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

coletado em ambiente controlado por perito treinado utilizando procedimento operacional padronizado.

Em regra, 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 sopesar as hipóteses de os registros, questionado e padrão, serem ou não do mesmo indivíduo.

Os levantamentos realizados por [1] e [2] 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 [3] 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 [4–7], mais voltados para a área forense, apresentam estudos baseados em características perceptivas, e.g., frequência fundamental e formantes.

Nesse nicho, o presente trabalho busca explorar o potencial de características não-perceptivas, como MFCC e suas variantes, em condições próximas às encontradas na prática forense, i.e., em áudios com codificação GSM, banda estreita e ruído de canal.

Também é objetivo do trabalho estudar as limitações das diferentes técnicas aplicadas em CFL no Português Brasileiro, apontar índices de validade, confiabilidade e os limites de cada técnica.

Dentro desse contexto, o presente trabalho avalia diferentes características cepstrais quando simuladas em canal GSM, contaminadas com relação sinal ruído de 25, 23, 20, 17, 15 e 12 dB. Em complementação, é apresentado um breve estudo dos efeitos da codificação GSM e uma proposta de combinação das características visando reduzir taxa de erros.

O principal método de avaliação utilizado nos experimentos será a taxa de mesmo erro EER (*equal error rate*), pois o mesmo apresenta um melhor equilíbrio na relação dos erros do tipo I e do tipo II.

A contribuição principal do presente trabalho é estabelecer limites da EER que podem ser esperados quando uma etapa da CFL é realizada utilizando um determinado descritor com os áudios contaminados por uma faixa de SNR. Este trabalho também dá os primeiros passos na exploração da combinação de descritores acústicos para aumentar a robustez de uma comparação de locutores em áudios contaminados.

A próxima seção apresenta as bases matemáticas do cálculo dos descritores acústicos, a modelagem por função densidade de probabilidade e o método de inferência utilizado. A seção 3 apresenta a base de dados utilizada, os tipos de ruído, descreve como foram realizados os experimentos e apresenta os principais resultados. Na seção 4 são apesentadas análise de ralação entre os descritores acústicos e três propostas de combinação para melhoria de robustez (i.e., redução da EER).

2. MODELAGEM ACÚSTICA

2.1. Características/Descritores Acústicos Utilizados

A extração de características 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 Prediction*), 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 [3], 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.

2.1.1. Pré-processamento

O pré-processamento é composto pelas etapas de pré-ênfase, divisão em quadros e o janelamento. A pré-ênfase consiste em aplicar de um filtro do tipo $y[n] = x[n] - \alpha x[n-1]$ para corrigir a declinação de 6 db/8^a do sinal de voz. Empiricamente utiliza-se o valor $0,95 \leq \alpha \leq 1$.

A divisão em quadros consiste na extração de trechos sobrepostos do sinal de voz. Neste trabalho, foram utilizados os quadros de $25\ ms$ deslocados a cada $10\ ms$. O janelamento é a multiplicação do quadro por uma função janela, que tem o efeito de suavizar os efeitos da duração do quadro.

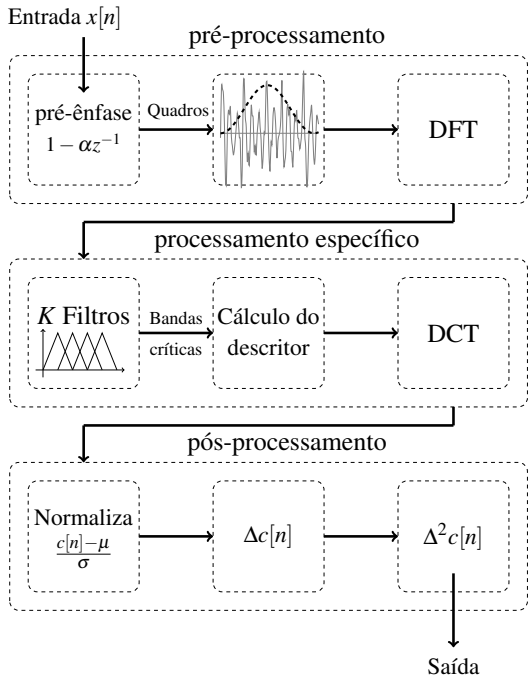

Figura 1: Diagrama de Blocos indicando as etapas, comuns e específicas, para o cálculo das características/descritores acústicos utilizados neste trabalho. No diagrama, o bloco DFT (*Discrete Fourier Transform*) corresponde a transformada discreta de Fourier, o bloco DCT (*Discrete Cosine Transform*) à transformada discreta de cossenos. Os blocos $\Delta c[n]$ e $\Delta^2 c[n]$ são definidos em [8].

2.1.2. 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 de cossenos (DCT-*Discrete Cossine Transform*) em cada quadro (vide 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 [9] e [10] o filtro utilizado possui formato espectral 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 [11] e [12]. 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 [13, 14], o PLP realiza a normalização com base na curva de intensidade perceptiva (*loudness*). Os filtros, separados na escala de Bark, possuem uma forma particular, são assimétricos, com região central plana e têm diferentes decaimentos exponenciais para frequências altas e baixas. A contribuição do RASTA-PLP é a aplicação de uma filtragem relativa a amplitude do espectrograma.

O MFEC e o TEOCC, explorado respectivamente por [15, 16] 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 [17] por banda. Aos valores calculados são aplicados à DCT.

Diferentemente dos descritores anteriores, o SSCH [18] e o ZCPA [19] 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, nos experimentos do presente trabalho, alguns descritores foram adaptados em relação à forma do filtro e à escala, como apresentado na Tabela 1.

Tabela 1: Parâmetros dos descritores que foram utilizados no experimento para cálculo das características.

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.

2.1.3. Pós-Processamento

O pós-processamento envolve a normalização das características utilizando a média e variância de todos os locutores [20] e o cálculo das variações temporais de primeira Δc (delta cepstrum) e segunda ordem $\Delta^2 c$, ao longo dos T quadros do áudio (mais detalhes vide [8, 20]).

2.2. Modelo do Locutor e Classificação

Para comparação de locutores, o modelo de mistura de gaussianas λ (GMM - *Gaussian Mixture Model*) representa a função densidade de probabilidade (FDP) das características $\vec{x}$ do locutor a partir da soma ponderada de distribuições normais (multidimensionais) com médias $\vec{\mu}_g$ e matriz de variância Σ_i conhecidas [8]:

$$\lambda = \{w_g, \vec{\mu}_g, \Sigma_g\} \; para \; g \; = \; 1, 2, ..., G, \qquad (1)$$

onde w_i é o peso ou fator de ponderação de cada gaussiana na mistura e G o número de distribuições.

Assim, $p(\vec{x}|\lambda)$, na Equação 2, é a probabilidade *a piori*, de os dados $\vec{x}$ terem sido gerados pelo modelo λ [8]:

$$p(\vec{x}|\lambda) = \sum_{g=1}^{G} \frac{w_g}{\sqrt{(2\pi)^D |\Sigma_g|)}} \cdot e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_g)\Sigma_g^{-1}(\vec{x}-\vec{\mu}_g)^T}, \quad (2)$$

onde D é a dimensionalidade de $\vec{x}$. Desta forma, para ajustar o modelo λ, utiliza-se a maximização da expectância (EM - *Expected Maximization*) [8, 21], ajustando o modelo paramétrico do locutor λ aos dados $\vec{x}$.

Para uma observação $\vec{x}_Q$ – que representa um conjunto de medições acústicas do áudio questionado –, e o modelo de locutor λ_P – obtido da amostra padrão $\vec{x}_P$ –, é possível formalizar as hipóteses[1]:

$$H : \begin{cases} H_0 : \theta \in \Theta_0 \rightarrow & \text{a amostra } \vec{x}_Q \text{ não provém do mesmo locutor que gerou o modelo } \lambda_P; \text{ e} \\ H_1 : \theta \notin \Theta_1 \rightarrow & \text{a amostra } \vec{x}_Q \text{ provém do mesmo locutor que gerou o modelo } \lambda_P. \end{cases} \qquad (3)$$

Entrando no contexto forense, têm-se a amostra questionada $\vec{x}_Q$, a amostra padrão $\vec{x}_P$ ou seu modelo λ_P e uma base de referência $\vec{x}_{UBM}$ ou seu modelo λ_{UBM}[2]. A utilização do UBM permite avaliar a ocorrência das características e melhora o desempenho na comparação baseada em GMM.

Uma métrica que permite decidir, a partir de $\vec{x}_Q$, $\vec{x}_P$ e $\vec{x}_{UBM}$, entre as hipóteses H_0 e H_1, definidas em 3, é a razão de verossimilhança $LR(\vec{x}_Q)$ (*Likelihood Ratio*):

$$LR(\vec{x}_Q) = \frac{p(\vec{x}_Q|H_0)}{p(\vec{x}_Q|H_1)} \begin{cases} \leq \zeta_0 & H_0 \text{ não é rejeitada;} \\ \geq \zeta_0 & H_0 \text{ é rejeitada.} \end{cases} \qquad (4)$$

[1] A Hipótese na Equação 3, assim como a Equação 4, são apresentadas de forma diferente de [8] principalmente para a adequação com o cenário forense proposto neste trabalho.

[2] As amostras questionada e padrão são conjuntos de medições acústicas da voz. A amostra $\vec{x}_{UBM}$, e o respectivo modelo de fundo λ_{UMB}, é obtido a partir de um banco de vozes de áudio padrão de diferentes locutores.

Na metodologia GMM-UBM tem-se como variável de decisão a "razão de verossimilhança" obtida entre o modelo de locutor comparado λ_P e o UBM λ_{UBM} que, para fins de formulação, é calculada como o logaritmo da razão de verosimilhança (LLR *log-likelihood ratio*). A estatística $LLR(\vec{x})$ fica, assim,

$$LR(\vec{x}_Q) = \frac{p(\vec{x}_Q|\lambda_P)}{p(\vec{x}_Q|\lambda_{UBM})}$$
$$\downarrow \log()$$
$$LLR(\vec{x}_Q) = \log\left(\frac{p(\vec{x}_Q|\lambda_P)}{p(\vec{x}_Q|\lambda_{UBM})}\right), \qquad (5)$$

onde o valor da estatística $LLR(\vec{x}_Q) \in [-\infty, \infty]$.

Para comparação de amostras de voz, Reynolds e colaboradores [8] propõem que no cômputo da Equação 5 seja utilizada a média das verossimilhanças computadas em cada um dos T quadros de voz. A Equação 5, para fins de comparação de voz, toma a forma

$$\log\left(p(\vec{x}_Q|\lambda_P)\right) = \frac{1}{T}\sum_{t=0}^{T-1} \log\left(p(x_Q[t]|\lambda_P)\right). \qquad (6)$$

Onde λ é o modelo do locutor e $x_Q[t]$ é a característica questionada no quadro t. Esta normalização pela quantidade de quadros de voz foi proposta para compensar o efeito da duração dos trechos de voz.

Nota-se que esta normalização transforma o resultado da Equação 5 em uma pontuação, que devidamente calibrada, indica se é mais provável (ou não) que a amostra $\vec{x}_Q$ seja oriunda do locutor λ_P.

3. DESEMPENHO FRENTE A SIMULAÇÃO DAS CONDIÇÕES FORENSES

3.1. Configuração Experimental para Análise das Características Acústicas

Nesta etapa, foi planejado um experimento para avaliar a EER da comparação automática de locutores, utilizando a metodologia GMM-UBM, e as amostras do Corpus Cefala-1 [22] obtidas pelo aparelho celular.

Na preparação dos áudios, primeiramente foi realizada uma subamostragem para 8 kHz precedida de uma filtragem com faixa de passagem entre 300 e 3500 Hz. Cada amostra do corpus foi separada de acordo com sua etapa de coleta: a fala espontânea, leitura de texto corrido e leitura de frases isoladas.

A concatenação da etapa de leitura de texto com 66% da etapa de fala espontânea originou o *áudio padrão*. Estes áudios foram utilizados para parametrizar os modelos GMM e UBM da denominada *Base de Treinamento Não Alterada*.

A versão do áudio padrão, codificada e decodificada pelo *codec* GGSM 06.60 [23], foi utilizada para parametrizar modelos os GMM e UBM da denominada *Base de Treinamento GSM*.

Os 34% restantes da etapa de fala espontânea com a etapa de leitura de frases foram concatenados para criar o *Áudio Teste*. Este áudio, que não é representativo para a comparação forense, foi utilizado como referência para o valor da EER do experimento. A versão do *Áudio Teste* processada pelo *codec* GSM 06.60 foi denominado *Áudio Questionado*.

Os *Áudios Contaminados* foram gerados a partir dos *Áudio Teste*. Tais amostras foram contaminadas por ruídos de *Trânsito*, *Disparos*, *Multidão*[3], e ruídos *Branco* e *Rosa*. Para cada ruído buscou-se a contaminação com SNR nos valores de 25, 23, 20, 17, 15 e 12 dB, resultando em um total de 30 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 GSM.

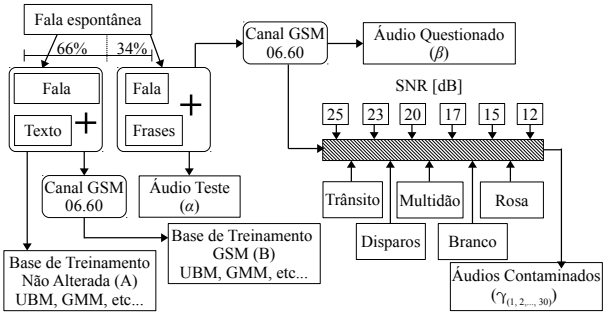

Figura 2: Diagrama apresentando como foram obtidas as amostras utilizadas no experimento de contaminação. As bases de treinamento são representadas pelas letras (A) e (B), e os conjuntos de áudio por (α), (β) e (γ).

Em resumo têm-se as seguintes amostras envolvidas no experimento:

Base de Treinamento Não Alterada Conjunto de áudios padrão utilizados para parametrizar os modelos GMM e UBM. É indicado na Figura 2 pelo quadro (A).

Base de Treinamento GSM Conjunto de áudios padrão, processados pelo *codec* GSM 06.60, utilizados para parametrizar os modelos GMM e UBM. É indicado na Figura 2 pelo quadro (B).

Áudio Teste Conjunto de amostras de áudio para comparação, retiradas diretamente do Corpus Cefala-1, utilizadas como referência nas comparações. Este conjunto é indicado na Figura 2 pelo quadro (α).

Áudio Questionado Áudios resultantes do processamento dos áudios teste pelo *codec* GSM 06.60. Este conjunto é indicado na Figura 2 pelo quadro (β).

[3]Este ruído pode ter a mesma classificação do *babble noise*.

Áudios Contaminados Áudios resultantes da contaminação e do processamento dos áudios teste pelo *codec* GSM 06.60. Este conjunto é indicado na Figura 2 pelo quadro (γ). Para cada amostra do Corpus Cefala-1 foram geradas 30 amostras contaminadas com diferentes tipos e níveis de ruído.

As 32 amostras por locutor (Teste, Questionada e Contaminadas) foram comparadas com os dois modelos GMM-UBM gerando um total de 64 resultados por locutor. Essas comparações ainda foram realizadas utilizando oito diferente descritores (ou características) que representam o sinal de voz:

- MFCC (*Mel-Frequency Component Cepstrum*);
- TEOCC (*Teager Enegy Operator Component Cepstrum*);
- PNCC (*Power Normalized Component Cepstrum*);
- ZCPA (*Zero-Crossing with Peak Amplitude*);
- SSCH (*Subband Spectral Centroid Histograms*);
- PLP (*Perceptual Linear Prediction*);
- RASTA-PLP (*Representations Relative Spectra*);
- MFEC (*Mel Frequency Entropy Cepstrum*).

3.2. Ruído Contaminante

O ruído contaminante foi escolhido com base em dois critérios. O primeiro é sobre a tipicidade de ruídos encontrados na CFL, sendo eles o ruído de trânsito, disparos e multidão. Os ruídos branco e rosa – de comportamento conhecido –, foram utilizados como parâmetros de referência.

O segundo critério foi a dinâmica temporal. Sendo que os ruído de trânsito, disparos e multidão possui uma maior variação da amplitude, enquanto os ruídos rosa e branco possuem amplitudes estacionárias.

Tabela 2: Tabela apresentando a caracterização de amplitude dos ruídos interferentes. O valor máximo que o sinal pode assumir é 1.0.

Descrição	RMS	Pico.
Trânsito de veículos.	0,057	0,412.
Disparos de arma de fogo.	0,082	0,886.
Multidão (*babble noise*).	0,065	0,502.
Ruído branco	0,339	1,000.
Ruído rosa.	0,211	0,960.

A Fig. 3 juntamente com a Tab. 2 apresentam algumas características dos ruídos contaminantes. Um detalhe observável na Fig. 3 é a amplitude não estacionária dos ruídos, com exceção dos ruídos branco e rosa[4].

[4]Os três primeiros ruídos foram obtidos em 24/12/2016 no sítio https://www.buscasons.com e são os arquivos "exterior_cidade_transito_-200905281114_longo-01.mp3", "Metrelhadora0001.mp3" e "publico_-area_grande.mp3". Estes arquivos foram convertidos para um canal e frequência de amostragem de 8 kHz. Os demais ruídos foram gerados pelo programa de edição de áudio Audacity.

A Figura 4 apresenta o espectro médio de longo termo (LTAS - long term averange spectrum) dos ruídos.

3.3. Resultados das Comparações

Passando para análise dos resultados, no gráfico RDI (*Raw-data Description and Inference*) 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 $\alpha = 0,05$.

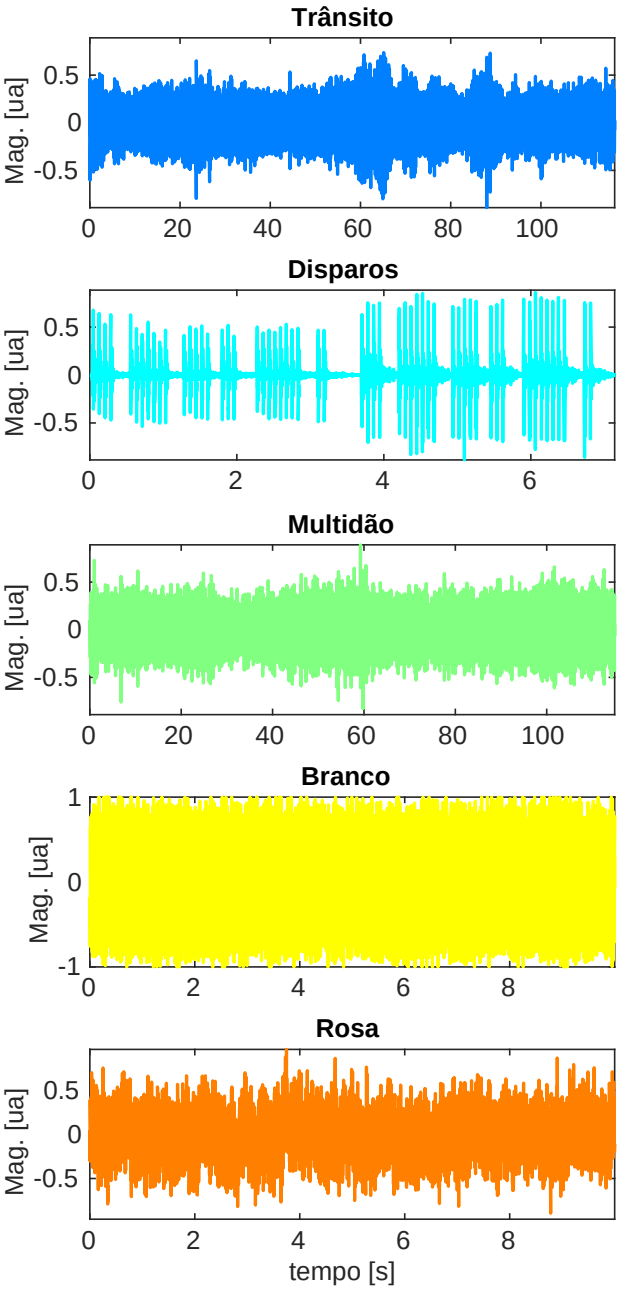

Figura 3: Caracterização dos tipos de ruído no domínio do tempo. Na imagem nota-se que os ruídos de transito, disparos e multidão não apresentam estacionariedade de amplitude

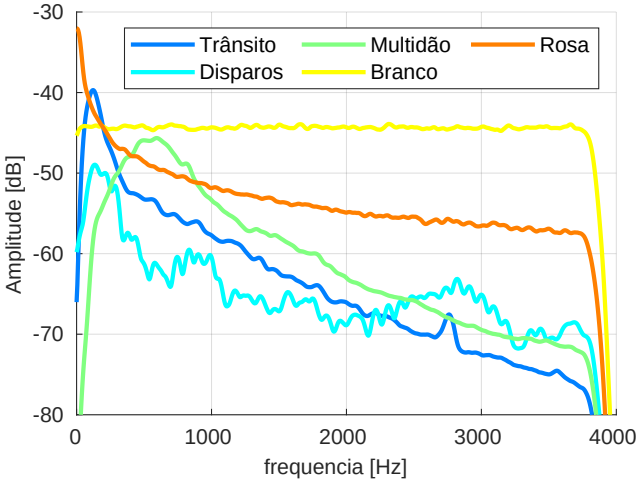

Figura 4: Espectro médio de longo termo (LTAS) dos tipos de ruído utilizados nos experimentos.

A figura 5 apresenta um recorte geral sobre a EER de cada característica, mostrando que o MFCC apresentou menor EER médio frente aos demais.

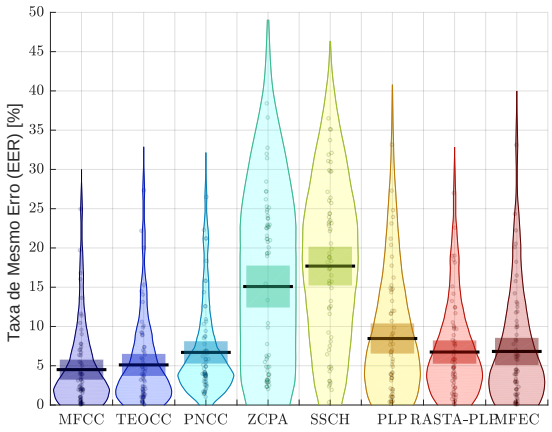

Figura 5: Gráfico RDI apresentando a diferença na taxa de mesmo erro (EER) apresentado pelas diferentes características na comparação de locutores. Em cada coluna, os pontos são as durações individuais, as curvas laterais são a distribuição de probabilidade empírica, a linha preta horizontal a média e o retângulo escuro é o intervalo de confiança da média para $\alpha = 0,05$. Nas colunas têm-se as características conforme indicado na Tabela 1 e descritos na Seção 3.1

Na análise de variância da figura 6, para um nível de significância de 5%, indica uma equivalência estatística entre as médias do MFCC com o TEOCC, PNCC, RASTA-PLP e MFEC.

Sob o ponto de vista do tipo de base de treinamento utilizada (Não-Alterada e GSM), a figura 7 mostra uma menor EER média de $12,5\%$ enquanto o GMM-UBM contaminado pelo canal GSM possui EER média de 6%. Este estes valores podem ser justificados pela influência do canal GSM no cálculo das características.

Nas comparações de locutor, 31 das 32 amostras questi-

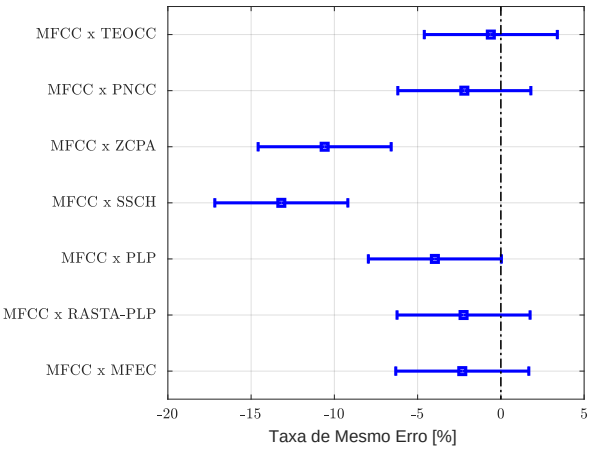

Figura 6: Analise de variância apresentando a EER do MFCC comparado com os demais descritores. Os pontos são as médias das diferenças e as linha horizontais o intervalo de confiança. A linha pontilhada vertical indica a diferença nula.

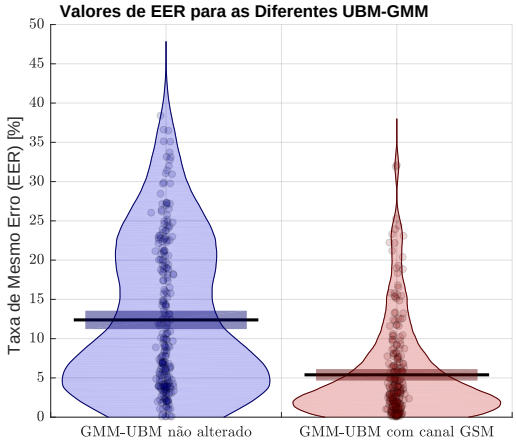

Figura 7: Gráfico RDI apresentando a diferença na taxa de mesmo erro (EER) de acordo a modelagem GMM-UBM. Vide detalhes sobre o gráfico RDI na legenda da Figura 5.

onadas por locutor foram codificadas e decodificadas pelo *codec* GSM 06.60. Desta forma, é plausível que um modelo de fundo que também foi contaminado pelo canal GSM apresente um menor valor da EER. Este resultado também deixa claro como a metodologia GMM-UBM é sensível aos dados utilizados como suporte.

Sobre os resultados que avaliam a contaminação por ruído, tem-se o resultado esperado do aumento da taxa de mesmo erro com a redução da relação sinal ruído (SNR), como apresentado na Figura 8. Entretanto, o importante deste resultado é que ele permite quantificar o quanto esperar de taxa de mesmo erro para uma determinada intensidade de ruído. Neste caso, o experimento mostrou que com uma relação sinal ruído de 17 dB tem-se uma taxa de mesmo erro médio menor que 10%.

Sobre o tipo de ruído, o branco destaca-se como o que apresenta as maiores taxas de mesmo erro (vide Figura 9). Os demais apresentam uma taxa de mesmo erro menor que 10%.

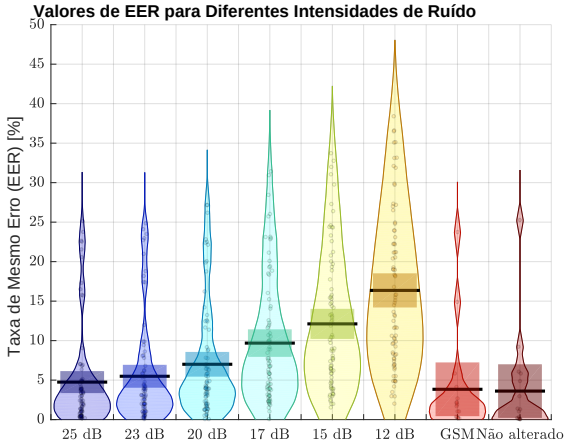

Figura 8: Gráfico RDI apresentando a diferença na taxa de mesmo erro (EER) pela relação sinal ruído (SNR). Vide detalhes sobre o gráfico RDI na legenda da Figura 5.

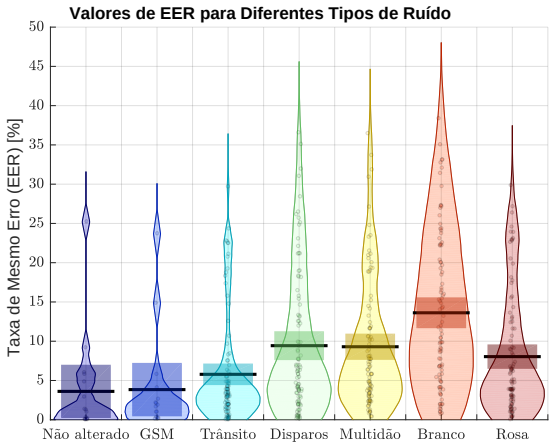

Figura 9: Gráfico RDI apresentando a diferença na taxa de mesmo erro (EER) de acordo com o tipo de contaminação. Vide detalhes sobre o gráfico RDI na legenda da Figura 5.

Nas figuras 10 a 11 são apresentados os recortes dos valores da taxa de mesmo erro para cada característica com alteração do tipo de ruído, SNR e tipo de base de treinamento apenas para modelo GMM-UBM contaminados pelo canal GSM.

Nos diagramas, cada círculo representa a combinação de uma característica, no eixo horizontal, com uma variável de controle, no eixo vertical. A abertura angular, a partir do eixo vertical no sentido horário, indica a taxa de mesmo erro na escala entre 0 e 40 $\%$. As figuras 10 e 11 apresentam resultados obtidos com UBM e GMM da base de treinamento GSM.

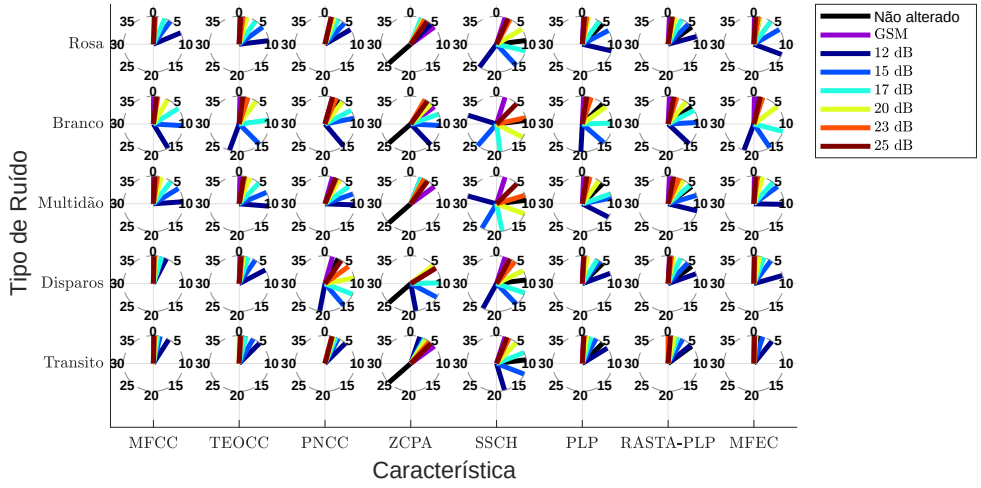

Figura 10: Gráfico apresentando a média da taxa de mesmo erro (EER) com recorte por características, tipo e intensidade de ruído para modelo GMM-UBM contaminados pelo canal GSM. Cada círculo representa a combinação de uma característica, no eixo horizontal, com uma variável de controle, no eixo vertical. A abertura angular, a partir do eixo vertical no sentido horário, indica a taxa de mesmo erro na escala entre 0 e 40 %.

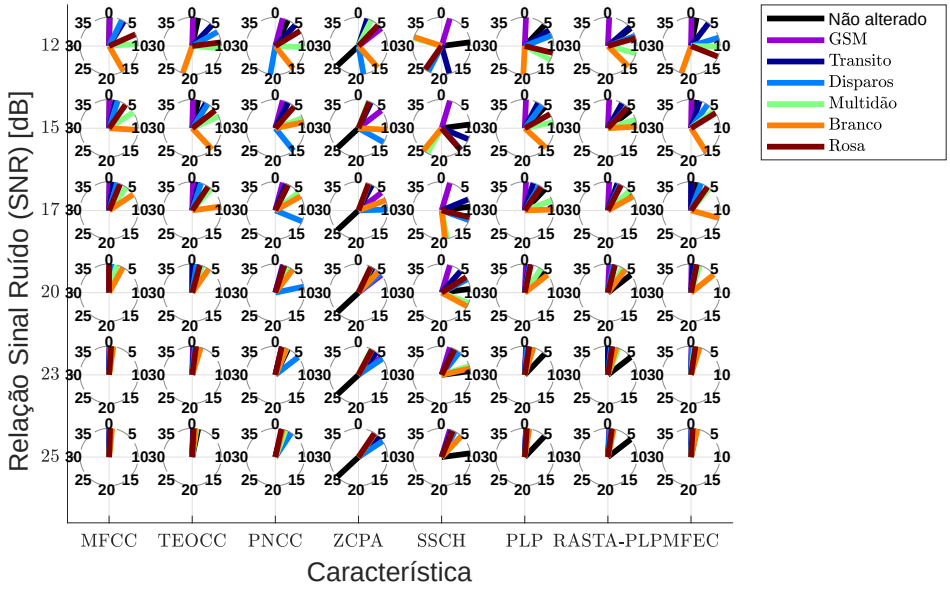

Figura 11: Gráfico apresentando a média da taxa de mesmo erro (EER) com recorte por características, tipo e intensidade de ruído para modelo GMM-UBM contaminados pelo canal GSM. Vide detalhes sobre o diagrama na legenda da Figura 10.

Na Figura 10, a variável de controle é o tipo de ruído, enquanto a SNR é representada pelas cores. Em todas os diagramas os áudios *não alterados* e *GSM* correspondem, respectivamente, a dados obtidos pelas amostras de áudio de teste e questionada.

Em um recorte específico, é possível analisar a taxa de mesmo erro quando os locutores são comparados por MFCC com os modelos GMM-UBM contaminados pelo canal GSM. A Figura 12 mostra que a EER média para uma contaminação de 15 dB é de 5%.

Sobre os resultados, o ruído branco é o que apresenta maior EER se comparado com as demais contaminações. Outro detalhe que se repete é a tendência de uma taxa de mesmo erro menor para a SNR menor.

Um resultado intuitivo é o fato de a amostra de teste apresentar uma taxa de mesmo erro elevada na base de treinamento GSM com o ZCPA (vide figuras 10 e 11).

As principais conclusões deste experimento, que agregam valor ao cenário forense, são:

- Em caso de comparação automáticas utilizando a metodologia GMM-UBM, onde a amostra questionada é contaminada pelo *codec* GSM 06.60, é recomendável que os modelos GMM e UBM sejam oriundos de áudios que também foram processados pelo *codec* GSM

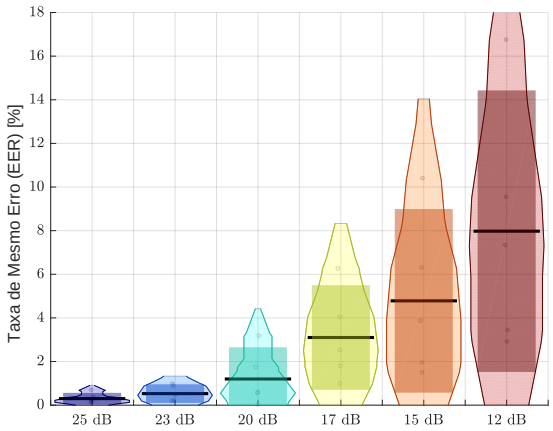

Figura 12: Gráfico RDI apresentando a diferença na taxa de mesmo erro (EER) de acordo com o tipo de contaminação – sendo a característica o MFCC GMM-UBM contaminados por GSM. Em cada coluna os pontos são as durações individuais, as curvas laterais são a distribuição de probabilidade empírica, a linha preta horizontal a média e o retângulo escuro é o intervalo de confiança da média para $\alpha = 0,05$.

06.60;

- As características MFCC, TEOCC, RASTA-PLP e MFEC foram as que apresentaram comportamento menos disperso; as últimas três apresentam menor EER para SNR acima de 17 dB com a base de treinamento GSM;
- O PNCC apresentou uma sensibilidade ao ruído de disparo de arma fogo (estampido);
- Com exceção da presença do ruído branco, o MFCC apresenta taxa de mesmo erro até 5 % para $SNR > 15\ dB$ se a base de treinamento for GSM.

É importante ressaltar que este estudo é um direcionamento que visa aproximar as condições da comparação à realidade que é encontrada nos laboratórios de áudio forense. Mais estudos precisam ser realizados e o próximo passo é apresentar uma investigação sobre os efeitos de interação entre as variáveis consideradas aqui.

4. ANÁLISES E ALTERNATIVAS PARA REDUÇÃO DE EER

Os resultados apresentados nas seções anteriores – em especial a equivalência estatística do MFCC com outros descritores (vide Fig. 6) –, motivaram uma análise exploratória dos dados para buscar responder algumas questões levantadas *a posteriori*.

A primeira delas é motivada pela menor EER quando utiliza-se o modelo GMM-UBM codificado pelo canal GSM (vide Fig. 7). A questão é qual o efeito da codificação GSM nas características acústicas da Tabela 3. Esta questão foi investigada observando as características calculadas antes e depois da codificação GSM.

A segunda questão é se a EER do MFCC pode ser superado ou melhorado pela combinação dos outros descritores. Em relação redução do EER, os autores adiantam que fora realizados apenas os estudos preliminares, com o conjunto de Áudios Questionados, e que o MFCC apresentou índices de superiores às estratégias propostas.

Diante do exposto acima os autores gostariam de informar que não foi encontrado na literatura especializada publicação com uma metodologia similar. O mais comum na literatura é a combinação do *score* obtidos por descritores ou algoritmos distintos, uma vez que os métodos de extração são semelhantes (vide Figura 1).

Em segundo lugar, este tipo de exploração não trás absolutamente nenhuma garantia que pode reduzir o EER da verificação de locutores (ou qualquer outro índice), por isso é uma análise exploratória e preliminar.

Em terceiro lugar, os descritores acústicos (vide seções 2.1 e 3.1) são consequência físico-causal de aspectos distintos da produção da fala, porém não é garantido que uma estratégia de combinação adicione informação relevante para o problema de verificar locutores ou reduza incertezas.

Adiantamos que as estratégias apresentadas para combinação das características acústicas não apresentaram redução no valor da EER. Porém, como foram resultados exploratórios e preliminares não é possível afirmar se estamos "caminhando em círculos" pois foi uma análise *a posteriori* de um experimento e, de forma alguma, as três estratégias sugeridas podem ser consideradas como únicas[5]

4.1. Relação entre as Características

Para estudar o efeito da codificação GSM no cálculo das características, foram utilizadas duas medidas: a correlação de Pearson e a informação mútua. Para cada um dos 104 locutores, foram calculadas a correlação e a informação mútua entre cada uma das 39 dimensões para cada um das 8 características extraídas antes (Áudio Teste) e após a codificação GSM (Áudio Questionado). Isto é, cada extrato de características calculada do Áudio Teste foi confrontada com sua equivalente calculada do Áudio Questionado.

As figuras 13 e 14 apresentam, respectivamente, os gráficos RDI para a correlação e para informação mútua. Nota-se nas imagens que as características que apresentam maior correlação também apresentam maior informação mútua. Este resultado não é uma confirmação definitiva, porém é um indicativo que o ZCPA, o PNCC e o MFCC são características

[5] Se considerarmos apenas a restrição de escolher 39 dimensões em um grupo de 8 descritores com 39 dimensões cada tem-se $\binom{8\times39}{39} \approx 7,7 \times 10^{49}$ combinações.

menos sensíveis à codificação GSM.

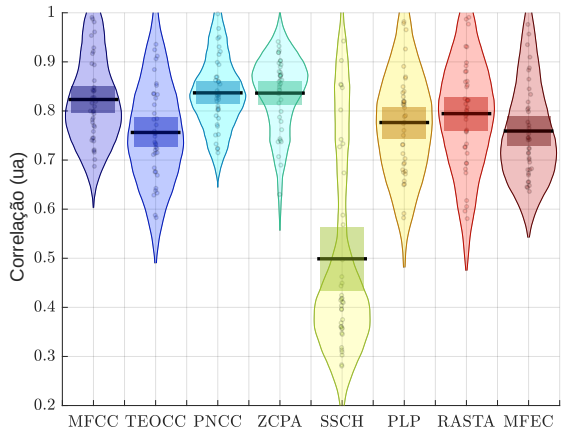

Figura 13: Gráfico RDI apresentando a correlação de Pearson entre as características calculadas antes (Áudio Teste) e após (Áudio Questionados) a codificação GSM. No gráfico, cada ponto nas colunas indica a correlação para cada um dos 104 locutores.

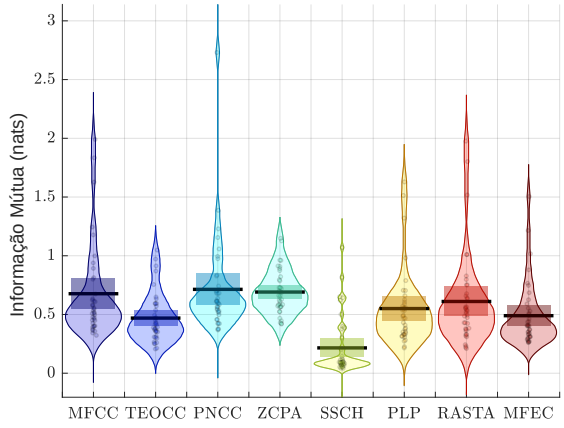

Figura 14: Gráfico RDI apresentando a informação mútua entre as características calculadas antes (Áudio Teste) e após (Áudio Questionados) a codificação GSM. No gráfico, cada ponto nas colunas indica a correlação para cada um dos 104 locutores.

Outra pergunta é o quanto dois descritores (ou características) podem estar correlacionados. Desta forma, foram calculadas a correlação de Pearson com a informação mútua entre as características extraídas do Áudio Questionado. Estas estatísticas foram calculadas ao longo das 39 dimensões relacionando as oito características par a par[6].

A Fig. 15 apresenta o perfil em cada ponto a relação da correlação de Pearson (eixo horizontal) com a informação mútua (eixo vertical) calculada utilizando logaritmo de base natural medido em unidades naturais de informação (nats - *natural unit of information*). A linha contínua – usada como referência–, indica a relação entre a correlação ρ e a informação mútua de uma variável aleatória gaussiana bidimensional. As médias par a par são indicadas por "$\times$" pretos.

Neste caso, nota-se que mesmo características que estão pouco correlacionadas possuem uma informação mútua acima da linha de referência. Naturalmente estas análises não respondem o questionamento de como as características calculadas após a codificação se relacionam com as calculadas dos áudios não codificados. Porém trata-se de um primeiro passo para entender como diferentes descritores se relacionam.

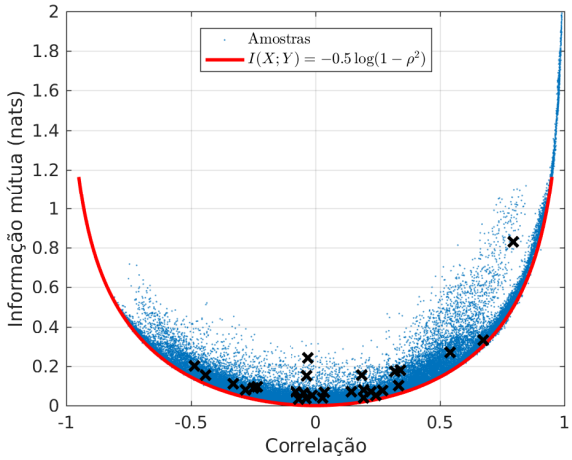

Figura 15: Gráfico indicando as ocorrências de correlação pela informação mútua (utilizando logaritmo de base natural) medido em unidades naturais de informação (nats - *natural unit of information*). Os pontos indicam cada relação para cada dimensão das diferentes características calculadas por locutor. A linha vermelha indica a relação entre a correlação de Pearson e a informação mútua de uma variável aleatória gaussiana bidimensional.

Os pares de descritores descorrelacionados e com pouca informação mútua têm potencial para serem combinados e motivaram o desenvolvimento de pelo menos uma estratégia para redução da EER descrita na próxima seção.

4.2. Estratégias para redução de EER

A presente seção propõe o estudo de combinação de características com objetivo de reduzir o EER médio na comparação de locutores. As formas de combinação das características fazem uso da premissa que as dimensões dos descritores acústicos podem ser pareadas, pois espera-se que uma dimensão em particular de alguma característica (descritor) pode apresentar um poder discriminante maior. Entretanto é importante ter em mente que as dimensões são obtidas de faixas espectrais diferentes (utilizando escala e formas de filtros diferentes, vide Tab. 1).

A hipótese então é: é possível obter uma combinação dos descritores que pode reduzir o EER médio da verificação

[6]Monta um total de oito características combinadas duas a duas em 39 dimensões dos 104 locutores. Em suma $\binom{8}{2} \times 39 \times 104 = 113568$ coeficientes de correlação.

de locutor. Para testar esta hipótese foram propostas três estratégias. A primeira, baseada na divergência de Kullback-Liebler [24], a segunda, na distância de Mahalanobis [21] e a terceira, na taxa de mesmo erro (min. EER) apresentada por cada dimensão.

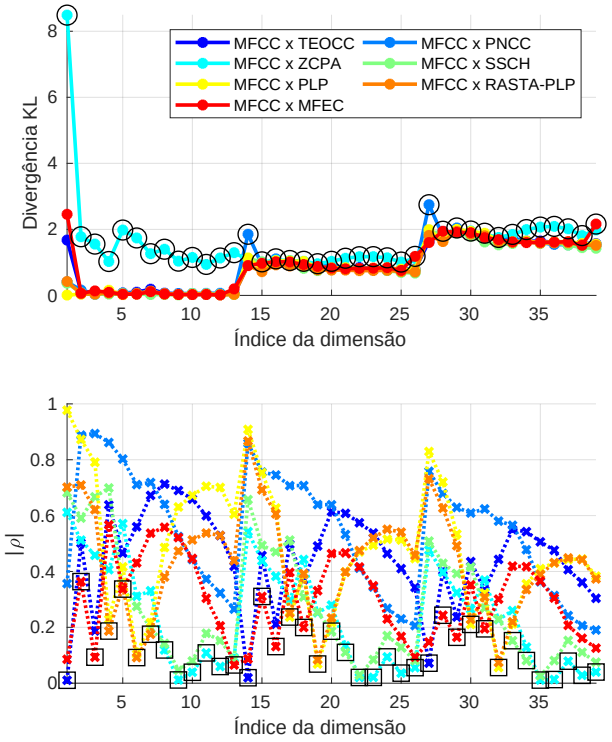

Figura 16: Divergência de Kullback-Liebler e correlação dos decritores com o MFCC. No painel superior tem-se a divergência de Kullback-Liebler, e os círculos indicam as maiores ocorrências. No painel inferior, o módulo da correlação de Pearson, e os quadrados indicam as ocorrências mais próximas de zero.

Tomando o MFCC como referência, é possível utilizar a divergência de Kullback-Liebler como uma medida de ganho de informação de uma característica em relação ao MFCC [24]. Esta estratégia, combinada com a correlação de Pearson, foi utilizada para selecionar as características para ocupar cada uma das 39 dimensões. Sendo representada por *KLD* nas figuras 18 e 19 e na Tab. 3.

O painel superior da Fig. 16 apresenta a divergência de Kullback-Liebler de cada descritor, em relação ao MFCC, para as 39 dimensões. Os descritores foram ranqueados pela maior divergência (decrescente) e menor módulo da correlação (crescente) e foram selecionados os descritores com menor soma dos ranques. A Tab. 3 indica qual descritor foi escolhido para cada dimensão utilizando a estratégia supra.

A segunda estratégia mediu a distância de Mahalanobis – em unidades arbitrárias (ua) –, entre os 104 locutores para cada dimensão dos descritores. A Fig. 17 apresenta, no painel superior, a distância de Mahalanobis média em cada dimensão. Os autores gostariam de chamar atenção para o fato de as distâncias de Mahalanobis apresentarem valores muito inferiores à unidade.

Esta estratégia visa selecionar os descritores por um processo interativo. Primeiramente, os descritores são ranqueados (ordem decrescente) em cada dimensão. Seleciona-se o descritor da primeira dimensão e este descritor forma dois pares utilizando os dois primeiros ranqueados da segunda dimensão. Destes pares seleciona-se o que apresenta a maior distância de Mahalanobis. O processo é repetido para cada dimensão (componente cepstral) até a 39^a. Esta estratégia é representada por *Maha* nas figuras 18 e 19 e na Tab. 3.

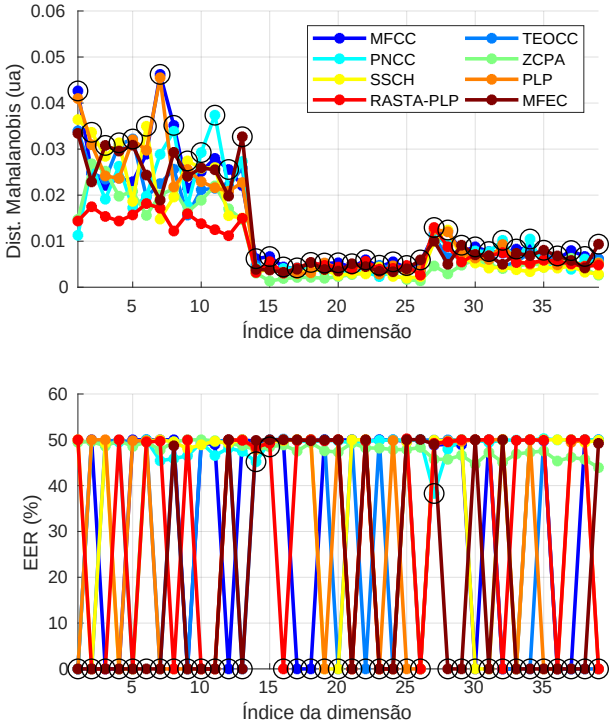

Figura 17: Distância de Mahalanobis (em unidades arbitrárias - ua) e taxa de mesmo erro (ERR) por descritor e por dimensão. Na figura do topo tem-se a distância de Mahalanobis média entre diferentes locutores. No gráfico da base tem-se a taxa de mesmo erro também para cada descritor e por dimensão.

Por fim, a estratégia de mínima EER realiza a comparação de locutor utilizando apenas uma dimensão de cada descritor por vez. Os descritores são selecionados por dois critérios, o primeiro é a menor taxa de mesmo erro e, em caso de empate, seleciona-se o descritor com menor EER geral (vide Fig. 5). Esta estratégia é representada por *min. EER* nas figuras 18 e 19 e na Tab. 3. A Fig. 17 apresenta no gráfico da base a a taxa de mesmo erro em cada dimensão dos descritores.

A EER de cada estratégia, em relação ao MFCC pode ser visualizado na Fig. 18. No painel superior tem-se a comparação da taxa de mesmo erro (EER) enquanto, no painel

inferior, o custo do logaritmo da razão de verossimilhança (C_{LLR}) calculado utilizando o logaritmo de base natural medido em nats [25].

Nota-se que o MFCC apresentou os menores índices. A comparação dos valores do logaritmo da razão de verossimilhança $LLR(\vec{x}_Q)$ de cada estratégia é apresentada no painel superior da Fig. 19. Na imagem, as barras representam o $LLR(\vec{x}_Q)$ médio enquanto as barras verticais os limites para três desvios padrões das ocorrências do $LLR(\vec{x}_Q)$. Na comparação entre áudios do mesmo locutor (barras azuis) espera-se o valor de $LLR(\vec{x}_Q)$ maior que zero e para locutores diferentes (barra vermelha), esperam-se valores inferiores a zero.

Nota-se que, no MFCC, uma menor faixa do espaço do LLR se sobrepõe (linhas verticais pretas sobre as barras), enquanto, para as demais estratégias, o intervalo de sobreposição do LLR é maior. Para fins ilustrativos, o painel inferior da Fig. 19 indica quantas dimensões de cada descritor foram utilizadas em cada estratégia.

Em relação a EER das estratégias propostas, não é uma surpresa inesperada que as combinações baseadas na divergência de Kullback-Liebler, na distância de Mahalanobis

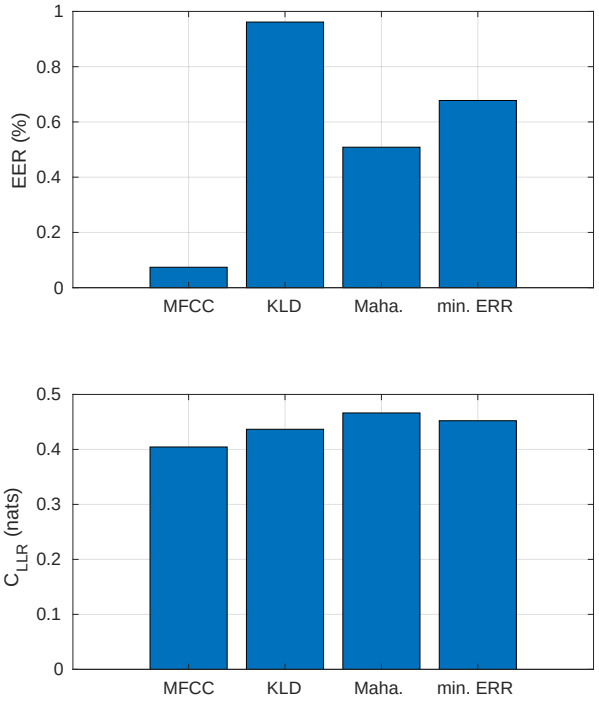

Figura 18: Desempenho das estratégias propostas frente ao MFCC. No gráfico do topo tem-se a comparação da taxa de mesmo erro (EER) enquanto no gráfico da base o custo do logaritmo da razão de verossimilhança (C_{LLR}).

Tabela 3: Características selecionadas para cada uma das estratégias de combinação de descritores.

Dim.	KLD	Maha	min-ERR
1	TEOCC	MFCC	MFCC
2	ZCPA	SSCH	TEOCC
3	MFEC	SSCH	MFCC
4	RASTA	SSCH	MFCC
5	SSCH	TEOCC	PNCC
6	RASTA	SSCH	MFEC
7	RASTA	PLP	TEOCC
8	ZCPA	MFCC	RASTA
9	ZCPA	PLP	MFCC
10	PNCC	PNCC	RASTA
11	ZCPA	MFCC	RASTA
12	SSCH	MFCC	MFCC
13	SSCH	PNCC	TEOCC
14	TEOCC	TEOCC	PNCC
15	SSCH	RASTA	ZCPA
16	TEOCC	PNCC	RASTA
17	PLP	PNCC	MFCC
18	PLP	MFEC	MFCC
19	SSCH	PLP	PLP
20	ZCPA	RASTA	TEOCC
21	PLP	MFEC	RASTA
22	PNCC	RASTA	TEOCC
23	SSCH	TEOCC	RASTA
24	SSCH	PLP	TEOCC
25	PNCC	RASTA	PLP
26	ZCPA	MFEC	MFCC
27	TEOCC	PLP	PNCC
28	ZCPA	SSCH	MFEC
29	TEOCC	MFEC	MFEC
30	PLP	PNCC	MFCC
31	SSCH	MFCC	MFEC
32	SSCH	PNCC	MFCC
33	ZCPA	MFEC	MFCC
34	SSCH	PNCC	TEOCC
35	PNCC	TEOCC	MFEC
36	SSCH	MFEC	MFCC
37	SSCH	ZCPA	MFCC
38	PNCC	PNCC	TEOCC
39	ZCPA	TEOCC	TEOCC

e na mínima EER não apresentassem melhorias. Vejamos alguns detalhes capazes de explicar estes resultados.

Primeiramente, a verificação de locutores é realizada como descrito na seção 2.2. Esta classificação é baseada na média do logaritmo da razão de verossimilhança entre o modelo de locutor padrão e o UBM (vide equações 5 e **??**). Basicamente, $LLR(\vec{x}_Q)$ maior que zero indica um grau de equivalência [26] de informação das amostras – questionada $\vec{x}_Q$ e padrão $\vec{x}_P$ –, com o modelo padrão λ_P.

Desta forma, espera-se que um descritor que apresente uma Kullback-Liebler maior que o MFCC apresente mais informação que o MFCC. Esta premissa não é falsa, mas o ganho de informação não necessariamente implica que esta informação é relevante para distinguir os locutores. Para obter sucesso com uma estratégia como esta é preciso isolar o ganho de informação que é relevante na distinção dos locutores. Esta proposta será estuda na continuidades dos trabalhos.

Sobre a distância de Mahalanobis entre os locutores, a estratégia foi baseada em medidas unidimensionais dos descritores. Como a inferência é multidimensional, a combinação de dimensões com uma distância de Mahalanobis elevada não implica no aumento desta distância entre os locutores. Mesmo que a estratégia de seleção sempre teste entre dois descritores para que a combinação agregada resulte em

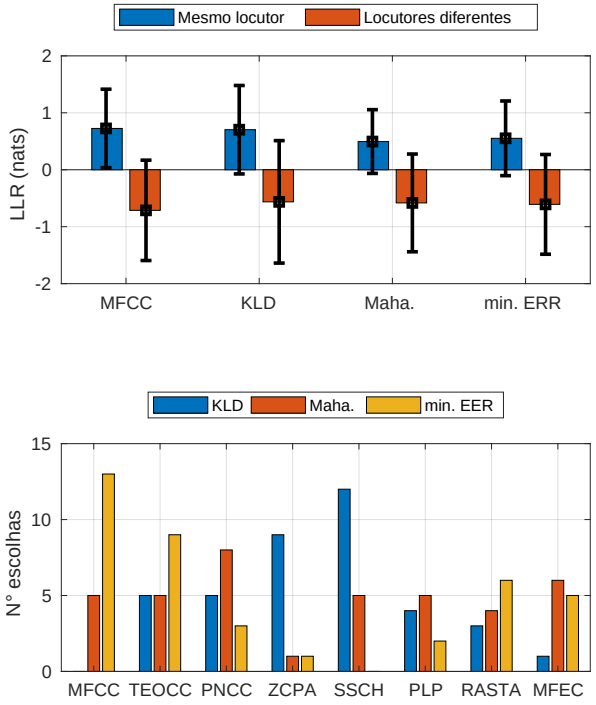

Figura 19: Desempenho das estratégias de comparação e ocorrência dos descritores. No gráfico do topo as barras representam o $LLR(\vec{x}_Q)$ médio enquanto as barras verticais os limites para três desvios padrões das ocorrências do $LLR(\vec{x}_Q)$ em comparação entre o mesmo locutor (barras azuis) e de locutores diferentes (barra vermelha). No gráfico da base tem-se a distribuição de ocorrências de ada descritor para cada estratégia proposta.

uma maior distância, o fato de combinar obrigatoriamente pelo menos um descritor por dimensão não implica em um aumento monotônico da distância de Mahalanobis.

Uma alternativa que talvez seja mais eficiente é testar as combinações que resultassem sempre no crescimento da distância de Mahalanobis média entre os locutores. Esta alternativa também será estuda na continuidades dos trabalhos.

Por fim, a estratégia de mínima EER também sofreu da mesma falha que a estratégia baseada na distância de Mahalanobis. Como as taxas foram calculadas por dimensão e por descritor, a combinação das dimensões não implica na minimização da EER das dimensões combinadas.

Por fim, ressalta-se que estas estratégias foram testadas apenas em áudios que não foram contaminados por ruído (apenas com a codificação GSM), fato que explica os valores de EER abaixo de 1%.

5. CONCLUSÕES

O estudo das características fornece informações que permitem balizar os próximos experimentos. A partir destes estudos é possível elencar as seguintes conclusões básicas:

- na média, os áudios codificados pelo *codec* GSM 06.60 apresentam menor EER se os modelos GMM-UBM são oriundos de áudios também codificados pelo *codec* GSM 06.60;
- utilizando MFCC's e UBM-GMM processados pelo *codec* GSM 06.60, a taxa de mesmo erro esperada é menor que 5% a uma SNR até 17 dB;
- apesar da equivalência estatística entre algumas características e o MFCC (vide Fig. 6), este descritor apresentou menor EER médio inclusive diante estratégias de combinação.

Ficam como propostas de continuidade uma exploração mais detalhada das técnicas de combinação dos descritores acústicos, inclusive para cenários com contaminação por ruído.

AGRADECIMENTOS

AGRADECIMENTOS SUPRIMIDOS PARA REVISÃO DUPLO CEGA.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] E. Gold; P. French. International practices in forensic speaker comparison. *International Journal of Speech Language and the Law.* **18**:293–307 (2011).

[2] G.S. Morrison; F.H. Sahito; G. Jardine; D. Djokic; S. Clavet; S. Berghs; C.G. Dorny. Interpol survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* **263**:92–100 (2016).

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

[4] Y. Kinoshita; S. Ishihara; P. Rose. Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *International Journal of Speech, Language & the Law* **16**:91–111 (2009).

[5] G.S. Morrison; C. Zhang; P. Rose. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International* **208**:59–65 (2011).

[6] E. Enzinger; 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* **277**:30–40 (2017).

[7] R.R.d. Silva; J.P.C.L. da Costa; R.K. Miranda; G. Del Grado. Aplicação do valor de base da frequência

fundamental via estatística MVKD em comparação forense de locutor. *Revista Brasileira de Criminalística* **5**:30–38 (2016).

[8] D.A. Reynolds; T.F. Quatieri; R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing* **10**:19–41 (2000).

[9] S.B. Davis; P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in Speech Recognition*, 65–74. Elsevier (1990).

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

[11] D.S. Kim; S.Y. Lee; R.M. Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on Speech and Audio Processing* **7**:55–69 (1999).

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

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

[14] H. Hermansky; N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* **2**:578–589 (1994).

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

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

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

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

[19] J. Kacur; M. Varga; G. Rozinaj. ZCPA features for speech recognition. In *Telecommunications (BIHTEL), 2012 IX International Symposium on*, 1–4. IEEE (2012).

[20] R. Togneri; D. Pullella. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits And Systems Magazine* **11**:23–61 (2011).

[21] R.O. Duda; P.E. Hart; D.G. Stork. *Pattern Classification*. John Wiley & Sons (2012), 53,136.

[22] A. Follador Neto; A. Pinheiro Silva; H. Camille Yehia. Corpus cefala-1: Base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia. *Revista de Estudos da Linguagem* **27** (2019).

[23] G. ITU. Gsm full rate speech transcoding. gsm rec 06.10 (1991).

[24] I. Guyon; S. Gunn; M. Nikravesh; L.A. Zadeh. *Feature Extraction: Foundations and Applications*, volume 207. Springer (2008), 100.

[25] J. Gonzalez-Rodriguez; P. Rose; D. Ramos; D.T. Toledano; J. Ortega-Garcia. Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **15**:2104–2115 (2007).

[26] G. Casella; R. Berger. *Inferência Estatística - Tradução da 2ª edição norteamericana*. Centage Learning (2011), (p. 259).