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.

TABELA I
DURAÇÃO (EM SEGUNDOS) DA AMOSTRA COMPLETA E DAS SUBCATEGORIAS, COM OS VALORES MÍNIMO, MÉDIO E MÁXIMO PARA O TEMPO TOTAL DE ÁUDIO E PARA O TEMPO DE FALA (I.E. EXCLUINDO PAUSAS).

	mínimo		médio		máximo	
	Total	Fala	Total	Fala	Total	Fala
Amostra	202	112	273	171	412	251
FESP	55	31	116	78	208	144
TEXT	51	36	66	46	132	70
FRAS	42	33	56	43	99	56

$H(z)$ com o refinamento das raízes de $A(z)$ utilizando a propriedade da integral de Cauchy [8].

O resultado da pesquisa de [1] mostra que os formantes em vogais são muito utilizados como parâmetros na comparação forense de locutores por vários profissionais.

Outra importância dos formantes reside na acústica de vogais e ditongos [9] e na análise de longo termo dos formantes. A medida de formantes, entretanto, é muito sensível à codificação de canal telefônico, principalmente os formantes F_1 e F_4[10].

B. Base de Dados Utilizada

O Corpus Cefala-1 foi gravado em quatro capturas de áudio e uma captura audiovisual. Três das capturas de áudio foram realizadas utilizando uma placa M-Audio FireWire modelo 1814, codificação PCM (*Pulse Code Modulation*) com frequência de amostragem $44,1\ kHz$ e 16 bits para caracterização da amplitude. A quarta captura de áudio foi realizada pelo microfone externo (viva-voz) de um aparelho celular marca Samsung modelo Galaxy S2 Lite GT-i9070 e a captura audiovisual foi realizada por uma câmera GoPro, modelo Hero 3 + Black Edition.

O protocolo utilizado possui três etapas distintas:

- Etapa de fala espontânea, contendo pelo menos 2 minutos de tempo de gravação. Nesta etapa solicitou-se ao sujeito que realizasse uma declaração contínua sem interrupção (referida como FESP na Tabela I);
- Etapa de leitura de texto, momento em que é apresentado ao sujeito um pequeno texto (referida como TEXT na Tabela I); e
- Etapa de leitura de frases, momento em que são apresentadas vinte frases de controle para pronúncia intervalada (referida como FRAS na Tabela I).

Para fins ilustrativos, a Tabela I apresenta estatísticas do tempo duração das amostras completas e das três subcategorias (etapas) de acordo com o protocolo de coleta (fala espontânea, leitura de texto e leitura de frases).

C. Metodologia MVKD

O trabalho de Morrison [6] propõe a avaliação da razão de verossimilhança (LR *Likelihood Ratio*) por meio de densidade do núcleo de multivariáveis (MVKD - *Multivariate Kernel Density*). O procedimento MVKD é definido por [5] e permite escrever a estatística $LR(\vec{x})$ como

$$LR(\vec{x}_Q) = \left[\frac{\sqrt{|C|}mh^p e^{-\frac{1}{2}(\bar{y}_1-\bar{y}_2)^T(D_1+D_2)^{-1}(\bar{y}_1-\bar{y}_2)}}{\sqrt{|D_1||D_2||D_1^{-1}+D_2^{-1}+h^2C^{-1}|}}\right] \times \left[\frac{\sum_{i=1}^{m} e^{-\frac{1}{2}(y^*-\bar{x}_i)^T((D_1^{-1}+D_2^{-1})^{-1}+h^2C)^{-1}(y^*-\bar{x}_i)}}{\prod_{l=1}^{2}\frac{\sum_{i=1}^{m} e^{-\frac{1}{2}(\bar{y}_l-\bar{x}_i)^T(D_l+h^2C)^{-1}(\bar{y}_l-\bar{x}_i)}}{\sqrt{|D_l^{-1}||D_l^{-1}+h^2C^{-1}|}}}\right], \quad (1)$$

onde m é o número de amostras que compõe o modelo de fundo (UBM), n_i o número de observações em cada amostra do modelo de fundo, n_l o número de observações nas amostras padrão e questionada, p a dimensionalidade de cada amostra (medição), x_{ij} são as medições que constituem as amostras de fundo, y_{ij} são as medições que constituem as amostras padrão e questionada, D_l são estimativas da matriz de covariância de grupo escalonada pelo número de observações da amostra padrão e questionada, U é a matriz de covariância de grupo empírica, C é a matriz de covariância empírica entre as amostras e h é o parâmetro de suavização do núcleo.

Os demais parâmetros são definidos [6] como:

$$y^* = (D_1^{-1}+D_2^{-1})^{-1}(D_1^{-1}\bar{y}_1 + D_2^{-1}\bar{y}_2), \quad (2a)$$

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad (2b)$$

$$x_{ij} = (x_{ij1},...,x_{ijp})^T,\ i \in \{1,...,m\},\ j \in \{1,...,n_i\}, \quad (2c)$$

$$\bar{y}_l = \frac{1}{n_l}\sum_{j=1}^{n_l} y_{lj}, \quad (2d)$$

$$y_{lj} = (y_{lj1},...,y_{ljp})^T,\ l \in \{1,2\},\ j \in \{1,...,n_l\}, \quad (2e)$$

$$D_l = \frac{U}{n_l}, \quad (2f)$$

$$U = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T}{\sum_{i=1}^{m}(n_i-1)}, \quad (2g)$$

$$C = \frac{\sum_{i=1}^{m}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^T}{(m-1)} - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T}{\sum_{i=1}^{m} n_i(n_i-1)}, \quad (2h)$$

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i, \quad (2i)$$

$$h = \left(\frac{4}{2p+1}\right)^{\frac{1}{p+4}} m^{-\frac{1}{p+4}}. \quad (2j)$$

D. Descrição das Comparações

Nesta etapa, foi planejado um experimento para avaliar o desempenho da comparação automática de locutores, utilizando a metodologia MVKD, e as amostras do Corpus Cefala-1 obtidas pelo aparelho celular.

Na preparação dos áudios, inicialmente foi realizada uma subamostragem para $8\ kHz$ seguido da aplicação de um filtro com largura de faixa entre 300 e 3500 Hz (simulando o canal telefônico). Cada amostra do corpus foi separada de acordo

com sua etapa de coleta: a fala espontânea, leitura de texto e leitura de frases isoladas.

A concatenação da etapa de leitura de texto com 66% da etapa de fala espontânea originou o *áudio padrão* de cada locutor, que foi codificado e decodificado pelo *codec* GSM 06.60[1].

Os 34% restantes da etapa de fala espontânea com a etapa de leitura de frases foram concatenados para criar o *Áudio Teste*. Os *Áudios Questionados* foram gerados a partir dos Áudio Teste. Tais amostras foram contaminadas 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 foram codificados e decodificados pelo *codec* GSM 06.60 para simular a influência do canal.

Os formantes foram extraídos dos trechos considerados vozeados pelo algoritmo proposto em [11] com presença de frequência fundamental computada pelo YAAPT (*Yet Another Algorithm for Pitch Tracking*[2]) [12].

Foram calculados quatro formantes e suas respectivas largura de banda $F[n] = \{F[0], F[1], ..., F[T-1]\}$ no espectro de 0 a 4kHz das amostras padrão e questionadas com pré-enfase de 0.95, quadros de 25 ms de duração e passo de tempo de 10 ms. As variações temporais $\Delta F[n]$ e $\Delta^2 F[n]$ foram calculadas, ao longo dos T quadros do áudio como

$$\Delta F[n] = \frac{\sum_{k=1}^{K} k(F[n+k] - F[n-k])}{2\sum_{k=1}^{K} k^2} \quad (3)$$

$$\Delta^2 F[n] = \frac{\sum_{k=1}^{K} k(\Delta F[n+k] - \Delta F[n-k])}{2\sum_{k=1}^{K} k^2}. \quad (4)$$

Sendo $K = 2$. Desta forma encontravam-se disponíveis as frequências e larguras de banda de 4 formantes, e suas respetivas variações temporais, totalizando 24 valores por quadro de voz.

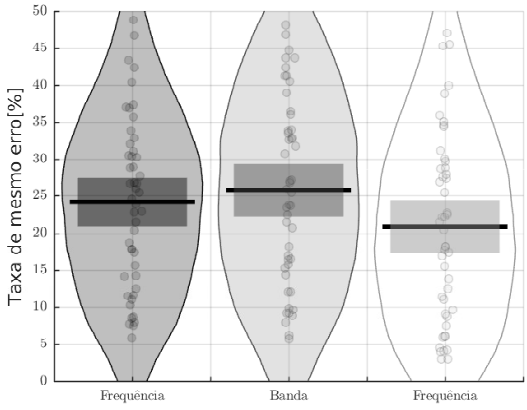

Fig. 1. Gráfico RDI apresentando a EER de acordo com a característica utilizada.

As observações foram reduzidas, por medidas amostrais de média (MED), mediana (MAN), desvio padrão (DPD), valor de base (VBS), curtose (CUR), assimetria (ASS), moda (MOD) e densidade modal (DSM) sobre intervalos de 5 segundos, i.e., 500 quadros.

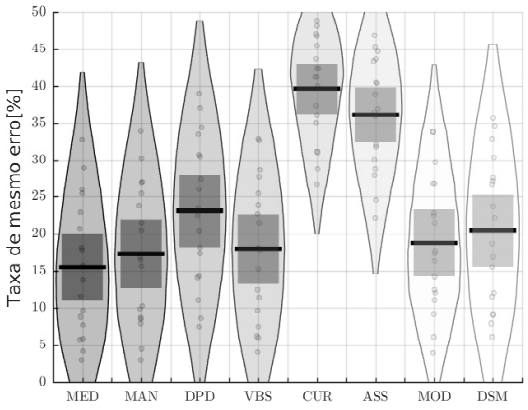

Fig. 2. Gráfico RDI apresentando a EER de acordo com a grandeza estatística de caracterização.

O valor da base de cada característica foi calculado conforme a proposta de [13], que indica o valor de base como o percentil equivalente a $7,64\%$ da distribuição empírica acumulada. Esta proposta é mais robusta e minimiza o impacto de valores extremos (*outliers*). A moda e densidade modal foram obtidas da densidade empírica de probabilidade.

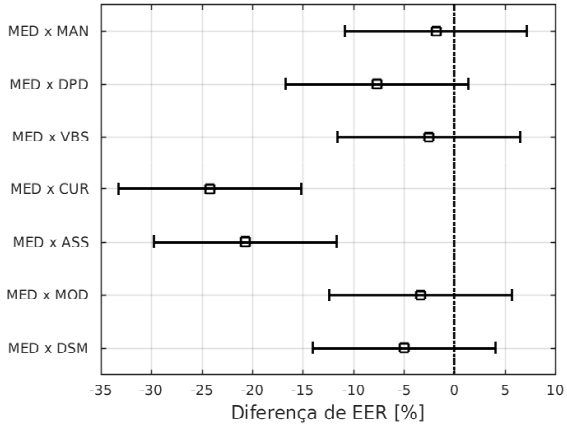

Fig. 3. Análise de variância, com significância $\alpha = 0,05$, indicando a diferença média entre o EER apresentado por cada grandeza estatística de redução.

Para avaliar o desempenho, utilizou-se a taxa de mesmo erro (*Equal Error Rate* - EER) por esta apresentar o equilíbrio entre os erros do tipo I e do tipo II para a proporção do teste. Na CFL, o erro do tipo I associa erroneamente um locutor a uma evidência de fato típico, enquanto, o erro do tipo II desassocia o autor de um áudio questionado. Em suma, o erro do tipo I pode condenar um inocente enquanto o erro do tipo II pode inocentar o culpado (ou falhar em condenar um culpado).

III. RESULTADOS

Realizadas as comparações par a par entre os 104 locutores do corpus Cefala-1 e foi computada a EER por descritor acústico (i.e. frequência e largura de banda dos formantes),

[1]Implementado no *software* SOund eXchange - sox.

[2]Mais um Algoritmo para Rastreamento de Frequência Fundamental - tradução livre.

grandeza estatística de redução (i.e. média, valor de base, etc...) e valor da SNR.

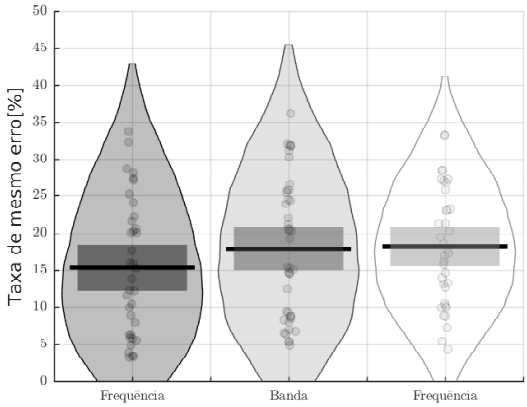

Fig. 4. Gráfico RDI apresentando a EER de acordo com a característica utilizada.

A Figura 1 apresenta o gráfico RDI (*Raw(data)- Description and Inference*) da EER de acordo com as características do formante – i.e. a frequência, largura de banda ou a combinação desses. No gráfico RDI, em cada coluna espalham-se (no eixo vertical) os valores individuais da EER obtida para cada SNR. As curvas laterais representam a distribuição empírica de probabilidade, a linha preta horizontal a média amostral e o retângulo escuro é o intervalo de confiança da média para significância $\alpha = 0,05$. No gráfico nota-se que a menor EER média ocorre quando combina-se as informações de frequência e largura de banda.

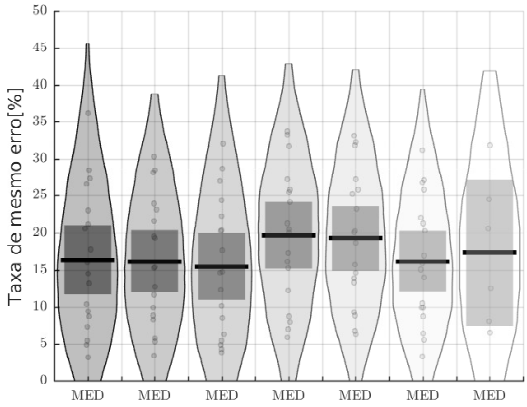

Fig. 5. Gráfico RDI apresentando a EER de acordo com a grandeza estatística de caracterização.

Observando as densidades empíricas desta mesma Figura nota-se que os valores de EER se espalham pelo domínio independentemente da característica do formante.

Do ponto de vista da grandeza estatística utilizada para reduzir as observações, nota-se que a média (MED) é a grandeza de melhor desempenho médio (Figura 2), seguida da mediana (MAN), valor de base (VBS) e moda (MOD). Estas quatro grandezas, juntamente com o desvio padrão (DPD) e a densidade modal (DSM) também são equivalentes do ponto de vista de desempenho médio, como apresenta a análise de variância (tomando como referência a média) para um nível de significância de $5\ \%$ na Figura 3.

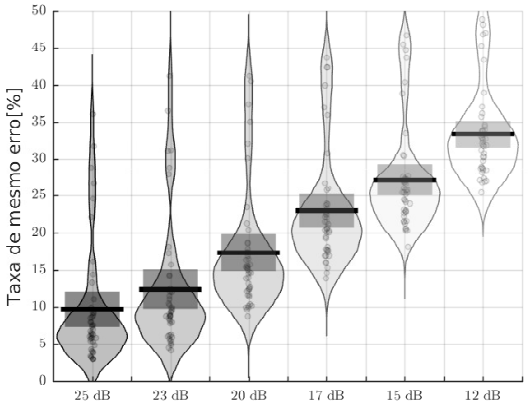

Fig. 6. Gráfico RDI apresentando a EER de acordo com a SNR.

A segunda etapa desta investigação combinou as grandezas estatísticas, construindo pares com a média, com intuito de verificar o desempenho combinado das grandezas estatísticas.

No recorte elaborado pelas características dos formantes, nota-se que a frequência do formante destaca-se com um menor EER médio. Inclusive, este desempenho médio reduziu de 20%, para a combinação da frequência com a largura de banda (vide Figura 1), para 15% com a frequência dos formantes como mostra a Figura 4.

Em relação ao par de grandezas estatísticas, o melhor desempenho médio foi apresentado pela combinação da média com o valor de base (MED + VBS). Entretanto, em relação ao desempenho isolado da média este valor não apresentou uma melhoria significativa.

A Figura 6 apresenta o gráfico RDI com a taxa de mesmo erro (EER) de acordo com a relação sinal ruído levando em consideração todos os resultados (i.e. grandezas acústicas isoladas e combinadas). Como esperado, o desempenho da comparação é inversamente proporcional a relação sinal ruído.

Apesar de ser uma informação relativa a experiência de um dos autores, uma SNR de 20 dB pode ser considerado otimista para áudio questionado. Desta forma, o valor médio de EER de $17,5\%$ nesta valor de SNR pode ser considerado elevado.

O menor valor de EER médio, em $12,3\%$, ocorreu para a média (MED) da combinação entre a frequência e a largura de banda dos formantes. A menor EER com valor de $2,9\%$ ocorreu para a combinação entre média e densidade modal (MED + DSM) da frequência dos formantes.

Um resumo dos valores obtidos de ERR, do ponto de vista dos três recortes, pode ser observado no diagrama da Figura 7. Neste gráfico cada círculo representa a combinação de um descritor (no eixo horizontal), com uma grandeza estatística (no eixo vertical). A abertura angular, a partir do eixo vertical no sentido anti-horário, indica a EER na escala entre 0 e $56\ \%$.

IV. Conclusões

Inicialmente é importante ressaltar que este foi um experimento empírico, exploratório e piloto.Do ponto de vista das características de redução, a equivalência percebida na Figura

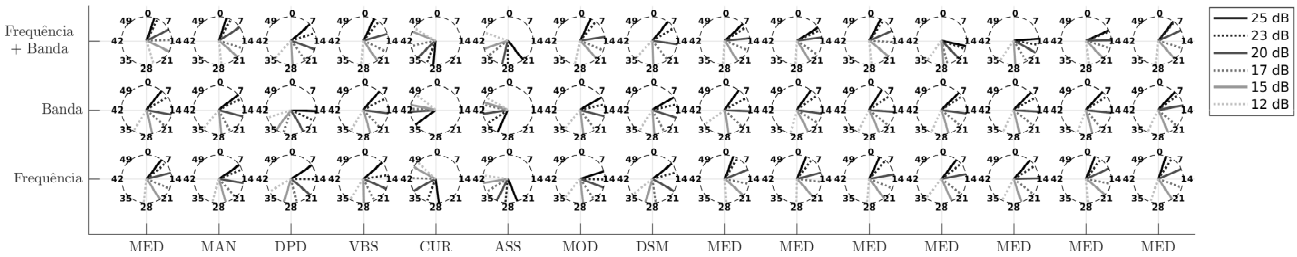

Fig. 7. Diagrama apresentando a EER com recorte por características, grandeza estatística e intensidade de ruído. A abertura angular de cada círculo, a partir do eixo vertical no sentido anti-horário, indica a EER na escala entre 0 e 56 %. A posição do círculo representa a frequência dos formantes, largura de banda ou a combinação (no eixo vertical), com uma grandeza estatística de redução (no eixo horizontal).

3 pode ter sido originada pela variação de desempenho que cada descritor apresenta em relação a característica do formante, como pode ser observado na Figura 7. No diagrama a EER varia (para o mesmo valor de SNR) tanto no eixo vertical quanto no horizontal da Figura 7. Este resultado necessita de uma exploração mais profunda antes de conclusões mais definitivas sobre as características dos formantes na CFL.

Em relação a grandeza estatística de redução, nota-se que a MED, MAN, VBS e MOD destacam-se por apresentar menor EER. Entretanto estas variáveis podem simplesmente estar fortemente correlacionadas. Sobre a relação sinal ruído, o resultado foi como esperado e comentado na seção anterior.

Em relação a metodologia MVKD, estes resultados permitem uma comparação com a UBM-GMM, que também fará parte das continuidades destas investigações.

Apesar do número de locutores na base de dados, este experimento foi relativamente pequeno. As avaliações restringiram-se a apenas um tipo de ruído, com tempo fixo para redução das grandezas estatísticas, e não explorou a correlação entre as características. Entretanto, para fins forenses, o resultado indica que as características dos formantes apresentam desempenho inferior a descritores não pragmáticas como os componentes mel cepstrais (MFCC).

AGRADECIMENTOS

Os autores gostariam de agradecer a G. Morrison, D. Ellis, C. Kim e N. Phillips por sempre compartilhar códigos. Agradecemos a equipe do SPAV do Instituto de Criminalística de Minas Gerais, em especial a Harley Cesar de Melo. Por fim o agradeço a todos colegas e professores do CEFALA.

O presente trabalho foi realizado com o apoio financeiro do Centro Universitário Newton Paiva.

REFERENCES

[1] E. Gold and P. French, "International practices in forensic speaker comparison.," *International Journal of Speech Language and the Law.*, vol. 18, no. 2, pp. 293–307, 2011.

[2] G. S. Morrison, C. Zhang, and P. Rose, "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system," *Forensic science international*, vol. 208, no. 1, pp. 59–65, 2011.

[3] E. Enzinger and G. S. Morrison, "Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case," *Forensic Science International*, vol. 277, pp. 30–40, 2017.

[4] R. R. d. Silva, "Aplicação do valor de base da frequência fundamental via estatística MVKD em comparação forense de locutor," 2017.

[5] C. G. Aitken and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 53, no. 1, pp. 109–122, 2004.

[6] G. S. Morrison, "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*, vol. 53, no. 2, pp. 242–256, 2011.

[7] V. Hughes, *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. PhD thesis, 2014.

[8] C. Kim, K.-d. Seo, and W. Sung, "A robust formant extraction algorithm combining spectral peak picking and root polishing," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 33–33, 2006.

[9] P. Rose, P. Warren, and C. Watson, "The intrinsic forensic discriminatory power of diphthongs," in *Proc. 11th Australian Int. Conf. Speech Sci. Technol*, pp. 64–69, 2006.

[10] P. Rose, *Forensic speaker identification*. CRC Press, 2003.

[11] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[12] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[13] J. Lindh and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.