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.

Nesse nicho, o presente trabalho busca explorar o potencial de características não-pragmáticas, como MFCC e suas variantes, 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 (Aitken and Lucy, 2004) e adaptado para a comparação de locutores por (Morrison, 2011). Em suma, essa metodologia mostra-se eficaz se poucas observações são disponíveis por amostra e quando essas observações são correlacionadas. Se comparada a metodologia UBM-GMM (*Universal Background Model-Gaussian Mixture Model*), a MVKD possui uma acurácia inferior (Morrison, 2011). Entretanto, por não necessitar de etapas treinamento, o MVKD é difundido em aplicações de CFL, como em (Hughes, 2014) e (Gold, 2014).

Dentro desse contexto, o presente trabalho tem por objetivo avaliar diferentes características cepstrais quando simuladas em canal GSM, com ruído rosa por SNR 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 (Silva et al., 2016), que utilizou a frequência fundamental (F_0) para realizar a comparação dos locutores.

2 Características/Descritores Acústicos Utilizados

O processo fundamental e comum das formas de comparação de locutores é a extração de características. Este processo consiste em transformar o sinal de áudio em um conjunto de vetores, igualmente espaçados no tempo, capazes de descrever uma característica presente no sinal de voz. Muitas vezes, em processamento de voz, a característica é denominada *descritor acústico*.

As características utilizadas neste experimento foram o MFCC (*Mel-Frequency Component Cepstrum*), MFEC (*Mel Frequency Entropy Cepstrum*), PLP (*Perceptual Linear Predictive*), PNCC (*Power Normalized Component Cepstrum*), RASTA-PLP (*Representations Relative Spectra*), SSCH (*Subband Spectral Centroid Histograms*), TEOCC (*Teager Energy Operator Component Cepstrum*) e ZCPA (*Zero-Crossing with Peak Amplitude*).

Não é o objetivo do presente texto detalhar a forma de extração de cada característica. Porém, tomando como referência o MFCC, mais difundida para a verificação automática de locutores (Tirumala et al., 2017) é possível obter uma visão ampla dos métodos de extração. Esses possuem etapas comuns e podem ser resumidos em pré-processamento, processamento específico e o pós-processamento, como indicado na Figura 1.

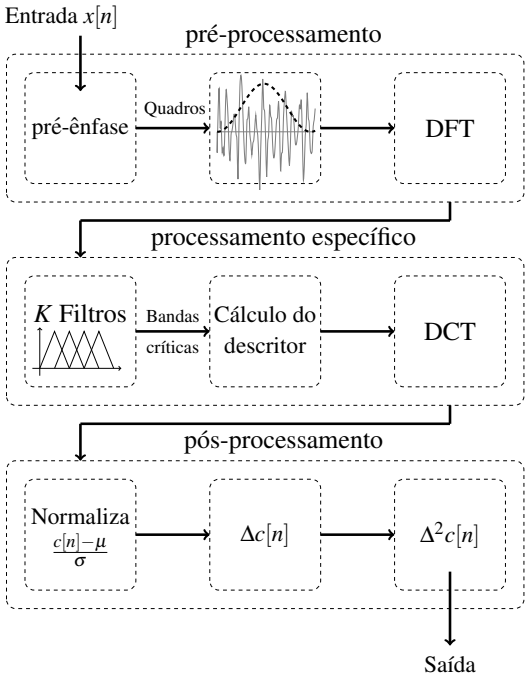

Figura 1: Diagrama de Blocos indicando as etapas, comuns e específicas, para o cálculo das características/descritores acústicos.

2.1 Pré-processamento

O pré-processamento das características acústicas é composto pelas etapas sequenciais de pré-ênfase, divisão em quadros e o janelamento. A pré-ênfase consiste na aplicação de um filtro passa alta, do tipo $y[n] = x[n] - \alpha x[n-1]$ para corrigir a declinação de 6 $db/8^a$ do sinal de voz. A pré-ênfase também reduz o intervalo de valores do espectro (Togneri and Pullella, 2011). No presente trabalho utilizou-se $\alpha = 0,97$.

A divisão em quadros consiste na extração de trechos sobrepostos do sinal de voz, com duração e deslocamento fixos. Neste trabalho foram utilizados os quadros de 25 ms deslocados a cada 10 ms, i.e. os quadros consecutivos possuem 25 ms e são gerados a cada 10 ms com superposição de 15 ms.

O janelamento é a multiplicação do quadro por uma função janela, que tem o efeito de suavizar os efeitos da duração do quadro. O resultado da função janela no domínio da frequência é convolutivo, criando um aspecto mais suave e com menos espalhamento espectral (Togneri and Pul-

lella, 2011). Neste trabalho foi utilizada a janela espectral de Hanning.

A etapa de pré processamento tem como saída uma sequência de quadros de áudio no domínio da frequência. Após o pré-processamento cada descritor utiliza uma técnica específica para extração de informação do sinal de voz.

2.2 Processamento Específico dos Descritores Acústicos

O processamento específico de cada descritor tem como entrada os quadros do sinal e como saída os C componentes cepstrais computados pela transformada discreta cosseno (DCT-*Discrete Cossine Transform*) em cada quadro (Figura 1). A primeira etapa do processamento específico é a filtragem em bandas críticas. Nessa etapa são extraídas as sub-bandas dos quadros de áudio para o cálculo do descritor. Essa etapa possui duas variantes básicas, o formato do filtro e a escala de distribuição. Nas definições do MFCC realizadas por (Davis and Mermelstein, 1990) e (Reynolds, 1994) o filtro utilizado possui formato triangular distribuído na escala mel. No MFCC são extraídos os logaritmos da energia de cada banda para em seguida ser calculada a DCT.

Um variação do MFCC é o PNCC, descrito em (Kim et al., 1999) e (Kim and Stern, 2016). Utilizado também para verificação de locutores, o PNCC possui uma etapa de normalização, que visa corrigir efeitos da variação de energia ao longo do registro de áudio. Outra diferença é a utilização do filtro gammatone em escala ERB (*equivalent rectangular bandwidth*).

O PLP e o RASTA-PLP também são descritores baseados na equalização de energia. Propostos respectivamente em (Hermansky, 1990) e (Hermansky and Morgan, 1994) o PLP realiza a normalização com base na curva de intensidade perceptiva (*loudness*). Os filtros, separados na escala de Bark possuem uma forma particular, é assimétrico, com região central plana e diferentes decaimentos exponenciais para frequências altas e baixas. A contribuição do RASTA-PLP é a aplicação de uma filtragem relativa amplitude do espectrograma.

O MFEC e o TEOCC, propostos respectivamente por (Jam and Sadjedi, 2009) e (Jabloun et al., 1999) possuem um processamento específico semelhante. Enquanto o primeiro foi definido por filtros triangulares na escala mel, o segundo utiliza filtros de Dauberchies(6^a ordem) em tipografia de árvore, ambos realizam o processamento da banda crítica no domínio do tempo. No caso do MFEC é calculada a entropia e no TEOCC a energia Teager (Holambe and Deshpande, 2012) por banda. Aos valores calculados são aplicados a DCT.

Diferentemente dos descritores anteriores, o SSCH (Gajic and Paliwal, 2001) e o ZCPA (Kacur et al., 2012) são utilizados para o reconhecimento de fala. O princípio do SSCH é separar o espectro em bandas na escala de Bark por filtros retangulares e calcular o centroide de energia de cada banda. A esses centroides são aplicados a DCT. O ZCPA utiliza filtros do tipo gammatone na escala ERB para – no domínio do tempo –, calcular um índice que pondera a taxa de cruzamento por zeros com a amplitude.

Por questões práticas, no experimento do presente trabalho alguns descritores foram discretamente adaptados em relação a forma do filtro e à escala, como apresentado na Tabela 1

Tabela 1: Parâmetros adaptados no experimento para obtenção dos descritores acústicos.

Descritor acústico	Formato Filtro	Escala
MFCC	Triangular	Mel
MFEC	Triangular	Mel
PLP	Particular	Bark
PNCC	Gamatone	ERB
RASTA-PLP	Particular	Bark
SSCH	Quadrado	Bark
TEOCC	Gammatone†	Mel†
ZCPA	Triangular†	Bark

† Alterações realizadas nos experimentos deste trabalho.

2.3 Pós-Processamento

A etapa de pós processamento envolve a normalização das características e o cálculo das variações temporais de primeira e segunda ordem. A normalização é realizada utilizando a média e variância de todos os locutores (Togneri and Pullella, 2011), enquanto as variações temporais, de primeira Δc (delta cepstrum) e segunda ordem $\Delta^2 c$, ao longo dos T quadros do áudio são obtidas como

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

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

A maioria dos autores utiliza $P = 1$ ou $P = 2$. Após a extração das características, o vetor de medidas do áudio $\vec{x} = \{x_1, x_2, ..., x_T\}$ é composto por T quadros com um número de características, ou dimensionalidade, definido pela quantidade de componentes cepstrais e a presença ou não dos Δc e $\Delta^2 c$. Assim $\vec{x}$ terá T observações de C dimensões na ausência de Δc e $\Delta^2 c$.

3 Cenário de Simulação das Condições Forenses

3.1 Base de Dados Utilizada

O Corpus Cefala-1 é uma base de vozes com objetivo de privilegiar a diversidade de falantes. O material consiste no registro de 104 locutores – 55 do sexo masculino e 49 do sexo feminino –, com coletas realizadas com ruído de fundo de 34 dBA.

O material foi gravado em quatro capturas de áudio e uma captura audiovisual, sendo três 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 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 fala bruta. Nesta etapa solicitou-se ao fornecedor que realizasse uma declaração contínua sem interrupção (referida como FESP na Tabela 2);
- Etapa de leitura de texto, momento em que é apresentado ao fornecedor um pequeno texto (referida como TEXT na Tabela 2); e
- Etapa de leitura de frases, momento em que são apresentadas vinte frases de controle para pronúncia intervalada (referida como FRAS na Tabela 2).

Para fins ilustrativos, a Tabela 2 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).

3.2 Metodologia MVKD

O trabalho de (Morrison, 2011) propõe a avaliação da LR por meio de densidade do núcleo de multivariáveis (MVKD - *Multivariate Kernel Density*). O procedimento MVKD é definido por (Aitken and Lucy, 2004) e permite escrever a estatística $LR(\vec{x})$ como

$$LR(\vec{x}_Q) = \left[\frac{\sqrt{|C|}mh^p e^{-\frac{1}{2}(\bar{y}_1-\bar{y}_2)^T(D_1+D_2)^{-1}(\bar{y}_1-\bar{y}_2)}}{\sqrt{|D_1||D_2||D_1^{-1}+D_2^{-1}+h^2C^{-1}|}}\right]\times \left[\frac{\sum_{i=1}^{m} e^{-\frac{1}{2}(y^*-\bar{x}_i)^T((D_1^{-1}+D_2^{-1})^{-1}+h^2C)^{-1}(y^*-\bar{x}_i)}}{\prod_{l=1}^{2}\frac{\sum_{i=1}^{m} e^{-\frac{1}{2}(\bar{y}_l-\bar{x}_i)^T(D_l+h^2C)^{-1}(\bar{y}_l-\bar{x}_i)}}{\sqrt{|D_l^{-1}||D_l^{-1}+h^2C^{-1}|}}}\right], \quad (3)$$

Tabela 2: 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

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 como

$$y^* = (D_1^{-1}+D_2^{-1})^{-1}(D_1^{-1}\bar{y}_1 + D_2^{-1}\bar{y}_2), \quad (4a)$$

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad (4b)$$

$$x_{ij} = (x_{ij1},...,x_{ijp})^T,\ i \in \{1,...,m\},\ j \in \{1,...,n_i\}, \quad (4c)$$

$$\bar{y}_l = \frac{1}{n_l}\sum_{j=1}^{n_l} y_{lj}, \quad (4d)$$

$$y_{lj} = (y_{lj1},...,y_{ljp})^T,\ l \in \{1,2\},\ j \in \{1,...,n_l\}, \quad (4e)$$

$$D_l = \frac{U}{n_l}, \quad (4f)$$

$$U = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T}{\sum_{i=1}^{m}(n_i-1)}, \quad (4g)$$

$$C = \frac{\sum_{i=1}^{m}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^T}{(m-1)} - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T}{\sum_{i=1}^{m} n_i(n_i-1)}, \quad (4h)$$

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i, \quad (4i)$$

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

3.3 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.

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.

As características extraídas dos trechos considerados vozeados pelo algoritmo de (Sohn et al., 1999). Essas características foram calculadas das amostras padrão e questionadas pelos descritores apresentados na seção 2, utilizando 13 bandas críticas, com quadros de 25 ms de duração e passo de tempo de 10 ms, gerando as observações $c[n]$, $\Delta c[n]$ e $\Delta^2 c[n]$.

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.

O valor da base de cada característica foi calculado conforme a proposta de (Lindh and Eriksson, 2007), 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.

Desta forma, cada amostra passa a possuir (para cada 5 segundos) suas grandezas estatísticas em 39 dimensões ($c[n]$, $\Delta c[n]$ e $\Delta^2 c[n]$).

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).

4 Resultados

Na comparação par a par entre os 104 locutores do corpus Cefala-1 foi computada a EER por descritor acústico (i.e. MFCC, PLP, etc...), grandeza estatística de redução (i.e. média, valor de base, etc...) e valor da SNR. A Figura 2 apresenta o gráfico RDI (*Raw(data)- Description and Inference*) com a taxa de mesmo erro (EER) de acordo com a relação sinal ruído. No gráfico RDI, em cada coluna espalham-se os valores individuais da EER obtida para cada SNR. As curvas (ou silhuetas) 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 $\alpha = 0,05$. Como esperado, o desempenho da comparação é inversamente proporcional a relação sinal ruído.

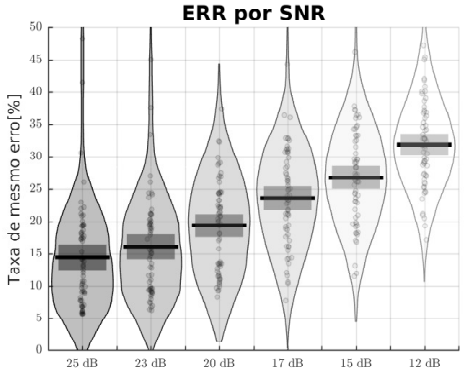

Figura 2: Gráfico RDI apresentando a EER de acordo com a SNR.

No recorte elaborado pelo descritor acústico, a Figura 3 ilustra que a menor EER média ocorre com o TEOCC seguido por PNCC, PLP e MFEC. Entretanto, na análise de variância para um nível de significância de 5 % (tomando como referência o TEOCC), todas os descritores acústicos apresentaram desempenho equivalente (com exceção ao MFCC) como pode ser constatado na Figura 4. Observando as densidades empíricas da Figura 3 nota-se que os valores de ERR se espalham pelo domínio independentemente da característica/descritor acústico.

Do ponto de vista da grandeza estatística utilizada para reduzir as observações, nota-se que a assimetria (ASS) é a grandeza de melhor desempenho médio (Figura 5), seguida da densidade modal (DSM), valor de base (VBS) e desvio padrão (DPD). Estas quatro grandezas também são equivalentes do ponto de vista de desempenho médio, como apresenta a análise de variância (tomando como referência a assimetria) para um nível de significância de 5 % na Figura 6.

A menor EER com valor de $5,6\%$ ocorreu para assimetria do PNCC com SNR de 25 dB. Um resumo dos valores obtidos de ERR, do ponto de

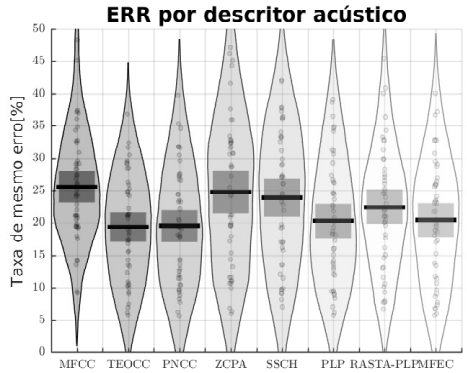

Figura 3: Gráfico RDI apresentando a EER de acordo com a característica utilizada.

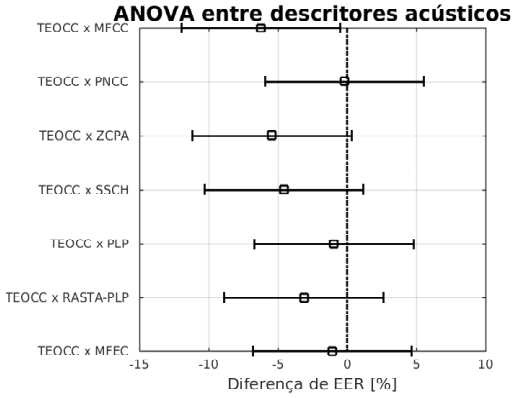

Figura 4: Análise de variância, com significância $\alpha = 0,05$, indicando a diferença média entre o EER apresentado por cada característica.

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 %.

5 Discussões e conclusão

Inicialmente é importante ressaltar que este foi um experimento empírico, exploratório e piloto. O experimento utilizou um recorte específico de características, tipo de ruído e SNR. Isso implica que os resultados permitem extrair uma quantidade limitada de informação.

Do ponto de vista dos descritores, a equivalência percebida na Figura 4 pode ter sido originada pela variação de desempenho que cada descritor apresenta em relação a grandeza estatística de redução, 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 aceitar o TEOCC como um descritor promissor para CFL.

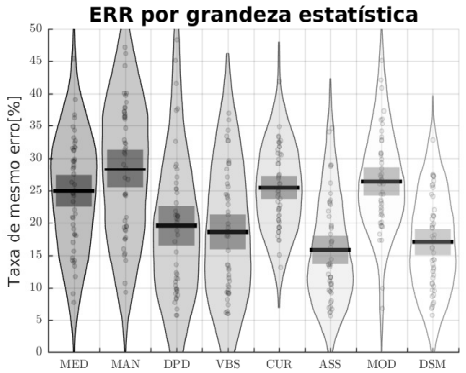

Figura 5: Gráfico RDI apresentando a EER de acordo com a grandeza estatística de caracterização.

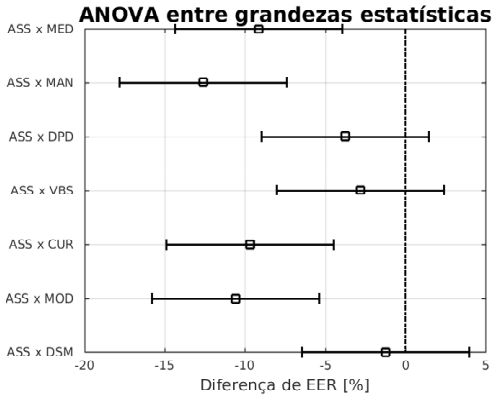

Figura 6: Análise de variância, com significância $\alpha = 0,05$, indicando a diferença média entre o EER apresentado por cada grandeza estatística de redução.

Em relação a grandeza estatística de redução, nota-se que a ASS, DPD, VBS e DSM destacam-se. Essas grandezas de redução destacam-se por apresentar menor EER. Uma proposta para continuidade dos trabalhos é a combinação de grandezas estatísticas e a variação do tamanho do intervalo de redução, pois nestes experimentos o valor foi fixado em cinco segundos.

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 combinações entre as variáveis. Entretanto, para fins forenses, o resultado indica valor máximo de EER em 11% para combinação do TEOCC com a assimetria em SNR maior que 17 dB.

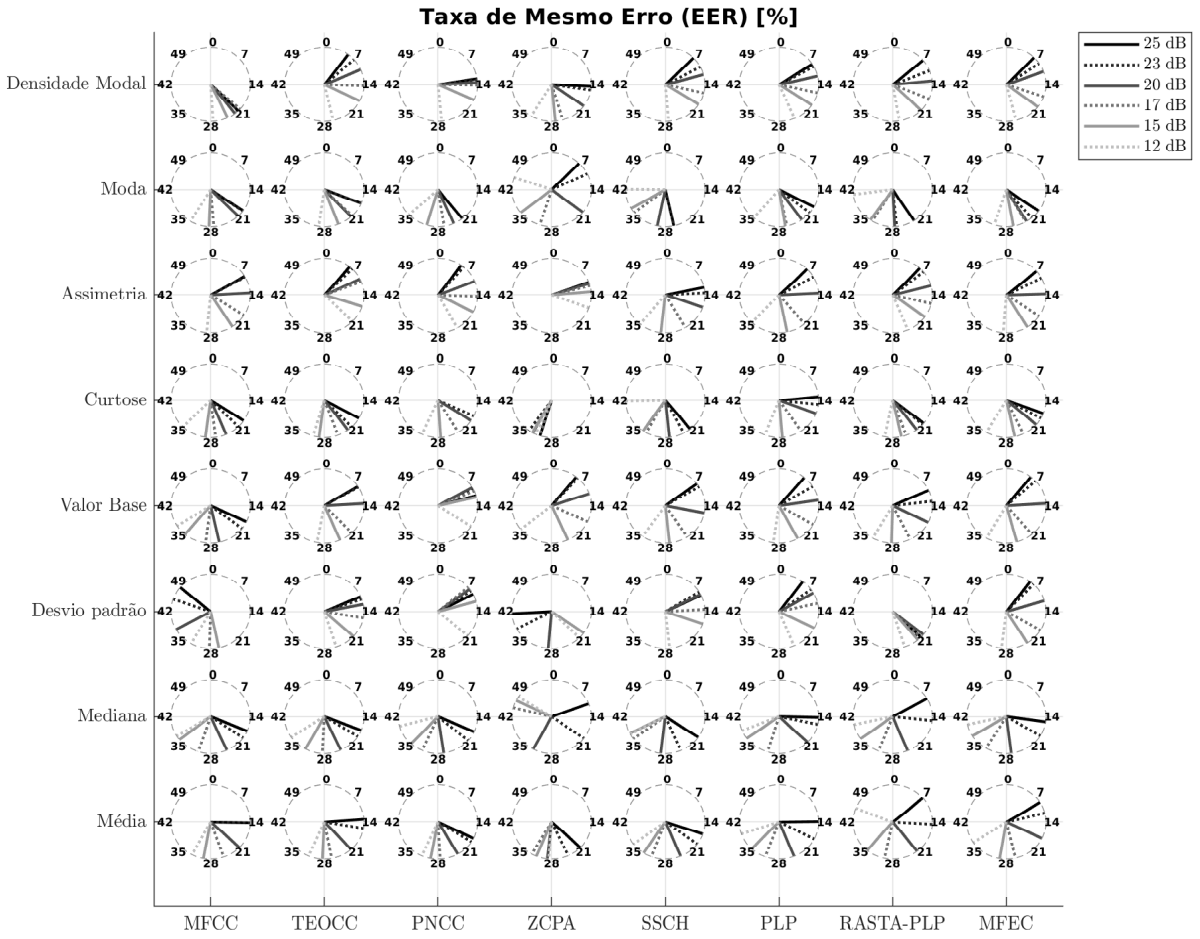

Figura 7: Diagrama apresentando a EER com recorte por características, grandeza estatística e intensidade de ruído. A abertura angular de cada círculo, a partir do eixo vertical no sentido anti-horário, indica a EER na escala entre 0 e 56 %. A posição do círculo representa a combinação de um descritor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

6 Agradecimentos

Os autores gostariam de agradecer a Geoffrey Morrison, Dan Ellis, Chanwoo Kim e Nathaniel Phillips por sempre compartilhar códigos. Agradecimentos também são direcionados a equipe do Setor de Perícias em Áudio e Vídeo do Instituto de Criminalística de Minas Gerais, em especial a Harley Cesar de Melo pelo apoio no desenvolvimento dos trabalhos. 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.

As bases de dados e códigos para reprodução dos experimentos estão disponíveis em https://bit.ly/2Kcd7ho.

Referências

Aitken, C. G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **53**(1): 109–122.

Davis, S. B. and Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *Readings in speech recognition*, Elsevier, pp. 65–74.

Enzinger, E. and Morrison, G. S. (2017). 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* **277**: 30–40.

Gajic, B. and Paliwal, K. K. (2001). Robust parameters for speech recognition based on subband spectral centroid histograms, *Seventh European Conference on Speech Communication and Technology*.

Gold, E. (2014). Calculating likelihood ratios in forensic speaker comparison cases using phonetic and linguistic features, *Unpublished PhD Thesis, University of York* .

Gold, E. and French, P. (2011). International practices in forensic speaker comparison., *In-*

ternational Journal of Speech Language and the Law. **18**(2): 293–307.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech, *the Journal of the Acoustical Society of America* **87**(4): 1738–1752.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech, *IEEE transactions on speech and audio processing* **2**(4): 578–589.

Holambe, R. S. and Deshpande, M. S. (2012). Noise robust speaker identification: using nonlinear modeling techniques, *Forensic Speaker Recognition - Law Enforcement and Counter-Terrorism*, Springer, pp. 153–182.

Hughes, V. (2014). *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*, PhD thesis, University of York.

Jabloun, F., Cetin, A. E. and Erzin, E. (1999). Teager energy based feature parameters for speech recognition in car noise, *IEEE Signal Processing Letters* **6**(10): 259–261.

Jam, M. M. and Sadjedi, H. (2009). Identification of hearing disorder by multi-band entropy cepstrum extraction from infant's cry, *Biomedical and Pharmaceutical Engineering, 2009. ICBPE'09. International Conference on*, IEEE, pp. 1–5.

Kacur, J., Varga, M. and Rozinaj, G. (2012). ZCPA features for speech recognition, *Telecommunications (BIHTEL), 2012 IX International Symposium on*, IEEE, pp. 1–4.

Kim, C. and Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* **24**(7): 1315–1329.

Kim, D.-S., Lee, S.-Y. and Kil, R. M. (1999). Auditory processing of speech signals for robust speech recognition in real-world noisy environments, *IEEE Transactions on speech and audio processing* **7**(1): 55–69.

Kinoshita, Y., Ishihara, S. and Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition., *International Journal of Speech, Language & the Law* **16**(1).

Lindh, J. and Eriksson, A. (2007). Robustness of long time measures of fundamental frequency, *Eighth Annual Conference of the International Speech Communication Association*.

Morrison, G. S. (2011). 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* **53**(2): 242–256.

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S. and Dorny, C. G. (2016). Interpol survey of the use of speaker identification by law enforcement agencies, *Forensic science international* **263**: 92–100.

Morrison, G. S., Zhang, C. and Rose, P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system, *Forensic science international* **208**(1): 59–65.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification, *IEEE Transactions on Speech and Audio Processing* **2**(4): 639–643.

Rose, P. (2003). *Forensic speaker identification*, CRC Press.

Silva, R. R. d., Costa, J. P. C. L. d., Miranda, R. K. et al. (2016). Aplicação do valor de base da frequência fundamental via estatística MVKD em comparação forense de locutor, *Revista Brasileira de Criminalística* **5**(3): 30–38.

Sohn, J., Kim, N. S. and Sung, W. (1999). A statistical model-based voice activity detection, *IEEE signal processing letters* **6**(1): 1–3.

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S. and Wang, R. (2017). Speaker identification features extraction methods: A systematic review, *Expert Systems With Applications* **90**: 250–271.

Togneri, R. and Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues., *IEEE Circuits And Systems Magazine* .