Adelino Pinheiro Silva

Intervalo de Evidência e Pareamento *Fuzzy* Utilizando Relação Sinal-Ruído Aplicados à Comparação Forense de Locutores

Belo Horizonte

2020

Universidade Federal de Minas Gerais

Escola de Engenharia

Programa de Pós-Graduação em Engenharia Elétrica

INTERVALO DE EVIDÊNCIA E PAREAMENTO FUZZY UTILIZANDO RELAÇÃO SINAL RUÍDO APLICADOS À COMPARAÇÃO FORENSE DE LOCUTORES

Adelino Pinheiro Silva

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do Título de Doutor em Engenharia Elétrica.

Orientador: Prof. Maurílio Nunes Vieira

Belo Horizonte - MG

Abril de 2020

Silva, Adelino Pinheiro.
S586i Intervalo de evidência e pareamento fuzzy utilizando relação sinal-ruído aplicados à comparação forense de locutores [recurso eletrônico] / Adelino Pinheiro Silva. - 2020.
1 recurso online (138 f. : il., color.) : pdf.

Orientador: Maurílio Nunes Vieira.
Coorientador: Adriano Vilela Barbosa.

Tese (doutorado) Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndices: f. 113-138.

Bibliografia: f. 105-112.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Fonética forense - Teses. 3. Processamento de sinais - Teses. 4. Reconhecimento de padrões - Teses. I. Vieira, Maurílio Nunes. II. Barbosa, Adriano Vilela. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 621.3(043)

Ficha catalográfica: Biblioteca Profº Mário Werneck, Escola de Engenharia da UFMG

"Intervalo de Evidência e Pareamento *Fuzzy* Utilizando Relação Sinal Ruído Aplicados à Comparação Forense de Locutores"

Adelino Pinheiro Silva

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 16 de abril de 2020.

Por:

Maurílio Nunes Vieira

Prof. Dr. Maurílio Nunes Vieira
DELT (UFMG)

Prof. Dr. Adriano Vilela Barbosa
DELT (UFMG)

Prof. Dr. Luiz Wagner Pereira Biscainho
Depto de Engenharia Eletrônica e de Computação/Escola Politécnica (UFRJ)

Prof. Dr. Plínio Almeida Barbosa
Instituto de Estudos da Linguagem (IEL) (UNICAMP)

Prof. Dr. Frederico Rodrigues Borges da Cruz
Departamento de Estatística/ICEx (UFMG)

Prof. Dr. Hani Camille Yehia
DELT (UFMG)

Profª. Drª. Zélia Myriam Assis Peixoto
Programa de Pós-Graduação Em Engenharia Elétrica (PUC Minas)

Este trabalho é dedicado às pessoas que utilizam evidências para fazer suas escolhas.

Agradecimentos

Os agradecimentos principais são direcionados a Maurílio Nunes Vieira, professor, orientador e companheiro de várias discussões; ao professor Adriano Vilela, espírito desbravador; ao nosso coordenador Hani Yehia; e aos amigos do CEFALA (Centro de Estudos da Fala, Acústica, Linguagem e Música), com destaque para Adrielle, Arlindo, Carla, João Pedro, Leandro e Leonardo; foram muitas mãos que contribuíram com este trabalho.

Agradecimentos especiais à minha família, Fabiana, Fernanda e Álvaro, pela inspiração, paciência e cuidado. É claro, aos meus pais, minha irmã e minha sogra por todo suporte. E também aos meus amigos Acauã, Fausto, Guilherme, João Carlos, Leonardo, Rafael, Reinaldo, Tiago e Vladmir pelas discussões em vários campos do conhecimento.

Gostaria de agradecer a meus colegas de trabalho, peritos criminais, que de forma direta e indireta fizeram sua parte para que este trabalho se consolidasse. Primeiramente aos diretores do Instituto de Criminalística, Marco Paiva, Sérgio Belas, Dário Lopes e Carla Rogéria, por gentilmente cederem os registros de áudio para realização dos experimentos. Aos amigos do Setor de Perícias em Áudio e Vídeo, Anamari, Bruno, Harley, Geovane, José Roberto, Júlia, Marcos, Marcelo, Nicola, Suelen, Thabita e Vanessa, pela pressão de um resultado prático. Não posso deixar de agradecer ao Alessandro, parceiro na realização de muitos estudos na comparação de locutor.

É importante agradecer aos membros da banca, os professores, Frederico da Cruz, Hani Yehia, Luiz Wagner Biscainho, Plínio Barbosa e Zélia Peixoto pelas contribuições ao presente trabalho.

Não posso deixar de agradecer ao professor coordenador Wellington Dutra, do Centro Universitário Newton Paiva, pelo apoio durante o período de doutorado.

Gostaria de fazer um agradecimento póstumo a Altivo Pereira Lima, que anteviu esta etapa, e ao professor Félix Carneiro Carvalho pelas dicas e discussões mais diversas e aleatórias.

Por fim, gostaria de agradecer a todos os pesquisadores que compartilham seus trabalhos, que de forma direta ou indireta contribuíram para este resultado.

“If fifty million people say a foolish thing,
it is still a foolish thing.”
Anatole France

Resumo

A Comparação Forense de Locutor (CFL) é o exame pericial que tem como tarefa analisar duas amostras de voz e inferir sobre a compatibilidade de suas características. Uma amostra de voz é vestígio de um fato típico penal enquanto a segunda é de um indivíduo conhecido. A CFL difere-se da biometria por voz em vários aspectos. A biometria permite o controle de algumas variáveis não disponíveis na CFL, entre elas, o dispositivo de gravação, o ruído de canal, a quantidade e duração das amostras e a cooperação dos locutores. Além disso, não existe o risco de associar incorretamente um inocente (erro do Tipo I) ou falhar em associar um culpado (erro do Tipo II). Neste cenário, a presente tese apresenta duas linhas de trabalho experimentais motivadas pelo paradigma das ciências forenses, o que inclui resultados quantitativos baseados em medidas estatísticas apoiadas por bancos de dados representativos. Os experimentos foram conduzidos para emular condições presentes na prática da CFL. Na primeira linha desenvolveu-se uma solução sintética para o Teste de Significância Genuinamente Bayesiano (FBST - *Full Bayesian Significance Test*) sobre a média com variância desconhecida propondo uma estimativa por intervalo, denominada intervalo de evidência, aplicável à CFL. Os experimentos com variação da SNR mostraram que o intervalo de evidência reduziu as taxas de erro (Tipo I e Tipo II) em torno de 6,4%, superando os demais métodos avaliados. A segunda linha propôs a utilização de medidas espectrais de relação sinal-ruído S^2NR (*Spectrographic Signal-to-Noise Ratio*) para separar as características do sinal de voz em conjuntos nebulosos e realizar a comparação de locutores considerando a influência destes conjuntos. Nesta linha de trabalho propôs-se ainda uma adequação para o cálculo das estatísticas de Baum-Welch. Os experimentos, baseados em conjuntos nebulosos, variaram a SNR e o tamanho das amostras. Os resultados, em relação as outras técnicas, reduziram as taxas de falso positivo (erro Tipo I) em $35,7\%$ em amostras contaminadas, e apresentaram uma acurácia 4,5% superior para amostras com limitação de tamanho. Do ponto de vista prático, os resultados são promissores e estão sendo utilizados de forma experimental no Instituto de Criminalística da Polícia Civil de Minas Gerais.

Palavras-chave: Comparação Forense de Locutor, Fonética Forense, Ciência e Tecnologia da Fala, Processamento de Sinais, Reconhecimento de Padrões.

Abstract

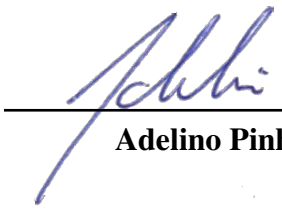
Forensic Speaker Comparison (FSC) is an analysis of two voice samples to infer the similarity of their features. One voice sample is a trace of a criminal fact, while the other is from a known individual. FSC differs from voice biometrics in several ways. Biometrics allows the control of some variables not controllable in FSC, among them, the recording device, the channel noise, the quantity and duration of the samples, and speaker cooperation. Also, there is no risk of associating (or failing to associate) an innocent to guilty part. The new paradigm shift of forensic sciences can be characterized as quantitative databased implementation of the likelihood-ratio framework with quantitative evaluation of the reliability of results. This thesis presents two lines of experimental work within the new paradigm shift. The experiments were conducted using two databases for training and validation. Parameters such as noise type and intensity of speech contamination, as well as the duration of the speech sample were evaluated. The first line developed a synthetic solution for the Full Bayesian Significance Test (FBST) over the mean with unknown variance, proposing an interval estimation, hereinafter referred to as evidence interval, applicable to the FSC. In the experiments, variation of the SNR showed the evidence interval reduced error rates (Type I and Type II) by approximately 6.4%, surpassing other evaluated methods. The second line of investigation proposed the use of Spectrographic Signal-to-Noise Ratio (S^2NR) measures to separate the signal into fuzzy sets for the comparison of speakers. Moreover, it was also proposed an adaptation for the calculation of the Baum-Welch statistics. Experiments with combinations of SNR and speech size showed the proposed method reduced false-positive rates by 35.7%, rising also accuracy in 4.5%, compared to other evaluated techniques. The results are promising and are being used experimentally at the Instituto de Criminalística of Polícia Civil de Minas Gerais.

Keywords: Forensic Speaker Comparison, Forensic Phonetics, Speech Science and Tecnology, Signal Processing, Pattern Classification.

Declaração de Originalidade

Declaro que este texto é de minha autoria e relata meus trabalhos originais. Todas as imagens e gráficos referentes aos resultados produzidos foram pela análise de gravações das bases de dados desenvolvidas durante a pesquisa.

A fonte de informação de cada material, imagens, ideias, gráficos e algoritmos que foram utilizados ou adaptados de outros pesquisadores estão devidamente referenciadas, de acordo com as normas requeridas. Este texto não foi submetido para avaliação no âmbito de qualquer Curso de Ensino do país ou do exterior.



Adelino Pinheiro Silva

LISTA DE ILUSTRAÇÕES

Figura 1 – Diagrama indicando os processos envolvidos na comunicação por voz. . 46
Figura 2 – Diagrama apresentando as etapas do modelo fonte-filtro para produção de vogais. 47
Figura 3 – Diagramação do processo de separação dos quadros de voz, indicando o comprimento da janela t_w, e o passo de tempo t_s. 49
Figura 4 – Etapas do algoritmo para detecção de atividade de voz. 50
Figura 5 – Etapas na obtenção dos componentes mel-cepstrais com filtro triangular. . 51
Figura 6 – Exemplo indicando as etapas do processamento para obtenção dos MFCC. 53
Figura 7 – Etapas do algoritmo S^2NR para obtenção da relação sinal-ruído por quadro. 54
Figura 8 – Exemplo indicando as primeiras etapas do processamento para obtenção do S^2NR. 55
Figura 9 – Exemplo indicando as etapas finais do processamento para obtenção do S^2NR. 56
Figura 10 – Predição linear com excitação por código (CELP - *Code-excited linear prediction*). 58
Figura 11 – Etapas da comparação de locutores utilizando GMM-UBM. 60
Figura 12 – Etapas da comparação de locutores utilizando *i-vector*. 64
Figura 13 – Exemplo da função densidade de Probabilidade $f_n(\mu,\rho|n,\bar{x},Q)$. 74
Figura 14 – Contorno e pontos extremos do conjunto tangente T^*. 75
Figura 15 – Exemplo de conjunto $T^{0,05}$ que delimita a superfície de integração. . . . 78
Figura 16 – Corte transversal sobre o conjunto tangente indicando os limites do intervalo de evidência. 79
Figura 17 – Exemplo da estimativa por intervalo sobre a média de $LLR(x_Q)$. 80
Figura 18 – Cenários de resultados da inferênca intervalar. 82
Figura 19 – Curva DET da etapa de treinamento das técnicas de estimativa por intervalo. 83

Figura 20 – Percentual de classificações corretas para cada método de estimativa por intervalo com diferentes valores de SNR na etapa de testes. 84
Figura 21 – Resultados da estimativa por intervalo apresentando o percentual de ocorrências dos cenários (b) e (c) da Figura 18. 84
Figura 22 – Gráfico RDI do tamanho de intervalo para cada método de estimativa por intervalo. 85
Figura 23 – Etapas da comparação de locutores utilizando *fi-vector*-S^{2}NR. 88
Figura 24 – Conjuntos fuzzy-SNR calculados a partir das amostras padrão da etapa de treinamento. 90
Figura 25 – Curva DET da etapa de treinamento das técnicas de verificação de locutor enumeradas na Tabela 3. 92
Figura 26 – Taxas de verdadeiro positivo, falso negativo e acurácia para cada metodologia e de acordo com a contaminação SNR do áudio questionado. . . . 93
Figura 27 – Análise de variância entre a metodologia de referência e as técnicas que apresentaram melhor desempenho. 93
Figura 28 – Curva RDI apresentando os valores de verdadeiro positivo, falso positivo e acurácia em relação à duração da amostra questionada. 94
Figura 29 – Análise de variância entre a metodologia de referência e "*fi-vector* 300-FLDA". 95
Figura 30 – Média das taxas de verdadeiro positivo, falso positivo e acurácia em relação à duração da amostra questionada pelo tipo de ruído. 96
Figura 31 – Médias das taxas de verdadeiro positivo, falso positivo e acurácia em relação à duração da amostra questionada e a intensidade do ruído. 97
Figura 32 – Análise de variância entre o método de referência e "*fi-vector* 300-FLDA". 98
Figura 33 – Gráfico RDI apresentando a duração dos áudios padrão, antes e depois da detecção de atividade de voz. 116
Figura 34 – Exemplo do resultado do alinhamento dos registros de áudio de cada microfone. 117
Figura 35 – Gráfico RDI apresentando a duração dos áudios do Corpus CEFALA-1 antes e depois do processamento VAD. 118
Figura 36 – Gráfico RDI apresentando a duração das etapas dos áudios do Corpus CEFALA-1 antes e depois do processamento VAD. 119

LISTA DE TABELAS

Tabela 1 – Diferenças técnicas entre o áudio questionado e o áudio padrão. 40
Tabela 2 – Resultado comparativo com os percentuais de classificação entre a estimativa pontual e a estimativa por intervalo. 83
Tabela 3 – Resumo das variações propostas sobre a metodologia *i-vector* de referência. 90
Tabela 4 – Resumo de resultados das etapas de treinamento e teste para os principais resultados dentre as variações da metodologia de referência. 92
Tabela 5 – Estatísticas temporais dos áudios padrão do corpus Criminal-Contínuo, em minutos, antes e após processamento VAD. 115
Tabela 6 – Estatísticas temporais do corpus CEFALA-1, em minutos, antes e após o processamento VAD. 119

LISTA DE ABREVIATURAS E SIGLAS

ACELP	*Algebraic* CELP
AMR	*Adaptive Multi Rate*
ANOVA	*Analysis of Variance*
CEFALA	Centro de Estudos da Fala, Acústica, Linguagem e Música
CELP	*Code-Excited Linear Prediction*
CFL	Comparação Forense de Locutor
CNC	Congressos Nacional de Criminalística
DCT	*Discrete Cossine Transform*
DET	*Detection Error Tradeoff*
DFT	*Discrete Fourier Transform*
EER	*Equal Error Rate*
EFR	*Enhanced Full Rate*
EM	*Expectation–Maximization*
ENFSI	*European Network of Forensic Science Institutes*
ETSI	*European Telecommunications Standards Institute*
FBST	*Full Bayesian Significance Test*
FDP	Função densidade de probabilidade

FLDA	*Fuzzy Linear Discriminant Analysis*
FFT	*Fast Fourier Transform*
GMM	*Gaussian Mixture Models*
GSM	*Global System for Mobile Communications*
HMM	*Hidden Markov Models*
HRSS	*Hightest Relative Surprise Set*
IC-MG	Instituto de Criminalística de Minas Gerais
IFT	*Inverse Fourier Transform*
IID	Independente e Identicamente Distribuída
ITU	*International Telecommunication Union*
KNN	*K-Nearest-Neighborhood*
LDA	*Linear Discriminant Analysis*
LLR	*Log-Likelihood Ratio*
LPC	*Linear Predictive Coding*
LR	*Likelihood Ratio*
LSP	*Line Spectral Pairs*
LTP	*Long Term Prediction*
MAP	*Maximum a Posteriori*
MEFCA	*Maximum Entropy Fuzzy Clustering Algorithm*
MEMS	*Micro-ElectroMechanical Systems*
MCMC	*Markov Chain Monte Carlo*
MFCC	*Mel-Frequency Cepstral Coefficients*
MFEC	*Mel-Frequency Entropy Ceptrum*
ML	*Maximum Likelihood*
MLP	*Multilayer Perceptron*

MMSE	*Minimum Mean Square Error*
MVKD	*Multivariate Kernel Density*
NB	*NarrowBand*
PCM	*Pulse Code Modulation*
PCMG	Polícia Civil de Minas Gerais
PLDA	*Probabilistic Linear Discriminant Analysis*
PLP	*Perceptual Linear Predictive*
PNCC	*Power Normalized Component Cepstrum*
POP	Procedimento Operacional Padrão
PTT	*Push to Talk*
RASTA	*Representations Relative Spectra*
RDI	*Raw(data), Description and Inference*
RNA	Rede Neural Artificial
RPE	*Regular Pulse Excitation*
SENASP	Secretaria Nacional de Segurança Pública
S^2NR	*Spectrographic Signal-to-Noise Ratio*
SNFF	Seminário Nacional de Fonética Forense
SNR	*Signal-to-Noise Ratio*
SPAV	Setor de Perícias em Áudio e Vídeo
SPTC	Superintendência de Polícia Técnico-Científica
SSCH	*Subband Spectral Centroid Histograms*
TEOCC	*Teager Enegy Operator Component Cepstrum*
UBM	*Universal Background Model*
VAD	*Voice Activity Detection*
ZCPA	*Zero-Crossing with Peak Amplitude*

Prefácio

Os estudos aqui apresentados tiveram como ponto de partida os desafios enfrentados dentro do Instituto de Criminalística de Minas Gerais a partir do ano de 2009, período que coincide com o primeiro treinamento que o autor fez em fonética forense junto à Secretaria Nacional de Segurança Pública (SENASP).

Após o fim do curso, em 2010, começaram os esforços, no Instituto de Criminalística, para aprimorar o exame de Verificação de Locutor. Nesta época foram desenvolvidas algumas rotinas para extração de características e um pequeno corpus composto por 28 (vinte e oito) falantes (colegas de trabalho) e 50 (cinquenta) pares de registros (padrões de voz, e seus respectivos áudios questionados) selecionados de perícias realizadas pelo autor desta tese. Uma série de leituras, estudos e experimentos não publicados foram realizados, com este material, antes do início do doutorado.

No início do doutorado, no segundo semestre de 2015, dividia atenção com minha primeira filha recém nascida, os plantões no Instituto de Criminalística e com o Centro Universitário Newton Paiva[1]. As pesquisas tiveram como ponto de partida os modelos de produção da voz e fala, incluindo o modelo fonte-filtro, o modelo dinâmico de pregas vocais e técnicas de extração de características acústicas.

Na sequência, foram objeto de estudo as técnicas de inferência para a comparação destas características, momento em que três linhas principais se consolidaram. A primeira propunha uma comparação de locutores pareada em "classes" de sons da fala, a segunda investigava uma estimativa intervalar baseada no Teste de Significância Genuinamente Bayesiano (FBST - *Full Bayesian Significance Test*), e a terceira explorava a contaminação por canal e ruído.

[1] Época em que dois grupos de alunos desenvolveram seus Trabalhos de Conclusão de Curso (TCC), que constam no apêndice, na área de ciência da fala.

No período, com ajuda do meu colega Arlindo Neto e contribuição de todos do CEFALA, também realizou-se a coleta das vozes para o corpus CEFALA-1[2] (NETO et al., 2019).

Essas três linhas de trabalho foram apresentadas no exame de qualificação realizado em 2017. A segunda abordagem, que propôs a estimativa por intervalo, foi desenvolvida após o professor Filipe Zabala sugerir o FBST quando nos conhecemos no XI Seminário Nacional de Fonética Forense[3]. Esta proposta ganhou um pouco de consistência e sua evolução foi discutida em Silva et al. (2018e) e recentemente aceita para publicação em Silva et al. (2020).

A linha de trabalho que propõe o pareamento começou com a ideia de dividir a Comparação Forense de Locutor em classes acústicas. Os primeiros resultados foram apresentados por Silva (2016), Silva et al. (2017) e Silva (2017). Esta ideia fundiu-se, em parte, com a terceira e evoluiu para pareamento utilizando os valores da relação sinal-ruído ao longo do sinal de voz[4]. Este desenvolvimento foi inspirado após uma conversa com meus orientadores e viabilizado pelo trabalho do meu colega João Pedro Sansão. Os resultados desta segunda abordagem foram apresentados por Silva et al. (2018c), Silva et al. (2018d) e Silva et al. (2018e). Posteriormente, foram agregadas técnicas baseadas em lógica *fuzzy*, e os resultados compõem esta tese e aguardam publicação[5].

Uma pequena evolução da terceira linha de trabalho, que explorou a contaminação por canal e ruído, foi recentemente publicada por Silva et al. (2019) antes de ser sobrestada dos esforços principais da pesquisa. Esta linha de trabalho também é apresentada nesta tese, as demais citadas estão apensadas ao final do texto.

Adelino Pinheiro Silva

---

[2] Mais detalhes do Corpus CEFALA-1 são apresentados no Apêndice A.

[3] O evento ocorreu em Florianópolis-SC em 2016, só pude comparecer no último dia, mas valeu a pena. Este é um exemplo de como é importante discutir seus trabalhos e ideias com a comunidade acadêmica.

[4] No presente trabalho assume-se que as terminologias "relação sinal-ruído" e "razão sinal-ruído" são equivalentes.

[5] Vide Apêndice E.

SUMÁRIO

1 INTRODUÇÃO 27
1.1 Contextualização 27
1.2 Objetivos 31
1.2.1 Geral 31
1.2.2 Específicos 31
1.3 Exames Forenses em Material Audiovisual e Equipamentos Eletrônicos 32
1.4 Organização do Texto 33

2 PARADIGMA DA COMPARAÇÃO FORENSE DE LOCUTOR 35
2.1 Ciência Forense 35
2.2 Cenário da Comparação Forense de Locutores 38
2.3 Principais Pontos 42

3 ANÁLISE ACÚSTICA PARA COMPARAÇÃO FORENSE DE LOCUTOR 45
3.1 Princípios da Produção da Voz e da Fala 45
3.2 Medição de Características de Registros Acústicos 48
3.2.1 Detecção de Atividade de Voz 49
3.2.2 Decomposição Mel-Cepstral 50
3.2.3 Medidas instantâneas de relação sinal-ruído 53
3.3 Codificação do Sistema de Telefonia Móvel no Brasil 57
3.4 Comparação Automática de Locutor 59
3.4.1 Modelagem e Classificação por GMM-UBM 59
3.4.2 Modelagem e Classificação por Vetor de Informação 64
3.5 Principais Pontos 66

4 CONTRIBUIÇÕES PARA A ROBUSTEZ E CONFIABILIDADE EM COMPARAÇÃO FORENSE DE LOCUTORES 69
4.1 Aplicação do FBST na Comparação de Locutores 70
4.1.1 Teste de Significância Genuinamente Bayesiano - FBST 70
4.1.2 Estimativa por Intervalo na CFL 72
4.1.3 FBST sobre Média com Variância Desconhecida 73

4.1.4 Definição do Intervalo de Evidência 77
4.1.5 Comparação com outros métodos 80
4.2 Comparação de Locutores Pareada pela relação sinal-ruído 85
4.2.1 Aplicação de Técnicas Fuzzy 86
4.2.2 Resultados Experimentais 89
4.3 Conclusões e Principais Pontos 98

5 CONSIDERAÇÕES FINAIS 101
5.1 Conclusões 101
5.2 Continuidade dos Trabalhos 102

REFERÊNCIAS 105

APÊNDICE A CONJUNTO DE DADOS E CORPUS 113
A.1 Corpus Criminal-Contínuo 114
A.2 Corpus Cefala-1 116

APÊNDICE B TRABALHO DE CONCLUSÃO DE CURSO - ENGENHARIA ELÉTRICA NEWTON PAIVA 121

APÊNDICE C TRABALHOS PUBLICADOS EM EVENTOS 125

APÊNDICE D TRABALHOS PUBLICADOS EM REVISTAS 133

APÊNDICE E TRABALHO AGUARDANDO REVISÃO 137

Capítulo 1

Introdução

"Emperor Qin's philosophy – the only one permitted – was called 'legalism', which is just what it sounded like, do as the law says or else. It's a philosophy that's not highly conducive to questioning authority."

— Neil deGrace Tyson

1.1 Contextualização

O uso de elementos presentes na voz e fala[1] para inferir ou obter informação, como naturalidade ou estado de saúde, é antigo. Por exemplo, relato presente em Juízes 12:5-6 apresenta como foi utilizado o termo "Xibolete" para distinguir indivíduos de dois grupos linguísticos, como transcrito a seguir:

> Para não deixar que os efraimitas passassem, os gileaditas tomaram os lugares onde o rio Jordão podia ser atravessado. Quando algum efraimita que estava tentando escapar pedia para atravessar o rio, os homens de Gileade perguntavam: — Você é efraimita? Se ele respondia que não, eles o mandavam dizer a palavra "Chibolete". Mas, se ele dizia "Sibolete" porque não podia falar direito a palavra, então o agarravam e matavam ali mesmo, na beira do rio Jordão. Naquela ocasião foram mortos quarenta e dois mil efraimitas. (Jz. 12:5-6, Nova Tradução na Linguagem de Hoje)

[1] O aparelho fonador, parte do trato vocal, é responsável pela produção de uma vasta gama de sons incluindo a voz, que no sentido estrito, refere-se apenas ao som produzido pela vibração das pregas vocais e não envolve a produção de consoantes surdas, cliques, assobios e sussurros. A fala faz uso dos sons produzidos no aparelho fonador (entre o pulmão e os lábios), para a comunicação, com aporte organizado de significado (TITZE, 1994, p. xviii).

Mais especificamente, no século XX, definiu-se o *reconhecimento de locutor* como o processo para inferir autoria de uma expressão oral, que incluiria as tarefas de identificação, verificação, discriminação e autenticação de locutores. A informação utilizada nestas tarefas estaria presente nas características do sinal acústico (ATAL, 1976). O reconhecimento de locutor iniciou-se com o advento das gravações em fita cassete e das ferramentas de análise espectrográfica. Juntamente com a espectrografia surgiu a ideia da impressão visual da voz – *voiceprint* – e suas comparações para reconhecimento de locutor, inclusive forenses, que logo foram consideradas controversas, pois não eram capazes de distinguir locutores de forma satisfatória (YOUNG; CAMPBELL, 1967; HOLLIEN, 1990; BROEDERS, 2001).

Na prática, a comparação forense de locutor (CFL) consiste no confronto de características de dois registros de áudio, com o objetivo de fornecer informações de autoria para elucidação criminal, como também em assuntos cíveis ou administrativos. Na biometria (HANSEN; HASAN, 2015), denomina-se *verificação de locutor* a comparação de dois registros de áudio para uma reivindicação de identidade. A *identificação de locutor* é a tarefa de localizar um falante dentro de um conjunto de registros de voz[2].

Atualmente, a comparação forense de locutor pode ser dividida em três abordagens distintas e suas combinações, conforme enumeradas por Figueiredo (1994), Broeders (2001), Hansen e Hasan (2015):

- a primeira, denominada perceptiva, é realizada por foneticista treinado para realizar análise fonética, qualitativa e quantitativa, e extrair, com ou sem suporte computacional, uma variedade de parâmetros fonéticos e linguísticos;
- a segunda, denominada semi-supervisionada, realiza análise de parâmetros da voz e fala – como formantes de vogais, taxa de articulação e estimação de medidas anatômicas, extraídas semi-automaticamente por profissionais ou através de ferramentas computacionais para análise de voz;
- e a terceira, completamente automática, verifica os locutores pelo cálculo de pontuações (*scores*) de similaridade de características extraídas dos registros de áudio por algoritmo especializado.

Os autores supracitados sugerem que todas as abordagens façam uso de uma base de dados como suporte. Na verificação automática, a análise da base de dados gera o *Universal Back-*

[2] No Capítulo 2 tem-se uma explicação mais detalhada da tarefa de CFL, bem como as definições e contribuições da verificação e identificação de locutor (incluindo a autenticação e a distinção de locutor) na CFL.

ground Model (UBM) ou modelo de falante de fundo universal[3] (REYNOLDS et al., 2000; TOGNERI; PULLELLA, 2011; HANSEN; HASAN, 2015).

Nos trabalhos de Gold e French (2011) e Morrison et al. (2016) têm-se relatos de práticas da CFL que combinam procedimentos automáticos e o uso de características tanto da fala, e.g. processos fonológicos e da análise de padrões de formantes em vogais, quanto de aspectos fonatórios, e.g., frequência fundamental e qualidade de voz.

A verificação automática de locutor é difundida e utilizada para reconhecimento biométrico, permitindo utilizar a fala como senha de acesso em sistemas de segurança ou em dispositivos com reconhecimento pela fala (TOGNERI; PULLELLA, 2011).

Em outra direção, a verificação de locutor realizada exclusivamente por especialista nativo do idioma (BROEDERS, 2001) – ou abordagem perceptiva – é demorada (podendo chegar a algumas semanas), a confiabilidade é questionável (MORRISON, 2014), e são necessárias várias ocorrências de cada evidência para dar suporte a um resultado não quantificado. Esta abordagem perceptiva também possui pouca repetibilidade e muita subjetividade, pois dois especialistas podem, a partir do mesmo corpo de delito e diferentes argumentações, chegar a conclusões distintas (FIGUEIREDO, 1994; MORRISON, 2014; HANSEN; HASAN, 2015).

O contraponto é que os métodos automáticos, na CFL, são mais sensíveis a efeito de canal, codificação, ruído, quantidade de áudio da amostra, mimetização, cooperação do locutor, e do banco de dados utilizado como suporte (MORRISON, 2014). Nos dois extremos das abordagens nota-se um problema com a consiliência entre resultados.

Frente às abordagens utilizadas na CFL, o presente trabalho tem como questão propor aprimoramentos às técnicas automáticas de verificação de locutor aplicadas em cenários característicos da CFL. Isto inclui comparação realizada em conjunto aberto de locutores, limitação da duração da amostra, contaminação por ruído e presença do efeito da codificação de canal de comunicação. Os aprimoramentos visam a melhoria dos índices de desempenho (taxa de verdadeiro positivo, taxa de falso positivo, acerto e erro).

Por atuar realizando exames periciais de Comparação Forense de Locutores, o autor da presente tese motivou-se na necessidade de aprimorar a robustez[4], a testabilidade e a repetibilidade do referido exame frente ao crescimento de sua demanda. Tal crescimento vem da combinação do atual modelo de comunicação telefônica e a facilidade de obtenção de registros de áudio para produção de provas no meio jurídico brasileiro.

[3] Explicações mais detalhadas dos termos relacionados a verificação de locutor serão apresentadas na Seção 3.4.

[4] No contexto da CFL, a robustez do exame implica em comparações com repetibilidade (independente de variáveis de confusão) e com maiores taxas de verdadeiro positivo e menores taxas de falso positivo.

Também motiva a presente pesquisa a necessidade de tornar o exame pericial de CFL mais transparente, quantitativo, célere, disponível ao contraditório[5], menos suscetível às variáveis de confusão e menos dependente do perito criminal ou auxiliar técnico.

A necessidade da robustez encaixa-se no novo paradigma para a CFL proposto por Morrison (2009), que inclui o uso da estrutura de razão de verossimilhança (*Likelihood Ratio* - LR). O novo paradigma propõe testabilidade, repetibilidade e transparência, com comparações baseadas em medidas quantitativas, apoiadas por bancos de dados representativos, modelos estatísticos e testes de validação, para serem confiáveis dentro das condições da investigação jurídica ou criminal.

Atualmente, na comparação de dois registros acústicos, é possível indicar evidências quantitativas que rejeitem (ou falhem em rejeitar) a hipótese de associá-los ao mesmo locutor (GONZALEZ-RODRIGUEZ et al., 2007). Entretanto, a verdadeira questão não é somente o quanto o registro padrão é compatível ou não com um registro questionado, mas também quão confiável e robusto é o resultado e qual a possibilidade de existir outros locutores não analisados também compatíveis. Este problema é considerado nesta pesquisa, em especial para evitar a falácia da individualização (SAKS; KOEHLER, 2008)[6].

A relevância da questão anterior está no aforismo jurídico do princípio da inocência *Absolvere nocentem satius est, quam condemnare innocentem*[7]. A CFL não se resume na associação entre um registro acústico e um falante, mas também na rejeição da possibilidade de outros falantes serem associados em situações onde os registros possuem semelhança do ponto de vista perceptivo (MARTINS et al., 2014).

Motivado pelo nicho, o presente trabalho visa a propor aprimoramentos e métodos que podem ser utilizados na CFL. Isto inclui obter informações sobre a robustez – taxas de acerto e erro na distinção e associação dos falantes – e confiabilidade – através de estimativa por intervalo[8] – com uma metodologia que permita a repetibilidade e transparência dentro do recorte do novo paradigma da ciência forense (COUNCIL et al., 2009; MORRISON, 2014).

---

[5] O contraditório visa permitir às posições jurídicas opostas entre si o acesso aos dados e às análises periciais, de modo que o trabalho forense não assuma posição, limitando-se a aplicar a ciência de maneira imparcial.

[6] A "falácia da individualização" consiste em uma afirmação categórica, muitas vezes sem suporte empírico, de que características comparadas nos exames periciais são individualizadoras, como no exemplo da grafotecnia indicado por Valente (2012 apud MENDES, 1999, p. 6) "A escrita é individual. A escrita é resultante de estímulos cerebrais que determinam movimentos e estes criam as formas gráficas. Muito embora os cérebros de todos sejam anatomicamente iguais, a sua função varia de pessoa para pessoa. O mesmo ocorre com o sistema somático. Vale dizer, portanto, que ambos tendem variar ao infinito. Como a escrita resulta do concurso desses dois sistemas, evidentemente ela também varia ao infinito. Se assim não fosse, a perícia grafotécnica, que é aceita universalmente, não teria o menor valor.".

[7] É melhor absolver um culpado do que condenar um inocente.

[8] A estimativa por intervalo (ou inferência intervalar) é a técnica estatística de estabelecer um intervalo de valores plausíveis para um parâmetro desconhecido.

O primeiro aprimoramento consiste na proposta de estimação intervalar para ser aplicado à metodologia de verificação de locutor GMM-UBM[9] em cenários de CFL. A estimação intervalar foi viabilizada pelo desenvolvimento do método para solucionar o Teste de Significância Genuinamente Bayesiano (FBST - *Full Bayesian Significance Test*) para o valor esperado (média) com variância desconhecida sem o uso de Monte Carlo e Cadeia de Markov (MCMC).

O segundo aprimoramento consiste em aplicar métodos de lógica *fuzzy*, com base em medidas de relação sinal-ruído (SNR - *Signal-to-Noise ratio*), à metodologia de verificação de locutor *i-vector*[10] em cenários de CFL. A lógica *fuzzy* foi aplicada em dois momentos, o primeiro para parear as amostras de voz de acordo com a SNR e, em um segundo momento, na etapa de análise discriminante linear.

Os aprimoramentos à verificação automática de locutor aplicada à CFL foram comparados por método experimental. Para emular o cenário forense utilizou-se duas bases de dados, uma de treinamento e uma para testes, codificação do canal de comunicação, contaminação por ruído e limitação do tamanho do áudio. Cada aprimoramento proposto foi comparado com as metodologias não aprimoradas, e os resultados indicam que as propostas apresentaram acurácia superior nos cenários que emulam a CFL.

1.2 Objetivos

1.2.1 Geral

O objetivo do presente trabalho é propor aprimoramento de métodos automáticos aplicados às condições da CFL utilizando de estimativa por intervalo e de comparação pareada por medidas de relação sinal-ruído com falantes do Português Brasileiro.

1.2.2 Específicos

- Avaliar as técnicas clássicas de extração e análise de características em condições que emulam a CFL para o Português Brasileiro;
- Descrever e avaliar o Teste de Significância Genuinamente Bayesiano (*Full Bayesian Significance Test* - FBST), apresentado por Pereira e Stern (1999), e propor uma metodologia de estimativa por intervalo baseada no FBST com aplicações na CFL.

[9] A metodologia GMM-UBM é detalhada na Seção 3.4.1.
[10] A metodologia *i-vector* é detalhada na Seção 3.4.2.

- Avaliar o desempenho da estimativa por intervalo aplicada ao reconhecimento automático de locutores em cenários de codificação e contaminação por ruído que emulam as condições da CFL.
- Aprimorar técnicas de verificação de locutor em conjuntos abertos, agregando técnicas *fuzzy* e pareamentos baseados na relação sinal-ruído.
- Comparar o aprimoramento supracitado com métodos de origem em cenários de codificação, contaminação por ruído e tamanho da amostra que emulam as condições da CFL.

1.3 Exames Forenses em Material Audiovisual e Equipamentos Eletrônicos

A prática forense em solo brasileiro pode ser dividida, a grosso modo, em dois grandes blocos: a criminalística e a medicina legal. Essa divisão é baseada na organização das Instituições Periciais e de Polícia Técnico-Científica que atuam em território brasileiro[11].

A criminalística é o ramo da ciência forense que analisa locais e objetos visando a rastrear, colher e explicitar provas e evidências. Desta análise são extraídas informações para dar suporte a diferentes hipóteses de causalidade. Na maioria dos casos, a hipótese de causalidade envolve uma infração da Lei Penal. As bases científicas da criminalística remontam ao trabalho de Locard (1939), que é uma das primeiras publicações que busca aplicar a metodologia científica na elucidação de infrações da lei penal. Uma curiosidade é que o trabalho de Locard (1939) não inclui material audiovisual em sua lista de vestígios de crimes, o que é justificável, uma vez que, na época, estas formas de registros eram caras e pouco difundidas e suas análises estavam em desenvolvimento.

A maioria dos Institutos de Criminalística brasileiros são divididos em setores relacionados a uma especialidade. Essa pode estar relacionada com a hipótese do crime – como setores de perícias em crimes contra a vida – ou com a natureza do vestígio ou área do conhecimento, como seção de documentoscopia ou seção de biologia legal.

O Instituto de Criminalística de Minas Gerais (IC-MG) é vinculado à Superintendência de Polícia Técnico Científica (SPTC), órgão pertencente à estrutura da Polícia Civil de Minas Gerais (PCMG). A divisão setorial do IC-MG segue a maioria dos Institutos de Crimina-

[11] No caso do Brasil, a Polícia Técnico-Científica não é prevista na constituição. Mesmo assim, a instituição Polícia Técnico-Científica, ou análoga, existe em dezoito estados. Em alguns casos a Polícia Técnico-Científica é parte da Polícia Judiciária, – como o caso de Minas Gerais, Distrito Federal e da União. Dos vinte e oito modelos de Polícia Técnico-Científica, todas incluem o Instituto de Criminalística, vinte e sete o Instituto de Medicina-Legal e vinte e sete, com exceção de Minas Gerais, incluem o Instituto de Identificação.

lística brasileiros. O exame forense de comparação por voz tem seus técnicos alocados no Setor de Perícias em Áudio e Vídeo (SPAV). O referido setor é responsável pelos exames em material audiovisual e equipamentos eletrônicos e as tarefas envolvem principalmente conhecimentos de eletrônica, *design* gráfico, artes visuais, linguística e estatística. Os exames periciais realizados pelo SPAV podem ser enumerados da seguinte forma:

1. **Exames em registros de áudio**: Análise do conteúdo e informações do registro de áudio; tratamento para melhoria da inteligibilidade; verificação de edição nos registros e comparação forense de locutores.
2. **Exames em registros de imagem**: Análise do conteúdo e informações dos registros; tratamento e extração de informações; comparação de características de cena com pessoas e objetos; comparação facial; retrato falado e verificação de edição.
3. **Exames em equipamentos eletrônicos**: Análise e extração de informações dos equipamentos; identificação ou constatação dos modos e meios de funcionamento.

Parte das técnicas utilizadas e exames periciais em material audiovisual e eletrônico provém de tratados, como Fernandes (2014), artigos, como Maher (2009) e Rodríguez et al. (2010), e de recomendações internacionais como as da ENFSI (*European Network of Forensic Science Institutes*). As experiências profissionais brasileiras são compartilhadas no "Congresso Nacional de Criminalística" (CNC), realizados a cada dois anos, no "Seminário Nacional de Fonética Forense" (SNFF), em anos intercalados ao CNC, e em grupos de discussão de amplitude nacional.

No meio forense, o ramo que envolve a comparação de registros acústicos de falantes foi denominado por Hollien (1990) como *Fonética Forense*. Tal denominação é amplamente utilizada por outros autores da área. No cenário brasileiro, encontram-se ainda termos como Fonoaudiologia Forense, Acústica Forense e Áudio Forense, além de Fonética Forense para referir-se à CFL.

O exame pericial de comparação por voz, realizado pelas polícias técnico-científicas, possui diferentes denominações, como *Identificação / Verificação de Locutor*[12], *Identificação por Voz e Fala* ou, ainda, *Comparação Forense de Locutor*[13].

1.4 Organização do Texto

Esse capítulo apresentou, de forma introdutória, a problemática da pesquisa dentro do novo paradigma da CFL. O paradigma implica em tornar a ciência forense mais rigorosa, baseada

[12] Terminologia utilizado pela Polícia Civil de Minas Gerais.
[13] Terminologia mais abrangente e será amplamente utilizado neste texto

nos princípios da falseabilidade, transparência, repetibilidade e consiliência entre resultados, e incorporando dados empíricos e estatísticas pertinentes.

Em continuidade, o presente trabalho apresenta, no Capítulo 2, uma exposição sobre o confronto forense de registros acústicos. Nesta etapa do trabalho, busca-se imergir o leitor no contexto, que envolve as nuances presentes no ambiente das ciências forenses. O autor desta tese, como profissional da área, buscou a imparcialidade na apresentação do contexto forense[14].

No Capítulo 3, tem-se um breve resumo das técnicas utilizadas para extração de informação, medidas acústicas e comparação de locutores. O foco mantém-se nos métodos que foram explorados nos experimentos apresentados nesta tese.

O Capítulo 4 apresenta as duas linhas de trabalho para onde a pesquisa convergiu. Em relação à contribuição quanto à confiabilidade, foi proposto o *intervalo de evidência*[15], que busca adicionar um grau de flexibilidade ao resultado de uma CFL. Em relação à robustez, propõe-se um aprimoramento de técnicas de verificação de locutor em conjuntos abertos, agregando técnicas *fuzzy* e pareamentos baseados na relação sinal-ruído. Nesse caso, a robustez aparece na forma da melhoria das taxas de verdadeiro positivo e de falso negativo em cenários com contaminação e limitação dos áudios questionados.

O Capítulo 5 apresenta as conclusões, uma breve discussão do trabalho e as propostas de continuidade.

[14] O trabalho no ambiente forense é desafiante e tenso. O desafio é aplicar a ciência nos ambientes mais hostis de obtenção de dados. Na maioria da ocasiões vestígios e evidências só podem ser coletados pela perseverança do perito criminal. Por outro lado, há ocasiões não raras em que a administração hierárquico-disciplinar – e às vezes política – confunde-se com a metodologia científica criando situações moralmente constrangedoras. No caso, a imparcialidade reside em despir-se do gosto e do desgosto com foco no fato científico.

[15] A denominação "intervalo de evidência" foi forjada pelo autor e foi inspirada pelas denominações utilizadas nas publicações sobre FBST, além de possuir apelo semântico às Ciências Forenses.

Capítulo 2

Paradigma da Comparação Forense de Locutor

"I was never a saint. I was just a guy chasing justice."

— Barry Allen

O Capítulo 1 apresentou o tema do presente trabalho listando alguns detalhes de forma geral, principalmente aqueles que tangem o meio forense. Como o assunto é amplo, dedica-se este capítulo a definir estritamente alguns conceitos e discutir brevemente o assunto, enumerando as particularidades do problema tratado na presente tese.

Primeiramente, são apresentados o conceito de ciência forense e as principais particularidades que envolvem a CFL. Em seguida apresenta-se como as técnicas e métodos utilizados no meio criminal – em especial na CFL – evoluíram frente à metodologia científica.

2.1 Ciência Forense

O termo forense refere-se a fórum, que no direito Romano é a praça, ou o lugar público para discursos, discussões e, inclusive, processos criminais. Etimologicamente, forense é tudo aquilo que pode ser apresentável ao público, podendo-se esperar objetividade, confiabilidade e transparência.

Ciência forense, em definição mais ampla, é a aplicação da ciência à lei, onde a tecnologia provida pelo conhecimento científico é utilizada na definição e execução da legislação imposta por agências de polícia em um sistema de justiça criminal (SAFERSTEIN, 2004, p. 27).

Recai sobre o Estado[1] o poder e o dever de impor uma sanção a quem praticar a infração da lei penal, que possui como elementos a conduta, o resultado (jurídico ou normativo), a tipicidade e, por fim, nexo de causalidade. A ciência forense busca unir estes elementos (conduta, resultado, tipicidade e nexo causal) a partir de informações objetivas e materiais – os vestígios – presentes na infração da lei penal.

Durante o processo de elucidação da infração penal são aplicados os princípios constitucionais da legalidade e publicidade, e cabe ao sistema judiciário[2] iniciar a investigação, colher provas, apreender objetos e realizar os exames de corpo de delito[3].

O exame de corpo de delito – ou exame forense – é a ferramenta legal de análise de vestígios. Tal exame é considerado indispensável e pode ser realizado em qualquer dia e a qualquer hora[4]. Os exames forenses são capazes de fornecer informações objetivas, podendo vincular a conduta criminosa ao resultado, explicitando a dinâmica dos fatos e o nexo de causalidade (SAFERSTEIN, 2004).

No entanto, a condução de uma elucidação penal é balizada por princípios como o de preservação do ônus da prova[5], onde a prova de cada alegação é incumbida a quem a fizer, presunção de inocência até que se prove o contrário[6] e o *nemo tenetur se detegere*[7] (que veda a auto-incriminação, que garante o direito ao silêncio[8]). Tais princípios conduzem à elucidação criminal e visam a impedir a utilização de provas e evidências obtidas por meios ilícitos[9].

O princípio empírico mais defendido pelos cientistas forenses é conhecido como princípio de Locard. Edmond Locard (1877–1966) era formado em medicina e direito e iniciou seus trabalhos na França. Seu princípio descreve como a interação entre o agente de um crime e a respectiva cena podem deixar evidências físicas:

> Em qualquer lugar que pise, o que quer que toque ou que deixe para trás – mesmo inconscientemente – servirá como testemunha silenciosa contra

1 Art. 4º do decreto-lei nº 3.689, de 03 de outubro de 1941.
2 Art. 100 do decreto-lei n° 2.848, de 07 de dezembro de 1940.
3 Art. 6º, incisos II, III, e VII do decreto-lei 3.689, de 03 de outubro de 1941.
4 Art. 158 e Art. 161 do decreto-lei 3.689, de 03 de outubro de 1941.
5 Art. 156 do decreto-lei 3.689, de 03 de outubro de 1941.
6 Art 5º, incido LVII da constituição federal de 1988.
7 Ninguém é obrigado a se mostrar.
8 Art 5º, inciso LXIII da constituição federal de 1988.
9 Art 5º, inciso LVI da constituição federal de 1988.

> o autor. Não apenas suas impressões digitais ou pegadas, mas também seu cabelo, fibras de tecido de suas roupas, vidro quebrado, marcas de ferramenta, arranhões na pintura, sangue ou sêmen que deposita ou recolhe. Esses e mais vestígios carregam o testemunho silencioso contra o autor. Essa é a evidência que não se esquece. Não se confunde pela emoção do momento. Não está ausente quando as testemunhas humanas estão. Essas são as evidências concretas. A evidência física não pode estar errada, não pode perjurar-se, não pode ausentar-se. Apenas a falha humana em encontrá-la, estudá-la e entendê-la pode diminuir o seu valor (PYREK, 2010 apud KIRK, 1974, p. 1) (tradução do autor).

Esse princípio não deve ser interpretado como "todo contato deixa um vestígio" e sim que um infrator da lei penal poderá deixar ou carregar vestígios do local de crime. O princípio de Locard, juntamente com outros dois princípios, da individualidade e da individualização[10], possuem uma longa história na ciência forense do século XX (ROBERTSON et al., 2016).

Estes conceitos, apesar de intuitivos, não são potencialmente falseáveis, e a possibilidade de excluir coincidências impede afirmações categóricas de individualização. Outra questão reside no número de semelhanças (características convergentes) necessárias para uma individualização, pois, como não existe um critério que indique o número (ou nível), qualquer valor ou nível escolhido passa a possuir um caráter arbitrário (SAKS; KOEHLER, 2008; ROBERTSON et al., 2016)[11]. Dentro do paradigma, o cientista forense é incumbido da responsabilidade de descrever o valor da evidência se afastando do "princípio" de individualização (SAKS; KOEHLER, 2008; ROBERTSON et al., 2016).

Outra particularidade da aplicação da ciência forense é que os sujeitos de interesse – tribunais, vítimas e sociedade – buscam nas informações o que pode ser inferido (quanto à culpa de um investigado) a partir de observações realizadas (vestígios coletados em decorrência do fato típico) sobre um fato a *priori* desconhecido (infração da lei penal). Sobre essa particularidade, Poincaré indicara que para poder calcular, a partir de um evento observado, a probabilidade de uma causa, faz-se necessário conhecer, a *priori*, a probabilidade de uma causa e, para cada causa possível, a probabilidade de ocorrência do evento observado (ROBERTSON et al., 2016). Em suma, para se inferir sobre um evento já ocorrido é preciso possuir informações prévias de como o evento se comporta[12].

[10] O princípio da individualidade indica que objetos podem ser indistinguíveis mas não existem dois objetos idênticos. O princípio da individualização, indica que se existem semelhanças suficientes entre dois objetos para excluir a possibilidade de coincidência, esses objetos devem ter vindo da mesma fonte.

[11] Um exemplo anedótico é o princípio empírico de que duas impressões digitais são consideradas como oriundas da mesma fonte se possuírem 14 (catorze) pontos característicos comuns. Dependendo do perito o valor pode variar entre 12 e 18 pontos característicos. O trabalho de Neumann et al. (2007) indica que um confronto com impressões digitais já pode ser viável a partir de 3 pontos característicos.

[12] Nesta indicação de Poincaré é possível fazer um paralelo com o teorema de Bayes, onde uma probabilidade a *posteriori* (probabilidade da causa dado um evento observado) $p(\omega|x)$ depende da probabilidade do evento observado $p(x)$ e da probabilidade de observar o evento, dado que a causa ocorreu $p(x|\omega)$.

Frente ao cenário forense, especificamente em infrações penais onde a evidência é um registro acústico, entra a aplicação técnico-científica a exames de áudio. Saferstein (2004) argumenta que as técnicas forenses devem ser submetidas aos rigores científicos, como: possibilidade de teste; revisão por pares e publicação; identificação da taxa potencial de erro; normatização para o funcionamento; e ampla aceitação dentro de uma comunidade científica relevante. Mais especificamente, em técnicas de análise forense de registros acústicos, Maher (2009) complementa que tais técnicas devem ser não-enviesadas[13], possuir confiabilidade estatística, ser não-destrutiva e aceitas por especialistas da área.

2.2 Cenário da Comparação Forense de Locutores

Esta seção delineará o presente trabalho, tentando deixar claro o cenário e as limitações presentes na comparação forense a partir de registros acústicos.

No estado de Minas Gerais, a comparação forense de registros acústicos é denominada no meio policial como Perícia de Identificação/Verificação de Locutor. Entretanto, uma vasta gama de denominações, como perícia fonética e perícia de voz, são utilizadas para referir-se aos exames realizados em dois registros acústicos com o objetivo de determinar se as vozes neles presentes podem ter sido realizações de um mesmo indivíduo.

O resultado de uma comparação forense de locutores entra no cenário social pois o resultado torna-se uma peça pública[14] utilizada na elucidação de uma infração da lei penal. Na prática processual, o resultado da comparação forense de locutores pode acabar determinando um autor para a infração da lei penal investigada. Entretanto, do ponto de vista técnico, o resultado apenas indica se as vozes possuem características compatíveis ou parecidas.

No presente texto será utilizada a denominação de Rose (2003) e Morrison et al. (2016) para origem – ou natureza – dos registros acústicos. Denomina-se *áudio questionado*[15] o registro acústico que é um vestígio físico de uma infração da lei penal. O registro acústico obtido para comparação com este *áudio questionado* é denominado *áudio padrão*.

O áudio questionado é um registro acústico não controlado que pode ter diferentes origens que alteram a codificação e a natureza do ruído, por exemplo, como uma gravação ambiental ou do sistema de radiodifusão. Entretanto, em sua grande maioria, o áudio questionado é

[13] Maher (2009) afirma que as técnicas forenses devem ser neutras em relação à hipótese da promotoria e à hipótese da defesa.

[14] Exceto os previstos no Art. 20 e Art. 201 parágrafo 6º do Decreto Lei nº 3.689, de 03 de Outubro de 1941.

[15] Muitos profissionais também utilizam a expressão áudio motivo ou áudio desconhecido.

oriundo de interceptações telefônicas ou de mensagens de voz *half-duplex*[16]. A interceptação telefônica é um procedimento previsto na condução penal[17] que consiste em registrar as conversas telefônicas que transitam por um determinado terminal.

No caso de o terminal telefônico ser de uso pessoal, sua interceptação configura invasão da vida privada do indivíduo que o utiliza. Muitas interceptações resultam em uma vasta gama de registros com os mais íntimos assuntos, e dentre eles, os registros de voz que configuram um vestígio de um ilícito penal.

Portanto, nos áudios questionados, o suposto locutor encontra-se nas mais diferentes situações psicológicas, estresse, vigília, ou até mesmo sob influência de substâncias psicoativas. Mesmo que desconfie ou saiba que está sendo interceptado o locutor apresentará voz – e consequentemente um discurso – pouco controlados.

As interceptações telefônicas ainda possuem o problema do sistema de comunicação. Como são registrados na operadora de telefonia – ou extraídos de aplicativos – os áudios sofrem influência da codificação e do meio de transmissão das chamadas. Entre elas está a amostragem a 8 kHz; a limitação da banda (NB - *NarrowBand*) entre as frequências de 300 e 3.400 Hz; e as codificações que dependem do tipo da comunicação como listados a seguir:

- ITU-T[18] G.711 PCM (*Pulse Code Modulation*) com 64 kb/s para telefonia fixa;
- EFR (*Enhanced Full Rate*) GSM (*Global System for Mobile Communications*) com 12,2 kb/s para telefonia móvel (também conhecida por codificação *GSM 06.60*); e
- AMR (*Adaptive Multi Rate*) de banda estreita para voz, definido pela ETSI (*European Telecommunications Standards Institute*) com 12,2 kb/s para sistemas 3G e posteriores como, por exemplo, aplicativos de comunicação.

Em especial, o sistema GSM para telefonia possui codificação ACELP (*Algebraic Code-Excited Linear Prediction*) com quadros (*frames*) de 20 ms, 244 bits por quadro, com atraso (*delay*) do algoritmo de 20 ms sem utilizar *look-ahead* (amostras à frente)[19].

Por outro lado, o áudio padrão é uma gravação fornecida espontaneamente[20] coletada por perito criminal treinado e seguindo um Procedimento Operacional Padrão (POP) para coleta

[16] Mensagens enviadas por dispositivos em apenas um sentido, sem simultaneidade, do tipo PoC (*push-to-talk over cellular*) como disponibilizado por alguns serviços telefônicos como Nextel ou aplicativos como Whatsapp e Telegram.

[17] Vide Art. 5° inciso XII da Constituição da República, Art. 13-A e 13-B do Decreto Lei nº 3.689, de 03 de Outubro de 1941 e Lei nº 9.296, de 24 de Julho de 1996.

[18] ITU (*International Telecommunication Union*) é o órgão das nações unidas responsável pelas tecnologias de comunicação e informação. A ITU-T é o setor da padronização das telecomunicações (*Telecommunication Standardization Sector*) da ITU.

[19] Mais detalhes da codificação GSM 06.60 são apresentados na Seção 3.3.

[20] Art 5º, inciso LXIII da Constituição Federal de 1988.

do áudio. O procedimento de coleta sugere que a gravação ocorra em ambiente de ruído controlado.

A dinâmica da coleta de áudio padrão é uma entrevista onde o fornecedor é instruído a dialogar. No diálogo, busca-se atingir o estado de linguagem habitual e uma amostra de pelo menos 5 minutos de atividade de voz dentro de um intervalo de coleta de pelo menos 20 minutos. O procedimento de coleta é realizado às cegas[21] para não enviesar o perito criminal[22]. A coleta às cegas busca reduzir o efeito do viés confirmatório tornando o estado de linguagem habitual do fornecedor uma variável desconhecida.

Existem outras variáveis que afetam as condições do fornecedor do áudio padrão. Entre elas pode-se citar que muitas vezes o fornecedor é suspeito de infração da lei penal, podendo ou não estar algemado, e pode não ser cooperativo, ou seja, o fornecedor pode não desejar ser associado ao áudio questionado e alterar ativamente parâmetros de sua voz habitual. Após a coleta, o áudio padrão torna-se peça pública e pode ser utilizada em futuros exames sob autorização do fornecedor.

Sobre as características técnicas, o áudio padrão é gravado com frequência de amostragem de 44.100 Hz por um canal com codificação PCM linear. A Tabela 1 resume alguns contrastes presentes na comparação forense de locutores.

Tabela 1 – Diferenças técnicas entre o áudio questionado e o áudio padrão.

Característica	Questionado	Padrão
Obtenção	Gravação ambiental, radiodifusão, interceptação telefônica, etc	Procedimento padronizado de coleta de áudio
Características do ambiente	Não controlado	Controlado
Interferência externa	Canal telefônico, ruído do ambiente	Mínima
Duração da gravação	Variável	Entre 5 e 20 minutos
Faixa de frequência	300 a 3.400 Hz	20 a 22.050 Hz
Microfone e dispositivos	Desconhecido ou indeterminado	Conhecido
Codificação	ITU-T G.711 PCM, EFR GSM (GSM 06.60) ou ETSI AMR	PCM-linear
Estado do Falante	Descontraído/habitual	Constrangido/controlado
Identidade do Falante	Indeterminada	Conhecida

[21] Colher o padrão de voz às cegas significa realizar o procedimento sem ter contato com o áudio questionado. A coleta às cegas ocorre por vários fatores, um deles é a demanda, que muitas vezes agenda coletas de voz antes do envio do áudio questionado. Outro fator é a separação intencional, onde um perito realiza a coleta, e outro o exame de CFL.

[22] O POP sugere que a coleta seja realizada sem conhecimento prévio do material questionado e, quando possível, por um profissional diferente do que realizará o exame.

Nas definições de nomenclatura da atividade de comparação de locutores, denomina-se reconhecimento de locutor o "(...) processo de tomada de decisão que utiliza características do sinal acústico para determinar se um determinado indivíduo é autor de um determinado registro acústico ... (tradução do autor)" (ATAL, 1976, p. 1). A literatura separa as tarefas de identificação e verificação de locutor como distintas, dentro da área de reconhecimento de locutor. Esta distinção tem como base a pergunta que se deseja responder e a natureza da tarefa decisória (ROSE, 2003).

Na identificação, compara-se uma amostra de voz (questionada) com várias amostras conhecidas (ou padrão) de uma base de dados conhecida, visando a determinar se ela foi produzida por algum dos falantes. Na verificação de locutor, compara-se uma amostra de voz com uma amostra da base de dados conhecida visando a atestar uma reivindicação de identidade (ROSE et al., 2009).

Outra distinção importante é se a base de dados é aberta ou fechada. Em uma base de dados fechada existe a expectativa, a *priori*, de que a amostra de voz a ser comparada está dentro da base de dados de locutores conhecidos, enquanto em uma base aberta esta informação não está disponível. No caso da base de dados aberta, faz-se necessário estabelecer um limite (ou *score*) a partir do qual dois registros de voz serão considerados como do mesmo falante (ROSE et al., 2009). Esse é o cenário mais comum na CFL uma vez que, em princípio, o áudio questionado é de origem desconhecida.

Em relação ao tipo de decisão, na identificação de locutor, a amostra questionada aponta qual dos falantes é o mais provável, enquanto a verificação indica se a identidade reivindicada é compatível ou não.

Tanto as técnicas de verificação quanto as de identificação de locutores podem ser utilizadas no cenário forense. Uma tarefa complementar é encontrar um peso ou uma "força da evidência" que suporta a (in)compatibilidade entre as amostras.

Sobre o resultado da CFL, alguns autores preferem a resposta categórica, que afirma se o áudio questionado é ou não uma execução do mesmo locutor que forneceu o áudio padrão (KERSTA, 1962). Entretanto, Rose (2003) afirma que é mais razoável indicar, a partir de uma abordagem baseada na razão de verossimilhança, qual hipótese seria reforçada e apresenta sugestões de como o resultado de uma razão de verossimilhança pode ser convertido em uma resposta categórica. Porém, a diferença prática é que a razão de verossimilhança não descarta a probabilidade de uma correspondência aleatória (falso positivo ou fatores de confusão) entre os áudios questionado e padrão que, em um conjunto aberto, geralmente é não nula.

Por fim, é importante diferenciar como são tratadas as hipóteses na comparação de locutor para fins biométricos e para fins forenses. Basicamente, existem duas alternativas: a hipótese de os locutores serem os mesmos ou a hipótese de os locutores serem diferentes. Na comparação biométrica, considera-se hipótese nula (H_0^{BIO}) o fato de os dados de voz serem de um determinado falante e como hipótese alternativa (H_1^{BIO}) o caso contrário (REYNOLDS et al., 2000).

O problema forense busca manter um alinhamento com a presunção de inocência, sendo que a hipótese nula (H_0^{CFL}) considera que os áudios padrão e questionado são provenientes de locutores diferentes, com o caso contrário na hipótese alternativa (H_1^{CFL}), como mostrado nas Hipóteses 2.1[23]. Nota-se que a definição de hipóteses deposita sobre o estado *onus probandi*, que é a busca de evidências para refutar a hipótese nula[24].

$$\begin{cases} H_0^{CFL}: & \text{A amostra questionada não provém do mesmo locutor que forneceu a amostra padrão,} \\ H_1^{CFL}: & \text{A amostra questionada provém do mesmo locutor que forneceu a amostra padrão.} \end{cases} \tag{2.1}$$

Sob essas hipóteses, o erro do Tipo I consiste em rejeitar a hipótese nula quando ela é verdadeira, que neste cenário equivale a imputar provas (ou evidências) a um inocente. O caso complementar, ou erro do Tipo II, consiste em não rejeitar a hipótese nula quando ela é falsa (MONTGOMERY; RUNGER, 2010). Nesse caso, deixar-se-ia de imputar uma prova (ou evidência) a um autor de um fato típico penal.

Na verificação de locutor aplicada à biometria, as hipóteses são invertidas. Nessa aplicação rejeitar o acesso de um locutor autêntico é menos grave que permitir acesso a um locutor estranho. Entretanto, alterar a ordem das hipóteses altera o significado dos erros do tipo I e do tipo II que estão respectivamente ligados à significância e ao poder do teste.

2.3 Principais Pontos

Neste capítulo, foram apresentados alguns fundamentos que envolvem as ciências forenses e como existe uma interface com o meio jurídico, principalmente no que tange a infrações da lei penal dentro do direito democrático.

[23] Alguns autores, como Gonzalez-Rodriguez et al. (2007), denominam a hipótese H_1^{CFL} como hipótese do promotor e H_0^{CFL} como hipótese da defesa.

[24] Geralmente a hipótese nula é formulada como uma igualdade. No caso, H_0^{CFL} afirma a inexistência de relação entre parâmetros da voz e fala extraídos dos dois registros de áudio.

Em seguida, definiu-se o objetivo da CFL e a origem dos registros acústicos envolvidos na comparação, introduzindo os conceitos de áudio padrão e questionado. Sobre o áudio questionado também foram apresentadas suas limitações técnicas devido às formas de obtenção dentro do processo de elucidação de fatos típicos de infração da lei penal.

No próximo capítulo serão abordadas técnicas para extração e análise de características acústicas da voz, assim como métodos de modelagem e inferência para CFL utilizadas nesta tese. As formalizações doravante consideram o cenário forense como premissa e apresentarão problemáticas mais estritas.

Capítulo 3

Análise Acústica para Comparação Forense de Locutor

"A vaidade de muita ciência é prova de pouco saber."
— Mariano José Pereira da Fonseca (Marquês de Maricá)

A questão forense e algumas de suas particularidades foram tratadas no Capítulo 2 e permearão o texto, principalmente na nomenclatura e nas características dos registros de áudio. O presente capítulo aborda as medidas e características acústicas pertinentes a este trabalho e os métodos automáticos utilizados para a comparação de locutores.

3.1 Princípios da Produção da Voz e da Fala

A comunicação falada é o principal mecanismo usado pelos seres humanos para transmitir informação. Ao mecanismo de manifestação da linguagem pode-se atribuir parte do desempenho de nossa civilização. A voz de um indivíduo é parâmetro comportamental particular que pode ser influenciada pela personalidade, classe social, humor e saúde (PIERANGELO; GIULIANI, 2007).

A comunicação falada inicia-se no cérebro do falante por um processo linguístico de geração do significado através de palavras e frases. Em ato contínuo, um comando fisiológico ativa

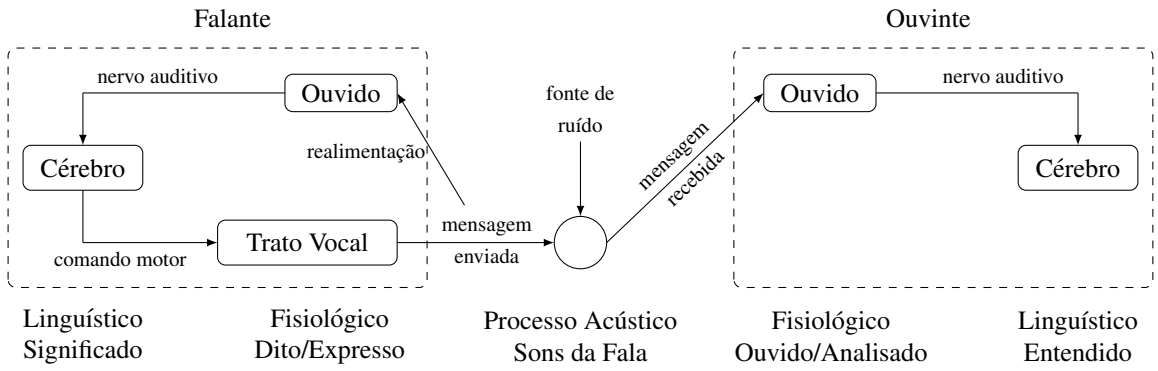


Figura 1 – Diagrama indicando os processos envolvidos na comunicação por voz. Adaptado de Niessen (2004), Furui (2000), Flanagan (2013) e Denes e Pinson (2015).

o trato vocal para gerar a manifestação acústica. O sinal acústico origina de flutuações da pressão de ar geradas pelo conjunto de ações no trato vocal (como vibrações quase periódicas das pregas vocais, ruídos turbulentos em estreitamentos, ou explosões pela liberação súbita de oclusões). Estas flutuações são, por sua vez, moduladas pelo movimento de articuladores (língua, lábios, palato mole, mandíbula) antes de serem radiados pela boca e/ou nariz.

A mensagem gerada é transmitida por um canal de comunicação – como o ar, por exemplo – até o ouvinte. Após ser detectada pelo ouvido, a informação é transmitida e interpretada pelo cérebro. A Figura 1 resume o processo de comunicação por um canal com ruído interferente.

Essa breve descrição da comunicação permite abstrair um modelo físico para o aparelho fonador para a produção de vogais, como proposto por Fant (1971), composto por uma fonte de energia e um conjunto de filtros modulantes como apresentado na Figura 2. O modelo, denominado fonte-filtro, a bateria representa a pressão pulmonar sobre os tubos da traqueia. A estrutura cartilaginosa que abriga dois lábios de ligamento e músculo denominadas pregas vocais. Nelas, a pressão subglótica é modulada e gera o fluxo de ar (sinal de excitação da fonte) que ressona nas cavidades superiores (filtros). Nos tubos da faringe, trato oral e nasal o fluxo de ar é modulado de acordo com os sons da fala (FLANAGAN, 2013).

O trato oral consiste de um tubo acústico com área da seção transversal não uniforme. O trato oral tem início nas pregas vocais na parte superior da traqueia e estende-se até os lábios. Em homens adultos o tubo oral tem cerca de 17 centímetros de comprimento e sua seção transversa é modificada pelo movimento dos articuladores ativos – lábios, mandíbula e língua (FLANAGAN, 2013).

O trato nasal permite um caminho auxiliar para a transmissão de som. Esse estende-se do véu palatino às narinas. No homem adulto a cavidade tem um comprimento de cerca de 12 cm. O acoplamento acústico entre as vias nasal e oral é controlado pela abertura no véu palatino. O

acoplamento nasal influencia substancialmente o caráter do som irradiado. Na produção de sons não nasais o véu palatino veda a entrada para a cavidade nasal (FLANAGAN, 2013).

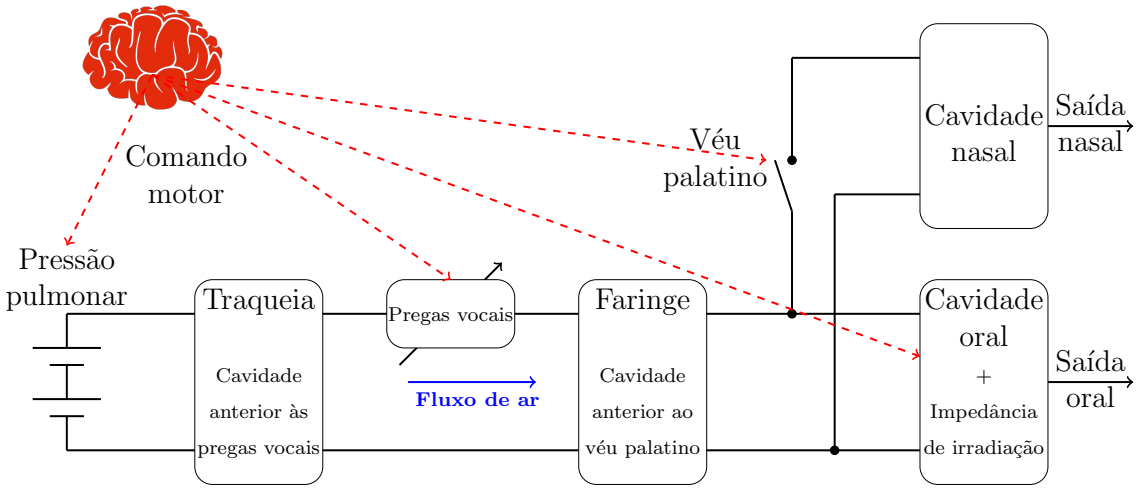


Figura 2 – Diagrama apresentando as etapas do modelo fonte-filtro para produção de vogais. Nesta modelagem a bateria representa a pressão pulmonar sobre os tubos da traqueia, as pregas vocais modulam a pressão subglótica gerando o fluxo de ar que excita os tubos da faringe, trato oral e nasal. Para a produção de outros sons (e.g., fricativas e plosivas), a fonte pode estar em outra região do trato vocal. Adaptado de Fant (1971).

Observados os processos envolvidos na produção da voz, é possível propor um modelo comportamental e biométrico individualizador que acopla os efeitos cognitivos ao modelo acústico fonte-filtro. Desta forma ter-se-iam os efeitos e condições de contorno descritos a seguir:

Efeito Cognitivo Efeito do sistema nervoso central e seu estado na produção da voz. Tal efeito é influenciado pela capacidade neuromotora do indivíduo, sua competência e desempenho na produção dos sons. Estes efeitos são limitados pela capacidade intelectual do indivíduo e pelos efeitos fisiológicos sobre o seu estado natural de vigília como fadiga, estresse ou uso de substância com efeito psicoativo (HOLLIEN, 1990, p. 223).

Efeito de Fonte De forma ampla, é a influência direta do modo de fonação na produção do som[1]. Tal efeito é oriundo dos parâmetros das pregas vocais, – dimensões e características mecânicas – e capacidade de ajuste muscular (FANT, 1971).

Efeito do Filtro Configuração espacial dos articulares do trato oral e nasal para produzir as ressonâncias que caracterizarão os sons da comunicação. Naturalmente o posicionamento dos articuladores possui limitações anatômicas do próprio falante (FANT, 1971).

Efeito do Meio Influência do meio de propagação e do instrumento sensibilizado pela onda acústica da voz. Esse fator modela as influências sobre a onda sonora desde o falante

[1] Entretanto, outras regiões do trato vocal podem funcionar como fonte de fluxo de ar turbulento, como, por exemplo, na produção de fricativas, onde a fonte é a região de estreitamento dos articuladores.

até o registro acústico. Um exemplo desta influência é a codificação GSM para transmissão telefônica (FURUI, 2000).

O modelo individualizador anteriormente descrito não abrange todas as variáveis presentes no registro acústico de um falante. Porém, pode abranger diversos aspectos capazes de influenciar as características medidas no sinal de voz que são abordados por Rose (2003), Campbell et al. (2009), French et al. (2010), Neustein e Patil (2011) e Rehder et al. (2015). Doravante, o presente trabalho adota como base elementos presentes nesse modelo abstrato de produção da voz que possui limitações.

3.2 Medição de Características de Registros Acústicos

Na Seção 3.1 buscou-se descrever sucintamente os princípios físicos e anatômicos que envolvem a produção da fala que são relevantes para o modelo acústico adotado no presente trabalho.

Esta seção apresentará as principais técnicas de medição utilizadas para a realização dos experimentos apresentados no presente projeto. Dentre as operações com sinais de voz destacam-se a detecção de atividade de voz, a medição da relação sinal-ruído e a decomposição *mel-cepstral* do sinal de voz[2].

O primeiro passo da análise de sinais da voz é sua divisão em quadros. De um registros acústico discreto $\mathbf{s}[t]$, representado por letra minúscula e em negrito, que contém T amostras, define-se o quadro no tempo t_0 como

$$\mathbf{s}_w[t,t_0] = \mathbf{s}[t] \cdot \mathbf{w}_r[t-t_0]. \tag{3.1}$$

Onde $\mathbf{s}_w[t,t_0]$ é o quadro de voz que começa no tempo t_0 e $\mathbf{w}_r[t]$ é uma janela retangular com t_w amostras sendo que $\mathbf{w}_r[t] = 1$ para $0 \leq t < t_w$ e $\mathbf{w}_r[t] = 0$ caso contrário. Em cada quadro assume-se a estacionariedade das características acústicas. Cada quadro de voz, que possuem comprimento de t_w amostras, são deslocadas entre si por um passo de t_s amostras, em geral $t_s \leq t_w$, como apresenta a Figura 3.

De cada quadro podem ser calculadas as matrizes de características – representadas por letras maiúsculas e em negrito. A extração de características converte um quadro de t_w amostras em um vetor de dimensão F. Assim uma matriz que representa os MFCC (*Mel-Frequency Ceps-*

[2] Ciente dos conceitos de voz e fala no modelo fonte-filtro, estes termos serão utilizados como sinônimos no restante do texto, exceto quando houver necessidade de diferenciação.

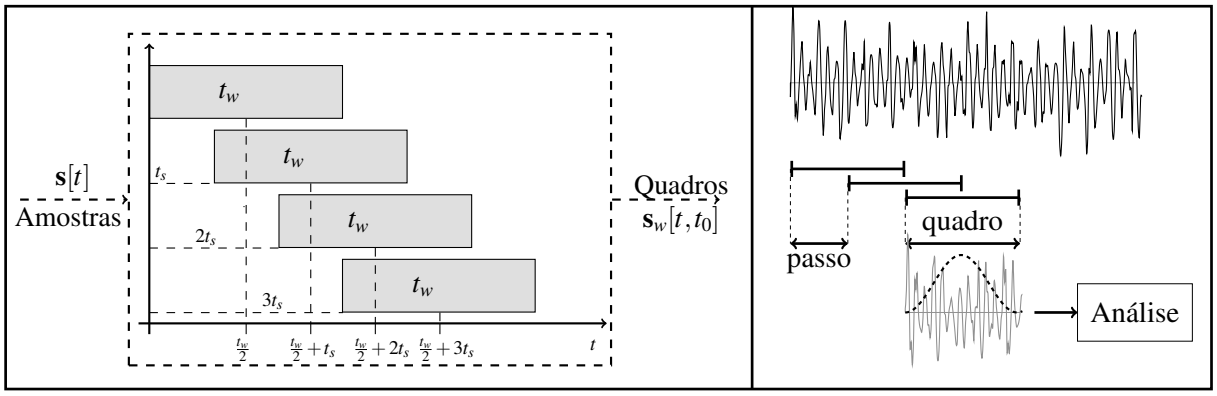


Figura 3 – Diagramação do processo de separação dos quadros de voz, indicando o comprimento da janela t_w, e o passo de tempo t_s.

tral Coefficients) $\mathbf{X}$ terá dimensão $F \times T$, onde T é numero de quadros de voz[3]. Doravante, o presente texto assume uma amostra de voz como um conjunto de quadros (elementos ou realizações) que são considerados independentes e igualmente distribuídos (IID)[4].

3.2.1 Detecção de Atividade de Voz

A detecção da atividade de voz (VAD - *Voice Activity Detection*) é a identificação dos trechos de um registro acústico onde a fala predomina sobre o ruído interferente. As regiões de fala são, categoricamente, os trechos de áudio onde o falante que desejamos estudar imprime seu registro acústico, ao passo que os trechos contrários podem conter registros de diferentes origens, como silêncio, ruído ambiente ou outros sinais de áudio sem interesse.

Nos experimentos apresentados nesta tese, a detecção de atividade de voz utiliza o método apresentado por Sohn et al. (1999)[5]. A abordagem assume que um trecho de voz está degradado por ruído aditivo não correlacionado[6].

A detecção da atividade de voz segue os passos apresentados na Figura 4, sendo que primeiramente o sinal de áudio $\mathbf{s}[t]$ é dividido em quadros para a estimação do espectro de potência a partir da Transformada Discreta de Fourier (DFT - *Discrete Fourier Transform*).

---

[3] É importante observar que o valor T indica o número de observações temporais de uma variável. Não significa, necessariamente, que o intervalo amostral no tempo de $\mathbf{s}[t]$ será o mesmo de sua matriz de características $\mathbf{X}$.

[4] Considera-se que variáveis aleatórias são IID se possuírem a mesma distribuição de probabilidade e forem mutuamente independentes. Essa premissa, porém, é pouco realista. Em uma amostra de voz, apesar de serem consideradas IID, as medidas obtidas de quadros adjacentes são correlacionadas.

[5] O método de Sohn et al. (1999) é amplamente utilizado para detecção de atividade de voz em algoritmos de compactação de sinais de áudio em telecomunicações. Uma implementação do método para MATLAB® está disponível em https://github.com/ImperialCollegeLondon/sap-voicebox (acessado em 14/01/2020).

[6] No trabalho de Sohn et al. (1999) não fica claro se "ruído aditivo não correlacionado" refere-se a ruído branco (AWGN - *Additive White Gaussian Noise*).

O espectro de potência pode ser representado pela matriz $\mathbf{S}_f = [\,\mathbf{s}_f[f,0], \mathbf{s}_f[f,1]), ..., \mathbf{s}_f[f,T-1]\,]$, onde $\mathbf{s}_f[f,t]$ é o espectro de potência ao longo da frequência f do t-ésimo quadro.

A partir do espectro de potência estima-se a variância do ruído (MARTIN, 2001) e a SNR – a *priori* e a *posteriori* –, minimizando o erro quadrático médio (MMSE - *Minimum Mean Square Error*) conforme descrito por Ephraim e Malah (1984).

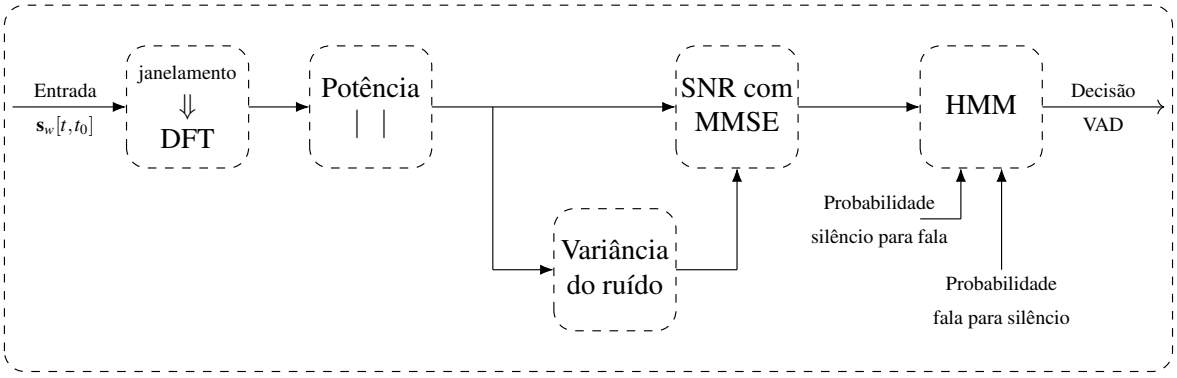


Figura 4 – Etapas do algoritmo para detecção de atividade de voz proposto por Sohn et al. (1999). O sinal $\mathbf{s}[t]$ é dividido em quadros para o cálculo da DFT, do espectro de potência estima-se a variância do ruído e a SNR minimizando o erro quadrático médio (MMSE). Estes parâmetros, juntamente com as probabilidades de transição entre fala para silêncio (e vice-versa), alimentam um HMM para tomada de decisão. Adaptado de (KIZHANATHAM, 2003, p. 10)

Para a etapa seguinte define-se como hipótese nula (H_0^{VAD}), que um determinado quadro de áudio não possui fala, e como hipótese alternativa (H_1^{VAD}), que o trecho de áudio possui fala e ruído, como estados em um modelo oculto de Markov (HMM - *Hidden Markov Model*) com as probabilidades de transição entre voz para silêncio (e vice-versa). Para tomada de decisão utiliza-se a razão de verossimilhança $\mathcal{L}(t)$ (ou o fator de Bayes) como

$$\mathcal{L}(t) \triangleq \frac{p(\mathbf{s}_f[f,t]|H_1^{VAD})}{p(\mathbf{s}_f[f,t]|H_0^{VAD})} = \frac{P(H_0^{VAD})}{P(H_1^{VAD})} \frac{P(H_1^{VAD}|\mathbf{s}_f[f,t])}{P(H_0^{VAD}|\mathbf{s}_f[f,t])} \begin{cases} < \xi & \Rightarrow H_0^{VAD} \\ \geq \xi & \Rightarrow H_1^{VAD} \end{cases} \tag{3.2}$$

onde ξ é o limiar de decisão. Como ferramenta forense, as técnicas de VAD separam, dos áudio questionado e padrão, o material que contém voz para ser utilizado na comparação.

Nos experimentos apresentados no Capítulo 4 o método de detecção de atividade de voz foi utilizado na etapa de pré-processamento para selecionar os trechos dos registros de áudio que serão utilizados nas tarefas de comparação de locutor.

3.2.2 Decomposição Mel-Cepstral

Os componentes mel-cepstrais ou MFCC (*Mel-Frequency Cepstral Coefficients*) são os parâmetros extraídos do sinal de voz mais utilizados em sistemas biométricos de verificação de

locutores. A transformada de Fourier expressa o sinal no domínio da frequência, e a transformação cepstral no domínio ciclos/hertz. Tal nomenclatura e outras similares para grandezas e operações no domínio cepstral foram propostas no trabalho de Bogert et al. (1963).

A escala mel, nome derivado da palavra melodia (*melody*), é uma escala logarítmica perceptual definida por Stevens et al. (1937)[7]. Na literatura, é possível encontrar diferentes propostas de transformação da frequência linear (f) na escala mel (m_{freq}), sendo mais comum a equação proposta por O'shaughnessy (1987)

$$m_{freq}(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right), \tag{3.3a}$$

donde

$$f(m) = 700 \cdot \left(e^{\frac{m_{freq}}{1127}} - 1\right). \tag{3.3b}$$

O objetivo da escala mel e outras, como a divisão em Barks, é escalonar e modelar as bandas críticas da audição. A separação das bandas críticas[8] pode ser realizada por diferentes formatos de filtros (triangulares, gammatone, etc), com ou sem superposição de banda (LYON et al., 2010).

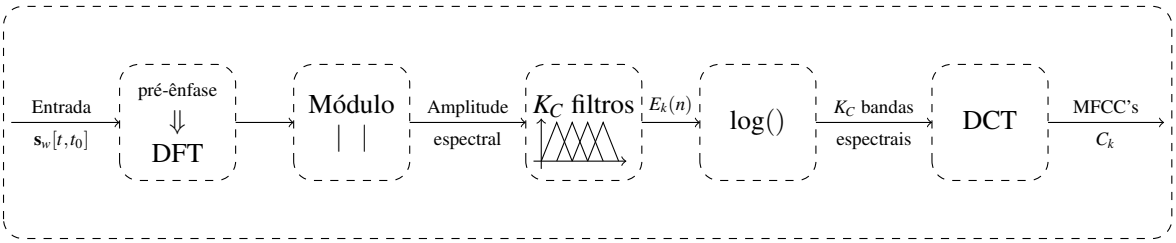


Figura 5 – Etapas na obtenção dos componentes mel-cepstrais com filtro triangular.

O cálculo do MFCC, conforme etapas apresentadas na Figura 5, começa com a correção da inclinação espectral de um sinal $\mathbf{s}[t]$ por meio de um filtro de pré-ênfase e sua separação em T quadros de comprimento de 25 ms intervalados por 10 ms. Para cada um dos $t \in 0, ..., T-1$ quadros é realizada a pré-enfase, o janelamento e, em seguida, calculada a transformada discreta de Fourier (DFT - *Discrete Fourier Transform*) e extraído o módulo.

Makhoul (1975) recomenda a aplicação de pré-enfase da forma $s_p[t] = s[t] - \alpha_p s[t-1]$. Empiricamente, utiliza-se $0,95 \leq \alpha_p \leq 1$[9]. Já o janelamento é utilizado para reduzir o espalhamento espectral.

[7] No presente texto utiliza-se a notação $\ln(\cdot)$ para o logaritmo natural e $\log_N(\cdot)$ para o logaritmo na base N. Por exemplo, $\log_{10}(\cdot)$ refere-se ao logaritmo na base 10.

[8] Bandas críticas são intervalos de frequência de aproximadamente uma terça de oitava ($\approx 19\%$) dentro dos quais ocorrem fenômenos de interação entre sons simultâneos. Estes fenômenos podem ser, entre outros, batimentos entre tons puros de mesma amplitude (ROEDERER, 2008, p. 42).

[9] O fator de correção ótimo (MAKHOUL, 1975) pode ser obtido por $\alpha_p^{Op} = \frac{\phi[1]}{\phi[0]}$, onde $\phi[t]$ é a função de autocorrelação de $s[t]$.

A Figura 6 ilustra os efeitos do processamento intermediário sobre o sinal no cálculo dos MFCC. No painel superior esquerdo tem-se o quadro de voz no domínio do tempo, logo abaixo a comparação entre o módulo espectral sem (painel central esquerdo) e com a aplicação da pré-ênfase de $\alpha_p = 0,975$ e a janela espectral de Hamming (painel inferior esquerdo).

O espectro de magnitude do sinal (referente à janela em análise) é então dividido em bandas por K_C filtros triangulares (vide painel superior direito da Figura 6), igualmente espaçados na escala mel e com sobreposição. Calcula-se a energia de cada banda e em seguida seu logaritmo – ficando $E_k(n)$ como o logaritmo da energia da k-ésima banda espectral –, indicado no painel central direito da Figura 6.

Em seguida, como apresentado no painel inferior direito da Figura 6, toma-se a transformada discreta cosseno do tipo 2 (DCT-II - *Discrete Cosine Transform* - type II) (TOGNERI; PULLELLA, 2011) como

$$c(k) = \sqrt{\frac{2}{C}}\beta(k)\sum_{n=0}^{N-1} E_k(n)cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \text{ para } k = 0,1,...,K_C-1, \tag{3.4}$$

onde K_C equivale ao número de filtros, N o número de amostras do sinal $E_k(n)$ e $\beta(k)$ é dado por

$$\beta(k) = \begin{cases} \frac{1}{\sqrt{2}}, & k = 0; \\ 1, & k = 1,2,...K_C-1. \end{cases} \tag{3.5}$$

A revisão de Togneri e Pullella (2011) sugere o uso de 13 filtros e as variações de primeira ordem Δc (delta cepstrum) e de segunda ordem $\Delta^2 c$ (delta-delta cepstrum) dos MFCC ao longo dos T quadros do áudio

$$\Delta c[t] = \frac{\sum_{p=1}^{P} p(c[t+p]-c[t-p])}{2\sum_{p=1}^{P} p^2} \tag{3.6}$$

$$\Delta^2 c[t] = \frac{\sum_{p=1}^{P} p(\Delta c[t+p]-\Delta c[t-p])}{2\sum_{p=1}^{P} p^2}. \tag{3.7}$$

A maioria dos autores utiliza $P = 1$ ou $P = 2$. Após a extração dos MFCC, a matriz com as medidas de MFCC do áudio $\mathbf{X} = \{\mathbf{x}(0), \mathbf{x}(1), ..., \mathbf{x}(T-1)\}$ é composta por T quadros com um número de características, ou dimensionalidade, definido pelo número de MFCCs e a presença ou não dos Δc e $\Delta^2 c$. Assim $\mathbf{X}$ terá T amostras de $3K_C$ dimensões na presença de Δc e $\Delta^2 c$.

A medida do MFCC está diretamente ligada à configuração espacial do trato vocal durante a produção da voz. As variações de primeira e segunda ordem modelam a dinâmica do trato vocal.

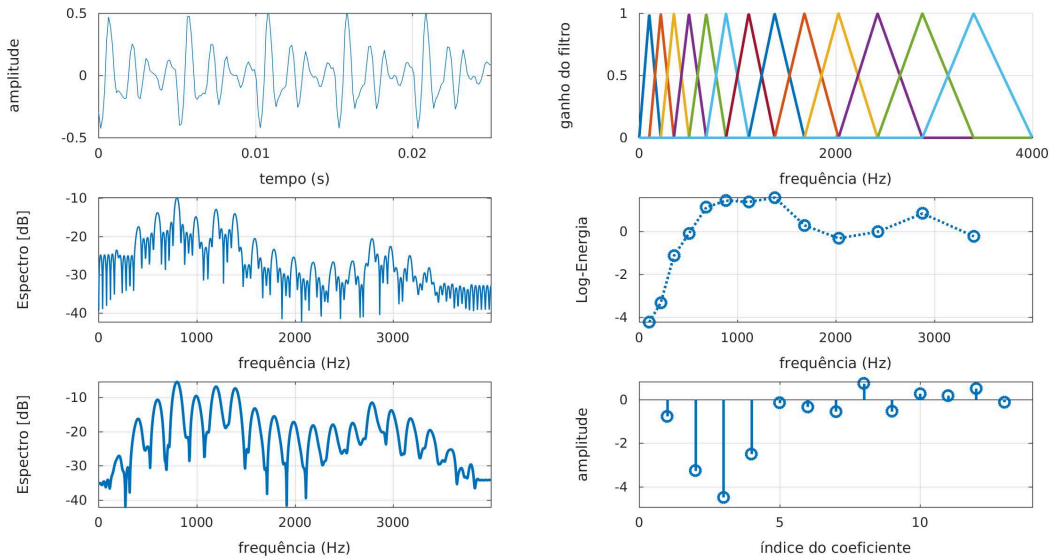


Figura 6 – Exemplo indicando as etapas do processamento para obtenção dos MFCC. No painel superior esquerdo tem-se o sinal de áudio, no painel logo abaixo tem-se o espectro de magnitude sem o janelamento e a pré-enfase, no painel inferior esquerdo o espectro de magnitude utilizando a janela de Hamming e pré-enfase. No painel superior direito têm-se os filtros triangulares, logo abaixo o logaritmo da energia para cada banda crítica, e, no painel inferior direito, os MFCC. Adaptado de (HANSEN; HASAN, 2015).

Em aplicações forenses, o MFCC é amplamente utilizado como característica para a modelagem de locutores em sistema de comparação automática. Um subproduto da presente pesquisa (SILVA et al., 2019), que pode ser conferida no apêndice, explorou o desempenho de diferentes características para cenários simulados de CFL. Em suma, comparado com outras características de tempo curto – como PNCC[10], TEOCC[11], PLP[12], RASTA-PLP[13], ZCPA[14], e SSCH[15] – o MFCC apresenta melhor desempenho.

No presente trabalho, o MFCC é utilizado como a medida acústica utilizada para a definição do modelo estatístico do locutor, tanto por modelo de mistura de gaussiana (GMM - *Gaussian Mixture Model*) quanto por vetor de informação (*i-vector*).

3.2.3 Medidas instantâneas de relação sinal-ruído

Nesta seção será descrito resumidamente o algoritmo S^{2}NR (*Spectrographic Signal-to-Noise Ratio*) proposto por Vieira et al. (2014) para a obtenção da relação sinal-ruído por quadro de voz. Basicamente, o S^{2}NR realiza o processamento da imagem espectrográfica para obter

[10] *Power Normalized Component Cepstrum* definido em Kim e Stern (2016).
[11] *Teager Enegy Operator Component Cepstrum*, definido em Niessen (2004, p. 167).
[12] *Perceptual Linear Predictive* definido em Hermansky (1990).
[13] *Representations Relative Spectra* definido em Hermansky e Morgan (1994).
[14] *Zero-Crossing with Peak Amplitude* definido em Kacur et al. (2012).
[15] *Subband Spectral Centroid Histograms* definido em Gajic e Paliwal (2001).

a relação sinal-ruído de um trecho acústico. O cálculo da S^2NR é aplicado na proposta de aprimoramento da robustez da CFL com sua aplicação mais detalhada na Seção 4.2.

A motivação para a utilização de medidas de SNR na CFL vem principalmente dos problemas de ordem prática, pois o nível de contaminação do áudio questionado afeta o desempenho dos métodos automáticos de comparação (SILVA et al., 2018a). A medida S^2NR foi escolhida por sua robustez e linearidade na faixa de contaminação em que se realizam os exames de CFL.

Os passos para o cálculo do S^2NR são apresentados na Figura 7 enquanto as figuras 8 e 9 apresentam exemplos de resultados do processamento intermediário[16].

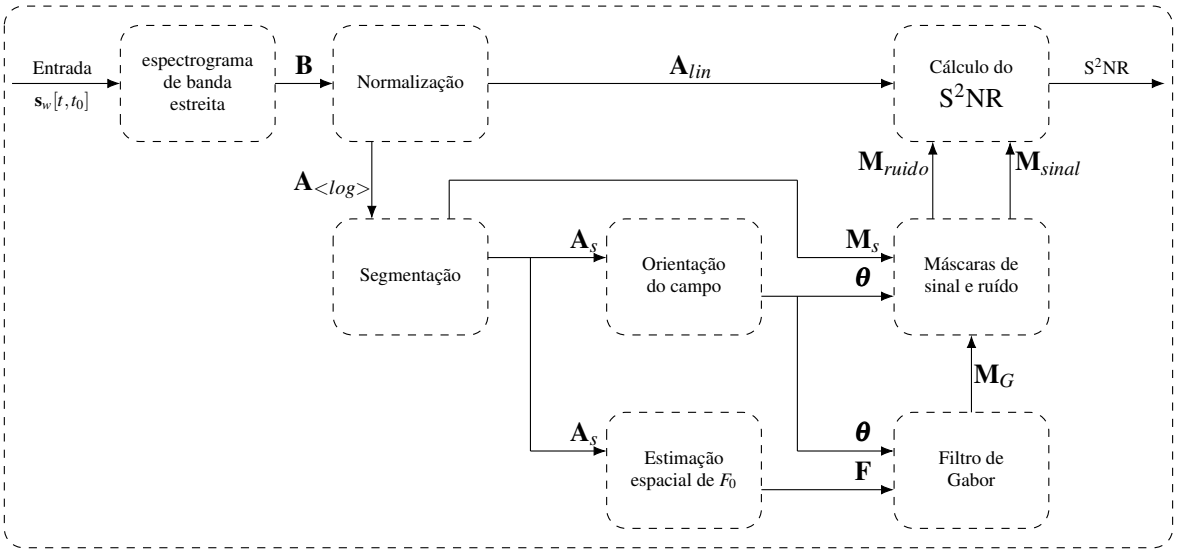


Figura 7 – Etapas do algoritmo S^2NR para obtenção da relação sinal-ruído por quadro $\mathbf{s}_w[t,t_0]$. Fonte: Adaptado de Vieira et al. (2014).

Dado um registro acústico de tempo discreto $\mathbf{s}[t]$, é possível obter seu espectrograma (magnitude) de banda estreita $\mathbf{B}$ em quadros consecutivos de 25 ms separados por passo de tempo de 10 ms. A imagem espectrográfica $\mathbf{B}(m,t)$ possui resolução de T quadros compostos por M pontos utilizados para o cálculo da transformada rápida de Fourier (FFT - *Fast Fourier Transform*). Na Figura 8 tem-se no painel superior esquerdo um trecho de áudio com a detecção da atividade de voz (linha tracejada preta) e o valor medido de S^2NR (linha tracejada laranja).

Os passos subsequentes são:

1. Após extração do espectrograma de banda estreita são realizadas a normalização linear e logarítmica. A linear gera $\mathbf{A}_{lin}(m,t)$ com valores entre 0 e 1 (vide painel superior

[16] Esta seção apresenta apenas um roteiro para o cálculo do S^2NR. Detalhes da implementação matemática, equacionamento, e heurísticas utilizadas são amplamente descritos e discutidos em (VIEIRA et al., 2014).

central da Figura 8). A logarítmica gera $\mathbf{A}_{<log>}(m,t)$, que é normalizada entre 0 e 1 (vide painel superior direito da Figura 8).

2. A segmentação busca traços de componentes harmônicos no espectrograma em blocos de 5*x*5 *pixels* que são classificados como sinal ou ruído. De $\mathbf{A}_{<log>}(m,t)$ deriva-se a imagem segmentada $\mathbf{A}_s(m,t)$, normalizada pela média e desvio padrão (vide painel inferior esquerdo da Figura 8), e em seguida se extrai a máscara $\mathbf{M}_S$ (painel central inferior da Figura 8). A máscara $\mathbf{M}_S$ contém apenas os blocos classificados como sinal a partir de um limiar σ_{th}.
3. De $\mathbf{A}_s(m,t)$, a partir de blocos 16*x*16 *pixels*, obtém-se as linhas harmônicas projetadas pelos ângulos de $\boldsymbol{\theta}(m,t)$ (painel inferior central da Figura 9) permitindo o cálculo da frequência entre as linhas harmônicas. Esses valores são armazenados na imagem $\mathbf{F}(m,t)$.
4. A máscara de Gabor $\mathbf{M}_G(m,t)$ (vide painel inferior direito da Figura 8) é calculada pelo filtro de Gabor com ângulo $\boldsymbol{\theta}(m,t)$ e frequência $\mathbf{F}(m,t)$.
5. A confiabilidade $\mathbf{R}$ (painel superior direito da Figura 9) é obtida a partir de uma combinação do gradiente de $\mathbf{A}_s(m,t)$ e da máscara de Gabor $\mathbf{M}_G(m,t)$. Da confiabilidade obtém-se a máscara de confiabilidade $\mathbf{M}_R(m,t)$, a partir de um limiar R_{th}, como apresentado no painel superior central da Figura 9.

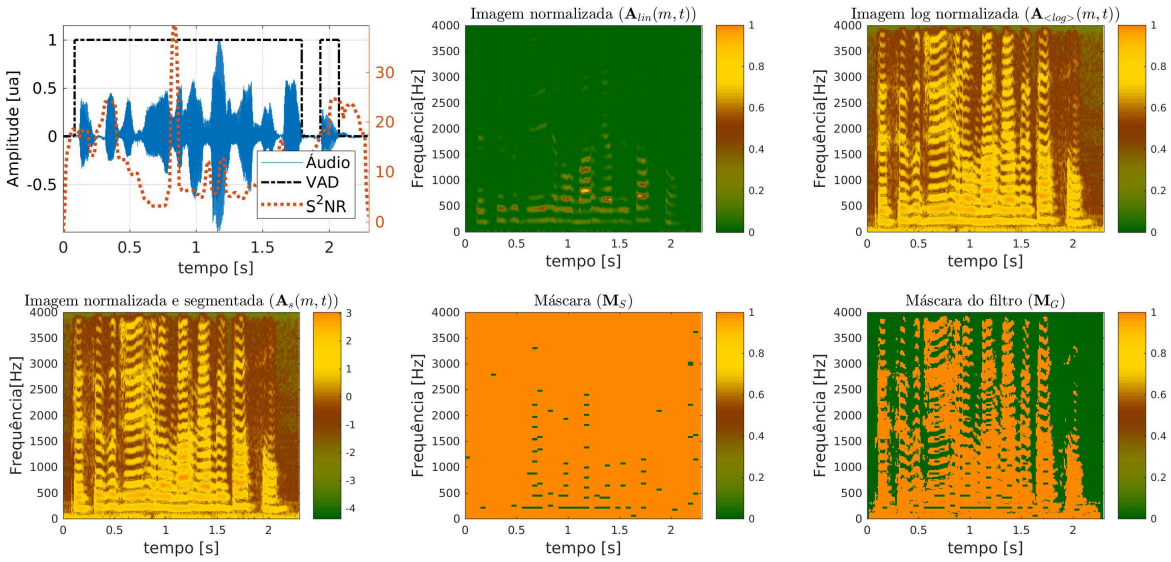


Figura 8 – Exemplo indicando as primeiras etapas do processamento para obtenção do S^2NR. No painel superior direito tem-se uma elocução de voz, o resultado da detecção da atividade de voz (linha tracejada preta) e o valor medido de S^2NR (linha tracejada laranja, eixo à direita em *dB*). Nos dois painéis à direita, têm-se, respectivamente os espectrogramas normalizados $\mathbf{A}_{lin}(m,t)$ (linear) $\mathbf{A}_{<log>}(m,t)$ (logarítmico). No painel inferior direito tem-se a imagem segmentada, na porção inferior central a máscara de segmentação $\mathbf{M}_S$ e à direita a máscara de Gabor $\mathbf{M}_G$.

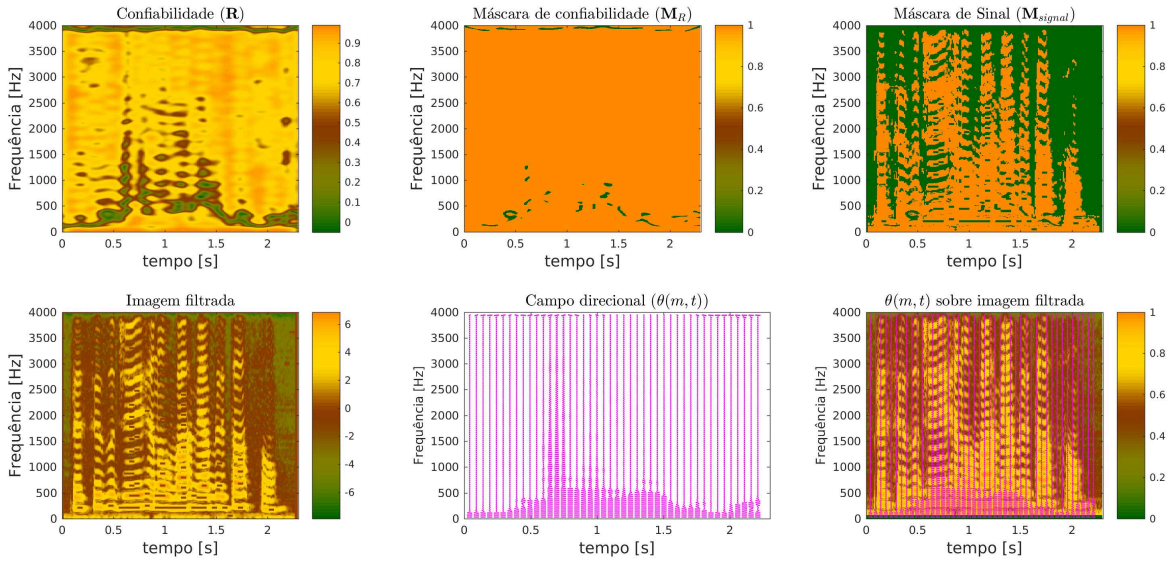


Figura 9 – Exemplo indicando as etapas finais do processamento para obtenção do S^{2}NR. No painel superior direito tem-se a matriz de confiabilidade $\mathbf{R}$ e, à sua direita, a máscara de confiabilidade $\mathbf{M}_R(m,t)$. No painel superior direito tem-se a máscara de sinal $\mathbf{M}_{sinal}$. Na porção inferior tem-se a imagem filtrada seguida da orientação ângulo $\boldsymbol{\theta}(m,t)$ e, no painel inferior direito, a orientação de ângulo sobre o espectrograma.

A máscara de sinal, apresentada no painel superior direito da Figura 9 é obtida da interseção entre as máscaras de Gabor, segmentação, e confiabilidade da forma $\mathbf{M}_{sinal} = \mathbf{M}_G \cap \mathbf{M}_R \cap \mathbf{M}_S$, sendo a máscara de ruído sua complementar. Assim, são obtidas as imagens de sinal $\mathbf{A}_{sinal} = \mathbf{M}_{sinal} \cap \mathbf{A}_{lin}$ e ruído $\mathbf{A}_{ruido} = \mathbf{M}_{ruido} \cap \mathbf{A}_{lin}$ que permitem calcular o valor de S^{2}NR para cada quadro da forma

$$\mathbf{S}_{sinal}[t] = \sum_{m=0}^{\frac{M}{2}-1} [\mathbf{A}_{sinal}(m,t)]^2, \qquad \mathbf{S}_{ruido}[t] = \sum_{m=0}^{\frac{M}{2}-1} [\mathbf{A}_{ruido}(m,t)]^2, \text{ e} \tag{3.8a}$$

$$\mathbf{S^2NR}[t] = 10\log_{10}\left(\frac{\mathbf{S}_{sinal}(t)}{\mathbf{S}_{ruido}(t)}\right). \tag{3.8b}$$

Sendo $\mathbf{S}_{sinal}[t]$ a estimativa de magnitude do sinal, $\mathbf{S}_{ruido}[t]$ a estimativa de magnitude do ruído, e $\mathbf{S^2NR}[t]$ a relação sinal-ruído ao longo dos quadros de voz.

Para o trabalho apresentado nesta tese, a medida espectrográfica de relação sinal-ruído, i.e. o S^{2}NR, permite estimar a contaminação para cada quadro de voz e parear a verificação de locutor. A comparação pareada, utilizando conjuntos *fuzzy*, processa separadamente os quadros de voz de acordo com a contaminação[17].

É importante observar que o método S^{2}NR foi desenvolvido para medidas em trechos vozeados, em especial em vogais. Vieira et al. (2014) coloca que o uso em fala corrida pode ser

[17] Uma implementação para MATLAB® do S^{2}NR pode ser obtida em https://github.com/jsansao/s2nr (acessado em 14/01/2020).

influenciado por fatores linguísticos, uma vez que trechos não vozeados poderiam apresentar medidas imprecisas de SNR. Para mitigar estes efeitos foram realizados ajustes no algoritmo S^2NR durante a realização dos experimentos apresentados neste trabalho. Os experimentos utilizam de registros de áudio intencionalmente contaminados por diferentes tipos de ruído[18]. Durante o processo de contaminação ajustou-se o S^2NR para apresentar, na media, a mesma SNR dos áudios contaminados.

3.3 Codificação do Sistema de Telefonia Móvel no Brasil

No cenário atual da comparação forense de locutor, a variável que carrega maior incerteza está relacionada ao dispositivo de captura e ao canal de transmissão. Sabe-se que, na grande maioria dos casos de CFL, o áudio questionado é oriundo de uma interceptação telefônica ou de áudios de aplicativos com função PTT (*Push-to-Talk*). Nesses casos, o dispositivo de captura costuma ser um microfone com tecnologia MEMS (*Micro-ElectroMechanical Systems*) e o canal de comunicação é o sistema de telecomunicações.

O sistema de comunicação por voz brasileiro utiliza o espectro faixa estreita (NB - *Narrow-Band*) entre 300 Hz e 3.400 Hz. Tanto a comunicação por telefonia digital fixa como as móveis, – GSM e 3G – utilizam esta banda de áudio com taxa de amostragem de 8000 Hz.

O padrão para a telefonia fixa é o ITU-T G.711 PCM (*Pulse-Code Modulation*) que foi proposto por ITU (1988)[19]. Esse foi o primeiro padrão de codificação da fala que opera entre 300 Hz e 3.400 Hz. Nesta norma definem-se os padrões de conversão/compactação da quantificação uniforme – de 14 bits por amostra – para quantificações logarítmicas de 8 bits por amostra, sendo elas a lei-A (Europa) e a lei-μ (EUA e Japão). O padrão ITU-T G.711 PCM é uma codificação baseada na forma de onda, e opera à taxa de transmissão de 64 kbit/s com atraso algorítmico de um quadro (20 ms).

O sistema GSM é especificado pela União Internacional de Telecomunicação (ITU - *International Telecommunication Union*) e sua codificação é baseada em excitação regular por pulsos (RPE - *Regular Pulse Excitation*) e predição de longo termo (LTP - *Long Term Prediction*). O comprimento do quadro é de 20 ms (160 amostras à taxa de amostragem de 8 kHz), e cada quadro de fala é dividido em 4 sub-quadros de 5 ms (ITU, 1991).

[18] A contaminação intencional realizada nos experimentos é descrita no Capítulo 4.

[19] Segundo informações disponibilizadas pela ITU, a última emenda desta norma ocorreu em novembro de 2009.

Nessa codificação, a análise de predição linear e a predição de longo termo são realizadas em todo quadro, enquanto a análise de excitação é realizada nos sub-quadros. O GSM melhorado (EFR - *Enhanced Full Rate*), ou GSM 06.60, foi definido na norma ITU (2000b). Baseia-se em ACELP (CELP algébrico) e opera a 12,2 kb/s[20].

O filtro de longo termo, também denominado filtro de síntese de frequência fundamental é dado por

$$\frac{1}{B(z)} = \frac{1}{1 - g_p z^{-T_p}}, \tag{3.9}$$

onde T_p é o atraso de frequência fundamental e g_p o respectivo ganho. O filtro de síntese de frequência fundamental é implementado junto com a abordagem que utiliza livro de códigos (*codebooks*) adaptativos (vide Figura 10).

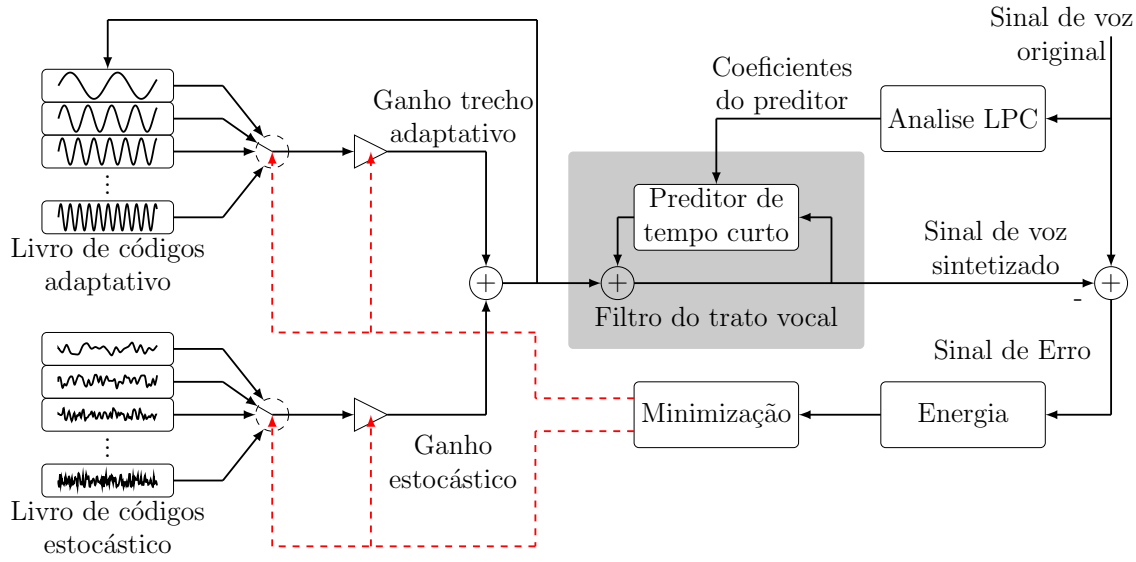


Figura 10 – Predição linear com excitação por código (CELP - *Code-excited linear prediction*). Adaptado de (HAUENSTEIN, 2003).

A comunicação por voz no sistema 3G – para aplicativos PPT – baseia-se no padrão ITU (2000a) que define a codificação multi-taxas adaptativa (AMR - *Adaptive Multi Rate*) de banda estreita, também baseada no ACELP. A diferença básica da codificação utilizada no GSM para o 3G (AMR) reside na flexibilidade para utilização de 8 diferentes taxas de bits, variáveis entre 4,75, 5,15, 5,9, 6,7, 7,4, 7,95, 10,2 e 12,2 kb/s. As tabelas de alocação de bits podem ser encontradas na norma ITU (2000a).

A predição linear com excitação por código, ou CELP (*Code-Excited Linear Prediction*) é um algoritmo de codificação baseado em modelo LPC (*Linear Predictive Coding*) como apresentado no diagrama da Figura 10. Como fonte de excitação para o modelo LPC do trato

[20] O pares de linhas espectrais (LSP - *Line Spectral Pairs*) são representações dos coeficientes LPC (*Line Predictive Coding*) por polinômios auto-recíprocos. Os LSP são mais estáveis à quantização e, por isso, utilizados na transmissão pelo canal telefônico (SUGAMURA; ITAKURA, 1981).

vocal o CELP utiliza um livro de códigos (*codebook*). A codificação ACELP (*Algebraic* CELP) ou CELP algébrico tem o livro de códigos definido algebricamente. Segundo ITU (2000b) a codificação ocorre em janelas de 20 ms (160 amostras) e utiliza LPC de ordem 10 e janela de Hamming.

Nos experimentos realizados nesta tese, a codificação do sistema de comunicação por voz foi utilizada para emular parte do efeito do canal telefônico. Os procedimentos de codificação e decodificação foram realizados utilizando bibliotecas *Sound eXchange*[21], para a codificação GSM, e *Opus*[22], para a codificação de voz em 3G.

3.4 Comparação Automática de Locutor

Na Seção 3.2 foram apresentadas as principais técnicas de extração de informação de registros acústicos utilizadas no presente trabalho. Nesta seção serão apresentadas as duas principais abordagens para comparação automática de locutores: a primeira baseada em mistura de gaussianas e a segunda em vetores de informação.

O princípio da comparação biométrica de locutores é confrontar medidas acústicas extraídas dos sinais de voz e inferir sobre sua compatibilidade. Essa comparação utiliza como suporte uma base de dados composta por diferentes amostras de voz. O objetivo da base de dados é fornecer informação sobre a ocorrência e variabilidade das medidas acústicas e como estas são mais relevantes no confronto dos sinais de voz. Uma comparação consistente possui uma base representativa da população e permite inferir sobre a compatibilidade entre dois sinais de voz.

As amostras de voz podem ser representadas diretamente pelos dados não tratados, que são suas medidas acústicas, ou por modelos paramétricos – caso deste capítulo – que são mais compactos, tanto para os locutores individuais quanto para a base de fundo.

3.4.1 Modelagem e Classificação por GMM-UBM

A metodologia GMM-UBM (*Gaussian Mixture Models-Universal Background Model*) com base na adaptação de locutores foi proposta por Reynolds et al. (2000). O método utiliza uma base de locutores para criar um modelo universal de fundo (UBM - *Universal Background Model*) paramétrico. O UBM é uma mistura de gaussianas (GMM - *Gaussian Mixture Model*) das características (MFCC), e cada locutor é modelado a partir de uma adaptação do UBM.

[21] Mais informações sobre a utilização da biblioteca *Sound eXchange* estão disponíveis em http://sox.sourceforge.net/ (acessado em 14/01/2020).

[22] Mais informações sobre a utilização da biblioteca *Opus Interactive Audio Codec* estão disponíveis em https://www.opus-codec.org/ (acessado em 14/01/2020).

Segundo Hansen e Hasan (2015), esta modelagem é robusta por ser independente de texto, e pelos modelos de locutores adaptados do UBM serem mais confiáveis que os treinados diretamente de cada locutor.

A modelagem paramétrica consiste em obter uma função densidade de probabilidade (FDP) que possua a máxima verossimilhança com os dados que se deseja modelar. A função densidade de probabilidade é aproximada por um modelo de mistura de gaussianas (GMM - *Gaussian Mixture Model*), ou seja, a soma ponderada de funções de distribuição normal.

As etapas da verificação de locutor utilizando a metodologia GMM-UBM são apresentadas na Figura 11. Para obter o GMM, considere que de um registro de áudio são extraídos os trechos de fala utilizando detecção por atividade de voz (VAD - *Voice Activity Detection*), proposta por Sohn et al. (1999) e, em seguida calcula-se a matriz de MFCC $\mathbf{X} = [\mathbf{x}[0], \mathbf{x}[1], ..., \mathbf{x}[T-1]]$, onde $\mathbf{X} \in \mathbb{R}^{F \times T}$ e $\mathbf{x}[t] \in \mathbb{R}^{F \times 1}$.

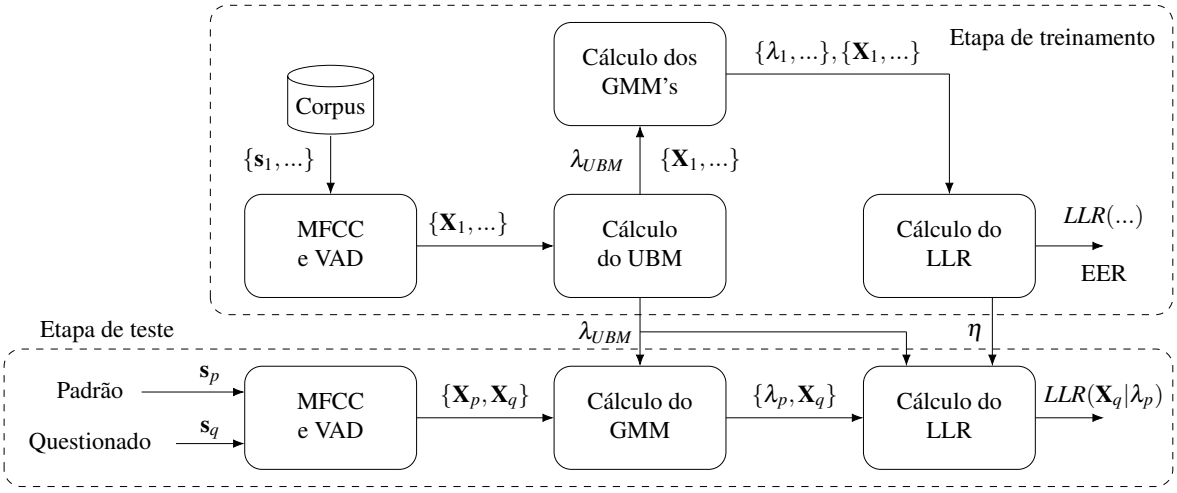


Figura 11 – Etapas da comparação de locutores utilizando GMM-UBM. Primeiramente as características MFCC são extraídas dos registros de voz que são selecionados pelo algoritmo de detecção de atividade de voz (VAD). Destas características são modelados o UBM e as GMM de cada locutor. Na etapa final a pontuação calculada pelo LLR.

O GMM λ de um locutor modela a matriz $\mathbf{X}$ a partir de um conjunto de G distribuições normais, com média $\boldsymbol{\mu}_g \in \mathbb{R}^{F \times 1}$ e matriz de variância $\boldsymbol{\Sigma}_g \in \mathbb{R}^{F \times F}$ como:

$$\lambda = \{p_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\} \text{ para } g = 1, 2, ..., G, \tag{3.10}$$

onde p_g é o peso ou fator de ponderação de cada gaussiana na mistura e G número de distribuições (REYNOLDS et al., 2000).

A probabilidade da matriz de MFCC (ou amostra) $\mathbf{X}$ ter sido gerada pelo modelo λ é definida por Reynolds et al. (2000) como

$$p(\mathbf{X}|\lambda) = \sum_{g=1}^{G} \frac{p_g}{\sqrt{(2\pi)^F |\Sigma_g|)}} \cdot e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_g)\Sigma_g^{-1}(\mathbf{X}-\boldsymbol{\mu}_g)^\top}, \tag{3.11}$$

Para ajustar o modelo λ, utiliza-se o algoritmo de maximização da expectância (EM - *Expectation–Maximization*), ajustando o modelo paramétrico do locutor λ aos dados $\mathbf{X}$, que é a matriz de MFCC. O algoritmo EM é um método iterativo para encontrar a máxima verossimilhança (ML - *Maximum Likelihood*) de parâmetros em modelos estatísticos, onde o modelo depende das variáveis observadas (DEMPSTER et al., 1977).

As etapas do algoritmo EM alternam entre os passos:

- cálculo da expectância (passo E), quando se utiliza uma função para o cálculo do logaritmo da verossimilhança avaliada a partir do modelo presente; e
- maximização (passo M), que calcula os parâmetros que maximizam a log-verossimilhança até obter-se o modelo de máxima verossimilhança.

Se considerarmos, por simplificação, que os dados $\mathbf{X}$ são independentes entre si (REYNOLDS et al., 2000), a probabilidade de observação do conjunto $\mathbf{X}$ dado o modelo λ é

$$p(\mathbf{X}|\lambda) = p(\mathbf{x}[0],...,\mathbf{x}[T-1]|\lambda) = p(\mathbf{x}[0]|\lambda)...p(\mathbf{x}[T-1]|\lambda) = \prod_{t=0}^{T-1} p(\mathbf{x}[t]|\lambda). \quad (3.12)$$

O logaritmo da verossimilhança $\mathcal{L}\{\lambda|\mathbf{X}\}$ pode ser obtido como

$$\mathcal{L}\{\lambda|\mathbf{X}\} = \ln\left(\prod_{t=0}^{T-1}\sum_{g=1}^{G} p_g N(\mathbf{x}[t],\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right) = \sum_{t=0}^{T-1}\ln\left(\sum_{g=1}^{G} p_g N(\mathbf{x}[t],\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right). \quad (3.13)$$

A ideia básica do algoritmo de maximização da verossimilhança consiste em obter, a partir de um modelo inicial λ^i, um novo modelo λ^{i+1} onde $\mathcal{L}\{\lambda^{i+1}|\mathbf{X}\} \geq \mathcal{L}\{\lambda^i|\mathbf{X}\}$, até atingir um limiar de convergência. Para o cálculo no modelo $i+1$ são utilizadas as condições descritas por Reynolds et al. (2000), que resultam nas equações a seguir, e garantem o incremento monotônico do modelo:

$$p_{i+1} = \frac{1}{N}\sum_{T=0}^{T-1} p(i|\mathbf{x}[t],\lambda^i), \quad (3.14)$$

$$\boldsymbol{\mu}_{i+1} = \frac{\sum_{T=0}^{T-1} p(i|\mathbf{x}[t],\lambda^i)\mathbf{x}[t]}{\sum_{t=0}^{T-1} p(i|\mathbf{x}[t],\lambda^i)}, \quad (3.15)$$

$$\boldsymbol{\sigma}_{i+1}^2 = \frac{\sum_{t=0}^{T-1} p(i|\mathbf{x}[t],\lambda^i)\mathbf{x}[t]^2}{\sum_{t=0}^{T-1} p(i|\mathbf{x}[t],\lambda^i)} - \boldsymbol{\mu}_i^2. \quad (3.16)$$

A probabilidade da mistura i, dados o vetor de dados $\mathbf{X}$ e o modelo de mistura de gaussianas λ^i, pode ser calculada como:

$$p(i|\mathbf{x}[t],\lambda^i) = \frac{p_i N(\mathbf{x}[t],\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{g=1}^{G} p_g N(\mathbf{x}[t],\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)}, \quad (3.17)$$

sendo que a abordagem GMM-UBM de Reynolds et al. (2000) utiliza matriz de variância Σ_g diagonal pelos seguintes motivos: a matriz de variância diagonal é mais eficiente, em termos computacionais; uma GMM que utiliza matriz de variância completa pode ser modelada por um conjunto maior de GMM's com variância diagonal; e, empiricamente, a modelagem por matriz de variância diagonal supera a modelagem por matriz de variância completa (REYNOLDS et al., 2000).

A modelagem de cada locutor, na metodologia GMM-UBM, é realizada por meio da estimação por máximo a *posteriori* (MAP) a partir do UBM e dos momentos de ordem zero, de primeira ordem e de segunda ordem das equações 3.18 a 3.20

$$n_i = \sum_{t=0}^{T-1} p(i|\mathbf{x}[t], \lambda^i), \tag{3.18}$$

$$E(\mathbf{X}) = \frac{1}{n_i} \sum_{t=0}^{T-1} p(i|\mathbf{x}[t], \lambda^i)\mathbf{x}[t], \tag{3.19}$$

$$E(\mathbf{X}^2) = \frac{1}{n_i} \sum_{t=0}^{T-1} p(i|\mathbf{x}[t], \lambda^i)\mathbf{x}^2[t], \tag{3.20}$$

onde $\mathbf{x}^2[t]$ é a diagonal da matriz $\mathbf{x}[t]\mathbf{x}^\mathsf{T}[t]$. A partir das estatísticas dos momentos calculam-se os parâmetros $\hat{p}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ de cada locutor como

$$\hat{p}_i = \left[\frac{\alpha_i^p n_i}{T} + (1 - \alpha_i^p) p_i \right] \gamma, \tag{3.21}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\mathbf{X}) + (1 - \alpha_i^m)\boldsymbol{\mu}_i, \tag{3.22}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\mathbf{X}^2) + (1 - \alpha_i^v)(\boldsymbol{\sigma}_i^2 - \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2, \tag{3.23}$$

onde $\hat{p}$, $\hat{\boldsymbol{\mu}}$, e $\hat{\boldsymbol{\Sigma}}$ são, respectivamente os pesos, média e variância do locutor modelado. Nas equações 3.21 à 3.23 os valores $\{\alpha_i^p, \alpha_i^m, \alpha_i^v\}$ referem-se a fatores de adaptação. Na prática,

$$\alpha^p = \alpha^m = \alpha^v = \frac{n_i}{n_i + r}, \tag{3.24}$$

onde r é um fator de relevância sugerido por Reynolds et al. (2000) como $r = 16$.

No contexto forense tem-se a amostra questionada $\mathbf{X}_Q$, a amostra padrão $\mathbf{X}_P$ ou seu modelo λ_P e uma base de referência $\mathbf{X}_{UBM}$ ou seu modelo λ_{UBM}[23]. A utilização do UBM permite

[23] As amostras questionada e padrão são conjuntos de medições acústicas da voz. A amostra $\mathbf{X}_{UBM}$, e o respectivo modelo de fundo λ_{UMB}, são obtidos a partir de um banco de vozes de referência, que podem ser amostras de áudio padrão de diferentes locutores.

avaliar a ocorrência das características e melhora o desempenho na comparação baseada em GMM.

Na metodologia GMM-UBM tem-se como variável de decisão uma pontuação (*score*) baseada na razão de verossimilhança obtida entre o modelo de locutor comparado (λ_P) e o UBM (λ_{UBM}) que, para fins de formulação, é calculada logaritmicamente (LLR *log-likelihood ratio*). A estatística $LLR(\mathbf{X_Q})$ é da forma

$$LR(\mathbf{X}_Q) = \frac{p(\mathbf{X}_Q|\lambda_P)}{p(\mathbf{X}_Q|\lambda_{UBM})} \quad \xrightarrow{\ln(\cdot)} \quad LLR(\mathbf{X}_Q) = \ln\left(\frac{p(\mathbf{X}_Q|\lambda_P)}{p(\mathbf{X}_Q|\lambda_{UBM})}\right), \tag{3.25}$$

onde $\mathbf{X}_Q$ representa a matriz de MFCC de um áudio questionado e o valor da estatística $LLR(\mathbf{X}_Q) \in [-\infty, \infty]$.

Para comparação de amostras de voz, Reynolds et al. (2000) propõem que no cálculo da Equação 3.25 seja utilizada a média das verossimilhanças computadas em cada um dos T quadros de voz, que toma a forma

$$\ln\left(p(\mathbf{X}_Q|\lambda_P)\right) = \frac{1}{T}\sum_{t=0}^{T-1} \ln\left(p(\mathbf{x}_Q[t]|\lambda_P)\right). \tag{3.26}$$

A média sobre os quadros de voz foi proposta para compensar o efeito da duração dos trechos de fala e permite calcular a estimativa por intervalo sobre a média, pois o logaritmo da razão de verossimilhança passa a ser

$$LLR(\mathbf{X}_Q) = \ln\left(\frac{p(\mathbf{X}_Q|\lambda_P)}{p(\mathbf{X}_Q|\lambda_{UBM})}\right) = \frac{1}{T}\sum_{t=0}^{T-1} \hat{\mathbf{x}}_Q[t], \tag{3.27}$$

onde

$$\hat{\mathbf{x}}_Q[t] = \left[\ln\left(p(\mathbf{x}_Q[t]|\lambda_P)\right) - \ln\left(p(\mathbf{x}_Q[t]|\lambda_{UBM})\right)\right], \tag{3.28}$$

e $LLR(\mathbf{X}_Q)$ é a média de $\hat{\mathbf{x}}_Q[t]$.

Tecnicamente, a premissa de que os dados da amostra de voz $\mathbf{X}_Q$ são independentes e identicamente distribuídos (IID), como apresentado na Equação 3.26, é muito frágil, pois ao longo dos quadros de voz as características acústicas são correlacionadas (MORRISON, 2011a). Dessa forma, a estatística $LLR(\mathbf{X}_Q)$ passa a ter o significado de pontuação (*score*), calculada a partir das amostras padrão e questionada, que indiscutivelmente não é uma razão de verossimilhança[24].

Um dos aprimoramentos propostos nesta tese recai sobre a Equação 3.26. Sobre o valor de $LLR(\mathbf{X}_Q)$ foi aplicado o intervalo de evidência com objetivo de melhorar a confiabilidade da metodologia GMM-UBM em condições que emulam a CFL, como descrito mais detalhadamente na Seção 4.1.

[24] Uma implementação para MATLAB® da metodologia GMM-UBM pode ser encontrada na *MSR Identity Toolbox* distribuída pelo sitio https://www.microsoft.com/en-us/research/publication/msr-identity-toolbox-v1-0-a-matlab-toolbox-for-speaker-recognition-research-2/ (acessado em 14/01/2020).

3.4.2 Modelagem e Classificação por Vetor de Informação

A modelagem por vetor de informação (*i-vectors*) (KENNY, 2012; KENNY et al., 2005; KENNY et al., 2008) é um aprimoramento da GMM-UBM e modela um locutor por um supervetor extraído do UBM. A metodologia *i-vector* também permite agregar a variabilidade dos locutores e dos canais de captação (transmissão) a partir da análise fatorial conjunta (JFA - *Joint Factor Analysis*) (DEHAK et al., 2011).

A Figura 12 apresenta resumidamente as etapas da metodologia *i-vector*. A etapa de treinamento utiliza um corpus (base de dados) com S locutores sendo que cada locutor possui C gravações (ou canais), resultando num conjunto $\mathcal{S} = \{\mathbf{s}_1, ...\mathbf{s}_{S\times C}\}$ com $S \times C$ vetores de áudio. De cada registro de áudio são extraídos os trechos com fala utilizando detecção por atividade de voz (VAD - *Voice Activity Detection*) proposta por Sohn et al. (1999) e, em seguida, a matriz de MFCC $\mathbf{X}$, de dimensionalidade $F \times T$. Os áudios da base de dados $\mathcal{S}$ são combinados para o cálculo do UBM com G gaussianas, gerando o modelo λ_{UBM}.

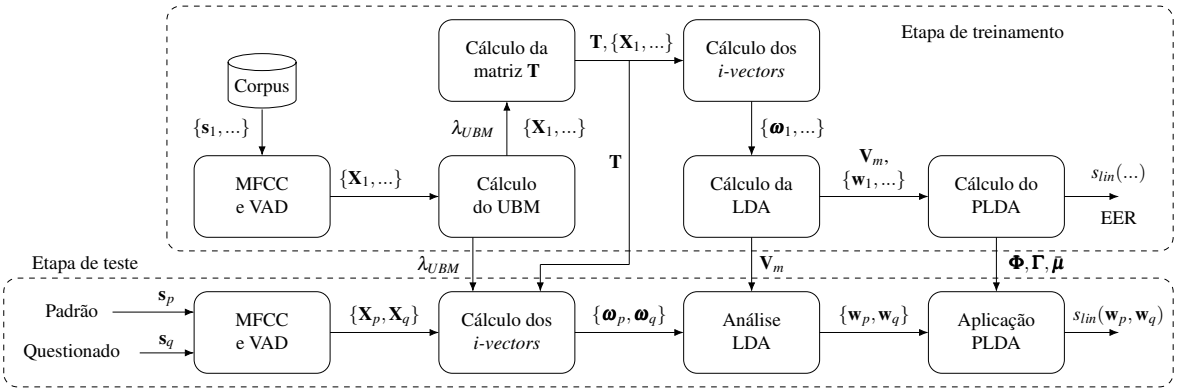


Figura 12 – Etapas da comparação de locutores utilizando *i-vector*. Primeiramente as características MFCC são extraídas dos registros de voz que são selecionados pelo algoritmo de detecção de atividade de voz (VAD). Destas características são modelados o UBM e a matriz de variabilidade total $\mathbf{T}$ e extraídos os *i-vector*s. Na etapa final a dimensionalidade é reduzida pela LDA e a pontuação calculada pela PLDA.

O passo seguinte é a estimação da matriz de variabilidade total $\mathbf{T}$. Considere um supervetor $\mathbf{m}_{ivec}$, que representa as informações presentes em um registro de voz de um locutor, modelado por

$$\mathbf{m}_{ivec} = \tilde{\boldsymbol{\mu}} + \mathbf{T}\boldsymbol{\omega}, \tag{3.29}$$

onde $\tilde{\boldsymbol{\mu}}$ é um supervetor independente do locutor e do canal, extraído das médias $\boldsymbol{\mu}_g$ do UBM, $\mathbf{T}$ é uma matriz retangular de baixo posto e $\boldsymbol{\omega}$ é o *i-vector* (vetor de informação) que representa o locutor. O *i-vector* é uma variável aleatória oculta que pode ser definida pela sua distribuição de probabilidade a *posteriori* condicionada às estatísticas de Baum-Welch (KENNY, 2012; DEHAK et al., 2011).

A modelagem por *i-vector* projeta a amostra de voz de um locutor em um vetor de alta dimensionalidade, no caso $GF \times 1$. O vetor de informação $\boldsymbol{\omega}$ passa a ser o modelo do locutor a ser testado. O cálculo do *i-vector* parte de uma matriz de MFCC com T quadros de voz $\mathbf{X} = [\mathbf{x}[0], \mathbf{x}[1], ..., \mathbf{x}[T-1]]$ – onde é $\mathbf{x}[t]$ tem dimensão $F \times 1$ – e um UBM λ_{UBM} com G gaussianas de dimensionalidade F, resultando nas estatísticas de Baum-Welch para cada uma das G gaussianas

$$\mathbf{n}_g = \sum_{t=0}^{T-1} P(g|\mathbf{x}[t], \lambda_{UBM}), \text{ e} \tag{3.30}$$

$$\tilde{\mathbf{f}}_g = \sum_{t=0}^{T-1} P(g|\mathbf{x}[n], \lambda_{UBM})(\mathbf{x}[t] - \boldsymbol{\mu}_g), \tag{3.31}$$

onde $\mathbf{n}_g$ e $\tilde{\mathbf{f}}_g$ são, respectivamente, os vetores com as estatísticas de ordem zero e de primeira ordem centralizadas, e $\boldsymbol{\mu}_g$ é a média da g-ésima gaussiana do UBM.

Um dos aprimoramentos propostos consiste na aplicação de lógica *fuzzy*, e altera o cálculo das estatísticas de Baum-Welch das equações 3.30 e 3.30. A alteração inclui a pertinência de cada quadro $\mathbf{x}[t]$ com um conjunto *fuzzy* relacionado à relação sinal-ruído[25].

A partir de $\mathbf{n}_g$ e $\tilde{\mathbf{f}}_g$, associadas à g-ésima gaussiana, calculam-se a matriz $\mathbf{N}_u$ e o supervetor $\tilde{\mathbf{f}}_u$ relacionados com o u-ésimo áudio de $\mathcal{S}$. A matriz $\mathbf{N}_u$ é diagonal por blocos, de dimensões $GF \times GF$, sendo que cada bloco $F \times F$ é o produto $I\mathbf{n}_g$. O supervetor $\tilde{\mathbf{f}}_u$, de dimensionalidade $GF \times 1$, é obtido pela concatenação vertical de cada $\tilde{\mathbf{f}}_g$. A partir destas matrizes estima-se o *i-vector* $\boldsymbol{\omega}_u$, relacionado com o u-ésimo áudio como

$$\boldsymbol{\omega}_u = \left(I + \mathbf{T}^\top \boldsymbol{\Sigma}^{-1} \mathbf{N}_u \mathbf{T}\right)^{-1} \mathbf{T}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{f}}_u, \tag{3.32}$$

onde $\mathbf{T}^\top$ é a transposta de $\mathbf{T}$. A matriz de covariância total $\boldsymbol{\Sigma}$ é estimada juntamente com a matriz de variabilidade total $\mathbf{T}$ (DEHAK et al., 2011). No passo seguinte, a LDA (*Linear Discriminant Analysis*) projeta os *i-vectors* em eixos ortogonais, reduz a dimensionalidade e minimiza as diferenças entre as gravações (ou canais) do mesmo locutor. O problema de otimização da LDA pode ser definido de acordo com o escalar J obtido de

$$J(\mathbf{v}) = \frac{\mathbf{v}^\top \mathbf{S}_b \mathbf{v}}{\mathbf{v}^\top \mathbf{S}_w \mathbf{v}}, \tag{3.33}$$

onde $\mathbf{S}_b$ e $\mathbf{S}_w$ são, respectivamente, as matrizes de variação inter e intra locutores. Tais matrizes são definidas para o total de S locutores a partir do *i-vector* $\boldsymbol{\omega}_{s,c}$ relacionado ao locutor s e canal c como

$$\mathbf{S}_b = \frac{1}{S} \sum_{s=1}^{S} (\bar{\boldsymbol{\omega}}_s - \bar{\boldsymbol{\omega}})(\bar{\boldsymbol{\omega}}_s - \bar{\boldsymbol{\omega}})^\top, \tag{3.34}$$

[25] Mais detalhes do aprimoramento estão presentes na Seção 4.2

$$\mathbf{S}_w = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{\omega}_{s,c}-\bar{\boldsymbol{\omega}}_s)(\boldsymbol{\omega}_{s,c}-\bar{\boldsymbol{\omega}}_s)^{\mathsf{T}}, \tag{3.35}$$

onde a média por locutor $\bar{\boldsymbol{\omega}}_s$ e a média dos locutores $\bar{\boldsymbol{\omega}}$ são

$$\bar{\boldsymbol{\omega}}_s = \frac{1}{C}\sum_{c=1}^{C}\boldsymbol{\omega}_{s,c}, \text{ e} \tag{3.36}$$

$$\bar{\boldsymbol{\omega}} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{C}\sum_{c=1}^{C}\boldsymbol{\omega}_{s,c}\ , \tag{3.37}$$

C é o número de gravações (ou canais) associadas ao locutor s. A matriz de projeção $\mathbf{V}_M$ é definida com os autovetores relacionados aos M autovalores de maior magnitude, mapeando os *i-vectors* (DEHAK et al., 2011)

$$\mathbf{w}_{s,c} = \mathbf{V}_M\boldsymbol{\omega}_{s,c}, \tag{3.38}$$

onde $\mathbf{w}_{s,c}$ é a projeção LDA do *i-vector* $\boldsymbol{\omega}_{s,c}$.

Um aprimoramento, que consiste na substituição da LDA pela análise discriminante linear *fuzzy* (FLDA - *Fuzzy Linear Discriminant Analysis*), também foi proposto. Neste caso, alteram-se as equações 3.34 e 3.35 para incluir o valor de pertinência de agrupamentos *fuzzy*[26].

Dados dois *i-vectors* mapeados, $\mathbf{w}_p$ (do áudio padrão) e $\mathbf{w}_q$ (do questionado), o modelo da PLDA (*Probabilistic Linear Discriminant Analysis*) permite calcular a pontuação (ou *score*) de verificação $s_{lin}(\mathbf{w}_p,\mathbf{w}_q)$, pela razão de verossimilhança

$$s_{lin}(\mathbf{w}_p,\mathbf{w}_q) = \frac{p(\mathbf{w}_p,\mathbf{w}_q|H_1^{PLDA})}{p(\mathbf{w}_p|H_0^{PLDA})p(\mathbf{w}_q|H_0^{PLDA})}, \tag{3.39}$$

onde a hipótese H_1^{PLDA} indica que os dois *i-vectors* possuem características compatíveis, ou seja, vêm do mesmo locutor, enquanto H_0^{PLDA} indica que eles vêm de locutores diferentes. Os experimentos realizados neste trabalho utilizaram a PLDA Gaussiana (G-PLDA) baseado nas matrizes de variabilidade de locutor $\boldsymbol{\Phi}$, de canal $\boldsymbol{\Gamma}$ e na média global $\bar{\boldsymbol{\mu}}$ (PRINCE; ELDER, 2007; GARCIA-ROMERO; ESPY-WILSON, 2011)[27].

3.5 Principais Pontos

O presente capítulo iniciou revisando conceitos da produção da voz e da fala e suas características acústicas pertinentes à CFL. Em seguida, foram abordados aspectos relacionados

[26] Mais detalhes do aprimoramento estão presentes na Seção 4.2.

[27] Uma implementação para MATLAB® da metodologia *i-vector* pode ser encontrada na *MSR Identity Toolbox* distribuída pelo sitio https://www.microsoft.com/en-us/research/publication/msr-identity-toolbox-v1-0-a-matlab-toolbox-for-speaker-recognition-research-2/ (acessado em 14/01/2020).

com os métodos automáticos de comparação de locutores. Na questão da classificação foi estabelecido como são colocadas as hipóteses na CFL e foram ainda discutidas as principais fontes de contaminação presentes na amostra de voz questionada.

Sobre a comparação automática, é importante indicar o contraste entre as metodologias GMM-UMB e *i-vector*. Se aplicada à CFL, a primeira utiliza a Equação 3.27 enquanto a segunda a Equação 3.39. Nota-se pela natureza da Equação 3.27 que a GMM-UBM permite a aplicação de uma estimativa por intervalo; porém, na literatura esta metodologia evoluiu para o cálculo com *i-vector*. Como este trabalho possui como foco a aplicação forense, nenhum dos métodos foi descartado. Sobre cada metodologia, foi proposta uma abordagem para contribuir com o objetivo do trabalho, isto é, acertar mais e errar menos na aplicação da ciência no estado democrático de direito.

O método de comparação de locutores proposto por Morrison (2011a) realiza a avaliação da razão de verossimilhança por meio de densidade por núcleo multivariável (MVKD - *Multivariate Kernel Density*) definido por Aitken e Lucy (2004). Este método foi avaliado ao longo das pesquisas realizadas durante o desenvolvimento desta tese, e os principais resultados foram discutidos em Silva et al. (2018a) e Silva et al. (2018b), disponíveis no Apêndice C.

Capítulo 4

Contribuições para a Robustez e Confiabilidade em Comparação Forense de Locutores

"The force of the temptation which urges us to seek for such evidence and appearances as are in favour of our desires, and to disregard those which oppose them, is wonderfully great."

— Michael Faraday

O capítulo anterior revisou os métodos para extração de características da voz e o princípio da comparação automática de locutores. O presente capítulo apresenta resultados experimentais de duas linhas de trabalho que buscam contribuir para robustez e confiabilidade da CFL.

Foram planejados experimentos para a realização da tarefa de verificação de locutor em conjunto aberto, ou seja, uma base de dados para treinamento e calibração e outra para testes. A primeira linha de trabalho consiste em aplicar o teste de significância genuinamente Bayesiano para obter uma estimativa por intervalo na comparação forense de locutores. Na próxima seção, descreve-se, de forma matemática, uma solução do teste para média com variância desconhecida. Em seguida, define-se o cálculo do intervalo de evidência a partir do FBST e apresenta-se os resultados comparativos com outros métodos de estimativa por intervalo aplicados à CFL.

A segunda linha de trabalho propõe aprimorar o método *i-vector* utilizando a medida espectrográfica de relação sinal-ruído para separar o sinal de voz em blocos semelhantes pareáveis e, dentro destes blocos, realizar a comparação de locutores. O objetivo desse pareamento é homogenizar os dados a serem comparados e reduzir a heterocedasticidade. Este direcionamento aplica medidas nebulosas (*fuzzy*) de pertinência para as amostras dentro de cada bloco. Os resultados, na maioria dos cenários, apresentam desempenho superior às técnicas não aprimoradas.

4.1 Aplicação do FBST na Comparação de Locutores

Nesta seção são apresentados os resultados da aplicação do Teste de Significância Genuinamente Bayesiano (FBST *Full Bayesian Significance Test*) para cálculo do intervalo de credibilidade da CFL utilizando a metodologia GMM-UBM.

Primeiramente, será apresentado o princípio do FBST e uma proposta de solução rápida sobre a média com variância desconhecida. Na sequência, apresenta-se a proposta de cálculo intervalar – o intervalo de evidência – utilizando o valor-*e* e, em seguida, os experimentos comparativos com estimativa por intervalo para CFL.

Esta seção usará letras romanas para representar variáveis aleatórias observáveis, e letras gregas para os parâmetros. Para representar funções, por exemplo, $g(\theta)$ ou $h(\theta)$, ou função densidade de probabilidade – como $p(x)$ ou $f_n(\theta|x)$ utilizaram-se letras minúsculas. O valor de probabilidade será representado pela notação $Pr(\cdot)$, por exemplo, a probabilidade de um evento A será $Pr(A)$.

4.1.1 Teste de Significância Genuinamente Bayesiano - FBST

Consideremos um espaço Θ – doravante denominado espaço paramétrico – subconjunto de $\mathbb{R}^n$. Cada dimensão deste espaço Θ representa um parâmetro de um modelo. Por exemplo, se o espaço Θ referir-se a um modelo gaussiano ele estará em $\mathbb{R}^2$ e uma dimensão representará a média e outra a variância. O FBST foi proposto por Pereira e Stern (1999) e tem como premissa inferir sobre este espaço paramétrico onde é definido o modelo probabilístico do problema.

A formulação do FBST e suas características, após bem compreendidas, mostram-se intuitivas, coerentes, além de possuírem uma interpretação geométrica. O FBST tem início na formulação de uma hipótese precisa H que gera um subconjunto através de definições de

igualdade e/ou desigualdade no espaço de parâmetros, sendo

$$H : \theta \in \Theta, \tag{4.1}$$

onde o subconjunto Θ_H é definido pela hipótese da Equação 4.1 como

$$\Theta_H = \{\theta \in \Theta \mid g(\theta) \leq 0 \wedge h(\theta) = 0\}. \tag{4.2}$$

$g(\theta)$ e $h(\theta)$ são funções que definem os limites do subconjunto Θ_H. Por exemplo, uma hipótese no modelo gaussiano pode definir um valor para a média e um limite superior para a variância da forma $H : \mu = 0 \wedge \sigma^2 \leq 1$, onde $\wedge$ é a conjunção lógica, o que automaticamente define o subespaço $\Theta_H = \left\{\theta \in \Theta \mid \sigma^2 - 1 \leq 0 \wedge \mu = 0\right\}$. Nota-se também que a definição precisa de igualdade cria uma redução de dimensionalidade no espaço Θ_H; no exemplo da gaussiana, o subespaço Θ_H possui dimensionalidade $\mathbb{R}^1$.

É possível formular uma hipótese precisa do tipo $H : \theta = \theta_0$ onde θ é uma variável do espaço paramétrico Θ e θ_0 o valor sobre o qual desejamos realizar o teste. Dada a observação de um experimento que gera os dados $\mathbf{x} \in \mathcal{X}$, a probabilidade a *posteriori* $f_n(\theta|\mathbf{x})$ de uma hipótese precisa condicionada à observação $\mathbf{x}$ pode ser definida como a proporção

$$f_n(\theta|\mathbf{x}) \propto f(\theta)\mathcal{L}(\theta|\mathbf{x}), \tag{4.3}$$

onde $f(\theta)$ é a probabilidade a *priori* da hipótese, $\mathcal{L}(\theta|\mathbf{x})$ a verossimilhança entre a observação e a hipótese, e $\mathcal{X}$ o espaço da variável aleatória $\mathbf{x}$. Assim, definem-se

$$\theta^* = \arg\max_{\theta \in \Theta_H} f_n(\theta|\mathbf{x}) \tag{4.4}$$

e

$$f_n^* = \max_{\theta \in \Theta_H} f_n(\theta^*|\mathbf{x}), \tag{4.5}$$

respectivamente, como o parâmetro θ^* que maximiza a probabilidade a *posteriori* e a probabilidade f_n^* máxima sobre a hipótese. Em consequência, tem-se o conjunto de maior surpresa relativa (HRSS - *Hightest Relative Surprise Set*) (MADRUGA et al., 2003), que é a região T^* onde

$$T^* = \{\theta \in \Theta \mid f_n(\theta|\mathbf{x}) > f_n^*\}. \tag{4.6}$$

O valor da evidência contra a hipótese $H : \theta \in \Theta$, considerando como suporte os dados $\mathbf{x}$, é definido como

$$\overline{ev} = Pr(\theta \in T^*|\mathbf{x}) = \int_{T^*} f_n(\theta|\mathbf{x}) d\theta, \tag{4.7}$$

onde $Pr(\theta \in T^*|\mathbf{x})$ é a probabilidade de o parâmetro θ estar contido na região T^*. No contexto do FBST, o trabalho de Petri (2007) define o valor-e associado ao FBST como sendo

$$\text{valor-}e = 1 - Pr(\theta \in T^*|\mathbf{x}). \tag{4.8}$$

Diferentemente da abordagem frequentista, o FBST obtém o valor-e, – ou a probabilidade da evidência $\overline{ev}$ –, no espaço dos parâmetros Θ, enquanto o valor-p é uma probabilidade calculada sobre espaço da amostra (MADRUGA et al., 2003).

4.1.2 Estimativa por Intervalo na CFL

A inferência pontual na verificação de locutores é baseada em uma pontuação (*score*) que indica a (dis)similaridade entre uma amostra padrão e uma questionada (REYNOLDS et al., 2000; MORRISON, 2011a; HANSEN; HASAN, 2015). Por outro lado, a estimativa por intervalo aparece como uma relação de compromisso entre precisão e confiança. Abre-se mão de alguma precisão da estimativa movendo-se de um ponto para um intervalo, obtendo alguma confiança, ou garantia, de que o resultado encontra-se naquele intervalo (CASELLA; BERGER, 2011, pp. 374).

Os trabalhos com estimativa por intervalo em reconhecimento automático de locutor começaram com Bisani e Ney (2004), que utilizou o método de *bootstrap* (EFRON; TIBSHIRANI, 1994) no cálculo de intervalos de confiança. Na sequência, Campbell et al. (2005) utilizou *multi-layer perceptron* (MLP) para obter intervalos de confiança baseados na entropia estatística. Posteriormente, Koval e Lokhanova (2011) utilizaram funções sigmoides para aproximar a probabilidade *a posteriori* $P(H_0^{PS}|\mathbf{X})$, onde $\mathbf{X}$ são as características da voz e H_0^{PS} a hipótese nula, utilizando *Platt scaling* (PLATT et al., 1999) na estimação de intervalos de credibilidade. O intervalo de credibilidade também pode ser calculado por métodos empíricos como o proposto por Morrison et al. (2011).

Na metodologia GMM-UBM, a pontuação (*score*) da comparação é obtida pela Equação 3.27, que trata de uma média amostral. Um método analítico para o cálculo do intervalo de confiança – sobre o espaço das amostras – utiliza a distribuição t-Student proposta por Gosset (STUDENT, 1908; ZABELL, 2008) como

$$t_{\left(\frac{\alpha}{2},T-1\right)}\sqrt{\frac{var(LLR(\mathbf{X}_Q))}{T}} \leq LLR(\mathbf{X}_Q) \leq t_{\left(\frac{\alpha}{2},T-1\right)}\sqrt{\frac{var(LLR(\mathbf{X}_Q))}{T}}. \tag{4.9}$$

Tal abordagem foi proposta por Morrison et al. (2010), onde $var(LLR(\mathbf{X}_Q))$ é a variância amostral do logaritmo da razão de verossimilhança (*log-likelihood ratio*) $LLR(\mathbf{X}_Q)$ da amostra questionada $\mathbf{X}_Q \in \mathcal{R}^{F\times T}$ (vide Equação 3.27), T o número de quadro, e $t_{\left(\frac{\alpha}{2},T-1\right)}$ é a distribuição t de Student com significância α e $T-1$ graus de liberdade.

As aplicações do FBST na estimação intervalar ou aplicadas à CFL não foram encontradas durante o levantamento bibliográfico desta pesquisa. Assim, a principal contribuição desta linha de trabalho é a proposta da estimação intervalar, sua aplicação à CFL e o método para

solucionar o FBST para o valor esperado (média) com variância desconhecida sem o uso de Monte Carlo e Cadeia de Markov (MCMC).

4.1.3 FBST sobre Média com Variância Desconhecida

Esta seção tem por objetivo detalhar o FBST para média de um modelo gaussiano com variância desconhecida. A formulação busca também uma solução para o cálculo de $\overline{ev}$, como indicado na Equação 4.7, sem utilizar o algoritmo de Monte Carlo e Cadeia de Markov (MCMC) proposto por Madruga et al. (2003).

Isto posto, considere uma amostra normal $\mathbf{x} \in \mathcal{X}$ com n observações independentes e igualmente distribuída da forma $\mathcal{N}(\mu, 1/\rho)$, sendo μ a média e $\rho = \frac{1}{\sigma^2}$ a precisão. A estatística mínima suficiente é obtida da média amostral $\bar{x}$ e da soma dos quadrados dos desvios $Q = \sum_{i=1}^{n} (x_i - \bar{x})^2$. A função de verossimilhança para os limites $\mu = (-\infty, \infty)$ e $\rho = (0, \infty)$ pode ser escrita da forma (MADRUGA et al., 2003; PETRI, 2007; OLIVEIRA et al., 2018; ASSANE et al., 2019)

$$\mathcal{L}(\mu, \rho | n, \bar{x}, Q) \propto \rho^{\frac{n}{2}} e^{-\rho \frac{Q}{2}\left(1 + \frac{n}{Q}(\mu - \bar{x})^2\right)}. \tag{4.10}$$

Tomando como probabilidade a *priori* não informativa do tipo $p(\mu, 1/\rho) = \frac{d\mu\, d\rho}{\rho}$ obtém-se a probabilidade a *posteriori* da forma (MADRUGA et al., 2003)

$$f_n(\mu, \rho | n, \bar{x}, Q) = c\rho^{\frac{n}{2}-1} e^{-\rho \frac{Q}{2}\left(1 + \frac{n}{Q}(\mu - \bar{x})^2\right)}, \tag{4.11}$$

onde

$$c = \frac{Q^{\frac{n-1}{2}} \sqrt{n}}{2^{\frac{n}{2}} \sqrt{\pi} \Gamma\left(\frac{n-1}{2}\right)}. \tag{4.12}$$

A constante c da Equação 4.12 foi obtida integrando a Equação 4.11 para obter área unitária. Para encontrar o valor máximo de $f_n(\mu, \rho)$[1] é possível tomar as derivadas

$$\begin{aligned} \frac{\partial f_n}{\partial \mu} &= -nc\rho^{\frac{n}{2}}(\mu - \bar{x}) e^{-\rho \frac{Q}{2}\left(1 + \frac{n}{Q}(\mu - \bar{x})^2\right)} \\ \frac{\partial f_n}{\partial \rho} &= c\left[\frac{n-2}{2}\rho^{\frac{n}{2}-2} - \left(\frac{Q}{2}\left(1 + \frac{n}{Q}(\mu - \bar{x})^2\right)\right)\rho^{\frac{n}{2}-1}\right] e^{-\rho \frac{Q}{2}\left(1 + \frac{n}{Q}(\mu - \bar{x})^2\right)}, \end{aligned} \tag{4.13}$$

levando aos pontos de máximo com $P_(\mu*, \rho*)$ da forma

$$\begin{aligned} \mu^* &= \bar{x} \\ \rho^* &= \frac{n-2}{Q}. \end{aligned} \tag{4.14}$$

[1] Doravante a função densidade de probabilidade $f_n(\mu, \rho | n, \bar{x}, Q)$ será representada apenas por $f_n(\mu, \rho)$.

Pode-se definir a hipótese precisa de que a variável aleatória $\mathbf{x} \in \mathcal{X}$ possui média igual a η, ou seja

$$H : \mu = \eta, \tag{4.15}$$

onde a variável η que representa um valor no espaço paramétrico da média (valor esperado) μ. A precisão ρ_A, que delimita o conjunto tangente T^*, é definida como

$$\rho_A = \arg\max f_n(\mu = \eta, \rho).$$

Para fins de exemplificação, consideremos uma amostra gaussiana unidimensional $\mathbf{x} \in \mathcal{X}$ em que se deseja testar se sua média é nula, ou seja, $\eta = 0$ e, consequentemente, $H : \mu = 0$. A superfície da Figura 13 mostra a função densidade de probabilidade a *posteriori* $f_n(\mu, \rho)$ da Equação 4.11. Ao aplicar a hipótese $H : \mu = 0$, tem-se a redução de dimensionalidade do espaço paramétrico – de uma superfície em $\mathbb{R}^2$ para a linha azul $f_n(\mu = 0, \rho)$ em $\mathbb{R}$ – imposta pela hipótese nula. O valor máximo sobre a linha azul é $f_n^* = f_n(\mu = 0, \rho_A)$ que define o contorno da linha amarela onde $f_n(\mu, \rho) = f_n^*$. A linha amarela delimita o conjunto tangente T^* onde $f_n(\mu, \rho) \geq f_n^*$. A linha vermelha indica o subconjunto de ρ que atravessa o valor máximo de $f_n(\mu, \rho)$, ou seja, $f_n(\mu, \rho = \rho^*)$[2].

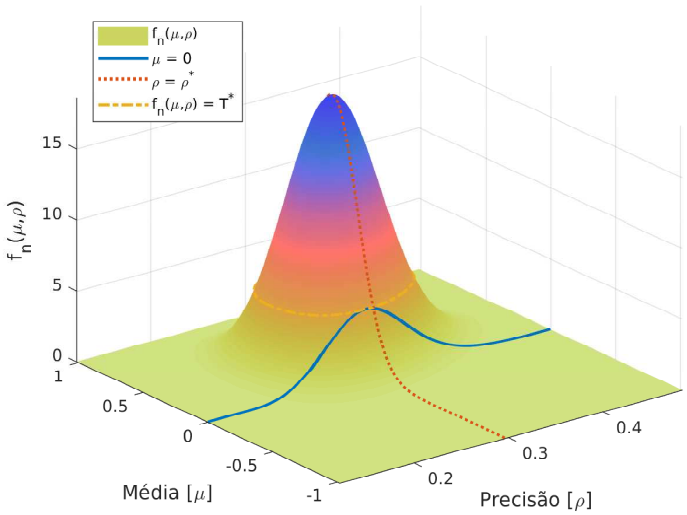


Figura 13 – Função densidade de Probabilidade $f_n(\mu, \rho | n, \bar{x}, Q)$. Na imagem a linha azul $f_n(\mu = 0, \rho)$ indica a hipótese nula. O valor máximo sobre a linha azul é $f_n^* = f_n(\mu = 0, \rho_A)$ que define o contorno da linha amarela onde $f_n(\mu, \rho) = f_n^*$ que delimita o conjunto tangente T^* onde $f_n(\mu, \rho) \geq f_n^*$. A linha vermelha indica o subconjunto $f_n(\mu, \rho = \rho*)$ que atravessa o valor ρ^*. Adaptado de Arruda (2012).

A evidência contra a hipótese precisa ($H : \mu = \eta = 0$) é a integral de $f_n(\mu, \rho)$ tendo como limites de integração o conjunto tangente T^*. Observando o contorno da Figura 14, nota-

[2] Nota-se que ρ_A não é necessariamente igual a ρ^*.

se que este é simétrico sobre o eixo da média μ e que possui pontos extremos nas duas dimensões (μ e ρ).

Os trabalhos pesquisados que tratam do teste sobre a média com variância desconhecida utilizando o FBST (MADRUGA et al., 2003; PETRI, 2007; OLIVEIRA et al., 2018; ASSANE et al., 2019), sugerem o algoritmo de Monte Carlo e Cadeia de Markov (MCMC) para realizar a integral de f_n como

$$\overline{ev} = Pr(\mu, \rho \in T^* | \mathbf{x}) = \int_{T^*} f_n(\mu, \rho) d\theta. \tag{4.16}$$

Para o desenvolvimento da proposta do intervalo de evidência – baseado no FBST sobre a média com variância desconhecida – fez-se importante evoluir na solução da Equação 4.11. O desenvolvimento a seguir apresenta uma tentativa de obter a solução do FBST sem uso do MCMC.

Devido à concavidade da superfície da Equação 4.11 nota-se que o conjunto tangente T^* possui os pontos extremos, na dimensão da precisão, como ρ_A, ρ_B, ρ_C e ρ_D, como apresentados na Figura 14, onde

$$\rho_A = \frac{n-2}{Q\left(1+\frac{n}{Q}(\eta-\bar{x})^2\right)},$$

e

$$\rho_B = \frac{n-2}{Q\left(1+\frac{n}{Q}(2\bar{x}-\eta-\bar{x})^2\right)} = \rho_A.$$

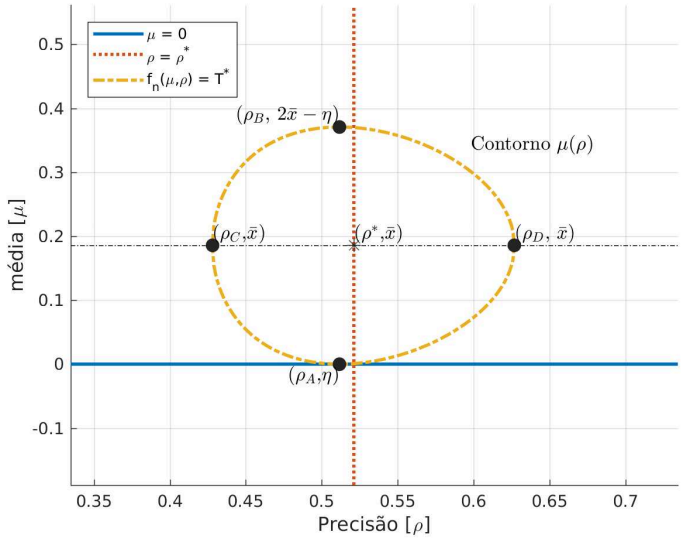


Figura 14 – Contorno e pontos extremos do conjunto tangente T^*.

Os pontos ρ_C e ρ_D não são tão intuitivos de demonstrar. Entretanto, fazendo $f_n(\bar{x}, \rho) = f_n(\eta, \rho_A)$ tem-se

$$c\rho^{\frac{n}{2}-1}e^{-\rho\frac{Q}{2}} = c\rho_A^{\frac{n}{2}-1}e^{-\rho_A\frac{Q}{2}\left(1+\frac{n}{Q}(\eta-\bar{x})^2\right)};$$

e agrupando-se as variáveis e tomando-se o logaritmo de ambos os lados, segue que

$$\left(\frac{\rho}{\rho_A}\right)^{\frac{n}{2}-1} e^{-\frac{Q}{2}\left[\rho-\rho_A\left(1+\frac{n}{Q}(\eta-\bar{x})^2\right)\right]} = 1$$

$$\downarrow \ln(\cdot)$$

$$\rho - \left(\frac{n-2}{Q}\right)\ln(\rho) + \left(\frac{n-2}{Q}\right)\ln(\rho_A) - \rho_A\left(1+\frac{n}{Q}(\eta-\bar{x})^2\right) = 0. \tag{4.17}$$

A equação 4.17 é do tipo:

$$\rho + a\ln(\rho) + b = 0,$$

onde

$$\begin{aligned} a &= -\frac{n-2}{Q} \\ b &= \left(\frac{n-2}{Q}\right)\ln(\rho_A) - \rho_A\left(1+\frac{n}{Q}(\eta-\bar{x})^2\right), \end{aligned} \tag{4.18}$$

e possui como raízes

$$\begin{aligned} \rho_D &= \exp\left[-W_{-1}\left(\frac{e^{-\frac{b}{a}}}{a}\right) + \frac{b}{a}\right]; \text{ e} \\ \rho_C &= \exp\left[-W_0\left(\frac{e^{-\frac{b}{a}}}{a}\right) + \frac{b}{a}\right] \end{aligned} \tag{4.19}$$

onde $W_k(\cdot)$ é a função Lambert-W[3]. Considerando que T^* é simétrico em relação ao eixo μ, o cálculo da evidência contra a hipótese nula pode ser avaliado como:

$$\overline{ev} = 2\int_{\rho_C}^{\rho_D}\int_{\bar{x}}^{\mu(\rho)} c\rho^{\frac{n}{2}-1} e^{-\rho\frac{Q}{2}\left(1+\frac{n}{Q}(\mu-\bar{x})^2\right)} d\mu\, d\rho, \tag{4.20}$$

onde $\mu(\rho)$ é a função que define o contorno, superior ou inferior, no eixo μ do conjunto tangente T^* (vide Figura 14). Sobre o contorno, tem-se:

$$f_n(\mu,\rho) = c\rho^{\frac{n}{2}-1} e^{-\rho\frac{Q}{2}\left(1+\frac{n}{Q}(\mu-\bar{x})^2\right)} = f_n^*$$

$$\downarrow \ln(\cdot)$$

$$\frac{n-2}{2}\ln(\rho) - \rho\frac{Q}{2}\left(1+\frac{n}{Q}(\mu-\bar{x})^2\right) = \ln\left(\frac{f_n^*}{c}\right)$$

$$\mu^2 - 2\bar{x}\mu + \bar{x}^2 + \frac{Q}{n} - \frac{2}{\rho n}\left[\frac{n-2}{2}\ln(\rho) - \ln\left(\frac{f_n^*}{c}\right)\right] = 0. \tag{4.21}$$

[3] A função Lambert-W que soluciona a equação $ye^y = x$, com solução apenas para $x \geq -\frac{1}{e}$. Para $x \geq 0$ a solução é $y = W_0(x)$. Para $-\frac{1}{e} \leq x \leq 0$ têm-se duas soluções, $y = W_0(x)$ e $y = W_{-1}(x)$ (VALLURI et al., 2000).

As raízes da equação 4.21, do segundo grau em μ, definem os contornos da forma:

$$\mu(\rho) = \bar{x} \pm \sqrt{\frac{2}{\rho n}\left[\frac{n-2}{2}\ln(\rho) - \frac{f_n^*}{c}\right] - \frac{Q}{n}}. \tag{4.22}$$

Pela simetria do problema, calcula-se a integral da equação 4.20 como:

$$\overline{ev} = 2c \int_{\rho_C}^{\rho_D} \rho^{\frac{n}{2}-1} e^{-\rho\frac{Q}{2}} \left[\int_{\bar{x}}^{\mu(\rho)} e^{-\frac{n\rho}{2}(\mu-\bar{x})^2} d\mu\right] d\rho \rightarrow$$

$$\overline{ev} = 2c \int_{\rho_C}^{\rho_D} \rho^{\frac{n}{2}-1} e^{-\rho\frac{Q}{2}} \sqrt{\frac{2\pi}{n\rho}} \left[\operatorname{erf}\left(\sqrt{\frac{n\rho}{2}}(\mu(\rho)-\bar{x})\right) - \cancelto{0}{\operatorname{erf}\left(\sqrt{\frac{n\rho}{2}}(\bar{x}-\bar{x})\right)}\right] d\rho, \tag{4.23}$$

onde $erf(\cdot)$ é a função erro. O argumento que aparece na função erro pode ser simplificado para a forma

$$v(\rho) = \frac{n-2}{2}\ln\left(\frac{\rho}{\rho_A}\right) - \frac{Q}{2}(\rho-\rho_A) + \frac{\rho_A n}{2}(\eta-\bar{x})^2, \tag{4.24}$$

onde ρ_A é o limite inferior de ρ, e η é o valor da média que se deseja testar na hipótese 4.15. Assim, a equação 4.20 pode ser reescrita como uma integral unidimensional com limites de integração conhecidos pela forma

$$\overline{ev} = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{Q}{2}\right)^{\frac{n-1}{2}} \int_{\rho_C}^{\rho_D} \rho^{\frac{n-3}{2}} e^{-\rho\frac{Q}{2}} \operatorname{erf}\left(\sqrt{v(\rho)}\right) d\rho. \tag{4.25}$$

A relevância desse desenvolvimento foi solucionar o FBST sobre a média com variância desconhecida sem a utilização do MCMC. Obviamente, exitem esforços numéricos para o cálculo da Equação 4.25 e das funções erro e Lambert-W. Mesmo assim viabiliza-se o cálculo do intervalo de evidência proposto na próxima seção.

4.1.4 Definição do Intervalo de Evidência

Esta seção propõe aplicar a formulação do FBST para definir uma estimativa por intervalo sobre o parâmetro da média com variância desconhecida que doravante será denominada *intervalo de evidência*. Observando as equações 4.24 e 4.25, nota-se que o valor da evidência $\overline{ev}$ depende dos parâmetros da amostra – $\bar{x}$, n e Q – e do valor η sobre o qual deseja-se testar a média.

Nota-se na Figura 13, que à medida que o valor de η percorre o eixo da média afastando-se da média amostral, que está no topo da superfície, o conjunto tangente T^* aumenta. Consequentemente, a área de integração e o valor da evidência $\overline{ev}$ contra a hipótese $H : \mu = \eta$ também aumenta.

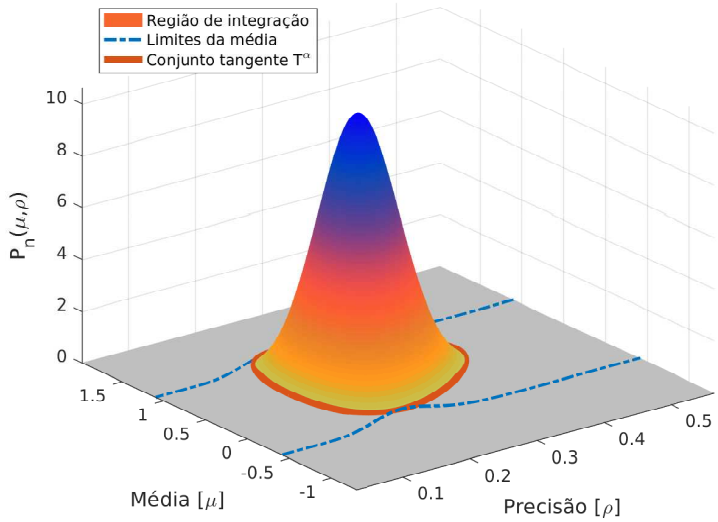


Figura 15 – Exemplo de conjunto $T^{0,05}$, ou seja $\alpha = 0,05$, que delimita a superfície de integração. Na imagem, a linha vermelha, paralela ao plano horizontal, indica o conjunto $T^{0,05}$, a porção da curva em degradê está contida no conjunto e as linhas azuis são os limites superior e inferior do conjunto no eixo da média μ.

A Equação 4.22 indica que o conjunto tangente T^* é simétrico em relação ao eixo da média μ. Assim define-se um valor $\alpha \in [0,1]$, análogo ao nível de significância, que permite obter um conjunto tangente T^α tal que o resultado da integral da Equação 4.7 tenha o valor 1 - α.

A Figura 15 apresenta um exemplo de conjunto $T^{0,05}$, ou seja $\alpha = 0,05$. A linha vermelha, paralela ao plano horizontal, indica o conjunto $T^{0,05}$, a porção da curva em degradê está contida no conjunto e as linhas azuis são os limites superior e inferior do conjunto tangente no eixo da média μ[4].

Sobre o eixo da média μ existem dois limites, η_H maior e η_L menor que a média amostral $\bar{x}$, e tangentes ao conjunto T^α, como indicado pelas linhas azuis nos exemplos das figuras 15 e 16. Dadas as localizações geométricas sobre o FBST define-se, dado o valor de α, o intervalo de evidência para a média com variância desconhecida como o intervalo entre η_L e η_H como indicado pela linha amarela, paralela ao eixo da média, no exemplo da Figura 16. Nesta proposta o valor α seria análogo ao nível de significância.

A evidência (valor-*e* calculado pelo FBST) de que a média paramétrica μ possui valores fora do intervalo $[\eta_L, \eta_H]$ é menor que α. É claro que a definição acima não se encaixa no intervalo tradicional de confiança (ou credibilidade), como definido por Bolstad (2013) ou por Casella e Berger (2011). No entanto, é uma metodologia analítica baseada no espaço

[4] Na imagem da Figura 15, a integral sob a curva degradê, dentro do conjunto $T^{0,05}$, vale 0,95 e a integral sob a curva cinza, fora do conjunto $T^{0,05}$, vale 0,05.

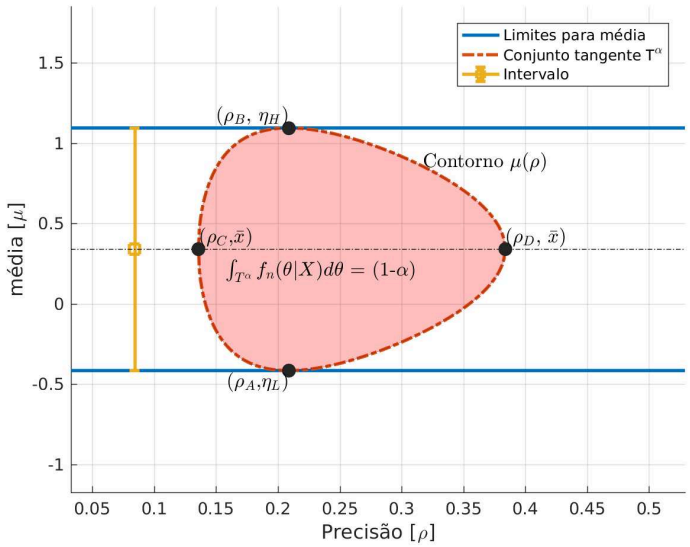


Figura 16 – Corte transversal sobre o conjunto tangente indicando os limites do intervalo de evidência.

de parâmetros, é transparente e representa os limites de evidência que a amostra é capaz de fornecer[5].

Na prática da CFL, considere uma comparação entre as características $\mathbf{X}_Q$ de uma voz questionada com um modelo λ_P conhecido, dado um UBM λ_{UBM}, utilizando a metodologia GMM-UBM. O resultado é uma série temporal $\hat{\mathbf{x}}_Q[t]$ com valor médio $LLR(\mathbf{X}_Q)$ (vide Equações 3.27 e 3.28), que é a pontuação (*score*) da comparação. A Figura 17 mostra um exemplo numérico destes dados. Na imagem da Figura 17 tem-se, no painel da esquerda, o histograma normalizado (Norm. Hist.) dos valores de $\hat{\mathbf{x}}_Q[t]$, a linha azul escura é a função de densidade de probabilidade empírica (FDP emp.) e o círculo sobre essa curva indica a localização do valor médio $LLR(\mathbf{X}_Q)$ no eixo horizontal. O retângulo pontilhado no gráfico à esquerda é a região no gráfico à direita.

No exemplo, o valor de $LLR(\mathbf{X}_Q) \approx -0.8\,Np$ (nepers)[6]. A avaliação da hipótese $H : LLR(\vec{x}_Q) = \eta$ ao longo da variável η no espaço LLR (eixo horizontal). A partir do FBST (Equação 4.25), é possível construir a *curva ev*.

A variação dos valores de η resultou na curva-ev, amarelo escuro sólido, indicada no gráfico à direita da Figura 17. Essa curva é calculada através da amostragem do espaço η e da solução da Equação 4.25 para cada amostra. No gráfico, a linha tracejada pontilhada horizontal (ev = 0,05) indica a evidência Bayesiana (análogo a significância) $\alpha = 0,05$ (valor de evidência

[5] Aparentemente essa metodologia também permite extrair uma estimativa por intervalo sobre a precisão ρ no intervalo $[\rho_C, \rho_D]$. Porém, esse não é foco deste trabalho e não foi aplicado a CFL.

[6] Neper é utilizado para expressar logaritmo natural de uma razão entre valores. A palavra é derivada do nome de John Napier. Valores em Np e em dB (decibel) têm uma razão constante, verificando-se que $1\,Np = 20\log_{10}(e) \approx 8,7\,dB$.

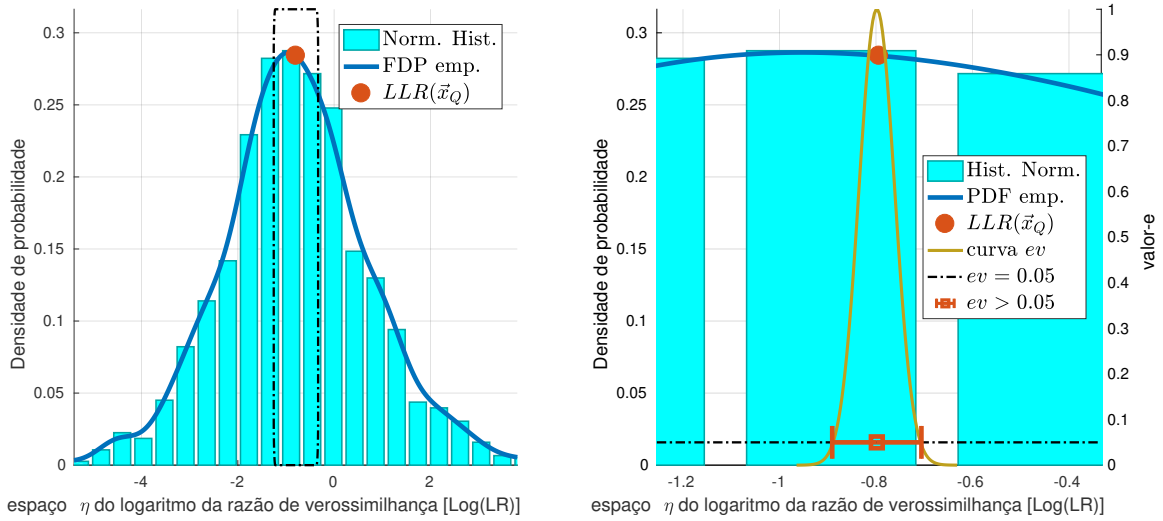


Figura 17 – Exemplo da estimativa por intervalo sobre a média de $LLR(x_Q)$. O painel à esquerda mostra o histograma normalizado de $LLR(x_Q)$ ocorrências e a função de densidade de probabilidade empírica (linha grossa). No mesmo painel, o retângulo tracejado indica a região mostrada no painel à direita. Neste painel, a linha sólida amarelo escuro em forma de sino indica *curva ev* e a barra de erro laranja na horizontal indicam intervalo de evidência e a média amostral.

contra hipóteses $\overline{ev} = 95\%$ ou valor-e = 0,05). A barra de erro horizontal (ev > 0,05) indica o intervalo de evidência e a média amostral[7].

4.1.5 Comparação com outros métodos

Nesta etapa, foi planejado um experimento para avaliar o desempenho da estimativa por intervalo na comparação automática de locutores, utilizando a metodologia GMM-UBM, e as amostras do Corpus CEFALA-1 (NETO et al., 2019) obtidas pelo aparelho celular. Em complementação, foram utilizadas 50 amostras de voz padrão gravadas para casos de comparação forense de locutor entre 2009 e 2013, cedidas pelo Instituto de Criminalística da Polícia Civil de Minas Gerais, Brasil[8].

Na preparação dos áudios foi realizada uma subamostragem para 8 kHz seguida da aplicação de um filtro com largura de faixa entre 300 e 3.400 Hz (simulando o canal telefônico). Para a etapa de treinamento, cada amostra de voz (áudio) do corpus CEFALA-1 foi dividido igualmente, metade da gravação sendo utilizada para emular o áudio padrão e o restante o áudio questionado.

[7] Uma implementação para MATLAB® do método para o cálculo do intervalo de evidência é disponibilizado pelo autor em https://github.com/adelinocpp/FBST_matlab (acessado em 14/01/2020).

[8] Mais detalhes e informações da composição do corpus CEFALA-1 e sobre os registros de áudio cedidos pelo IC-MG podem ser encontradas no Apêndice A.

Os áudios questionados foram contaminados por ruído rosa com SNR nos valores de 25, 23, 20, 17, 15 e 12 *dB*, resultando em um total de 6 áudios contaminados para cada amostra. Após a contaminação, os áudios padrão (não contaminado) e questionados foram codificados e decodificados pelo *codec* GSM 06.60 para simular a influência do canal. A etapa de testes utilizou os modelos e as calibrações obtidas na etapa de treinamento e comparou as vozes das 50 amostras complementares com os mesmos procedimentos de divisão da amostra, contaminação por ruído rosa e SNR. No experimento todos os áudios padrão são comparados com os áudios questionados.

Como características foram calculadas os MFCC utilizando 13 bandas críticas, com quadros de 25 *ms* de duração e passo de tempo de 10 *ms*, gerando as observações $\mathbf{X} \in \mathcal{R}^{F \times T}$. Os trechos de fala foram extraídos pelo algoritmo de Sohn et al. (1999).

Os métodos usados para calcular a estimativa por intervalo (significância, ou análogo, $\alpha = 0,05$) foram:

FBST A metodologia proposta que calcula o intervalo de evidência como um subespaço de espaço paramétrico em que valor-*e* é α;

Gosset O intervalo de confiança calculado pela distribuição *t*-Student (ou *t*-Gosset) como indicado na Equação 4.9.

Morrison Intervalo de credibilidade empírica computado combinando a técnica KNN (*K-Nearest-Neighborhood*) com regressão linear, como descrito por Morrison (2011b).

O método proposto por Morrison (2011b) foi originalmente usado para calcular o intervalo de credibilidade sobre os dados, e não sobre a média. No presente trabalho, o método de Morrison foi adaptado para calcular a média de 50 subamostras com substituição (similar ao *bootstrap* proposto por Efron e Tibshirani (1994)).

Para a comparação dos métodos de estimativa por intervalo definiram-se três cenários básicos como resultados da comparação de locutor com um limiar de decisão e uma estimativa por intervalo. O primeiro cenário é uma comparação correta, ou seja, associam-se os registros de áudio do mesmo locutor ou desassociam-se registros de áudio de locutores diferentes. No caso da inferência pontual, basta a pontuação obtida estar, respectivamente, acima ou abaixo do limiar de decisão. Na estimativa por intervalo, o limiar de decisão não está contido no intervalo.

Um segundo cenário, a comparação incorreta, desassocia os registros de áudio de um mesmo locutor ou associa registros de áudio de locutores diferentes. Na inferência pontual, o *score*

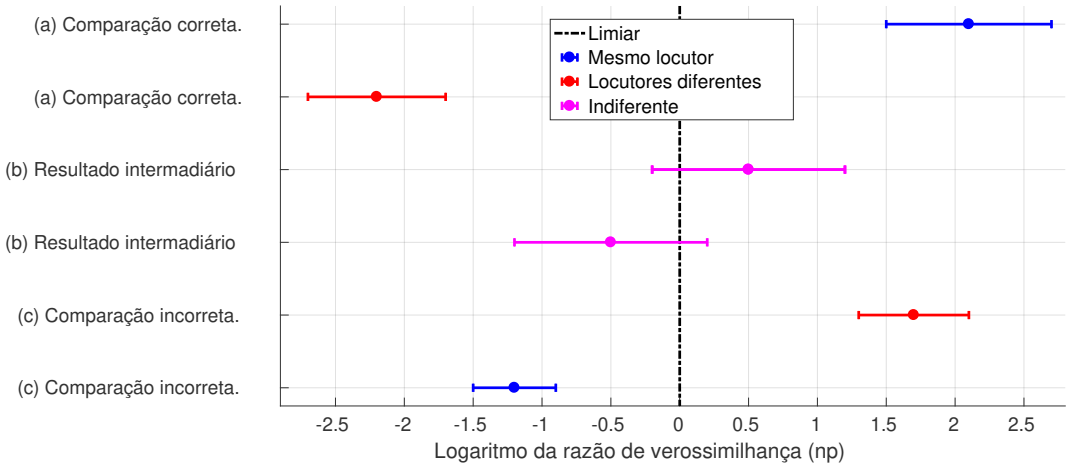


Figura 18 – Cenários de resultados da estimativa por intervalo. A linha pontilhada vertical indica o limiar de decisão em $LLR = 0$, pontuação acima do limiar associa locutores e abaixo desassocia. As linhas azuis representam as comparações entre áudios de um mesmo falante e as vermelhas entre falantes diferentes. Os cenários são o de comparação correta (a), de resultados intermediários (b), e de comparações incorretas (c).

obtido aparece, respectivamente, abaixo e acima do limiar de decisão. Assim como no primeiro cenário, o limiar de decisão está fora do intervalo. No resultado intermediário, que ocorre apenas na estimativa por intervalo, o limiar de decisão aparece dentro do intervalo.

A Figura 18 apresenta exemplos destes cenários de estimativa por intervalo, com o eixo horizontal indicando o valor de LLR. Na figura, o círculo indica o valor médio, a linha horizontal a estimativa por intervalo, e a linha vertical preta o limiar de decisão sobre $LLR = 0$. Os cenários das duas primeiras linhas (a), são exemplos de comparação correta, onde o mesmo locutor (linha azul) possui pontuação e intervalo maiores que o limiar e locutores diferentes (vermelho) possuem pontuação e intervalo menores que o limiar. No resultado intermediário (b), o intervalo (magenta) estende-se sobre o limiar de decisão. Na comparação incorreta (c), ocorre o oposto da comparação correta.

Os resultados intermediários definem uma região de transição entre a decisão de associar ou desassociar dois registros de voz ao mesmo locutor, o que não acontece na inferência pontual. Na CFL, esses cenários são "resultados inconclusivos" ou um *in dubio pro reo*, que é alinhado ao aforismo jurídico do princípio da inocência.

A Figura 19 apresenta as curvas *Detection Error Tradeoff* (DET) da etapa de treinamento. A curva DET apresenta, em escala não linear, a variação das taxas de falso positivo (FP) e falso negativo (FN) com o limiar de decisão. O cruzamento da curva DET com a linha preta pontilhada (FP = FN) indica a taxa de mesmo erro (EER - *Equal Error Rate*) da técnica. Quanto mais próximo da origem, menores são as taxas de FP e FN na etapa de treinamento.

A etapa de treinamento foi calibrada utilizando a inferência pontual e apresentou uma taxa de mesmo erro de 8,1%.

Na Figura 19, a curva vermelha foi obtida da inferência pontual e as curvas limites foram obtidas considerando as inferências intervalares. Na imagem é possível notar o intervalo de evidência, calculado pela formulação do FBST, apresenta uma maior dispersão da inferência pontual, fato que também pode ser constatado na Figura 22.

Na Tabela 2 têm-se os resultados de cada um dos cenários de comparação ilustrados na Figura 18 para as inferências intervalares pelos métodos de Gosset, Morrison e FBST nas etapas de treinamento e teste.

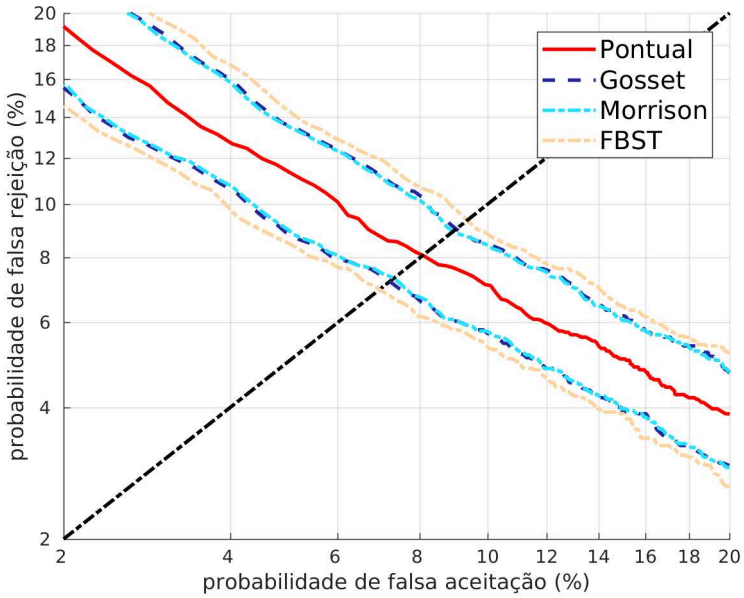


Figura 19 – Curva DET (*detection error tradeoff*) da etapa de treinamento das técnicas de estimativa por intervalo.

Tabela 2 – Resultado comparativo apresentando os valores médios dos percentuais de classificação, das etapas de treinamento e teste, entre a estimativa pontual e a estimativa por intervalo.

		Classificação correta (%)	Classificação incorreta (%)	*In dubio pro reo* (%)
Treinamento	Pontual	91,9	8,1	
	Gosset	91,0	7,2	1,8
	Morrison	91,0	7,3	1,7
	FBST	90,4	7	2,6
Teste	Pontual	87,95	12,05	
	Gosset	87,5	11,4	1,1
	Morrison	87,5	11,3	1,2
	FBST	87,45	11,25	1,3

Nota-se, nos resultados da etapa de teste, que a redução do percentual de classificações incorretas, com a aplicação da estimativa por intervalo, está entre 0,6% e 0,8%. Já a redução

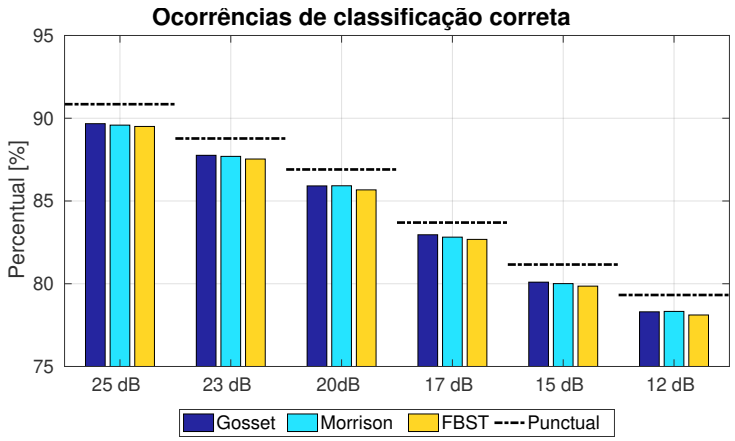


Figura 20 – Percentual de classificações corretas, cenário (a), para cada método de estimativa por intervalo com diferentes valores de SNR na etapa de testes.

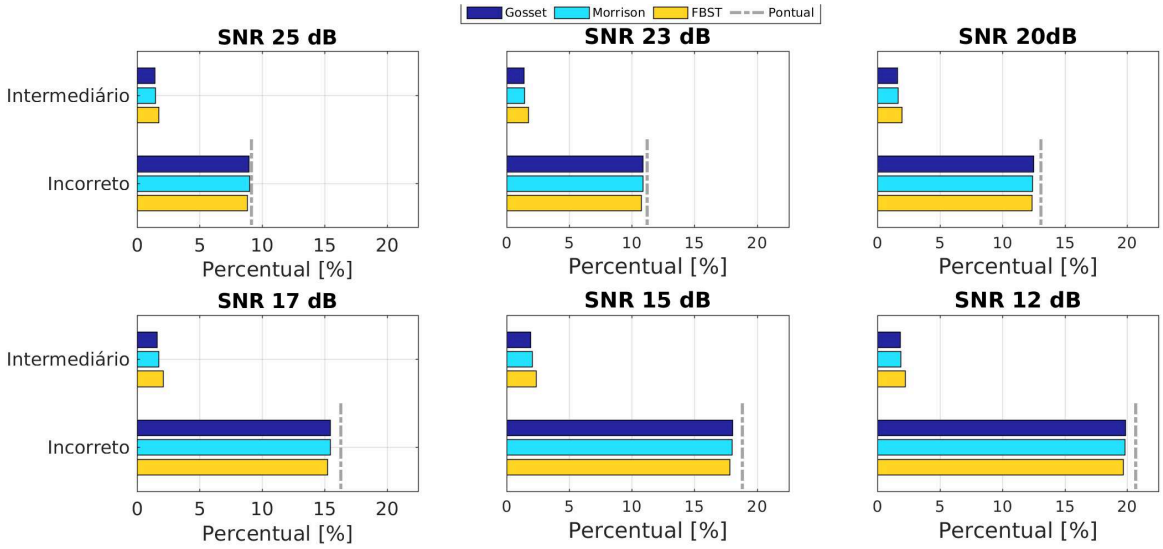


Figura 21 – Resultados da estimativa por intervalo apresentando o percentual de ocorrências dos cenários (b) e (c) da Figura 18. As barras horizontais indicam as porcentagens por método, enquanto a linha vertical tracejada em cinza escuro indica a porcentagem de erro da inferência pontual.

que ocorre nas classificações corretas está entre 0,45% e 0,5%. Em valores percentuais comparados com a estimativa pontual, a aplicação do intervalo de evidência reduziu em 6,4% o número de comparações incorretas. Por outro lado, quando aplicado o intervalo de evidência, o número de comparações corretas foi reduzido em 0,6%.

No recorte realizado pela intensidade da ruído tem-se que quanto maior a SNR maior é o percentual de classificação corretas. Outra tendência que pode ser observada na Figura 21 é o aumento na proporção de resultados intermediários com a diminuição da SNR. Na média, tem-se, respectivamente, 1,5% e 2,1% de resultados intermediários para contaminação de 25 *dB* e 12 *dB*.

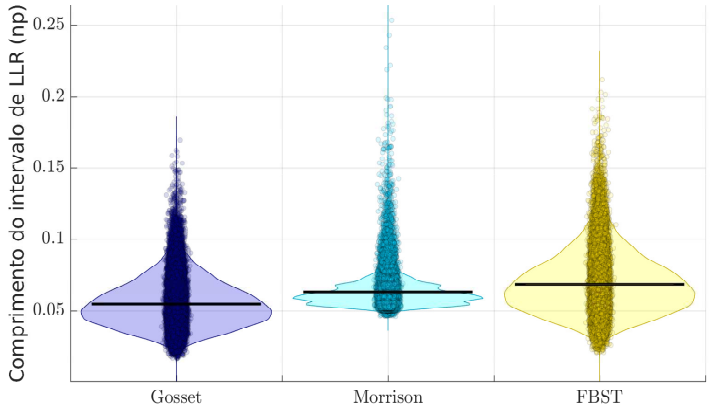


Figura 22 – Gráfico RDI (*Raw data, Description and Inference* do tamanho de intervalo para cada método de estimativa por intervalo. Nos gráficos, em cada coluna, os pontos são os valores individuais, as curvas laterais indicam a distribuição de probabilidade empírica, a linha preta horizontal a média.

A Figura 22 apresenta o gráfico RDI (*Raw data, Description and Inference*) com as ocorrências do tamanho do intervalo de acordo com o método. Nos gráficos, em cada coluna, os pontos são os valores individuais de cada comprimento de intervalo, as curvas laterais indicam a distribuição de probabilidade empírica, a linha preta horizontal a média amostral.

Nota-se que, na média, o intervalo de evidência, dados em amarelo, é maior que as demais e sua distribuição é tão dispersa quanto a de Gosset. Uma particularidade que apareceu no método de Morrison foi a dispersão dos valores (largura da FDP empírica), menor que dos demais métodos, e a extensão (cauda) dos intervalos maiores.

Sobre as melhorias no cálculo do FBST para distribuição de média com variância desconhecida, os cálculos propostos permitem eliminar o uso de técnicas de Monte Carlo e Cadeia de Markov (MCMC - *Markov chain Monte Carlo*) para calcular a integral FBST.

O trabalho também propõe o uso do intervalo de evidência na CFL, que, comparado com outros métodos, reduziu a quantidade de ocorrências de erro Tipo I, em detrimento do aumento de resultados intermediários (*In dubio pro reo*), em cenários com baixa SNR.

4.2 Comparação de Locutores Pareada pela relação sinal-ruído

O objetivo da presente seção é apresentar os resultados obtidos na aplicação de técnicas de lógica *fuzzy* à metodologia *i-vector* de verificação de locutor, em conjunto aberto e independente de texto, em cenários que emulam as condições forenses de contaminação por canal GSM, ruído e limitação da duração da amostra questionada.

O presente trabalho inova ao aplicar lógica *fuzzy* em dois pontos independentes da metodologia *i-vector*:

- o primeiro é a substituição da etapa análise discriminante linear (LDA - *Linear Discriminant Analysis*) pela análise discriminante linear *fuzzy* (FLDA - *Fuzzy Linear Discriminant Analysis*) proposta por Zhi et al. (2013);
- o segundo – doravante denominado *fuzzy*-S^2NR –, é o pareamento dos quadros de áudios em conjuntos *fuzzy*. Os conjuntos são definidos por medidas espectrográficas de relação sinal-ruído (S^2NR - *Spectrographic Signal-to-Noise Ratio*) e cada conjunto possui uma modelagem e comparação próprias. Além disso, propõe-se a modificação no cálculo das estatísticas de Baum-Welch para incluir a pertinência dos conjuntos de *fuzzy*-S^2NR.

O pareamento por S^2NR na verificação de locutor foi explorado na metodologia GMM-UBM por Silva et al. (2018a). O Pareamento fragmenta as amostras em conjuntos de acordo o valor da S^2NR de cada quadro de voz. O pareamento calcula um modelo de verificação de locutor para cada conjunto (ou faixa) de S^2NR e recombina os resultados utilizando uma rede neural artificial (RNA).

A proposta é agregar uma função de pertinência a cada conjunto, e obter um modelo completo incluindo UBM, matriz de variabilidade **T**, *i-vector* e o modelo PLDA para cada conjunto. A pontuação resultante da comparação é uma combinação, por RNA, da pontuação obtida por cada conjunto. A aplicação da técnica *fuzzy* dá-se na inclusão do valor de pertinência no cálculo das estatísticas de Baum-Welch.

A contribuição principal da presente linha de trabalho é a obtenção de uma melhoria da técnica *i-vector*, aplicada em cenário forense. Também foram estimados, sobre a técnica proposta, limites de acurácia, taxas de verdadeiro positivo e de falso positivo. O cenário da comparação é independente de texto, em conjunto aberto, com variação de ruído e tamanho de amostra questionada.

A próxima seção apresenta onde são inseridas as modificações e soluções propostas para o cálculo dos *i-vectors*. A Seção 4.2.2 apresenta os resultados da etapa de treinamento e de testes utilizando bases de dados diferentes e discute o desempenho de cada modificação proposta.

4.2.1 Aplicação de Técnicas Fuzzy

A análise discriminante linear *fuzzy* (FLDA - *Fuzzy Linear Discriminant Analysis*) utilizada no presente trabalho é baseada na maximização da entropia do algoritmo de agrupamento

fuzzy (MEFCA - *Maximum Entropy Fuzzy Clustering Algorithm*) proposto em (ZHI et al., 2013), que considera um conjunto de S *i-vectors* da forma $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_S\}$. A função objetivo do MEFCA para obter a matriz de associação *fuzzy* pode ser definida como

$$\begin{aligned} J_{MEFCA}(\mathbf{U}) = &\sum_{k=1}^{K}\sum_{s=1}^{S} u_{ks}||\boldsymbol{\omega}_s - \bar{\boldsymbol{\omega}}_k||^2 \\ &+ \beta^{-1}\sum_{k=1}^{K}\sum_{s=1}^{S} u_{ks}\ln(u_{ks}), \end{aligned} \tag{4.26}$$

onde $||\cdot||$ é a norma vetorial, $\mathbf{U} = \{u_{ks}\}_{K\times S}$ é a matriz de associação *fuzzy* para K fatores com pertinência entre o s-ésimo *i-vector* e o k-ésimo fator $u_{ks} \in [0,1]$ e $\sum_{k=1}^{K} u_{ks} = 1$, $\bar{\boldsymbol{\omega}}_k$ é a média (centro) do k-ésimo fator (agrupamento) e β um parâmetro de regularização positivo. As condições que minimizam J_{MEFCA} são

$$\bar{\boldsymbol{\omega}}_k = \frac{\sum_{s=1}^{S} u_{ks}\boldsymbol{\omega}_s}{\sum_{s=1}^{S} u_{ks}}, \tag{4.27}$$

$$u_{ks} = \frac{\exp(-\beta||\boldsymbol{\omega}_s - \bar{\boldsymbol{\omega}}_k||^2)}{\sum_{k=1}^{K}\exp(-\beta||\boldsymbol{\omega}_s - \bar{\boldsymbol{\omega}}_k||^2)}. \tag{4.28}$$

A FLDA é obtida do agrupamento obtido pela MEFCA considerando a pertinência u_{kn} ficando as matrizes de variação intra e inter locutores como

$$\mathbf{S}_{fw} = \sum_{k=1}^{K}\sum_{s=1}^{S} u_{ks}(\boldsymbol{\omega}_s - \bar{\boldsymbol{\omega}}_k)(\boldsymbol{\omega}_s - \bar{\boldsymbol{\omega}}_k)^{\mathsf{T}}, \tag{4.29}$$

$$\mathbf{S}_{fb} = \sum_{k=1}^{K}\sum_{s=1}^{S} u_{ks}(\bar{\boldsymbol{\omega}}_k - \bar{\boldsymbol{\omega}})(\bar{\boldsymbol{\omega}}_k - \bar{\boldsymbol{\omega}})^{\mathsf{T}}, \tag{4.30}$$

onde $\bar{\boldsymbol{\omega}}$ é a média dos *i-vectors* de $\boldsymbol{\Omega}$. Nos experimentos realizados, foi utilizado $\beta = 50$ de acordo com os resultados apresentados em (ZHI et al., 2013).

Na técnica de pareamento baseada na contaminação *fuzzy*-S^{2}NR tem-se que para cada matriz de MFCC é definido um vetor $\boldsymbol{\sigma} = [\sigma[0], \sigma[1], ..., \sigma[T-1]]$ que armazena a medida S^{2}NR para cada quadro de voz, sendo o espaço S^{2}NR dividido em P conjuntos como indicado na Figura 23. Para cada um dos P conjuntos *fuzzy* é definida uma função de pertinência $\rho_p[t] = \mu(\sigma[t])$ onde $\rho_p[t]$ é a pertinência do quadro t ao p-ésimo conjunto e $\mu_p(\sigma[t])$ a função de pertinência do p-ésimo conjunto. A matriz de pertinência $\mathbf{P} = \{\rho_p[t]\}_{P\times T}$ associa o quadro t ao p-ésimo conjunto *fuzzy*-S^{2}NR.

Os índices T^p, que indicam quais quadros de $\mathbf{X}$ pertencem ao conjunto p, podem ser obtidos como $T^p = \{0 \leq t \leq T-1 | \arg\max_{t,p}(\mu_p(\sigma[t]))\}$, ou seja, o t-ésimo quadro de voz de $\mathbf{X}$

estará contido no conjunto *fuzzy*-S^{2}NR p de maior pertinência $\rho_p[t]$. Na etapa de treinamento definem-se P funções de pertinência sobre o espaço S^{2}NR para que cada conjunto contenha o mesmo percentil de quadros.

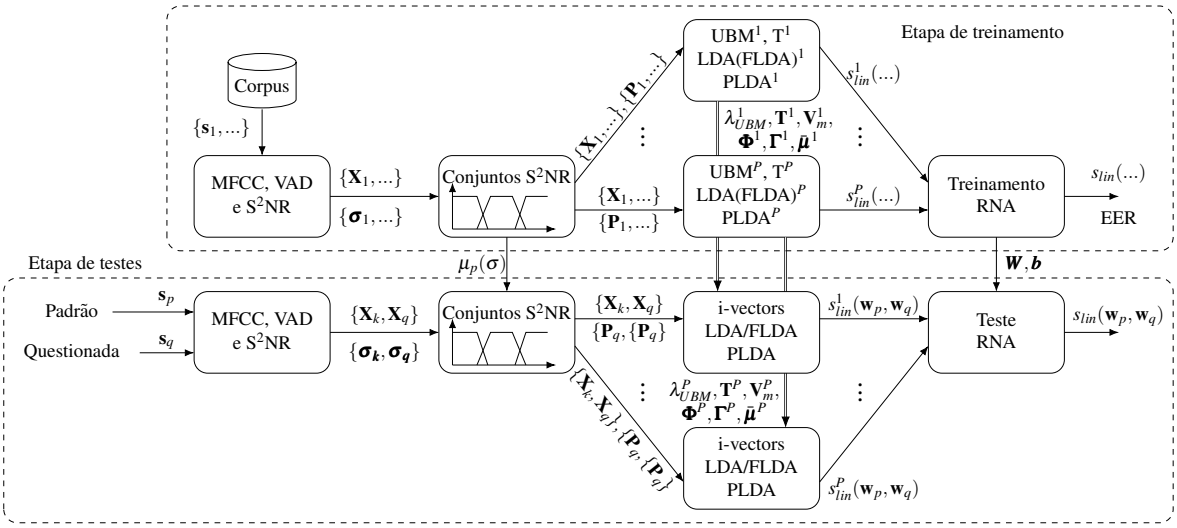


Figura 23 – Etapas da comparação de locutores utilizando *fi-vector*-S^{2}NR. Primeiramente as características – MFCC e o S^{2}NR – são extraídas dos registros de voz que são selecionados pelo algoritmo VAD. Destas características são modelados os P conjuntos. Para cada conjunto são modelados o UBM e a matriz de variabilidade total $\mathbf{T}$, os *i-vector*s, a LDA (ou FLDA) e a pontuação calculada pela PLDA. Uma rede neural consolida o resultado.

Com o pareamento, realiza-se a comparação entre locutores dentro de cada conjunto *fuzzy*-S^{2}NR. Para cada conjunto foram construídos um UBM, uma matriz de variabilidade total $\mathbf{T}$ e os *i-vectors*, da mesma maneira que na metodologia de referência. Na metodologia *fuzzy*-S^{2}NR as estatísticas *fuzzyficadas* de Baum-Welch de ordem zero $\mathbf{n}^p_{fg}$ e de primeira ordem centralizada $\tilde{\mathbf{f}}^p_{fg}$, relacionada ao conjunto *fuzzy*-S^{2}NR p e a gaussiana g, ficam da forma

$$\mathbf{n}^p_{fg} = \sum_{t \in T^p} \frac{\rho_p[t]}{\sum_{j=1}^{P} \rho_j[t]} P(g|\mathbf{x}[t], \lambda^p_{UBM}), \tag{4.31}$$

$$\tilde{\mathbf{f}}^p_{fg} = \sum_{t \in T^p} \frac{\rho_p[t]}{\sum_{j=1}^{P} \rho_j[t]} P(g|\mathbf{x}^p[t], \lambda^p_{UBM})(\mathbf{x}^p[t] - \boldsymbol{\mu}_g), \tag{4.32}$$

sendo $P(g|\mathbf{x}[t], \lambda^p_{UBM})$ a probabilidade da g-ésima gaussiana $(g = 1,...,G)$ condicionada ao quadro de voz $\mathbf{x}[t]$ e ao UBM λ^p_{UBM} da classe p $(p = 1,...,P)$ e $\rho_j[t]$ a pertinência do t-ésimo quadro $\mathbf{x}[t]$ da matriz de MFCC com o conjunto *fuzzy*-S^{2}NR j. Doravante, a técnica de pareamento utilizando conjuntos *fuzzy*-S^{2}NR combinada com as estatísticas de Baum-Welch *fuzzyficadas* das equações 4.31 e 4.32 será referenciada como *fi-vectors* (*fuzzy i-vectors*).

4.2.2 Resultados Experimentais

A base de dados CEFALA-1 (NETO et al., 2019) apresenta um total de 104 locutores gravados por cinco dispositivos de capturas (canais) distintos. O corpus apresenta 55 locutores do sexo masculino e 49 do sexo feminino com coletas realizadas em cabine isolada acusticamente com ruído de fundo de 34 *dBA*. Em complementação, foram utilizadas 50 amostras de voz padrão gravadas para casos de comparação forense de locutor entre 2009 e 2013, cedidas pelo Instituto de Criminalística da Polícia Civil de Minas Gerais, Brasil[9].

Para a etapa de treinamento, cada amostra de voz (áudio) do corpus CEFALA-1 foi dividida igualmente, metade da gravação sendo utilizada para emular o áudio padrão e a outra metade o áudio questionado. As amostras questionadas foram contaminadas por ruídos de *trânsito*, *disparos* e *multidão* (*babble noise*), *branco* e *rosa* com SNR médio de 25, 23, 20, 17, 15 e 12 *dB*, resultando em um total de 30 áudios questionados para cada áudio padrão.

A simulação do canal telefônico foi realizada com subamostragem para 8 kHz, limitação da banda entre 300 e 3.400 Hz e codificação/decodificação pelo *codec* GSM 06.60 (ITU, 1991). A etapa de testes utilizou os modelos e as calibrações obtidas na etapa de treinamento e comparou as vozes das 50 amostras complementares com os mesmos procedimentos de divisão da amostra, contaminação por tipo de ruído e intensidade.

Os MFCC foram calculados usando janela de Hamming de 25 ms com cálculo de 13 coeficientes a cada 10 ms, incluindo as variações de primeira e de segunda ordem para produzir um vetor de características com 39 dimensões. A relação sinal-ruído (SNR) foi extraída alinhada ao vetor de características utilizando o método S^2NR (VIEIRA et al., 2014) e a atividade de voz com o método de Sohn et al. (1999).

O UBM foi calculado, independente de gênero, com 512 gaussianas. A matriz de variabilidade total $\mathbf{T}$ e os *i-vectors* foram extraídos com 400 fatores. Os *i-vectors* foram normalizados como sugerido por (GARCIA-ROMERO; ESPY-WILSON, 2011; HATCH et al., 2006).

Na metodologia de referência ("*i-vector* LDA"), o UBM, a matriz de variabilidade total $\mathbf{T}$ e os *i-vectors* são calculados utilizando toda amostra de voz com as estatísticas de Baum-Welch (equações 3.30 e 3.31), a LDA realizada com 103 fatores e a pontuação obtida com PLDA Gaussiana (KENNY, 2012). A primeira variação dessa metodologia substituiu a etapa LDA pela FLDA (ZHI et al., 2013; WU; ZHOU, 2006) utilizando as matrizes de variabilidade inter e intra classe, respectivamente $\mathbf{S}_{fb}$ e $\mathbf{S}_{fw}$, como indicado nas equações 4.29 e 4.30. Os experimentos calcularam a FLDA variando o número de fatores K (50, 100, 150, 200, 300 e 400).

[9] Mais detalhes e informações da composição do corpus CEFALA-1 e sobre os registros de áudio cedidos pelo cedidas pelo IC-MG podem ser encontradas no Apêndice A.

Tabela 3 – Resumo das variações propostas sobre a metodologia *i-vector* de referência (indicada pelo índice 1). As variações que substituem a etapa de LDA pelo FLDA são indicadas entre os índices 2 ao 7. Dos índices 8 ao 14 têm-se as metodologias que aplicam a técnica fuzzy-S^{2}NR com as estatísticas de Baum-Welch *fuzzyficadas*.

Índice	Técnica	Indicação
1	UBM, a matriz **T** e *i-vectors* calculados com toda amostra de voz, estatísticas de Baum-Welch e LDA	*i-vector* LDA (ref.)
2-7	UBM, a matriz **T** e *i-vectors* calculados com toda amostra de voz, estatísticas de Baum-Welch e FLDA com K fatores	*i-vector* K-FLDA
8	UBM, a matriz **T** e *i-vectors* calculados em cada conjunto baseado em SNR, estatísticas de Baum-Welch *fuzzy* e LDA	*fi-vector* LDA
9-14	UBM, a matriz **T** e *i-vectors* calculados em cada conjunto baseado em SNR, estatísticas de Baum-Welch *fuzzy* e FLDA com K fatores	*fi-vector* K-FLDA

’ Foram calculados três conjuntos *fuzzy* $(P = 3)$ baseados na S^{2}NR, de forma que cada conjunto contenha um terço de todos quadros vozeados do áudio padrão. Na Figura 24 pode-se observar as funções de pertinência $\mu_p(\sigma)$, sendo a “Z” entre -10,5 *dB* e 12 *dB*, a “S” entre 19,5 *dB* e 38 *dB* e a Gaussiana centrada em 15,7 *dB* com largura (análoga ao desvio padrão) de 3,7 *dB* (MEDASANI et al., 1998).

O S^{2}NR foi calculado com 512 pontos na DFT e limiares $\sigma_{th} = 0,082$ e $R_{th} = 0,1$. No trabalho original (VIEIRA et al., 2014), tais parâmetros foram calibrados em $\sigma_{th} = 0,1$ e $R_{th} = 0,6$ para a vogal /a/ sintetizada com frequência fundamental de 220 Hz e SNR de 30 *dB*. Para os experimentos desta tese, o S^{2}NR foi ajustado em condições distintas do trabalho original, i.e., fala corrida com SNR entre 12 e 25 *dB*.

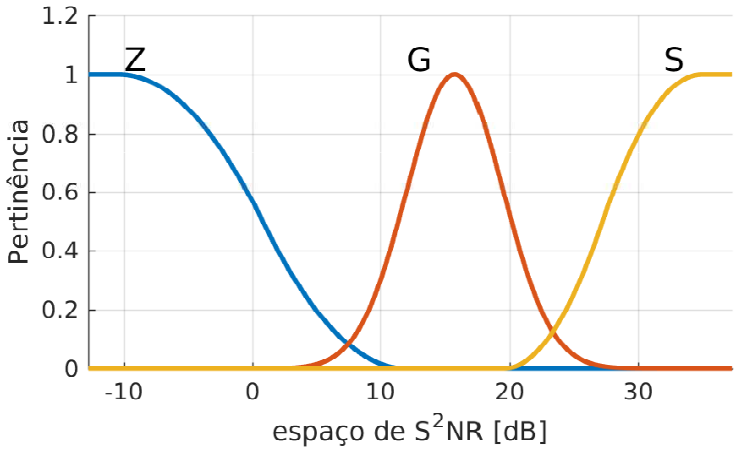


Figura 24 – Conjuntos fuzzy-SNR calculados a partir das amostras padrão da etapa de treinamento. No gráfico a curva “Z” (azul) é localizada entre -10,5 e 12 *dB*, a “S” (amarela) entre 19,5 e 38 *dB*, e a Gaussiana (“G” vermelha) centrada em 15,7 com largura (desvio padrão) de 3,7 *dB*.

Ao aplicar os conjuntos *fuzzy*-S^{2}NR, calculou-se, para cada conjunto, um modelo completo com UBM, matriz **T** e *i-vectors*. Esta comparação, pareada por conjunto de S^{2}NR, calcula as estatísticas de Baum-Welch considerando a pertinência de cada quadro em relação ao conjunto *fuzzy*-S^{2}NR pelas equações 4.31 e 4.32.

As pontuações $s^p_{lin}(\cdot)$ obtidas por cada conjunto utilizando a PLDA são combinadas por uma Rede Neural Artificial (RNA). As variações que aplicam a FLDA também podem ser aplicadas nos conjuntos *fuzzy*-S^{2}NR. A Tabela 3 apresenta um resumo das variações da metodologia de referência.

A RNA utilizada foi uma *multilayer perceptron* (MLP), com três camadas, sendo três neurônios na camada de entrada, nove na camada oculta e um na saída. O treinamento da RNA utilizou algoritmo genético com todos os resultados que utilizavam conjuntos *fuzzy*-S^{2}NR (índices 8 a 14 da Tabela 3), com o objetivo de minimizar a taxa de mesmo erro (EER - *Equal Error Rate*) média da etapa de treinamento. A avaliação da etapa de testes foi baseada nas taxas:

1. verdadeiro positivo (VP): associar duas amostras de voz quando são oriundas do mesmo locutor, a taxa de falso negativo é sua medida complementar;
2. falso positivo (FP): associar duas amostras de voz quando são oriundas de locutores diferentes, a taxa de verdadeiro negativo é sua medida complementar;
3. acurácia (ACC): taxa de associações corretas dentre todas as comparações.

Em um cenário ideal de verificação de locutor, o treinamento apresentaria EER igual a zero. Já em casos forenses, um falso positivo, que gera uma evidência que pode associar locutores diferentes (levar a condenar um inocente), é mais grave que um falso negativo.

A Figura 25 apresenta as curvas *Detection Error Tradeoff* (DET) da etapa de treinamento. A curva DET apresenta, em escala não linear, a variação das taxas de FP e FN com o limiar de decisão. O cruzamento da curva DET com a linha preta pontilhada (FP = FN) indica a EER da técnica. Quanto mais próximo da origem, menores são as taxas de FP e FN na etapa de treinamento.

Na imagem é possível notar que várias técnicas apresentam EER entre 3,5% e 4,5% sendo que apenas três superam a técnica de referência (EER = 4%) são elas "*fi-vector* 200-FLDA", "*i-vector* 200-FLDA" e "*fi-vector* 300-FLDA".

Na Tabela 4 são destacadas em negrito as técnicas que obtiveram uma taxa de falso positivo abaixo de 16,8% (referência). Dentre elas, o "*fi-vector* 300-FLDA" apresenta uma taxa de falso positivo de 10,8% e o "*i-vector* 300-FLDA", de 15,6%.

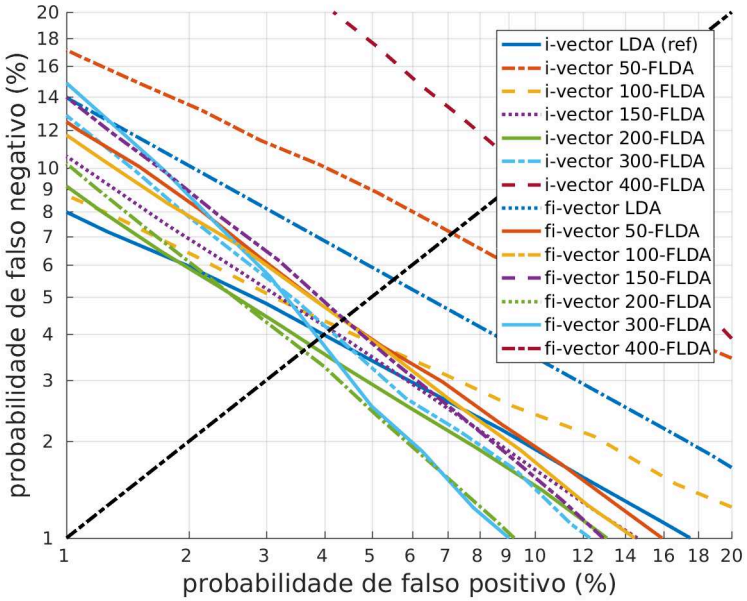


Figura 25 – Curva DET (*Detection Error Tradeoff*) da etapa de treinamento das técnicas de verificação de locutor enumeradas na Tabela 3. Várias técnicas apresentam EER entre 3,5% e 4,5% sendo que apenas três apresentam EER inferior a referência com 4%, sendo elas "*fi-vector* 200-FLDA", "*i-vector* 200-FLDA" e "*fi-vector* 300-FLDA".

Tabela 4 – Resumo de resultados das etapas de treinamento e teste para os principais resultados dentre as variações da metodologia de referência. Na segunda coluna tem-se a EER da etapa de treinamento e nas colunas seguintes a taxa de verdadeiro positivo (VP), de verdadeiro negativo (VN), de falso positivo (FP), de falso negativo (FN) e a acurácia (AC) da etapa de teste.

	Treinamento	Testes				
Técnica	EER (%)	VP (%)	VN (%)	FP (%)	FN (%)	AC (%)
i-vector LDA (ref)	4,0	95,5	83,2	16,8	4,5	83,4
i-vector 100-FLDA	4,2	94,4	85,7	**14,3**	5,6	85,8
i-vector 150-FLDA	4,1	95,4	84,7	**15,3**	4,6	84,9
i-vector 200-FLDA	**3,8**	94,6	84,4	**15,6**	5,4	84,6
i-vector 300-FLDA	4,1	92,1	86,5	**13,5**	7,9	86,6
fi-vector 50-FLDA	4,4	95,5	85,3	**14,7**	4,5	85,5
fi-vector 100-FLDA	4,4	95,7	85,9	**14,1**	4,3	86,1
fi-vector 200-FLDA	**3,6**	96,2	84,3	**15,7**	3,8	84,5
fi-vector 300-FLDA	**3,9**	93,7	89,2	**10,8**	6,3	89,3

Comparando as técnicas da Tabela 4 com o "*i-vector* LDA", nota-se, no treinamento, uma diferença máxima de EER 0,4%, pois a EER de referência é 4,0% com os máximos e mínimo, respectivamente, em 4,4% e 3,6%. Entretanto, a redução na taxa de falso positivo da referência para o "fi-vector 300-FLDA" em 6% (16,8% - 10,8%) é um resultado considerável. Em valores percentuais, comparados com a metodologia de referência, "fi-vector 300-FLDA"

apresentou um incremento de acurácia de 7,1% e reduziu as taxas de verdadeiro e falso positivo, respectivamente, em 1,9% e 35,7%.

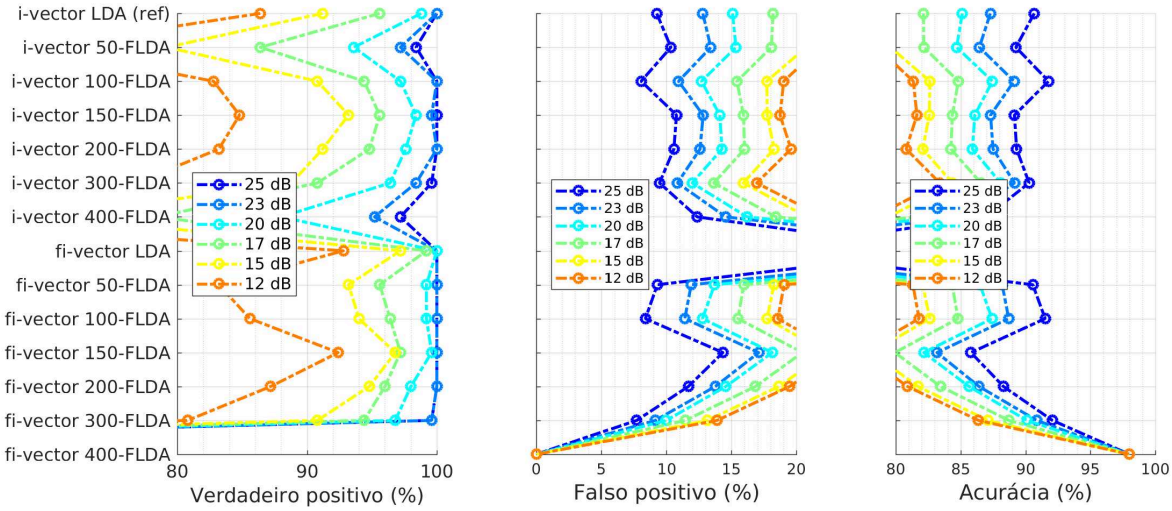


Figura 26 – Taxas de verdadeiro positivo, falso negativo e acurácia para cada metodologia e de acordo com a contaminação SNR do áudio questionado.

Ainda observando a etapa de testes, é possível fazer um recorte mais específico em relação à SNR dos registros de áudio utilizados como questionados na etapa de teste. A Figura 26 apresenta a taxa de verdadeiro positivo (à esquerda), a taxa de falso positivo (ao centro) e a acurácia (à direita). Nota-se que a contaminação do áudio questionado é fundamental para estabelecer os limites de desempenho da comparação de locutores. Vê-se que tanto a técnica de referência quanto as técnicas que apresentam um desempenho comparável possuem dispersão das taxas de acordo com a intensidade da relação sinal-ruído do áudio questionado.

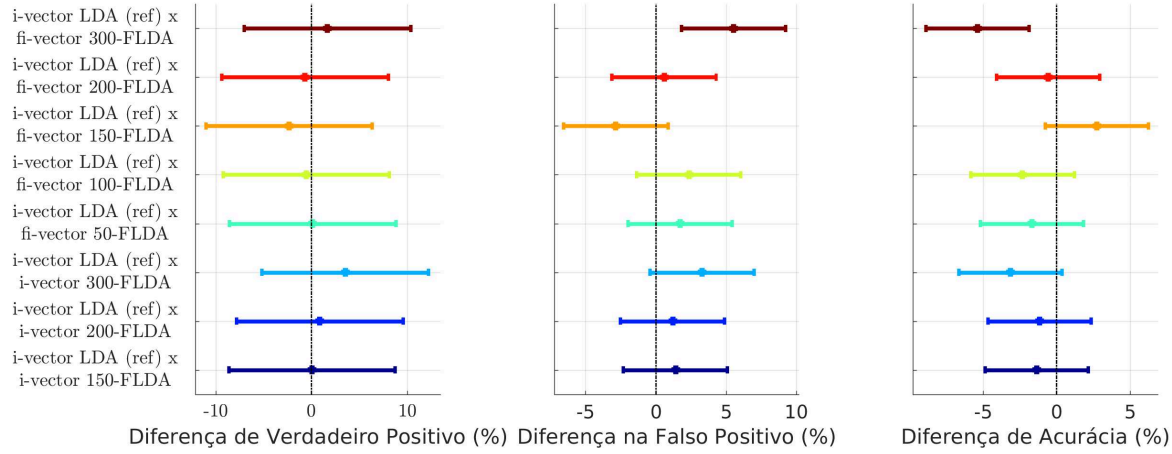


Figura 27 – Análise de variância entre a metodologia de referência e as técnicas que apresentaram melhor desempenho para as taxas de verdadeiro positivo (à esquerda), falso positivo (centro) e da acurácia (à direita) com $\alpha = 0,05$. Em cada linha a diferença média é representada pelo quadrado e o intervalo de confiança, pela linha horizontal.

Considerando que na etapa de testes foram utilizadas 30 variações do áudio questionado (sendo 5 tipos de ruído com seis níveis de SNR), é possível utilizar a análise de variância para comparar o resultado das amostras de teste. A Figura 27 apresenta a diferença média e o intervalo de confiança, para $\alpha = 0,05$, das taxas de verdadeiro positivo, falso positivo e da acurácia entre a metodologia de referência e às técnicas que apresentaram melhor desempenho. Para a taxa de verdadeiro positivo, nota-se no painel da esquerda que todas as metodologias foram estatisticamente equivalentes. Entretanto, em relação à taxa de falso positivo (painel central), a metodologia "*fi-vector* 300-FLDA" (na primeira linha) apresentou um desempenho superior, assim como para a acurácia, como pode ser constatado no painel da direita da Figura 26 e na Tabela 4.

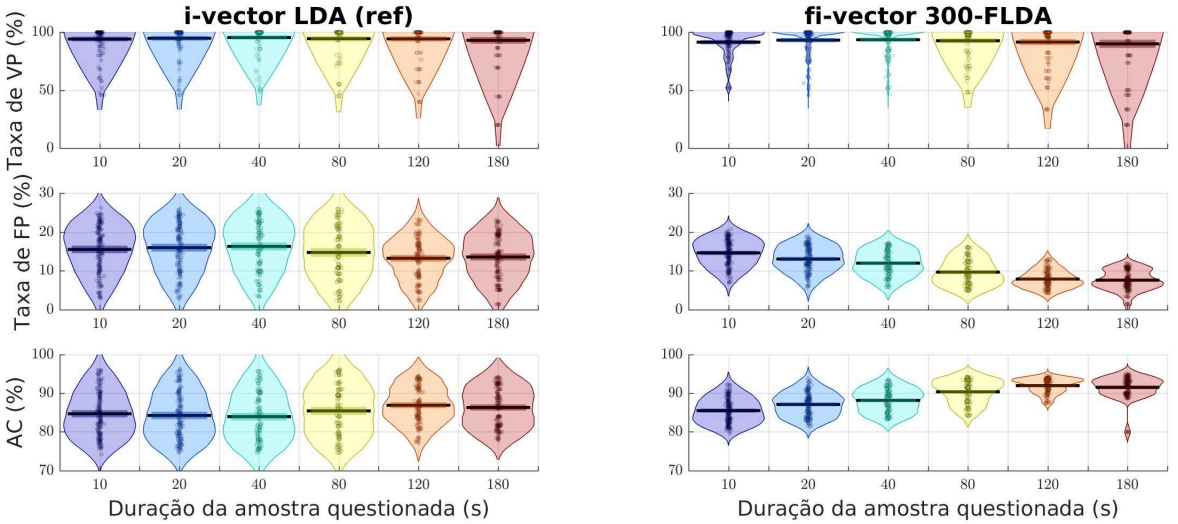


Figura 28 – Curva RDI (*Raw, Description and Inference*) apresentando os valores de verdadeiro positivo (acima), falso positivo (no centro) e acurácia (na base) em relação à duração da amostra questionada. Nos gráficos, em cada coluna, os pontos são os valores individuais, as curvas laterais indicam a distribuição de probabilidade empírica, a linha preta horizontal a média.

Ainda para a etapa de testes, foi elaborado um segundo experimento para avaliar a influência da duração da amostra questionada. A duração do tempo de elocução, medido por após a VAD, foi variada em 10, 20, 40, 80, 120 e 180 segundos em cinco instâncias de subamostragem aleatória para cada duração. Está análise foi comparativa entre a técnica de referência ("i-vectors LDA") e "*fi-vector* 300-FLDA".

A Figura 28 apresenta a curva RDI (*Raw data, Description and Inference*) comparando as taxas de verdadeiro positivo (acima), falso positivo (no centro) e acurácia (na base) em relação à duração da amostra questionada. No painel da esquerda têm-se os resultados para a técnica de referência ("i-vectors LDA") enquanto na direita têm-se os resultados para "*fi-vector* 300-FLDA". Nos gráficos, os pontos em cada coluna são os valores individuais, as curvas laterais apresentam a distribuição de probabilidade empírica e a linha preta horizontal a média.

A primeira diferença entre as metodologias é a dispersão dos resultados que pode ser observado pela distribuição de probabilidade empírica nas curvas laterais. Enquanto a técnica de referência espalha seus resultados em torno da média, com uma variância maior, a técnica "*fi-vector* 300-FLDA"apresenta sempre uma dispersão comparável ou menor. Ainda na Figura 28 as linhas verticais em cada coluna representam as médias. A Figura 29 apresenta a análise de variância da diferença entre as médias.

Nota-se, na análise de variância da Figura 29, que a principal diferença entre a técnica de referência e a "*fi-vector* 300-FLDA" está na taxa de falso positivo e na acurácia. A técnica de "*fi-vector* 300-FLDA" possui um desempenho médio superior à referência (exceto para duração de 10 segundos, que apresenta equivalência) com menor dispersão em torno da média.

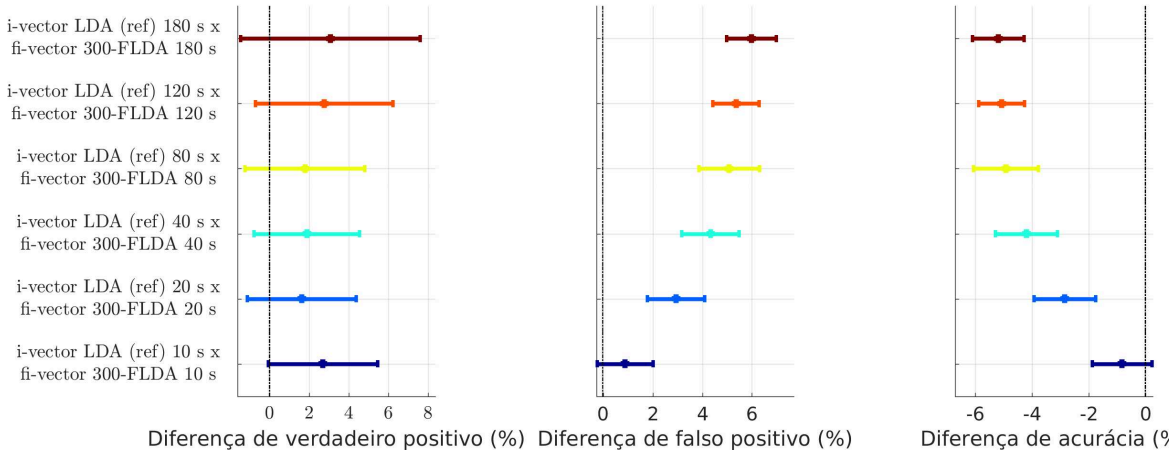


Figura 29 – Análise de variância entre a metodologia de referência e "*fi-vector* 300-FLDA" para as taxas de verdadeiro positivo (à esquerda), falso positivo (centro) e da acurácia (à direita) com $\alpha = 0,05$. Em cada linha a diferença média é representada pelo quadrado e o intervalo de confiança, pela linha horizontal.

Outro recorte do resultado pode ser realizado pelo tipo de ruído como apresentado na Figura 30. O gráfico apresenta as médias das taxas de verdadeiro positivo (acima), falso positivo (no centro) e acurácia (na base) em relação a duração da amostra questionada e ao tipo de ruído para o método de referência e para "*fi-vector* 300-FLDA". Nos gráficos, da esquerda para direita, são agrupados os resultados de acordo com o tamanho da amostra questionada (eixo horizontal inferior) para cada tipo de ruído contaminante (eixo horizontal superior).

Os resultados do gráfico apresentam três pontos de destaque. O primeiro, é o desempenho semelhante das duas técnicas quando contaminadas por ruído branco, pois se nota que o ruído branco afeta em maior grau a taxa de verdadeiro positivo, sendo o desempenho do método de referência para a taxa de falso positivo superior neste tipo de ruído. O segundo, é a queda de desempenho na taxa de verdadeiro positivo da técnica "*fi-vector* 300-FLDA" quando contaminada por ruído de multidão (*babble noise*). O terceiro, é o desempenho médio

superior da técnica de referência ("*i-vector*-LDA") para ruído rosa com amostra questionada de 10 segundos. Estes pontos indicam que a melhoria obtida pela técnica "*fi-vector* 300-FLDA" não é generalizada e não ocorre em alguns dos cenários pesquisados.

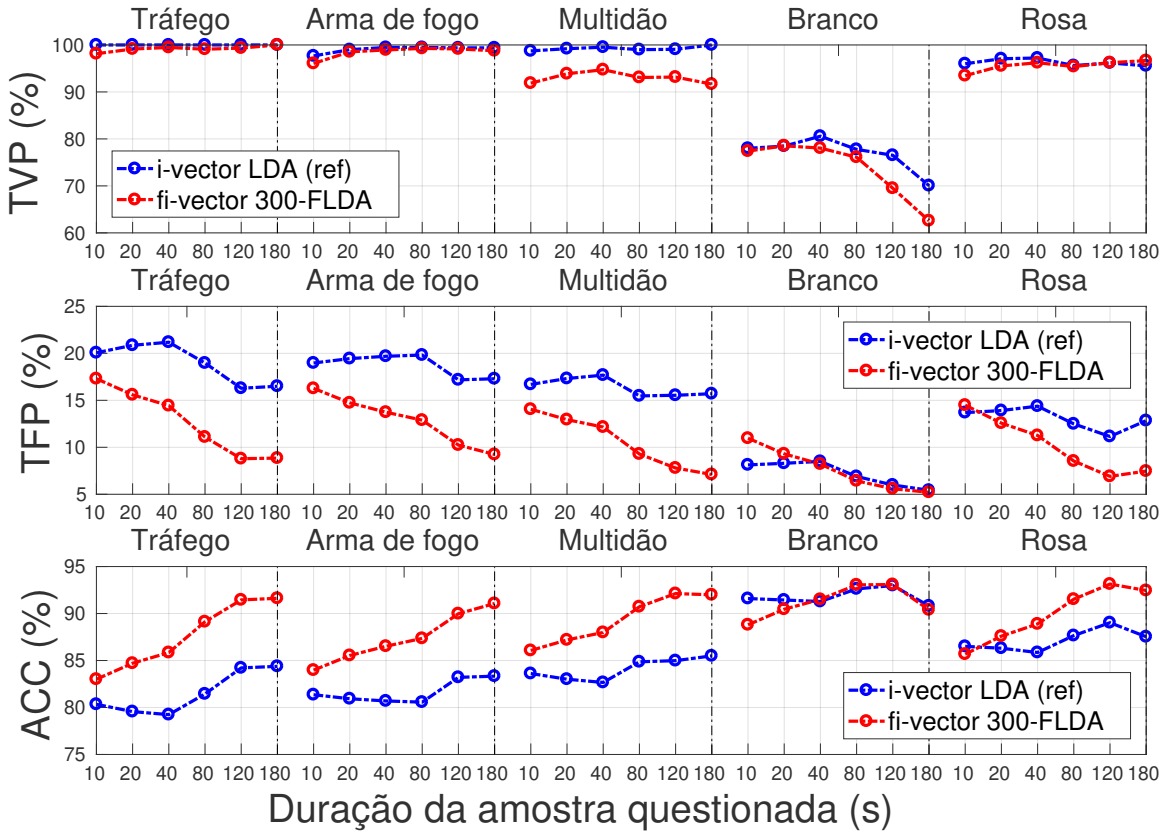


Figura 30 – Média das taxas de verdadeiro positivo (TVP) no painel acima, falso positivo (TFP) ao centro e acurácia (ACC) abaixo em relação à duração da amostra questionada pelo tipo de ruído. Da esquerda para direita são agrupados os resultados de acordo com o tamanho da amostra questionada (eixo horizontal inferior) para cada tipo de ruído contaminante (eixo horizontal superior).

O recorte de acordo com a intensidade do ruído contaminante é apresentado na Figura 31, que foi construída nos mesmos moldes da Figura 30. Na Figura 31, tem-se que as taxas de falso positivo e a acurácia melhoram com o incremento da SNR e com o aumento da amostra questionada. Especificamente para este experimento e para a metodologia "*fi-vector* 300-FLDA", obtém-se acurácia superior a 90% e taxa de falso positivo inferior a 10% em todos os cenários de SNR. Para um áudio questionado com SNR de 17 *dB* obtêm-se essas taxas a partir de 80 segundos de vozeamento. Com SNR de 25 *dB* o mesmo desempenho é obtido a partir de 20 segundos de vozeamento.

A análise de variância, considerando todas as comparações independentemente do tamanho da amostra questionada, é apresentada na Figura 32. Na imagem, nota-se que a técnica de referência (*i-vector*-LDA) apresentou, na média, uma taxa de verdadeiro positivo superior à

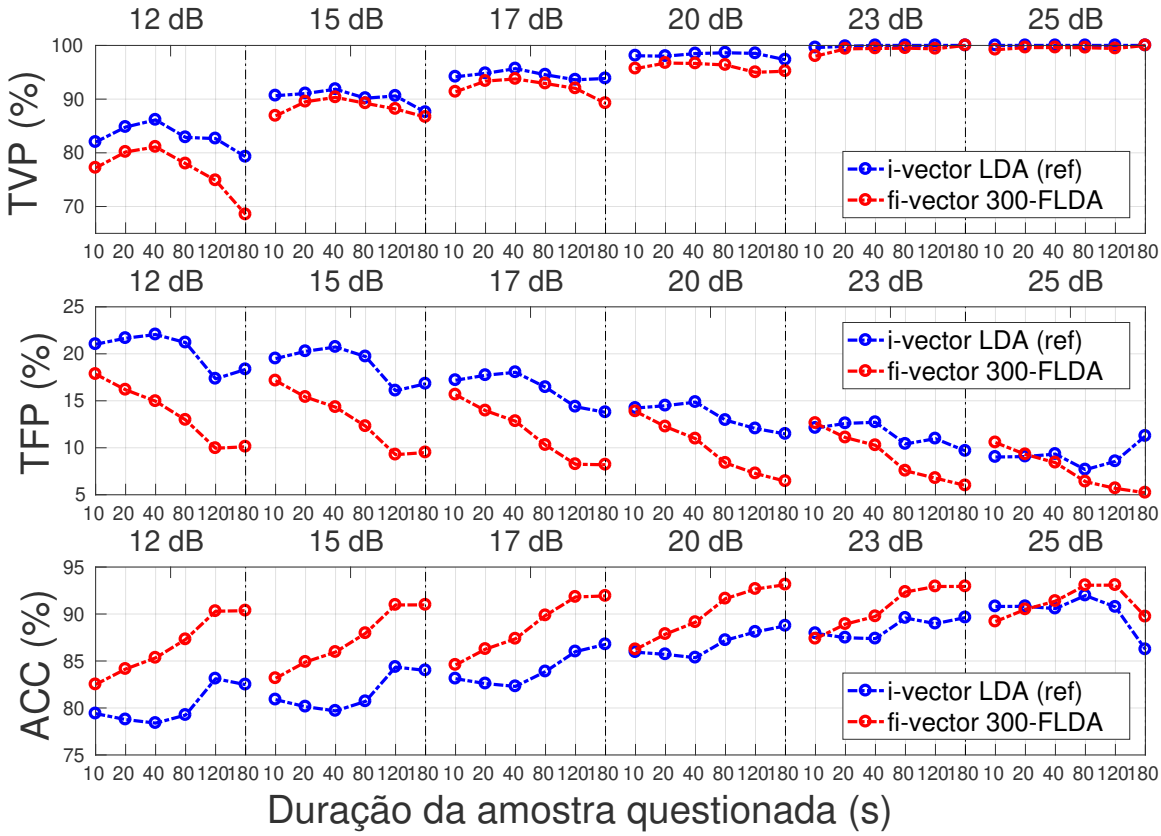


Figura 31 – Médias das taxas de verdadeiro positivo (acima), falso positivo (no centro) e acurácia (na base) em relação à duração da amostra questionada e a intensidade do ruído para o método de referência e para "*fi-vector* 300-FLDA". Da esquerda para direita são agrupados os resultados de acordo com o tamanho da amostra questionada (eixo horizontal inferior) para a intensidade do ruído contaminante (eixo horizontal superior).

técnica de "*fi-vector* 300-FLDA" em 2,3%. Por outro lado, a técnica "*fi-vector* 300-FLDA" apresentou uma acurácia média superior em 3,8% e uma taxa de falso positivo média inferior em 4,1%. Em valores percentuais, comparados com a metodologia de referência, "fi-vector 300-FLDA" apresentou um incremento de acurácia de 4,5% e reduziu as taxas de verdadeiro e falso positivo, respectivamente, em 2,4% e 27,5%.

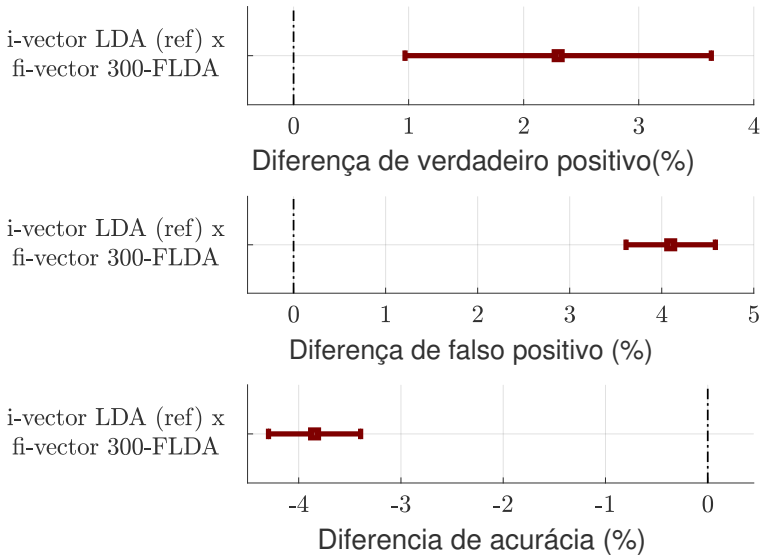


Figura 32 – Análise de variância entre o método de referência e "*fi-vector* 300-FLDA" as taxas de verdadeiro positivo (acima), falso positivo (centro) e da acurácia (abaixo), com $\alpha = 0,05$, considerando apenas os resultados com a duração da amostra questionada. Em cada linha a diferença média é representada pelo quadrado e o intervalo de confiança, pela linha horizontal.

4.3 Conclusões e Principais Pontos

O presente capítulo apresentou os resultados experimentais obtidos nas duas linhas de trabalho que despontaram da pesquisa. O primeiro é a estimativa por intervalo por FBST e o segundo a aplicação de pareamento e técnicas fuzzy a comparação por voz.

A estimativa por intervalo utilizando a formulação do FBST, aqui denominada de intervalo de evidência, mostrou-se prática e aplicável. A formulação apresentada neste trabalho aplica-se a média com variância desconhecida e pode ser obtida para diferentes tamanhos de amostras sem utilizar métodos MCMC (*Markov chain Monte Carlo*). A formulação parece permitir definir uma estimativa por intervalo para o parâmetro da precisão, porém esse não foi o foco do trabalho.

O intervalo de evidência e outras inferências intervalares, quando aplicado à metodologia GMM-UBM de verificação de locutor em conjunto aberto, apresentaram redução na taxa de erro da ordem de $0,8\%$. Este resultado pode ser do fato que as amostras utilizadas na etapa de teste possuírem, em média, uma duração maior de as amostras do corpus CEFALA-1[10], e consequentemente, suas inferências intervalares terem intervalos menores[11].

[10] As gravações cedidas pelo IC-MG têm, em média, 6,8 minutos (vide Tabela 5 no Apêndice A). Já as gravações do corpus CEFALA-1 têm, em média, 2,9 minutos (vide Tabela 6 no Apêndice A).

[11] Nota-se que a Equação 3.27 possui um T no denominador e que a Equação 4.11 uma função gama de n no denominador.

Isto posto gostaria de colocar que o intervalo de evidência, e outras inferências intervalares, ainda podem ser explorados considerando o tamanho da amostra de voz questionada.

Sobre a robustez da CFL, agregar a medida S^2NR combinada com técnicas *fuzzy* melhorou resultados da metodologia *i-vector*, inclusive seus limites e índices de desempenho foram apresentados tanto em relação à contaminação quanto para a duração da amostra.

Nos resultados, a combinação da técnica de FLDA com a aplicação das estatísticas de Baum-Welch *fuzzy*, mais especificamente a técnica "*fi-vector* 300-FLDA", apresentou a maior redução na taxa de falso positivo e aumento da acurácia na etapa de testes sem perdas significativas na taxa de verdadeiro positivo.

Para aplicação de casos de exigência forense, apesar de a utilização de uma base de dados verdadeiramente prática, as propostas ainda precisam passar por mais validação e revisão.

Capítulo 5

Considerações Finais

"How many roads must a man walk down, before you can call him a man?"

— Bob Dylan

5.1 Conclusões

O objetivo principal do trabalho propôs "(...) aprimoramento de métodos automáticos aplicados às condições da CFL utilizando de estimativa por intervalo e de comparação pareada por medidas de relação sinal-ruído (...)". As principais contribuições foram:

1. O intervalo de evidência e outras inferências intervalares, quando aplicados à metodologia GMM-UBM de verificação de locutor em conjunto aberto, apresentaram redução na taxa de erro da ordem de 6,4% (0,8% no valor absoluto). Este valor é entre um e dois exames realizados por anos no SPAV. Neste caso, o intervalo de evidência e outras inferências intervalares ainda podem ser explorados considerando o tamanho da amostra de voz questionada.
2. O intervalo de evidência ainda é uma proposta que depende de mais experimentação e aplicação. Porém as taxas de classificação corretas e incorretas mostraram-se coerentes com a SNR e com o comportamento em conjunto aberto. O método do intervalo de evidência está sendo aplicado, de forma experimental, nos exames realizados pelo Setor de Perícias em Áudio e Vídeo do IC-MG.

3. Em relação ao pareamento na comparação por voz, os experimentos compararam aplicações de técnicas *fuzzy* buscando uma melhoria da metodologia *i-vectors* na CFL, ou seja, em conjunto aberto e independente de texto. Nos experimentos, fixaram-se as características (MFCC), a base de dados de treinamento (CEFALA-1) e os tipos de ruídos e intensidade de contaminação (SNR médio). Isto implica que a informação presente nos resultados é limitada por estes parâmetros.
4. Nos resultados, a combinação da técnica de FLDA com a aplicação das estatísticas de Baum-Welch *fuzzy*, mais especificamente a técnica "*fi-vector* 300-FLDA", apresentou redução da taxa de falsos positivos, de 16,8% para 10,8% (35,7% relativo à metodologia de referência), além de uma melhoria da acurácia em 5,9% (89,3% - 83,4%, ou 7,1% relativo à referência) que, em contrapartida, perde, na média, 1,8% (1,9% relativo à metodologia de referência) na taxa de verdadeiro positivo.
5. Na avaliação que considera o tamanho da amostra questionada, o método "fi-vector 300-FLDA" mostra melhorias em relação à acurácia e à taxa falso-positivo, respectivamente, em 4,1% (4,5% relativo à metodologia de referência) e 3,8% (27,5% relativo à metodologia de referência), perdendo 2,3% (2,4% relativo à metodologia de referência) na taxa de verdadeiro positivo.

Por fim, ressalta-se a contribuição da medida de S^2NR, em especial pela robustez da técnica na faixa de SNR entre 10 dB e 40 dB. A medida confiável permite estabelecer os conjuntos *fuzzy*-S^2NR para um pareamento dos quadros da matriz de MFCC em um espaço contínuo. O pareamento reduz a heterocedasticidade dos dados, permitindo a comparação de medidas (MFCC) obtidas em ambiente de ruído semelhante.

Listam-se os seguintes pontos, desenvolvidos no doutorado, que contribuíram direta ou indiretamente com a prática da CFL:

1. O desenvolvimento do corpus CEFALA-1, que foi amplamente utilizado.
2. O desenvolvimento da solução do FBST sobre a média com variância desconhecida.
3. A redução de 6,4%, na taxa de erros oriundos da aplicação do intervalo de evidência.
4. O desenvolvimento e formulação do pareamento *fuzzy* pela medida S^2NR.
5. A redução na taxa de falso positivo apresentado pela técnica *fi-vector* 300-FLDA.

5.2 Continuidade dos Trabalhos

Os trabalhos aqui apresentados carecem de avaliações de outros efeitos que podem influenciar a voz de um falante, como a contemporaneidade das gravações, a mimetização, e alterações temporárias ou definitivas como doenças, estado mental ou cirurgias.

As duas linhas de trabalho apresentadas neste texto foram as que contribuíram para melhorar os resultados. Algumas abordagens foram experimentadas e sobrestadas ao longo do período de doutorado. Como nem todas as ideias foram colocadas em prática, seguem algumas reflexões que podem dar continuidade ao apresentado nesta tese.

Relacionado ao desenvolvimento a partir da formulação do FBST:

- Notou-se, no desenvolvimento do intervalo de evidência, que também é possível propor uma estimativa por intervalo para a precisão.
- Aplicar o intervalo de evidência para extrair uma estimativa por intervalo sobre a calibração (etapa de treinamento) do reconhecimento de locutor.
- Pode-se realizar experimentos com diferentes características acústicas, como PNCC (*Power Normalized Cepstrum Coefficients*), PLP (*Perceptual Linear Predictive*) ou frequência fundamental, incluindo também elementos segmentais supra-segmentais como, por exemplo, a prosódia.
- Pode-se avaliar a influência do comportamento temporal do ruído (impulsivos como estampidos de arma de fogo ou contínuo como de trânsito) e diferentes durações da amostra questionada.
- Expansão da solução sintética para o problema de Behrens-Fisher[1] e para amostra multidimensionais.

Sobre o pareamento utilizando a medida S^2NR:

- A inclusão de áudios contaminados por ruído no treinamento do UBM e da matriz de variabilidade total **T** para modelar as interferências junto com os *i-vectors*.
- Variação do número de conjuntos *fuzzy*-S^2NR, pois neste trabalho utilizaram-se apenas três conjuntos.
- Estudo mais profundo em relação ao número de fatores mais adequado na etapa da análise discriminante linear *fuzzy*.
- Pareamento orientado na medida S^2NR por banda de frequência.

[1] Problema de estimativa por intervalo sobre a diferença entre as médias de duas amostras independentes com variância não são consideradas iguais.

Referências

AITKEN, C. G.; LUCY, D. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 53, n. 1, p. 109–122, 2004. Citado na página 67.

ARRUDA, M. L. d. *FBST seqüencial*. Tese (Doutorado em Estatística) — Instituto de Matemática e Estatística, University of de São Paulo, 2012. Citado na página 74.

ASSANE, C. C.; PEREIRA, B. d. B.; PEREIRA, C. A. d. B. Model choice in separate families: A comparison between the FBST and the cox test. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 48, n. 9, p. 2641–2654, 2019. Citado 2 vezes nas páginas 73 e 75.

ATAL, B. S. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, IEEE, v. 64, n. 4, p. 460–475, 1976. Citado 2 vezes nas páginas 28 e 41.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics: Investigating language structure and use*. [S.l.]: Cambridge University Press, 1998. Citado na página 113.

BISANI, M.; NEY, H. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: IEEE. *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. [S.l.], 2004. v. 1, p. I–409. Citado na página 72.

BOGERT, B. P.; HEALY, M. J.; TUKEY, J. W. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In: CHAPTER. *Proceedings of the symposium on time series analysis*. [S.l.], 1963. v. 15, p. 209–243. Citado na página 51.

BOLSTAD, W. M. *Introduction to Bayesian Statistics*. [S.l.]: John Wiley & Sons, 2013. Citado na página 78.

BROEDERS, A. Forensic speech and audio analysis forensic linguistics. In: *13th INTERPOL Forensic Science Symposium, Lyon, France*. [S.l.: s.n.], 2001. v. 26. Citado 2 vezes nas páginas 28 e 29.

CAMPBELL, J. P.; SHEN, W.; CAMPBELL, W. M.; SCHWARTZ, R.; BONASTRE, J. F.; MATROUF, D. Forensic speaker recognition, a need for caution. *IEEE Signal Processing Magazine*, n. 95, March 2009. Citado na página 48.

CAMPBELL, W. M.; REYNOLDS, D. A.; CAMPBELL, J. P.; BRADY, K. Estimating and evaluating confidence for forensic speaker recognition. In: IEEE. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. [S.l.], 2005. v. 1, p. I–717. Citado na página 72.

CASELLA, G.; BERGER, R. Inferência estatística. *Centage Learning*, 2011. Citado 2 vezes nas páginas 72 e 78.

COUNCIL, N. R. et al. *Strengthening forensic science in the United States: a path forward*. [S.l.]: National Academies Press, 2009. Citado na página 30.

DEHAK, N.; KENNY, P. J.; DEHAK, R.; DUMOUCHEL, P.; OUELLET, P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 19, n. 4, p. 788–798, 2011. Citado 3 vezes nas páginas 64, 65 e 66.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, JSTOR, p. 1–38, 1977. Citado na página 61.

DENES, P. B.; PINSON, E. N. *The speech chain: the physics and biology of spoken language*. [S.l.]: Waveland Press, 2015. Citado na página 46.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994. Citado 2 vezes nas páginas 72 e 81.

EPHRAIM, Y.; MALAH, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, IEEE, v. 32, n. 6, p. 1109–1121, 1984. Citado na página 50.

FANT, G. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. [S.l.]: Walter de Gruyter, 1971. v. 2. Citado 2 vezes nas páginas 46 e 47.

FERNANDES, J. R. *Perícias em áudios e imagens forenses*. [S.l.]: Domingos Toccheto, 2014. (Tratado de Perícias de Criminalística). Citado na página 33.

FIGUEIREDO, R. M. D. *Identificação de falantes: aspectos teóricos e metodológicos*. Tese (Doutorado) — Programa de Pós-Graduação em Linguística, Universidade Estadual de Campinas, 1994. Citado 2 vezes nas páginas 28 e 29.

FLANAGAN, J. L. *Speech analysis synthesis and perception*. [S.l.]: Springer Science & Business Media, 2013. v. 3. Citado 2 vezes nas páginas 46 e 47.

FRENCH, P.; NOLAN, F.; FOULKES, P.; HARRISON, P.; MCDOUGALL, K. The UK position statement on forensic speaker comparison; a rejoinder to Rose and Morrison. *International Journal of Speech Language and the Law.*, v. 17, n. 1, p. 143–152, 2010. Citado na página 48.

FURUI, S. *Digital speech processing: synthesis, and recognition*. [S.l.]: Taylor & Francis, 2000. ISBN 9781420002669. Citado 2 vezes nas páginas 46 e 48.

GAJIC, B.; PALIWAL, K. K. Robust parameters for speech recognition based on subband spectral centroid histograms. In: *Seventh European Conference on Speech Communication and Technology*. [S.l.: s.n.], 2001. Citado na página 53.

GARCIA-ROMERO, D.; ESPY-WILSON, C. Y. Analysis of i-vector length normalization in speaker recognition systems. In: *Twelfth Annual Conference of the International Speech Communication Association*. [S.l.: s.n.], 2011. Citado 2 vezes nas páginas 66 e 89.

GOLD, E.; FRENCH, P. International practices in forensic speaker comparison. *International Journal of Speech Language and the Law.*, v. 18, n. 2, p. 293–307, 2011. Citado na página 29.

GONZALEZ-RODRIGUEZ, J.; ROSE, P.; RAMOS, D.; TOLEDANO, D. T.; ORTEGA-GARCIA, J. Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 15, n. 7, p. 2104–2115, 2007. Citado 2 vezes nas páginas 30 e 42.

HANSEN, J. H.; HASAN, T. Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Processing Magazine*, IEEE, v. 32, n. 6, p. 74–99, 2015. Citado 5 vezes nas páginas 28, 29, 53, 60 e 72.

HARSANYI, Z. *A vida de Galileu:(o contemplador de estrelas)*. [S.l.]: Livraria José Olympio, 1957. Citado na página 117.

HATCH, A. O.; KAJAREKAR, S.; STOLCKE, A. Within-class covariance normalization for SVM-based speaker recognition. In: *Ninth international conference on spoken language processing*. [S.l.: s.n.], 2006. Citado na página 89.

HAUENSTEIN, M. *Speech Coding*. 2003. <http://http://www.markus-hauenstein.de>. [Online; acessado em 12-Março-2016]. Citado na página 58.

HERMANSKY, H. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, ASA, v. 87, n. 4, p. 1738–1752, 1990. Citado na página 53.

HERMANSKY, H.; MORGAN, N. RASTA processing of speech. *IEEE transactions on speech and audio processing*, IEEE, v. 2, n. 4, p. 578–589, 1994. Citado na página 53.

HOLLIEN, H. *The acoustics of crime: the new science of forensic phonetics*. [S.l.]: Springer Science & Business Media, 1990. Citado 3 vezes nas páginas 28, 33 e 47.

ITU, G. *Pulse code modulation (PCM) of voice frequencies. ITU-T G.711*. [S.l.]: International Telecommunications Union (ITU), 1988. Citado na página 57.

ITU, G. *GSM full rate speech transcoding. GSM Rec 06.10*. [S.l.]: International Telecommunications Union (ITU), 1991. Citado 2 vezes nas páginas 57 e 89.

ITU, G. *Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding. ETSI-EN-301-704 V7.2.1*. [S.l.]: International Telecommunications Union (ITU), 2000. Citado na página 58.

ITU, G. *Digital cellular telecommunications system (phase 2+); Enhanced Full Rate (EFR) speech transcoding. ETSI-EN-300-726 V8.0.1*. [S.l.]: International Telecommunications Union (ITU), 2000. Citado 2 vezes nas páginas 58 e 59.

KACUR, J.; VARGA, M.; ROZINAJ, G. ZCPA features for speech recognition. In: IEEE. *Telecommunications (BIHTEL), 2012 IX International Symposium on*. [S.l.], 2012. p. 1–4. Citado na página 53.

KENNY, P. A small footprint i-vector extractor. In: *Odyssey 2012-The Speaker and Language Recognition Workshop*. [S.l.: s.n.], 2012. Citado 2 vezes nas páginas 64 e 89.

KENNY, P.; BOULIANNE, G.; DUMOUCHEL, P. Eigenvoice modeling with sparse training data. *IEEE transactions on speech and audio processing*, IEEE, v. 13, n. 3, p. 345–354, 2005. Citado na página 64.

KENNY, P.; OUELLET, P.; DEHAK, N.; GUPTA, V.; DUMOUCHEL, P. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 16, n. 5, p. 980–988, 2008. Citado na página 64.

KERSTA, L. G. Voiceprint identification. *Nature*, v. 34, p. 1253, 1962. Citado na página 41.

KIM, C.; STERN, R. M. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, IEEE Press, v. 24, n. 7, p. 1315–1329, 2016. Citado na página 53.

KIRK, P. L. *Crime investigation*. [S.l.]: Wiley, 1974. Citado na página 37.

KIZHANATHAM, A. R. *Detection of co-channel speech and usable speech*. Tese (Doutorado) — Temple University, 2003. Citado na página 50.

KOVAL, S.; LOKHANOVA, A. Confidence bounds curves as a tool for evaluation of automatic speaker recognition results uncertainty. In: *Proc. 14th Intern. Conf. on Speech and Computer. SPECOM*. [S.l.: s.n.], 2011. p. 284–289. Citado na página 72.

LOCARD, E. *A investigação criminal e os métodos científicos*. [S.l.]: Saraiva, 1939. Tradução de Fernando de Miranda. Citado na página 32.

LYON, R. F.; KATSIAMIS, A. G.; DRAKAKIS, E. M. History and future of auditory filter models. In: IEEE. *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. [S.l.], 2010. p. 3809–3812. Citado na página 51.

MADRUGA, M. R.; PEREIRA, C. d. B.; STERN, J. M. Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference*, Elsevier, v. 117, n. 2, p. 185–198, 2003. Citado 4 vezes nas páginas 71, 72, 73 e 75.

MAHER, R. C. Audio forensic examination. *IEEE Signal Processing Magazine*, IEEE, v. 26, n. 2, p. 84–94, 2009. Citado 2 vezes nas páginas 33 e 38.

MAKHOUL, J. Linear prediction: a tutorial review. *Proceedings of the IEEE*, IEEE, v. 63, n. 4, p. 561–580, 1975. Citado na página 51.

MARTIN, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, IEEE, v. 9, n. 5, p. 504–512, 2001. Citado na página 50.

MARTINS, F.; SIMÕES, D. R.; BRISSOS, F.; RODRIGUES, C. A fonética forense na produção de prova do ordenamento jurídico português: O parâmetro do pré-vozeamento. *Revista Virtual de Estudos da Linguagem-ReVEL*, v. 12, n. 23, p. 44–70, 2014. Citado na página 30.

MEDASANI, S.; KIM, J.; KRISHNAPURAM, R. An overview of membership function generation techniques for pattern recognition. *International Journal of approximate reasoning*, Elsevier, v. 19, n. 3-4, p. 391–417, 1998. Citado na página 90.

MENDES, L. B. *Documentoscopia*. [S.l.]: Sagra Luzzatto, 1999. Citado na página 30.

MONTGOMERY, D. C.; RUNGER, G. C. *Applied statistics and probability for engineers*. [S.l.: s.n.], 2010. Citado na página 42.

MORRISON, G. S. Forensic voice comparison and the paradigm shift. *Science & Justice*, Elsevier, v. 49, n. 4, p. 298–308, 2009. Citado na página 30.

MORRISON, G. S. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: multivariate kernel density (MVKD) versus gaussian mixture model–universal background model (GMM–UBM). *Speech Communication*, Elsevier, v. 53, n. 2, p. 242–256, 2011. Citado 3 vezes nas páginas 63, 67 e 72.

MORRISON, G. S. Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, Elsevier, v. 51, n. 3, p. 91–98, 2011. Citado na página 81.

MORRISON, G. S. Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science and Justice*, n. 54, p. 245–256, 2014. Citado 2 vezes nas páginas 29 e 30.

MORRISON, G. S.; SAHITO, F. H.; JARDINE, G.; DJOKIC, D.; CLAVET, S.; BERGHS, S.; DORNY, C. G. Interpol survey of the use of speaker identification by law enforcement agencies. *Forensic science international*, Elsevier, v. 263, p. 92–100, 2016. Citado 2 vezes nas páginas 29 e 38.

MORRISON, G. S.; THIRUVARAN, T.; EPPS, J. Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. In: *Odyssey*. [S.l.: s.n.], 2010. p. 12. Citado na página 72.

MORRISON, G. S.; ZHANG, C.; ROSE, P. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic science international*, Elsevier, v. 208, n. 1, p. 59–65, 2011. Citado na página 72.

NETO, A. F.; SILVA, A. P.; YEHIA, H. C. Corpus CEFALA-1: Base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia / corpus CEFALA-1: Audiovisual database of speakers for biometric, phonetic and phonology studies. *REVISTA DE ESTUDOS DA LINGUAGEM*, Faculdade de Letras da UFMG, v. 27, n. 1, p. 191, jan 2019. Citado 4 vezes nas páginas 24, 80, 89 e 116.

NEUMANN, C.; CHAMPOD, C.; PUCH-SOLIS, R.; EGLI, N.; ANTHONIOZ, A.; BROMAGE-GRIFFITHS, A. Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of forensic sciences*, Wiley Online Library, v. 52, n. 1, p. 54–64, 2007. Citado na página 37.

NEUSTEIN, A.; PATIL, H. A. *Forensic speaker recognition: law enforcement and counter-terrorism*. [S.l.]: Springer Science & Business Media, 2011. Citado na página 48.

NIESSEN, M. Speaker specific features in vowels. *Yayınlanmamış Doktora Tezi. Groningen*, Citeseer, 2004. Citado 2 vezes nas páginas 46 e 53.

OLIVEIRA, L. P. de. Linguística de corpus: teoria, interfaces e aplicações. 2009. Citado na página 113.

OLIVEIRA, N. L.; PEREIRA, C. A. d. B.; DINIZ, M. A.; POLPO, A. A discussion on significance indices for contingency tables under small sample sizes. *PloS one*, Public Library of Science, v. 13, n. 8, 2018. Citado 2 vezes nas páginas 73 e 75.

O'SHAUGHNESSY, D. *Speech communication: human and machine*. [S.l.]: Universities press, 1987. Citado na página 51.

PEREIRA, C. A. de B.; STERN, J. M. Evidence and credibility: full bayesian significance test for precise hypotheses. *Entropy*, Molecular Diversity Preservation International, v. 1, n. 4, p. 99–110, 1999. Citado 2 vezes nas páginas 31 e 70.

PETRI, C. *Relação entre níveis de significância Bayesiano e freqüentista: e-value e p-value em tabelas de contingência*. Dissertação (Mestrado) — Universidade de São Paulo, 2007. Citado 3 vezes nas páginas 71, 73 e 75.

PIERANGELO, R.; GIULIANI, G. *Special education eligibility: A step-by-step guide for educators*. [S.l.]: Corwin Press, 2007. Citado na página 45.

PLATT, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, Cambridge, MA, v. 10, n. 3, p. 61–74, 1999. Citado na página 72.

PRINCE, S. J.; ELDER, J. H. Probabilistic linear discriminant analysis for inferences about identity. In: IEEE. *2007 IEEE 11th International Conference on Computer Vision*. [S.l.], 2007. p. 1–8. Citado na página 66.

PYREK, K. *Forensic science under siege: the challenges of forensic laboratories and the medico-legal investigation system*. [S.l.]: Academic Press, 2010. Citado na página 37.

REHDER, M. I.; CAZUMBÁ, L. F.; CAZUMBÁ, M. *Identificação de falantes - uma introdução à fonoaudiologia forense*. [S.l.]: REVINTER, 2015. Citado na página 48.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, Elsevier, v. 10, n. 1, p. 19–41, 2000. Citado 8 vezes nas páginas 29, 42, 59, 60, 61, 62, 63 e 72.

ROBERTSON, B.; VIGNAUX, G. A.; BERGER, C. E. *Interpreting evidence: evaluating forensic science in the courtroom*. [S.l.]: John Wiley & Sons, 2016. Citado na página 37.

RODRÍGUEZ, D. P. N.; APOLINÁRIO, J. A.; BISCAINHO, L. W. P. Audio authenticity: detecting ENF discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security*, IEEE, v. 5, n. 3, p. 534–543, 2010. Citado na página 33.

ROEDERER, J. G. *The physics and psychophysics of music: an introduction*. [S.l.]: Springer Science & Business Media, 2008. Citado na página 51.

ROSE, P. *Forensic Speaker Identification*. [S.l.]: CRC Press, 2003. Citado 3 vezes nas páginas 38, 41 e 48.

ROSE, P.; MORRISON, G. et al. A response to the UK position statement on forensic speaker comparison. *The international journal of speech, language and the law*, v. 16, n. 1, p. 139, 2009. Citado na página 41.

SAFERSTEIN, R. *Criminalistics: an introduction to forensic science*. [S.l.]: Prentice Hall, 2004. Citado 2 vezes nas páginas 36 e 38.

SAKS, M. J.; KOEHLER, J. J. The individualization fallacy in forensic science evidence. *Vanderbilt Law Review*, v. 61, n. 1, p. 199–219, 2008. Citado 2 vezes nas páginas 30 e 37.

SILVA, A. P. Avaliação do reconhecimento de falantes em classes de sons semelhantes [evaluation of speech recognition in classes of similar sounds]. In: SINPOSC (Ed.). *Anais do XI Seminário Nacional de Fonética Forense*. [S.l.: s.n.], 2016. p. 11–13. Citado na página 24.

SILVA, A. P. *Contribuições para robustez e confiabilidade na comparação forense de locutores - Exame de qualificação*. 2017. Universidade Federal de Minas Gerais. Citado na página 24.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Aplicação de estatísticas de formantes e MVKD à comparação forense de locutor. In: SOCIEDADE BRASILEIRA DE TELECOMUNICAÇÕES. *XXXVI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. [S.l.], 2018. v. 1, p. 56–60. Citado 3 vezes nas páginas 54, 67 e 86.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Avaliação de descritores acústicos utilizando estatística MVKD aplicada à comparação forense de locutor. In: *XXII Congresso Brasileiro de Automática*. [S.l.: s.n.], 2018. Citado na página 67.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Comparação forense de locutor pareada por medidas instantâneas de relação sinal ruído. In: *XXVIII ENCONTRO DA SOBRAC*. [S.l.: s.n.], 2018. Citado na página 24.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Resultados da busca por robustez e confiabilidade na comparação forense de locutores. In: SINDPECO - SINDICATO DOS PERITOS OFICIAIS CRIMINAIS DO ESTADO DO MATO GROSSO. *Anais do XII Seminário Nacional de Fonética Forense*. [S.l.], 2018. Citado na página 24.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Robustness in forensic speaker comparison: the great quest. In: *6th EICEFALA*. [S.l.: s.n.], 2018. Citado na página 24.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Avaliação de descritores acústicos em simulação de condições forenses de verificação de locutor. *Revista Brasileira de Crimianlística*, v. 8, n. 2, p. 22–35, 2019. Citado 2 vezes nas páginas 24 e 53.

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Forensic speaker comparison using evidence interval in full bayesian significance test. *Matheamtical Problems in Enginner*, 2020. Citado na página 24.

SILVA, A. P.; VIEIRA, M. N.; CAMPOLINA, T. d. A. M. Análises de classes acústicas semelhantes da fala e aplicações na comparação de falantes. In: SOBRAC. *Anais do Encontro da Sociedade Brasileira de Acústica*. [S.l.], 2017. p. 1168–1176. Citado na página 24.

SOHN, J.; KIM, N. S.; SUNG, W. A statistical model-based voice activity detection. *IEEE signal processing letters*, IEEE, v. 6, n. 1, p. 1–3, 1999. Citado 8 vezes nas páginas 49, 50, 60, 64, 81, 89, 115 e 119.

STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 8, n. 3, p. 185–190, 1937. Citado na página 51.

STUDENT. The probable error of a mean. *Biometrika*, Journal Storage (JSTOR), p. 1–25, 1908. Citado na página 72.

SUGAMURA, N.; ITAKURA, F. Speech data compression by LSP speech analysis-synthesis technique. *transactions of Institute of Electronics, Information and Communication Engineers (IECE)*, v. 81, n. 8, p. J64A, 1981. Citado na página 58.

TITZE, I. R. *Principles of voice production*. [S.l.]: Prentice-Hall, 1994. Citado na página 27.

TOGNERI, R.; PULLELLA, D. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits And Systems Magazine*, Second Quarter 2011. Citado 2 vezes nas páginas 29 e 52.

VALENTE, C. R. Perspectivas da fonética forense num cenário de quebra do dogma da unicidade. In: *Anais da Conferência Internacional de Ciências Forenses em Multimídia e Segurança Eletrônica*. [S.l.: s.n.], 2012. p. 7–27. Citado na página 30.

VALLURI, S. R.; JEFFREY, D. J.; CORLESS, R. M. Some applications of the lambert w function to physics. *Canadian Journal of Physics*, NRC Research Press, v. 78, n. 9, p. 823–831, 2000. Citado na página 76.

VIEIRA, M. N.; SANSAO, J. P. H.; YEHIA, H. C. Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms. *Speech Communication*, Elsevier, v. 61, p. 17–32, 2014. Citado 5 vezes nas páginas 53, 54, 56, 89 e 90.

WU, X.-H.; ZHOU, J.-J. Fuzzy discriminant analysis with kernel methods. *Pattern Recognition*, Elsevier, v. 39, n. 11, p. 2236–2239, 2006. Citado na página 89.

YOUNG, M. A.; CAMPBELL, R. A. Effects of context on talker identification. *The Journal of the Acoustical Society of America*, v. 42, n. 6, p. 1250–1254, 1967. Citado na página 28.

ZABELL, S. L. On student's 1908 article "the probable error of a mean". *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 481, p. 1–7, 2008. Citado na página 72.

ZHI, X.-b.; FAN, J.-l.; ZHAO, F. Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm. *Pattern Recognition*, Elsevier, v. 46, n. 6, p. 1604–1615, 2013. Citado 3 vezes nas páginas 86, 87 e 89.

APÊNDICE A

Conjunto de Dados e Corpus

> *"If I had to choose, I would prefer to be a descendant of a humble monkey rather than of a man who employs his knowledge and eloquence in misrepresenting those who are wearing out their lives in the search for truth."*
>
> — Thomas Huxley

Este apêndice tem como objetivo descrever brevemente algumas características das bases de dados utilizadas nos experimentos descritos no presente trabalho. Para o caso de gravações de voz, a base de dados também é denominada corpus linguístico e definida por Oliveira (2009 apud BIBER et al., 1998, p. 49)

> (...) coleções de textos que ocorrem naturalmente na língua, organizadas sistematicamente para representar áreas de uso da língua, e das quais podemos extrair novas informações.

As bases de dados utilizadas nos experimentos partem de dois conjuntos distintos:

- O primeiro, doravante denominado Corpus Criminal-Contínuo, trata de amostras de voz que foram utilizados em exames periciais de identificação/verificação de locutor que tramitaram pelo Setor de Perícias em Áudio e Vídeo (SPAV) do Instituto de Criminalística de Minas Gerais (IC-MG);
- o segundo é o Corpus CEFALA-1, coletado pela equipe do laboratório CEFALA[1].

[1] Centro de Estudos da Fala, Acústica, Linguagem e Música, Escola de Engenharia, UFMG.

Isto posto segue a descrição detalhada de cada corpus.

A.1 Corpus Criminal-Contínuo

Como supra citado, o Corpus Criminal-Contínuo é composto por amostras de voz que foram utilizadas para CFL. Do ponto de vista legal, muitas metainformações deste corpus são legalmente protegidas[2] observando os princípios da constituição[3] e do sigilo de acordo com o interesse da sociedade[4]. Esse corpus engloba não apenas casos criminais, mas também casos administrativos e eleitorais.

Do ponto de vista técnico, o Corpus Criminal-Contínuo é dinâmico e agrega gravações assim que são disponibilizadas pelo IC-MG. Dentro de todo o material, parte dos registros são áudios apresentados como questionados, de diferentes fontes (e.g., gravação ambiente, conteúdo multimídia, radiodifusão, denúncias anônimas), mas a maioria é oriunda de interceptação telefônica judicialmente autorizada. Essa porção do Corpus Criminal-Contínuo não é controlada e sofre influência de uma enorme variedade de fatores, sendo o canal de comunicação o mais comum.

Outra parte do Corpus Criminal-Contínuo é áudio padrão que foi coletado por profissionais seguindo um protocolo padronizado de coleta. O procedimento de coleta é realizado utilizando microfone da marca SHURE, modelo SM58, digitalizado por placa de áudio marca EDIROL modelo UA-25 em um canal (*mono*), codificação PCM (*Pulse Code Modulation*) com frequência de amostragem de 44,1 kHz e 16 bits de profundidade para caracterização da amplitude.

Dentro do protocolo de coleta do material padrão pode-se elencar as seguintes características:

- Buscar o posicionamento sociolinguístico, como o local de nascimento, criação e quanto tempo residiu em tais localidades. Caso os pais do fornecedor tenham locais de nascimento e criação distintos, ou múltiplos, verificar quanto tempo residiram em tais localidades;
- Verificar se o fornecedor tem um histórico de patologias ou doença à época dos exames, principalmente em relação ao trato vocal, pulmão, garganta boca ou nariz ou se já realizou algum tipo de intervenção cirúrgica ou tratamento fonoaudiólogo;
- Verificar se o fornecedor toma, regularmente, algum remédio controlado. Em caso positivo, solicitar o nome do remédio e se este apresenta algum efeito colateral;

[2] Art. 31 da Lei nº 12.527 de 18 de novembro de 2011.
[3] Art. 93, inciso IX da Constituição da República Federativa do Brasil de 1988.
[4] Art. 20 do Decreto-Lei nº 3.689 de 3 de outubro de 1941

- Verificar se o fornecedor apresenta todos os dentes. Em caso negativo, verificar quais os dentes omissos ou se utiliza algum tipo de prótese dentária ou realizou tratamentos dentários;
- Verificar se o fornecedor é ou foi fumante, usuário de substância tóxica ou de bebida alcoólica. Em caso positivo, indicar quanto tempo e qual a frequência e quantidade diária do consumo. E em caso de já não utilizar estas substâncias, verificar a quanto tempo.
- A linguagem a ser apresentada pelo fornecedor deve ser o vernáculo, informal e descuidado. Para tanto, é necessário que o fornecedor esteja relaxado. Como a linguagem utilizada pelo coletor influencia sobremaneira a escolha da linguagem a ser utilizada pelo fornecedor, o coletor deve sempre que possível, utilizar-se também de linguagem vernacular. Caso o fornecedor não utilize do vernáculo, o coletor deve sempre estender a coleta no intuito de direcionar o fornecedor às proximidades da fala descuidada e informal. E caso o fornecedor insista em não atingir tal linguagem, o coletor deve lançar mão de artifícios emocionais para que o fornecedor se aproxime do vernáculo.

É importante observar que está em estudo a viabilização da publicação do Corpus Criminal-Contínuo. Para tal, todas as amostras de voz (questionada e padrão) precisam obedecer critérios legais que podem incluir a remoção de metainformações, incluindo a rastreabilidade do locutor em relação ao procedimento pericial relacionado.

O Corpus Criminal-Contínuo apresenta um total de 83 amostras coletadas para serem utilizadas como padrão de voz, sendo que 70 foram fornecidas por indivíduos do sexo masculino e 13 por indivíduos do sexo feminino.

Sobre a duração dos registros de áudio, os áudios do tipo padrão possuem maior duração devido ao controle sobre a gravação, entretanto, a média líquida é de 6,8 minutos e metade dos áudios possuem tempo de fala líquida superior a 6,3 minutos, como apresentado na Tabela 5. A duração dos registros de áudio já separados, antes e após o processamento por VAD (SOHN et al., 1999) é apresentado na Figura 33.

Tabela 5 – Estatísticas temporais dos áudios padrão do corpus Criminal-Contínuo, em minutos, antes e após processamento VAD.

	Tempo do áudio (minutos)	Tempo do áudio após VAD (minutos)
Mínimo	2,1	1,5
Máximo	27,6	18,4
Média	8,6	6,8
Mediana	8,0	6,3
Desvio padrão	4,5	3,4

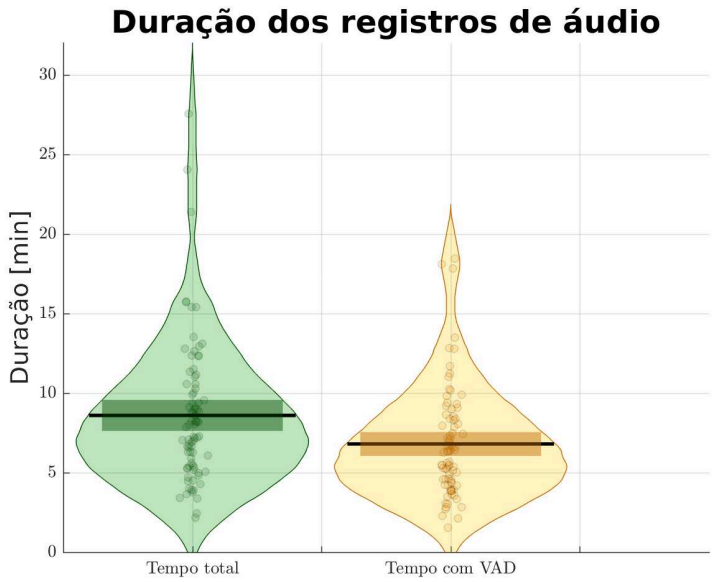


Figura 33 – Gráfico RDI apresentando a duração dos áudios padrão, antes e depois da detecção de atividade de voz. Em cada coluna os pontos são as durações individuais, as curvas laterais são a distribuição de probabilidade empírica, a linha preta horizontal a média e o retângulo escuro é o intervalo de confiança da média para $\alpha = 0,05$.

A.2 Corpus Cefala-1

O Corpus CEFALA-1 é uma base de vozes com com objetivo de privilegiar a diversidade de falantes. O material consiste no registro de 104 participantes, sendo 55 do sexo masculino e 49 do sexo feminino. As coletas foram realizadas em um ambiente controlado, com ruído de fundo de 34 dBA de pressão sonora[5].

A instrumentação de gravação consistiu de quatro capturas de áudio e uma captura audiovisual. As três principais capturas de áudio foram realizadas com uma placa de aquisição sincronizada marca M-Audio FireWire modelo 1814 em três canais, codificação PCM (*Pulse Code Modulation*) com frequência de amostragem 44,1 kHz e 16 bits de profundidade para caracterização da amplitude. Para o primeiro canal foi utilizado um microfone sem fio marca STANER modelo SW-481 para captura do áudio ambiente. No segundo, um microfone de lapela, transdução eletreto-condensador marca DYLAN modelo DL-09 posicionado no peitoral do locutor a aproximadamente 20 cm dos lábios. No terceiro, utilizou-se um microfone condensador marca Brüel & Kjær número de série 1430698 posicionado na altura dos lábios a uma distância de 15 *cm* com 45^o à direita do plano sagital. A quarta captura de áudio foi realizada utilizando o microfone externo (viva-voz) de um aparelho celular marca Samsung modelo Galaxy S2 Lite GT-i9070, em um canal, codificação PCM (*Pulse Code Modulation*)

[5] Mais informações e resultados oriundos do processamento do Corpus CEFALA-1 podem ser obtidas em (NETO et al., 2019). Para obter acesso aos dados do corpus acesse https://corpus.cefala.org (acessado em 14/01/2020).

com frequência de amostragem de 44,1 kHz e 16 bits de profundidade para caracterização da amplitude. O aparelho ficava localizado a aproximadamente 80 cm a frente do plano coronal a altura do pescoço.

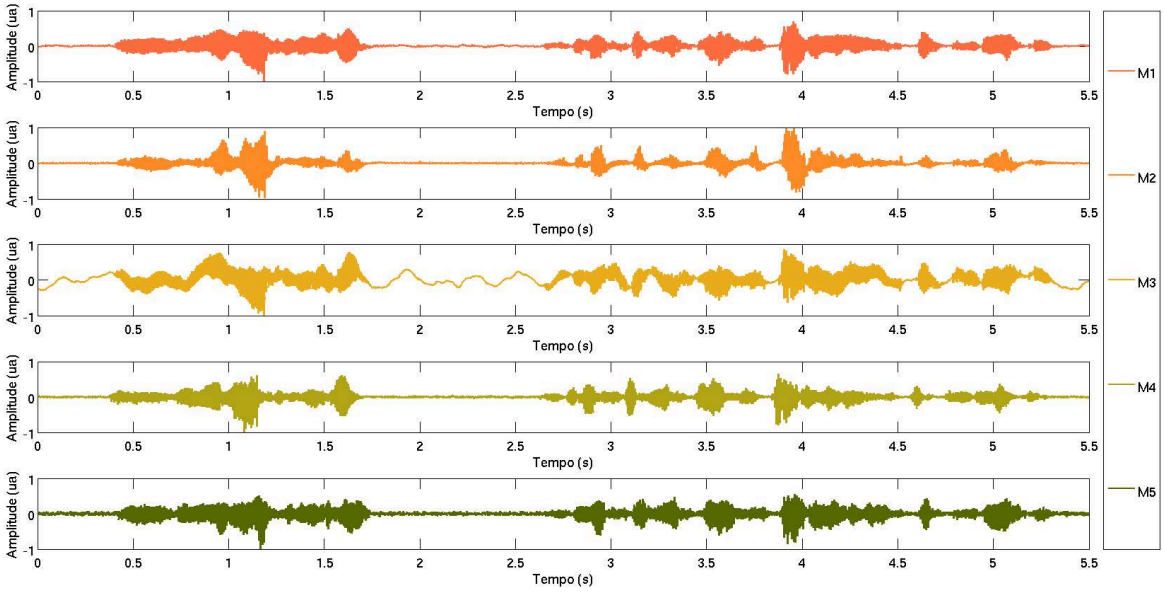


Figura 34 – Detalhe de um trecho de áudio com duração de 5,5 segundos, captado por cada um dos microfones do corpus CEFALA-1, exemplificando o resultado do processo de alinhamento dos registros de áudio de cada microfone.

A captura audiovisual foi realizada através de uma câmera marca GoPro, modelo Hero 3 + Black Edition. A captura de vídeo teve resolução de 1280x720 *pixels* a 60 quadros por segundo (FPS - *Frames per Second*), codificação de vídeo H264. O áudio foi gravado em um canal, codificação AAC (*Advanced Audio Coding*) com frequência de amostragem de 48 kHz e 32 bits de profundidade para caracterização da amplitude.

O protocolo utilizado consiste em três etapas distintas, sendo:

- Fala espontânea: nesta etapa o locutor foi orientado a discorrer a respeito de um assunto de seu interesse, por cerca de 2 minutos. A finalidade desta primeira etapa foi que o locutor atingisse o estado habitual, imprimindo no registro acústico aspectos emocionais, seu sotaque, gírias, entre outros;
- Leitura de texto: nesta etapa foi solicitado aos locutores que realizassem a leitura de um mesmo trecho, com 153 palavras, do livro A vida de Galileu: o contemplador de estrelas (HARSANYI, 1957); e
- Leitura de frases: momento em que foi solicitado aos locutores que realizassem a leitura de vinte frases com pronúncia intervalada. A lista de frases apresenta uma média de sete palavras por frase, sendo três o mínimo e 19 o máximo de palavras por frase.

Realizada a coleta do material audiovisual de todos os locutores, a tarefa seguinte focou-se no processamento e organização dos dados obtidos. O primeiro passo foi a extração do conteúdo de áudio dos registros obtidos pela câmera e sua conversão para o formato padrão da base de dados: formato WAV PCM com taxa de amostragem de 44,1 kHz e 16 bits por amostra. A etapa seguinte foi o alinhamento dos registros de áudio através de um pulso de sincronia utilizado para marcar o início da gravação em cada um dos microfones. O resultado do processo é exemplificado na Figura 34.

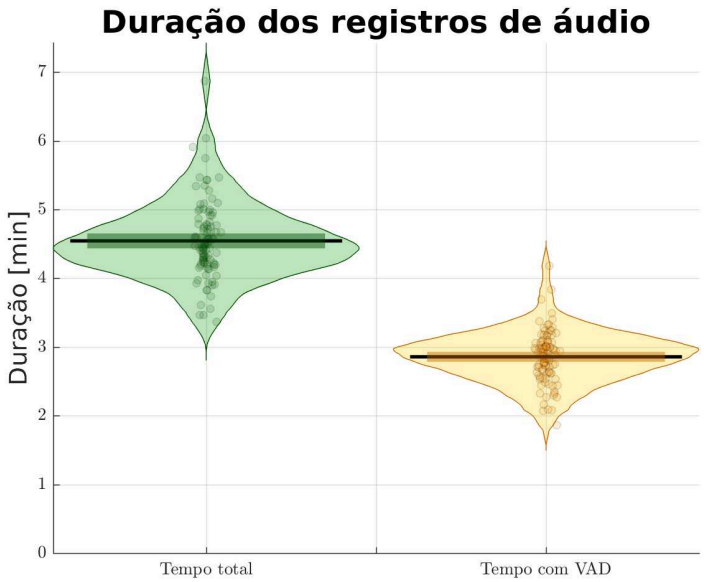


Figura 35 – Gráfico RDI apresentando a duração dos áudios do Corpus CEFALA-1 antes e depois do processamento VAD. Em cada coluna os pontos são as durações individuais, as curvas laterais são a distribuição de probabilidade empírica, a linha preta horizontal a média e o retângulo escuro é o intervalo de confiança da média para $\alpha = 0,05$.

O material sonoro bruto, que consiste no período completo de gravação, foi padronizado em comprimento através de um sinal de sincronia, de forma que para todos os locutores os cinco registros de áudio tratado possuem a mesma duração. A distribuição da duração dos áudios pode ser observada na Figura 35.

Em seguida foi feita a divisão nas três subcategorias (etapas) de acordo com o protocolo de coleta (fala espontânea, leitura de texto e leitura de frases). A duração de cada amostra de cada etapa pode ser observada na Figura 36.

Sobre a duração dos registros de áudio, tem-se que, na média, os arquivos possuem 4,5 minutos com tempo de fala de 2,9 minutos. Outro fato observado é que, na média, a etapa de fala espontânea apresentou uma quantidade de fala maior. A Tabela 6 indica os valores mínimos, máximos, média, mediana e desvio padrão dos arquivos de áudio e de cada uma

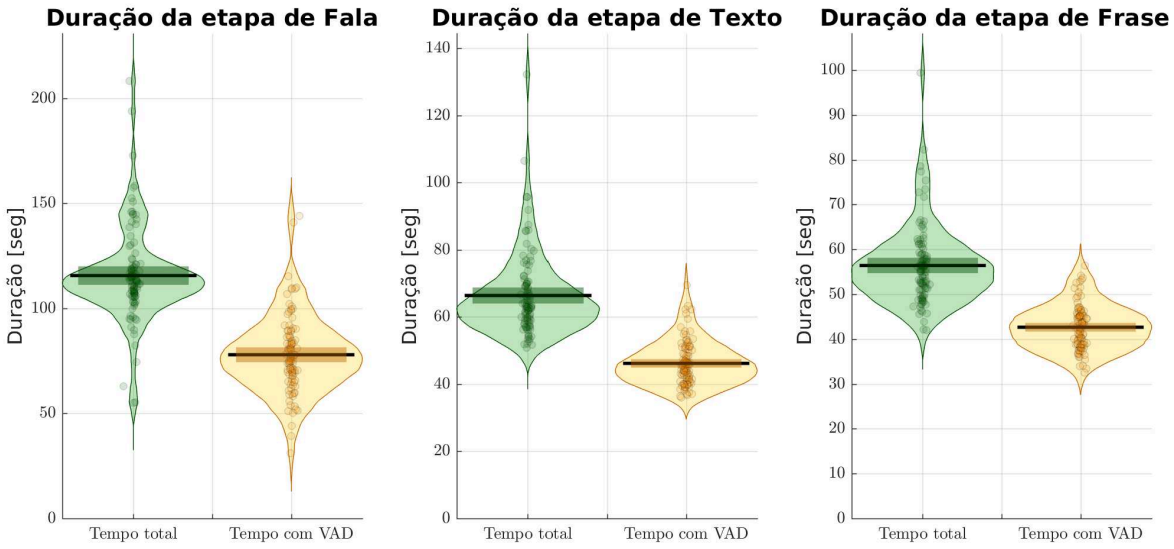


Figura 36 – Gráfico RDI apresentando a duração das etapas dos áudios do Corpus CEFALA-1 antes e depois do processamento VAD. Em cada coluna os pontos são as durações individuais, as curvas laterais são a distribuição de probabilidade empírica, a linha preta horizontal a média e o retângulo escuro é o intervalo de confiança da média para $\alpha = 0,05$.

das etapas da gravação do corpus CEFALA-1. Na Tabela 6 tem-se as estatísticas para os valores antes e depois do processamento por VAD (SOHN et al., 1999).

Tabela 6 – Estatísticas temporais do corpus CEFALA-1, em minutos, antes e após o processamento VAD.

		Mínimo	Máximo	Média	Mediana	Desvio padrão
Tempo do áudio (minutos)	Etapa de fala	0,9	3,5	1,9	1,9	0,4
	Leitura de texto	0,8	2,2	1,1	1,1	0,2
	Leitura de fases	0,7	1,7	0,9	0,9	0,1
	Áudio completo	3,4	6,9	4,5	4,5	0,6
Tempo do áudio após VAD (minutos)	Etapa de fala	0,5	2,4	1,3	1,3	0,3
	Leitura de texto	0,6	1,2	0,8	0,8	0,1
	Leitura de fases	0,5	0,9	0,7	0,7	0,1
	Áudio completo	1,9	4,2	2,9	2,9	0,4

APÊNDICE B

Trabalho de Conclusão de Curso - Engenharia Elétrica Newton Paiva

RECORTE DAS FRICATIVAS /S/ E /Z/ EM REGISTROS ACÚSTICOS POR INFERÊNCIA BAYESIANA

Adelino Pinheiro Silva[1]
Daniel Gonçalves Gomes[2]
Elizângela Mara Rodrigues de Oliveira[3]
Nathália Amorim Zolini[4]

RESUMO: Na sociedade contemporânea é cada vez maior a interação entre o homem e a máquina, oferecendo inúmeras possibilidades de exploração. A rápida informatização dos processos propicia à comunicação realizada através da fala torna-se uma alternativa viável para a melhoria desta interface e rápida adequação dos envolvidos na interação homem-máquina. Neste contexto, presente estudo tem como objetivo a composição de corpus coletado por procedimento padronizado para estudo de classificação das fricativas /s/ e /z/ em contexto de palavras. O texto introduz ainda ao leitor conceitos básicos na produção e análise de fala, tanto do ponto de vista descritivo, acústico e estatístico.

PALAVRAS-CHAVE: Análise de Voz e Fala. Identificação de Vogais. Redes neurais Artificiais. Análise Cepstral. Reconhecimento de Padrões.

1 - INTRODUÇÃO

Na sociedade contemporânea é cada vez maior a interação entre o homem e a máquina. Com o advento das mais avançadas de tecnologias, esta comunicação tem se tornado cada vez melhor, entretanto mais dependente e complexa. Diante deste contexto, a comunicação homem-máquina ainda oferece inúmeras possibilidades de exploração, e devido à informatização dos processos e à difícil adequação dos envolvidos, a comunicação realizada através da fala torna-se uma alternativa viável para a melhoria e expansão desta interface.

Editores de textos e softwares por comando de voz são uma realidade, embora necessitem aperfeiçoamento. O reconhecimento por comando de voz funciona, porém não é completamente robusto, devido a particularidades como gírias, chavões e regionalismos. Desta forma estes sistemas de reconhecimento de elementos de fala ainda possuem potencial de melhoria, tornando-se assim um vasto campo a ser explorado (MÜLLER, 2002), em especial como uma etapa do processo de identificação de falantes baseadas em características de alto nível (REYNOLDS et al., 2003).

Muitas empresas de tecnologia investem em técnicas de decodificação e quantificação de sinal da locução, sempre com o objetivo de preservar a informação de voz e fala e, em consequência, seu reconhecimento. A partir desta perspectiva o presente trabalho irá propor a identificação de padrões simbilantes em registros acústicos.

Inicialmente será realizado a composição de um corpus, coletado por procedimento padronizado, e a partir deste foram separadas as fricativas /s/ e e /z/ de contexto de palavras e posteriormente analisado padrões acústicos, utilizando uma combinação de de técnicas de reconhecimento de fala por padrões estatísticos e métodos quantitativos para codificação de sinais.

2 - FISIOLOGIA DO TRATO VOCAL

O discurso é o produto acústico final de movimentos voluntários, formalizados dos aparelhos respiratório e mastigatórios. O comportamento motor da produção da fala é adquirido, desenvolvido, controlado e mantido pela realimentação (***feedback***) acústica do mecanismo de audição e pela realimentação (***feedback***) sinestésica da musculatura da fala. A informação oriunda destes sentidos é organizada e coordenada pelo Sistema Nervoso Central e usados para conduzir a função da fala (FLANAGAN, 2013). Qualquer prejuízo ao mecanismo de controle, normalmente degrada o desempenho do aparelho vocal o que atrapalha o processo natural de comunicação, como apresentado pela figura 1.

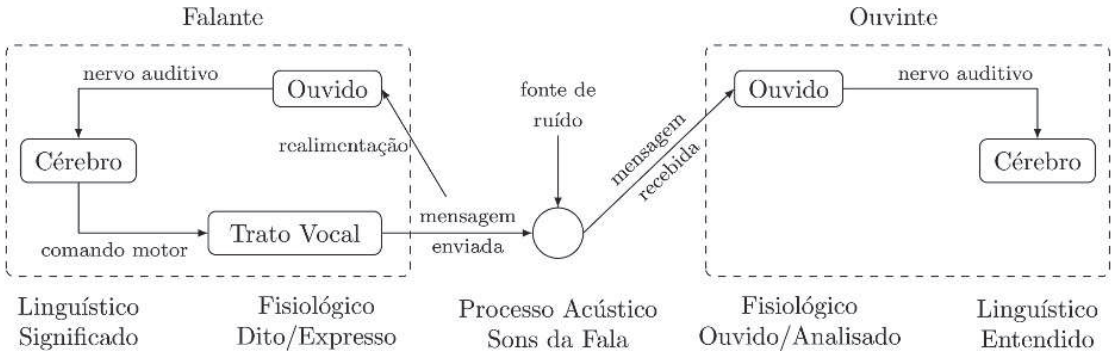


Figura 1 – Processo de comunicação.

Fonte: Elaborado pelos autores, adaptado de Flanagan (2013).

IDENTIFICAÇÃO DE PADRÕES DE VOGAIS EM REGISTROS ACÚSTICOS: análise por componentes cepstrais e redes neurais

Adelino Pinheiro Silva
Flávio Lúcio de Souza
Vinícius Rodrigo May

RESUMO: Os estudos referentes à biometria vocal são as principais fontes motivadoras desta pesquisa acadêmica. A identificação de vogais presentes em registros acústicos do português brasileiro é o primeiro passo na proposta de técnicas alternativas de identificação de locutores, permitindo consolidar informações referentes a condução da fala, em especial, para falantes perceptualmente próximos. A partir de amostras de corpus, trechos de áudios foram isolados e analisados com intuito de encontrar características acústicas capazes de caracterizar as vogais do português brasileiro em posição tônica. A etapa final de classificação proposta foi por meio de redes neurais artificiais.

PALAVRAS-CHAVE: Análise de Voz e Fala. Identificação de Vogais. Redes neurais Artificiais. Análise Cepstral. Reconhecimento de Padrões.

1 - INTRODUÇÃO

Plataformas que oferecem suporte aos comandos de voz já são realidade; os atuais dispositivos eletrônicos são prova disso. Entretanto, o reconhecimento de padrões de voz e de fala em registros acústicos é uma tecnologia ainda em desenvolvimento, sendo empregada em diferentes áreas desde o entretenimento, passando por aplicações de segurança, avaliação de saúde e sistemas de telefonia. Maiores avanços mostram-se possíveis; interfaces interativas são essenciais para aproximarem, ainda mais, homens e máquinas.

O reconhecimento biométrico por comandos de voz e de fala faz parte dessa tecnologia em ascensão. Estudos realizados à cerca deste tema proporcionam integração com diversos ramos do conhecimento e, prova disto é a quantidade significativa de áreas multidisciplinares envolvidas no processo, como por exemplo a fonética, a microeletrônica e o processamento de sinais (CAMBELL, 2009; TOGNERI; PULLELLA, 2011)

As técnicas à serem apresentadas neste trabalho, visam fazer parte de uma tarefa maior que é a implementação de um sistema de identificação de locutor. Tais técnicas, baseadas em redes neurais artificiais (RNA), mais especificamente o Perceptron Multi Camadas (MLP - Multi Layer Perceptron), são os modelos de reconhecimento de padrões de base para este estudo.

O objetivo do presente estudo é realizar a classificação, reconhecimento e análise de padrões acústicos em vogais cardinais do português brasileiro, em posição tônica, buscando identificar padrões capazes de recortar as vogais dos trecho de áudio. Para esta tarefa realizou-se: levantamento de corpus de falantes do português, com aleatorização de gênero e idade, utilizando protocolo padronizado; análise e identificação dos trechos de áudio de interesse e implementar redes neurais com base nas características relevantes para separação dos grupos de vogais.

2 - PRINCÍPIOS DA PRODUÇÃO DA VOZ E FALA

O processo de comunicação inicia-se no cérebro falante por um processo linguístico de geração do significado através de palavras e frases, em seguida um comando fisiológico ativa o trato vocal para gerar os sinais acústicos, que consistem em flutuações da pressão de ar que são geradas pelas pregas vocais são moduladas pelo trato vocal e irradiadas pela boca. A mensagem gerada é transmitida através de um canal de comunicação, como o ar por exemplo, até o ouvinte. A mensagem é detectada pelo ouvido, a flutuação de pressão no ar presente na mensagem transmitida é interpretada pelo cérebro do ouvinte. A figura 1 a seguir ilustra de forma resumida o processo de comunicação por um canal com ruído interferente.

Nos estudos dos processos de comunicação é importante ainda definir o conceito de linguagem como a base de transmissão de significado através de sinais, sons, gestos, ou marcas entendidas dentro de um grupo ou comunidade (PIERANGELO; GIULIANI, 2007). Com o passar dos tempos a humanidade esforçou-se para comunicar-se através de grandes distâncias, utilizando diferentes linguagens como batidas de tambores, sinais de fogo ou os telégrafos, ótico de Chappe e elétrico de Morse (GLEICK, 2013).

A transmissão elétrica dos registros acústicos ocorreu a partir dos desenvolvimentos de Bell que realizou estudos dos mecanismos da fala e da audição para aprimorar o processo de comunicação por voz em longas distâncias. Inicialmente as telecomunicações eram realizadas preservando a forma de onda acústica, entretanto, o desenvolvimento de técnicas matemáticas de análise e processamento de sinais permitiu realizar a codificação da onda acústica alcançar longas distâncias com mais eficiência (GLEICK, 2013).

APÊNDICE C

Trabalhos Publicados em Eventos

XI Seminário Nacional de Fonética Forense, VIII Seminário Nacional de Perícias em Crimes de Informática e III Seminário Nacional de Análise Forense de Imagens
Florianópolis-SC, 16 a 18 de novembro de 2016.

Avaliação do Reconhecimento de Falantes em Classes de Sons Semelhantes

A. P. Silva[1,2]
[1] *Instituto de Criminalística – Polícia Civil de Minas Gerais – IC/MG*
[2] Centro de Estudos da Fala, Acústica, Linguagem e músicA - Universidade Federal de Minas Gerais – *CEFALA/UFMG*

1 – Introdução

O presente trabalho tem como objetivo principal realizar a comparação de dois métodos de reconhecimento de locutores, por meio de seus registros de áudio. O primeiro método (doravante denominado algoritmo-01), reconhece falantes independente de texto, é difundido na literatura (Reynolds 1995) e utiliza como características as componentes mel cepstrais (MFCC) e realiza inferência bayesiana modelando os locutores através de modelos de misturas de gaussianas (GMM). Naturalmente, para aplicações forenses alguns autores discutem a aplicação destes métodos (Campbell 2009; Hollien, Bahr, and Harnsberger 2014). O segundo método (doravante denominado algoritmo-02) propõe utilizar o mesmo método de comparação descrito por Reynolds (1995), entretanto, a comparação das amostras de voz dos falantes será realizada entre classes de sons acusticamente semelhantes. Esta metodologia tem como ponto de partida o trabalho de Campolina (2012).

Assim tem-se a hipótese experimental se, na média, a acurrácia do algoritmo-02 é maior a acurácia do algoritmo-01, para aplicações em identificação de locutores em ambiente controlado.

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_0: \mu_1 - \mu_2 < 0 \end{cases}$$

Onde μ_1 é a acurácia média do algoritmo-01, μ_2 a acurácia média do algoritmo-02. O teste foi realizado com nível de significância $\alpha = 0,05$; mínimo efeito prático para acurácia: $\delta^* = 0,05$ ou d de Colen de $d^* = 0,9$; potência do teste desejada $\pi = 0,80$; e erro tipo II $\beta = 0,20$. A comparação dos dois algoritmos foi realizada em um conjunto de N = 28 locutores, de forma a acurácia do algoritmo será medida como a capacidade de distinguir um determinado locutor de um conjunto de vários locutores. Desta forma, cada observação independente do problema será a comparação dos dois algoritmos em identificar corretamente um determinado locutor do conjunto.

2 - Material e Métodos

O material sonoro dos falantes utilizados no presente estudo foram coletados em ambiente controlado (com pouca reverberação), utilizando microfone da marca SHURE, modelo SM58, com cápsula de captação em cardioide, digitalizado por placa de áudio marca EDIROL, modelo UA-25, em um canal (*mono*) com frequência de amostragem de 44.100 kHz e 16 bits de profundidade. O protocolo utilizado é a leitura de 22 (vinte e duas) frases de controle, com objetivo de distribuir as classes de sons anteriormente apresentadas. A procedimento de coleta dos dados procura catalogar e aleatorizar variáveis como a idade do falante, sexo, o horário da coleta, condição de saúde, conhecimento prévio de idiomas, entre outros fatores capazes de influenciar a relação competência/desempenho fonológico.

Para realização das análises o material sonoro dos locutores foi coicionado com base nas análises apresentadas no texto de Furui (2000), sendo elas: subamostragem para uma frequência de 8kHz, normalização da amplitude para manter o valor RMS em torno de -1 dB, cálculo das características $\boldsymbol{x_i}$ (MFCC's) do áudio de cada locutor, com 39 filtros, janela de 40 ms e passo de 20 ms. Na sequencia são aplicados os algoritmos conforme fluxograma da figura 1.

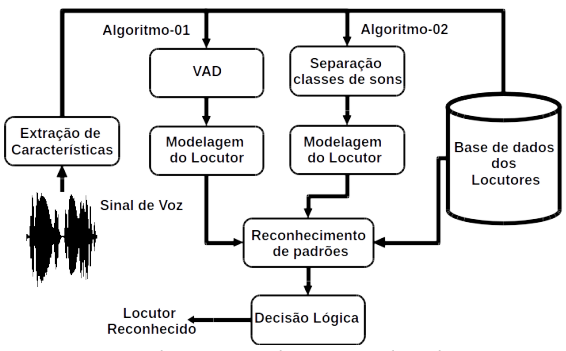


Figura 1: Fluxograma dos passos dos algoritmos comparados.

2.3 – Definições das Classes de sons Semelhantes

Para o presente trabalho o conteúdo de voz foi separado nas seguintes classes acústicas:

ANÁLISES DE CLASSES ACÚSTICAS SEMELHANTES DA FALA E APLICAÇÕES NA COMPARAÇÃO DE FALANTES

SILVA, Adelino P.[1,3]; VIEIRA, Maurílio N.[2]; CAMPOLINA; Thiago de A. M.[1].
(1) Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais;
(2) Departamento de Engenharia Eletrônica - Universidade Federal de Minas Gerais.
(3) Instituto de Criminalística – Polícia Civil de Minas Gerais.

RESUMO

O presente trabalho possui duas frentes, a primeira trata da classificação de sons da fala e como eles podem ser agrupados em classes semelhantes. A segunda frente utiliza os resultados do agrupamento da em classes semelhantes para realizar a comparação de falantes por pareamento de classes acústicas semelhantes. Na fonética e fonologia as principais ferramentas para a classificação dos sons da fala estão em parâmetros articulatórios, em função do significado linguístico ou em função da representação subjacente. A proposta do agrupamento em classes acústicas semelhantes na fala utiliza como premissa eventos da produção dos sons da fala e a tentativa de reduzir a heterocedasticidade das características utilizadas na comparação de falantes. Na área de comparação forense de locutor (CFL) os autores veem acompanhando a necessidade de aprimorar a robustez do referido exame frente ao crescimento de sua demanda e novos paradigmas. Dentre tais paradigmas pode-se incluir a análise semi-supervisionada baseada em parâmetros quantitativos com base em análise Bayesiana (taxa de verossimilhança).

Palavras-chave: Processamento de sinais, comparação de locutores, processamento de voz, classificação de padrões.

ABSTRACT

Present work discusses about two distincts subjects, the first about classification of speech sounds and how this sounds can be grouped in similar classes. Second front use results of analysis to classify speech sounds to improve speaker comparison paired by similar acoustics classes. In phonetics and phonology, the main tools for the classification of speech sounds are in articulatory parameters, depending on the linguistic meaning or the underlying representation. The propose of similar acoustics classes used as premise the events involved on production of speech sounds, and the attempt to reduce the heteroscedasticity of the characteristics used in the speaker comparison pairied by similar acoustics classes. In forensic speaker comparison (FSC), the authors see accompanying the need to improve the robustness of that examination against two reasons; the growth of its demand; and the new paradigms in FSC. This paradigms include the proposal of a semi-supervised analysis based on quantitative aproach over Bayesian analysis (likelihood ratio).

Keywords: Signal processing, Speaker comparation, speech processing, pattern classification.

1. INTRODUÇÃO

Atualmente, a comparação forense de locutor pode ser dividida em três abordagens distintas conforme enumeradas por Broeders (2001) e Hansen e Hasan (2015): a primeira denominada perceptiva, é realizada por foneticista treinado para realizar análise fonética e extrair uma variedade de parâmetros fonéticos, linguísticos e acústicos incluindo classificação da

XXII Congresso Brasileiro de Automática - CBA, 8-12 de setembro 2018, João Pessoa-PA

AVALIAÇÃO DE DESCRITORES ACÚSTICOS UTILIZANDO ESTATÍSTICA MVKD APLICADA À COMPARAÇÃO FORENSE DE LOCUTOR

Adelino Pinheiro Silva*† Maurílio Nunes Vieira‡ Adriano Vilela Barbosa‡

*Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais
Avenida Antônio Carlos, 6627 - CEP: 31270-901, Belo Horizonte, MG, Brasil

†Centro Universitário Newton Paiva
Rua José Cláudio Rezende, 420 - CEP: 30494-225, Belo Horizonte, MG, Brasil

‡Departamento de Engenharia Eletrônica - Universidade Federal de Minas Gerais
Avenida Antônio Carlos, 6627 - CEP: 31270-901, Belo Horizonte, MG, Brasil

Email: adelinocpp@yahoo.commaurilionunesv@gmail.comadriano.vilela@cefala.org

Abstract— Forensic speaker comparison (FSC) consists of comparing an unknown audio recording to a known one with the aim of determining whether both recordings come from the same individual. In most cases, the unknown recording comes from telephone interception, which means it is narrowband, GSM-encoded and corrupted by channel noise. Two surveys on international practices used in FSC published by the University of York in 2011 and by the INTERPOL in 2016 show that most of forensic experts carry out analyses based on perceptual and acoustic methodologies. On the other hand, automatic systems (assisted or not by an expert) have experienced little adoption. This work examines the discriminating power of descriptive statistics computed from acoustic features, such as Mel Frequency Cepstral Coefficients (MFCC), extracted from recordings in the CEFALA-1 Corpus. In an attempt to emulate forensic conditions, the recordings were narrowband filtered, GSM encoded, and contaminated with six levels of pink noise. Comparisons were performed by a log-likelihood ratio (LLR) framework using the Multivariate Kernel Density (MVKD). The best equal error rate (EER) obtained was 5.6% combining Power Normalized Component Cepstrum (PNCC) with skewness.

Keywords— Forensic Speaker Comparison, Cepstral analyze, Multivariate Kernel-Density, Equal Error Rate.

Resumo— Na prática, a comparação forense de locutor (CFL) consiste no confronto entre características de dois áudios, com o objetivo de associar as falas do áudio questionado a um indivíduo conhecido. Esse áudio, na maioria dos casos, é oriundo de interceptações telefônicas e possui codificação GSM, banda estreita e ruído de canal. Levantamentos do cenário mundial em CFL, realizados em 2011 e 2016, respectivamente pela Universidade de York e INTERPOL, indicaram que muitos peritos forenses baseavam-se em análises perceptuais e acústicas. Em contrapartida, a utilização de metodologias automáticas e assistidas são menos utilizadas. Nesse nicho, o presente trabalho busca explorar o potencial de características/descritores acústicos, como Componentes Mel Cepstrais e analisar o poder discriminante de grandezas estatísticas descritivas calculadas de características acústicas extraídas do Corpus CEFALA-1. Os experimentos utilizaram seis níveis de relação sinal ruído. Os cenários das comparações visam aproximar as condições forenses considerando a codificação GSM, a banda do sinal e o ruído de canal. O cálculo do logaritmo da razão de verossimilhança extraído por meio da densidade do núcleo de multivariáveis. A menor taxa de mesmo erro obtida foi de $5,6\%$ combinando PNCC com a assimetria.

Palavras-chave— Comparação Forense de Locutor, Análise cepstral, Densidade de núcleo de multivariáveis, Taxa de mesmo erro.

1 Introdução

Nas amostras confrontadas na prática da Comparação Forense de Locutor (CFL), tem-se os áudios questionados, vestígios de algum fato típico, e o áudio padrão. Em regra, áudio questionado é de autoria desconhecida e oriundo de interceptação telefônica. Esse áudio é comparado com o áudio padrão, que é fornecido espontaneamente por indivíduo suspeito. O áudio padrão é coletado em ambiente controlado por perito treinado utilizando procedimento operacional padronizado (Rose, 2003).

Os áudios questionado e padrão não possuem similaridade de contexto e, em muitos casos, o fornecedor do registro padrão não deseja ser vinculado ao áudio questionado. Em suma, a CFL busca evidências a favor da hipótese de os registros, questionado e padrão, serem ou não do mesmo indivíduo (Rose, 2003).

Os levantamentos realizados por (Gold and French, 2011) e (Morrison et al., 2016) indicam que a metodologia mais adotada para CFL combina análises perceptuais e acústicas. Por outro lado, a utilização de metodologias completamente automáticas e assistidas são menos utilizadas. Esses estudos também mostram que características como componentes cepstrais são menos exploradas em análises periciais.

A metanálise realizada por (Tirumala et al., 2017) indica que a maioria dos métodos de extração de características para verificação de locutores utiliza componentes cepstrais, em especial o MFCC (*Mel Frequency Cepstral Coefficient*) e variações. Por outro lado, trabalhos como de (Kinoshita et al., 2009; Morrison et al., 2011; Silva et al., 2016; Enzinger and Morrison, 2017), mais voltados para a área forense, apresentam estudos baseados em características pragmáticas, e.g., frequência fundamental e formantes.

Aplicação de Estatísticas de Formantes e MVKD à Comparação Forense de Locutor

Adelino Pinheiro Silva, Maurílio Nunes Vieira e Adriano Vilela Barbosa

***Resumo*— A comparação forense de locutor (CFL) consiste no confronto entre características de dois áudios, visando associar um indivíduo conhecido as falas do áudio questionado, que em geral, é oriundo de interceptações telefônicas e possui codificação GSM, banda estreita e ruído de canal. Nesse cenário, o presente trabalho busca explorar o potencial dos formantes de trechos vozeados na CFL. Os experimentos utilizaram o corpus CEFALA1 com seis níveis de ruído rosa emulando as condições da CFL. A variável de avaliação foi a razão de verossimilhança obtida por meio da densidade do núcleo de multivariáveis (MVKD). A menor taxa de mesmo erro média foi de 12,3% combinando frequência e banda dos formantes.**

***Palavras-Chave*— Comparação Forense de Locutor, Análise de Formantes, Densidade de núcleo de multivariáveis, Taxa de mesmo erro.**

***Abstract*— Forensic speaker comparison (FSC) consists of comparing an unknown audio recording to a known one with the aim of determining whether both recordings come from the same individual. The unknown recording comes from phone lawful interception, which means it is narrowband, GSM-encoded and corrupted by channel noise. This work examines the discriminating power of descriptive statistics computed from formants in short segments of speech. In an attempt to emulate forensic conditions, the recordings were contaminated with six levels of pink noise. Comparisons were performed by a likelihood ratio (LR) framework using the Multivariate Kernel Density (MVKD). The best mean equal error rate (EER) obtained was 12.3% combining formant frequency and band.**

***Keywords*— Forensic Speaker Comparison, Formant analyze, Multivariate Kernel-Density, Equal Error Rate.**

I. Introdução

Nas amostras confrontadas na prática da Comparação Forense de Locutor (CFL), tem-se os áudios questionados, vestígios de algum fato típico, e o áudio padrão. Em regra, áudio questionado é de autoria desconhecida e oriundo de interceptação telefônica. Esse áudio é comparado com o áudio padrão, que é fornecido espontaneamente por indivíduo suspeito. O áudio padrão é coletado em ambiente controlado por perito treinado utilizando procedimento operacional padronizado.

Os áudios questionado e padrão não possuem similaridade de contexto e, em muitos casos, o fornecedor do registro padrão não deseja ser vinculado ao áudio questionado. Em suma, a CFL busca evidências para confirmar ou refutar a hipótese de que os áudios questionado e padrão são provenientes do mesmo indivíduo.

A metodologia mais adotada para CFL combina análises perceptuais e acústicas [1] e trabalhos mais voltados para a área forense [2–4], apresentam estudos baseados em características pragmáticas, e.g., frequência fundamental e formantes.

Nesse nicho, o presente trabalho busca explorar o potencial de características pragmáticas, em especial os formantes de segmentos vozeados, representadas por estatísticas descritivas em condições próximas às encontradas na prática forense, i.e., em áudios com codificação GSM, banda estreita e ruído de canal. A inferência é baseada no logaritmo da razão de verossimilhança (LLR - *log-likelihood ratio*) calculada por *Multivariate Kernel-Density* (MVKD).

O MVKD foi proposto por [5] e adaptado para a comparação de locutores por [6]. Em suma, essa metodologia mostra-se eficaz se poucas observações são disponíveis por amostra e quando essas observações são correlacionadas. Se comparada a metodologia GMM-UBM (*Gaussian Mixture Model-Universal Background Model*), a MVKD possui uma acurácia inferior [6]. Entretanto, por não necessitar de etapas de treinamento, o MVKD é difundido em aplicações de CFL, como em [7].

Dentro desse contexto, o presente trabalho tem por objetivo avaliar o potencial da medidas de formantes de segmentos vozeados simuladas em canal GSM, com ruído rosa pelas relação sinal-ruído (SNR - *signal-to-noise ratio*) de 25, 23, 20, 17, 15 e 12 dB. Basicamente o experimento compara as características da amostra de voz após um procedimento de redução das observações por grandezas estatísticas descritivas. O resultado da redução é utilizado para o cálculo do LLR via MVKD. As grandezas estatísticas utilizadas para redução de observações foram a média, mediana, desvio padrão, valor de base, curtose, assimetria, moda e densidade modal. Estas grandezas foram computadas nos moldes do experimento de [4], que utilizou a frequência fundamental (F_0) para realizar a comparação dos locutores.

A Seção II apresenta o método para extração de características, a base de dados utilizada e descreve detalhadamente as condições em que o experimento supradescrito foi realizado. A Seção III discute os resultados obtidos. AS conclusões e propostas de continuidade são apresentadas na Seção IV.

II. Cenário de Simulação das Condições Forenses

A. Cálculo dos Formantes

A extração dos formantes foi realizada a partir da análise LPC composto pelo rastreamento dos picos do módulo de

Adelino Pinheiro Silva, Intituto de Criminalística da Polícia Civil de Minas Gerais, Programa de Pós Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais e Centro Universitário Newton Paiva. Maurílio Nunes Vieira e Adriano Vilela Barbosa, Departamento de Eletrônica da Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil. E-mails: adelinocpp@yahoo.com, mauriolionunesv@gmail.com, adriano.vilela@cefala.org.

XXVIII Encontro da Sociedade Brasileira de Acústica
3 a 5 de outubro de 2018
Porto Alegre - RS

COMPARAÇÃO FORENSE DE LOCUTOR PAREADA POR MEDIDAS INSTANTÂNEAS DE RELAÇÃO SINAL RUÍDO

SILVA, Adelino P.[1,2,3]; VIEIRA, Maurílio N.[4]; BARBOSA, Adriano V.[4];

(1) Instituto de Criminalística da Polícia Civil de Minas Gerais, Av. Augusto de Lima 1833, 31190-131., Belo Horizonte, MG, adelinocpp@yahoo.com.

(2) Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG.

(3) Centro Universitário Newton Paiva, R. José Cláudio Rezende, 420, 30494-225, Belo Horizonte, MG.

(4) Departamento de Engenharia Eletrônica, Av. Antônio Carlos, 6627, 31270-901, Belo Horizonte, MG.

RESUMO

A comparação forense de locutor (CFL) consiste no confronto de características de dois áudios com o objetivo de associar as falas do áudio questionado a um indivíduo conhecido. O áudio questionado é oriundo de interceptações telefônicas, i.e., possui codificação GSM (*Global System for Mobile*), banda estreita e ruído de canal. O presente trabalho propõe uma metodologia para CFL pareada pela relação sinal ruído instantânea visando melhorar o desempenho para áudios questionados, contaminados por ruído, utilizando como característica o MFCC (*Mel Frequency Cepstral Coefficient*), extraídos de gravações no Corpus CEFALA-1. A premissa é realizar um pareamento dos trechos com base na relação sinal ruído e realizar a comparação utilizando a metodologia GMM-UBM (*Gaussian Mixture Model-Universal Background Model*). Em uma tentativa de simular condições forenses, as gravações foram filtradas em banda estreita, codificadas em GSM e contaminadas com seis níveis de ruído rosa. Os resultados apresentam uma redução incremental na taxa de mesmo erro (EER - *Equal Error Rate*), na média, de $2,2\%$ para $1,7\%$.

Palavras-chave: Comparação forense de locutores, relação sinal ruído, taxa de mesmo erro.

ABSTRACT

Forensic speaker comparison (FSC) consists of comparing a questioned and known audio recordings with the aim of determining whether both recordings come from the same individual. In most cases, the questioned áudio comes from telephone interception, which means it is narrowband, GSM-encoded and corrupted by channel noise. This work proposes a methodoly of data paring by SNR (signal-to-noise ratio) to improve results for questioned áudio high contamination. This work uses MFCC (Mel Frequency Cepstral Coefficient), extracted from recordings of the CEFALA-1 Corpus. The premise is to pair data sets based on SNR, performing the comparison using the GMM-UBM methodology. In an attempt to emulate forensic conditions, the recordings were narrowband filtered, GSM encoded, and contaminated by six levels of pink noise. The results show an incremental decrease in equal error rate (EER), on average, from 2.2% to 1.7%.

Keywords: Forensic speaker comparison, signal-to-noise ratio, equal error rate.

Resultados da Busca por Robustez e Confiabilidade na Comparação Forense de Locutores

A.P. Silva[a,b,c], M.N. Viera[d], A.V. Barbosa[d]

[a]*Instituto de Criminalística de Minas Gerais - PCMG, Av. Augusto de Lima, 1833, Belo Horizonte - MG.*
[b]*Programa de pós Graduação em Engenharia Elétrica - UFMG, Av. Antônio Carlos, 6627, Belo Horizonte - MG.*
[c]*Centro Universitário Newton Paiva - Rua José Cláudio Rezende, 420, Belo Horizonte - MG.*
[d]*Departamento de Eletrônica - UFMG, Av. Antônio Carlos, 6627, Belo Horizonte - MG.*

Resumo

A comparação forense de locutores (CFL) é diferente da biometria de voz em muitos aspectos. Por exemplo, na biometria variáveis como o ambiente e o microfone ou o número de tentativas não são um problema, enquanto a CFL lida com gravações de diferentes origens. Na CFL, a regra é comparar uma amostra padrão, de um locutor conhecido, com uma (ou mais) amostras questionadas, que são vestígios de um fato típico. Além disso, existem os riscos de associar erroneamente a amostra de um inocente ao áudio questionado ou falhar em associar o áudio questionado a uma pessoa culpada. Neste trabalho são apresentadas duas propostas para melhoria de robustez e confiabilidade. A primeira é realizar a comparação pareada pela relação sinal ruído espectral (S^2NR - *Spectral signal-to-noise ratio*) e a segunda é a proposta de uma inferência intervalar, *intervalo de evidência*, baseado no Teste de Significância Genuinamente Bayesiano (FBST - *Full Bayesian Significance Test*). Na comparação pareada, os resultados apresentam uma redução incremental na taxa de mesmo erro (EER - *Equal Error Rate*), na média, de $2,2\%$ para $1,7\%$. O *intervalo de evidência* foi mais conservador, se comparado com outros métodos, com comprimento 49% maior que a abordagem de Gosset. O *intervalo de evidência* também apresentou uma redução em 71% na taxa de falso-positivo se comparado com aa inferência pontual.

Palavras-chaves: Comparação Forense de Locutor, Inferência intervalar, Razão de verossimilhança, Blocagem. Taxa de mesmo erro.

Abstract

Forensic speaker comparison (FSC) is different from voice biometry in many aspects. For example, the latter has more control of some variables while the former deals with recordings from different origins. In FSC, the rule is to compare a sample pattern with a crime vestige. Also, the stakes are higher in FSC, because of the risk of wrongly convicting an innocent person or acquiting a guilty person. On this papper two approaches have been used. The first is based on the separation of the speech signal in blocks with similar signal-to-noise ratios (SNR). The second approach applies the Full Bayesian Significance Test (FBST) to find the credibility interval of forensic speaker comparison. The main objective of block separation is to allow comparisons between more homogeneous data to be made. Blocking by SNR further decreased the EER from 2.2% to 1.7%, on average. In the FBST approach, the main challenge was to apply the test to a large number of observations and to formulate an equation to solve the test faster. Comparisons with other interval inference methodologies indicated that evidence interval (computed by FBST) is more conservative than others confidence intervals, with length 49% greater than Gosset approach. The evidence interval present 71% less false positive then punctual inference. On the other hand, the loss of accuracy by using evidence interval is less than 0.5%.

Key-words: Forensic Speaker Comparison, Intervalar inference, Likelihood ratio, Blocking, Equal error rate.

APÊNDICE D

Trabalhos Publicados em Revistas

Rev. Estud. Ling., Belo Horizonte, v. 27, n. 1, p. 191-212, 2019

Corpus CEFALA-1: Base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia

Corpus CEFALA-1: Audiovisual Database of Speakers for Biometric, Phonetic and Phonology Studies

Arlindo Follador Neto
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, Minas Gerais / Brasil
arlindo.neto@ict.ufvjm.edu.br

Adelino Pinheiro Silva
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
Centro Universitário Newton Paiva, Belo Horizonte, Minas Gerais / Brasil
Instituto de Criminalística de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
adelinocpp@yahoo.com

Hani Camille Yehia
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais / Brasil
hani@cpdee.ufmg.br

Resumo: A fala humana tem sido estudada em diferentes áreas do conhecimento, as quais incluem desde biometria até fonética e fonologia. Nas pesquisas realizadas em tais áreas, amostras da fala são recursos necessários para a obtenção de resultados e validação de hipóteses. Para isso, amostras de diferentes locutores e conteúdos são armazenadas em arquivos de áudio e organizadas em bases de dados. Tais bases de dados permitem a continuidade, praticidade e confiabilidade de pesquisas, eliminando a difícil e demorada etapa de coleta de dados. Além disso, permitem comparações consistentes entre estudos diferentes. Entretanto, bases de acesso livre na língua portuguesa ou gravadas em ambiente controlado são raramente encontradas. Dessa forma, o objetivo deste trabalho foi construir uma base de dados pública e gratuita do português brasileiro, nomeada Corpus CEFALA-1. A base de dados reúne 104 locutores orientados por um protocolo específico para coleta de amostras audiovisuais de fala gravadas em estúdio.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.27.1.191-212

v. 8, n. 2, p. 22-35, 2019
ISSN 2237-9223

REVISTA BRASILEIRA DE CRIMINALÍSTICA

DOI: http://dx.doi.org/10.15260/rbc.v8i2.344

Avaliação de Descritores Acústicos em Simulação de Condições Forenses de Verificação de Locutor

A.P. Silva[a,b,c,*], M.N. Vieira[d], A.V. Barbosa[d]

[a]*Instituto de Criminalística, Polícia Civil de Minas Gerais - Av Augusto de Lima 1833, Belo Horizonte, MG, Brasil.*
[b]*Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, Belo Horizonte, MG, Brasil.*
[c]*Faculdade de Ciências Exatas e Tecnológicas, Centro Universitário Newton Paiva - Rua José Cláudio Resende 420, Belo Horizonte, MG, Brasil.*
[d]*Departamento de Eletrônica, Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, Belo Horizonte, MG, Brasil.*

[*]*Endereço de e-mail para correspondência: adelinocpp@yahoo.com . Tel.: +55-31-988013605*

Recebido em 25/11/2018; Revisado em 11/06/2019; Aceito em 15/08/2019

Resumo

A comparação forense de locutor (CFL) consiste no confronto entre características de dois áudios, com o objetivo avaliar cientificamente se o resultado desse confronto fortalece ou enfraquece a hipótese de que as falas nesses áudios foram produzidas pelo mesmo indivíduo. O áudio, na maioria dos casos, é oriundo de interceptações telefônicas e possui codificação GSM (*Global System Mobile*), banda estreita e ruído de canal. Nesse nicho, o presente trabalho busca explorar o potencial de características/descritores acústicos, como Componentes Mel Cepstrais e analisar o poder discriminante destas características acústicas extraídas de corpus. Os experimentos utilizaram cinco tipos de ruído em seis níveis de relação sinal ruído. Os cenários das comparações visam aproximar as condições forenses considerando a codificação GSM, a banda do sinal e o ruído de canal. Um resultado observado é uma menor taxa de erro na utilização de componentes mel-cepstrais (5% em relação sinal ruído de 15 dB), sua equivalência com outros descritores, e o efeito da presença da codificação GSM. Na análise dos descritores, percebeu-se que alguns preservam mais informação e correlação após a codificação GSM porém este fato não reflete na redução de erros na comparação dos locutores.

Palavras-chaves: Comparação Forense de Locutor, Análise cepstral, Taxa de mesmo erro.

Abstract

Forensic speaker comparison (FSC) consists of comparing unknown an known speaker audio recordings with the aim of strengthening or weakening the hypothesis that both recordings come from the same individual. In most cases, the unknown recording comes from telephone interception, which means it is narrowband, GSM-encoded and corrupted by channel noise. This work examines the discriminating power of descriptive statistics computed from acoustic features, such as Mel Frequency Cepstral Coefficients (MFCC). In an attempt to emulate forensic conditions, the recordings were narrowband filtered, GSM encoded, and contaminated with six levels of noise. The scenarios of the comparisons aim to emulate the forensic conditions considering the GSM-encoded, the narrowband and the channel noise. An observed result is a lower error rate in the use of mel-frequency cepstrum (5 % in signal-to-noise ratio of 15 dB), its equivalence with other descriptors, and the effect of the presence of the GSM coding. In the analysis of the descriptors, it is noted that some preserve more information and correlation after the GSM-encoded, but this fact does not reflect in the reduction of errors in the speaker comparison.

Key-words: Forensic Speaker Comparison, Cepstral analyzis, Equal Error Rate.

1. INTRODUÇÃO

Na prática da Comparação Forense de Locutor (CFL) têm-se os áudios questionados, vestígios de algum fato típico, e o áudio padrão. Em regra, áudio questionado é de autoria desconhecida e oriundo de interceptação telefônica. Esse áudio é comparado com o áudio padrão, que é fornecido espontaneamente por indivíduo suspeito. O áudio padrão é

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2914942, 9 pages
https://doi.org/10.1155/2020/2914942

Research Article

Forensic Speaker Comparison Using Evidence Interval in Full Bayesian Significance Test

Adelino P. Silva [1,2,3] Maurílio N. Vieira,[4] and Adriano V. Barbosa [4]

[1]*Graduate Program in Electrical Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil*
[2]*Institute of Criminalistics of Minas Gerais, Av. Augusto de Lima 1833, 30110-017, Belo Horizonte, MG, Brazil*
[3]*Centro Universitário Newton Paiva, Rua José Cláudio Resende 420, 30494-230, Belo Horizonte, MG, Brazil*
[4]*Department of Electronic Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil*

Correspondence should be addressed to Adriano V. Barbosa; adriano.vilela@cefala.org

Received 10 July 2019; Accepted 7 January 2020; Published 24 March 2020

Academic Editor: Arturo J. Fernández

This paper describes the application of a full Bayesian significance test (FBST) to compute evidence intervals in forensic speaker comparison (FSC). In the FBST approach, the challenge is to apply the test to a large number of observations and to formulate an equation to solve the test quickly. The contribution of the present work is that it proposes an application of the FBST to FSC and develops a method to calculate the FBST for the distribution of expected values (mean) with unknown variance without using Monte Carlo Markov chains (MCMC). Comparisons with other interval inference methodologies indicate that the evidence interval size is 49% greater than that computed with the Gosset approach. The evidence interval presented 71% fewer classification errors than the punctual inference did for the signal-to-noise ratio (SNR) of 17 dB.

1. Introduction

The main task in forensic speaker comparison (FSC) is to analyze two or more voice records to infer whether they come from the same speaker. FSC differs from biometric voice recognition in the hypothesis test approach and in the nature of the voice samples. In the FSC scenario, a *questioned-voice* is compared to a *known-voice*, whereas in biometric recognition, the comparison is made among multiple speakers [1, 2].

The questioned-voice (or voice evidence) is an audio recording accepted as a vestige or evidence in a criminal investigation. The questioned-voice may be recorded in different situations, such as lawful phone interception (wiretapping), recordings of face-to-face conversation, or audio broadcasting.

In FSC, the hypothesis H_0 considers that both the questioned- and known-voices come from different speakers, whereas H_1 assumes that the questioned- and known-voices come from the same speaker.

However, the "individualization" that the hypotheses above propose has been considered a fallacy. This individualization assumes that the result of the confrontation between the questioned and standard voice is unique, without *a priori* probability and without repeating the test for the entire population [3, 4]. According to Saks and Koehler [3], the most reasonable hypotheses would be

$$H: \begin{cases} H_0: \text{the features of the questioned} - \text{voice are not compatible} \\ \text{with the features of the known} - \text{voice}, \\ H_1: \text{the features of the questioned} - \text{voice are compatible with} \\ \text{the features of the known} - \text{voice}. \end{cases} \quad (1)$$

APÊNDICE E

Trabalho Aguardando Revisão

JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 1

Improvements in Forensic Speaker Comparison by i-vector with Fuzzy Techniques using Image-Based Signal-to-Noise Ratio Measurements

A.P. Silva, M.N. Vieira and A.V. Barbosa

***Abstract*—Forensic speaker comparison (FSC) consists of comparing questioned and known audio recordings with the aim of determining whether both recordings come from the same individual. In most cases, the questioned audio comes from lawful phone interception (wiretapping), which means it is narrowband, GSM-encoded and corrupted by channel noise. This work proposes the application of Fuzzy Linear Discriminant Analysis (FLDA) and a methodology of data paring and fuzzification by image-based signal-to-noise ratio (SNR) measurements to improve results for questioned audio with high contamination. The premise is to pair data sets based on SNR, performing the comparison using the i-vector methodology. In an attempt to emulate forensic conditions, the recordings were narrowband filtered, GSM encoded, and contaminated by six levels of five type of noise. The main contribution of this work is: improve the performance of automatic steps in FSC by reducing false positive rates without losing accuracy. Taking i-vector as a reference, the results show an incremental decrease in equal error rate (EER), on average, from 4% to 3.6% on training step. In the testing stage, the improvement with fuzzy techniques reduced the false positive rate from 16.8% to 10.8%.**

***Index Terms*—Forensic speaker comparison, i-vectors, fuzzy linear discriminant analysis, signal-to-noise ratio, equal error rate.**

I. Introdução

O objetivo da Comparação Forense de Locutores (CFL) é obter evidências que reforcem ou rejeitem a hipótese de dois registros de áudio terem como autoria o mesmo locutor. O áudio questionado é um vestígio de fato típico penal, de autoria desconhecida e, usualmente, é oriundo de interceptação telefônica e contaminado por ruído. O áudio padrão é fornecido espontaneamente por indivíduo conhecido (suspeito) e coletado – por perito treinado –, em ambiente controlado utilizando procedimento operacional padronizado.

Os áudios padrão e questionado não possuem similaridade de contexto, contemporaneidade e, em muitos casos, o fornecedor do registro padrão não deseja ser vinculado ao áudio questionado. Na prática, a CFL é uma tarefa de verificação de locutor independente de texto e em conjunto aberto [1], [2]. Entretanto, a tarefa de identificação de locutor para mais de um áudio padrão pode ser solicitada dependendo da linha investigativa ou processual.

A.P. Silva, Programa de Pós Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais (PPGEE-UFMG), Instituto de Criminalística - Polícia Civil de Minas Gerais, Centro Universitário Newton Paiva, Belo Horizonte, MG, Brazil (adelinocpp@yahoo.com).

M.N. Vieira e A.V. Barbosa, Departamento de Engenharia Eletrônica - Universidade Federal de Minas Gerais (DELT-UFMG), Belo Horizonte, MG, Brazil.

A CFL resulta em relatório forense (laudo pericial) cuja problemática é preservar o princípio da inocência, mantendo a acurácia do exame. Em outras palavras, busca-se não associar erroneamente dois falantes distintos (falso positivo) frente ao princípio do ônus da prova e do estado democrático de direito.

Para contribuir na resolução deste problema o presente artigo busca melhorias em técnicas automáticas para compor etapas automáticas da CFL. Na verificação biométrica de locutores, as principais técnicas são baseadas em vetores de informação (*i-vector - Information Vector*) [3] ou em *x-vector*, vetores extraídos por redes neurais profundas (DNN - *Deep Neural Network*) [4]. Para a verificação de locutor, é amplamente divulgada [5] a metodologia *i-vector* utilizando componentes mel-cepstrais (MFCC - *Mel-Frequency Cepstral Coefficients*) com a análise discriminante linear probabilística (PLDA - *Probabilistic Linear Discriminant Analysis*).

O objetivo do presente trabalho é aplicar técnicas de lógica *fuzzy* à metodologia de verificação de locutor, em conjunto aberto e independente de texto, baseada em *i-vector* em cenários que emulam as condições forenses de contaminação por canal GSM (*Global System for Mobile*) [6] e ruído. As melhorias obtidas e o conhecimento das taxas de erro podem contribuir em etapas automáticas da CFL.

A inovação do presente trabalho consiste da aplicação de lógica *fuzzy* em três pontos que podem ser combinados:

- o primeiro é a substituição da etapa análise discriminante linear (LDA - *Linear Discriminant Analysis*) pela análise discriminante linear *fuzzy* (FLDA - *Fuzzy Linear Discriminant Analysis*) [7];
- o segundo – doravante denominado conjuntos *fuzzy*-S^2NR –, é o pareamento dos quadros de áudios em conjuntos *fuzzy*, estabelecidos por medidas espectrográficas de relação sinal ruído (S^2NR - *Spectrographic Signal-to-Noise Ratio* [8]), sendo que cada conjunto possui uma modelagem independente;
- o terceiro, é uma proposta de alteração no cálculo das estatísticas de Baum-Welch para considerar a pertinência dos conjuntos de *fuzzy*-S^2NR.

O pareamento por S^2NR na verificação de locutor (segundo ponto) foi explorado na metodologia GMM-UBM (*Gaussian Mixture Models-Universal Background Model*) em [9] e, basicamente, fragmenta as amostras em conjuntos considerando o valor da S^2NR de cada quadro de voz. O pareamento calcula um modelo de verificação de locutor para cada faixa (ou conjunto) de S^2NR e recombina os resultados utilizando uma rede neural artificial (RNA).