AVALIAÇÃO DE METODOLOGIA FORENSE DE COMPARAÇÃO AUTOMÁTICA DE LOCUTORES EM VOZES SINTETIZADAS

Adelino Pinheiro Silva
http://lattes.cnpq.br/8373538496107754 – https://orcid.org/0000-0002-2796-4841
adelino.pinheiro@policiacivil.mg.gov.br
Polícia Civil de Minas Gerais, Belo Horizonte, MG, Brasil

Gerson Albuquerque Silva
http://lattes.cnpq.br/4649063081893795 - https://orcid.org/0009-0000-9376-164X
gerson.gas@policiacientifica.sp.gov.br
Superintendência da Polícia Técnico-Científica de São Paulo, SP, Brasil

Ronaldo Silva
http://lattes.cnpq.br/5889828570974736 - https://orcid.org/0000-0003-1263-4572
rrrrodrigues70@gmail.com
Departamento de Polícia Federal, DF, Brasil

Rafaello Virgilli
http://lattes.cnpq.br/0625389300835349 - https://orcid.org/0009-0002-5040-5869
rvirgilli@gmail.com
Superintendência da Polícia Técnico-Científica de Goiás, GO, Brasil

Avante
REVISTA ACADÊMICA DA POLÍCIA CIVIL DE MINAS GERAIS

RESUMO

A comunicação oral carrega informações identificadoras além da mensagem transmitida, permitindo o desenvolvimento de sistemas biométricos vocais e protocolos científicos para a comparação forense de locutores (CFL). Com a evolução da síntese de voz por inteligência artificial (IA), surgem preocupações sobre a segurança e a capacidade de detecção humana. Apesar do desempenho dos Sistemas de Reconhecimento Automático de Locutores (SRAL), estes ainda precisam evoluir para contornar as tecnologias de síntese por IA, especialmente no contexto do português brasileiro. Frente a esse desafio, o presente trabalho visa comparar o desempenho de SRAL aplicados com metodologias da CFL em vozes sintetizadas por IA questionando como os SRAL, utilizando ECAPA-TDNN implementado no SpeechBrain, reagem à comparação de vozes clonadas. A metodologia exploratória quantitativa utilizou o Corpus Forense do Português Brasileiro (CFPB) para calibração e o *corpus* CEFALA1 para experimento, empregando os serviços de clonagem ElevenLabs® e Coqui-TTS®. Os resultados mostraram que o *framework* apresentou desempenho ótimo em vozes naturais (precisão balanceada > 95%), mas demonstrou vulnerabilidades às vozes sintetizadas, com 67% das vozes clonadas classificadas como do mesmo locutor. Frente a este resultado, recomenda-se o desenvolvimento de protocolos específicos para análises forenses com suspeita de clonagem vocal.

Palavras-chave: Vozes sintetizadas; Comparação forense de locutor; Clonagem de voz; Sistema de reconhecimento automático de locutor; D*eepfake*.

EVALUATION OF FORENSIC METHODOLOGY FOR AUTOMATIC COMPARISON OF SPEAKERS IN SYNTHESIZED VOICES

ABSTRACT

Oral communication carries identifying information beyond the transmitted message, enabling the development of vocal biometric systems and scientific protocols for forensic speaker comparison (FSC). With the evolution of artificial intelligence voice synthesis, concerns arise about security and human detection capability. Despite the performance of Automatic Speaker Recognition Systems (ASRS), these still need to evolve to overcome AI synthesis technologies, especially in the Brazilian Portuguese context. This work aims to compare the performance of ASRS applied with FSC methodologies on AI-synthesized voices, questioning how ASRS using ECAPA-TDNN implemented in SpeechBrain reacts to cloned voice comparison. The quantitative exploratory methodology used the Brazilian Portuguese Forensic Corpus (CFPB) for calibration and the CEFALA-1 *corpus* for experimentation, employing the SpeechBrain ECAPA-TDNN model and ElevenLabs® and Coqui-TTS®

cloning services. Results showed that the framework presented optimal performance on natural voices (balanced accuracy > 95%), but vulnerabilities to synthesized voices, with all cloned voices classified as the same speaker. Given this result, the development of specific protocols for forensic analyses with suspected voice cloning is recommended.

Keywords: Synthesized voices; Forensic speaker comparison; Voice cloning; Automatic speaker recognition system; Deepfake

DOI: https://doi.org/10.70365/2764-0779.2025.164

Recebido em: 02/08/2025.
Aceito em: 21/10/2025.

1 INTRODUÇÃO

A comunicação oral é um dos métodos mais antigos de troca de informação presentes na humanidade e permite descrever tanto eventos concretos como conceitos abstratos. A capacidade de comunicação oral é plástica de forma que um mesmo locutor apresenta uma variabilidade própria de suas características e também pode voluntariamente alterar sua forma habitual de falar (Kreiman *et al.*, 2015). O sinal acústico que é produto da vocalização carrega, além da mensagem transmitida, informações do locutor, por exemplo, seu estado de saúde – e.g., rouco, gripado –, e sua fisiologia – e.g. cansado (Flanagan, 2008). Da mesma forma, variações na fonologia podem identificar tanto a região de origem – e.g., sotaque – quanto fatores sociais – e.g., profissão, classe social, etnia e idade (Labov, 1972). Dessa forma, além das variações entre os falantes de um idioma (extrafalante), existem também particularidades dentro da fala de um mesmo indivíduo (intrafalante) que permitem, entre outras coisas, inferir características sobre ele a partir de um registro oral (Virgilli, 2024).

O conhecimento da variabilidade intra e extrafalante fornece base para inferências sobre raridade vocal e tipicidade, permitindo a identificação de falantes por meio da voz (Kreiman & Sidtis, 2011). Em complementação, a evolução da computação digital viabilizou o desenvolvimento de sistemas biométricos que utilizam a voz como traço identificador. Paralelamente, o novo paradigma em ciências forenses estabeleceu protocolos científicos robustos para a atribuição de origem vocal (Saks & Koehler, 2008; Morrison, 2009). Nesse contexto, a autenticação biométrica e a análise forense criminal diferenciam-se quanto a: (i) objetivos primários (controle de acesso vs. evidência jurídica); (ii) protocolos de decisão (limiares automatizados vs. análise interpretativa); (iii) quadros de referência (probabilístico vs. de validação metodológica).

Na comparação forense de locutores (CFL), confrontam-se duas categorias de amostras: o material questionado, que é o vestígio acústico de origem desconhecida associado a fato típico penal, coletado sem controle de parâmetros de qualidade; e o material padrão, que corresponde às amostras obtidas mediante consentimento livre de suspeito, frequentemente não colaborativo devido aos riscos de associação criminal (Maher, 2009; Silva, 2020).

Os primeiros métodos sugeridos para a CFL, como de Kersta (1962), evoluíram de perfis auditivo instrumentais (Gfrörer, 2003) para sistemas baseados em redes neurais profundas (Sztahó & Fejes, 2023). O paradigma contemporâneo sugere a quantificação dos resultados via Razão de Verossimilhança (LR – *likelihood ratio*), calculada estatisticamente em bancos de dados representativos que capturem a variabilidade populacional (raridade e tipicidade fonética) e modelagens de características acústicas compatíveis com métricas de divergência baseadas em LR (Campbell *et al.*, 2009; Kabir *et al.*, 2021) e com confiabilidade mensurável. Tipicamente, a CFL confronta características acústicas extraídas dos áudios questionados e padrão, utilizando

a base de dados como fonte dos modelos estatísticos para inferência. A comparação gera uma pontuação de similaridade e o intervalo de confiança/credibilidade associados (Silva, 2020).

Em etapas analíticas da CFL, é possível utilizar sistemas de reconhecimento automático de locutores (SRAL) para obter uma pontuação derivada do processamento não linear via ECAPA-TDNN (*Emphasized Channel Attention, Propagation and Aggregation* aplicada a *Time-Delay Neural Networks*). Esse processamento mapeia amostras de durações variáveis em vetores *embeddings* de dimensionalidade fixa (Desplanques *et al.*, 2020) extraídos do espaço latente da TDNN (SNYDER *et al.*, 2018) e otimizados para minimizar distâncias intrafalantes e maximizar as extrafalantes. Uma implementação aberta é disponibilizada pelo *SpeechBrain*, treinada com a bases de dados VoxCeleb 1 e 2 (Ravanelli *et al.*, 2021; Nagrani *et al.*, 2017).

O estudo de Sztahó e Fejes (2023) reportaram Taxa de Erro Igual (EER – *equal error rate*) de 3,1% (geral) e 1,1% (segmentos de 10 segundos) com ECAPA-TDNN adaptada ao húngaro – idioma com inventário consonantal similar ao português, porém inventário vocálico mais extenso. Basu e colaboradores (2022, 2023) compararam sistemas TDNN com métodos auditivo-acústico-fonéticos, constatando superioridade dos primeiros em amostras de aproximadamente 15 segundos (Custo Logarítmico da Razão de Verossimilhança – C_{LLR}: 0,42 vs. 0,51 em ouvintes humanos). Resultados equivalentes (C_{LLR} = 0,52) foram replicados por Silva *et al.* (2023) com gêmeos monozigóticos usando segmentos de 10 segundos. Entretanto, apesar do desempenho, os SRAL ainda precisam evoluir para contornar as tecnologias de síntese de voz por inteligência artificial (IA). A síntese de voz (ou clonagem) por IA tem suscitado preocupações em segurança da informação e capacidade humana de detecção. Galyashina e Nikishin (2021) reportam que a qualidade da síntese evoluiu drasticamente de 20 minutos de amostra necessários em 2016 para apenas 3,7 segundos em 2018, com casos documentados de fraudes financeiras superiores a 220 mil euros.

A mimetização biométrica pela voz iniciou-se com modelos ocultos de Markov (HMM - *Hidden Markov Models*) até a utilização das redes adversariais generativas (GANs – *generative adversarial networks*) conforme documentado por Tolosana e colaboradores (2020). Essas tecnologias exploram vulnerabilidades fundamentais na percepção humana. O estudo de Barrington e colaboradores (2025) revelou que, de forma perceptual, participantes identificaram como da mesma pessoa aproximadamente 80% de vozes clonadas e uma precisão de 67,4% e 60,8%, respectivamente, na identificação de vozes reais e sintéticas.

Dos trabalhos recentes que buscam analisar a robustez das metodologias de CFL para detecção de voz sintetizada por IA, destacam-se Barrington e colaboradores (2023), que, em uma base de 1.472 vozes reais e sintetizadas (50-50) do inglês, utilizaram três abordagens para comparação: (1) estatísticas de

amplitude e pausas com EER entre 24,9 e 48,5%; (2) estatísticas espectrais extraídas por *software* de processamento de áudio com EER entre 0,5% e 19,7%; (3) por e vetores *embeddings* utilizando o modelo TitaNet (KOLUGURI *et al.*, 2022) com EER entre 0,0 e 4,4%. O estudo de Kudera de colaboradores (2024) utilizou SRAL para avaliar a clonagem de voz por IA a partir de aproximadamente 2 minutos de fala com 4 locutores. Os resultados mostram que todas as vozes clonadas foram classificadas como do mesmo locutor e indicou a necessidade de desenvolvimento de protocolos específicos para detecção prévia dessas condições antes da aplicação de SRAL na CFL. Adicionalmente, nota-se uma lacuna na avaliação de vozes clonadas no recorte do português brasileiro.

Isso posto, o presente trabalho tem como objetivo comparar o desempenho de SRAL aplicados com metodologias da CFL em vozes sintetizas por IA. Mais especificamente, o trabalho busca criar uma base inicial de vozes sintetizadas; aplicar etapas automáticas da CFL em vozes reais e sintetizadas comparando o desempenho, com atenção nas comparações entre o mesmo locutor com recorte pela codificação, OPUS[1] ou PCM (*Pulse-code modulation*), e por sexo. A metodologia utilizada foi exploratória quantitativa. Foi planejado um experimento de comparação utilizando o *Corpus* Forense do Português Brasileiro (CFPB) para a calibração e as vozes dos *corpora* CEFALA1 (Neto, Silva, Yenia; 2019) para o experimento utilizando o modelo ECAPA-TDNN presente no *framework* do *SpeechBrain* em *python* e os serviços de mimetização de voz da ElevenLabs® e Coqui-TTS® – doravante referenciados, respectivamente, como ElevenLabs e Coqui-TTS.

A principal questão experimental é como os SRAL, utilizando ECAPA-TDNN implementado no *SpeechBrain* e calibrado pelo CFPB, reagem à comparação de vozes sintetizadas/clonadas e quais os cuidados para o uso na CFL. O método utilizado encaixa-se na aplicação forense, pois atende os requisitos do atual paradigma ao permitir expressar o resultado da comparação como uma razão de verossimilhança (LR - *likelihood ratio*), avaliando a correspondência e seu alcance, por meio de uma base de dados representativa (Saks & Koehler, 2008; Morrison, 2009).

O trabalho apresenta como limitação o recorte do *corpus* CEFALA1 – predominante em falantes do dialeto mineiro com 104 unidades experimentais contemporâneas; a limitação de banda entre 20 e 20 kHz dos microfones; a predominância de vozes masculinas (80%) no CFPB; a limitação de sintetização, utilizando amostras entre 1 e 3,6 minutos de fala; e a diferença entre os áudios do experimento (gravação controlada) e os áudios questionados, tipicamente provenientes de interceptação telefônica ou serviços mensagens com degradação por ruído e codificação.

O presente trabalho divide-se em mais quatro seções. Na próxima,

[1]OPUS é um *codec* de áudio com alta versatilidade aberto e livre de royalties muito utilizado em comunicação. Optou-se por testar a sensibilidade ao codec OPUS por este ser o mais utilizado em aplicativos de mensagem instantânea de voz como o *WhatsApp* da Meta®.

detalha-se o desenho experimental, apresentando os bancos de dados, o *framework Speechbrain,* os métodos utilizados para a geração da amostra experimental e os métodos utilizados na comparação das amostras de voz. Na terceira seção, são apresentados os resultados do experimento, e, na quarta seção, a discussão. Na última seção, enumeram-se as principais conclusões e as propostas de continuidade.

2 MATERIAIS E MÉTODOS

O presente experimento foi planejado visando verificar o desempenho do método de comparação de locutores via ECAPA-TDNN na comparação por distância cosseno implementada pelo *SpeechBrain*. No experimento, foram utilizadas duas bases de dados, o CFPB para calibração e o CEFALA1 para geração das vozes sintetizadas e testes. A implementação dos métodos foi realizada em *python* utilizando os pacotes de cálculo científico, redes neurais, aprendizado de máquina, e de manipulação e visualização de dados[2].

O CFPB consiste em um *corpus* com 257 locutores, sendo 206 masculinos, de diferentes regiões do Brasil. Para todos os locutores, tem-se uma seção de entrevista – aproximadamente 5 minutos – e leitura de sentenças – aproximadamente 1 minuto. Para 36 locutores, tem-se uma segunda seção não contemporânea com fala espontânea de aproximadamente 2 minutos. As gravações possuem uma homogeneidade em relação a duração e estilo de fala, gravadas por microfone Shure® modelo SM58, conectado a uma placa de áudio marca Edirol®, modelo UA25EX via Adobe Audition® 3.0 com o *codec* PCM. Do *corpus* CEFALA1, foram utilizadas amostras dos 104 locutores gravadas pelo microfone condensador da marca Brüel & Kjœr®, modelo 1065. No CEFALA1, cada locutor foi gravado em uma seção de fala espontânea, uma de leitura de texto com 153 palavras, e uma com 20 frases.

O experimento é dividido em duas etapas: a primeira, de calibração do *SpeechBrain,* e a segunda, de testes. Na etapa de calibração, as falas dos locutores do CFPB foram divididas em dois grupos, sendo o primeiro composto pelas 206 vozes masculinas com entrevista – divididas em trechos de 10 segundos para cálculo dos vetores *embeddings* –, e o segundo grupo composto pelo trecho de fala espontânea dos 36 locutores, contidos no primeiro grupo. Também foram extraídos os vetores de *embeddings* de trechos de 10 segundos.

Nos dois grupos do CFPB acima descritos, os locutores foram comparados, trecho a trecho, pela distância cosseno entre os vetores *embeddings*. A partir das distâncias, ajustou-se uma regressão logística de calibração entre as classes "mesmo locutor" e "locutores diferentes". A distribuição do logaritmo da razão de verossimilhança (LLR – *log-likelihood ratio*) das classes "mesmo locutor" e "locutores diferentes" no CFPB são apresentadas como as linhas contínuas, respectivamente, em azul e laranja nas Figuras 1 a 6.

[2]Respectivamente os pacotes *scipy*, *pytorch*, *scikit-learn*, *pandas* e *matplotlib*.

A calibração resultou em uma EER de 1,5% e C_{LLR} de 0,05 Np (nepers), com limiar de decisão do LLR em 0,08 Np.

O resultado da regressão logística indica a probabilidade condicional $p(H_1 | x)$ da distância x entre os vetores serem do "mesmo locutor" (H_1) e $p(H_0 | x) = 1 - p(H_1 | x)$ a probabilidade de ser da classe "locutores diferentes". A razão de verossimilhança (LR) para a classe "mesmo locutor" e seu logaritmo LLR são

$$LR = \frac{p(H_1|x)}{p(H_0|x)} = \frac{p(H_1|x)}{1-p(H_1|x)} \rightarrow LLR = log\left(\frac{p(H_1|x)}{1-p(H_1|x)}\right). \qquad (1)$$

O LLR é uma medida estatística que quantifica a evidência em favor de uma hipótese *versus* outra – e.g. hipótese de duas gravações de fala serem do "mesmo locutor" *versus* serem de "locutores diferentes". Nesse contexto, o custo é definido como

$$C_{LLR} = \frac{1}{2}\, log(\frac{1}{N_1}\sum \qquad log(1 + LR_1) + \frac{1}{N_0} log(1 + \frac{1}{LR_0})), \qquad (2)$$

onde N_1, N_0, LR_1 e LR_0 são, respectivamente, o número de comparações e as razões de verossimilhança entre "mesmo locutor" e "locutores diferentes". O C_{LLR} indica o quanto o método de classificação está longe de um sistema perfeito. Quanto menor o valor do C_{LLR}, melhor o sistema está em distinguir entre as duas classes. Um valor de C_{LLR} próximo a zero indica uma separação mais efetiva em termos de distinção entre as classes. Por outro lado, um valor próximo da unidade indica uma distinção mais próxima da decisão aleatória[3].

Para a etapa de testes de comparação de locutor, utilizou-se a amostra do *corpus* CEFALA1 dividida também em dois grupos. O primeiro composto apenas pelo trecho de fala espontânea de todos os locutores, e o segundo pelos trechos de leitura de texto e frases. A partir dessa divisão, foram geradas, para cada locutor, vozes sintetizadas com o conteúdo do texto e das frases utilizando como fonte de clonagem o trecho de fala espontânea. As vozes foram sintetizadas pelos serviços da ElevenLabs e da Coqui IA® (Coqui-TTS)[4]. Em seguida, a amostra de fala espontânea de cada locutor foi comparada com as amostras com texto e fases de todos os demais locutores nas versões naturais (gravadas no CEFALA1) e sintetizadas pelo ElevenLabs e Coqui-TTS em formato PCM e codificadas pelo *codec* OPUS em um total de seis cenários de comparação.

No processo de clonagem/síntese, as gravações do CEFALA1 foram subamostradas para 22 kHz para serem sintetizadas. Essa é a taxa de

[3]O uso do valor logaritmo justifica-se para comprimir a escala de análise, uma vez que os valores da razão de verossimilhança (LR) podem variar desde muito elevados (e.g. 10^6) até próximos de zero (e.g. 10^{-6}). No presente trabalho, utilizou-se o logaritmo natural tanto no cálculo do LLR quanto no C_{LLR}, resultando em unidades de medida em nepers (Np).

[4]Para o ElevenLabs, utilizou-se a função de clonagem de voz instantânea que utiliza trechos inferiores a 30 minutos, para a Coqui IA utilizou-se a função de conversão de texto para voz (TTS – *text to speech*).

amostragem recomendada nas configurações dos modelos[5]. Para serem comparados, todos os registros de áudio foram subamostrados a 16 kHz, que é a frequência utilizada pelo *SpeechBrain*.

Nas comparações dentro do CEFALA1, dividiram-se os áudios comparados em trechos de 10 segundos. Dos trechos foram extraídos os vetores *embeddings* e calculada a mediana da distância cosseno entre os vetores referentes às gravações de cada locutor. O valor da mediana da distância foi avaliado pela função de regressão logística ajustada pelo CFPB.

Da avaliação, foi calculado o LLR dividindo-se as comparações em duas classes: as realizadas entre o "mesmo locutor" e as realizadas entre "locutores diferentes". Em seguida, utilizando a análise de variância, foram analisadas as diferenças nos valores do LLR apenas entre as comparações na classe "mesmo locutor".

3 RESULTADOS

O trecho de fala espontânea foi comparado com o trecho com leitura de texto e frases da base CEFALA1 e sintetizadas pelo ElevenLabs e Coqui-TTS com a codificação original, PCM, e com a codificação OPUS. Na comparação, foram definidas as duas classes/condições fundamentais (*ground-truth*) de "mesmo locutor" e "locutores diferentes". No resultado inferencial, optou-se por definir um intervalo de incerteza no valor de LLR de ± 2,94 Np em torno do limiar de decisão de 0,08 Np. Esse valor equivale ao logaritmo da razão de verossimilhança equivalente a uma razão de 0,95/0,05 (vide Equação 1), i.e., a probabilidade empírica de uma das classes é de pelo menos 95%. Optou-se por inserir um intervalo de alcance, conforme sugerido por Saks e Koehler (2008), para tornar a decisão mais suave a favor do *in dubio pro reo*.

Nessas condições, ao comparar dois locutores, tem-se que o valor previsto pela inferência pode ser de "mesmo locutor", para um valor de LLR acima do limiar superior de 3,03 Np, e de "locutores diferentes", para um LLR inferior ao limiar de - 2,86 Np ou cair no intervalo de - 2,86 < LLR < 3,03. Na Tabela 1, são apresentadas as matrizes de confusão expandidas para os resultados. Em cada grupamento da Tabela 1, há os diferentes cenários de comparações, desde a realizada entre as vozes naturais até as sintetizadas codificadas por OPUS. Em cada grupo de resultado, as linhas indicam as classes fundamentais das comparações entre "mesmo locutor" e "locutores diferentes". Nas colunas, são indicadas a inferência. Destacam-se, em negrito, os valores e as taxas percentuais de falsos positivos, falsos negativos, precisão balanceada[6] e o valor total de inferências no intervalo.

Tabela 1 – Matriz de confusão expandida indicando os resultados dos cenários das comparações.

[5]Foi utilizado o modelo eleven_multilingual_v2 da ElevenLabs® e XTTSv2 da Coqui-TTS®.

[6]A precisão balanceada é a média entre as taxas de verdadeiro positivo e de verdadeiro negativo.

Para cada cenário, são indicadas as classes fundamentais nas linhas e nas colunas as inferências. Destacam-se, em negrito, os valores e as taxas de falsos positivos e negativos, precisão balanceada e o valor total de inferências no intervalo.

		Comparação entre vozes naturais - previsto				
		mesmo locutor	locutor diferente	intervalo	total	precisão
condição	mesmo locutor	104 (100,0%)	**0 (0,0%)**	0	104 (1,0%)	**10127 (96,8%)**
	locutor diferente	**33 (0,3%)**	10023 (93,6%)	656	10712 (99,0%)	C_{LLR} (Np)
	total	137	10023	**656**	10816	0,03
		Comparação entre voz natural e Coqui-TTS - previsto				
		mesmo locutor	locutor diferente	intervalo	total	precisão
condição	mesmo locutor	34 (32,7%)	**8 (7,7%)**	62	104 (1,0%)	**10350 (64,5%)**
	locutor diferente	**7 (0,1%)**	10316 (96,3%)	389	10712 (99,0%)	C_{LLR} (Np)
	total	41	10324	**451**	10816	0,54
		Comparação entre voz natural e ElevenLabs - previsto				
		mesmo locutor	locutor diferente	intervalo	total	precisão
condição	mesmo locutor	101 (97,1%)	**0 (0,0%)**	3 (2,9%)	104 (1,0%)	**10205 (95,7%)**
	locutor diferente	**19 (0,2%)**	10104 (94,3%)	589 (5,5%)	10712 (99,0%)	C_{LLR} (Np)
	total	120	10104	**592 (5,5%)**	10816	0,04
		Comparação entre vozes naturais codificadas por OPUS - previsto				
		mesmo locutor	locutor diferente	intervalo	total	precisão
condição	mesmo locutor	104 (100,0%)	**0 (0,0%)**	0 (0,0%)	104 (1,0%)	**10006 (96,2%)**
	locutor diferente	**39 (0,4%)**	9902 (92,4%)	771 (7,2%)	10712 (99,0%)	C_{LLR} (Np)
	total	143	9902	**771 (7,1%)**	10816	0,04
		Voz natural vs. Coqui-TTS codificadas por OPUS - previsto				
		mesmo locutor	locutor diferente	intervalo	total	precisão
condição	mesmo locutor	42 (40,4%)	**4 (3,8%)**	58 (55,8%)	104 (1,0%)	**10204 (67,5%)**
	locutor diferente	**12 (0,1%)**	10162 (94,9%)	538 (5,0%)	10712 (99,0%)	C_{LLR} (Np)
	total	54	10166	**596 (5,5%)**	10816	0,39
		Voz natural vs. ElevenLabs codificadas por OPUS - previsto				
		mesmo locutor	locutor diferente	intervalo	total	precisão
condição	mesmo locutor	101 (97,1%)	**0 (0,0%)**	3 (2,9%)	104 (1,0%)	**10112 (95,3%)**
	locutor diferente	**23 (0,2%)**	10011 (93,5%)	678 (6,3%)	10712 (99,0%)	C_{LLR} (Np)
	total	124	10011	**681 (6,3%)**	10816	0,03

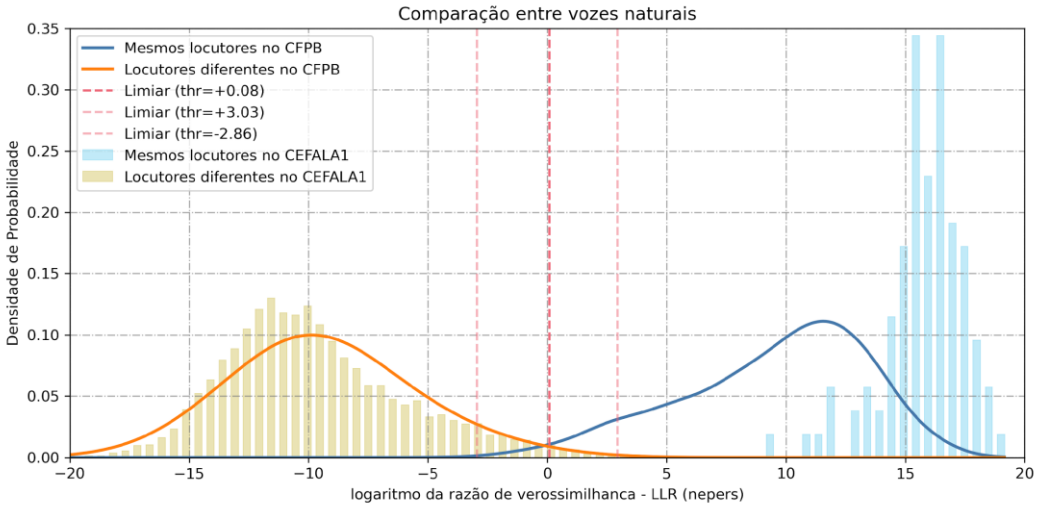
Fonte: Elaborado pelos autores.

Analisando a matriz de confusão, nota-se que a prevalência da classe "mesmo locutor" é de 1%. Essa prevalência é um desafio para a inferência, uma

vez que se busca separar uma quantidade muito pequena de resultados dentro do conjunto. Devido a esse fato, utilizou-se a métrica da EER para calibração visando equilibrar as taxas de falsos positivos e de falsos negativos. Em concordância, optou-se também por utilizar a métrica da precisão balanceada e do C_{LLR}, que facilita a comparação entre os cenários, pois também equilibra as taxas de falsos positivos e negativos.

Nas Figuras 1 a 6, são apresentadas a densidade de probabilidade dos valores de LLR obtidos em cada grupo de comparações. Em todas as figuras, têm-se as linhas contínuas azul e laranja que indicam a densidade de probabilidade empírica do LLR, respectivamente, para as classes de "mesmo locutor" e "locutores diferentes" obtidas pela calibração do *SpeechBrain* com o CFPB. As linhas pontilhadas verticais indicam o limiar de decisão de calibração em LLR = 0,08 Np e os limites superior e inferior da inferência no intervalo de incerteza entre - 2,86 e 3,03 Np. As barras horizontais royal e areia indicam a densidade de probabilidade empírica do LLR, respectivamente, para as classes de "mesmo locutor" e "locutores diferentes", obtidas na comparação do CEFALA1.

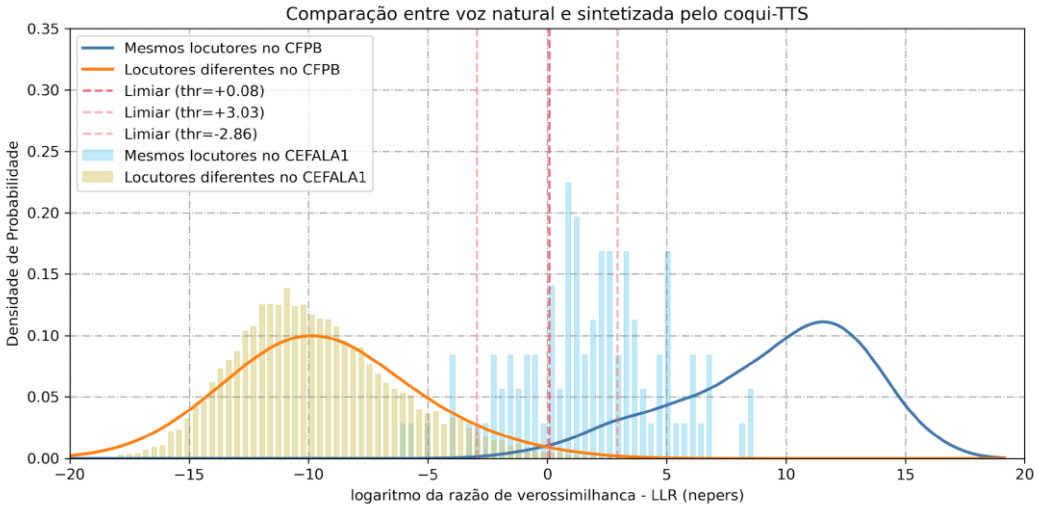
Figura 1 – Distribuição do LLR para as comparações realizadas entre os locutores com as vozes naturais. As linhas contínuas representam a densidade de probabilidade empírica calibrada pelo CFPB. As barras representam o resultado das comparações no CEFALA1. As linhas pontilhadas verticais representam o limiar de decisão e os limites superior e inferior do intervalo de transição.



Fonte: Elaborado pelos autores.

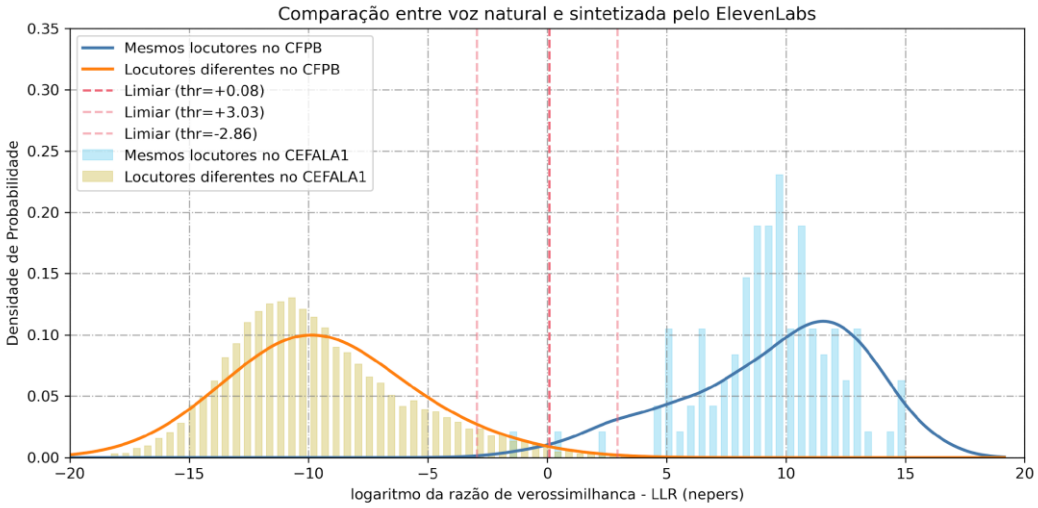
Comparando primeiramente os cenários das Figuras 1, 2 e 3, nos quais os áudios estavam em sua codificação original (PCM), nota-se o melhor desempenho das vozes naturais com maior precisão balanceada (96,8%), menor C_{LLR} (0,03 Np), menor taxa do falso negativo (0%), porém a maior taxa do falso negativo (0,3%) e o maior número de comparações dentro do intervalo (656). Nota-se, na Figura 1, que a média do LLR (15,5 Np) para as comparações entre o mesmo locutor é muito superior ao limiar de 3,03 Np.

Figura 2 – Distribuição do LLR para as comparações realizadas entre os locutores com as vozes gravadas e as sintetizadas pelo Coqui-TTS nos mesmos padrões da Figura 1.



Fonte: Elaborado pelos autores.

Figura 3 – Distribuição do LLR para as comparações realizadas entre os locutores com as vozes gravadas e as sintetizadas pelo ElevenLabs nos mesmos padrões da Figura 1.

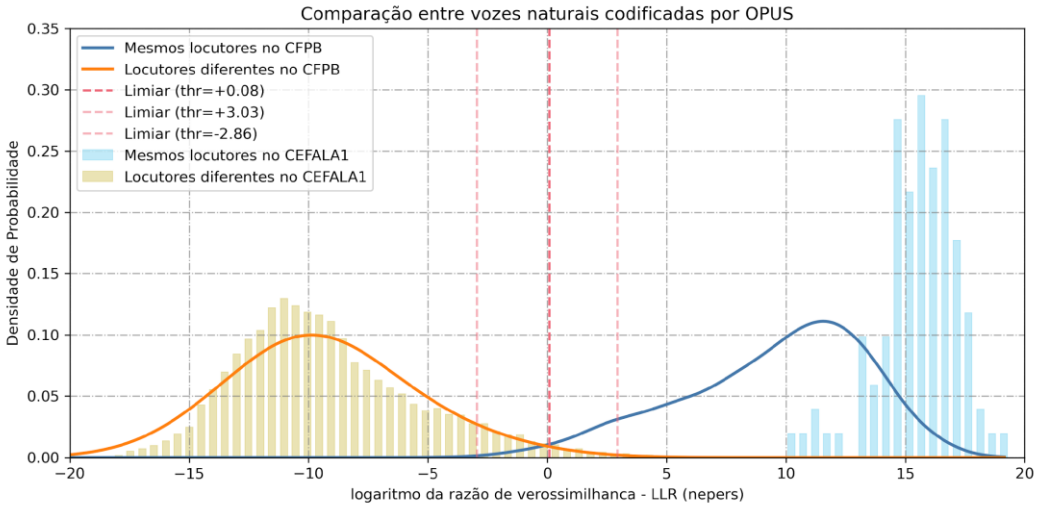


Fonte: Elaborado pelos autores.

Entretanto, nota-se que a comparação entre as vozes naturais e as sintetizadas pelo ElevenLabs atingiu um desempenho semelhante, com precisão balanceada de 95,7% e C_{LLR} de 0,04, uma taxa de falso positivo de 0,2% e um menor número de ocorrências no intervalo (592). Na Figura 3, nota-se que a média do LLR (9,1 Np) também é muito superior ao limiar de 3,03 Np. A comparação das vozes sintetizadas pelo Coqui-TTS apresentou o pior desempenho em relação à precisão balanceada (64,5%), porém apresentou a menor taxa de falso positivo (0,1%), o menor valor de ocorrências no intervalo (451), apesar de a média do LLR para o mesmo locutor (1,7 Np) cair dentro do intervalo (vide Figura 2).

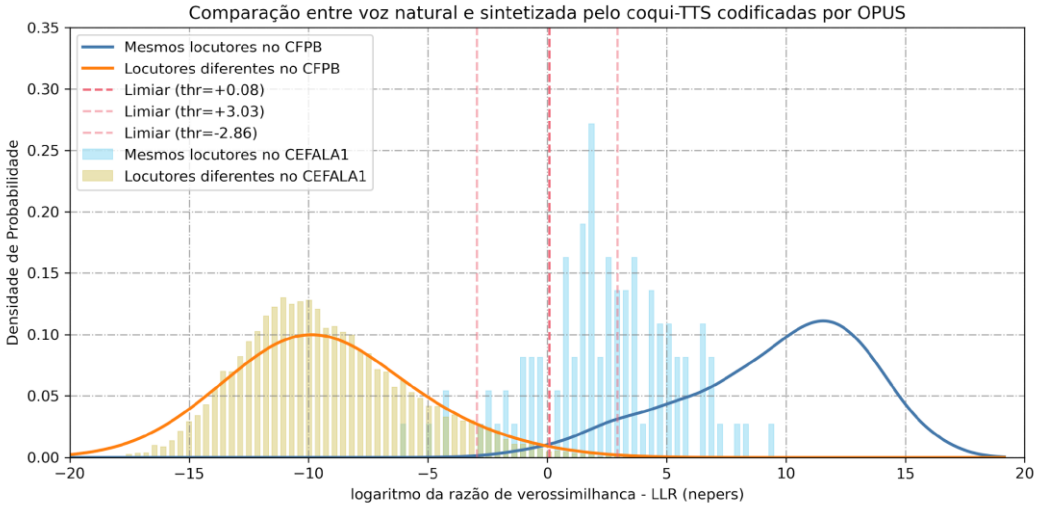
Para os cenários em que os áudios foram codificados pelo *codec* OPUS, indicados nas Figuras 4, 5 e 6, a comparação entre as vozes naturais também apresentou maior precisão balanceada (96,2%), menor taxa de falso negativo (0%), porém a maior taxa de falso negativo (0,4%), o maior número de comparações dentro do intervalo (771) e o C_{LLR} intermediário (0,04 Np). A média do LLR (15,3 Np) para as comparações entre o mesmo locutor ficou um pouco inferior comparando com a codificação PCM, porém ainda é muito superior ao limiar de 3,03 Np.

Figura 4 – Distribuição do LLR para as comparações realizadas entre os locutores com as vozes gravadas após a codificação OPUS nos mesmos padrões da Figura 1.



Fonte: Elaborado pelos autores.

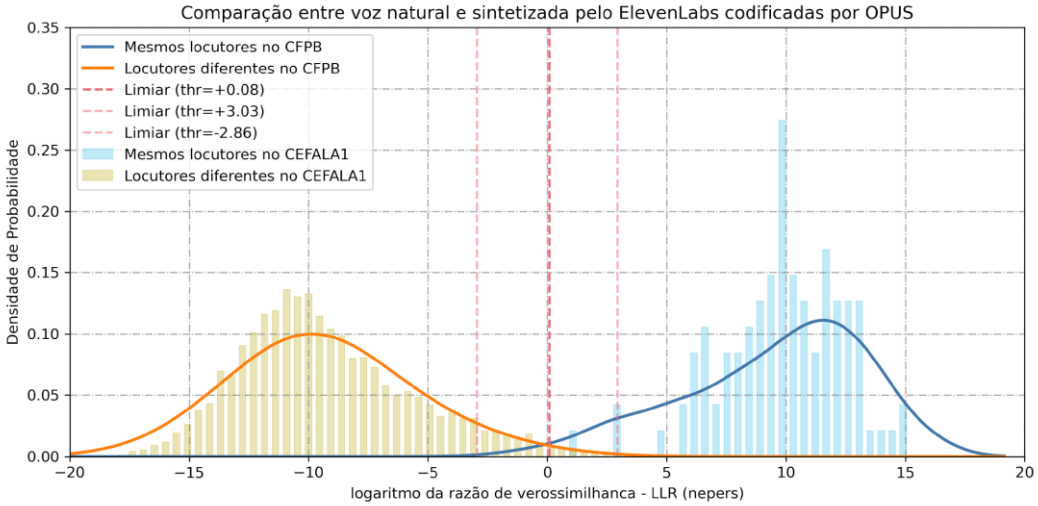
Figura 5 – Distribuição do LLR para as comparações realizadas entre os locutores com as vozes gravadas e as sintetizadas pelo Coqui-TTS após a codificação OPUS nos mesmos padrões da Figura 1.



Fonte: Elaborado pelos autores.

Com a codificação OPUS, a comparação entre as vozes naturais e as sintetizadas pelo ElevenLabs obteve o menor C_{LLR} (0,03) e uma taxa de falso positivo de 0,2%. Na Figura 6, nota-se que a média do LLR (9,7 Np) elevou-se em relação à codificação PCM e também é muito superior ao limiar de 3,03 Np. Na comparação das vozes sintetizadas pelo Coqui-TTS, notou-se uma melhoria discreta na precisão balanceada (67,5%), a menor taxa de falso positivo (0,1%), o menor valor de ocorrências no intervalo (596), C_{LLR} de 0,39 Np, e a média do LLR para o mesmo locutor (2,3 Np) foi mais elevada, porém ainda dentro do intervalo (vide Figura 5).

Figura 6 – Distribuição do LLR para as comparações realizadas entre os locutores com as vozes gravadas e as sintetizadas pelo ElevenLabs após a codificação OPUS nos mesmos padrões da Figura 1.



Fonte: Elaborado pelos autores.

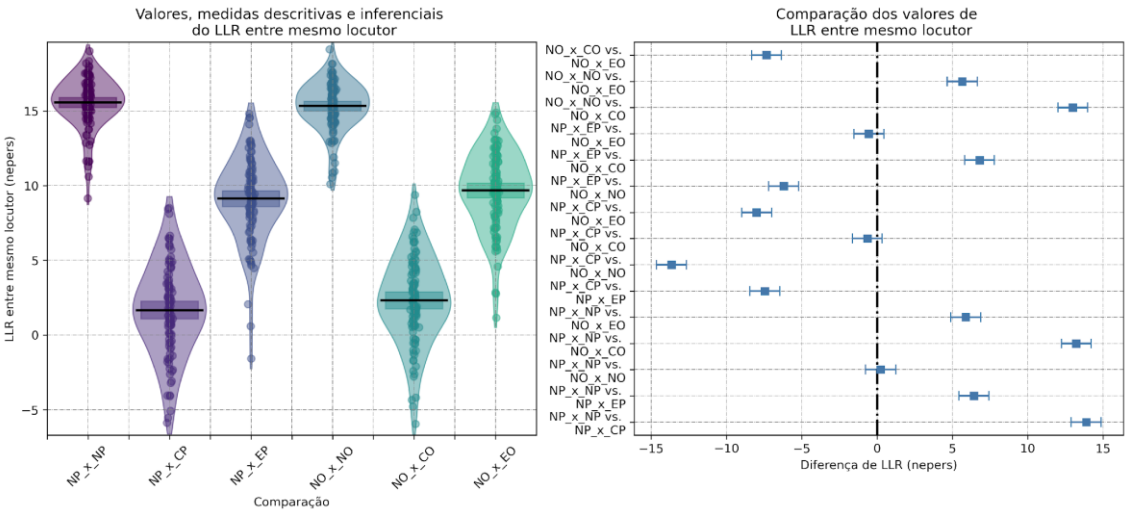
Em seguida, foram comparados os valores de LLR obtidos na comparação na classe de "mesmo locutor". Na Figura 7, o gráfico à esquerda apresenta o gráfico RDI (*Raw data, Description and Inference*) com os valores de LLR da mesma classe com recorte pelos cenários de comparação. No gráfico RDI, em cada coluna, os pontos são os valores individuais de LLR, as curvas laterais indicam a distribuição de probabilidade empírica, e a linha preta horizontal mostra a média amostral com o intervalo de confiança da média representado pelo retângulo (Phillips, 2017). As legendas no eixo horizontal indicam os cenários da forma:

- (NP_x_NP): Gravação natural com *codec* PCM *versus* a gravação natural com *codec* PCM;
- (NP_x_CP): Gravação natural com *codec* PCM *versus* a sintetizada pelo Coqui-TTS com *codec* PCM;
- (NP_x_EP): Gravação natural com *codec* PCM *versus* a sintetizada pelo ElevenLabs com *codec* PCM;

- (NO_x_NP): Gravação natural com *codec* opus *versus* a gravação natural com *codec* opus;
- (NO_x_CP): Gravação natural com *codec* opus *versus* a sintetizada pelo Coqui-TTS com *codec* opus; e
- (NO_x_EO): Gravação natural com *codec* opus *versus* a sintetizada pelo ElevenLabs com *codec* opus.

Apesar de algumas amostras se sobreporem, na comparação, observa-se uma diferença de desempenho em relação às vozes naturais sempre com LLR médio maior que o obtido pelas vozes sintetizadas.

Figura 7 – Distribuição do LLR nas comparações realizadas com o mesmo locutor e suas vozes sintetizadas. À esquerda, tem-se o gráfico RDI onde, em cada coluna, os pontos são os valores individuais de LLR, as curvas laterais indicam a distribuição de probabilidade empírica, e a linha preta horizontal a mostra média amostral. À direita, observa-se o resultado da análise de variância com intervalo calculado pelo teste de Tukey.



Fonte: Elaborado pelos autores.

Legenda: (NP_x_NP): Gravação natural com *codec* PCM *versus* a gravação natural com *codec* PCM. (NP_x_CP): Gravação natural com *codec* PCM *versus* a sintetizada pelo Coqui-TTS com *codec* PCM. (NP_x_EP): Gravação natural com *codec* PCM *versus* a sintetizada pelo ElevenLabs com *codec* PCM. (NO_x_NP): Gravação natural com *codec* opus *versus* a gravação natural com *codec* opus. (NO_x_CP): Gravação natural com *codec* opus *versus* a sintetizada pelo Coqui-TTS com *codec* opus. (NO_x_EO): Gravação natural com *codec* opus *versus* a sintetizada pelo ElevenLabs com *codec* opus.

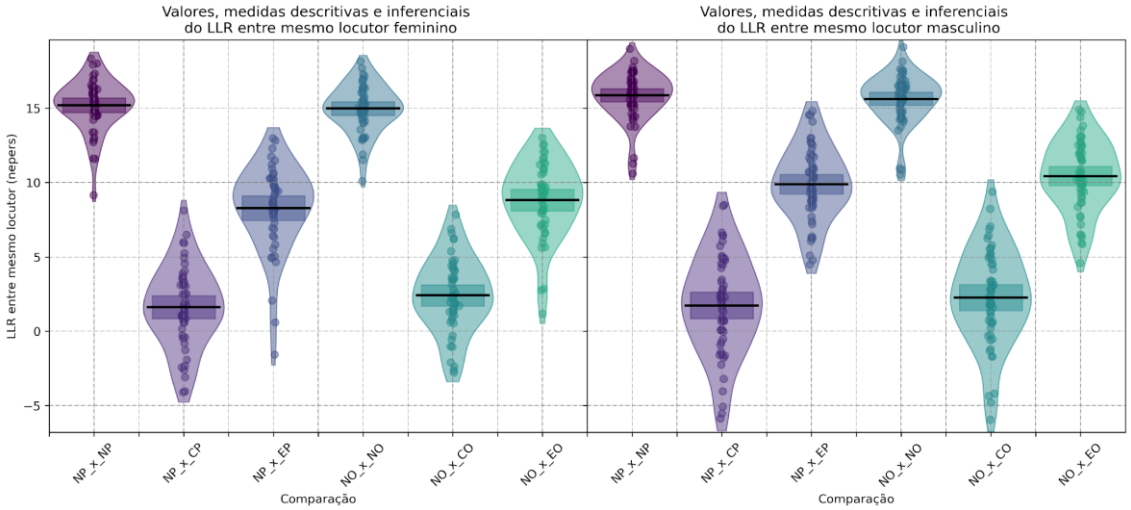
No gráfico à direita da Figura 7, observa-se, nos quadrados, a diferença média entre os LLR obtidos em cada cenário. As linhas horizontais indicam o intervalo de confiança da diferença entre as médias calculado pelo teste de Tukey. A linha vertical pontilhada indica o valor de diferença igual a zero. A análise de variância entre os diferentes cenários apresentou diferença significativa entre quase todos os cenários, com exceção nos cenários em que

a diferença é a codificação.

Na imagem, nota-se que todas as médias estão distantes da origem, indicando diferença significativa, exceto nos cenários em que a única diferença é a codificação (codec) PCM ou OPUS. Esse resultado indica que o conjunto de calibração do *SpeechBrain* com o CFPB foi pouco sensível à diferença de codificação nas vozes do CEFALA1.

No recorte pelo sexo do locutor, masculino ou feminino, notou-se que, em alguns cenários, a diferença no valor médio de LLR foi significativa. A Figura 8 apresenta, à esquerda, o gráfico RDI com os valores de LLR das comparações entre mesmo falante do sexo feminino e, à direita, o análogo para falantes do sexo masculino. Na imagem, nota-se que os dados dos pares de mesmo cenário se sobrepõem com pouca diferença na variação pelo *codec*.

Figura 8 - Distribuição do LLR nas comparações realizadas com o mesmo locutor e suas vozes sintetizadas com recorte referente ao sexo dos locutores. À esquerda, o gráfico RDI para locutores femininos e, à direita, para locutores masculinos.



Fonte: Elaborado pelos autores.

Legenda: (NP_x_NP): Gravação natural com *codec* PCM *versus* a gravação natural com *codec* PCM. (NP_x_CP): Gravação natural com *codec* PCM *versus* a sintetizada pelo Coqui-TTS com *codec* PCM. (NP_x_EP): Gravação natural com *codec* PCM *versus* a sintetizada pelo ElevenLabs com *codec* PCM. (NO_x_NP): Gravação natural com *codec* opus *versus* a gravação natural com *codec* opus. (NO_x_CP): Gravação natural com *codec* opus *versus* a sintetizada pelo Coqui-TTS com *codec* opus. (NO_x_EO): Gravação natural com *codec* opus *versus* a sintetizada pelo ElevenLabs com *codec* opus.

O resultado do teste sobre a diferença entre as médias do valor do LLR no recorte por sexo é apresentado na Tabela 2, que indica o valor da diferença entre as médias, o intervalo de confiança para um nível de confiança de 95% e o valor-*p*. Nota-se uma diferença significativa entre os sexos nas vozes sintetizadas pelo ElevenLabs e uma diferença marginalmente significativa entre as médias nas vozes naturais.

A diferença medida, apesar de significativa, não pode ser perfeitamente

isolada, pois existem efeitos tanto da base de dados de treinamento, Voxceleb, quanto do CFPB, em que se utilizaram apenas falantes masculinos. Não foi encontrada correlação significativa entre o tempo de áudio disponível para síntese e o valor do LLR obtido na comparação, sendo o valor máximo de correlação inferior a 0,25.

Tabela 2 – Resultado do teste-*t* da diferença entre as médias do LLR obtido na comparação entre "mesmo locutor" no recorte por sexo. Nas colunas, tem-se o valor da diferença entre as médias o intervalo de confiança para um nível de confiança de 95% e o valor-*p*.

	Diferença entre médias (Np)	intervalo de confiança (Np)	valor-*p*
vozes naturais (NP_x_NP)	0,67	[0,02; 1,31]	**0,04**
voz natural vs. Coqui-TTS (NP_x_CP)	0,12	[-1,07; 1,31]	0,84
voz natural vs. ElevenLabs (NP_x_EP)	1,6	[0,57; 2,65]	**< 0,01**
vozes naturais vs. naturais OPUS (NO_x_NO)	0,66	[0,02; 1,29]	**0,04**
voz natural vs. Coqui-TTS OPUS (NO_x_CO)	-0,15	[-1,28; 0,98]	0,79
voz natural vs. ElevenLabs OPUS (NO_x_EO)	1,61	[0,65; 2,58]	**< 0,01**

Fonte: Elaborado pelos autores.

4 DISCUSSÃO

O processo de síntese de voz por IA, do texto para a fala, tem apresentado uma série de aplicações comerciais, como, por exemplo, o acesso de obras literárias a deficientes visuais. Em adição, o processo de mimetização (ou clonagem) de uma voz conhecida possibilita uma série de aplicações em áudio, desde o entretenimento à infração da lei penal, possibilitando uma falsidade ideológica convincente no meio digital (BBC, 2024). Nesse ponto, vozes clonadas por IA constituem entidades que existem apenas como instanciações digitais sem referente físico original (Floridi, 2018) e emulam duplos funcionais que são difíceis de serem distinguidos por humanos (Barrington *et al.*, 2025), mas ontologicamente inautênticos.

Entretanto, como mostraram os resultados do experimento, o *framework* que combina o *SpeechBrain* calibrado pelo CFPB aplicado a vozes naturais mostrou um desempenho com precisão balanceada acima de 95% e C_{LLR} inferior a 0,05, que pode ser classificado como ótimo, segundo Morrison (2021). Esse resultado corrobora com o paradigma contemporâneo da CFL, estabelecido entre 2020-2025, que determina uma transição do método espectrográfico tradicional para *frameworks* Bayesianos. Ferragne e colaboradores (2024) apontam que a adoção de SARL pelos laboratórios forenses cresceu de 17% para 40% entre 2011-2019.

Os resultados do experimento – em concordância com Barrington *et al.* (2023) e Kudera *et al.* (2024) – mostraram que, apesar de apresentar diferenças nos resultados, o *SpeechBrain* calibrado com CFPB apresenta vulnerabilidades. No experimento, a ocorrência de vozes clonadas classificadas como do

"mesmo locutor" indica a necessidade de desenvolvimento de protocolos específicos em suspeita de clonagem de voz – como o adotado por Silva *et al.* (2023) para gêmeos univitelinos. Em consonância, sugere-se a inserção da possibilidade de clonagem na etapa de avaliação de risco, conforme os passos sugeridos pela UK *Forensic Science Regulator* (Morrison, 2021).

Outros dois pontos preocupantes são a facilidade de obtenção de gravações públicas e o viés algorítmico. Embora a Lei Geral de Proteção de Dados Pessoais (LGPD) brasileira (Brasil, 2018) classifique dados biométricos como sensíveis e exija consentimento explícito para seu uso, a disseminação voluntária de gravações públicas em redes sociais – como fontes para a clonagem de voz – é uma lacuna na proteção deste dado biométrico. Por outro lado, a escassez de pesquisas que contemplem a realidade do português brasileiro permite que emerja o viés algorítmico inerente a bancos de dados como o VoxCeleb – base de treinamento do ECAPA-TDNN implementado no *SpeechBrain*. Tal viés, caracterizado pela sub-representação de dialetos regionais, compromete significativamente o desempenho dos SRAL (Müller *et al.*, 2022).

5 CONSIDERAÇÕES FINAIS

Conforme proposto, o experimento permitiu comparar o desempenho dos SRAL composto pelo *framework* que combina o *SpeechBrain* calibrado pelo CFPB com metodologias de CFL em vozes sintetizas por IA incluindo uma variação com a codificação OPUS. O trabalho também permitiu a expansão do *corpus* CEFALA1 agregando vozes sintetizadas.

O resultado mostrou a efetividade do *framework* que combina o *SpeechBrain* calibrado pelo CFPB aplicado a vozes naturais. Por outro lado, em resposta à questão experimental, ficou patente a necessidade de mais estudos no que tange à distinção de vozes reais e sintetizadas, no português brasileiro. O *framework* com desempenho de EER de 1,5%, C_{LLR} de 0,05 Np não foi capaz de distinguir as vozes sintetizadas. Em aspectos pragmáticos, sugere-se que análises forenses com suspeita de clonagem de voz utilize uma variação de conjuntos de amostras e um protocolo de calibração específico.

O trabalho apresentou como limitação o recorte dialetal do *corpus* CEFALA1 e a contemporaneidade das gravações, o tipo único de microfone, e a limitação de serviços de clonagem de voz disponíveis em português brasileiro. Como propostas de continuidade, sugerem-se a expansão do trabalho com análises de estatísticas espectrais, a avaliação de outros modelos com a TitaNet, a sensibilidade da comparação e a síntese na presença de ruído.

REFERÊNCIAS

BARRINGTON, Sarah *et al.* Single and multi-speaker cloned voice detection: From perceptual to learned features. In: **2023 IEEE International Workshop on Information Forensics and Security** (WIFS). IEEE, 2023. p. 1-6.

BARRINGTON, Sarah; COOPER, Emily A.; FARID, Hany. People are poorly equipped to detect AI-powered voice clones. **Scientific Reports**, v. 15, n. 1, p. 11004, 2025.

Basu, N., Bali, A. S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K. A., & Morrison, G. S. Speaker identification in courtroom contexts–part i: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. **Forensic science international,** Elsevier, v. 341, p. 111499, 2022.

Basu, N., Bali, A. S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K. A., & Morrison, G. S. Speaker identification in courtroom contexts–part ii: Investigation of bias in individual listeners' responses. **Forensic Science International**, Elsevier, p. 111768, 2023.

BBC. 'Eram meu rosto e minha voz, mas era golpe': como criminosos 'clonam pessoas' com inteligência artificial. **BBC News Brasil**, 28 fev. 2024. Disponível em: https://www.bbc.com/portuguese/articles/cd1jv45dq3go. Acesso em: 30 jul. 2025.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018.

CAMPBELL, J. P.; SHEN, W.; CAMPBELL, W. M.; SCHWARTZ, R., BONASTRE, J. F.; MATROUF D. Forensic speaker recognition. **IEEE Signal Processing Magazine**, v. 26, n. 2, p. 95-103, 2009.

DESPLANQUES, B.; THIENPONDT, J.; DEMUYNCK, K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: MENG, H.; XU, B.; ZHENG, T. F. (Ed.). **Interspeech 2020**. [S.l.]: ISCA, 2020. p. 3830–3834.

FERRAGNE, E. *et al.* The adoption of automatic speaker recognition in forensic contexts: A survey. **Forensic Science International**, v. 346, 111635, 2024.

FLANAGAN, J. L. **Speech analysis synthesis and perception**. Springer Science & Business Media, 2013.

FLORIDI, Luciano. Artificial Intelligence, Deepfakes and a Future of Ectypes. **Philosophy & Technology**, v. 31, n. 3, p. 317-321, 2018.

GALYASHINA, E.; NIKISHIN, V. AI generated fake audio as a new threat to information security: legal and forensic aspects. In: **Proceedings of the**

international scientific and practical conference on computer and information security, Yekaterinburg, Russia. 2021. p. 17-21.

GFRÖRER, S. G. Auditory-instrumental forensic speaker recognition. In: **INTERSPEECH**. 2003. p. 705-708.

KERSTA, L. G. Voiceprint identification, **Nature**, vol. 196, no. 4861, pp. 1253-1257, 1962.

KOLUGURI, Nithin Rao; PARK, Taejin; GINSBURG, Boris. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In: **ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing** (ICASSP). IEEE, 2022. p. 8102-8106.

KREIMAN, j.; SIDTIS, D. **Foundations of voice studies: an interdisciplinary approach to voice production and perception**. Malden, MA: Wiley-Blackwell, 2011.

KREIMAN, J.; PARK, S. J.; KEATING, P. A.; ALWAN, A. The relationship between acoustic and perceived intraspeaker variability in voice quality. In: **Sixteenth Annual Conference of the International Speech Communication Association**. 2015.

KUDERA, Jacek *et al.* Voice Cloning and Mismatch Conditions in Forensic Automatic Speaker Recognition. In: **International Conference on Speech and Computer**. Cham: Springer Nature Switzerland, 2024. p. 171-184.

LABOV, W. **Sociolinguistic patterns**. University of Pennsylvania press, 1973.

MAHER, R. C. Audio forensic examination. **IEEE Signal Processing Magazine**, v. 26, n. 2, p. 84-94, 2009.

MIRSKY, Yisroel; LEE, Wenke. The Creation and Detection of Deepfakes: A Survey. **ACM Computing Surveys**, v. 54, n. 1, p. 1-41, 2021.

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

MORRISON, G. S.; Enzinger, E.; Hughes, V.; Jessen, M.; Meuwly, D.; Neumann C.; Planting S.; Thompson, W. C.; Vloed D.; Ypma, R. J.F.; Zhang C.; Anonymous A.; Anonymous, B. Consensus on validation of forensic voice comparison. **Science & Justice**, v. 61, n. 3, p. 299-309, 2021.

MÜLLER, Nicolas *et al.* Does Audio Deepfake Detection Generalize? In: I**NTERSPEECH 2022**, 2022, Incheon. Proceedings... Incheon: ISCA, 2022. p. 2783-2787.

NAGRANI, Arsha; CHUNG, Joon Son; ZISSERMAN, Andrew. Voxceleb: a large-scale speaker identification dataset. **arXiv preprint arXiv:1706.08612**, 2017.

NETO, A. F.; SILVA, A. P.; YEHIA, H. C. Corpus CEFALA-1: base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia/corpus CEFALA-1: audiovisual database of speakers for biometric, phonetic and phonology studies. **Revista de Estudos da Linguagem**, v. 27, n. 1, p. 191-212, 2019.

Phillips, N. D. (2017). **Yarrr! The pirate's guide to R.** ***APS Observe****r*, *30*.

RAVANELLI, Mirco *et al.* SpeechBrain: A general-purpose speech toolkit. **arXiv preprint arXiv:2106.04624**, 2021.

SAKS, M. J.; KOEHLER, J. J. The individualization fallacy in forensic science evidence. **Vanderbilt Law Review**, v. 61, p. 199, 2008.

SILVA, A. P. **Intervalo de evidência e pareamento fuzzy utilizando relação sinal ruído aplicados à comparação forense de locutores**. 2020. Tese de Doutorado. Universidade de Federal de Minas Gerais.

SILVA, R. R. da; CAVALCANTI, J. C.; ERIKSSON, A. Avaliação de sistema de rec. de locutor em uma base de vozes de gêmeos idênticos. In: **Anais 4° Interforensics – ICMedia**. [S.l.: s.n.], 2023. p. 264.

SNYDER, D. *et al.* X-vectors: Robust DNN embeddings for speaker recognition. In: IEEE. 2018 **IEEE international conference on acoustics, speech and signal processing** (ICASSP). [S.l.], 2018. p. 5329–5333.

SZTAHÓ, D.; FEJES, A. Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. **Journal of forensic sciences**, v. 68, n. 3, p. 871-883, 2023.

TOLOSANA, R., VERA-RODRIGUEZ, R., FIERREZ, J., MORALES, A., ORTEGA-GARCIA, J. Deepfakes and beyond: A survey of face manipulation and fake detection. **Information Fusion**, v. 64, p. 131-148, 2020.

VIRGILLI, R., CANDIDO Jr., A., ROSA, A.S., OLIVEIRA, F.S., SOARES, A.d.S. Dual-Bandwidth Spectrogram Analysis for Speaker Verification. In: **Brazilian Conference on Intelligent Systems**. Cham: Springer Nature Switzerland, 2024. p. 340-351.