos que os clusters dos dados em sua dimensão original (maior dimensão), e à geração de modelos GMM comparáveis ou melhores que os modelos GMM calculados pelo algoritmo EM (sem projeções aleatórias) (DASGUPTA). Entretanto, como mostra o histograma da FIG. 4.1, obtido a partir de 100 projeções aleatórias do conjunto de vetores de formantes extraído dos trechos não-sonoros da base IME2001[16], nem sempre é garantido (DASGUPTA) que a matriz $\mathbf{W}$ calculada pela projeção aleatória leve a uma redução de excentricidade (melhor modelo GMM), pois o procedimento do cálculo da matriz não é sistemático (é, na verdade, um sorteio). Como prova disso, na FIG. 4.1, percebe-se que o valor da verossimilhança alcançada pelo algoritmo EM sem projeções aleatórias, indicado pela barra vermelha e pelo losango no eixo horizontal, está compreendido entre o mínimo e o máximo valor de verossimilhança alcançados pelo algoritmo EM com projeções aleatórias, indicados pelas barras verticais na figura. Destes valores, houve um aumento de verossimilhança em 87% dos casos.

Caso se queira tentar obter valores de verossimilhança maiores que os computados com o algoritmo EM e maiores do que os calculados por apenas uma execução do algoritmo acima, pode-se efetuar uma repetição de projeções aleatórias conforme abaixo, doravante chamada de EM-RP (RP de *Random Projection*), tornando o mecanismo de projeção aleatória mais sistemático (FLORES):

a) Inicializar o número de execuções do algoritmo descrito acima. Uma vez escolhido o número de repetições, os passos de "b" a "d" seguintes podem ser observados na FIG. 4.1;

b) Repetir n_{RP} vezes, para cada execução, os itens de "a" a "g" da seqüência de passos enumerada no início desta seção;

c) Execução – computar o valor da verossimilhança (barras azuis da FIG. 4.1);

d) Armazenar, para o máximo valor de verossimilhança alcançado, os parâmetros da mistura de gaussianas e o valor da verossimilhança. O valor da verossimilhança

[16] A base IME2001 será descrita no Capítulo 5 deste trabalho

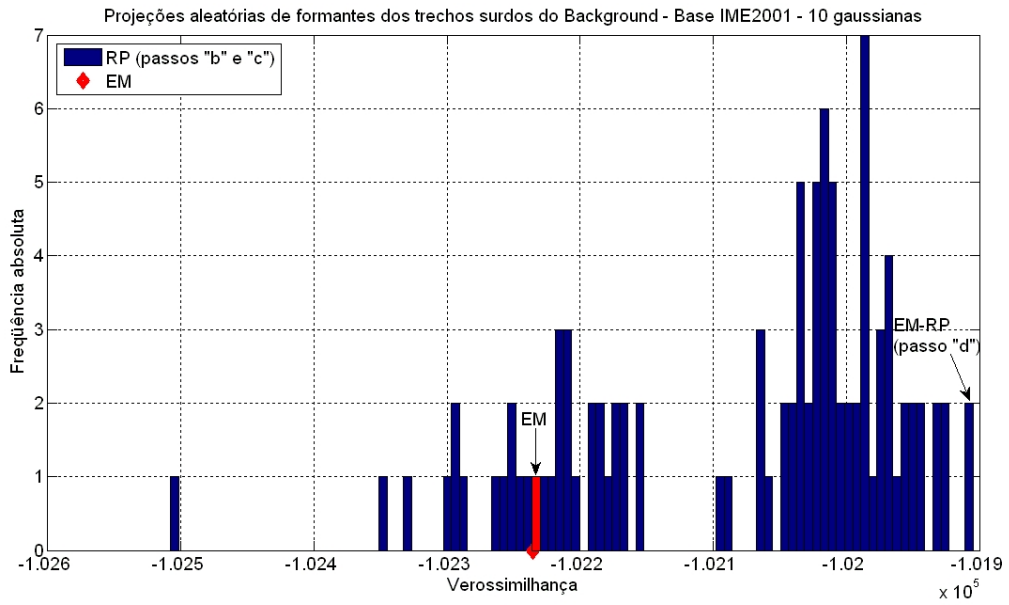


FIG.4.1: Projeções aleatórias. O histograma indica no eixo horizontal os valores de verossimilhança alcançados para cada projeção. O eixo vertical mostra o número de valores de verossimilhança correspondente a cada faixa de valores do eixo horizontal. A verossimilhança do modelo GMM estimada pelo algoritmo EM simples está indicada pela seta rotulada com "EM". Os passos de "b" a "d" do algoritmo EM-RP podem também ser facilmente observados no gráfico.

correspondente ao modelo GMM estimado pelo algoritmo EM-RP é representada pela última barra azul da esquerda para a direita na FIG. 4.1.

4.2.4 EMPREGO DE ALGORITMOS GENÉTICOS

Devido ao caráter não-sistemático do algoritmo de projeções aleatórias comentado na Seção 4.2.3, torna-se necessário definir uma estratégia de escolha da matriz $\mathbf{W}$, tendo em vista melhorar a eficácia da estimação ML. É proposta a escolha da matriz $\mathbf{W}$ através do algoritmo genético (LIN; FLORES) que, de forma diferente da busca aleatória comentada na Seção 4.2.3, busca otimizar a verossimilhança (função-custo) por meio de um esquema evolucionário análogo à teoria da evolução dos seres vivos. Esses algoritmos empregam operadores que imitam os processos de seleção natural de indivíduos, migração populacional, mutação e recombinação de genes. A terminologia dada às variáveis do processo também segue essa idéia – costuma-se dar à variável vetorial da função-custo o nome de cromossomo; cada elemento constituinte deste cromossomo recebe o nome de gene. No algoritmo proposto, o cromossomo $\mathbf{V}$ é um vetor composto pelos elementos (genes) da matriz $\mathbf{W}$.

O algoritmo EM-GA proposto (FLORES) se encontra descrito abaixo:

a) Inicializar os parâmetros peculiares do algoritmo genético: número de gerações (n_g), quantidade de elementos da população inicial ($n_{\mathbf{W}}$), tipo de seleção dos indivíduos, tipo de codificação do cromossomo, fator de recombinação ou *crossover* (k_c). Neste caso, a fração de elementos destinada a sofrer mutação é dada por $(1 - k_c)$. Devem ser escolhidos também o valor da variância da população inicial e o fator de redução da variância ao decorrer das gerações;

b) Escolher a dimensão d de projeção dos dados originais e computar a população inicial: $n_{\mathbf{W}}$ matrizes de redução de dimensionalidade (vide itens "a" e "b" do procedimento enumerado no início da Seção 4.2.3);

c) Computar o algoritmo genético regido pela população inicial e pelos parâmetros n_g, $n_{\mathbf{W}}$ e k_c (vide itens acima). A função-objetivo verossimilhança será otimizada em função da variável vetorial $\mathbf{V}$ e calculada de acordo com os itens de "b" a "g" do procedimento enumerado no início da Seção 4.2.3;

d) Armazenar os valores finais da estimação ML e da verossimilhança.

4.2.5 ESTUDO DE CASO COM OS ALGORITMOS EM-RP, EM-GA E A MEDIDA BIC

Para comprovar a eficácia da técnica de algoritmos genéticos na melhoria da estimação ML (FLORES), foram utilizadas bases de dados reais, extraídas do repositório público *UCI Machine Learning Repository*[17], citadas abaixo. As matrizes $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N]$ são, por construção, formadas pelos N vetores de características de cada classe das bases *SatImage* (informações de tipo de terreno referentes a pixels de imagens de satélite), *Pima-indians-diabetes* (informações relativas à não-incidência ou incidência de diabetes em uma localidade) e *Statlog-Segment* (informações de classificação de tipo de imagem) chamadas, respectivamente, de LANDSAT, PIMA e SEGMENT. A TAB. 4.1 fornece informações sobre as bases de dados.

Foram comparadas em (FLORES) três técnicas diferentes – o algoritmo EM aplicado diretamente aos dados originais, o algoritmo EM-RP e o algoritmo EM-GA proposto. Em todos os casos, são fornecidas tabelas com os valores finais de verossimilhança. Os maiores valores de verossimilhança de cada conjunto de vetores (linha da tabela) são representados

[17] http://www.ics.uci.edu/∼mlearn/MLRepository.html

TAB.4.1: Bases de dados

	Dimensões (D)	**Total de vetores**	**Classes**
LANDSAT	36	4453	6
PIMA	8	768	2
SEGMENT	19	2310	7

em negrito. Inicialmente, foi estimado um modelo GMM (com 5 gaussianas) para cada conjunto de vetores. Posteriormente, os resultados comparativos foram decorrentes da seleção da ordem de modelo GMM pela medida BIC (*Bayesian Information Criterion*) (HALBE, 2005; STOICA, 2004; AJMERA, 2004).

O algoritmo EM-RP comprime os conjuntos de vetores da sua dimensão original a sub-espaços de dimensões menores com $n_{RP} = 28002$. O motivo do emprego desse número de projeções será apresentado adiante.

O treinamento do algoritmo EM-GA foi implementado com $k_c = 0,8$ (é interessante manter a fração de mutação muito mais baixa que 1,0 – no caso 0,2 – de forma a retirar do algoritmo proposto a tendência pela busca aleatória pura) e $n_{\mathbf{W}} = 30$. Foi empregada mutação gaussiana cuja variância σ_i^2 foi reduzida de forma controlada, para cada geração i, pela EQ. 4.10.

$$\sigma_i^2 = \sigma_{i-1}^2 \left(1 - S\frac{i}{n_g}\right); \sigma_0^2 = \frac{1}{2}; 1 \leq i \leq n_g \tag{4.10}$$

Foi usada seleção estocástica uniforme. Todos os cromossomos utilizados foram codificados como números reais. Para o número fixo de gaussianas e para a estimação da ordem do modelo pela medida BIC, o valor de n_g foi fixado em 1000, ou seja, adotaram-se mil gerações. Foram considerados 2 elementos vencedores em cada geração. Isto forma um total de $30 + (30 - 2)(1000 - 1) = 28002$ projeções no total (justificando o valor n_{RP} acima).

Além do acima exposto, vale lembrar que algoritmos genéticos possuem outras estratégias de mutação e recombinação. Porém, para efeito de investigação inicial de desempenho do algoritmo proposto, foi considerado o caso particular da EQ. 4.10. Nessa situação empregou-se o parâmetro de controle da redução da variância $S = 0,75$.

É importante mencionar que os operadores genéticos de mutação e de cruzamento não produzem uma geração de matrizes ortonormais, mesmo que a geração anterior seja composta de matrizes ortonormais. Nessa situação, poder-se-ia obter uma geração de matrizes ortonormais mediante uso do procedimento de Gram-Schmidt (STRANG, 1988). Isso, porém, prejudicaria a filosofia de evolução que caracteriza os algoritmos genéticos.

Assim sendo, foi relaxada a restrição de as matrizes **W** serem ortonormais.

Caso o algoritmo proposto usasse a estratégia descrita por (LIN), o cromossomo possuiria $(2D + 1)M$ elementos (M pesos, MD elementos de vetores de médias e MD elementos de matrizes de covariâncias diagonais). Com a otimização da função-objetivo do algoritmo EM-GA feita apenas em função da matriz **W**, empregam-se apenas dD elementos. O número de elementos do cromossomo do algoritmo proposto sempre será menor que o do algoritmo em (LIN) se $(2D + 1)\, M > dD$, que leva a $M > \frac{dD}{2D+1}$. Substituindo o valor de D para as bases PIMA, SEGMENT e LANDSAT (vide Tabela 4.1), tem-se, para $d = 2$ e $d = 3$, os resultados da TAB. 4.2.

TAB.4.2: Condição para economia de parâmetros

Base de dados	**Condição**	**Valor de M mínimo**
PIMA – $d = 2$	$M > 16/17$	1
PIMA – $d = 3$	$M > 24/17$	2
SEGMENT – $d = 2$	$M > 38/39$	1
SEGMENT – $d = 3$	$M > 57/39$	2
LANDSAT – $d = 2$	$M > 72/73$	1
LANDSAT – $d = 3$	$M > 108/73$	2

Nota-se, da TAB. 4.3 à TAB. 4.5 que, para todas as bases de dados testadas, o algoritmo proposto EM-GA foi tão ou mais eficaz do que o algoritmo EM na estimação ML, uma vez que os valores de verossimilhança obtidos pelo EM-GA foram maiores ou iguais aos valores de verossimilhança pelo algoritmo EM. O mesmo não ocorreu com o EM-RP, pois este não superou o algoritmo EM em um dos 18 conjuntos de vetores analisados, revelando uma maior fragilidade da busca aleatória. Além disso, nenhum valor de verossimilhança calculada pelo EM-RP superou os valores calculados pelo algoritmo EM-GA.

TAB.4.3: Verossimilhança da base PIMA usando um modelo GMM com 5 gaussianas, com projeção em $d = 2$

Classe	**EM**	**EM-RP**	**EM-GA**
1	-12568	**-11871**	**-11871**
2	-6980	-6432	**-6420**
Todas	-19716	-18671	**-18466**

A FIG. 4.2 mostra outra qualidade do algoritmo EM-GA sobre o EM-RP: a convergência mais rápida à média dos valores da função-objetivo. Percebe-se que, na média

TAB.4.4: Verossimilhança da base LANDSAT usando um modelo GMM com 5 gaussianas, com projeção em $d = 8$

Classe	**EM**	**EM-RP**	**EM-GA**
1	-118843	-118757	**-118756**
2	-59181	-58841	**-58840**
3	-104547	**-104008**	**-104008**
4	-45632	**-45089**	**-45089**
5	-57151	**-57122**	**-57122**
6	-113098	**-111790**	**-111790**
Todas	**-566986**	-566988	**-566986**

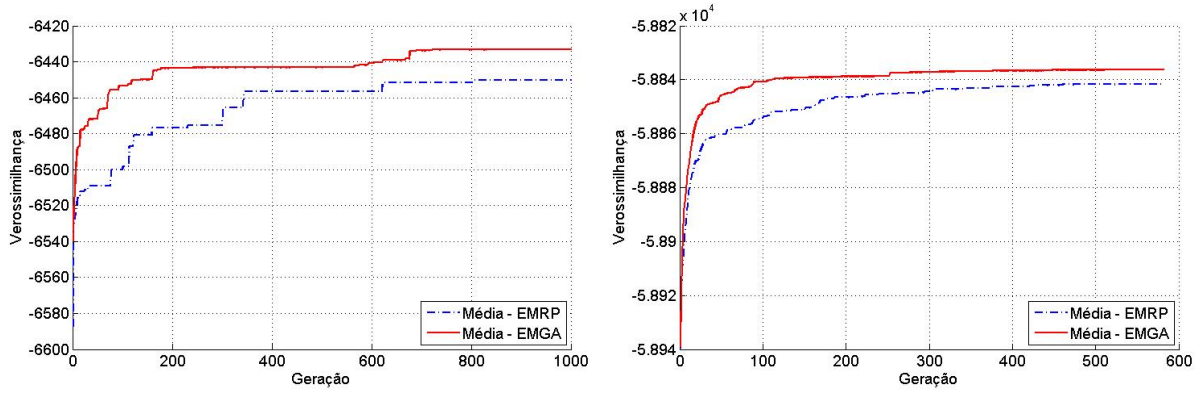
TAB.4.5: Verossimilhança da base SEGMENT usando um modelo GMM com 5 gaussianas, com projeção em $d = 4$

Classe	**EM**	**EM-RP**	**EM-GA**
1	-6740	**-6474**	**-6474**
2	-9008	**-8808**	**-8808**
3	-11296	**-9820**	**-9820**
4	-12872	**-12357**	**-12357**
5	-8801	**-8341**	**-8341**
6	-9743	**-9349**	**-9349**
7	-10247	**-9562**	**-9562**
Todas	-99438	-96065	**-96041**

das execuções dos algoritmos EM-RP e EM-GA, o algoritmo EM-GA apresenta maior vantagem em termos dos valores de verossimilhança apresentados e maior velocidade de convergência do que o EM-RP. Cabe destacar que a complexidade computacional do algoritmo genético é governada basicamente pelo ajuste dos indivíduos da população na função-custo, tendo em vista a extrema simplicidade dos operadores genéticos. Como esse ajuste é realizado tanto para o algoritmo genético quanto para a busca aleatória, no que diz respeito ao cálculo da função-custo, a complexidade computacional do algoritmo EM-GA é comparável à do EM-RP.

O valor da medida BIC (HALBE, 2005) em função da verossimilhança (EQ. 4.4) é dado por:

$$BIC\left(M\right) = -2\ell\left(\mathbf{X};\lambda\right) + \gamma\nu\log\left(n\right) \tag{4.11}$$



(a) PIMA (classe 2) (b) LANDSAT (classe 2)

FIG.4.2: Comparação entre as médias das verossimilhanças calculadas pelos algoritmos EM-RP e EM-GA. Foi computada a média de 5 e 30 realizações para as bases PIMA e LANDSAT, respectivamente. Pode ser percebida a maior eficiência do algoritmo EM-GA na média das realizações.

A segunda parcela na EQ. 4.11 representa a penalidade aplicada em função do número de vetores n de cada conjunto de vetores originais e do coeficiente ν, expresso na EQ. 4.12 em termos do número de gaussianas M da mistura e da dimensão D do conjunto de vetores de dados originais. A segunda parcela é adequada ao caso do modelo GMM com matrizes de covariância diagonais diferentes para cada gaussiana da mistura (HALBE, 2005). A constante γ pode ser ajustada a um maior ou a um menor valor (AJMERA, 2004) caso se queira penalizar a quantidade n de vetores com maior ou menor intensidade, respectivamente. A FIG. 4.3 mostra o efeito desta penalidade, refletida na escolha de uma menor ordem de modelo para maiores valores de γ.

$$\nu = 2MD + M - 1 \tag{4.12}$$

Da EQ. 4.11 resulta que, para dois algoritmos que estimem, a partir do mesmo conjunto de vetores $\mathbf{X}$, modelos GMM distintos com verossimilhanças $\mathcal{L}_1$ e $\mathcal{L}_2$ ($M_1 = M_2 = M$), tem-se:

$$\mathcal{L}_1 > \mathcal{L}_2 \Leftrightarrow BIC_1(M) < BIC_2(M) \tag{4.13}$$

Em termos práticos, a EQ. 4.13 mostra que, para um mesmo número de gaussianas, e conjunto de vetores, uma estimação ML mais bem ajustada implica em menor medida BIC.

Nesta subseção foram calculados, além das verossimilhanças, os números de gaussianas

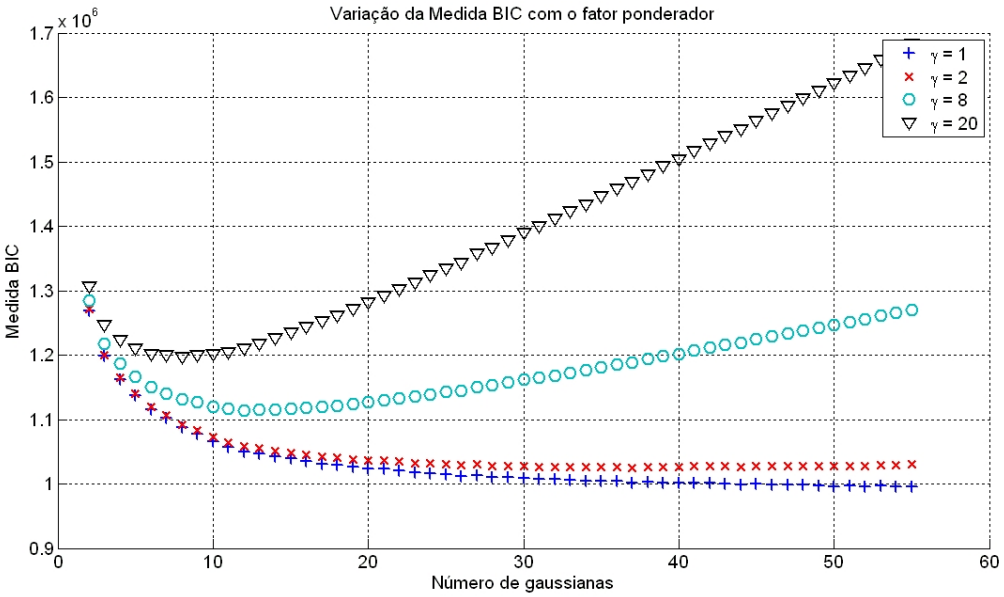


FIG.4.3: Comportamento do valor da medida BIC em função do coeficiente γ.

resultantes da seleção de ordem de modelo realizada pela medida BIC. Estes dados constam da TAB. 4.6 à TAB. 4.8 (as ordens dos modelos GMM selecionadas pela medida BIC se encontram entre parênteses; todas obedecem às condições da TAB. 4.2).

TAB.4.6: Verossimilhança - medida BIC - PIMA - $d = 2$ / $d = 3$

Classe	**EM**	**EM-RP**	**EM-GA**	n_g	γ
1	-12977(2)	-11904(5)	**-11873(5)**	2	2
		-12658(3)	**-11873(5)**		
2	-7121(2)	-6887(3)	**-6593(4)**	5	2
		-7121(2)	**-6965(2)**		
Todas	-19778(4)	-19525(3)	**-18321(6)**	30	2
		-19314(4)	**-18354(6)**	30	2

O algoritmo EM-GA proposto foi, individualmente, o que apresentou os melhores resultados. Da TAB. 4.6 à TAB. 4.8 mostrou-se que, em 34 dos 36 casos analisados (94%), o EM-GA superou o desempenho do algoritmo EM-RP, ao passo que o EM-RP foi individualmente mais eficaz apenas em 4 dos 36 casos analisados (11%). Em relação ao algoritmo EM convencional, o algoritmo proposto foi mais eficaz em 100% dos casos. O EM-GA foi mais eficiente que o EM-RP em 32 dos 36 casos analisados (89%). As tabelas e a FIG. 4.4 mostram que o algoritmo EM-GA produz uma estimação ML com melhor ajuste em comparação com o EM. Esta figura mostra o comportamento da medida

TAB.4.7: Verossimilhança - medida BIC - LANDSAT - $d = 2$ / $d = 3$

Classe	**EM**	**EM-RP**	**EM-GA**	n_g	γ
1	-117686(4)	-117355(6)	**-117283(6)**	5	4
		-117288(6)	**-117284(6)**		
2	-59820(4)	-59811(4)	**-58901(5)**	2	4
		-59814(4)	**-58901(5)**		
3	-105269(4)	-102952(6)	**-102920(6)**	5	4
		-102960(6)	**-102927(6)**		
4	-46102(4)	-45852(4)	**-45850(4)**	5	4
		-45850(4)	**-45849(4)**		
5	-59177(3)	-57127(5)	**-57123(5)**	5	4
		-57122(5)	**-57122(5)**		
6	-115639(3)	-110773(6)	**-110734(6)**	5	4
		-110725(6)	**-110708(6)**		
Todas	-533642(11)	-518954(13)	**-518677(13)**	30	8
		-518636(13)	-518779(13)		

BIC em função do número de gaussianas. Nota-se, na Figura FIG. 4.4(a), a redução da medida BIC e maior suavidade da função quando a mesma é computada em função dos modelos GMM estimados pelo algoritmo proposto. Na FIG. 4.4(b), percebe-se a maior ordem do modelo selecionado quando estimada pela técnica EM-GA. Além disso, a figura mostra na prática o resultado da EQ. 4.13 — ocorre, de fato, a melhor estimação ML para cada valor da ordem do modelo.

Outros exemplos de comparação de desempenho do algoritmo EM com o algoritmo EM-GA por meio da medida BIC são representadas na série de figuras da FIG. 4.5 à FIG. 4.10. Percebe-se a tendência à maior eficiência da estimação da ordem dos modelos GMM pelas medidas BIC para os exemplos indicados em cada figura. Uma vez que, na EQ. 4.13, foi mostrada a relação direta entre redução da medida BIC e aumento da verossimilhança para a mesma ordem do modelo GMM, pode-se concluir que a estimação dos modelos GMM é, de fato, mais eficaz pelo algoritmo EM-GA proposto.

4.3 SELEÇÃO DE CARACTERÍSTICAS

Na Subseção 3.5.2 foi comentado que não existe nenhuma característica extraída dos sinais de voz que garanta 100% de certeza acerca da identidade de um locutor. Mesmo

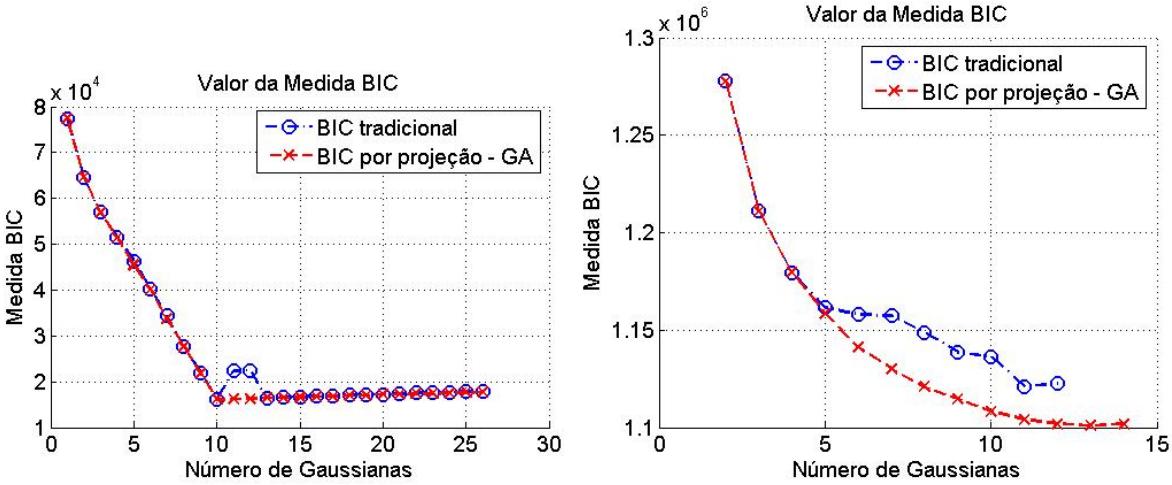
TAB.4.8: Verossimilhança - medida BIC - SEGMENT - $d = 2$ / $d = 3$

Classe	**EM**	**EM-RP**	**EM-GA**	n_g	γ
1	-6640(5)	-5915(7)	**-5910(7)**	5	2
		-5908(7)	**-5905(10)**		
2	-8833(5)	-8554(6)	**-8553(6)**	2	2
		-8596(6)	**-8330(7)**		
3	-11673(3)	**-8690(9)**	-8933(7)	30	2
		-9942(5)	**-8746(8)**		
4	-12745(4)	-11798(7)	**-11744(7)**	5	2
		-12051(6)	**-11530(8)**		
5	-8001(6)	-7641(7)	**-7274(8)**	10	2
		-7295(8)	**-6784(10)**		
6	-9789(4)	**-8744(7)**	**-8744(7)**	10	2
		-9018(6)	**-8752(7)**		
7	-9337(6)	-9369(6)	**-9314(6)**	5	2
		-9577(5)	**-9318(6)**		
Todas	-89466(8)	-87929(8)	**-85734(9)**	30	8
		-84255(10)	**-82034(11)**		

com essa constatação é possível, dentro de um esquema discriminatório, selecionar quais características são mais ou menos relevantes para a verificação automática de locutores de uma base de sinais de voz qualquer. A este tipo de técnica se denomina *seleção de características*. Um dos objetivos da seleção é permitir um maior conhecimento de quais características são mais participativas na discriminação entre locutores. Outro objetivo, decorrente do anterior, é permitir o descarte das características menos relevantes e, em conseqüência, propiciar menor tempo de verificação.

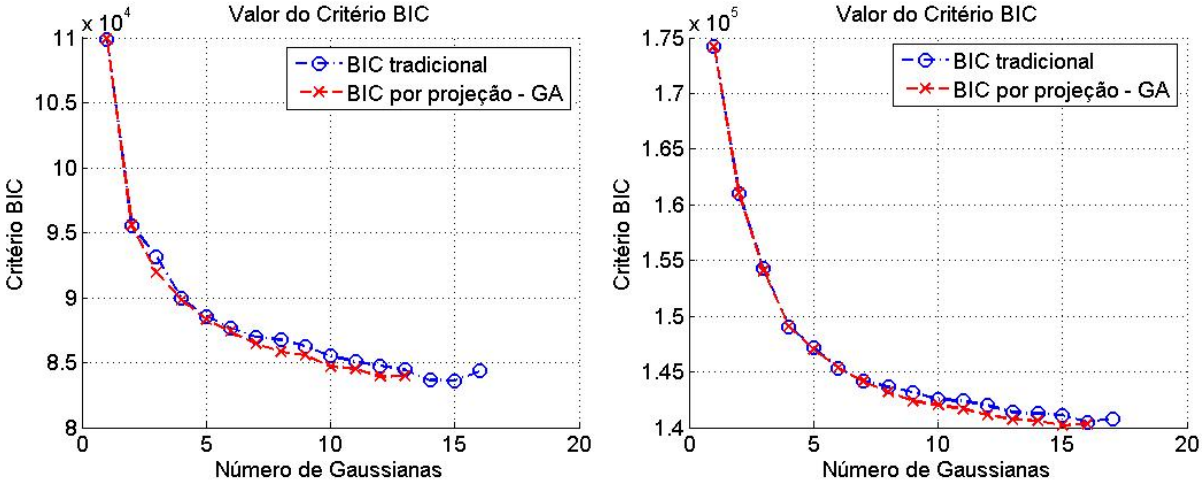
Uma das funções mais conhecidas para a discriminação de locutores é o *Discriminante de Fisher*, também conhecido como *Razão F* (BISHOP, 2006; ZENG). Para uma aplicação qualquer envolvendo K características, em particular um sistema de verificação automática com M locutores, o valor do Discriminante de Fisher Δ_f é a razão entre a variância interlocutor e a variância intralocutor, representadas respectivamente por $\sigma^{2(k)}_{inter}$ e $\sigma^{2(k)}_{intra}$ $(k = 1, \ldots, K)$:

$$\Delta_f^{(k)} = \frac{\sigma^{2(k)}_{inter}}{\sigma^{2(k)}_{intra}} \tag{4.14}$$



(a) medida BIC – modelo sintético($d = 2$) (b) medida BIC – LANDSAT ($d = 2$)

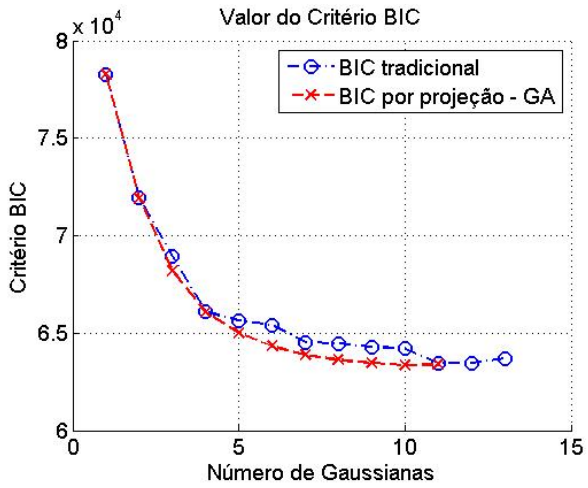
FIG.4.4: Comparação entre as Medidas BIC obtidas dos modelos GMM estimados pelo algoritmo EM e pelo algoritmo genético (EM-GA) até 25 gaussianas. Em (a) a medida BIC foi calculada sobre um modelo GMM sintético de 10 gaussianas (4000 vetores do $\mathbb{R}^{10}$ projetados ao $\mathbb{R}^2$). Notar os valores de medida BIC menores e de comportamento mais suave devido à implementação do algoritmo proposto. Além disso, o mínimo local (HALBE, 2005) (coincidente com o mínimo global) da medida BIC em (a) é de 10 gaussianas para ambos os algoritmos, exatamente a ordem arbitrada para o modelo sintético.



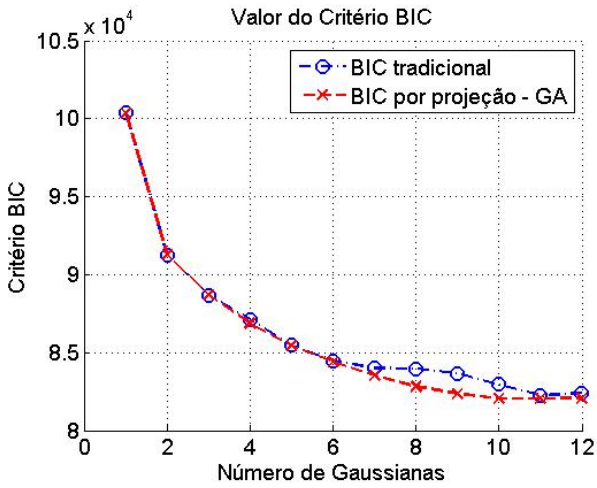
(a) LANDSAT – Classe 2 ($d = 3$) (b) LANDSAT – Classe 3 ($d = 3$)

FIG.4.5: Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 2 e 3 da base de dados LANDSAT do $\mathbb{R}^{36}$ ao $\mathbb{R}^3$.

A EQ. 4.14, conceitualmente, define a capacidade de discriminação de uma característica k do sistema de verificação baseada em estatísticas de segunda ordem. Quanto maior a variância interlocutor, mais espalhadas estarão as características de locutores di-

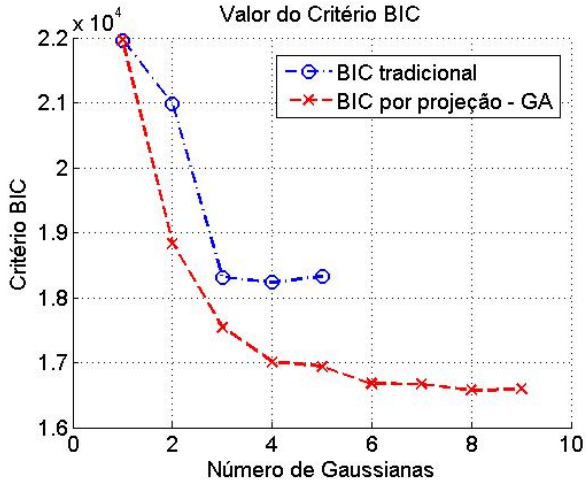


(a) LANDSAT – Classe 4 ($d = 3$)

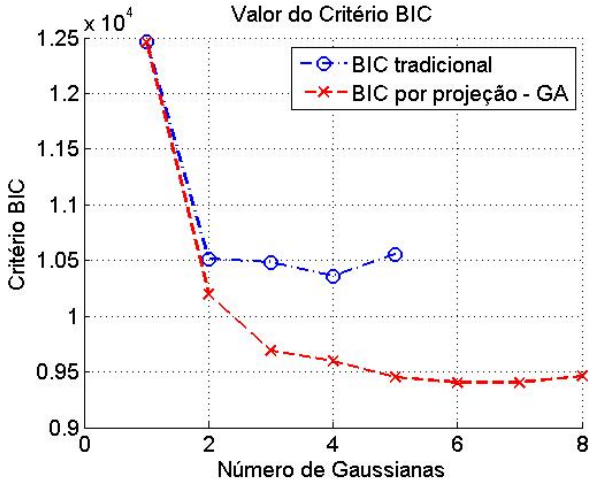


(b) LANDSAT – Classe 5 ($d = 3$)

FIG.4.6: Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 4 e 5 da base de dados LANDSAT do $\mathbb{R}^{36}$ ao $\mathbb{R}^3$.



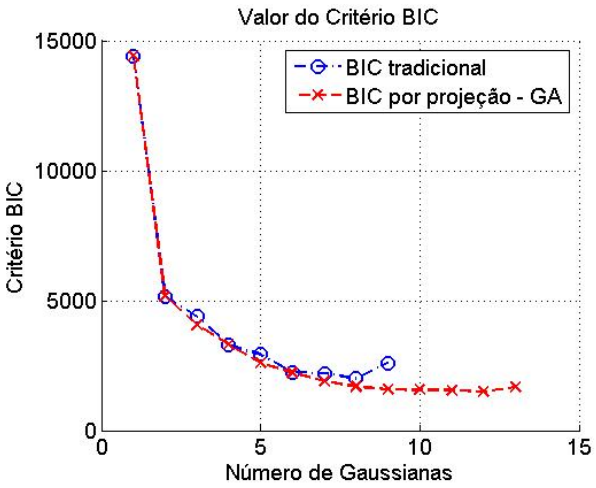
(a) PIMA – Classe 1 ($d = 3$)



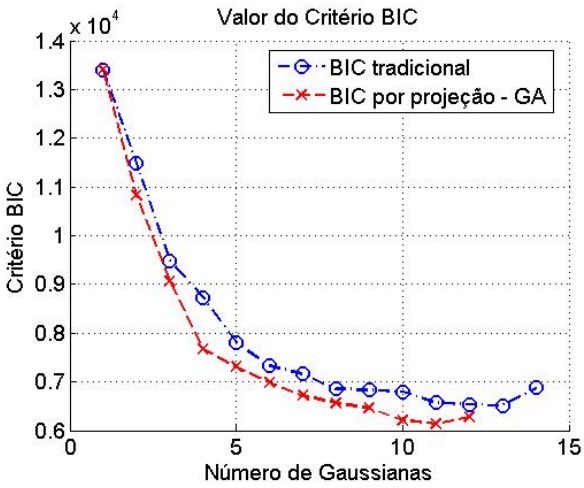
(b) PIMA – Classe 2 ($d = 3$)

FIG.4.7: Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 1 e 2 da base de dados PIMA do $\mathbb{R}^8$ ao $\mathbb{R}^3$.

ferentes, ao passo que, quanto menor a variância intralocutor, mais compactadas estarão as características dos mesmos locutores. Em outras palavras, as classes de características estarão mais separadas entre si. Estas variâncias (EQ. 4.15 e EQ. 4.16) são definidas com base no número de características N_p de cada locutor ($p = 1, \ldots, C$), nas características $x_{pj}^{(k)}$ de cada locutor ($p = 1, \ldots, C_p$) e nas médias intralocutor e interlocutor, designadas respectivamente por $m_p^{(k)}$ e $m^{(k)}$.

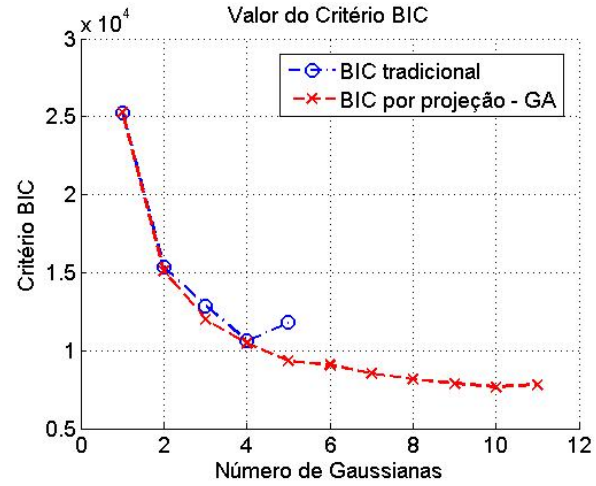


(a) SEGMENT – Classe 1 ($d = 3$)

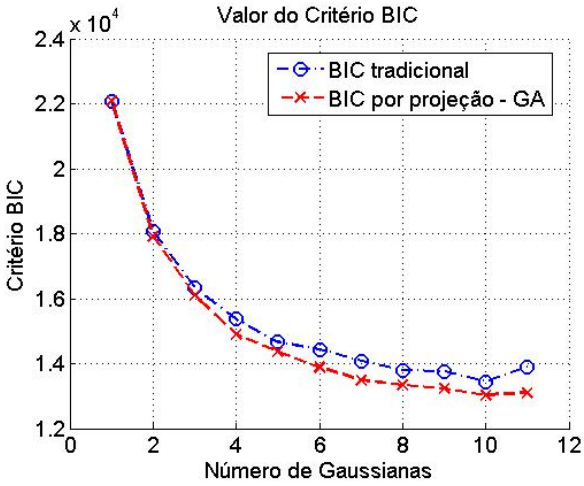


(b) SEGMENT – Classe 2 ($d = 3$)

FIG.4.8: Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 1 e 2 da base de dados SEGMENT do $\mathbb{R}^{19}$ ao $\mathbb{R}^3$.



(a) SEGMENT – Classe 3 ($d = 3$)



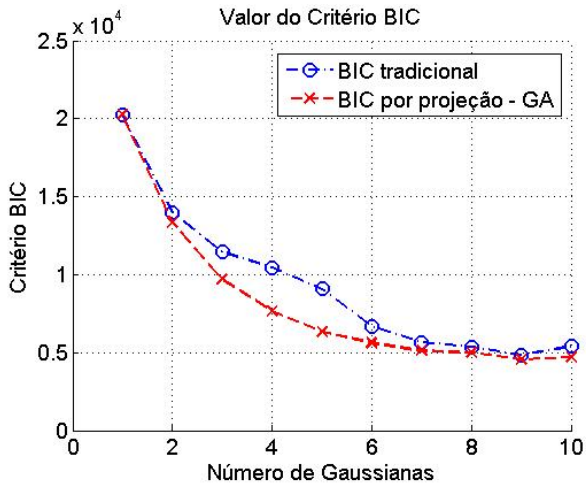
(b) SEGMENT – Classe 4 ($d = 3$)

FIG.4.9: Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 3 e 4 da base de dados SEGMENT do $\mathbb{R}^{19}$ ao $\mathbb{R}^3$.

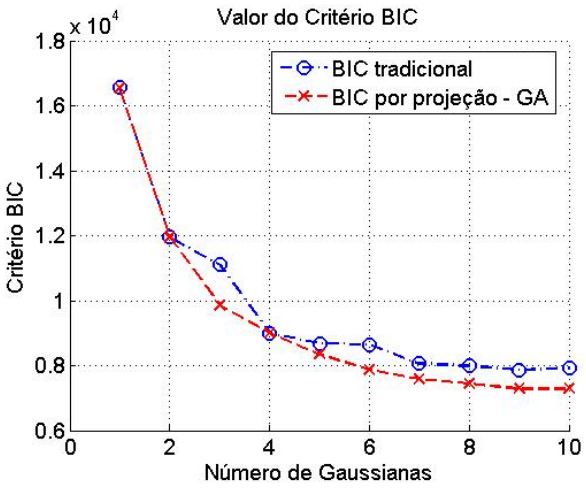
$$\sigma^{2(k)}_{intra} = \frac{1}{C-1}\sum_{p=1}^{C}\left[\frac{1}{N_p}\sum_{j=1}^{N_p}\left(x^{(k)}_{pj} - m^{(k)}_{p}\right)^2\right] \tag{4.15}$$

$$\sigma^{2(k)}_{inter} = \frac{1}{C-1}\sum_{i=1}^{M}\left(m^{(k)}_{i} - m^{(k)}\right)^2 \tag{4.16}$$

Para o caso de vetores de características, o valor do Discriminante de Fisher é dado



(a) SEGMENT – Classe 5 ($d = 3$)



(b) SEGMENT – Classe 6 ($d = 3$)

FIG.4.10: Comparação entre as Medidas BIC calculadas pelo algoritmo EM e pelo algoritmo genético (EM-GA) nos moldes da anterior. Neste caso, projetam-se as classes 5 e 6 da base de dados SEGMENT do $\mathbb{R}^{19}$ ao $\mathbb{R}^3$.

pela equação EQ. 4.17, em função da matriz de covariância intralocutor $\mathbf{\Sigma}_{intra}$ e interlocutor $\mathbf{\Sigma}_{inter}$ definidas pelas equações EQ. 4.18 e EQ. 4.19, respectivamente. A p-ésima classe de características é representada na EQ. 4.18 por $\mathcal{C}_p$,$p = 1, \ldots, C$. Em ambas as equações EQ. 4.18 e EQ. 4.19, os vetores-médios intralocutor e interlocutor são denotados, respectivamente, por $\boldsymbol{\mu}_p$ e $\boldsymbol{\mu}$. O número de vetores de características extraídos de cada locutor é dado por N_p.

$$\Delta_f = \mathrm{Tr}\left\{\mathbf{\Sigma}_{intra}^{-1}\mathbf{\Sigma}_{inter}\right\} \tag{4.17}$$

$$\mathbf{\Sigma}_{intra} = \sum_{p=1}^{C}\sum_{n\in\mathcal{C}_p}\left(\mathbf{y}_n - \boldsymbol{\mu}_p\right)\left(\mathbf{y}_n - \boldsymbol{\mu}_p\right)^T \tag{4.18}$$

$$\mathbf{\Sigma}_{inter} = \sum_{p=1}^{C} N_p\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_p - \boldsymbol{\mu}\right)^T \tag{4.19}$$

A seleção de características pode exigir uma complexidade computacional elevadíssima, dependendo do número de características envolvidas. Por exemplo, se forem testadas 3 características e todas as suas combinações (individualmente, em duplas e o trio), haverá 8 combinações diferentes. Se esse número de características for elevado para 20, serão 1048575 combinações diferentes ($2^K - 1$ combinações[18] para K características). Uma solução pode ser a escolha das características que exibirem os melhores coeficientes

[18]Subtrai-se o valor 1 de 2^K porque não se considera o subconjunto vazio das características.

individualmente ou em grupos de mesmo tipo de característica (por exemplo, se forem extraídos dos sinais de voz 15 coeficientes MFCC e 4 coeficientes SSC, computar dois valores de Discriminante de Fisher — um para os 15 coeficientes MFCC e outro para os 4 coeficientes SSC). A seleção de características é imprescindível quando se deseja manter algumas características originais inalteradas (por exemplo, o perfil de formantes de um locutor em fones específicos). Além disso, é interessante empregar a seleção de características quando o número de características irrelevantes pelo critério do Discriminante de Fisher for muito maior do que o número das características relevantes.

4.4 CONCLUSÃO

Neste capítulo foi abordado o importante conceito de modelos de misturas de gaussianas (GMM), enfocando as técnicas que permitem um melhor ajuste dos modelos estimados. De fato, comprova-se experimentalmente (DASGUPTA; LIN; FLORES) que técnicas de estimação de modelos GMM envolvendo projeções aleatórias e computação evolucionária tendem a uma estimação global em máxima verossimilhança. Também foi abordado sinteticamente o Discriminante de Fisher, parâmetro importante na determinação das características de melhor discriminação entre locutores, visando a redução de dimensões dos vetores extraídos de cada locutor. A melhoria da eficiência das técnicas de VAL, em conseqüência do emprego das técnicas abordadas neste capítulo, será objeto de análise do capítulo seguinte.

5 AVALIAÇÃO DE RESULTADOS EM CONTRIBUIÇÃO À METODOLOGIA DE FONÉTICA FORENSE

5.1 INTRODUÇÃO

Este capítulo trata das contribuições práticas à metodologia de perícia em âmbito forense, focada no contexto de verificação automática de locutor (VAL). Serão discutidas as ferramentas matemáticas e estatísticas que dão subsídio a esta teoria. A Seção 5.2 aborda sinteticamente um sistema genérico de verificação por GMM. A Seção 5.3 descreve as avaliações de VAL realizadas neste capítulo. A Seção 5.4 aborda as taxas de erros no contexto da VAL. A Seção 5.5 expõe os resultados das avaliações de VAL empregando a técnica LTF comentada previamente na Subseção 3.5.2. A Seção 5.6 fornece os resultados das avaliações empregando outras características, tais como SSC, SCF, SFM, MFCC e tonalidade. A Seção 5.7 avalia o ganho de desempenho da VAL decorrente da inserção das projeções aleatórias e dos algoritmos genéticos na estimação dos modelos GMM dos locutores. Por fim, na Seção 5.8, são tecidas as conclusões sobre as técnicas abordadas no presente capítulo.

5.2 ESQUEMA DE VERIFICAÇÃO POR GMM

Todas as ferramentas estatísticas abordadas nas Seções 4.2 e 4.3 servem de suporte para aumentar a eficiência das técnicas de verificação de locutor empregando modelos de misturas de gaussianas. A verificação está inserida no contexto da perícia em Fonética Forense por se tratar de uma aplicação em que se deseja inferir se o locutor que proferiu um conjunto de falas contido nas peças-padrão coincide com o locutor que proferiu as falas da peça-motivo. A verificação difere da identificação (REYNOLDS, 1995b) pois esta engloba locutores de um conjunto predefinido (também chamado de conjunto fechado); ao contrário, a verificação é um teste realizado sobre um conjunto aberto de locutores (o locutor em questão, locutores de teste e locutores desconhecidos). Os locutores desconhecidos formam o modelo universal de falsos locutores (ou UBM, de *Universal Background Model*).

É necessário ressaltar que a verificação consiste em um teste de hipótese, no qual se quer decidir se as locuções de um determinado locutor realmente são do locutor em

questão, cujo modelo GMM pode ser estimado pelas técnicas da Seção 4.2, ou são de outros locutores, cada qual definido por um modelo GMM (potenciais suspeitos) ou definidos por um modelo GMM único (caso do GMM para o UBM, que estima um modelo GMM único como se todos os locutores fossem um falso locutor único). Na teoria, as hipóteses em questão H_0 e H_1 são:

- **Hipótese** H_0: A locução X é do locutor em questão (locutor verdadeiro). Assume como modelo do locutor $\lambda_0 = \{\pi_{j,0}, \boldsymbol{\mu}_{j,0}, \boldsymbol{\Sigma}_{j,0} \mid j = 1, \ldots, M\}$;
- **Hipótese** H_1: A locução X não é do locutor em questão (locutor falso). Assume como modelo do locutor $\lambda_1 = \{\pi_{j,1}, \boldsymbol{\mu}_{j,1}, \boldsymbol{\Sigma}_{j,1} \mid j = 1, \ldots, M\}$.

Empregando o critério da máxima probabilidade *a posteriori* (MAP, de *Maximum a posteriori*) (VANTREES, 1968), a fala X corresponderá ao locutor verdadeiro caso a probabilidade *a posteriori* do modelo do locutor verdadeiro $P(\lambda_0 \mid X)$ supere a probabilidade *a posteriori* do locutor falso $P(\lambda_1 \mid X)$, ou seja:

$$P(\lambda_0 \mid X) > P(\lambda_1 \mid X). \tag{5.1}$$

A EQ. 5.1 pode ser reescrita com o auxílio da Regra de Bayes, evidenciando as probabilidades *a priori* das hipóteses $P(H_0)$ e $P(H_1)$, a probabilidade de ocorrência da fala $P(X)$ e as funções de verossimilhança $P(X \mid \lambda_0)$ e $P(X \mid \lambda_1)$:

$$\frac{P(X \mid \lambda_0)P(H_0)}{P(X)} > \frac{P(X \mid \lambda_1)P(H_1)}{P(X)}. \tag{5.2}$$

Cancelando os termos $P(X)$ e assumindo as probabilidades *a priori* $P(H_0)$ e $P(H_1)$ iguais, o critério MAP pode ser reescrito como uma razão de funções de verossimilhança

$$\frac{P(X \mid \lambda_0)}{P(X \mid \lambda_1)} > 1, \tag{5.3}$$

que é, na verdade, um caso particular do teste de razão de verossimilhança (LRT, de *Likelihood Ratio Test*), expresso em função da razão de verossimilhança LR, dada por:

$$LR = \frac{P(X \mid \lambda_0)}{P(X \mid \lambda_1)} \tag{5.4}$$

Reescrevendo a EQ. 5.4 na forma logarítmica, tem-se a expressão da razão logarítmica de verossimilhança $\Lambda(X)$:

$$\Lambda(X) = \log p(X \mid \lambda_0) - \log p(X \mid \lambda_1) \tag{5.5}$$

Os termos da EQ. 5.5 são as funções de verossimilhança das falas supondo que tenham sido proferidas pelo locutor em questão ou por um locutor desconhecido, respectivamente. O esquema de verificação que ilustra a razão de verossimilhança se encontra representado pela FIG. 5.1. Decide-se pela hipótese λ_0 caso a razão de verossimilhanças da EQ. 5.5 seja maior que um dado limiar θ. Caso contrário, opta-se pela hipótese λ_1.

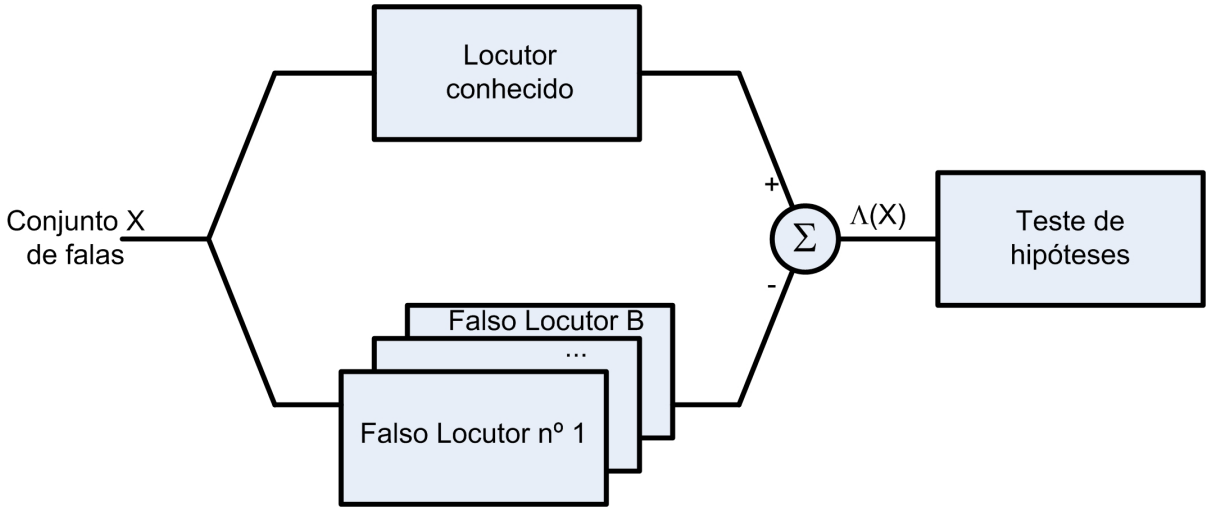


FIG.5.1: Sistema de verificação de locutor. Pode ser percebida a razão de verossimilhanças logarítmica $\Lambda(X)$ na entrada do bloco do teste de hipóteses

Assumindo T vetores de características $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, a verossimilhança referente ao modelo λ_0 pode ser calculada como

$$\log p(X \mid \lambda_0) = \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_0), \tag{5.6}$$

ao passo que a verossimilhança referente ao modelo λ_1 pode ser calculada de acordo com o modelo de *background* escolhido:

- **Modelo *cohort***: Utilizando um modelo GMM λ_b para cada um dos B locutores que compõe o *background*, tal que $\lambda_b = \{\pi_{j,b}, \boldsymbol{\mu}_{j,b}, \boldsymbol{\Sigma}_{j,b} \mid j = 1, \ldots, M; b = 1, \ldots, B\}$, com características mais próximas de um locutor verdadeiro. Nesse caso, a verossimilhança $P(X \mid \lambda_1)$ será dada pela EQ. 5.7 (REYNOLDS, 1995a)

$$\ln P(X \mid \lambda_1) = \frac{1}{B} \sum_{b=1}^{B} \ln P(X \mid \lambda_b), \tag{5.7}$$

  cujo fator $\frac{1}{B}$ funciona como normalização das verossimilhanças com respeito a cada locutor do *background*;

- **Modelo de *background* universal (UBM, de *Universal Background Model*)**: Empregando um modelo GMM treinado como se o *background* fosse composto por um único locutor. Conseqüentemente, a verossimilhança $P(X \mid \lambda_1)$ será calculada pela EQ. 5.8 (REYNOLDS, 2000)

$$\log p(X \mid \lambda_1) = \frac{1}{T_b} \sum_{t=1}^{T_b} \log p(\mathbf{x}_t \mid \lambda_1), \tag{5.8}$$

  assumindo extraídos T_b vetores de características do *background*.

O fator $\frac{1}{T}$ contido na EQ. 5.6 serve como normalização, necessária para compensar efeitos de variação de duração das sentenças dos diferentes locutores. Já o fator $\frac{1}{B}$ serve, no caso do modelo de *background* do tipo *Cohort*, para normalizar a verossimilhança do modelo de locutores falsos de forma a torná-la insensível ao número de modelos falsos.

5.2.1 COMPOSIÇÃO DO *BACKGROUND*

Neste trabalho foi escolhido, por simplicidade, o modelo de *background* do tipo UBM. Nesse caso, o modelo dos locutores falsos deve ser treinado como um modelo GMM único para todos os B locutores a descaracterizar do locutor conhecido. É importante salientar que o modelo de locutores falsos deve ser compatível com o tipo de sistema de verificação a implementar. Por exemplo, se o sistema de verificação só abranger locutores do sexo masculino, o modelo de locutores falsos deverá também abranger locutores homens. Se houver mistura de sexos, é necessário que o modelo de locutores falsos possua um percentual de locutores ou locutoras o mais próximo possível da proporção dos locutores conhecidos. Até mesmo as condições de canal de comunicações e de ambiente de gravação devem respeitar as mesmas condições quanto ao nível de ruído, se a gravação transcorreu em estúdio ou por conversação telefônica; no caso telefônico, se a chamada ocorreu por canal de telefonia fixa ou móvel.

Não existe uma regra geral para a composição do *background*. Define-se por senso comum na literatura (REYNOLDS, 2000) que a composição seja a mais balanceada possível para o modelo do *background* não ser tendencioso para um determinado conjunto de locutores falsos (conhecido como sub-população) e para uma determinada ordem (número de componentes gaussianas) de modelo GMM. Nesse caso, para os testes de VAL por modelos GMM abordados nas Seções 5.6 e 5.7, a ordem do modelo GMM do *background* será sempre maior do que a ordem dos modelos GMM das falas de treinamento dos locutores envolvidos nas avaliações.

5.3 DESCRIÇÃO SUMÁRIA DOS TESTES REALIZADOS

A base de voz utilizada (doravante denominada IME2001) é constituída do total de 7.252 arquivos em formato *wav* contendo sinais de voz de 40 locutores homens, gravados em laboratório com microfones de eletreto. Para cada locutor, os sinais foram subdivididos em um sinal de voz de treinamento de 60 segundos, e vários sinais de voz de teste com 3, 10 e 30 segundos. Como o número de repetições de arquivos de áudio de teste era variável para cada locutor, optou-se por realizar os testes subseqüentes sobre uma seleção aleatória de 10 arquivos de cada duração por locutor. Os sinais de voz de todos os arquivos de áudio foram capturados com amostragem de 8 KHz. Esta taxa de amostragem foi escolhida por ser bastante consagrada na comunidade científica como o padrão empregado em telefonia fixa.

Além dos 40 locutores de treinamento e teste, foi empregado um *background* de 10 locutores adicionais (totalizando 4 minutos de fala), cujas vozes não foram utilizadas no treinamento e nos testes. O *background*, como mencionado na seção anterior, foi treinado com 64 componentes gaussianas.

A FIG. 5.2 esquematiza o plano de testes a ser seguido neste trabalho. O primeiro teste realizado (a ser discutido na Seção 5.5) envolveu a avaliação da técnica LTF (vide a Subseção 3.5.2). Os formantes dos locutores foram extraídos em ambiente assíncrono (como visto na Seção 2.6) e *pitch*-síncrono. Os testes com sincronismo de *pitch* consistiram na extração dos formantes a cada 1, 2, 3 e 4 períodos de *pitch*; exceto no caso com um período, foi considerada a superposição de apenas um período de *pitch*. Os resultados são exibidos em duas etapas:

- **Etapa 1**: o conjunto de formantes que obteve o maior discriminante de Fisher (razão-F). Este conceito foi abordado no capítulo anterior, na Seção 4.3;

- **Etapa 2**: as taxas de FR (falsa rejeição), FA (falsa aceitação) e o erro total para o conjunto de formantes de melhor discriminação da etapa 1, variando o tempo de teste de 3 s, 10 s e 30 s com resolução de 200 bandas (ou *bins*) de resolução do histograma, igualmente espaçadas;

- **Etapa 3**: as taxas de FR, FA e o erro total para o mesmo conjunto acima, com tempo de teste fixo em 30 s e resolução dos histograma assumindo os valores de 200, 100, 50, 20 e 10 bandas de resolução do histograma, igualmente espaçadas.

Para os testes de LTF não foi empregado *background*, pois ocorre comparação direta de

pares de histogramas de sinais de voz.

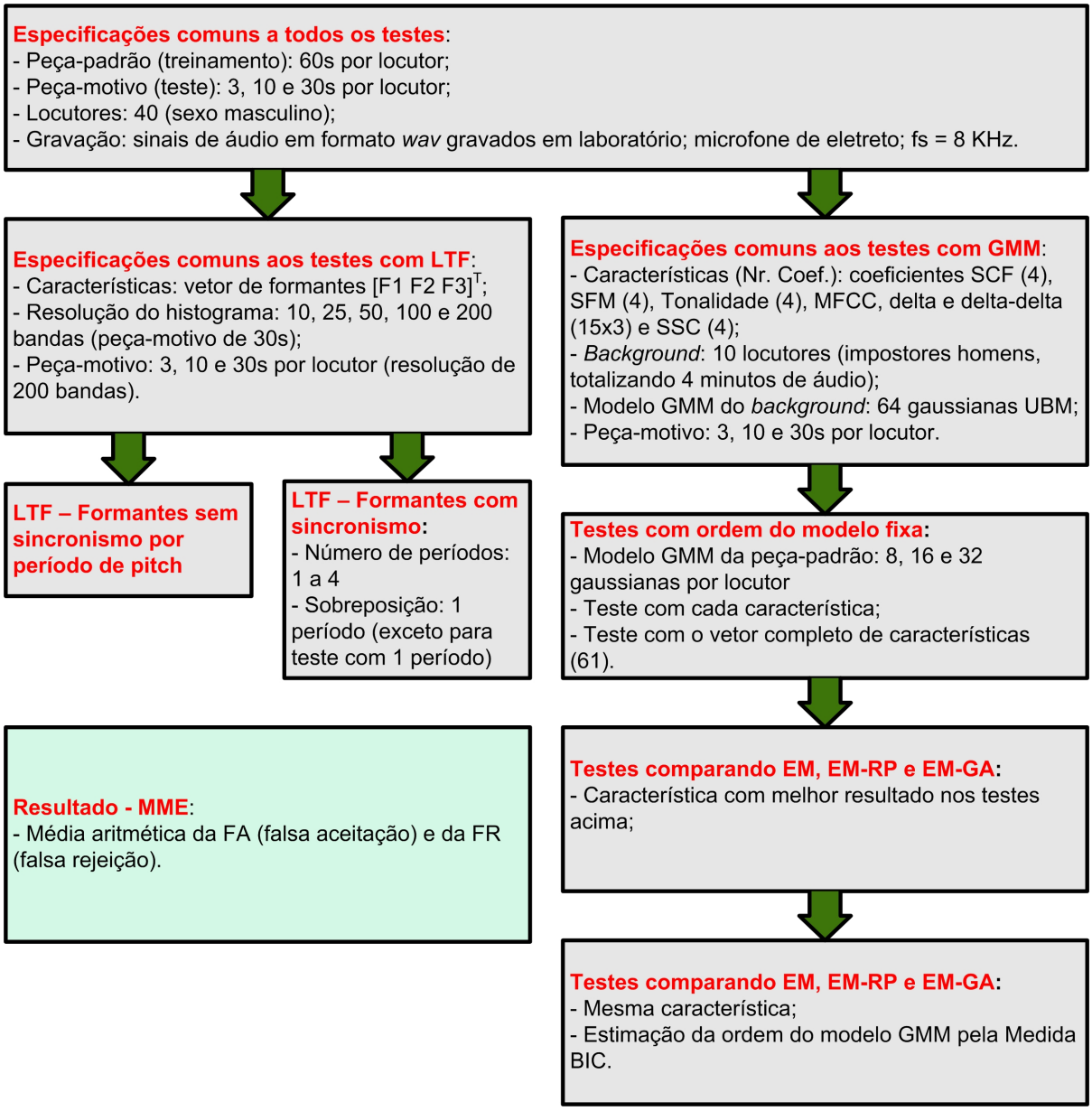


FIG.5.2: Plano de testes a ser seguido neste trabalho.

O segundo teste realizado (a ser discutido na Seção 5.6) buscou a seleção das características com maior poder de discriminação através do discriminante de Fisher. Em outras palavras, levantou-se uma tabela com as características mais discriminativas dentre SFM, SCF, Tonalidade, SSC e coeficientes MFCC completos (incluindo coeficientes MFCC, Delta e Delta-Delta) (vide Capítulo 2). Após o treinamento do modelo GMM de todas as características para cada locução de treinamento, foi efetuada a verificação do GMM das características isoladas e das características mais discriminativas. Para as características isoladas, computou-se o teste de verificação para testes de 3 s, 10 s e 30 s,

com modelos GMM de 8, 16 e 32 gaussianas. Para as características mais discriminativas, realizou-se o mesmo teste a 3 s, 10 s e 30 s.

O terceiro teste realizado (a ser discutido na Seção 5.7) avaliou comparativamente a VAL empregando os modelos GMM estimados pelos algoritmos EM simples, EM-RP e EM-GA (enunciados na Seção 4.2). Para este teste, a verificação foi efetuada sobre as características de melhor discriminação do teste anterior para 3 s, 10 s e 30 s de teste, com modelos GMM de 8, 16 e 32 gaussianas.

Para todos os testes de VAL envolvendo GMM (Seções 5.6 e 5.7), foi empregada a ordem de 64 gaussianas para o modelo GMM do *background*. Além disso, foi utilizado *background* do tipo UBM.

5.4 MODELAGEM DOS ERROS

Para cada teste de similaridade de LTF ou de verificação por GMM, foram armazenados dois vetores: um vetor contendo os *scores* relativos aos testes verdadeiros (locutores iguais para modelo de verificação e sinal de voz de teste entrante) e outro vetor contendo os *scores* relativos aos testes falsos (locutores distintos para o modelo e o sinal entrante).

Para caracterizar o sistema de verificação, é necessário introduzir dois tipos de erros, já citados na Subseção 3.3.5. O primeiro erro, conhecido como taxa de falsa aceitação (FA), ou erro Tipo-II, mede o percentual de ocorrência de aceitações do locutor falso pelo sistema. O segundo, conhecido como taxa de falsa rejeição (FR), ou erro Tipo-I, mede o percentual de ocorrência de falhas do sistema de verificação em detectar o locutor verdadeiro. Em sistemas de perícia, ambos são especialmente preocupantes. Um alto índice de falsa aceitação pode culpar um inocente em uma aplicação forense. Contudo, o ajuste do sistema de verificação com o intuito de reduzir a taxa de FA pode representar o aumento da taxa de culpados sendo inocentados (em termos práticos, ocorre o aumento da taxa de FR). Deve haver, portanto, um compromisso entre os valores das taxas de FA e FR.

Neste trabalho, dentre as várias técnicas existentes — Critério de Neyman-Pearson, mínima probabilidade de erro e Teste de Bayes (VANTREES, 1968) — foi considerada a mínima probabilidade de erro como o critério de determinação do limiar θ. Esse critério consiste no cômputo do erro médio[19] E como sendo a média aritmética dos erros de falsa aceitação E_{FA} e falsa rejeição E_{FR},

[19] O "erro total" mencionado na Etapa 2 da seção anterior.

$$E = \frac{E_{FR} + E_{FA}}{2}, \tag{5.9}$$

e na conseqüente escolha do limiar θ ajustado para o valor mínimo de E, doravante denominado MME (de *Minimum Mean Error*).

O NIST (*National Institute of Standards and Technology*) (PRZYBOCKI) prevê custos relativos à falsa aceitação e à falsa rejeição, denotados respectivamente como C_{FA} e C_{FR}. Além disso, são previstas probabilidades de falsa rejeição dado que o locutor verdadeiro é detectado e de falsa aceitação dado que o locutor verdadeiro não é detectado, denotadas respectivamente como $P(FR \mid \lambda_0)$ e $P(FA \mid \lambda_1)$. As probabilidades *a priori* do locutor verdadeiro, $P(\lambda_0)$, e do locutor falso, $P(\lambda_1)$, também são contemplados por esse modelo. Esses dados permitem a composição de uma forma alternativa de erro, conhecida como DCF (de *Detection Cost Function*, função de custo de detecção), expressa pela EQ. 5.10. Neste trabalho, para propósito de avaliação do MME, empregou-se $C_{FA} = 1$, $C_{FR} = 1$, $P(\lambda_0) = P(\lambda_1) = \frac{1}{2}$, condição para a qual a EQ. 5.10 se torna equivalente à EQ. 5.9 como verificado abaixo:

$$\begin{aligned} E &= C_{FA}P(FA \mid \lambda_1)P(\lambda_1) + C_{FR}P(FR \mid \lambda_0)P(\lambda_0) \\ &= C_{FA}E_{FA}P(\lambda_1) + C_{FR}E_{FR}P(\lambda_0) \\ &= E_{FA}\frac{1}{2} + E_{FR}\frac{1}{2} \\ &= \frac{E_{FR} + E_{FA}}{2} \quad \square \end{aligned} \tag{5.10}$$

5.5 TESTES DE FORMANTES POR LTF

Independente do algoritmo de aquisição de formantes implementado sobre cada locução, será atribuída a cada sinal de voz uma matriz de formantes de N linhas e N_f colunas. Cada linha e cada coluna estão associadas, respectivamente, a uma amostra do sinal de voz e a um formante específico. Ou seja, serão usados os N_f primeiros formantes de cada sinal de voz. Contudo, nada impede que a implementação contemple submatrizes desta matriz de formantes com algumas colunas selecionadas, de acordo com as ordens de formantes escolhidas para a análise.

Com respeito aos histogramas de formantes pelo método LTF, foram escolhidos os seguintes parâmetros básicos mínimos:

- **Quanto à escolha dos formantes**: foram escolhidos os 3 primeiros formantes,

pois são os menos afetados pelos efeitos do canal telefônico em uma situação real (KÜNZEL, 2001).

- **Quanto à resolução freqüencial do histograma**: foram comparados resultados com a resolução igual ao comprimento total da escala (no caso, 4 KHz) dividido por 200, 100, 50, 25 e 10 bandas, respectivamente.

Quanto à medida de distância entre os histogramas, foi implementada a distância euclidiana. O intuito da distância euclidiana é calcular a dissimilaridade entre os contornos dos histogramas, como foi ressaltado anteriormente na Subseção 3.5.2.

Para a geração de resultados, assumindo os histogramas de treinamento do locutor L representados como $f_L^{(trein)}$ e os histogramas de teste do locutor l como sendo $f_l^{(teste)}$, então a distância pode ser modelada pela EQ. 5.11:

$$d\left(f_L^{(trein)}, f_l^{(teste)}\right) = \sum_{i=1}^{N} \left[\left(f_{L,i}^{(trein)} - f_{l,i}^{(teste)}\right)^T \left(f_{L,i}^{(trein)} - f_{l,i}^{(teste)}\right) \right] \tag{5.11}$$

Nota-se na EQ. 5.11 que a distância é calculada em função da ordem dos formantes F_i, $i = 1, \ldots, N$. Caso se queira, em vez dos N primeiros formantes, obter a distância de formantes não-consecutivos, basta alterar os índices do somatório na equação.

No sistema implementado, são armazenados dois vetores distintos de distâncias para os casos $L = l$ (locutores iguais) e $L \neq l$ (locutores diferentes). O algoritmo do NIST descrito na Seção 5.4 processa ambos os vetores e, como resultado, gera a taxa de erro média (MME), a falsa aceitação (FA) e a falsa rejeição (FR). Foram avaliados os formantes extraídos com sincronismo temporal (janelas de 20 ms com sobreposição de 10 ms, sem sincronismo por *pitch*) e com sincronismo por *pitch*. Os formantes *pitch*-síncronos foram extraídos a cada 1, 2, 3 ou 4 períodos, com sobreposição de 1 período para os três últimos casos.

Para todos os casos avaliados, os testes foram efetuados para o vetor de formantes que apresentou o maior índice de Discriminante de Fisher, cujos valores estão listados na TAB. 5.1 em negrito.

Nota-se que, em todos os casos listados na TAB. 5.1, o terceiro formante obteve o maior valor de razão-F. Além disso, dos três casos analisados, o terceiro formante apareceu em todos. Este fato ressalta que os formantes de maior ordem são, de fato, mais importantes na discriminação entre locutores.

A TAB. 5.2 serve de legenda para as abreviações utilizadas para os testes indicados na TAB. 5.1 e indica as tabelas de referência desta seção.

TAB.5.1: Avaliação do discriminante de Fisher para os testes de LTF (vide legenda na TAB. 5.2.)

Vetor de formantes	Razão F						
	FSS	FS1P	FS1PS	FS2P	FS2PS	FS3P	FS4P
$[F_1\ F_2\ F_3]^T$	**0,157**	**0,170**	**0,116**	**0,308**	**0,185**	**0,285**	**0,322**
$[F_1\ F_3]^T$	0,133	0,142	0,098	0,254	0,157	0,236	0,263
$[F_2\ F_3]^T$	0,135	0,130	0,097	0,227	0,157	0,216	0,239

TAB.5.2: Legenda auxiliar para a TAB. 5.1.

Abreviação	**Tipo de sincronismo empregado no LTF**	**Tabelas**
FSS	Sem sincronismo por *pitch*	5.3, 5.4
FS1P	Sincronismo por 1 período de *pitch*, trechos sonoros	5.5, 5.6
FS1PS	Sincronismo por 1 período de *pitch*, trechos sonoros e surdos	5.7, 5.8
FS2P	Sincronismo por 2 períodos de *pitch*, trechos sonoros	5.9, 5.10
FS2PS	Sincronismo por 2 períodos de *pitch*, trechos sonoros e surdos	5.11, 5.12

A TAB. 5.3 mostra a avaliação dos testes LTF para os formantes extraídos sem sincronismo (FSS) — por janelas fixas de 20 ms com sobreposição de 10 ms. Foram realizados testes de 3 s, 10 s e 30 s com resolução do histograma fixa a 200 bandas. Nota-se a redução progressiva das taxas de FR e FA e do MME à medida que o tempo de teste das locuções é aumentado.

TAB.5.3: Teste FSS — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas.

Tempo	**FR(%)**	**FA(%)**	**MME(%)**
3 s	21,77	27,31	24,54
10 s	11,12	15,56	13,34
30 s	**7,77**	**8,87**	**8,32**

Além disso, a TAB. 5.4 mostra que, para o tempo de teste fixado a 30 s, a taxa de FA e o MME tendem a ser menores para uma condição intermediária de resolução do histograma (em 100 bandas). Apesar da taxa de FA ser significativamente menor para 100 bandas de resolução, a variação do MME não parece ser significativa para 50, 100 e 200 bandas, mostrando uma variação muito grande da "resolução ótima" para o MME.

Para as tabelas seguintes (TAB. 5.5 a TAB. 5.12), listadas na TAB. 5.2, também ocorre o mesmo fenômeno observado na TAB. 5.3 e na TAB. 5.4: os erros de FA, de FR

TAB.5.4: Teste FSS — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s.

Resolução	**FR(%)**	**FA(%)**	**MME(%)**
200 bandas	**7,77**	8,77	8,32
100 bandas	10,76	**5,81**	**8,28**
50 bandas	9,36	7,53	8,45
20 bandas	10,76	7,62	9,19
10 bandas	12,75	10,54	11,65

e o MME diminuem progressivamente com o aumento do tempo de teste de 3 s para 30 s (fato confirmado pelas figuras FIG. 5.3 e FIG. 5.5), bem como ocorre um menor MME para números de bandas intermediários do histograma (vide FIG. 5.4 e FIG. 5.6). Ao contrário do observado na TAB. 5.4, nem sempre o menor erro de FA foi observado no mesmo valor de resolução do histograma cujo MME foi menor, como pode ser observado nas tabelas TAB. 5.8 e TAB. 5.10. Contudo, pode ser observada a tendência a valores baixos de erro de FA na região de menor MME, embora esses valores não sejam os mínimos.

TAB.5.5: Teste FS1P — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas.

Tempo	**FR(%)**	**FA(%)**	**MME(%)**
3 s	30,85	28,89	29,87
10 s	13,25	20,04	16,65
30 s	**8,37**	**11,46**	**9,91**

TAB.5.6: Teste FS1P — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s.

Resolução	**FR(%)**	**FA(%)**	**MME(%)**
200 bandas	8,37	11,46	9,91
100 bandas	6,77	11,57	9,17
50 bandas	8,76	**9,16**	**8,96**
20 bandas	**5,38**	15,24	10,31
10 bandas	7,97	15,20	11,58

TAB.5.7: Teste FS1PS — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas.

Tempo	**FR(%)**	**FA(%)**	**MME(%)**
3 s	24,85	31,58	28,21
10 s	12,67	17,15	14,91
30 s	**8,76**	**7,20**	**7,98**

TAB.5.8: Teste FS1PS — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s.

Resolução	**FR(%)**	**FA(%)**	**MME(%)**
200 bandas	8,76	**7,20**	7,98
100 bandas	7,77	7,38	7,57
50 bandas	6,77	7,78	**7,28**
20 bandas	**5,98**	10,37	8,17
10 bandas	6,97	13,76	10,36

TAB.5.9: Teste FS2P — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas.

Tempo	**FR(%)**	**FA(%)**	**MME(%)**
3 s	21,93	31,81	26,87
10 s	13,70	15,40	14,55
30 s	**10,96**	**8,03**	**9,49**

TAB.5.10: Teste FS2P — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s.

Resolução	**FR(%)**	**FA(%)**	**MME(%)**
200 bandas	10,96	**8,03**	9,49
100 bandas	10,16	8,36	9,26
50 bandas	10,36	8,09	**9,22**
20 bandas	9,16	10,40	9,78
10 bandas	**6,18**	16,14	11,16

Na FIG. 5.3, percebe-se claramente que os testes de LTF para o caso FSS (formantes sem sincronismo por *pitch*) obtiveram melhores resultados para os testes de 3 s, 10 s e 30 s para 200 bandas de resolução do histograma.

Na FIG. 5.4, que representa os testes de LTF para o caso FSS e formantes *pitch*-

TAB.5.11: Teste FS2PS — Taxas de FR, FA e MME com variação do tempo de treinamento — resolução de 200 bandas.

Tempo	**FR(%)**	**FA(%)**	**MME(%)**
3 s	23,16	28,64	25,60
10 s	10,21	18,07	14,14
30 s	**8,17**	**10,11**	**9,14**

TAB.5.12: Teste FS2PS — Taxas de FR, FA e MME com variação da resolução do histograma — teste de 30 s.

Resolução	**FR(%)**	**FA(%)**	**MME(%)**
200 bandas	8,17	10,11	9,14
100 bandas	**7,77**	**9,68**	**8,72**
50 bandas	**7,77**	9,83	8,80
20 bandas	8,76	10,43	9,60
10 bandas	8,76	13,47	11,12

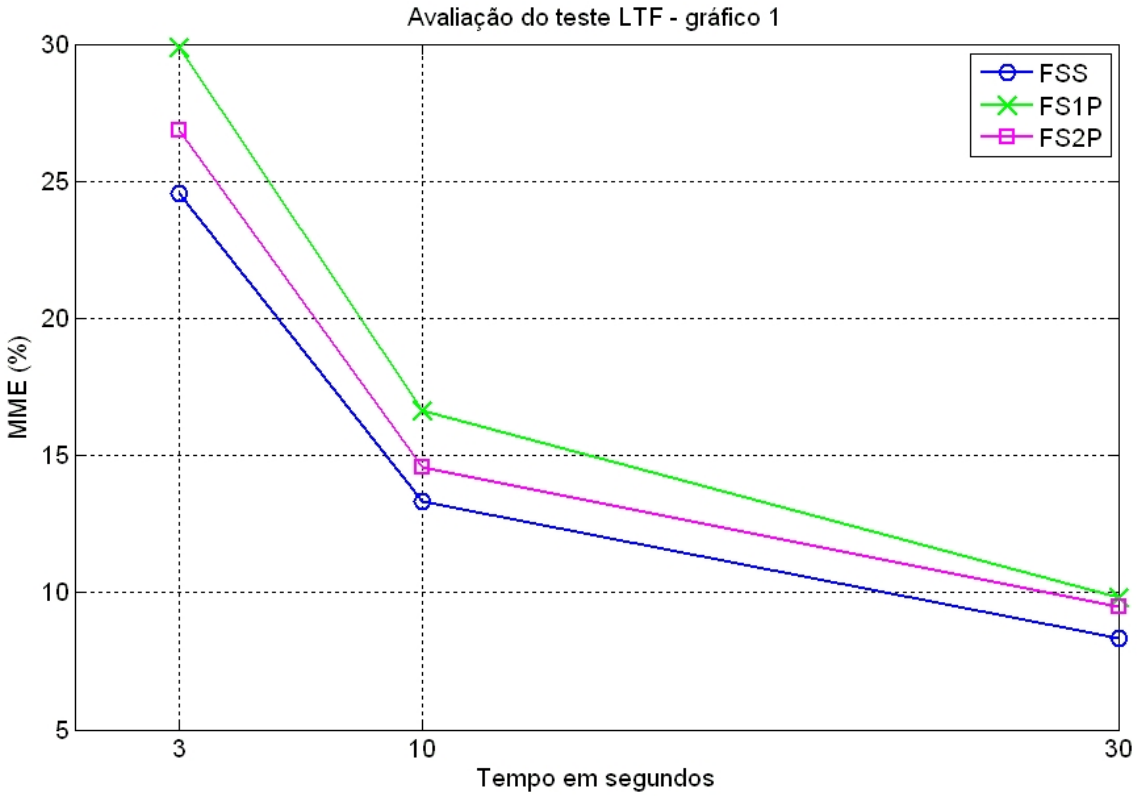


FIG.5.3: Gráfico do MME em função do tempo do teste de LTF para os casos indicados na legenda — resolução de 200 bandas — vide tabelas TAB. 5.3,TAB. 5.5 e TAB. 5.9.

síncronos extraídos apenas dos trechos sonoros, percebe-se claramente não haver um padrão de MME por resolução do histograma de LTF com o tempo de teste fixo em 30 s, com exceção do teste FSS, que gerou melhor resultado para as resoluções do histograma de 20, 50, 100 e 200 bandas; o teste FS2P gerou melhor resultado para 10 bandas de resolução.

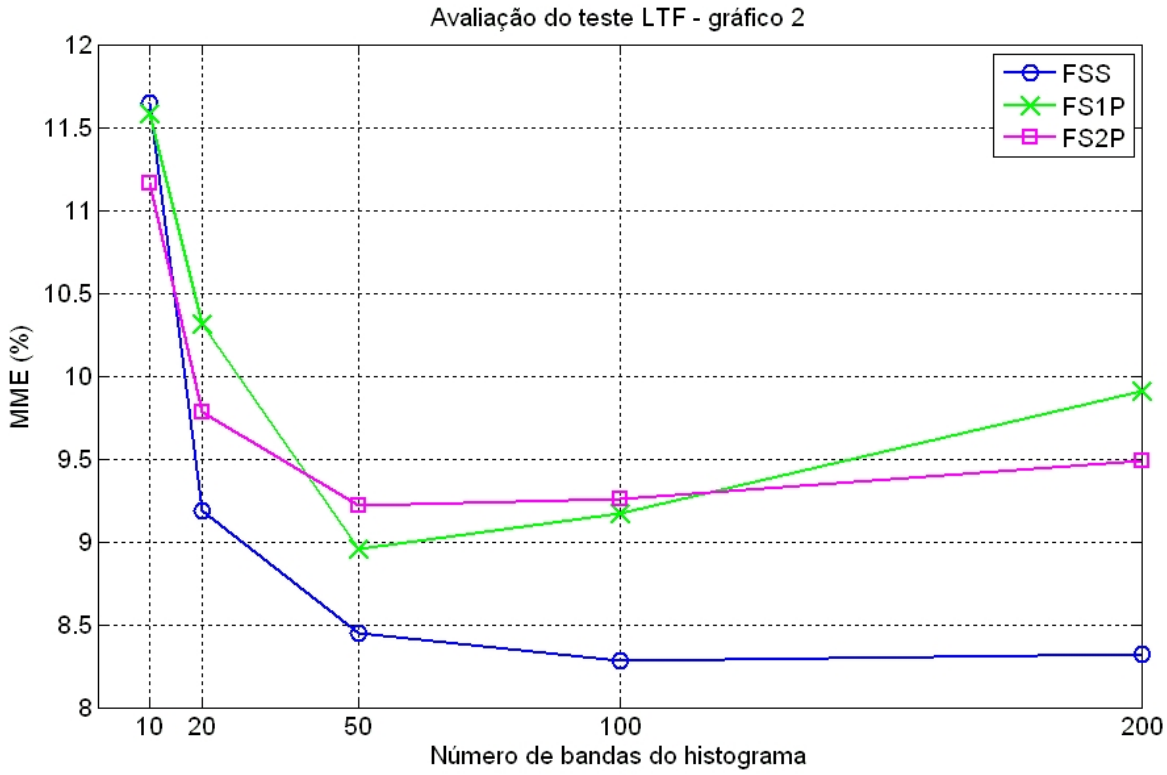


FIG.5.4: Gráfico do MME em função da resolução do histograma para os casos indicados na legenda — teste de 30 s — vide tabelas TAB. 5.4, TAB. 5.6 e TAB. 5.10.

Entretanto, ao se realizar o mesmo teste com a combinação dos formantes extraídos dos trechos surdos (sem sincronismo) aos formantes *pitch*-síncronos extraídos dos trechos sonoros, há a tendência de menor MME para o caso FS1PS (vide FIG. 5.6 e tabelas TAB. 5.8, TAB. 5.12, TAB. 5.6 TAB. 5.10 e TAB. 5.4). Comparando diretamente as tabelas TAB. 5.12 e TAB. 5.10, observa-se que os testes FS2PS apresentam também melhores resultados que os testes FS2P, confirmando o benefício da incorporação dos formantes dos fones surdos. O caso FS2PS, embora não supere o teste com extração dos formantes sem sincronismo na maior parte dos casos, foi melhor que essa para 10 bandas de resolução do histograma, da mesma forma que os testes FS2P e FS1P (registrados na FIG. 5.4) foram melhores que o teste FSS para o caso de 10 bandas. Percebe-se que os formantes

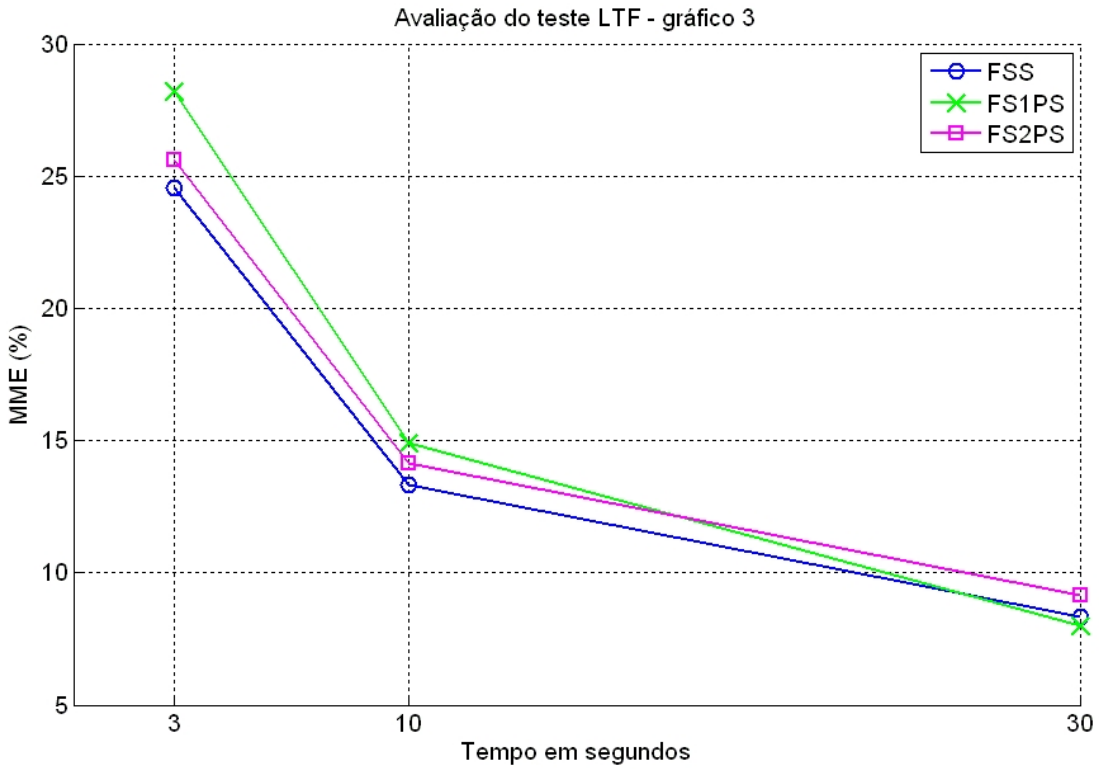


FIG.5.5: Gráfico do MME em função do tempo do teste de LTF para os casos indicados na legenda — resolução de 200 bandas — vide tabelas TAB. 5.3,TAB. 5.7 e TAB. 5.11.

pitch-síncronos com maior tempo de teste (30 s), de fato, se beneficiam de uma situação de melhor resolução espectral, como é sugerido na Subseção 2.6.2.

5.6 TESTES DAS CARACTERÍSTICAS PELO DISCRIMINANTE DE FISHER

Foram avaliados os discriminantes de Fisher das seguintes características extraídas de sinais de voz, totalizando 61 coeficientes por locutor:

- Coeficientes SFM, SCF e tonalidade (características de sinais de áudio) — 4 coeficientes de cada característica para cada locutor, totalizando 12 coeficientes por locutor;
- Coeficientes MFCC, delta e delta-delta (características perceptuais) — 15 coeficientes de cada característica para cada locutor, totalizando 45 coeficientes por locutor;
- Coeficientes SSC (características de reconhecimento de voz) — 4 coeficientes para cada locutor.

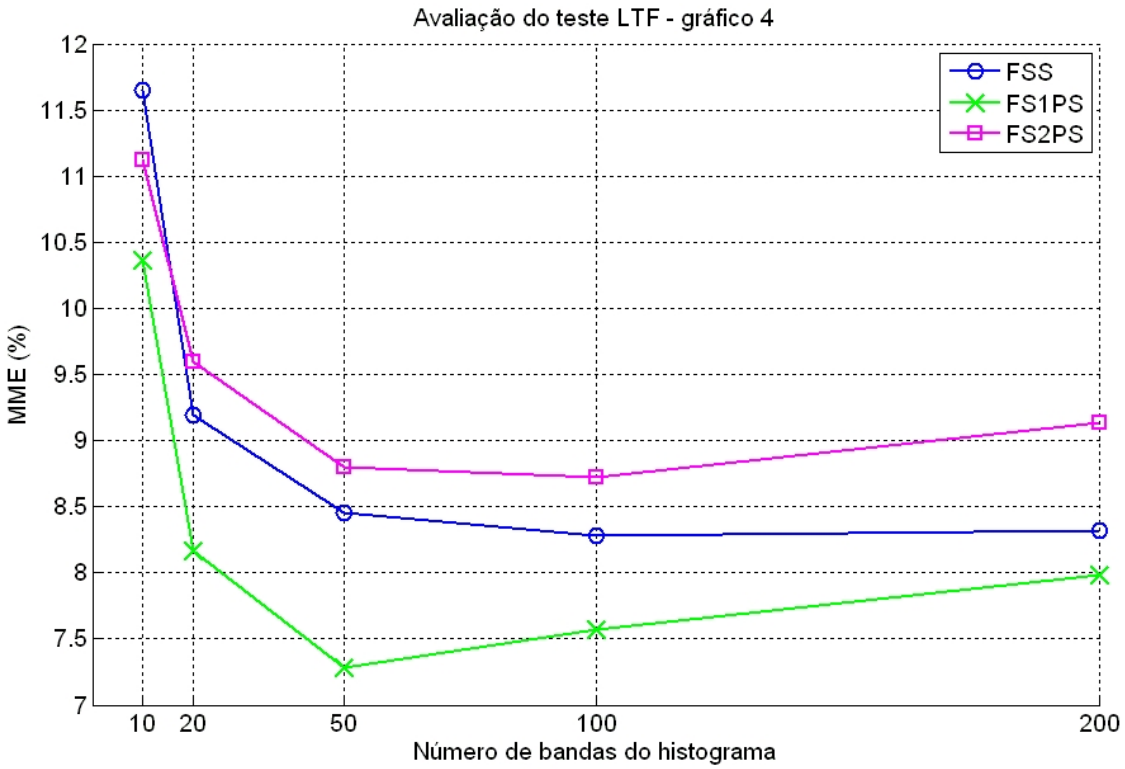


FIG.5.6: Gráfico do MME em função da resolução do histograma para os casos indicados na legenda — teste de 30 s — vide tabelas TAB. 5.4,TAB. 5.8 e TAB. 5.12. Nota-se que, ao contrário da FIG. 5.4, o teste FSS não obtém os melhores resultados, e sim o teste FS1PS.

A TAB. 5.13 foi levantada para os discriminantes de Fisher de cada conjunto de características:

TAB.5.13: Discriminantes de Fisher para as características extraídas dos sinais de voz

Característica	**Razão-F**
Coeficientes MFCC, delta e delta-delta	1,430
Coeficientes SSC	0,369
Tonalidade	0,345
Coeficientes SFM	0,328
Coeficientes SCF	0,208

Os valores do Discriminante de Fisher da TAB. 5.13 foram calculados pela EQ. 4.17 para cada vetor de características isolado (por exemplo, no caso do SSC, para o vetor completo de coeficientes SSC). As razões F foram dispostas em ordem decrescente, de modo a ressaltar as características mais relevantes em conjunto.

Para cada característica, de forma isolada (TAB. 5.14 à TAB. 5.18), foi levantada uma tabela de avaliação de VAL nos tempos de teste de 3 s, 10 s e 30 s, para modelos GMM de 8, 16 e 32 componentes gaussianas. As tabelas se encontram expostas na mesma ordem das características mostrada na TAB. 5.13.

TAB.5.14: Avaliação da VAL para os coeficientes MFCC de ordem 15, delta e delta-delta

Número de gaussianas	FR, FA e MME (%)								
	Teste de 3 s			Teste de 10 s			Teste de 30 s		
	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	6,25	7,43	6,84	3,50	4,78	4,14	1,25	4,37	2,81
16	3,50	4,15	3,82	0,25	3,50	1,87	0	3,08	1,54
32	3,25	0,56	1,91	0,25	0,44	0,35	0	0,21	**0,11**

TAB.5.15: Avaliação da VAL para os coeficientes SSC

Número de gaussianas	FR, FA e MME (%)								
	Teste de 3 s			Teste de 10 s			Teste de 30 s		
	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	7,25	14,62	10,93	4,50	11,15	7,83	1,50	10,81	6,16
16	5,00	13,73	9,36	2,25	9,44	5,85	0	9,07	4,53
32	5,25	12,14	8,69	2,50	7,92	5,21	0,50	6,94	**3,72**

TAB.5.16: Avaliação da VAL para os coeficientes de tonalidade

Número de gaussianas	FR, FA e MME (%)								
	Teste de 3 s			Teste de 10 s			Teste de 30 s		
	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	17,75	38,28	28,02	15,25	41,21	28,23	29,82	23,61	**26,71**
16	26,75	31,42	29,08	38,00	19,49	28,74	35,59	20,66	28,12
32	32,00	27,49	29,75	40,75	19,25	30,00	36,84	21,75	29,30

TAB.5.17: Avaliação da VAL para os coeficientes SFM

Número de gaussianas	FR, FA e MME (%)								
	Teste de 3 s			Teste de 10 s			Teste de 30 s		
	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	8,75	20,23	14,49	11,25	11,13	11,19	8,02	10,32	**9,17**
16	10,50	17,97	14,24	9,75	12,33	11,04	8,52	9,91	9,22
32	9,50	18,24	13,87	8,75	13,30	11,02	8,02	10,33	**9,17**

TAB.5.18: Avaliação da VAL para os coeficientes SCF

Número de gaussianas	FR, FA e MME (%)								
	Teste de 3 s			Teste de 10 s			Teste de 30 s		
	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	10,25	24,92	17,58	7,50	14,18	10,84	4,01	14,44	9,22
16	17,50	16,38	16,94	4,75	15,38	10,07	3,76	13,19	8,48
32	15,25	17,50	16,37	10,00	8,70	9,35	8,02	7,72	**7,87**

Para consolidar os resultados, também foi avaliada na TAB. 5.19 a VAL para o vetor completo de características SCF, SFM, tonalidade, MFCC, delta, delta-delta e SSC.

TAB.5.19: Avaliação da VAL para todos os 61 coeficientes

Número de gaussianas	FR, FA e MME (%)								
	Teste de 3 s			Teste de 10 s			Teste de 30 s		
	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	3,25	8,42	5,84	2,00	4,93	3,46	0,75	5,08	2,92
16	2,25	4,68	3,46	0	3,61	1,81	0	3,18	1,59
32	3,75	0,33	2,04	0,75	0,20	0,48	0,25	0,24	0,25

Analisando os resultados das avaliações de VAL da TAB. 5.14 à TAB. 5.18, observa-se que as características MFCC (associadas aos coeficientes delta e delta-delta) e SSC lograram maior êxito nos testes, fato coincidente com o resultado da TAB. 5.13, na qual se pode perceber que as características MFCC e SSC obtiveram, em conjunto, os maiores valores de Discriminante de Fisher.

Analisando o resultado da avaliação de VAL para o vetor completo de características na TAB. 5.19, nota-se que os valores de erro obtidos são muito próximos aos valores de

erro encontrados para os coeficientes MFCC, delta e delta-delta na TAB. 5.14. Este fato comprova a grande contribuição dos coeficientes MFCC, delta e delta-delta no conjunto de características. Por este motivo, os testes doravante apresentados serão realizados apenas com os coeficientes MFCC, delta e delta-delta. **Ressalta-se que, comparando as tabelas TAB. 5.19 e TAB. 5.14 para 30 s de teste, os coeficientes MFCC, delta e delta-delta produziram erros MME menores, para 8, 16 e 32 gaussianas, que os testes realizados nas mesmas condições para o vetor completo de 61 coeficientes.**

5.7 TESTES DE VAL ENGLOBANDO OS ALGORITMOS EM, EM-RP E EM-GA PARA ESTIMAÇÃO DOS PARÂMETROS DOS MODELOS GMM DAS CARACTERÍSTICAS

Seguindo a metodologia proposta na Seção 5.3, é realizada uma avaliação de VAL usando os parâmetros dos modelos GMM da base IME2001 estimados pelas projeções aleatórias (EM-RP) e pelo algoritmo proposto em (FLORES) (EM-GA). As características selecionadas para o teste foram os quinze coeficientes MFCC, acrescidos dos 15 coeficientes delta e dos 15 delta-delta, por ter sido o conjunto de coeficientes com melhor discriminante de Fisher. O critério de avaliação se subdividiu em três etapas independentes:

- **Etapa 1**: treinamento dos modelos GMM pelo algoritmo EM simples. O resultado desta etapa consta da TAB. 5.14;
- **Etapa 2**: treinamento dos modelos GMM pelo algoritmo EM-RP (por projeção aleatória). Para cada locutor, foi escolhido um modelo GMM tal que a verossimilhança de cada modelo estimado fosse maior que a verossimilhança estimada pelo modelo EM simples para um conjunto de $(7 + 5k)$ projeções, $k > 1$;
- **Etapa 3**: treinamento dos modelos GMM pelo algoritmo EM-GA (por algoritmo genético). O treinamento do algoritmo genético transcorreu em duas fases — por um número de gerações mínimo tal que as verossimilhanças dos modelos GMM estimados pelo algoritmo EM-GA fossem maiores que as verossimilhanças dos modelos estimados pelo algoritmo EM simples (etapa 3a), e por 50 gerações (etapa 3b). Em ambas as etapas, o algoritmo genético foi implementado com população inicial de 7 matrizes de projeção à dimensão $d = 2$, seleção dos 2 melhores indivíduos (equivalente ao utilizado na Etapa 2), fator de mutação de 0,2 (ou seja, dos 5 elementos

restantes, pelo menos um deles sofrerá mutação e os restantes, cruzamento), fator de redução de variância de 0,75, seleção estocástica uniforme;

Os resultados das três etapas de teste constam da TAB. 5.20. O tempo de teste considerado para a avaliação comparativa de desempenho foi de 3 s. O motivo do emprego do menor tempo de teste foi justamente verificar, para um pior caso, a queda dos erros de FA, de FR e total para as técnicas de projeção e computação evolucionária. A avaliação foi realizada para modelos GMM com ordem K de 8, 16 e 32 gaussianas.

TAB.5.20: Avaliação comparativa de desempenho dos algoritmos EM, EM-RP e EM-GA para os coeficientes MFCC, delta e delta-delta

K	FR, FA e MME (%)											
	EM (etapa 1)			EM-RP (etapa 2)			EM-GA (etapa 3a)			EM-GA (etapa 3b)		
	FR	FA	MME	FR	FA	MME	FR	FA	MME	FR	FA	MME
8	6,25	7,43	6,84	5,50	6,94	6,22	5,75	7,02	6,38	5,25	6,54	5,90
16	3,50	4,15	3,82	3,00	4,65	3,82	3,25	4,01	3,63	5,50	0,59	3,04
32	3,25	0,56	1,91	2,00	0,72	1,36	1,50	1,06	1,28	1,50	0,62	1,06

A TAB. 5.20 mostra que o algoritmo genético tende a melhores resultados de verificação. Comparando os erros obtidos das três variantes do algoritmo EM, percebe-se a melhora de desempenho[20] (redução percentual do MME) do algoritmo EM-RP (etapa 2) de 9,06% para 8 gaussianas e 28,80% para 32 gaussianas, no que diz respeito ao MME, comparado ao algoritmo EM simples (etapa 1 — mesmos resultados de testes de 3 s da TAB. 5.14); para 16 gaussianas, não há melhora no MME da etapa 2 em comparação com a etapa 1. O algoritmo EM-GA (etapa 3b) produz melhora de 13,74% para 8 gaussianas, 20,42% para 16 gaussianas e 44,50% para 32 gaussianas, quando comparado com o EM simples. Outro resultado expressivo do EM-GA (etapa 3b) sobre o EM-RP e o EM simples diz respeito ao erro de falsa aceitação — o algoritmo EM-RP produz aumento no erro FA para 16 e 32 gaussianas e a melhora com 8 gaussianas é pequena (6,59%), ao passo que o EM-GA (etapa 3b) produz um aumento no erro de FA quase inexpressivo e melhora bem mais a condição de falsa aceitação do que o EM-RP nos casos de 8 e 16 gaussianas, nos quais obtém, respectivamente, 11,98% e 85,78% de aprimoramento.

Confirmando o resultado explanado na TAB. 5.20, as figuras FIG. 5.7, FIG. 5.8 e FIG. 5.9 mostram o ganho de verossimilhança dos algoritmos EM-RP na etapa 2 (linhas com marcadores em "X") e EM-GA na etapa 3b (linhas com marcadores em círculo)

[20]Redução percentual dos erros de uma técnica para outra.

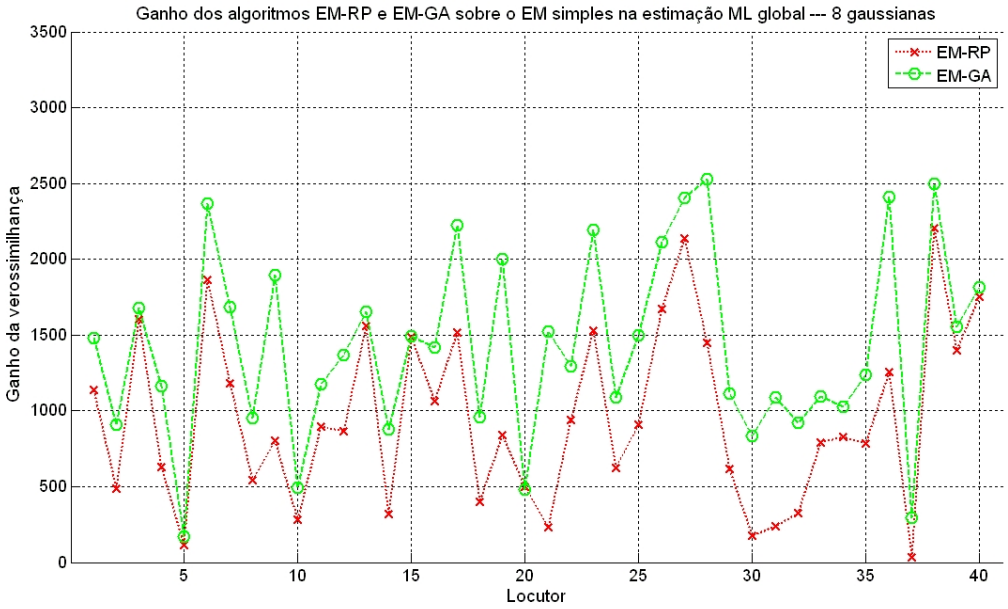


FIG.5.7: Ganho de verossimilhança dos modelos GMM estimados pelos algoritmos EM-GA ($\mathcal{L}_{i,EM-GA} - \mathcal{L}_{i,EM}$) e EM-RP ($\mathcal{L}_{i,EM-RP} - \mathcal{L}_{i,EM}$) indicados pela legenda (8 gaussianas) para cada locutor i, sendo $i = 1, \ldots, 40$.

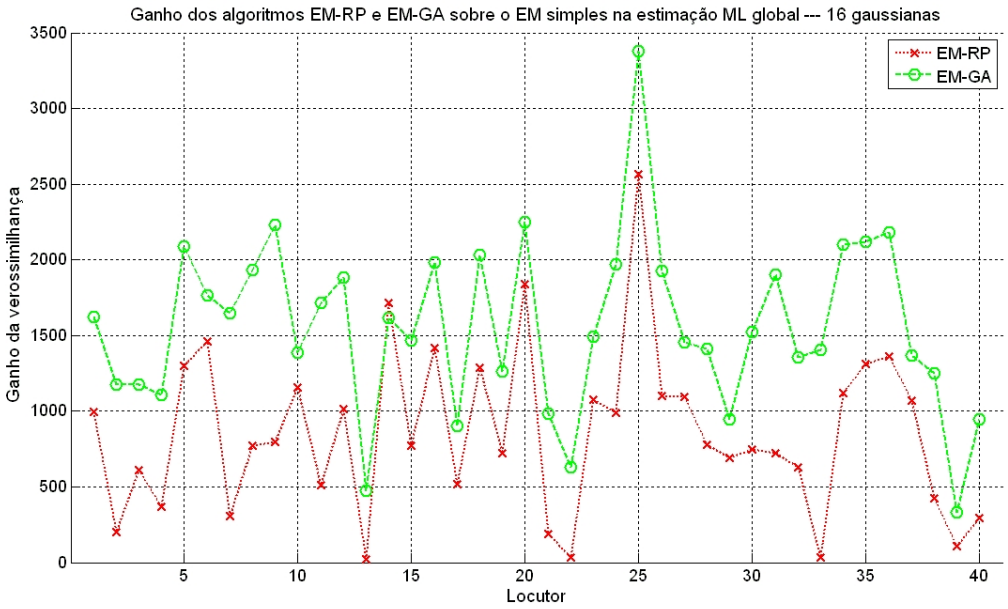


FIG.5.8: Ganho de verossimilhança dos modelos GMM estimados pelos algoritmos EM-GA ($\mathcal{L}_{i,EM-GA} - \mathcal{L}_{i,EM}$) e EM-RP ($\mathcal{L}_{i,EM-RP} - \mathcal{L}_{i,EM}$) indicados pela legenda (16 gaussianas) para cada locutor i, sendo $i = 1, \ldots, 40$.

sobre o algoritmo EM simples na determinação dos modelos GMM de cada locutor. Esse ganho, em termos numéricos, é expresso pela diferença entre as verossimilhanças do

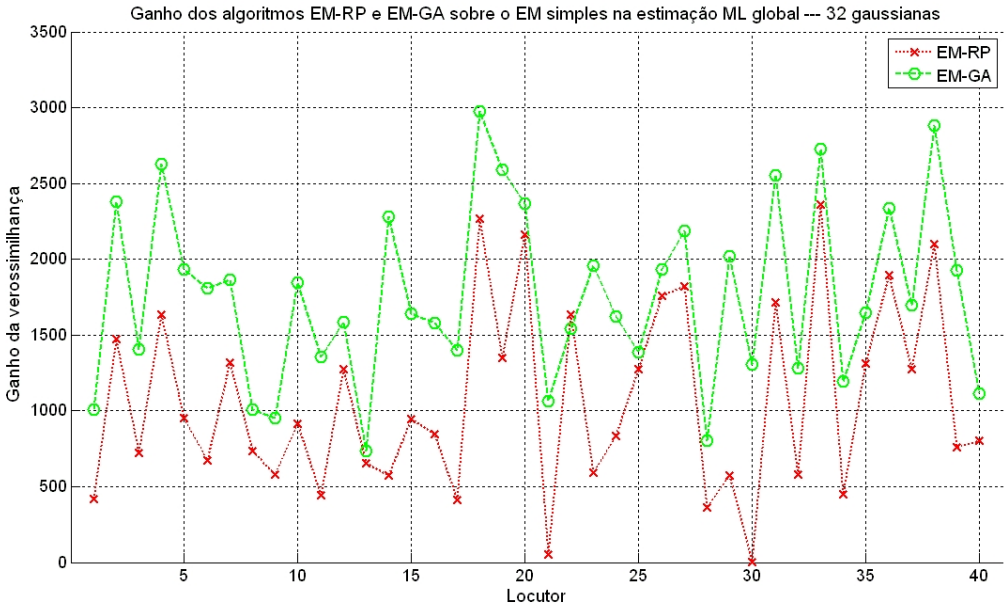


FIG.5.9: Ganho de verossimilhança dos modelos GMM estimados pelos algoritmos EM-GA ($\mathcal{L}_{i,EM-GA} - \mathcal{L}_{i,EM}$) e EM-RP ($\mathcal{L}_{i,EM-RP} - \mathcal{L}_{i,EM}$) indicados pela legenda (32 gaussianas) para cada locutor i, sendo $i = 1, \ldots, 40$.

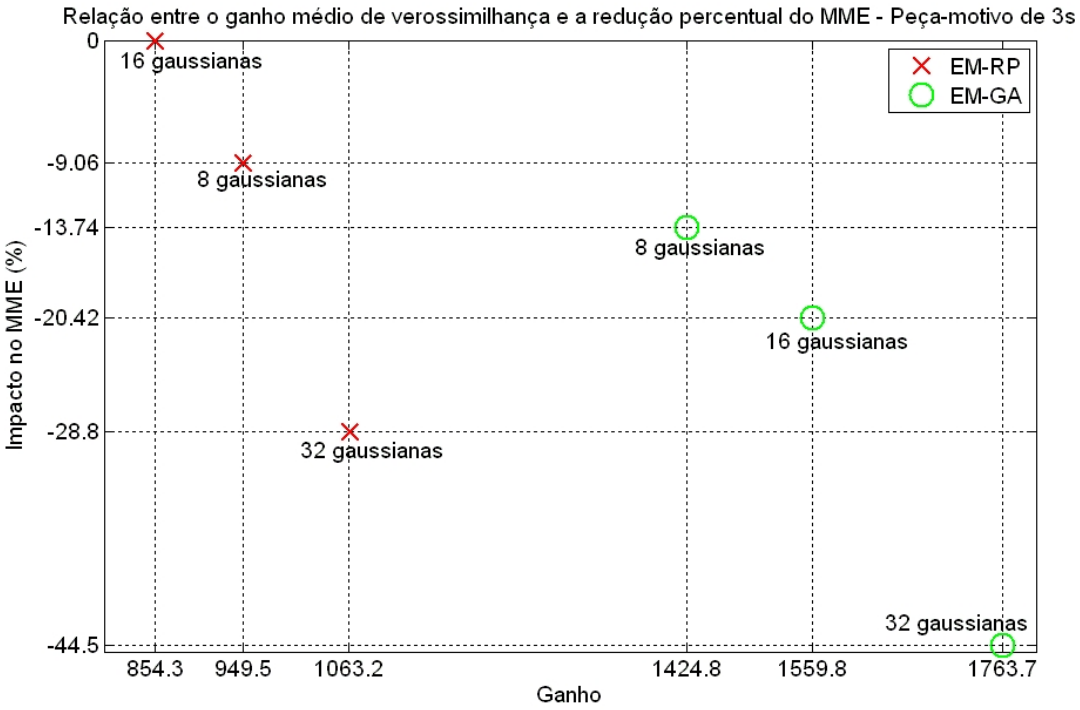


FIG.5.10: Impacto no MME dos modelos GMM estimados pelos algoritmos indicados pela legenda *versus* ganho médio de verossimilhança por locutor (testes de 3 s).

modelo obtido pelo algoritmo indicado pela legenda e a verossimilhança do modelo EM simples. Pode ser percebida a tendência do algoritmo EM-GA a produzir modelos com verossimilhanças maiores, ressaltando na prática a tendência dos algoritmos genéticos à estimação ML global, tal como abordado na Subseção 4.2.5. Além disso, comprova-se que os modelos que tenderam à estimação ML de forma mais eficiente conseguiram, na prática, melhores resultados de VAL. Comprova-se também, pela FIG. 5.10, que o ganho médio de verossimilhanças por locutores, para cada ordem de modelo de gaussianas, está associado de forma monotônica à redução do MME.

Também foram efetuados os testes de VAL para 10 s e 30 s de peça-motivo. A FIG. 5.11 combina os testes de 3 s e 10 s, e a FIG. 5.12 ilustra o teste de VAL com 30 s. Nota-se que para 10 s, os modelos treinados pelo algoritmo EM-GA implicam em melhores resultados. Com o aumento do tempo de teste para 30 s, o algoritmo EM-GA passa a não exibir mais os melhores resultados para ordens de modelos GMM intermediárias (como é possível perceber na FIG. 5.12), abrindo margem ao emprego da Medida BIC na estimação da ordem dos modelos na tentativa de compensar esse problema.

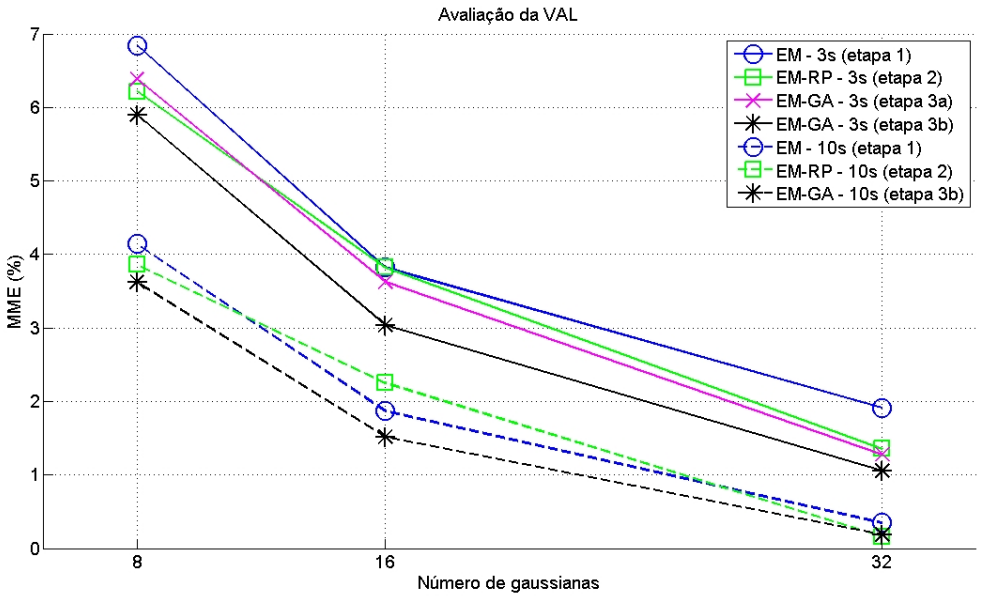


FIG.5.11: Teste de VAL para 3 s e 10 s de peça-motivo.

Com a implementação da Medida BIC até 15 gerações do algoritmo EM-GA, respeitando as mesmas configurações do teste EM-GA dos outros parâmetros genéticos supracitados, obteve-se o resultado da FIG. 5.13. Percebe-se que os modelos treinados pelo algoritmo EM-GA, embora não tenham produzido o melhor resultado, propiciaram resultados melhores que os modelos treinados pelo algoritmo EM simples, destacadamente

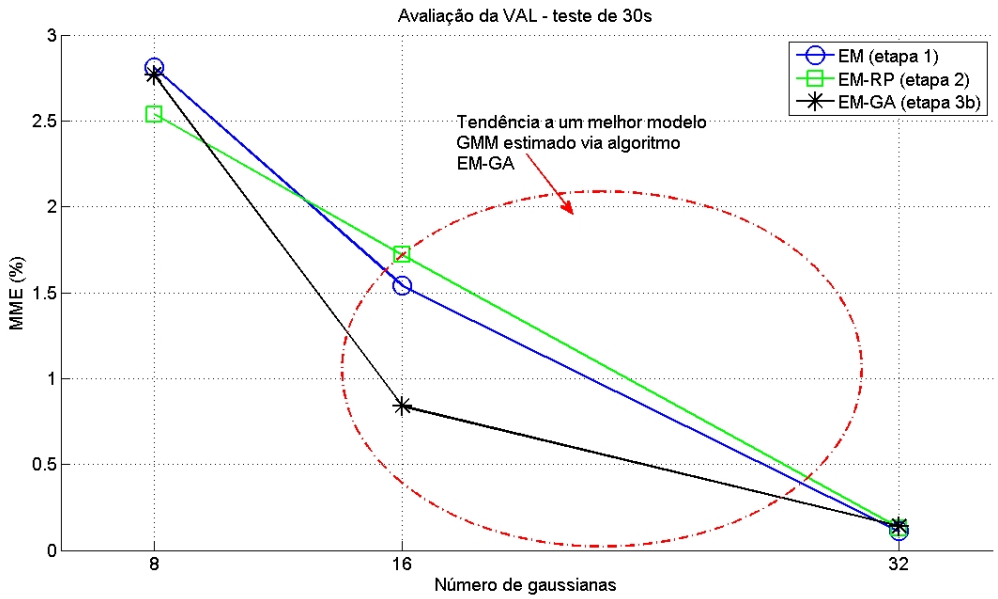


FIG.5.12: Teste de VAL para 30 s de peça-motivo.

para 3 s de teste.

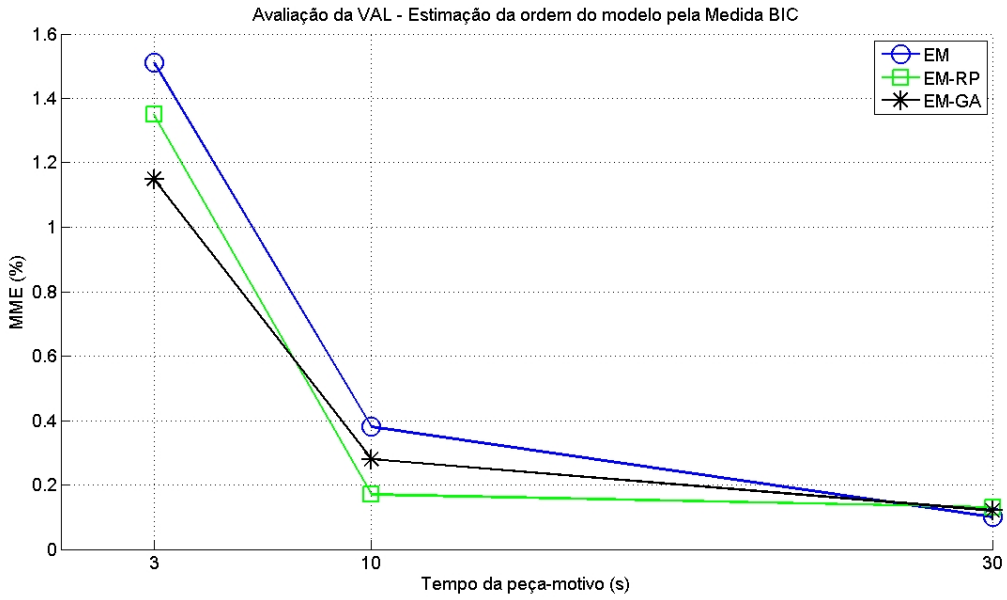


FIG.5.13: Teste de VAL para 3 s, 10 s e 30 s de peça-motivo. O treinamento dos modelos GMM foi efetuado pela Medida BIC.

5.8 RESUMO E CONCLUSÃO

Neste capítulo foi apresentada a contribuição prática à VAL para fins forenses. Foi mostrada na Seção 5.2 uma abordagem sintética de um sistema de verificação automática

de locutor, com ênfase em testes de hipóteses e na montagem dos modelos de locutores conhecidos e falsos locutores (*background*).

Quanto aos testes de VAL, foram constatados os seguintes fatos:

- O bom desempenho dos experimentos de VAL, tanto para os testes de LTF quanto para os testes de modelos GMM, depende fortemente do tempo de duração das falas de teste. Foi novamente comprovado, a exemplo de (LIMA, 2001), que os resultados sempre apresentam melhora quando os tempos das falas de teste são aumentadas de 3 s para 10 s e 30 s;

- No caso dos testes de LTF, os formantes extraídos de forma *pitch*-síncrona produzem melhores resultados de VAL quando analisados por histogramas de maior resolução. Entretanto, pode-se notar que, mesmo nos casos com maior escassez de vetores de formantes (a 3 e 4 períodos com sobreposição de um período de *pitch*), os MME não foram muito maiores que os MME dos casos com maior número de vetores de formantes (1 e 2 períodos de *pitch*). Além disso, foi percebido que a combinação dos formantes extraídos dos trechos surdos e sonoros aprimorou a VAL dos testes em comparação aos testes realizados com os formantes extraídos apenas dos trechos sonoros;

- As características que, em conjunto, exibiram um maior coeficiente de Fisher, também produziram melhor desempenho na VAL. A citar, os coeficientes SSC e os coeficientes MFCC acrescidos dos coeficientes delta e delta-delta. A proximidade numérica dos resultados decorrentes da avaliação da VAL com o vetor completo de características (SFM, SCF, tonalidade, MFCC, delta, delta-delta e SSC) aos resultados do MFCC acrescido dos coeficientes delta e delta-delta permitiu perceber que os coeficientes MFCC, delta e delta-delta são preponderantes nos bons resultados de VAL, sobretudo nos testes de 30 s e nos testes com 32 gaussianas (testes de maior duração e maior ordem do modelo GMM);

- Uma vez escolhidos os coeficientes MFCC, delta e delta-delta para gerar os testes de VAL com estimação dos parâmetros GMM via algoritmos EM simples, EM-RP e EM-GA, pode ser percebida a tendência do algoritmo genético de buscar a melhor configuração de modelos GMM de forma a produzir menores erros de verificação, bem como foi notada a associação entre a busca pela estimação ML global (para locutores isolados e na média entre locutores) e a busca pela maior redução percentual do MME;

- A estimação de ordem do modelo pela Medida BIC e pelo algoritmo EM-GA para tempos maiores de peça-motivo (10 s e 30 s de duração) necessita de um melhor ajuste em trabalhos futuros.

6 CONCLUSÃO E SUGESTÕES PARA TRABALHOS FUTUROS

Este trabalho apresentou uma breve teoria sobre o mecanismo de produção da fala, as características extraídas dos sinais de voz, a extração de características com sincronismo por períodos da *pitch* e o estado da arte da perícia em fonética forense praticada no Brasil. Foram introduzidas técnicas inovadoras, tais como a verificação automática de locutor por LTF — observa-se que o trabalho descrito em (NOLAN, 2005) não fornece taxas de VAL para uma base de dados de voz (*corpus* — e a sugestão dos algoritmos EM-RP (baseado em projeções aleatórias) e EM-GA (baseado em algoritmos genéticos) em contribuição à VAL em ambiente forense. Foram explicitadas as taxas de VAL obtidas pelos testes de LTF (em vários contextos, sem sincronismo e com sincronismo, em função dos períodos da *pitch*) e pela análise das demais características, com o suporte do Discriminante de Fisher e da computação evolucionária.

Outra contribuição deste trabalho foi a inserção de características de reconhecimento de áudio — SCF, SFM e tonalidade — e reconhecimento de voz (SSC) na avaliação da VAL.

Convém ressaltar que, embora as técnicas apresentadas busquem o aprimoramento das técnicas de verificação automática, tal procedimento, na esfera forense, consiste numa contribuição a um sistema de apoio à decisão, uma vez que a perícia em âmbito forense é desempenhada de forma semi-automática, ou seja, os resultados obtidos dão suporte à decisão final da autoridade competente.

Quanto aos testes realizados por LTF, mostrou-se que o emprego do sincronismo pela *pitch* produz melhores resultados, desde que sejam combinados os formantes dos trechos surdos e sonoros produz melhores resultados do que somente os formantes dos trechos sonoros. O desempenho dos testes se mostrou dependente do número de bandas do histograma: os melhores resultados se concentraram em valores intermediários de número de bandas.

Quanto aos testes envolvendo GMM, percebeu-se que os coeficientes MFCC, delta e delta-delta, beneficiados dos aspectos perceptuais dos sinais de voz, obtiveram os melhores resultados de VAL, chegando a valores de erro bem próximos dos testes que abrangeram todas as características, sendo os MME menores para os testes de 30s (8, 16 e 32 gaussianas), 3s a 32 gaussianas e 10s a 32 gaussianas. Pela análise do Discriminante de Fisher,

pôde ser percebido que os dois conjuntos de características que obtiveram maiores valores — MFCC (acrescidos dos coeficientes delta e delta-delta) e SSC — produziram os melhores desempenhos de VAL para 3s, 10s e 30s de teste. Ressalta-se o bom desempenho das características de reconhecimento de voz (SSC) e áudio (SCF e SFM) para VAL, obtendo MME abaixo de 10% para 30s de teste.

Comparando os algoritmos EM simples, EM-RP e EM-GA, pôde ser comprovado que os modelos GMM produzem melhor desempenho de VAL quando treinados por projeções aleatórias, e ainda melhor desempenho quando treinado por algoritmos genéticos. Em outras palavras, quanto maior o aspecto evolucionário (que implica em busca pela estimação ML global dos parâmetros dos modelos GMM), menores serão os MME produzidos. Em números, para testes de 3s, o algoritmo EM-RP produziu 9% e 29% de ganho de desempenho (redução percentual do MME) para 8 e 32 gaussianas, respectivamente, sem ganho no treinamento realizado por 16 gaussianas, ao passo que o algoritmo EM-GA exibiu ganhos de desempenho de 14%, 20% e 45% para 8, 16 e 32 gaussianas, respectivamente. Mostrou-se também uma associação entre o ganho de verossimilhança médio dos locutores e a redução do MME.

6.1 LIMITAÇÕES DO TRABALHO

A extração dos formantes foi realizada de forma automática com o emprego do aplicativo *Praat*, mencionado na Seção 3.3.4. Observou-se, em diversas instâncias, que os valores dos formantes extraídos apresentavam problemas de estimação, demonstrando ser fortemente dependentes de uma estimação do número de formantes mais adequada. Outros aplicativos como o *SFS* e o *Wavesurfer* apresentaram o mesmo problema. Outros fatores também influenciaram na dificuldade de extração de formantes de forma automática, tais como a falta de praticidade no uso de *scripts* observada no manuseio do aplicativo *Wavesurfer* (usa linguagem de programação pouco difundida — *Tcl/Tk*) e até mesmo a falta de aquisição automática, observada no aplicativo *WinPitchPro*, que depende da análise semi-automática (observação dos valores em tela) pelo perito.

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

Ficam como sugestões para trabalhos futuros:

- Verificar uma melhor forma de emprego das características SSC, SCF, SFM e tonalidade;

- Verificar o desempenho das técnicas propostas EM-RP e EM-GA com maior tempo de duração de sinal de teste para verificar se ocorrem menores taxas de erro;

- Melhorar o uso do algoritmo EM-GA com o emprego de vetores de parâmetros de matrizes ortogonais (ângulos de uma matriz de rotação de Givens, por exemplo);

- Realizar a análise comparativa da VAL pelo emprego da estimação de ordem dos modelos GMM pela medida do Critério de Informação Bayesiana (BIC, de *Bayesian Information Criterion*) para os algoritmos EM simples, EM-RP e EM-GA, permitindo a seleção da ordem dos modelos de forma automática em vez do emprego dos modelos fixos de 8, 16 e 32 gaussianas;

- A fusão de classificadores: combinar os *scores* provenientes dos testes de VAL empregando características diferentes (por exemplo, MFCC e SSC) para melhorar as taxas de teste;

- Efetuar um melhor ajuste dos parâmetros da técnica LTF;

- Realizar a avaliação comparativa da VAL por modelos gerados pelo critérios da Máxima Informação Mútua (MMI, de *Maximum Mutual Information*), da Análise de Componentes Principais (PCA, de *Principal Component Analysis*) e da Análise de Discriminantes Lineares (LDA, de *Linear Discriminant Analysis*), em busca de uma melhor discriminação entre características de locutores diferentes;

- Verificar o efeito das técnicas propostas em ambientes com ruído aditivo e canal (fixo e móvel/celular);

- Aumentar o tempo do sinal de treinamento para o caso de características com sincronismo de *pitch*;

- Estudar outras características usadas em VAL (por exemplo, o VOT) em aplicações de fonética forense;

- Propor uma nova metodologia de perícia em fonética forense que conjugue a experiência do perito com os resultados de um sistema de VAL.

7 REFERÊNCIAS BIBLIOGRÁFICAS

AJMERA, J., MCCOWAN, I. e BOURLARD, H. Robust speaker change detection. **IEEE Signal Processing Letters**, v. 11, p. 649-651, ago. 2004.

ALLEN, J. Cochlear modeling. **IEEE Acoustics, Speech and Signal Processing Magazine**, v. 2, n. 1, p. 3-29, jan. 1985.

BALDWIN, J. e FRENCH, P. **Forensic Phonetics**. Londres e Nova Iorque: Pinter, 1990.

BESACIER, L., MAYORGA, P., BONASTRE, J. F., FREDOUILLE, C. e MEIGNIER, S. Overview of compression and packet loss effects in speech biometrics. **IEEE Proceedings on Visual, Image and Signal Processing**, v. 150, n. 6, p. 372-376, 2003.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Springer, 2006.

BOERSMA, P. Praat, a system for doing phonetics by computer. **Glot International**, v. 5, n. 9, p. 341-345, 2001.

BRAID, A. C. M. **Fonética Forense**. 2. ed. Campinas: Millenium, 2003.

BROEDERS, A. P. A. Some observations on the use of probability scales in forensic identification. **Forensic Linguistics**, v. 6, n. 2, p. 228-241, 1999.

DASGUPTA, S. Experiments with random projection. In: UAI-2000: THE SIXTEENTH CONFERENCE IN ARTIFICIAL INTELLIGENCE. jun. 2000, Stanford, CA, EUA. **Anais...** San Francisco, EUA: Morgan Kaufmann Publishers Inc., 2000. p. 143-151.

DAVIS, S. B. e MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech and Signal Processing**, v. 28, n. 4. p. 357-366, ago. 1980.

DELLER, J. R., HANSEN, J. H. L. e PROAKIS, J. G. **Discrete-Time Processing of Speech Signals**. Nova Iorque:Wiley-Interscience-IEEE, 2000.

DUNN, R. **Speech Signal Processing and Speech Recognition**. IEEE Signal Processing Society, abr. 2003. (Technical Report RBD-5/13/2003).

EZZAIDI, H. e ROUAT, J. Pitch and MFCC dependent GMM models for speaker identification systems. In: CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING. 17., 2004, Niagara Falls, Canadá. **Anais...** Institute of Electrical and Electronic Engineers, Inc. v. 1, p. 43-46, maio 2004.

FANT, C. G. M. Acoustic description and classification of phonetic units. **Ericsson Technics**, n. 1, 1959.

FLANAGAN, J. L., SCHROEDER, M. R., ATAL, B. S., CROCHIERE, R. E., JAYANT, N. S. e TRIBOLET, J. S. Speech coding. **IEEE Transactions in Communications**, v. 27, n. 4, p. 710-737, abr. 1979.

FLORES, J. F. V. C., PINTO, E. L., GALDINO, J. F. e SILVA, D. G. Computação evolucionária aplicada à estimação de parâmetros de modelos GMM. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES. 25., set. 2007, Recife. **Anais eletrônicos...** Recife: Sociedade Brasileira de Telecomunicações, 2007. 1 DVD.

HALBE, Z. Model-based mixture discriminant analysis — an experimental study. **Pattern Recognition**, v. 38, n. 3, p. 437-440, 2005.

HERRE, J., ALLAMANCHE, E. e HELLMUTH, O. Robust matching of audio signals using spectral flatness features. In: 2001 IEEE WORKSHOP ON APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS. out. 2001, New Platz, NY, EUA. **Anais...** IEEE Signal Processing Society, 2001. p. 127-130.

KIM, S., ERIKSSON, T., KANG, H. G. e YOUN, D. H. A pitch synchronous feature extraction method for speech recognition. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. 2002, Orlando, FL, EUA. **Anais...** IEEE Signal Processing Society, 2002. v.4, p. 4076-4079.

KÜNZEL, H. J. Beware of the "telephone effect": the influence of telephone transmission on the measurement of formant frequencies. **Forensic Linguistics**, v. 8, n. 1, p. 80-99, 2001.

LEE, K. e PARK, K. A new pitch synchronous V/U/M/N/S classification algorithm. In: THE 1998 IEEE ASIA-PACIFIC CONFERENCE ON CIRCUITS AND SYSTEMS. nov. 1998, Chiangmai, Tailândia. **Anais...** The Institute of Electrical and Electronics Engineers, Inc., 1998. p. 315-318.

LIMA, C. B. **Sistemas de Verificação de Locutor Independente do Texto Baseados em GMM e AR-Vetorial Utilizando PCA**. Dissertação (mestrado), Programa de Pós-Graduação em Engenharia Elétrica, Instituto Militar de Engenharia, Rio de Janeiro, RJ, 2001.

LIN, L. e WANG, S. Genetic algorithms and fuzzy approach to gaussian mixture model for speaker recognition. In: IEEE INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND KNOWLEDGE ENGINEERING. out. 2005, Wuhan, China. **Anais...** Pequim, China: Publishing House BUPT, 2005. p. 142-146.

MASON, J. S. e BRAND, J. D. The role of dynamics in visual speech biometrics. In: Proceedings of the. **Speech, Language and the Law**, v. 12, n. 2, 2005.

MORGAN, D. P. e SILVERMAN, H. F. An event-synchronous signal processing system for connected speech recognition. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. abr. 1988, Nova Iorque, EUA. **Anais...** 1988. v.1, p. 299-302.

MORISSON, A. L. D. C. Verificação de locutor. **Perícia Federal**, n. 16, p. 19-23, nov. 2003.

NOLAN, F. e GRIGORAS, C. A case for formant analysis in forensic speaker identification. **Speech, Language and the Law**, v. 12, n. 2, p. 143-173, 2005.

PALIWAL, K. K. Spectral subband centroid features for speech recognition. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. maio 1998, Seattle, EUA. **Anais...** IEEE Signal Processing Society, 1998. v. 2, p. 617-620:617–620.

PEETERS, G. **A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project**. Institut de Recherche et Coordination Acoustique/Musique, abr. 2004. (Relatório Técnico versão 1.0).

POH, N. e BENGIO, S. Using chimeric users to construct fusion classifiers in biometric authentication tasks: an investigation. In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. 2006, Toulouse, França. **Anais...** IEEE Signal Processing Society, 2006. v. 5, p. 1077-1080.

PRATOR JR, C. H. e ROBINETT, B. W. **Manual of American English Pronunciation**. 3. ed. Holt, Rinehart and Winston, Inc., 1972.

PRZYBOCKI, M. A. e MARTIN, A. F. The 1999 nist speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking. In: PROCEEDINGS OF THE SIXTH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. set. 1999, Budapeste, Hungria. **Anais...** ISCA Archive, 1999, p. 2215-2218.

QUATIERI, T. F. **Discrete-Time Speech Signal Processing: Principles and Practice**. 1. ed. Upper Saddle River: Pearson Education Inc., 2002.

REYNOLDS, D. A. Speaker identification and verification using gaussian mixture speaker models. **Speech Communication**, v. 17, n. 1-2, p. 91-108, 1995a.

REYNOLDS, D. A., QUATIERI, T. F. e DUNN, R. B. Speaker verification using adapted gaussian mixture models. **Digital Signal Processing**, v. 10, n. 1-3, p. 19-41, 2000.

REYNOLDS, D. A. e ROSE, R. C. Robust text-independent speaker identification using gaussian mixture speaker models. **IEEE Transactions on Speech and Audio Processing**, v. 3, n. 1, p. 72-83, 1995b.

ROMITO, L. e GALATÀ, V. Towards a protocol in speaker recognition analysis. **Forensic Science International**, v. 146S, p. S107-S111, 2004.

ROSE, P. **Forensic Speaker Identification**. Londres e Nova Iorque: Taylor and Francis, 2002.

SILVA, T. C. **Fonética e Fonologia do Português: roteiro de estudos e guia de exercícios**. 8. ed. São Paulo: Contexto, 2005.

STOICA, P. e SELÉN, Y. Model order selection - a review of information criterion rules. **IEEE Signal Processing Magazine**, v. 21, n. 4, p. 36-47, jul. 2004.

STRANG, G. **Linear Algebra and its Applications**. 3. ed. Brooks & Cole, 1988.

TITZE, I. R. **Principles of Voice Production**. Englewood Cliffs: Prentice Hall, 1994.

TONACO, N. L. D. A. Cuidados com a gravação de material sonoro. **Perícia Federal**, n. 16, p. 24, nov. 2003.

VAN TREES, H. L. **Detection, Estimation and Modulation Theory - Part I**. John Wiley and Sons, Inc., 1968.

WET, F., WEBER, K., BOVES, L., CRANEN, B., BENGIO, S. e BOURLARD, H. Evaluation of formant-like features on an automatic vowel classification task. **Journal of the Acoustical Society of America**, v. 116, n. 3, p. 1781-1792, 2004.

XAFOPOULOS, A. **Speaker Verification (an Overview)**. Tampere International Center for Signal Processing (TICSP), Tampere University of Technology, Finlândia, ago. 2001. (Relatório Técnico).

ZENG, Y., WU, H. e GAO, R. Pitch synchronous analysis method and Fisher criterion based speaker identification. In: THIRD INTERNATIONAL CONFERENCE ON NATURAL COMPUTATION. aug. 2007, Haikou, Hainan, China. **Anais...** IEEE Computer Society, 2007. v. 2, p. 691-695.