UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Júlio de Mesquita Filho - campus de São José do Rio Preto

LEONARDO MENDES DE SOUZA

APRIMORANDO A DETECÇÃO DE FALSIFICAÇÃO DE VOZ POR MEIO DE ANÁLISE EXTENSIVA DE REDUÇÃO DE CARACTERÍSTICAS MULTICEPSTRAIS

São José do Rio Preto

2025

Leonardo Mendes de Souza

APRIMORANDO A DETECÇÃO DE FALSIFICAÇÃO DE VOZ POR MEIO DE ANÁLISE EXTENSIVA DE REDUÇÃO DE CARACTERÍSTICAS MULTICEPSTRAIS

Tese, apresentada à Universidade Estadual Paulista (UNESP), Júlio de Mesquita Filho, São José do Rio Preto, para obtenção do título de Doutorem Ciência da Computação.

Área de Concentração: Computação Aplicada

Orientador: Profº Dr. Rodrigo Capobianco Guido

São José do Rio Preto
2025

S729a

Souza, Leonardo Mendes de

Aprimorando a Detecção de Falsificação de Voz por Meio de Análise Extensiva de Redução de Características Multicepstrais / Leonardo Mendes de Souza. -- São José do Rio Preto, 2025

71 f. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Rodrigo Capobianco Guido

1. Redução de Dimensionalidade. 2. Spoofing Detection. 3. Reconhecimento de Padrões. 4. Análise Cepstral. 5. Aprendizado de Máquina. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

A presente pesquisa propõe um framework para detecção de falsificação de voz com técnicas multicepstrais e redução de dimensionalidade. O estudo contribui para o avanço da segurança biométrica e da ciência de dados, fortalecendo a proteção digital, promovendo inovação em inteligência artificial e qualificando profissionais.

POTENTIAL IMPACT OF THIS RESEARCH

This research proposes a framework for voice spoofing detection using multicepstral techniques and dimensionality reduction. The study contributes to advances in biometric security and data science, strengthening digital protection, fostering innovation in artificial intelligence, and enhancing professional qualification.

LEONARDO MENDES DE SOUZA

APRIMORANDO A DETECÇÃO DE FALSIFICAÇÃO DE VOZ POR MEIO DE ANÁLISE EXTENSIVA DE REDUÇÃO DE CARACTERÍSTICAS MULTICEPSTRAIS

Tese apresentado(a) à Universidade Estadual Paulista (UNESP), Júlio de Mesquita Filho, São José do Rio Preto, para obtenção do título de Doutor em Ciência da Computação.

Área de Concentração: Computação Aplicada

Data de defesa: 26/09/2025

BANCA EXAMINADORA

Profº Dr. Rodrigo Capobianco Guido
UNESP – Campus de São José do Rio Preto

Profº Dr. Diego Renan Bruno
UNESP - Campus de São José do Rio Preto

Profº Dr. Carlos Magnus Carlson Filho
FATEC- Campus de São José do Rio Preto

Profº Dr. Osvaldo Ishizawa
FATEC- Campus de Catanduva- SP

Profº Dr. Marcio Andrey Teixeira
IFSP– Campus de Catanduva- SP

AGRADECIMENTOS

A realização deste trabalho só foi possível graças à presença, apoio e inspiração de muitas pessoas ao longo da minha jornada acadêmica. Em primeiro lugar, agradeço a Deus, fonte de sabedoria, força e discernimento, e a Jesus Cristo, meu Salvador, por me sustentar nos momentos de dúvida, cansaço e desafio. Sem a fé que me acompanha, eu não teria alcançado este marco. À minha esposa Mariangela, minha companheira de vida, meu amor e porto seguro, agradeço por sua paciência, incentivo constante e por compartilhar comigo cada etapa desta caminhada. À minha filha Manuela, por ser minha luz diária e motivo maior de superação — seu sorriso sempre me deu forças para continuar. Aos meus pais, Jason e Rosineide, minha eterna gratidão por todo amor, educação e valores que me transmitiram. Vocês são as raízes que sustentam cada conquista da minha vida. Aos meus irmãos, Karina e Thiago, pelo carinho, apoio e amizade que me acompanharam desde sempre. Ao meu orientador, Professor Rodrigo, agradeço profundamente pela orientação segura, pelos ensinamentos valiosos e pela confiança em meu trabalho. Sua dedicação foi essencial para o desenvolvimento desta pesquisa. Ao IBILCE, por oferecer a infraestrutura necessária para a realização deste estudo, e por proporcionar um ambiente acadêmico que estimula o crescimento científico e pessoal. A todos que, de alguma forma, contribuíram para que esta tese se concretizasse, meu sincero muito obrigado.

"A ciência é construída de fatos, assim como uma casa é construída de pedras; mas uma coleção de fatos não é mais ciência do que um monte de pedras é uma casa."

(Poincaré, 1905, p. 12)

RESUMO

Sistemas biométricos de voz desempenham um papel crítico em diversas aplicações de segurança, incluindo autenticação de dispositivos eletrônicos, verificação de transações bancárias e comunicações confidenciais. Apesar de sua ampla utilidade, esses sistemas são cada vez mais alvo de ataques de *spoofing* sofisticados que utilizam técnicas avançadas de inteligência artificial para gerar fala sintética realista. Abordar as vulnerabilidades inerentes aos sistemas de autenticação por voz tornou-se, portanto, uma tarefa urgente e essencial. Este estudo propõe uma nova análise experimental que explora extensivamente diversas estratégias de redução de dimensionalidade em conjunto com modelos de *machine learning* supervisionados para identificar de forma eficaz sinais de voz falsificados. Nosso *framework* envolve a extração de características *multicepstrais* seguida da aplicação de diversos métodos de redução de dimensionalidade, tais como Análise de Componentes Principais (*Principal Component Analysis* — PCA), Decomposição em Valores Singulares Truncada (*Truncated Singular Value Decomposition* — SVD), seleção estatística de características (valor F do ANOVA, Informação Mútua), Eliminação Recursiva de Características (*Recursive Feature Elimination* — RFE), seleção baseada em regularização via LASSO, importância de características por *Random Forest* e técnicas de Importância por Permutação. A avaliação empírica, utilizando o conjunto de dados *ASVSpoof 2017 v2.0*, mede o desempenho da classificação com a métrica de Taxa de Erro Igual (*Equal Error Rate* — EER), atingindo valores em torno de 10%. Nossa análise comparativa demonstra ganhos significativos de desempenho quando os métodos de redução de dimensionalidade são aplicados, ressaltando seu valor no aumento da segurança e eficácia dos sistemas de verificação biométrica de voz contra ameaças emergentes de *spoofing*.

Palavras-Chave: *Spoofing Detection*; Redução de Dimensionalidade; Reconhecimento de Padrões; Análise Cepstral; Aprendizado de Máquina.

ABSTRACT

Voice biometric systems play a critical role in numerous security applications, including electronic device authentication, banking transaction verification, and confidential communications. Despite their widespread utility, these systems are increasingly targeted by sophisticated spoofing attacks that leverage advanced artificial intelligence techniques to generate realistic synthetic speech. Addressing the vulnerabilities inherent to voice-based authentication systems has thus become both urgent and essential. This study proposes a novel experimental analysis that extensively explores various dimensionality reduction strategies in conjunction with supervised machine learning models to effectively identify spoofed voice signals. Our framework involves extracting multicepstral features followed by the application of diverse dimensionality reduction methods such as Principal Component Analysis (PCA), Truncated Singular Value Decomposition (SVD), statistical feature selection (ANOVA F-value, Mutual Information), Recursive Feature Elimination (RFE), regularization-based LASSO selection, Random Forest feature importance, and Permutation Importance techniques. Empirical evaluation using the ASVSpoof 2017 v2.0 dataset measures the classification performance with the Equal Error Rate (EER) metric, achieving values of approximately 10%. Our comparative analysis demonstrates significant performance gains when dimensionality reduction methods are applied, underscoring their value in enhancing the security and effectiveness of voice biometric verification systems against emerging spoofing threats.

Keywords: Spoofing Detection; Dimensionality Reduction; Pattern Recognition; Cepstral Analysis; Machine Learning..

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AE	*Autoencoder* — Autoencoder
ASV	*Automatic Speaker Verification* — Verificação Automática de Locutor
ASVspoof	*Automatic Speaker Verification Spoofing and Countermeasures Challenge* — Desafio Internacional de Detecção de Falsificação de Voz
BFCC	*Bark Frequency Cepstral Coefficients* — Coeficientes Cepstrais na Escala de Bark
CFCCIF	*Constant Q Cepstral Coefficients with Instantaneous Frequency* — Coeficientes Cepstrais de Q Constante com Freqüência Instantânea
CNN	*Convolutional Neural Network* — Rede Neural Convolucional
CQCC	*Constant-Q Cepstral Coefficients* — Coeficientes Cepstrais de Q Constante
DNN	*Deep Neural Network* — Rede Neural Profunda
Dev	*Development Set* — Conjunto de Desenvolvimento
EER	*Equal Error Rate* — Taxa de Erro Igual
ESA	*Energy Spectral Analysis* — Análise Espectral de Energia
Eval	*Evaluation Set* — Conjunto de Avaliação
FPR	*False Positive Rate* — Taxa de Falsos Positivos
FNR	*False Negative Rate* — Taxa de Falsos Negativos
GA	*Genetic Algorithm* — Algoritmo Genético
GWO	*Grey Wolf Optimizer* — Otimização do Lobo Cinzento
HNR	*Harmonics-to-Noise Ratio* — Relação Harmônico-Ruído
LFCC	*Linear Frequency Cepstral Coefficients* — Coeficientes Cepstrais de Freqüência Linear
LASSO	*Least Absolute Shrinkage and Selection Operator* — Método de Regularização e Seleção de Atributos
LSTM	*Long Short-Term Memory* — Memória de Longo e Curto Prazo

MFCC	*Mel Frequency Cepstral Coefficients* — Coeficientes Cepstrais na Escala de Mel
MI	*Mutual Information* — Informação Mútua
PCA	*Principal Component Analysis* — Análise de Componentes Principais
PIS	*Permutation Importance Score* — Importância por Permutação
PSO	*Particle Swarm Optimization* — Otimização por Enxame de Partículas
QCFCCIF	*Quasi-Constant Frequency Cepstral Coefficients with Instantaneous Frequency* — Coeficientes Cepstrais de Freqüência Quase Constante com Freqüência Instantânea
RFE	*Recursive Feature Elimination* — Eliminação Recursiva de Atributos
RNN	*Recurrent Neural Network* — Rede Neural Recorrente
ROC	*Receiver Operating Characteristic* — Curva Característica de Operação do Receptor
SVD	*Singular Value Decomposition* — Decomposição em Valores Singulares
TTS	*Text-to-Speech* — Síntese de Voz a partir de Texto
VC	*Voice Conversion* — Conversão de Voz
VITS	*Variational Inference with Adversarial Learning for End-to-End Text-to-Speech* — Modelo End-to-End para Síntese de Voz com Inferência Variacional e Aprendizado Adversarial
WGAN	*Wasserstein Generative Adversarial Network* — Rede Geradora Adversarial de Wasserstein

SUMÁRIO

1 INTRODUÇÃO

Sistemas de autenticação por voz têm se tornado cada vez mais integrados à identificação segura de usuários devido à sua praticidade, unicidade e caráter não invasivo, aproveitando as características inerentes dos traços vocais individuais (Yudin et al., 2019; Senk; Dotzler, 2011; Memon et al., 2017). Apesar dessas vantagens, a autenticação baseada em voz continua vulnerável a ataques sofisticados de *spoofing*, que envolvem a manipulação ou síntese artificial de sinais de voz utilizando técnicas avançadas, especialmente aquelas baseadas em inteligência artificial (Khan et al., 2023; Shaheed et al., 2024). Esses métodos de *spoofing* representam uma ameaça significativa à confiabilidade e segurança de sistemas biométricos de autenticação por voz (Yamagishi et al., 2017).

Entre as categorias de atributos mais bem-sucedidas para a detecção de sinais de voz falsificados estão as características cepstrais, em especial os Coeficientes Cepstrais na Frequência Mel (MFCCs) e suas variantes (Godino-Llorente; Gómez-Vilda, 2004). Esses atributos são capazes de capturar detalhes espectrais sutis da fala, tornando-os altamente eficazes na distinção entre fala genuína e sinais sintéticos ou manipulados. Mais recentemente, abordagens multicepstrais — baseadas na agregação ou fusão de múltiplas representações cepstrais — têm surgido como uma estratégia poderosa para melhorar ainda mais o desempenho (Contreras; Viana; Guido, 2023). Essas abordagens oferecem perspectivas complementares do sinal por meio de estratégias diversas de projeção e mapeamento. Entretanto, a riqueza representacional ampliada das características multicepstrais tem um custo: altíssima dimensionalidade. Isso não apenas acarreta aumento no uso de memória e demanda computacional, como também introduz desafios como *overfitting*, representações esparsas dos dados e desempenho degradado dos classificadores — um fenômeno comumente conhecido como maldição da dimensionalidade (Bellman, 1961; Beyer et al., 1999; Gorban; Makarov; Tyukin, 2020). Para lidar com isso, a redução de dimensionalidade torna-se uma etapa crucial de pré-processamento, permitindo condensar os componentes mais informativos do espaço de atributos, ao mesmo tempo em que se minimiza a redundância e a sobrecarga computacional.

Este trabalho é uma versão estendida de um estudo preliminar (Contreras et al., 2025) no qual foram investigadas algumas técnicas de redução de dimensionalidade aplicadas a características multicepstrais no contexto de detecção de *spoofing* de voz. Enquanto a versão anterior focava em um conjunto mais restrito de métodos, este trabalho expande essa base de diversas maneiras importantes. Primeiramente, foi introduzida uma seção dedicada à revisão de trabalhos relacionados para contextualizar melhor as contribuições. Em seguida, o escopo do arcabouço experimental foi ampliado significativamente, com a inclusão de estratégias adicionais de redução de dimensionalidade. Estas agora abrangem técnicas como Análise de Componentes Principais (PCA), Decomposição de Valores Singulares Truncada (SVD) (Wall; Rechtsteiner; Rocha, 2003), seleção pelo valor-F da ANOVA, seleção por Informação Mútua, Eliminação

Recursiva de Atributos (RFE), seleção baseada em regressão LASSO, seleção de importância baseada em *Random Forest* e análise de Importância por Permutação. Por fim, também foram incorporados atributos não-cepstrais à análise, enriquecendo ainda mais a representação dos sinais de fala e permitindo uma avaliação mais abrangente das capacidades discriminativas.

As principais contribuições científicas deste trabalho estendido podem ser resumidas da seguinte forma:

- A incorporação de atributos não-cepstrais para enriquecer o poder descritivo da representação do sinal de voz;
- Uma análise experimental mais ampla, compreendendo técnicas adicionais de redução de dimensionalidade e um conjunto mais diverso de configurações;
- Validação extensiva no conjunto de dados de referência *ASVSpoof 2017 v2.0*, demonstrando a eficiência das técnicas propostas em diversas condições experimentais.

A estrutura deste trabalho está delineada da seguinte forma: o Capítulo 2.1 apresenta a fundamentação teórica que sustenta o estudo, fornecendo conceitos essenciais e definições necessárias para a compreensão do problema. O Capítulo 2.2 discute os avanços recentes e a literatura relevante sobre detecção de *spoofing* baseada em voz por meio de métodos de inteligência artificial. O Capítulo 2.3 apresenta um resumo conceitual e procedimental da extração de coeficientes cepstrais em formato matricial. A metodologia e o arcabouço experimental adotados neste estudo são descritos no Capítulo 2.4. O Capítulo 2.5 descreve as configurações e parametrizações específicas derivadas do arcabouço generalizado. Os resultados empíricos, juntamente com interpretações detalhadas, são apresentados no Capítulo 3. Por fim, o Capítulo 4 oferece reflexões finais e propõe direções para pesquisas futuras.

2 DESENVOLVIMENTO

2.1 FUNDAMENTAÇÃO TEÓRICA

2.1.1 Autenticação por Voz e Desafios de Segurança

Os sistemas biométricos de autenticação por voz utilizam características únicas do trato vocal humano para verificar a identidade de um indivíduo (Reynolds, 2002). A biometria vocal apresenta vantagens significativas sobre outros métodos de autenticação, como a não necessidade de contato físico, a possibilidade de uso remoto e a aceitação social mais ampla (Yu, 2015).

Esses sistemas são amplamente aplicados em diferentes domínios, incluindo autenticação de transações bancárias (Tian et al., 2016), desbloqueio de dispositivos móveis e controle de acesso a informações sensíveis (Yamagishi et al., 2019a). Em geral, um Sistema Automático de Verificação de Locutor (*Automatic Speaker Verification* — ASV) opera comparando uma amostra de voz fornecida com um modelo previamente cadastrado, emitindo uma decisão de aceitação ou rejeição (Campbell, 1997).

Apesar de sua utilidade, sistemas ASV são vulneráveis a ataques de *voice spoofing*, nos quais um agente mal-intencionado tenta enganar o sistema utilizando voz sintética, convertida ou reproduzida a partir de gravações legítimas (Wu et al., 2015a). O avanço das tecnologias de síntese e conversão de voz — impulsionado por técnicas de aprendizado profundo — tem tornado esses ataques cada vez mais realistas e difíceis de detectar (Todisco et al., 2019).

Nesse contexto, a detecção automática de *spoofing* vocal (*Spoofing Countermeasure*, ou CM) surge como uma camada adicional de defesa, atuando em paralelo ao sistema ASV tradicional e responsável por identificar se a amostra de voz é autêntica ou fraudulenta (Kinnunen et al., 2017). A combinação dessas duas camadas é conhecida como *tandem detection cost function* (t-DCF), métrica que avalia o desempenho conjunto de verificação e detecção (Kinnunen et al., 2018).

2.1.2 Tipos de Ataques em Voice Spoofing

Ataques de *voice spoofing* podem ser classificados em quatro categorias principais: ataques de reprodução (*replay attacks*), síntese de voz (*text-to-speech* — TTS), conversão de voz (*voice conversion* — VC) e ataques híbridos que combinam múltiplas técnicas (Wu et al., 2015a; Wang et al., 2020).

2.1.2.1 Replay Attacks

O ataque de reprodução (*replay attacks*) consiste na reprodução de uma gravação legítima da voz de um usuário para enganar o sistema de autenticação (Kinnunen et al., 2012). Apesar de ser considerado o mais simples de implementar, ainda é uma ameaça relevante, especialmente contra sistemas sem detecção de *liveness*. A gravação pode ser feita com dispositivos simples e

reproduzida por alto-falantes próximos ao microfone do sistema-alvo. A Figura 1 apresenta um exemplo ilustrativo do processo de ataque por repetição, comparando a fala genuína com a fala reproduzida por um invasor, evidenciando as etapas envolvidas nesse tipo de fraude.

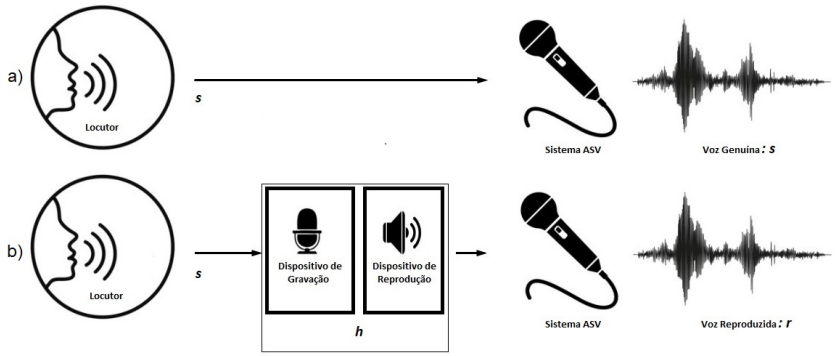

Figura 1 – Uma ilustração dos *replay attacks*. (a) é o processo de geração de fala genuína, a fala genuína é uma gravação de um acesso de boa-fé ao microfone do sistema ASV. (b) é o processo de geração de voz de repetição. A fala de repetição é uma gravação de falsificação na qual um invasor grava a voz do falante alvo e depois a reproduz em um sistema ASV

Fonte: Adaptado de Lin et al. (2020)

2.1.2.2 Síntese de Voz (TTS)

Os ataques por síntese de voz utilizam sistemas *text-to-speech* para gerar fala artificial a partir de texto escrito (Zen; Tokuda; Black, 2009). Com os avanços em redes neurais profundas, modelos como Tacotron 2 (Shen et al., 2018) e VITS (Kim et al., 2021) conseguem produzir fala altamente natural, dificultando a detecção de falsificações. A Figura 2 demonstra graficamente um ataque por síntese de voz, no qual um texto é transformado em fala artificial utilizando um sistema TTS.

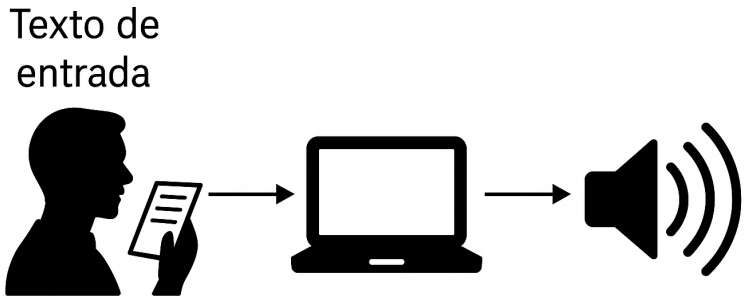

Figura 2 – Ilustração de um ataque por sintese de voz onde o texto é convertido em voz por meio de um sistema *text-to-speech*.

Fonte: Elaborado pelo autor (2025)

2.1.2.3 Conversão de Voz (VC)

Na conversão de voz, um áudio de fala de um locutor-fonte é transformado para soar como se tivesse sido produzido pelo locutor-alvo (Toda; Black; Tokuda, 2007). Técnicas modernas de VC empregam modelos generativos como *CycleGAN* e *Autoencoders* condicionais para preservar o conteúdo linguístico enquanto alteram características vocais. Na Figura 3, observa-se uma representação esquemática dos ataques de conversão de voz, em que a voz do locutor original é modificada para se assemelhar à de outro indivíduo, mantendo o conteúdo linguístico.

Figura 3 – Ataques de conversão de voz manipulam a voz do locutor para produzir vozes sintéticas.

Fonte: Elaborado pelo autor (2025)

2.1.2.4 Ataques Híbridos

Ataques híbridos combinam múltiplas técnicas, por exemplo, um áudio sintetizado que passa por um processo de conversão de voz ou um ataque de replay aplicado sobre voz artificial. Essa abordagem busca explorar vulnerabilidades específicas de cada camada do sistema de verificação (Wu et al., 2015a).

2.1.3 Técnicas de Extração de Características para Detecção de *Voice Spoofing*

A etapa de extração de características é fundamental na detecção de *voice spoofing*, pois define como o sinal de voz será representado para posterior análise e classificação. Nesta seção, abordamos as principais técnicas cepstrais e espectrais utilizadas na literatura, incluindo MFCC, LFCC, CQCC, BFCC e representações baseadas em espectrogramas. A Figura 4 oferece uma comparação visual entre diferentes representações de atributos cepstrais (MFCC, LFCC, CQCC, etc.) e espectrogramas, extraídos de um mesmo sinal de voz, evidenciando como cada técnica representa o conteúdo acústico.

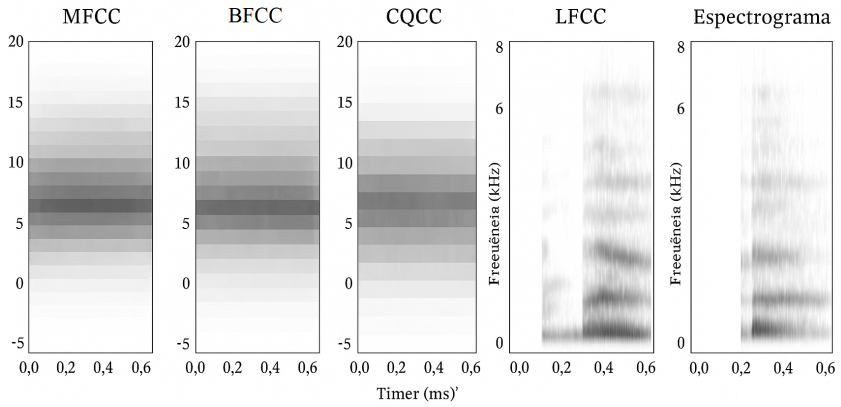

Figura 4 – Comparação ilustrativa entre diferentes representações cepstrais e espectrogramas para um mesmo sinal de voz.
Fonte: Elaborado pelo autor (2025)

2.1.3.1 MFCC – Mel-Frequency Cepstral Coefficients

Os *Mel-Frequency Cepstral Coefficients* (MFCCs) são amplamente utilizados em reconhecimento de fala e detecção de falsificação de voz (Davis; Mermelstein, 1980). Eles modelam a percepção auditiva humana ao aplicar a escala de Mel, extraindo informações do espectro de potência logarítmico por meio de uma transformada discreta do cosseno (DCT). Apesar de sua popularidade, MFCCs podem ser vulneráveis a ataques por síntese e conversão de voz devido à perda de detalhes espectrais finos.

2.1.3.2 LFCC – Linear Frequency Cepstral Coefficients

Os *Linear Frequency Cepstral Coefficients* (LFCCs) diferem dos MFCCs por utilizarem uma escala de frequência linear, preservando informações uniformemente ao longo do espectro (Mon et al., 2023). Essa abordagem pode capturar nuances de alta frequência que são atenuadas na escala Mel, oferecendo vantagens contra certos tipos de *spoofing*.

2.1.3.3 CQCC – Constant-Q Cepstral Coefficients

Os *Constant-Q Cepstral Coefficients* (CQCCs) (Todisco; Delgado; Evans, 2017a) utilizam a Transformada Constant-Q, que oferece resolução variável: alta em baixas frequências e baixa em altas frequências. Isso permite uma análise mais precisa de componentes harmônicos e características de fala gerada por sistemas de síntese e conversão.

2.1.3.4 BFCC – Bark Frequency Cepstral Coefficients

Os *Bark Frequency Cepstral Coefficients* (BFCCs) utilizam a escala de Bark, baseada em estudos psicoacústicos sobre a percepção de bandas críticas do ouvido humano (Skowronski; Har-

ris, 2004). Essa representação pode capturar de forma mais fiel aspectos perceptuais relevantes para a detecção de falsificações.

2.1.3.5 Embeddings de Espectrogramas

Além das abordagens tradicionais baseadas em coeficientes cepstrais, representações de espectrogramas — especialmente os *log-mel spectrograms* — têm sido empregadas como entrada para redes neurais convolucionais (CNNs) e modelos baseados em *transformers* (Purwins et al., 2019). Esses métodos permitem que a rede aprenda automaticamente filtros discriminativos, explorando padrões temporais e espectrais de forma conjunta.

2.1.4 Redução de Dimensionalidade na Detecção de *Voice Spoofing*

O problema da alta dimensionalidade em sistemas de detecção de *voice spoofing* é amplamente discutido na literatura. Representações multicepstrais e combinações de diferentes tipos de atributos podem gerar vetores de características com milhares de dimensões, resultando em maior custo computacional, risco de *overfitting* e na chamada *maldição da dimensionalidade* (Bellman, 1961). Métodos de redução de dimensionalidade visam mitigar esses problemas, preservando a informação mais relevante para a tarefa de classificação. A Figura 5 apresenta um diagrama de fluxo comparando os principais métodos de redução de dimensionalidade aplicados à detecção de *spoofing*, destacando como cada técnica transforma o espaço original de atributos.

2.1.4.1 PCA – Principal Component Analysis

A *Principal Component Analysis* (PCA) é um método linear que transforma o conjunto de dados original em um novo sistema de coordenadas, onde as primeiras dimensões (componentes principais) retêm a maior variância possível (Jolliffe, 2002). Ao reduzir o número de componentes, é possível manter boa parte da informação enquanto se diminui a dimensionalidade, reduzindo também o custo computacional.

2.1.4.2 SVD – Truncated Singular Value Decomposition

A *Singular Value Decomposition* (SVD) é uma técnica de fatoração de matrizes que decompõe os dados em três matrizes — U, Σ e V^T —, permitindo identificar direções principais de variação (Golub; Loan, 2013). Na forma truncada, mantém-se apenas os maiores valores singulares, o que resulta em uma representação mais compacta.

2.1.4.3 Seleção Estatística de Atributos

Técnicas como *ANOVA F-value* e *Mutual Information* são aplicadas para selecionar atributos com maior poder discriminativo (Guyon; Elisseeff, 2003). O *ANOVA F-value* compara a variância

entre classes e dentro de classes, enquanto a *Mutual Information* mede a dependência estatística entre cada atributo e o rótulo da classe.

2.1.4.4 RFE – Recursive Feature Elimination

O *Recursive Feature Elimination* (RFE) utiliza um modelo preditivo — como SVM ou regressão logística — para atribuir pesos aos atributos, removendo iterativamente aqueles com menor importância (Guyon et al., 2002). Esse método tende a encontrar subconjuntos de atributos mais otimizados para o classificador escolhido.

2.1.4.5 LASSO – Least Absolute Shrinkage and Selection Operator

O *Least Absolute Shrinkage and Selection Operator* (LASSO) aplica regularização L1 em modelos lineares, penalizando coeficientes e forçando alguns a zero, o que resulta em seleção automática de atributos (Tibshirani, 1996). Essa técnica é útil quando se busca uma solução esparsa, reduzindo o risco de *overfitting*.

2.1.4.6 Importância de Atributos via Random Forest

O *Random Forest* pode estimar a importância de cada atributo calculando a redução média de impureza (*mean decrease in impurity*) ou por meio de permutação (Breiman, 2001). Essa abordagem não linear é robusta e capaz de capturar interações complexas entre atributos.

2.1.4.7 Permutation Importance

O *Permutation Importance* (Fisher; Rudin; Dominici, 2019) mede o impacto de cada atributo embaralhando seus valores e observando a degradação no desempenho do modelo. Essa técnica é independente do algoritmo de treinamento, tornando-a aplicável a qualquer modelo treinado.

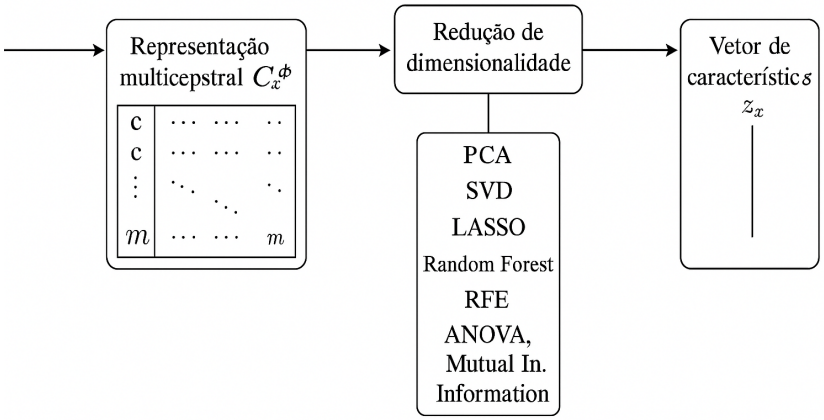

Figura 5 – Fluxo ilustrativo dos principais métodos de redução de dimensionalidade aplicados à detecção de *voice spoofing*.

Fonte: Elaborado pelo autor (2025)

2.1.5 Modelos de Aprendizado Supervisionado para Detecção de *Voice Spoofing*

A escolha de modelos de aprendizado supervisionado é crucial para o desempenho de sistemas de detecção de *voice spoofing*. Esses modelos aprendem padrões discriminativos a partir de um conjunto de treinamento rotulado, mapeando vetores de características para classes como “autêntico” ou “spoofado”. Nesta seção, são discutidos os principais modelos aplicados na literatura para esta tarefa.

2.1.5.1 Support Vector Machines (SVM)

As *Support Vector Machines* (SVM) são classificadores lineares ou não-lineares que buscam maximizar a margem entre as classes no espaço de características (Cortes; Vapnik, 1995). Em *voice spoofing*, SVMs são frequentemente combinadas com atributos cepstrais como MFCC e CQCC, usando núcleos como RBF para capturar fronteiras complexas (Villalba; Lleida, 2010).

2.1.5.2 Redes Neurais Profundas (DNN)

As *Deep Neural Networks* (DNN) também conhecidas como Redes Neurais Profundas consistem em múltiplas camadas densamente conectadas que aprendem representações hierárquicas dos dados (LeCun; Bengio; Hinton, 2015). Em *spoofing detection*, DNNs são capazes de explorar correlações não-lineares entre atributos acústicos, mas exigem grande volume de dados para evitar *overfitting*.

2.1.5.3 Redes Neurais Convolucionais (CNN)

As *Convolutional Neural Networks* (CNN) ou Redes Neurais Convolucionais são amplamente utilizadas para processar espectrogramas como imagens bidimensionais (Abdel-Hamid et al., 2014). Sua capacidade de extrair padrões locais de frequência e tempo as torna eficazes na detecção de artefatos gerados por síntese ou conversão de voz (Lai; Chen; Lee, 2019).

2.1.5.4 Redes Recorrentes (RNN) com LSTM

As *Recurrent Neural Networks* (RNN) baseadas em *Long Short-Term Memory* (LSTM) são adequadas para dados sequenciais, permitindo capturar dependências temporais de longo alcance em sinais de fala (Hochreiter; Schmidhuber, 1997). Na detecção de *spoofing*, LSTMs podem identificar inconsistências temporais típicas de vozes artificiais.

2.1.5.5 Random Forest

O *Random Forest* combina múltiplas árvores de decisão treinadas sobre subconjuntos de dados e atributos (Breiman, 2001). É robusto a ruídos e menos suscetível a *overfitting* em comparação a modelos de árvore única, sendo útil para atributos acústicos heterogêneos.

2.1.5.6 Extreme Gradient Boosting (XGBoost)

O *XGBoost* é uma implementação otimizada de *gradient boosting* que constrói modelos aditivos de forma sequencial (Chen; Guestrin, 2016). Apresenta alta acurácia e eficiência computacional, sendo competitivo em tarefas de classificação binária de *spoofing*.

2.1.5.7 Considerações sobre Seleção de Modelo

A seleção do classificador ideal depende de múltiplos fatores, incluindo:

- tipo e dimensionalidade dos atributos;
- disponibilidade de dados para treinamento;
- requisitos de tempo de execução;
- robustez a ruídos e variações de canal.

Na prática, avaliações comparativas são conduzidas utilizando métricas como EER (*Equal Error Rate*) e min-tDCF (*minimum detection cost function*) (Kinnunen et al., 2020).

A Figura 6 sintetiza os principais modelos supervisionados empregados na detecção de *spoofing*, juntamente com seus respectivos fluxos de entrada e saída de dados.

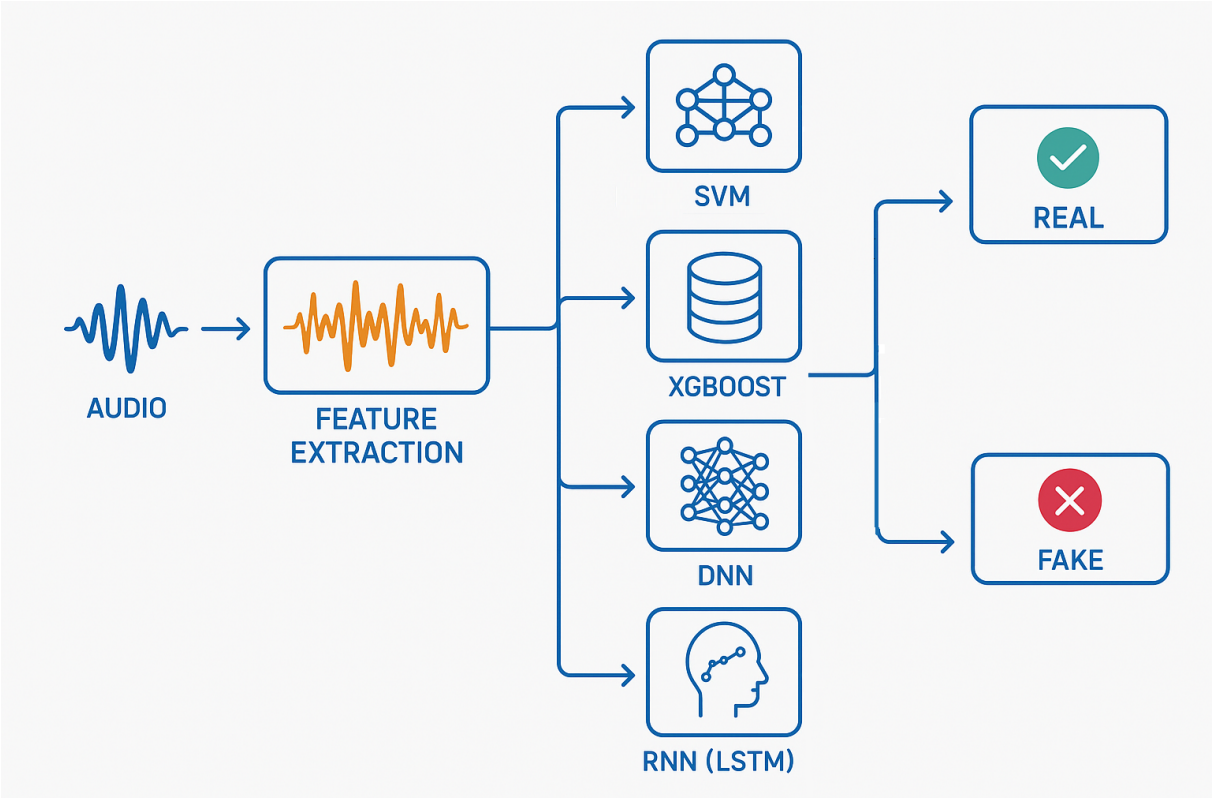

Figura 6 – Principais modelos supervisionados aplicados à detecção de *voice spoofing* e seus fluxos de entrada/saída.

Fonte: Elaborado pelo autor (2025)

2.1.6 Técnicas de *Data Augmentation* e Normalização de Atributos

A variabilidade nas condições de gravação, no canal de transmissão e nas características do locutor impõe desafios significativos aos sistemas de detecção de *voice spoofing*. Nesse contexto, o uso de *data augmentation* e técnicas de normalização é fundamental para melhorar a generalização dos modelos e reduzir a sensibilidade a variações irrelevantes.

2.1.6.1 *Data Augmentation*

O *data augmentation* consiste na criação de amostras adicionais de treinamento por meio da aplicação de transformações controladas sobre os dados originais (Shorten; Khoshgoftaar, 2019). Essa abordagem é particularmente relevante em tarefas de detecção de *spoofing*, onde os conjuntos de dados podem ser limitados ou apresentar desbalanceamento entre classes.

Entre as técnicas mais comuns aplicadas a sinais de voz, destacam-se:

- **Ruído aditivo**: inserção de ruídos brancos ou ambientais para simular diferentes condições acústicas (Ko et al., 2015).
- **Variação de velocidade e tom**: alteração do tempo e da frequência fundamental para aumentar a robustez frente a distorções temporais (Cai; Cai; Li, 2020).
- ***SpecAugment***: mascaramento temporal e espectral diretamente no espectrograma (Park et al., 2019).
- **Reverberação simulada**: convolução do sinal com respostas ao impulso (RIRs) para simular ambientes reverberantes (Ko et al., 2017).

2.1.6.2 Normalização de Atributos

A normalização de atributos visa padronizar a escala dos vetores de características, reduzindo o impacto de variações não relacionadas ao conteúdo da fala (Liu; Sahidullah; Kinnunen, 2021). Entre os métodos mais comuns estão:

- **Cepstral Mean and Variance Normalization (CMVN)**: subtração da média e divisão pelo desvio-padrão das características cepstrais (Viikki; Laurila, 1998).
- **Feature Warping**: mapeamento não linear dos atributos para aproximar uma distribuição normal (Pelecanos; Sridharan, 2001).
- **Instance Normalization**: normalização dos atributos por instância, útil para espectrogramas de entrada em redes neurais (Ulyanov; Vedaldi; Lempitsky, 2016).

2.1.6.3 Impacto na Detecção de *Voice Spoofing*

Estudos recentes mostram que a combinação de *data augmentation* com normalização apropriada pode reduzir significativamente o *Equal Error Rate* (EER) em *benchmarks* como o *ASVspoof* (Tak et al., 2021). A robustez obtida é atribuída à redução de *overfitting* e à maior diversidade estatística no conjunto de treinamento.

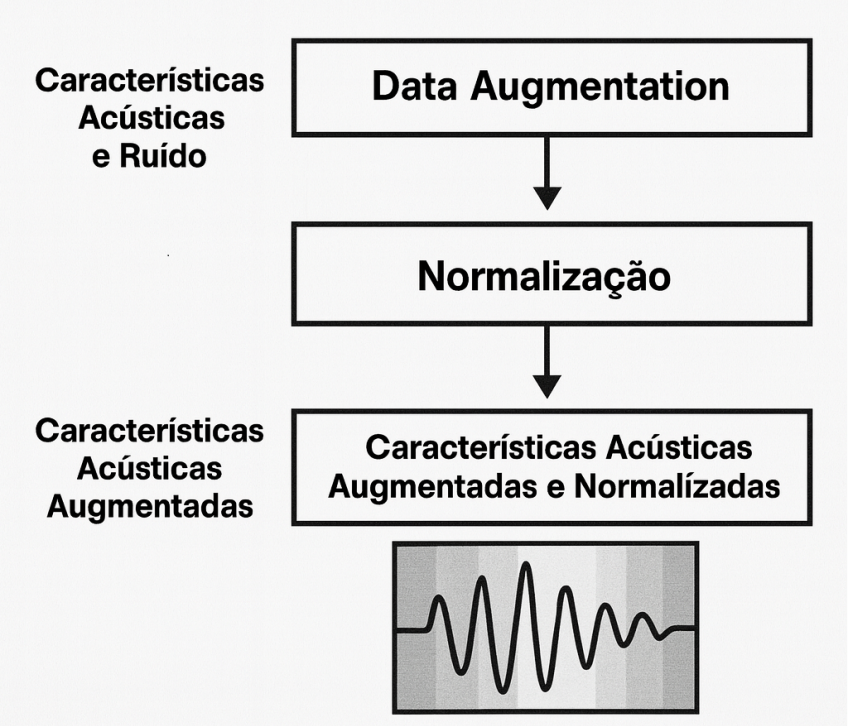

Figura 7 – Fluxo de pré-processamento de dados com técnicas de *data augmentation* e normalização aplicadas a atributos acústicos para detecção de *voice spoofing*.
Fonte: Elaborado pelo autor (2025)

2.1.7 Extração de Atributos Cepstrais e Não-Cepstrais

A escolha e a extração de atributos adequados é um passo crítico no desempenho de sistemas de detecção de *voice spoofing*. Esses atributos representam, de forma compacta, informações relevantes sobre o conteúdo espectral e temporal da fala, possibilitando que modelos de aprendizado de máquina distingam entre amostras legítimas e falsificadas.

2.1.7.1 Atributos Cepstrais

Os atributos cepstrais descrevem a estrutura espectral do sinal de voz e são amplamente utilizados devido à sua eficácia e compacidade (Davis; Mermelstein, 1980). Os principais exemplos incluem:

- **MFCC (Mel-Frequency Cepstral Coefficients)**: obtidos a partir da escala Mel, modelam a percepção humana de frequências e são padrão em reconhecimento automático de fala (Muda; Begam; Elamvazuthi, 2010).

- **LFCC (Linear-Frequency Cepstral Coefficients)**: semelhantes aos MFCC, mas calculados na escala linear, sendo mais sensíveis a detalhes de alta frequência, úteis em detecção de *spoofing* (Todisco; Delgado; Evans, 2017a).
- **CQCC (Constant-Q Cepstral Coefficients)**: extraídos a partir da Transformada de Q Constante, oferecem resolução variável de frequência, com melhor discriminação em baixas frequências (Todisco; Delgado; Evans, 2016).
- **BFCC (Bark-Frequency Cepstral Coefficients)**: baseados na escala Bark, aproximam outra função psicoacústica de percepção auditiva (Hermansky, 1990).

2.1.7.2 Atributos Não-Cepstrais

Atributos não-cepstrais podem capturar características complementares aos coeficientes cepstrais, incluindo aspectos prosódicos, temporais e espectrais que melhoram a robustez do sistema (Gobl; Chasaide, 2003). Entre eles:

- **Formantes**: frequências de ressonância do trato vocal, úteis para capturar padrões fisiológicos e patológicos (Deller; Proakis; Hansen, 2010).
- **Medidas de ruído**: como *Harmonics-to-Noise Ratio* (HNR) e *Jitter/Shimmer*, associadas à qualidade vocal (Gobl; Chasaide, 2003).
- **Atributos espectro-temporais**: como *Spectral Flux*, *Spectral Centroid* e *Zero-Crossing Rate*, que fornecem informações dinâmicas sobre a energia e periodicidade (Tzanetakis; Cook, 2002).

2.1.7.3 Combinação de Atributos

Pesquisas recentes indicam que a combinação de atributos cepstrais e não-cepstrais pode melhorar significativamente a performance dos detectores, reduzindo a vulnerabilidade a ataques de síntese e conversão de voz (Sahidullah et al., 2016). O uso de *feature-level fusion* permite a criação de vetores de atributos mais ricos, enquanto a fusão em nível de decisão (*score-level fusion*) pode explorar modelos especializados para diferentes subconjuntos de características.

No contexto dessa abordagem, a Figura 8 ilustra um *pipeline* conceitual de extração de atributos cepstrais e não-cepstrais aplicado a sinais de voz para detecção de *spoofing*. O fluxo inicia com o pré-processamento do sinal de entrada, onde técnicas de segmentação e normalização são aplicadas para garantir consistência e qualidade dos quadros de análise. Em seguida, ocorre uma bifurcação em duas vias complementares: a primeira dedicada à extração de coeficientes cepstrais (como MFCCs e variantes baseadas em *wavelets*), responsáveis por capturar informações espectrais compactas; e a segunda voltada para atributos não-cepstrais, como medidas prosódicas, estatísticas espectrais e *embeddings* de espectrogramas, que exploram padrões temporais e espaciais mais amplos.

Por fim, os vetores resultantes podem ser combinados por meio de estratégias de fusão em nível de atributos (*feature-level fusion*), formando representações mais discriminativas, ou integrados tardiamente em nível de decisão (*score-level fusion*), permitindo que classificadores especializados operem em paralelo e contribuam de forma complementar para a decisão final. Essa estrutura evidencia visualmente como a integração de diferentes tipos de atributos potencializa a robustez dos sistemas contra ataques de síntese e conversão de voz, reforçando a relevância da combinação de múltiplas representações acústicas no contexto da detecção de *spoofing*.

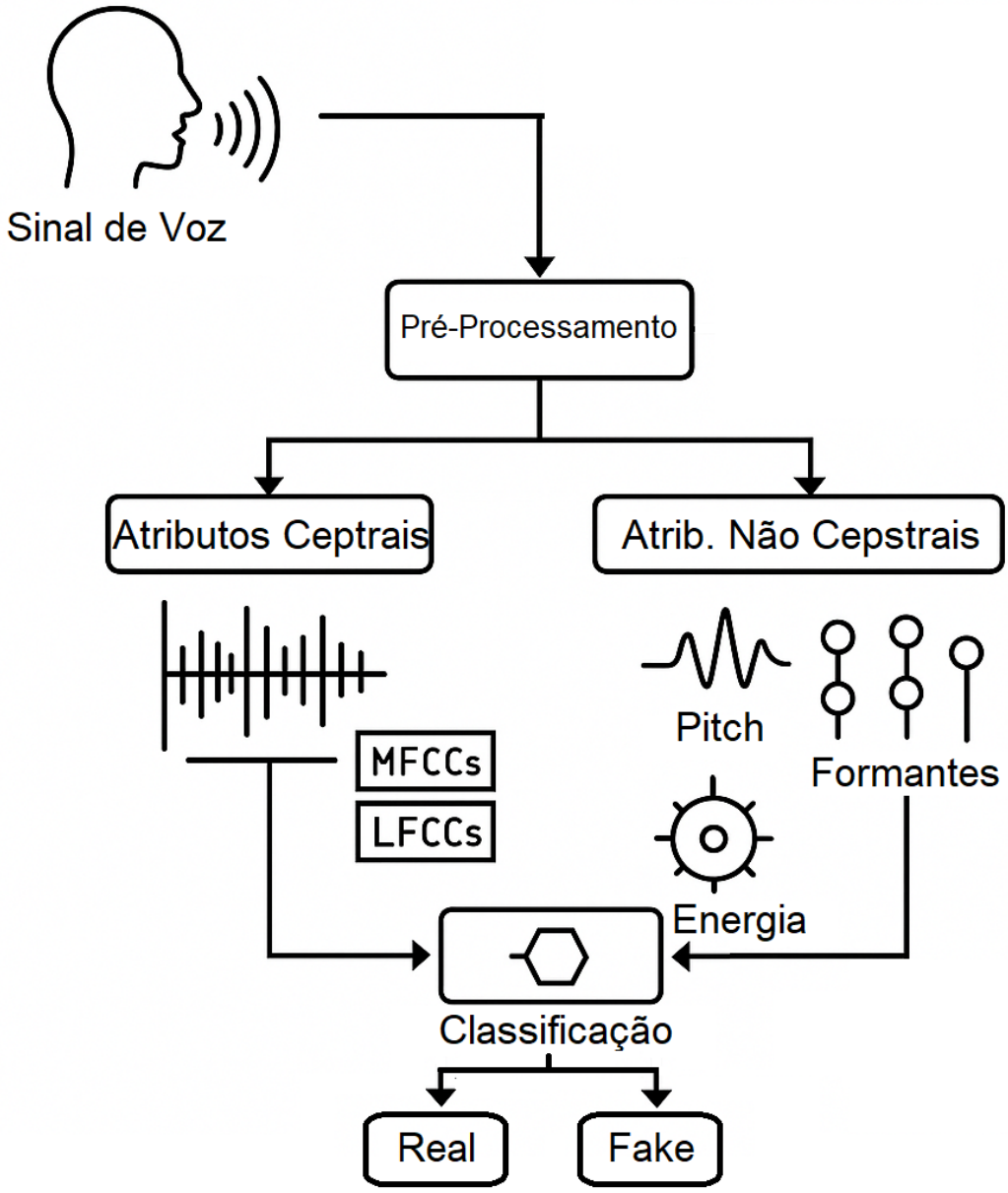

Figura 8 – Pipeline de extração de atributos cepstrais e não-cepstrais aplicado a sinais de voz para detecção de *spoofing*.

Fonte: Elaborado pelo autor (2025)

2.2 TRABALHOS RELACIONADOS

Para promover a inovação em estratégias *anti-spoofing*, o *ASVspoof Challenge* foi lançado em 2015 (Wu et al., 2015b), seguido por edições subsequentes em 2017, 2019 e 2021 (Kinnunen et al., 2017; Yamagishi et al., 2019b; Yamagishi et al., 2021). Essas iniciativas forneceram conjuntos de dados de referência contendo enunciados genuínos e falsificados, apoiando o desenvolvimento de contramedidas padronizadas. Na edição de 2015, Patel & Patil (2015) propuseram um Modelo de Mistura Gaussiana (GMM) alimentado por características derivadas dos Coeficientes Cepstrais de Filtro Coclear (CFCCs) combinados com frequência instantânea (IF). Sua abordagem simulava mecanismos de processamento auditivo e obteve desempenho notável, com EERs de 0,4079% para ataques conhecidos e 2,0132% para tipos desconhecidos. Sahidullah, Kinnunen & Hanilçi (2015) compararam sistematicamente 19 conjuntos de características agrupadas em espectrais de curto prazo, baseadas em fase e descritores de longo prazo. Seu trabalho combinou essas características com classificadores GMM e SVM, alcançando EERs tão baixos quanto 0,07% em condições combinadas e 1,67% em cenários não combinados.

A Transformada *Constant-Q* (CQT), um método inicialmente aplicado ao processamento de sinais musicais, foi adaptada para detecção de *spoofing* por Todisco, Delgado & Evans (2017a). Oferecendo melhor resolução em faixas de baixa e alta frequência, as características baseadas em CQT atingiram EERs de 0,048% para ataques conhecidos. Yang & Das (2019) aprimoraram essa técnica introduzindo normalização de baixa frequência baseada em quadros, o que aumentou a robustez à variabilidade ambiental. Seu modelo, utilizando SVMs e arquiteturas *ResNet*, atingiu EER de 10,31% no *ASVspoof 2017*. Descritores tradicionais como vetores de identidade (*i-vectors*) também mostraram-se promissores. Hanilçi (2018) demonstraram que a combinação de *i-vectors* com GMMs reduziu pela metade a EER em relação ao uso isolado de GMMs, validando sua eficácia para detecção de *spoofing* de voz.

Metodologias de *deep learning* ganharam destaque por sua capacidade de modelar padrões complexos. Qian, Chen & Yu (2016) exploraram cinco arquiteturas neurais — incluindo autoencoders, DNNs básicas e RNNs baseadas em LSTM — alcançando EERs próximas de 0% para condições conhecidas de *spoofing*. Huang & Pun (2019) introduziram um modelo híbrido *DenseNet-LSTM* utilizando características CQCC e MFCC, reduzindo a EER para 7,34% no *ASVspoof 2017*. Gomez-Alanis et al. (2019) abordaram a robustez ao ruído com uma Rede Neural Convolucional Recorrente com Portas (GRCNN), superando modelos existentes em três conjuntos de dados *ASVspoof*. Mais recentemente, Himawan et al. (2019a) enfrentaram desafios de *overfitting* utilizando *transfer learning*. Eles usaram o *AlexNet*, uma CNN pré-treinada em mais de um milhão de imagens, para extrair *embeddings* de espectrogramas, que foram então classificados via SVMs. Essa abordagem estabeleceu novos *benchmarks* e obteve uma EER de 11,21%, superando vários métodos de última geração.

Nossos estudos anteriores estabeleceram uma base sólida para a exploração de representações multicepstrais em sistemas biométricos baseados em voz. Em uma investigação inicial, introduzi-

mos estratégias de mapeamento para combinar diferentes coeficientes cepstrais — como MFCC, CQCC, BFCC e LFCC — em representações vetoriais enriquecidas para detecção de *spoofing*, enfatizando o potencial das técnicas de multi-projeção para aumentar a discriminabilidade do sinal (Contreras et al., 2023). Subsequentemente, estendemos essa abordagem ao domínio médico, onde aplicamos características multicepstrais à detecção de disfonia e analisamos o efeito de várias configurações de projeção, destacando a generalização dessas representações além da detecção de *spoofing* (Contreras; Viana; Guido, 2023). Com base nessa estrutura, propusemos o uso de algoritmos meta-heurísticos — como Algoritmo Genético (GA), Otimização por Enxame de Partículas (PSO) e Otimização do Lobo Cinzento (GWO) — para otimizar a seleção de atributos em espaços multicepstrais, obtendo ganhos de desempenho em relação às projeções de base (Contreras et al., 2024). Mais recentemente, conduzimos uma comparação sistemática de três estratégias de redução de dimensionalidade — Decomposição em Valores Singulares (SVD), Autoencoders (AE) e Algoritmos Genéticos — demonstrando sua eficácia em mitigar a maldição da dimensionalidade e melhorar a EER em até 7,11% no conjunto de dados *ASVSpoof 2017 v2.0* (Contreras et al., 2025). Esses estudos ressaltam nosso esforço contínuo em otimizar o uso de características multicepstrais por meio de redução de dimensionalidade e otimização meta-heurística no contexto de autenticação de voz segura.

Para fornecer uma visão consolidada das técnicas discutidas nesta seção, a Tabela 1 apresenta um resumo comparativo de estudos relevantes sobre detecção de *spoofing* de voz. A tabela destaca as principais metodologias adotadas, os conjuntos de dados utilizados, os melhores EERs reportados e as contribuições centrais de cada trabalho. Essa síntese comparativa evidencia a evolução das abordagens na literatura, desde características projetadas manualmente e classificadores tradicionais até redes neurais profundas e modelos híbridos. Além disso, posiciona as investigações anteriores dentro desse panorama mais amplo, ilustrando a exploração progressiva de representações multicepstrais, redução de dimensionalidade e otimização meta-heurística em sistemas biométricos baseados em voz.

Apesar dos avanços significativos na detecção de *spoofing*, a maioria dos estudos permanece centrada em representações baseadas em características cepstrais, particularmente MFCCs, CQCCs e seus derivados. Embora tais descritores tenham demonstrado notável poder discriminativo, eles frequentemente omitem aspectos complementares do sinal que poderiam ser capturados por atributos não-cepstrais. Além disso, a crescente complexidade das representações multicepstrais — impulsionada por fusões, projeções e transformações — tem introduzido espaços de características de alta dimensionalidade, que não são apenas computacionalmente exigentes, mas também propensos a problemas de *overfitting* e esparsidade. Dessa forma, as técnicas de redução de dimensionalidade não receberam atenção proporcional na literatura, apesar de seu conhecido potencial para mitigar a maldição da dimensionalidade. Métodos como PCA, SVD, ranqueamento por informação mútua e seleção de atributos guiada por modelos permanecem subexplorados no contexto da detecção de *spoofing* de voz. Da mesma forma, o uso conjunto de atributos cepstrais e não-cepstrais ainda não foi sistematicamente investigado nesse domínio,

Tabela 1 – Resumo comparativo de trabalhos relacionados sobre detecção de *spoofing* de voz.

Autores	**Técnica/Abordagem**	**Conjunto de dados**	**Melhor EER**	**Principais contribuições**
Patel & Patil (2015)	GMM com CFCC + IF	ASVspoof 2015	0,4079% (conhecido)	Simulação do processamento auditivo humano usando filtros cocleares e frequência instantânea.
Sahidullah, Kinnunen & Hanilçi (2015)	Comparação de 19 conjuntos de características + GMM/SVM	ASVspoof 2015	0,07% (conhecido)	Avaliação abrangente de características espectrais, de fase e de longo prazo.
Todisco, Delgado & Evans (2017b)	Transformada *Constant-Q*	ASVspoof 2015	0,048% (conhecido)	Adaptação de técnica de processamento de sinais musicais para detecção de *spoofing*.
Yang & Das (2019)	CQT + normalização de baixa frequência por quadro + SVM/ResNet	ASVspoof 2017	10,31%	Maior robustez a ruído e variabilidade ambiental.
Hanilçi (2018)	*i-vectors* + GMM	ASVspoof 2015	2,39%	Redução da EER ao combinar modelagem tradicional com *i-vectors*.
Qian, Chen & Yu (2016)	DNN, AE, LSTM-RNN	ASVspoof 2015	$\approx$ 0% (conhecido)	Comparação de cinco arquiteturas neurais profundas.
Huang & Pun (2019)	*DenseNet* + LSTM com MFCC + CQCC	ASVspoof 2017	7,34%	Fusão de duas redes profundas com características multicepstrais.
Gomez-Alanis et al. (2019)	*Gated Recurrent CNN*	*ASVspoof 2015/2017/2019*	–	Maior robustez a ruídos do mundo real.
Himawan et al. (2019b)	*AlexNet* (*transfer learning*) + SVM	ASVspoof 2017	11,21%	Uso de CNN pré-treinada para mitigar *overfitting* e extrair *embeddings*.
Contreras, Viana & Guido (2023)	Fusão de MFCC, CQCC, BFCC e LFCC com múltiplas projeções	ASVspoof 2017	15,02%	Introdução de mapeamentos vetoriais multicepstrais enriquecidos.
Contreras et al. (2024)	GA, PSO, GWO para seleção em espaço multicepstral	ASVspoof 2017	17,42%	Otimização meta-heurística em representações multicepstrais
Contreras et al. (2025)	SVD, AE, GA para redução de dimensionalidade	ASVspoof 2017	13,41%	Redução de dimensionalidade em espaços multicepstrais com melhoria de EER.

Fonte: Elaborada pelo autor.

deixando uma lacuna na compreensão de como diferentes descritores de sinal interagem e se complementam em cenários com dimensionalidade reduzida. Este trabalho busca preencher essas lacunas propondo um arcabouço experimental estendido que incorpora uma combinação rica de atributos cepstrais e não-cepstrais, avaliados sistematicamente sob diversas estratégias de redução de dimensionalidade. Ao fazer isso, buscamos revelar representações mais compactas, discriminativas e robustas, capazes de aprimorar o desempenho e a capacidade de generalização de sistemas de autenticação por voz na presença de ataques de *spoofing*.

2.3 FUNDAMENTOS DA EXTRAÇÃO DE CARACTERÍSTICAS CEPSTRAIS

Em sistemas de classificação baseados em voz, uma etapa inicial crucial envolve a extração de descritores significativos a partir de sinais de áudio brutos. Esses descritores, ou atributos, servem como representações compactas e informativas que facilitam o reconhecimento de padrões e a tomada de decisão. No contexto da análise de áudio, vários atributos de baixo nível derivados do domínio temporal — como energia do sinal (Guido, 2016a), entropia (Guido, 2018b), taxa de cruzamento por zero (ZCR) (Guido, 2016b) e o Operador de Energia de Teager (TEO) (Guido, 2018a) — são amplamente utilizados devido à sua simplicidade e eficiência computacional. No entanto, tais descritores podem ser insuficientes quando uma compreensão mais detalhada das características vocais é necessária, particularmente em tarefas que envolvem detecção de fala patológica ou autenticação biométrica em condições adversas.

Para lidar com isso, abordagens espectrais e baseadas em *cepstrum* são frequentemente preferidas. Esses métodos capturam a estrutura harmônica e as propriedades de modulação da fala, fornecendo informações mais ricas tanto sobre o trato vocal quanto sobre a fonte de excitação. Entre eles, os MFCCs e os coeficientes cepstrais de predição linear (LPCCs) são particularmente populares, oferecendo alto poder discriminativo em uma variedade de aplicações relacionadas à fala. Essas características cepstrais são normalmente extraídas por meio de um *pipeline* de múltiplas etapas que inclui (Contreras; Viana; Guido, 2023; Contreras et al., 2023):

- (i) um filtro de pré-ênfase para compensar a atenuação de altas frequências,
- (ii) segmentação do sinal em quadros curtos e sobrepostos,
- (iii) aplicação de funções de janela,
- (iv) transformação para o domínio da frequência — geralmente por meio da Transformada Rápida de Fourier (FFT) — e
- (v) mapeamento em bancos de filtros perceptuais ou lineares. A etapa final envolve o cálculo dos coeficientes cepstrais a partir dos invólucros espectrais resultantes.

Formalmente, um método de extração cepstral Φ é aplicado a um sinal de entrada $x \in \mathbb{R}^n$, produzindo uma matriz $\mathbf{C}_x^{\Phi} \in \mathbb{R}^{n_{\text{ceps}} \times n_{\text{frames}}}$, onde n_{ceps} é o número de coeficientes cepstrais

calculados por quadro e n_{frames} é o número total de quadros considerados. Essa matriz captura a evolução dinâmica das características cepstrais ao longo da duração do sinal:

$$\mathbf{C}_x^{\Phi} = \begin{bmatrix} — & \vec{c}_1 & — \\ — & \vec{c}_2 & — \\ & \vdots & \\ — & \vec{c}_{n_{\text{ceps}}} & — \end{bmatrix} \in \mathbb{R}^{n_{\text{ceps}} \times n_{\text{frames}}}, \tag{1}$$

onde cada vetor coluna $\vec{c}_i \in \mathbb{R}^{n_{\text{ceps}}}$ corresponde ao conjunto de características cepstrais extraídas do i-ésimo quadro de x. Detalhes adicionais sobre a formulação matemática e a implementação algorítmica dessas técnicas cepstrais podem ser encontrados em Prabakaran & Shyamala (2019), Alim & Rashid (2018).

2.4 MATERIAIS E MÉTODOS

Nesta seção, descreve-se como criar um vetor de características a partir da projeção dos coeficientes cepstrais (CCs) e de métricas adicionais não-cepstrais extraídas de um sinal de voz. O procedimento é inspirado no trabalho de Contreras, Viana & Guido (2023), Contreras et al. (2023), originalmente proposto para detecção de *spoofing* e disfonia. No entanto, neste estudo, foram introduzidas três inovações metodológicas principais:

- O uso de estratégias de aumento de dados (*data augmentation*) para dobrar o tamanho do conjunto de treinamento por meio de ruído aditivo;
- A integração de atributos acústicos não cepstrais para complementar as representações cepstrais;
- A experimentação sistemática com técnicas de redução de dimensionalidade aplicadas ao vetor final de características, a fim de otimizar a compacidade e a capacidade discriminativa da representação.

Existem várias técnicas disponíveis para o cálculo de CCs. Neste trabalho, considerou-se um conjunto diverso de descritores, denotado pelo conjunto P, definido a seguir:

$$\mathbf{P} = \{\Phi_1, \Phi_2, \ldots, \Phi_{n_{\mathbf{P}}}\}, \tag{2}$$

onde cada Φ_i é uma função de extração de coeficientes cepstrais.

Para resumir a informação contida em cada matriz cepstral e torná-la adequada para tarefas de classificação, aplica-se uma função de projeção. Enquanto a Análise de Componentes Principais (PCA) foi utilizada na formulação original, aqui esse processo foi generalizado para incluir outras estratégias, como sumarização estatística (média, desvio padrão, assimetria), bem como suas combinações. Seja $\text{Proj}(\cdot)$ uma função genérica de projeção ou composição de funções (por exemplo, $\text{Proj} = \{\text{PCA}, \text{MEAN}, \text{STD}, \text{SKEW}\}$).

Um dos objetivos centrais deste trabalho é investigar o impacto de diferentes técnicas de redução de dimensionalidade e projeção sobre o poder discriminativo dos vetores de características. Ao contrário de estudos anteriores que adotam um único método (geralmente PCA), propomos uma avaliação massiva e sistemática de uma ampla gama de estratégias de projeção e compressão, tanto no nível dos componentes cepstrais quanto no vetor final. Nossa hipótese é que a compacidade e a composição do vetor final — incluindo elementos multicepstrais e não-cepstrais — desempenham um papel crucial na otimização do desempenho para detecção de *spoofing* de voz.

Para um sinal de voz x e um método de extração cepstral ϕ, computamos a matriz de coeficientes cepstrais $\mathbf{C}_x^\phi$, bem como suas derivadas de primeira e segunda ordem, Δ_x^ϕ e $\Delta\Delta_x^\phi$. O vetor de características $\vec{v}_x^\phi$ é então definido como:

$$\vec{v}_x^\phi = \left(\text{Proj}(\mathbf{C}_x^\phi), \text{Proj}(\Delta_x^\phi), \text{Proj}(\Delta\Delta_x^\phi)\right), \tag{3}$$

onde cada $\text{Proj}(\cdot)$ mapeia uma matriz $\in \mathbb{R}^{n_{\text{ceps}} \times n_{\text{frames}}}$ para um vetor em $\mathbb{R}^{n_{\text{ceps}} \times n_{\text{proj}}}$, com n_{proj} dependendo do número e do tipo de funções de projeção utilizadas. O vetor de fusão descrito na Equação (3) é composto por múltiplas representações cepstrais, incluindo MFCC, CQCC e outras. Embora os princípios físicos e perceptuais subjacentes a essas representações não sejam detalhados neste trabalho por concisão, o leitor interessado pode consultar obras clássicas como Davis & Mermelstein (1980) e Hermansky (1990), que fornecem modelagens aprofundadas e motivações para seu uso no processamento de sinais de fala.

Como visto na Seção 2.2, existem medidas não-cepstrais que são extremamente úteis em sistemas de reconhecimento de padrões de fala. Isso se deve ao fato de que tais medidas fornecem informações adicionais sobre características da fala que podem não ser capturadas por medidas cepstrais. Por exemplo, *jitter* e *shimmer* refletem variações de amplitude e frequência, frequentemente associadas a anomalias vocais. Outras medidas, como valores médios de formantes, intensidade e estatísticas de duração, capturam aspectos de articulação, prosódia e estrutura temporal da fala.

Portanto, propõe-se que o sinal de voz também seja representado por métricas obtidas por meio de atributos não-cepstrais. Seja $\mathcal{N}$ o conjunto composto por $n_\mathcal{N}$ métricas não-cepstrais capazes de representar características específicas do sinal de voz:

$$\mathcal{N} := \{\mu_1, \mu_2, ..., \mu_{n_\mathcal{N}}\}, \tag{4}$$

onde $\mu_i : \mathbb{R}^n \to \mathbb{R}^{n_i}, \forall i \in \{1, 2, .., n_\mathcal{N}\}$.

Define-se o vetor de características $\vec{v}_{x,\text{non-cepstral}}$ de um sinal de fala $x \in \mathbb{R}^n$ como:

$$\vec{v}_{x,\text{non-cepstral}} := (\mu_1(x), \mu_2(x), ..., \mu_{n_\mathcal{N}}(x)). \tag{5}$$

Por fim, para representar o sinal de voz com informações tanto cepstrais quanto não-cepstrais,

propõe-se o vetor de características completo $\vec{v}_x$ definido como:

$$\vec{v}_x := (\vec{v}_{x,\text{cepstral}}, \vec{v}_{x,\text{non-cepstral}}) . \tag{6}$$

O processo completo para obtenção do vetor final $\vec{v}_x$ a partir de um sinal $x \in \mathbb{R}^n$ pode ser descrito nos seguintes passos:

1. **Geração de Aumento de Dados:** Com o objetivo de aumentar a robustez das representações de características e melhorar a generalização, ruído gaussiano aditivo é aplicado a cada sinal de voz original. Para cada amostra x, um novo sinal $\tilde{x}$ é gerado por $\tilde{x} = x + \epsilon$, onde $\epsilon \sim \mathcal{N}(0, \sigma^2)$, dobrando efetivamente o tamanho do conjunto de dados. A variância do ruído σ^2 é calibrada para preservar a inteligibilidade enquanto introduz variabilidade.

2. **Extração dos CCs:** Para cada $\Phi_i \in P$, computa-se a matriz de características cepstrais: $C_x^{\Phi_i} := \Phi_i(x)$.

3. **Cálculo Diferencial:** Calculam-se as derivadas temporais de primeira e segunda ordem: $\Delta_x^{\Phi_i}$ e $\Delta\Delta_x^{\Phi_i}$.

4. **Projeção das Características:** Aplica-se uma ou mais técnicas de projeção do conjunto candidato (por exemplo, PCA, MEAN, STD, SKEW ou combinações). Esta etapa é objeto de experimentação sistemática, e as características projetadas $\vec{v}_x^{\Phi_i}$ são geradas usando cada estratégia isoladamente ou em combinação, conforme definido na Equação 3.

5. **Fusão das Representações Cepstrais:** Concatenam-se todos os vetores intermediários $\vec{v}_x^{\Phi_i}$ em um único vetor de características representando todas as técnicas cepstrais selecionadas:

$$\vec{v}_{x,\text{cepstral}} := (\vec{v}_x^{\phi_1}, \vec{v}_x^{\phi_2}, \ldots, \vec{v}_x^{\phi_{n_P}}). \tag{7}$$

6. **Fusão com Características Não Cepstrais:** Concatena-se o vetor de características cepstrais $\vec{v}_{x,\text{cepstral}}$ com o vetor de características não-cepstrais $\vec{v}_{x,\text{non-cepstral}}$, formando a representação final $\vec{v}_x$, conforme a Equação 6.

7. **Normalização dos Dados:** Padroniza-se $\vec{v}_x$ utilizando normalização *z-score* sobre o conjunto de treinamento: $z_j = (v_j - \mu_j)/\sigma_j$, assegurando média zero e variância unitária para cada dimensão.

8. **Redução Final de Dimensionalidade:** Aplica-se uma técnica final de redução de dimensionalidade ao vetor normalizado $\vec{v}_x$, produzindo uma versão mais compacta $\vec{v}_x^{\text{final}}$. Esta etapa é crucial para avaliar o equilíbrio entre dimensionalidade e desempenho do modelo, constituindo o objetivo central deste estudo.

9. **Treinamento do Classificador:** O vetor reduzido final $\vec{v}_x^{\text{final}}$ serve como entrada para um modelo de aprendizado supervisionado, treinado para discriminar entre sinais de voz autênticos (*bonafide*) e falsificados (*spoofed*). O classificador é ajustado utilizando dados rotulados $(\vec{v}_x^{\text{final}}, y_x)$, onde y_x denota o rótulo verdadeiro. Diferentes modelos de classificação, como Máquinas de Vetores de Suporte (SVM), *Random Forests* e Regressão Logística, são explorados sob condições idênticas para avaliar sua compatibilidade e desempenho em conjunto com as estratégias de redução de dimensionalidade.

O processo geral é resumido visualmente na Figura 9.

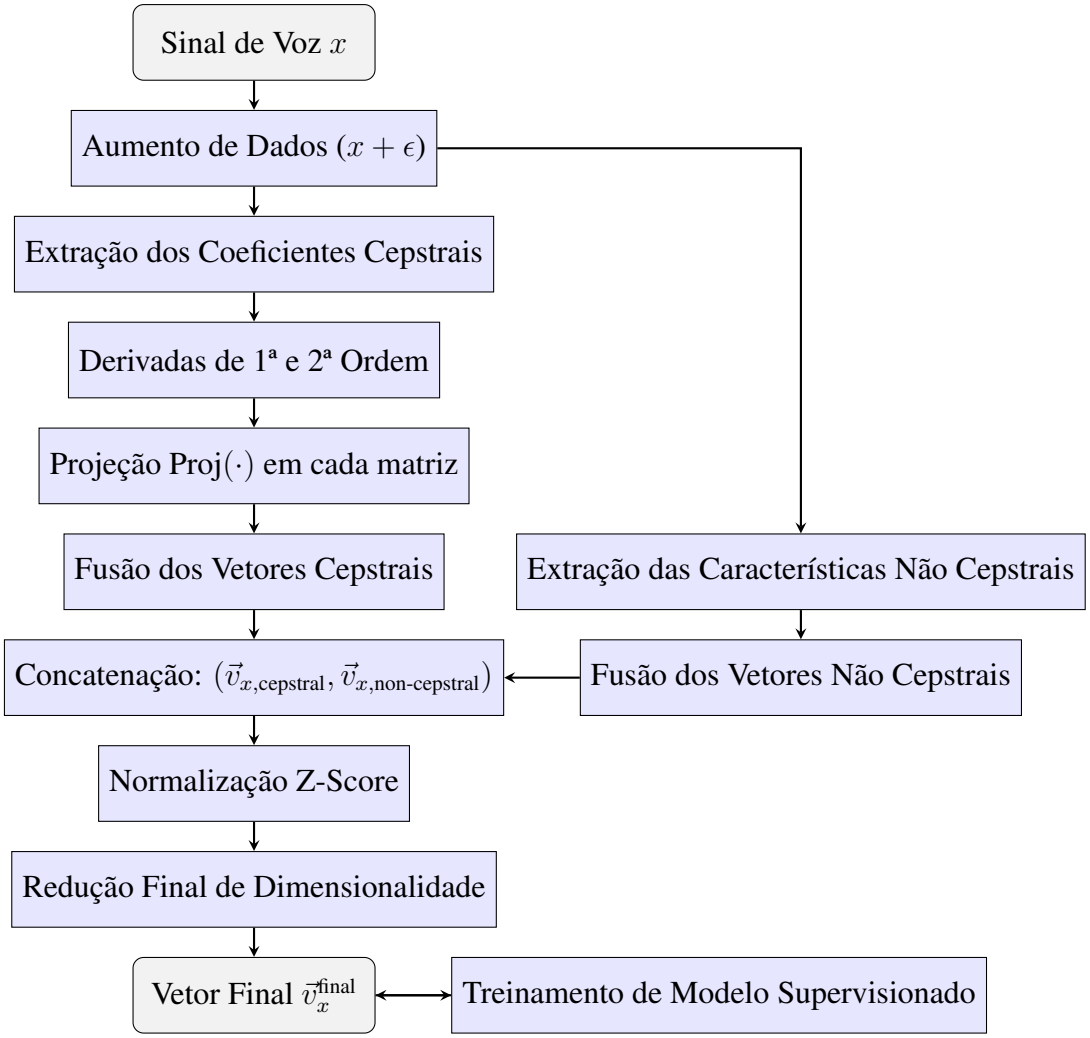

Figura 9 – Fluxograma do método de extração e representação de características com informações multicepstrais e não-cepstrais.
Fonte: Elaborado pelo autor (2025)

2.5 PARÂMETROS DO MÉTODO PROPOSTO E INSTÂNCIAS PRÁTICAS

Inicialmente, é importante enfatizar que todas as contribuições propostas por Contreras, Viana & Guido (2023) foram formuladas como estruturas generalizadas. Nesse sentido, o *pipeline* de

representação é parametrizado por um conjunto de funções de mapeamento para as características da matriz de CCs e por uma combinação configurável de técnicas de extração cepstrais. Da mesma forma, o arcabouço baseia-se em conjuntos predefinidos de características cepstrais e não-cepstrais, mecanismos de agregação de informações, procedimentos de normalização e métodos de balanceamento de dados.

Devido a essa flexibilidade, uma instância concreta do sistema deve ser definida para a execução das avaliações experimentais. Para isso, a estratégia proposta nesta tese, inspirada no trabalho citado, adota configurações propositalmente simplificadas para cada componente, com o objetivo de isolar e avaliar as contribuições individuais das estratégias de modelagem propostas, utilizando configurações acessíveis e interpretáveis. Adiante, detalham-se as técnicas específicas escolhidas para instanciar os componentes do arcabouço utilizados na fase experimental.

- **Funções de mapeamento**: No contexto deste estudo, o termo *função de mapeamento* refere-se a uma função de transformação $m : \mathbb{R}^{d \times T} \rightarrow \mathbb{R}^d$ aplicada a cada matriz de características cepstrais extraídas do sinal de voz. Essas matrizes normalmente representam uma sequência de T quadros, em que cada linha corresponde a um coeficiente cepstral. O objetivo do mapeamento é resumir a evolução temporal de cada coeficiente em um vetor de tamanho fixo. Essa etapa é crucial para converter os mapas temporais de características em um formato compatível com modelos de reconhecimento de padrões estáticos. Definimos essa etapa de projeção como a função Proj$(\cdot)$ usada na Equação (3). Para os experimentos realizados, foram consideradas quatro funções fundamentais de mapeamento, definidas como operações por coluna: (i) projeção via PCA (m_{PCA}), que reduz a dimensão temporal por meio de transformação ortogonal; (ii) média por coluna (MEAN), que resume a tendência central ao longo do tempo; (iii) desvio padrão por coluna (STD), que captura a variabilidade temporal; e (iv) assimetria por coluna (SKEW), que reflete a assimetria na distribuição temporal. Essas funções podem ser aplicadas individualmente ou combinadas em sequência para compor uma estratégia de projeção mais rica.

  Em conformidade com o conceito de multi-projeção proposto neste trabalho, foram avaliados os conjuntos de funções de mapeamento listados na Tabela 2. Cada configuração corresponde a uma estratégia diferente para projetar matrizes cepstrais em vetores compactos de características. O objetivo é comparar o potencial discriminativo de diferentes projeções e suas combinações no contexto da detecção de *spoofing* de voz.

- **Características não cepstrais:** Para enriquecer a representação do sinal de voz além das informações cepstrais, foi incorporado um conjunto abrangente de medidas não cepstrais, denotado por $\mathcal{N}$. Essas características capturam pistas acústicas, prosódicas e articulatórias complementares que podem evidenciar inconsistências ou artefatos introduzidos por ataques de *spoofing*. Com base na literatura prévia (Ladefoged; Johnson, 2014; Teixeira; Fernandes, 2014; Yang et al., 2014; Puts; Apicella; Cárdenas, 2012; Pisanski; Rendall,

Tabela 2 – Conjuntos de funções de mapeamento $\mathcal{M}$ avaliados para projeção de matrizes cepstrais.

ID	Estratégias de projeção
Proj_1	{PCA}
Proj_2	{MEAN}
Proj_3	{STD}
Proj_4	{SKEW}
Proj_5	{PCA, MEAN}
Proj_6	{MEAN, STD}
Proj_7	{STD, SKEW}
Proj_8	{MEAN, STD, SKEW}
Proj_9	{PCA, MEAN, STD, SKEW}

Fonte: Elaborada pelos autores.

2011; Little et al., 2008; Reby; McComb, 2003; Fitch, 1997), foram consideradas as seguintes métricas:

- *Estatísticas de frequência fundamental:* Média e desvio padrão de F_0, refletindo a dinâmica de entoação ao longo do enunciado.
- *Relação Harmônicos-Ruído (HNR):* Indica a razão entre a energia periódica (harmônica) e aperiódica (ruído) no sinal, frequentemente alterada por artefatos de síntese ou reprodução.
- *Características baseadas em jitter:*
  - **Jitter Local:** Mede a variabilidade ciclo a ciclo no período de entoação.
  - **Jitter Local Absoluto:** Captura desvios absolutos nas durações dos períodos de entoação.
  - **Jitter RAP:** Perturbação média relativa entre ciclos consecutivos.
  - **Projeção PCA do Jitter:** Projeção única que resume todas as características baseadas em jitter via PCA.
- *Características baseadas em shimmer:*
  - **Shimmer Local em dB:** Variação de amplitude por quadro em dB.
  - **Shimmer APQ3 / APQ5 / APQ11:** Quocientes de perturbação da amplitude calculados sobre 3, 5 e 11 ciclos adjacentes, respectivamente.
  - **Shimmer DDA:** Diferença absoluta média da amplitude entre ciclos vocais.
  - **Projeção PCA do Shimmer:** Projeção única que resume as características relacionadas ao shimmer.
- *Métricas baseadas em formantes:*
  - **Média e mediana de F_1–F_4:** Estatísticas descritivas das quatro principais frequências de formantes.

  - **Dispersão de Formantes:** Um terço da distância entre as medianas de F_4 e F_1.
  - **Média Aritmética dos Formantes:** Média dos valores medianos de F_1–F_4.
  - **Posição dos Formantes:** Média padronizada das medianas de F_1–F_4.
  - **Espaçamento de Formantes (ΔF):** Menor espaçamento entre formantes adjacentes, estimado via regressão linear.
  - **Comprimento Virtual do Trato Vocal (VTL) com base em ΔF:** Comprimento estimado do trato vocal calculado a partir do espaçamento ΔF.
  - **Frequência Média de Formantes (MFF):** Raiz quarta do produto das medianas de F_1–F_4.
  - **Comprimento Virtual do Trato Vocal de Fitch (FVTL):** Estimativa do comprimento do trato vocal derivada de modelagem espectral.

- **Técnicas de extração de coeficientes cepstrais:** Neste estudo, adotou-se o mesmo conjunto diverso de métodos de processamento de sinais para extração de representações cepstrais utilizado por Contreras, Viana & Guido (2023), os quais são amplamente empregados em tarefas de identificação de locutor e detecção de *spoofing*. As técnicas consideradas incluem: Coeficientes Cepstrais baseados em *Constant-Q* (CQCC), Coeficientes Cepstrais na Frequência Mel (MFCC), sua variante invertida (iMFCC), Coeficientes Cepstrais de Frequência Linear (LFCC) (Sahidullah; Kinnunen; Hanilçi, 2015), características baseadas em *Gammatone* (GFCC) (Valero; Alias, 2012), representações cepstrais com escala de *Bark* (BFCC) (Herrera; Rio, 2010), Coeficientes Cepstrais de Predição Linear (LPCC) (Rao; Reddy; Maity, 2015) e Coeficientes Cepstrais de *Gammachirp* Normalizados (NGCC) (Zouhir; Ouni, 2016). Para cada técnica, o processo de extração gera uma sequência de vetores de características compostos por 20 coeficientes estáticos ($n_{\text{ceps}} = 20$), juntamente com suas respectivas derivadas temporais de primeira e segunda ordem (Δ e $\Delta\Delta$), resultando em um total de 60 dimensões por quadro. Todas as características passam por normalização da média e variância cepstrais (CMVN), que padroniza a distribuição de cada coeficiente entre os quadros para média zero e variância unitária, reduzindo a variabilidade de canal e de sessão. Essas representações cepstrais são avaliadas tanto individualmente quanto em combinações predefinidas, conforme organizado na Tabela 3. O objetivo é investigar como diferentes escalas espectrais e estruturas de *filterbank* influenciam o poder discriminativo do vetor final de características quando utilizadas em conjunto com estratégias de projeção e fusão voltadas à detecção de *spoofing* de voz.

  Devido a limitações de espaço no texto, não foi possível considerar todas as combinações existentes das oito técnicas de extração de CCs. No entanto, as combinações mais relevantes foram analisadas para avaliar o desempenho do material desenvolvido. Foram consideradas as versões de $\mathcal{P}_1$ a $\mathcal{P}_8$, nas quais cada representação de $\mathcal{P}$ possui apenas um elemento. Essas versões permitiram avaliar o quanto o arcabouço está potencializando a capacidade das técnicas na detecção de disfonia em sinais de voz. Além disso, as demais

Tabela 3 – Configurações consideradas para os conjuntos de técnicas de extração de CCs utilizadas nos experimentos.

Conjunto	Técnica(s) de extração de CCs
$\mathbf{P}_1$	MFCC
$\mathbf{P}_2$	CQCC
$\mathbf{P}_3$	iMFCC
$\mathbf{P}_4$	BFCC
$\mathbf{P}_5$	LFCC
$\mathbf{P}_6$	LPCC
$\mathbf{P}_7$	GFCC
$\mathbf{P}_8$	NGCC
$\mathbf{P}_9$	CQCC, MFCC
$\mathbf{P}_{10}$	CQCC, LFCC
$\mathbf{P}_{11}$	CQCC, LPCC
$\mathbf{P}_{12}$	CQCC, GFCC
$\mathbf{P}_{13}$	CQCC, NGCC
$\mathbf{P}_{14}$	CQCC, MFCC, LFCC
$\mathbf{P}_{15}$	CQCC, MFCC, BFCC
$\mathbf{P}_{16}$	CQCC, MFCC, GFCC
$\mathbf{P}_{17}$	CQCC, MFCC, NGCC
$\mathbf{P}_{18}$	CQCC, MFCC, BFCC, NGCC
$\mathbf{P}_{19}$	CQCC, MFCC, BFCC, LFCC
$\mathbf{P}_{20}$	CQCC, MFCC, BFCC, GFCC

Fonte: Elaborada pelos autores.

versões das técnicas de extração de CCs devem servir para confirmar a capacidade do material proposto em representar diferentes características da amostra sonora, o que deve contribuir para a melhoria de sua capacidade de detectar disfonia nessas amostras.

- **Normalização:** Para avaliar o impacto do escalonamento das características na representação final $\vec{v}_x$, foram consideradas duas estratégias alternativas: normalização padrão e ausência de normalização. No primeiro caso, cada característica é padronizada para ter média zero e variância unitária no conjunto de treinamento, procedimento comumente utilizado para garantir escala uniforme e estabilidade numérica. No segundo caso, nenhuma transformação é aplicada aos valores originais, preservando a escala bruta das características. Matematicamente, o vetor padronizado $\text{NORM}_{\text{STD}}(y)$ é dado por:

$$\text{NORM}_{\text{STD}}(y) := \left(\frac{y_1 - \mu_1}{\sigma_1}, \ldots, \frac{y_m - \mu_m}{\sigma_m} \right),$$

onde μ_i e σ_i denotam, respectivamente, a média e o desvio padrão da i-ésima característica no conjunto de treinamento. A versão não normalizada é obtida aplicando-se a função identidade:

$$\text{NORM}_{\text{UnS}}(y) := y.$$

Essas duas estratégias são comparadas experimentalmente a fim de avaliar se o escalonamento influencia o desempenho da detecção de *spoofing* quando aplicado após a fusão e antes da redução de dimensionalidade.

- **Redução de dimensionalidade:** Como contribuição central deste trabalho, foi investigado como diferentes estratégias de redução de dimensionalidade impactam o poder discriminativo do vetor final fundido $\vec{v}_x$ no contexto da detecção de *spoofing*. Essa redução é aplicada após a concatenação e normalização das características cepstrais e não-cepstrais, com o objetivo de eliminar redundância, aumentar a capacidade de generalização e melhorar a eficiência computacional da etapa de classificação.

  Um total de oito técnicas foi considerado, abrangendo abordagens baseadas em projeção e em seleção:

  - **PCA:** Um método clássico de projeção linear que transforma o espaço original de características em um conjunto de componentes ortogonais ordenados pela sua capacidade de capturar a variância dos dados. Os primeiros componentes são mantidos para representar as direções mais informativas dos dados.
  - **SVD:** Semelhante ao PCA, o SVD decompõe os dados em vetores e valores singulares, retendo apenas os componentes mais significativos. Diferentemente do PCA, não requer o centramento dos dados e pode ser aplicado diretamente a matrizes esparsas ou de alta dimensionalidade.
  - **Seleção pelo valor-F da ANOVA:** Método estatístico univariado que seleciona características com base na razão entre a variância entre classes e a variância dentro das classes. As características com maior poder discriminativo, conforme medido pela estatística F, são selecionadas.
  - **Ranqueamento por Informação Mútua (MI):** Esta técnica quantifica a quantidade de informação compartilhada entre cada característica e a classe alvo. As características são ranqueadas com base em suas pontuações de informação mútua, sendo retidas as mais informativas.
  - **Eliminação Recursiva de Características (RFE):** Método *wrapper* que treina iterativamente um modelo preditivo e remove, a cada etapa, as características menos importantes. Esse processo continua até que um número desejado de características seja mantido, priorizando aquelas mais influentes para o desempenho do modelo.
  - **Seleção baseada em LASSO:** Abordagem baseada em regularização que realiza simultaneamente seleção de características e regressão. Ao aplicar uma penalização

ℓ_1 durante o treinamento do modelo, o LASSO força alguns coeficientes a zero, eliminando efetivamente características menos relevantes.

- **Importância baseada em Random Forest:** Estratégia de seleção baseada em *ensemble* que avalia a relevância das características com base na contribuição de cada uma para a acurácia de um conjunto de árvores de decisão. As características com maiores pontuações de importância acumulada são selecionadas.
- **Importância por Permutação:** Método independente de modelo que estima a importância de cada característica medindo a queda no desempenho preditivo quando os valores da característica são embaralhados aleatoriamente. Apenas as características cuja permutação leva a uma degradação significativa no desempenho são mantidas.

Todas as técnicas são aplicadas de forma independente ao vetor completo e normalizado $\vec{v}_x$, produzindo uma versão comprimida $\vec{v}_x^{\text{final}}$ que é utilizada como entrada para o classificador. Ao comparar esses métodos em diferentes níveis de redução, este estudo busca entender como a compacidade da representação influencia a robustez e a eficácia dos sistemas de detecção de *spoofing*.

Cada estratégia de redução de dimensionalidade apresenta vantagens e limitações distintas que podem influenciar o desempenho da detecção de *spoofing* de diferentes maneiras. Métodos baseados em projeção, como PCA e SVD, são particularmente eficientes em capturar a variância global e projetar os dados em um subespaço denso de menor dimensionalidade. Suas principais vantagens incluem simplicidade, rapidez e a capacidade de decorrelacionar características; no entanto, tratam-se de técnicas lineares e podem não explorar completamente relações não lineares complexas presentes no espaço de atributos. Métodos baseados em seleção, por outro lado, oferecem maior flexibilidade e frequentemente produzem modelos mais interpretáveis. ANOVA e MI dependem de associações estatísticas univariadas com a variável alvo, o que os torna rápidos e independentes de modelo, mas potencialmente limitados na captura de interações multivariadas. Métodos *wrapper*, como RFE e importância por permutação, são mais potentes para considerar tais interações, mas tendem a ser computacionalmente custosos e sensíveis à escolha do modelo subjacente. A seleção baseada em regularização via LASSO introduz esparsidade e robustez contra *overfitting*, mas pode ter desempenho inferior quando há alta correlação entre características. A importância baseada em *Random Forest* se beneficia da robustez de modelos *ensemble*, mas pode introduzir viés em favor de características com maior variabilidade ou cardinalidade.

Ao explorar este conjunto diverso de técnicas de redução de dimensionalidade, o objetivo é identificar quais métodos são mais compatíveis com as representações fundidas cepstrais e não-cepstrais utilizadas neste trabalho, e como suas propriedades intrínsecas afetam a generalização, a eficiência e a capacidade de detecção. Para leitores interessados em uma

análise técnica mais aprofundada de cada método, recomenda-se as revisões especializadas de Jia et al. (2022), Guyon & Elisseeff (2003) e Maaten et al. (2009).

- **Classificador:** Para a etapa de classificação, adotou-se um modelo de Máquina de Vetores de Suporte (SVM) com kernel Gaussiano, também conhecido como kernel de Função de Base Radial (RBF) (Noble, 2006). Essa escolha reflete sua ampla adoção em tarefas de verificação de locutor e detecção de *spoofing*, devido à sua capacidade de modelar eficazmente fronteiras de decisão não lineares. O classificador é treinado com estimativa de probabilidade de classe habilitada, o que permite extrair pontuações de probabilidade posterior para o cálculo de métricas de avaliação, como a EER. Para melhorar a estabilidade do treinamento e mitigar a sensibilidade à escala, as características de entrada são opcionalmente padronizadas antes do ajuste do modelo, utilizando média zero e variância unitária. Adicionalmente, aplica-se um mecanismo de ajuste de pesos de classe para reduzir a sensibilidade ao desbalanceamento, assegurando robustez mesmo em configurações experimentais que não incorporam estratégias de *oversampling*.

3 RESULTADOS EXPERIMENTAIS E DISCUSSÃO

Nesta seção, apresenta-se a avaliação experimental projetada para verificar o desempenho do método proposto descrito na Seção 2.4. Para isso, foi adotado um protocolo de referência descrito em detalhe na Seção 3.0.1, amplamente reconhecido na área de detecção de *spoofing* baseada em voz. Os experimentos têm como objetivo analisar a eficácia da abordagem proposta nas diversas configurações discutidas na Seção 2.5, incluindo diferentes combinações de características, projeções, normalizações e técnicas de redução. Como parte dessa avaliação, foi conduzida uma análise interna aprofundada da contribuição, explorando como diferentes fatores influenciam o desempenho. Isso inclui a avaliação de múltiplas representações cepstrais e suas estratégias de projeção, a inclusão ou omissão de características não cepstrais, o impacto da aplicação ou não de normalização e — mais criticamente — o efeito das técnicas de redução de dimensionalidade incorporadas à representação final de características. Esses experimentos são projetados para identificar quais configurações geram as representações mais compactas e discriminativas para detecção de *spoofing*.

Para comparar a eficácia das diferentes configurações, o foco foi direcionado para duas métricas principais de avaliação. A primeira é a *Acurácia* (ACC), que mede a proporção geral de instâncias corretamente classificadas. A segunda e mais importante métrica é a *Equal Error Rate* (EER), que corresponde ao ponto de operação em que a taxa de falsas aceitações (FAR) é igual à taxa de falsas rejeições (FRR). Essa métrica é particularmente relevante em contextos de segurança biométrica, pois reflete o equilíbrio entre rejeitar usuários legítimos e aceitar amostras falsas. A EER é calculada por meio da análise da curva *Receiver Operating Characteristic* (ROC), identificando o limiar τ no qual a taxa de falsos positivos (FPR) e a taxa de falsos negativos (FNR) são iguais ou o mais próximas possível. Na prática, ela é estimada encontrando o valor de FPR para o qual a diferença absoluta entre FPR e FNR é minimizada:

$$\text{EER} \approx \text{FPR}(\tau^*) \quad \text{onde} \quad \tau^* = \arg\min_{\tau} \left|\text{FPR}(\tau) - \text{FNR}(\tau)\right|. \tag{8}$$

Por fim, após a conclusão da análise interna, as configurações com melhor desempenho serão contrastadas com métodos do estado da arte reportados na literatura, permitindo contextualizar a eficácia da nossa abordagem e destacar sua relevância prática no domínio da detecção de *spoofing*. Todos os experimentos foram implementados na linguagem de programação Python versão 3.13.5[1], utilizando bibliotecas como Scikit-learn (Pedregosa et al., 2011), Spafe (Malek et al., 2023) e Parselmouth (Jadoul; Thompson; Boer, 2018), e executados em uma estação de trabalho pessoal equipada com 8 GB de RAM e um processador Intel(R) Core(TM) i5-4460 rodando a 3,20 GHz.

[1] https://www.python.org/

3.0.1 Conjunto de Dados

O conjunto de dados adotado para os experimentos é a segunda edição do *ASVspoof 2017* (Delgado et al., 2018), que compreende uma coleção diversificada de enunciados genuínos e falsificados. As gravações genuínas originam-se do corpus RedDots, capturadas por dispositivos móveis com sistema Android e compostas por dez frases secretas predefinidas. As amostras falsificadas são geradas por meio da reprodução desses enunciados originais em diferentes cenários de reprodução e gravação, simulando ataques de *spoofing* realistas sob diversas condições acústicas. Esse corpus é organizado em três partições padrão — Treinamento, Desenvolvimento e Avaliação — cada uma refletindo a variabilidade inerente introduzida por 177 sessões distintas e 61 configurações diferentes de gravação. A Tabela 4 resume as principais características do conjunto de dados e apresenta a distribuição das amostras *bona fide* e falsificadas entre os três subconjuntos.

Tabela 4 – Detalhes do Benchmark *ASVSpoof 2017 v2.0*.

Subconjunto	**Locutores**	**Sessões**	**Configurações**	**Genuínas**	**Falsificadas**
Treinamento	10	6	3	1507	1507
Desenvolvimento	8	10	10	760	950
Avaliação	24	161	57	1298	12008
Total	**42**	**177**	**70**	**3565**	**14465**

Fonte: Adaptado de (Todisco; Delgado; Evans, 2017b).

3.0.2 Análise Interna

Conforme descrito na Seção 2.5, o método proposto oferece suporte a uma ampla gama de parâmetros configuráveis, possibilitando a construção de múltiplas instâncias experimentais. Neste estudo, foram avaliadas sistematicamente milhares de combinações resultantes da variação nas técnicas de extração de características, estratégias de projeção, inclusão ou não de atributos não-cepstrais, abordagens de normalização e método final de redução de dimensionalidade. Essas variações geraram um grande volume de observações de desempenho sobre o *benchmark* selecionado para detecção de *spoofing* de voz, o que nos permitiu explorar o comportamento do sistema sob diferentes escolhas de projeto. Nesta seção, são resumidos os principais achados desse extenso processo de avaliação. Foram destacadas as configurações com melhor desempenho segundo as métricas consideradas — principalmente a EER e, secundariamente, a ACC — e apresentou-se uma interpretação estruturada dos fenômenos observados. O objetivo não é apenas identificar combinações eficazes de parâmetros, mas também caracterizar o impacto de cada componente metodológico na capacidade de detecção de *spoofing* do arcabouço proposto.

A Tabela 5 apresenta um resumo dos melhores resultados obtidos em termos de acurácia e EER, agrupados por conjunto (Desenvolvimento ou Avaliação), uso ou não de atributos não-cepstrais, e se os vetores de características foram normalizados. Para cada cenário, foi relatada

a acurácia máxima e a EER mínima obtidas entre todas as configurações testadas. Esse agrupamento permite uma compreensão geral de como a normalização e a inclusão de atributos não-cepstrais afetam o desempenho em cada condição. Os resultados revelam que o impacto da normalização e da inclusão de atributos não-cepstrais varia significativamente entre os conjuntos de Desenvolvimento e Avaliação. No conjunto de Desenvolvimento, o melhor desempenho — tanto em termos da menor EER ($10{,}32\%$) quanto da maior acurácia ($87{,}57\%$) — foi alcançado sem normalização e utilizando apenas atributos cepstrais, sugerindo que a representação não alterada preserva padrões discriminativos de forma mais eficaz nesse subconjunto. Por outro lado, a introdução de atributos não-cepstrais ou a aplicação da normalização levou a uma queda consistente no desempenho, indicando possível *overfitting* ou sensibilidade ao ruído. Entretanto, no conjunto de Avaliação, a inclusão de atributos não-cepstrais mostrou-se benéfica, resultando na menor EER ($10{,}64\%$) e maior acurácia ($85{,}58\%$), independentemente da normalização. Isso sugere que os atributos não-cepstrais contribuem positivamente para a generalização, especialmente em condições não vistas. A normalização, por sua vez, teve efeito negligenciável nos resultados do conjunto de Avaliação, já que o desempenho permaneceu inalterado com ou sem sua aplicação. Em resumo, esses achados indicam que, embora representações brutas compostas apenas por atributos cepstrais funcionem melhor na fase de desenvolvimento, representações mais ricas que incorporam pistas não-cepstrais são mais robustas na avaliação, e que a normalização oferece vantagem limitada em ambos os casos.

Tabela 5 – Melhor Acurácia (máx.) e menor EER (mín.) por configuração de avaliação.

Conjunto	Normalizado	Não-cepstrais	Acurácia (máx.)	EER (mín.)
Desenvolvimento	Não	Não	87,57	**10,32**
	Sim	Não	86,58	10,37
	Não	Sim	80,56	11,53
	Sim	Sim	79,42	12,05
Avaliação	Não	Não	83,31	11,07
	Sim	Não	83,31	11,07
	Não	Sim	85,58	**10,64**
	Sim	Sim	85,58	**10,64**

Fonte: Elaborada pelo autor.

A Figura 10 resume os menores valores de EER obtidos para cada combinação de conjuntos de atributos cepstrais, agrupados por cenário e pela presença ou ausência de atributos não-cepstrais. Os quatro cenários avaliados correspondem aos subconjuntos de *Development* e *Evaluation*, cada um testado com e sem normalização.

Na condição *Dev + Unscaled*, os melhores desempenhos são consistentemente observados para configurações que utilizam apenas atributos cepstrais, com menores EERs concentrados em técnicas como MFCC, CQCC e LFCC. Isso reforça achados anteriores de que representações mais simples e não normalizadas proporcionam melhor desempenho discriminativo em ambientes de treinamento controlados. Em contraste, no subconjunto de *Evaluation* (com e sem normalização),

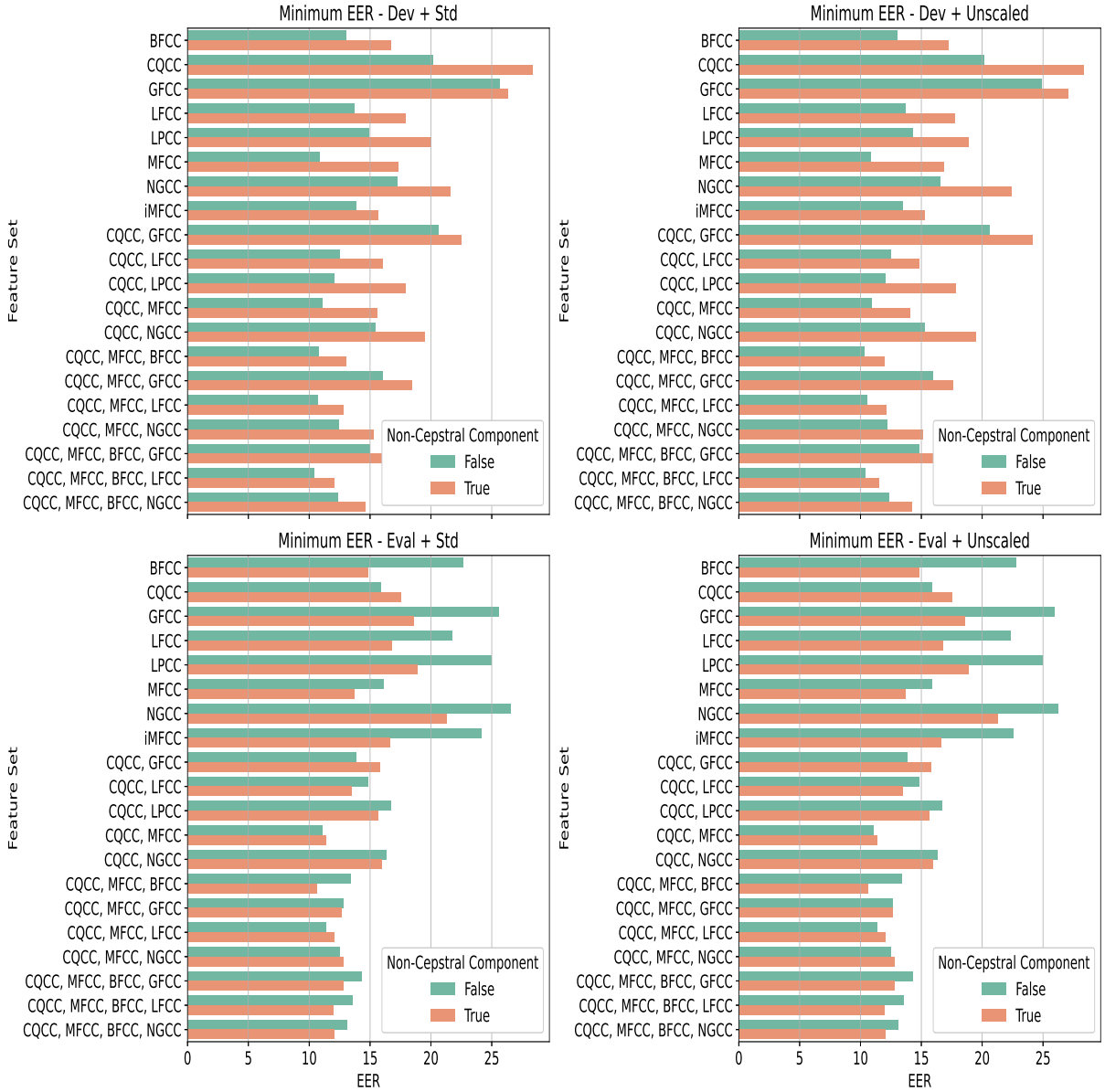

Figura 10 – Menores EERs obtidos para cada conjunto de atributos analisado, organizados conforme quatro cenários experimentais distintos. A configuração considera o tipo de conjunto de avaliação utilizado (*development* ou *evaluation*) e se a normalização dos dados foi aplicada. As cores das barras indicam a presença ou ausência de componentes não-cepstrais.

Fonte: Elaborado pelo autor (2025)

a adição de atributos não-cepstrais resulta em uma redução perceptível no EER, sugerindo que esses atributos contribuem positivamente para a generalização — especialmente quando o modelo é exposto a condições não vistas ou variações de gravação. O impacto da normalização em si mostrou-se desprezível, uma vez que valores de EER similares são observados nas variantes normalizadas e não normalizadas dentro de cada grupo.

Além disso, vale destacar que combinações mais complexas de técnicas cepstrais (aquelas que envolvem três ou quatro métodos) não necessariamente produzem os melhores resultados. Em

muitos casos, configurações mais enxutas superam as mais elaboradas, indicando que adicionar informações redundantes pode não melhorar — e até prejudicar — a capacidade discriminativa do modelo. De forma geral, essa visualização corrobora a interpretação de que representações somente cepstrais e não normalizadas são mais eficazes durante o treinamento, enquanto a integração de atributos não-cepstrais aumenta a robustez durante a avaliação, com influência mínima da normalização. Esses achados destacam a importância de avaliar os modelos tanto nos conjuntos de *Dev* quanto de *Eval* para garantir a generalização dos padrões observados.

Em vez de enumerar cada configuração individual, o foco foi direcionado na descrição dos padrões mais salientes e generalizáveis observados nos *heatmaps*, de forma a garantir concisão sem comprometer a profundidade analítica. Dessa forma, a Figura 11 fornece uma visão detalhada dos menores valores de EER observados para cada combinação de conjuntos de coeficientes cepstrais e estratégias de projeção, em quatro condições experimentais: subconjuntos de *Development* e *Evaluation*, com e sem normalização. Os *heatmaps* revelam padrões consistentes quanto à influência tanto da composição dos atributos quanto das técnicas de sumarização das matrizes.

A Figura 11 revela múltiplos *insights* sobre as interações entre estratégias de projeção e representações de atributos. No cenário *Dev* + *Unscaled*, os menores valores de EER concentram-se em torno de técnicas cepstrais tradicionais como MFCC e CQCC quando combinadas com projeções elementares como MEAN ou STD. Isso reforça achados anteriores que sugerem que, em ambientes semelhantes ao treinamento, representações mais simples e sem normalização tendem a preservar a estrutura discriminativa dos dados de forma mais eficaz. No entanto, à medida que se avança para os cenários *Eval*, especialmente *Eval* + *Norm*, uma mudança se torna evidente: combinações mais ricas — especialmente aquelas que envolvem estratégias de multi-projeção (por exemplo, MEAN + STD ou PCA + MEAN) — começam a superar as mais simples, particularmente quando aplicadas sobre conjuntos cepstrais mais amplos. Notavelmente, embora PCA isolado não seja dominante, sua combinação com outras projeções apresenta resultados competitivos em *Eval*. Além disso, conjuntos cepstrais híbridos como ‘CQCC, MFCC‘ ou ‘CQCC, LFCC‘ mostram melhor generalização quando acoplados a projeções múltiplas. O impacto da normalização, mais uma vez, parece limitado — modulando levemente o desempenho em alguns casos, mas sem alterar o ranking global das combinações. Esses achados sugerem que, enquanto representações brutas e mínimas são ideais para otimização em domínio (Dev), projeções aprimoradas e conjuntos cepstrais compostos oferecem robustez em contextos de avaliação com maior variabilidade.

A Tabela 6 apresenta um resumo comparativo de oito estratégias de redução de dimensionalidade e seleção de atributos aplicadas ao vetor final de atributos, incluindo métodos baseados em projeção (por exemplo, PCA, SVD) e métodos baseados em seleção (por exemplo, LASSO, RFE, RF). As métricas reportadas são a média, o desvio padrão e o menor EER (%) observados em todas as configurações avaliadas sob quatro condições experimentais: subconjuntos *Dev* e *Eval*, com e sem normalização padrão.

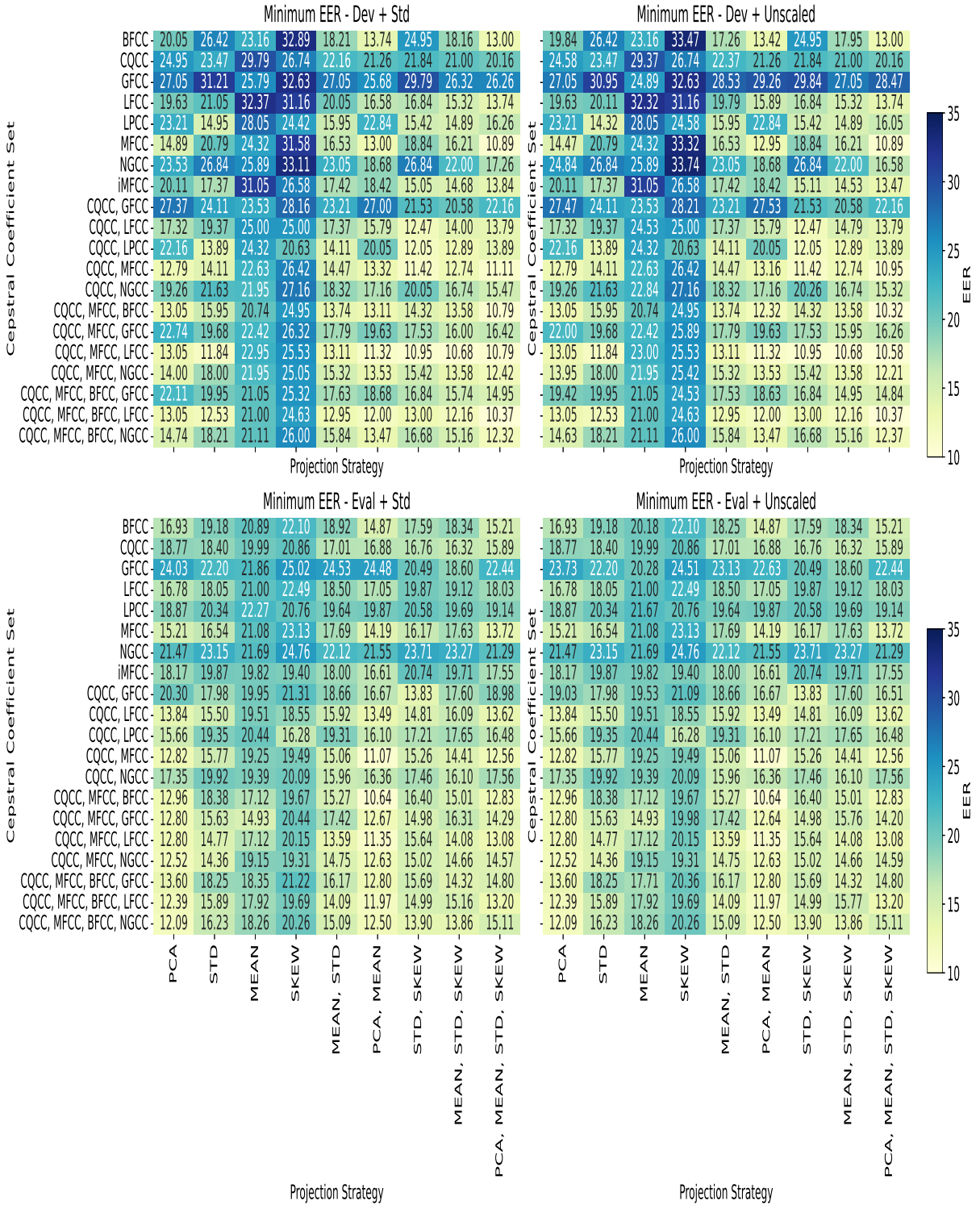

Figura 11 – *Heatmaps* mostrando os menores valores de EER para cada combinação de estratégia de projeção e método de redução de dimensionalidade. Cada subfigura corresponde a um cenário de avaliação diferente. Esta figura permite a identificação das combinações de métodos mais eficazes dentro de cada condição experimental.

Fonte: Elaborado pelo autor (2025)

Em todos os cenários, os melhores resultados estão consistentemente associados ao uso da regularização LASSO, que atinge o menor EER médio tanto em *Dev* quanto em *Eval*,

Tabela 6 – Comparação das técnicas de redução de dimensionalidade e seleção de atributos com base na média, mínimo e desvio padrão do EER (%) em todas as configurações.

Cenário	Técnica	Média do EER	Desvio Padrão	Mínimo do EER
Dev + Std	*LASSO*	22.31	6.39	10.89
	PCA	23.67	5.78	13.21
	SVD	24.57	6.66	11.63
	RFE	24.79	6.55	10.37
	PIS	27.99	9.41	10.53
	RF	28.69	7.46	10.79
	MI	31.12	8.18	13.63
	ANOVA	31.27	7.76	13.74
Dev + Unscaled	*LASSO*	22.31	6.39	10.89
	PCA	22.83	6.76	10.32
	SVD	23.74	7.04	10.63
	RFE	24.80	6.56	10.37
	PIS	27.99	9.41	10.53
	RF	28.63	7.34	10.58
	MI	31.13	8.17	13.63
	ANOVA	31.27	7.76	13.74
Eval + Std	*LASSO*	21.08	5.94	11.35
	PIS	22.05	6.37	10.64
	RFE	22.79	6.02	12.76
	PCA	23.58	5.90	13.39
	SVD	24.03	5.59	13.82
	RF	25.86	4.67	15.90
	MI	28.84	4.90	18.10
	ANOVA	29.71	5.12	18.15
Eval + Unscaled	*LASSO*	21.08	5.94	11.35
	PIS	22.05	6.36	10.64
	PCA	22.39	5.98	12.64
	RFE	22.79	6.02	12.76
	SVD	23.49	5.46	13.80
	RF	25.97	4.76	16.76
	MI	28.83	4.91	17.30
	ANOVA	29.71	5.12	18.15

Fonte: Elaborada pelo autor.

independentemente da normalização. Também é notável que LASSO atinge o menor desvio padrão, sugerindo alta estabilidade. PCA e RFE também demonstram desempenho competitivo, particularmente no conjunto *Dev* sem escala, onde PCA atinge o menor EER absoluto (10.32%). Em contraste, técnicas baseadas em ranqueamento univariado — ANOVA e *Mutual Information* — apresentam os maiores EERs médios e mínimos mais altos, indicando menor efetividade geral para a tarefa. Interessantemente, PIS e RF obtêm bons mínimos de EER, próximos das melhores configurações (por exemplo, 10.53%, 10.58%), mas com variâncias significativamente maiores,

sugerindo menor consistência entre os experimentos. Esses resultados reforçam que métodos embutidos como LASSO e RFE são mais adequados para cenários com alta dimensionalidade e atributos mistos, como o proposto neste trabalho. Adicionalmente, a semelhança geral entre variantes normalizadas e não normalizadas confirma que a normalização tem um papel limitado nesta etapa final do fluxo, em consonância com os achados anteriores.

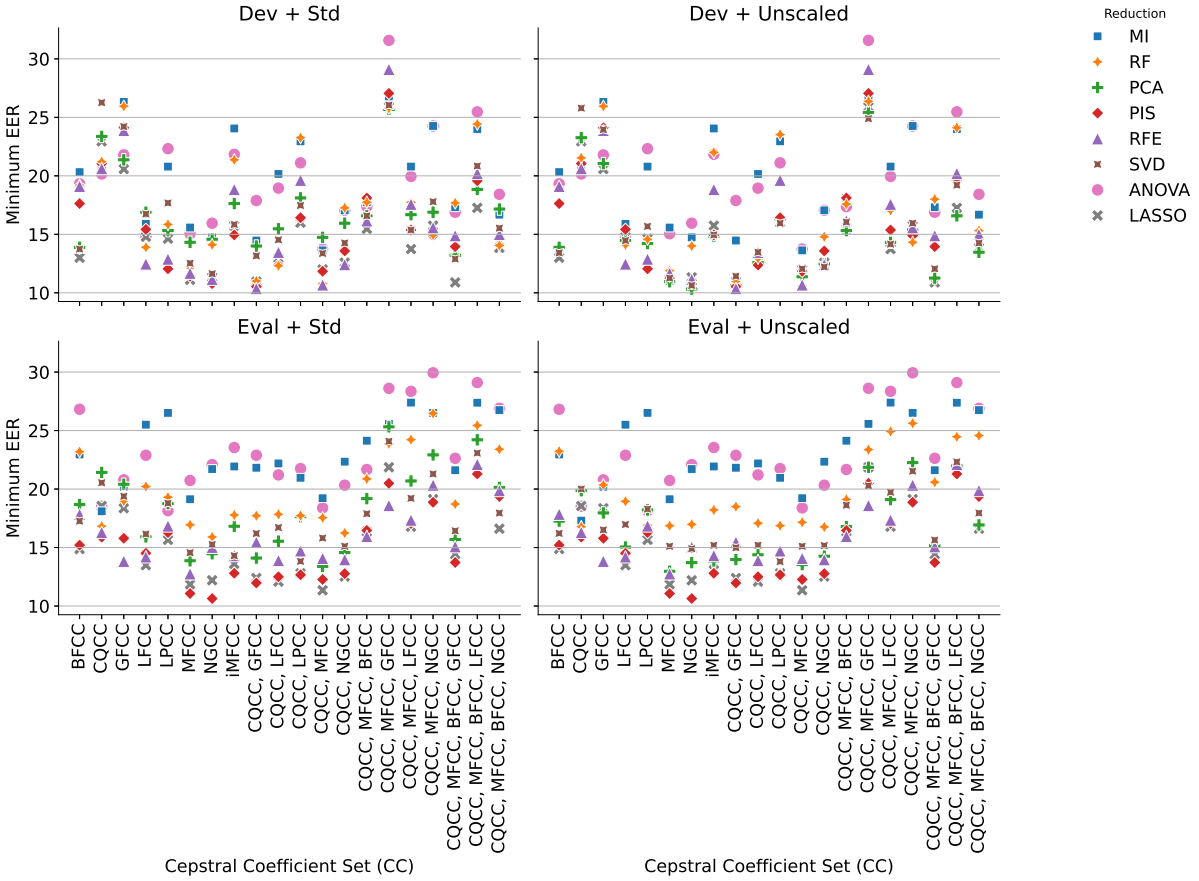

Figura 12 – EER obtido para cada conjunto de CC, considerando diferentes técnicas de redução de dimensionalidade em quatro cenários experimentais. Cada subfigura corresponde a uma configuração específica de conjunto de avaliação e normalização dos dados. Cores e formas identificam a técnica de redução utilizada. O gráfico destaca a influência de cada método sob diferentes condições, permitindo a comparação da robustez e consistência.

Fonte: Elaborado pelo autor (2025)

A Figura 12 permite uma comparação detalhada de como cada técnica de redução de dimensionalidade se comporta em diferentes conjuntos de coeficientes cepstrais e cenários. Uma inspeção visual revela que a seleção baseada em ANOVA (representada por círculos rosas) tende a resultar nos maiores valores de EER, independentemente do conjunto cepstral ou cenário, indicando que técnicas de seleção univariadas não são adequadas para este problema. Por outro lado, RFE (representado por triângulos roxos) apresenta desempenho competitivo nos cenários de *Development*, frequentemente ocupando a faixa inferior de EER, especialmente quando combinada com MFCC, CQCC ou suas combinações. Isso sugere que a eliminação recursiva é

eficaz para ajuste de desempenho em condições de treinamento. Em contraste, PIS (representado por losangos vermelhos) se destaca nos conjuntos de *Evaluation*, alcançando repetidamente baixos valores de EER, especialmente quando combinado com conjuntos cepstrais mais ricos. Isso implica que PIS proporciona melhor generalização em ambientes de teste mais diversos, provavelmente devido à sua seleção de atributos consciente do modelo. Adicionalmente, métodos baseados em LASSO (marcadores $\times$ cinzas) aparentam ser estáveis em todos os cenários, mantendo EER relativamente baixos e com menor variância, corroborando as observações anteriores da tabela-resumo. A estrutura em facetas torna evidente que, embora algumas estratégias de redução se destaquem em *Dev*, suas vantagens nem sempre se mantêm em *Eval*. Essa visualização reforça a necessidade de avaliar técnicas de redução não apenas pelo desempenho em treinamento, mas também por sua estabilidade e capacidade de generalização sob condições de avaliação.

A Figura 13 apresenta *boxplots* ilustrando a distribuição dos valores de *EER* obtidos por cada técnica de redução de dimensionalidade nos quatro cenários experimentais. Essas visualizações destacam não apenas a tendência central (medianas), mas também a variabilidade e a presença de *outliers* para cada método em cada condição.

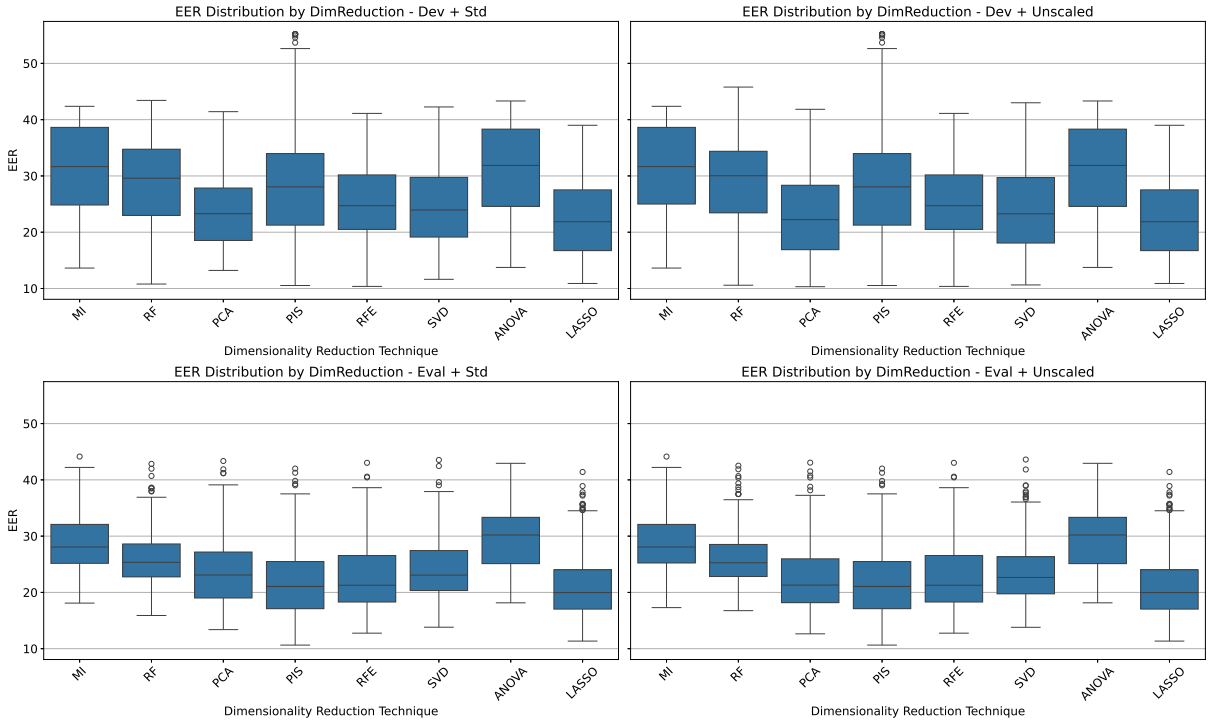

Figura 13 – EER para cada técnica de redução de dimensionalidade, analisado separadamente por cenário experimental. Cada *boxplot* resume a tendência central, dispersão e *outliers* dos valores de EER em diferentes combinações de projeção e conjuntos de atributos.

Fonte: Elaborado pelo autor (2025)

Em ambos os cenários de *Development*, LASSO exibe a distribuição mais compacta, com medianas consistentemente baixas e variância reduzida, confirmando sua estabilidade já obser-

vada. RFE também apresenta medianas competitivas em *Dev*, embora com uma dispersão um pouco maior. Interessantemente, PCA e SVD, apesar de serem métodos clássicos baseados em projeção, mostram desempenho moderado com maior variabilidade, indicando sensibilidade à configuração dos atributos e possivelmente robustez limitada. Por outro lado, os cenários de *Evaluation* revelam uma mudança de comportamento. PIS apresenta uma das menores medianas de EER em *Eval* — particularmente na condição sem normalização — mantendo ainda assim uma faixa interquartil relativamente estreita. Isso corrobora o achado anterior de que PIS generaliza bem para dados não vistos. LASSO permanece entre os melhores também em *Eval*, validando ainda mais sua consistência em diferentes configurações. Técnicas como ANOVA e *Mutual Information* exibem consistentemente tanto medianas mais altas quanto maior dispersão em todos os cenários, reforçando sua menor efetividade e maior instabilidade nesta aplicação. Notavelmente, a seleção baseada em *Random Forest*, embora competitiva em alguns casos isolados, mostra grande variabilidade no desempenho de EER, tornando-a menos previsível como estratégia geral. De forma geral, os *boxplots* corroboram os resultados quantitativos anteriores e enfatizam que LASSO e PIS são as opções mais equilibradas, oferecendo tanto precisão competitiva quanto comportamento estável entre os diferentes cenários de avaliação. Por outro lado, filtros univariados (ANOVA, MI) continuam sendo as escolhas menos confiáveis no contexto de redução de dimensionalidade para detecção de *spoofing*.

3.0.3 Comparação com o Estado da Arte

Para avaliar se as representações multicepstrais de atributos adotadas no *framework* proposto são competitivas em relação às soluções de estado da arte, conduzimos uma análise comparativa entre os melhores resultados obtidos pela abordagem proposta e aqueles reportados em publicações recentes que avaliam modelos no conjunto de dados "*ASVSpoof 2017 v2.0*". As Tabelas 7 e 8 resumem os menores valores de EER relatados pelos principais métodos da literatura, considerando tanto o conjunto de desenvolvimento — quando aplicável — quanto o conjunto de avaliação do *benchmark*. Adicionalmente, a última coluna apresenta a média dessas duas métricas, fornecendo uma medida agregada de desempenho para fins comparativos.

Como mostrado nas Tabelas 7 e 8, o método proposto alcança resultados competitivos quando comparado a uma ampla gama de abordagens recentes de estado da arte aplicadas ao corpus *ASVSpoof 2017 v2.0*. Em termos de EER, foram alcançados valores de 10.32% no conjunto de desenvolvimento e 10.64% no de avaliação, resultando em uma média de 10.48%, posicionando o método proposto entre as técnicas de melhor desempenho no geral, especialmente considerando que muitas das abordagens alternativas dependem de modelos profundos, combinações artesanais ou classificadores em *ensemble*.

Embora algumas abordagens, como aquelas baseadas em *ensembles*, tais como CQCC + CFCCIF + ESA, ou híbridos cepstrais ajustados, tais como CQCC + QCFCCIF - ESA, apresentem médias de EER mais baixas, é importante destacar que várias delas apresentam alta variância entre os resultados de desenvolvimento e avaliação, sugerindo uma possível falta de

Tabela 7 – Comparação do método proposto com abordagens de detecção de *spoofing* de estado da arte no conjunto de dados *ASVSpoof 2017 v2.0*. A tabela reporta os menores valores de EER (%) obtidos por cada método nos subconjuntos de desenvolvimento e avaliação, juntamente com a média entre eles.

Algorithm	**EER (%)**		
	Dev	**Eval**	**Mean**
Proposed Method	10.32	10.64	10.48
2D-ILRCC (Jelil; Sinha; Prasanna, 2024)	11.90	10.87	11.38
iMFCC/SCMC (Chutia; Bhattacharjee, 2024)	4.83	11.49	8.16
OMSp-HASC-Canny (Meriem; Messaoud; Bahia, 2024)	3.52	22.86	13.19
CQCC, MFCC (PCA) (Contreras et al., 2024)	-	17.42	17.42
CQCC, MFCC, LFCC (Contreras et al., 2023)	-	15.02	15.02
LFCC+CQCC+LTAS (Neamtu; Mihalache; Burileanu, 2023)	8.60	26.10	17.35
SWM-CQCC-DA (Li; Xia, 2023)	-	10.27	10.27
LCNN/smallCNN (Chettri, 2023)	7.37	10.70	9.03
CQCC+CFCC+CFCCIF+CFCCIF-ESA/QESA (Gupta; Chodingala; Patil, 2023)	1.88	10.99	6.43
LFDCC/CDOC (Xu et al., 2023)	13.03	11.63	12.33
CQCC + QCFCCIF-ESA (Gupta; Chodingala; Patil, 2022)	2.19	11.00	6.59
CQCC + JS (Woubie; Bäckström, 2022)	7.60	10.90	9.25
EMD-HS-RFCC+CQCC+APGDF (Bharath; Kumar, 2022)	4.89	10.84	7.86
ResNet/LCNN (Süslü; Eren; Demiroğlu, 2022)	7.85	8.20	8.02
CQCC+LFCC+MFCC+u-vector (Patil et al., 2022)	6.66	11.99	9.32
TECC (Kamble; Patil, 2021)	10.80	11.41	11.10
CQCC + AT-IMFCC + AT-MelRP (Liu et al., 2021b)	2.08	10.75	6.41

Fonte: Elaborada pelo autor.

Tabela 8 – Continuação: comparação do método proposto com abordagens de detecção de *spoofing* de estado da arte no conjunto de dados *ASVSpoof 2017 v2.0*. A tabela reporta os menores valores de EER (%) obtidos por cada método nos subconjuntos de desenvolvimento e avaliação, juntamente com a média entre eles.

Algorithm	**EER (%)**		
	Dev	**Eval**	**Mean**
Proposed Method	10.32	10.64	10.48
(LPCC+LFCC)-GMM (Jana et al., 2021)	5.26	22.65	13.95
CQCCs+AFCCsFAF+ARPDBF (Liu et al., 2021a)	1.72	11.22	6.47
Multicepstral features + ResNet/GMM (Avila et al., 2021)	7.57	11.64	9.60
CQCC-CNN (Chettri; Benetos; Sturm, 2020)	7.37	10.70	9.03
TDNN+2PLDA - CQCC+MFCC (Li et al., 2020)	-	11.16	11.16
CQSPIC (Das; Yang; Li, 2020)	-	11.34	11.34
ESA-IACC-IFCC (Kamble; Tak; Patil, 2020)	7.03	10.12	8.57
GMM+CQCC-MP_{aware} DNN + (MFCC+MGDCC) (Phapatanaburi et al., 2020)	5.86	24.14	15.00
CQCC (90-D (SDA)) - AWFCC (Kamble; Patil, 2020a)	5.75	10.42	8.08
LFCC (Patil; Patil, 2020)	7.02	14.83	10.92
(eCQCC-STSSI)-DA (Yang; Das, 2020)	-	10.07	10.07
CVOC-DA (Yang et al., 2020)	-	11.46	11.46
SFCC-QCN (Lin et al., 2020)	8.38	10.11	9.24
Spectrogram-CNN (Chettri; Kinnunen; Benetos, 2020)	10.82	16.03	13.42
CQCC+VTECC (Kamble; Patil, 2020b)	5.85	10.94	8.39
CQNCC-D DNN (Yang; Das, 2019)	-	10.31	10.31
CQCC + PNCC (Tapkir et al., 2018)	8.73	12.98	10.85
CQCC-CMVN (Delgado et al., 2018)	9.06	13.74	11.40

Fonte: Elaborada pelo autor .

generalização. Por exemplo, OMSp-HASC-Canny demonstra degradação significativa nos dados de avaliação. Em contraste, o método proposto mantém um desempenho equilibrado em ambas as partições, com uma diferença negligenciável entre *Dev* e *Eval*. Tal resultado reforça a robustez da estratégia baseada em multicepstrais e projeções adotada neste estudo, indicando que a fusão de atributos cepstrais e não cepstrais, combinada com um *pipeline* sistemático de redução de dimensionalidade, pode alcançar desempenho equivalente ao de arquiteturas mais complexas, ao mesmo tempo em que oferece transparência e flexibilidade para adaptação futura a diferentes cenários de *spoofing*.

4 CONCLUSÕES E TRABALHOS FUTUROS

Garantir uma detecção robusta de *spoofing* em sistemas biométricos de voz é um passo crítico para fortalecer a segurança de estruturas de autenticação. Neste estudo, foi proposto e avaliado sistematicamente um *framework* configurável e extensível que integra representações multicepstrais, características acústicas não cepstrais e um conjunto diverso de técnicas de redução de dimensionalidade. O objetivo foi analisar o impacto desses componentes sobre a EER sob diferentes condições de normalização e avaliação, utilizando o *benchmark ASVSpoof 2017 v2.0*.

Nossos resultados revelam que nenhuma combinação única de técnicas supera universalmente as demais; em vez disso, o desempenho depende do cenário. No entanto, algumas tendências foram evidentes: a redução de dimensionalidade baseada em LASSO demonstrou consistentemente forte generalização com baixa EER e baixa variância, enquanto filtros univariados como ANOVA e *Mutual Information* tiveram desempenho inferior em todos os casos. PIS mostrou-se particularmente eficaz nos subconjuntos de avaliação, sugerindo adaptabilidade superior a dados não vistos. Além disso, RFE apresentou resultados competitivos em cenários de desenvolvimento. Esses resultados enfatizam que a fusão de características multicepstrais e não cepstrais, quando combinada com estratégias adequadas de projeção e redução, pode igualar ou superar *pipelines* de *deep learning* mais complexos, oferecendo maior transparência e flexibilidade.

O método proposto alcançou uma das menores EERs médias entre os trabalhos recentes da literatura, reforçando sua relevância em aplicações práticas de segurança biométrica. O *framework* também permite fácil adaptação e experimentação, possibilitando a investigação de novas combinações sem a necessidade de reengenharia de todo o *pipeline*.

Com base nesses resultados, diversas direções de pesquisa emergem. Primeiramente, trabalhos futuros podem explorar a inclusão de características adicionais não cepstrais, como representações baseadas em fase ou em articuladores, que podem oferecer poder discriminativo complementar. Em segundo lugar, dado os resultados promissores de PIS e LASSO, seria valioso avaliar estratégias híbridas de redução que combinem importância baseada em modelo e restrições de regularização.

Outro caminho promissor envolve a integração de representações de textura sonora derivadas da análise de espectrogramas, inspiradas na detecção de *spoofing* baseada em textura em outras modalidades biométricas, como impressões digitais. Técnicas como *Local Binary Patterns* ou filtros de *Gabor* aplicados sobre representações tempo-frequência podem codificar distorções perceptuais causadas por *replay attacks*. Além disso, uma vez que este estudo focou em modelos rasos e interpretáveis, trabalhos futuros podem envolver a comparação desses *pipelines* artesanais com variantes explicáveis de modelos profundos, avaliando o balanço entre transparência e desempenho bruto.

Por fim, considerando a efetividade da otimização baseada em algoritmos genéticos em tra-

balhos anteriores, pretende-se investigar técnicas avançadas de metaheurística, como operadores massivos de busca local e esquemas evolutivos híbridos, com o objetivo de refinar a fase de redução de dimensionalidade e aprimorar ainda mais o desempenho da detecção em condições complexas e adversariais.

REFERÊNCIAS

ABDEL-HAMID, O. et al. Convolutional neural networks for speech recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 22, n. 10, p. 1533–1545, 2014.

ALIM, S. A.; RASHID, N. K. A. **Some commonly used speech feature extraction algorithms**. [S.l.]: IntechOpen London, UK:, 2018.

AVILA, A. R. et al. On the use of blind channel response estimation and a residual neural network to detect physical access attacks to speaker verification systems. **Computer Speech & Language**, Elsevier, v. 66, p. 101163, 2021.

BELLMAN, R. **Adaptive Control Processes: A Guided Tour**. [S.l.]: Princeton University Press, 1961.

BEYER, K. et al. When is "nearest neighbor" meaningful? In: **Proceedings of the 7th International Conference on Database Theory**. [S.l.: s.n.], 1999.

BHARATH, K.; KUMAR, M. R. Replay spoof detection for speaker verification system using magnitude-phase-instantaneous frequency and energy features. **Multimedia Tools and Applications**, Springer, v. 81, n. 27, p. 39343–39366, 2022.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

CAI, D.; CAI, W.; LI, M. Within-sample variability-invariant loss for robust speaker recognition under noisy environments. In: **ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2020. p. 6469–6473.

CAMPBELL, J. Speaker recognition: a tutorial. **Proceedings of the IEEE**, v. 85, n. 9, p. 1437–1462, 1997.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.]: Association for Computing Machinery, 2016. p. 785–794.

CHETTRI, B. The clever hans effect in voice spoofing detection. In: IEEE. **2022 IEEE Spoken Language Technology Workshop (SLT)**. [S.l.], 2023. p. 577–584.

CHETTRI, B.; BENETOS, E.; STURM, B. L. Dataset artefacts in anti-spoofing systems: a case study on the asvspoof 2017 benchmark. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 28, p. 3018–3028, 2020.

CHETTRI, B.; KINNUNEN, T.; BENETOS, E. Deep generative variational autoencoding for replay spoof detection in automatic speaker verification. **Computer Speech & Language**, Elsevier, v. 63, p. 101092, 2020.

CHUTIA, B. J.; BHATTACHARJEE, U. Effectiveness of different spectral features in replay attack detection-an experimental study. **Grenze International Journal of Engineering & Technology (GIJET)**, v. 10, 2024.

CONTRERAS, R. C. et al. Dimensionality reduction in multicepstral features for voice spoofing detection: Case studies with singular value decomposition, genetic algorithm, and auto-encoder. In: SPRINGER. **International Conference on Artificial Intelligence and Soft Computing**. [S.l.], 2025. p. 227–244.

CONTRERAS, R. C. et al. Metaheuristic algorithms for enhancing multicepstral representation in voice spoofing detection: An experimental approach. In: SPRINGER. **International Conference on Swarm Intelligence**. [S.l.], 2024. p. 247–262.

CONTRERAS, R. C.; VIANA, M. S.; GUIDO, R. C. An experimental analysis on mapping strategies for cepstral coefficients multi-projection in voice spoofing detection problem. In: RUTKOWSKI, L. et al. (Ed.). **Artificial Intelligence and Soft Computing - 22nd International Conference, ICAISC 2023, Zakopane, Poland, June 18-22, 2023, Proceedings, Part II**. Springer, 2023. (Lecture Notes in Computer Science, v. 14126), p. 291–306. Disponível em: https://doi.org/10.1007/978-3-031-42508-0_27.

CONTRERAS , R. C. et al. An experimental analysis on multicepstral projection representation strategies for dysphonia detection. **Sensors**, v. 23, n. 11, 2023. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/23/11/5196.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

DAS, R. K.; YANG, J.; LI, H. Assessing the scope of generalized countermeasures for anti-spoofing. In: IEEE. **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.], 2020. p. 6589–6593.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE transactions on acoustics, speech, and signal processing**, IEEE, v. 28, n. 4, p. 357–366, 1980.

DELGADO, H. et al. Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements. In: ISCA. **The Speaker and Language Recognition Workshop**. [S.l.], 2018. p. 296–303.

DELLER, J. R.; PROAKIS, J. G.; HANSEN, J. H. **Discrete-time processing of speech signals**. [S.l.]: John Wiley & Sons, 2010.

FISHER, A.; RUDIN, C.; DOMINICI, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. **Journal of Machine Learning Research**, v. 20, n. 177, p. 1–81, 2019.

FITCH, W. T. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 102, n. 2, p. 1213–1222, 1997.

GOBL, C.; CHASAIDE, A. N. The role of voice quality in communicating emotion, mood and attitude. **Speech Communication**, Elsevier, v. 40, n. 1, p. 189–212, 2003.

GODINO-LLORENTE, J. I.; GóMEZ-VILDA, P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 51, n. 2, p. 380–384, 2004.

GOLUB, G. H.; LOAN, C. F. V. **Matrix computations**. [S.l.]: JHU press, 2013.

GOMEZ-ALANIS, A. et al. A gated recurrent convolutional neural network for robust spoofing detection. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, 2019.

GORBAN, A. N.; MAKAROV, V. A.; TYUKIN, I. Y. High-dimensional brain in a high-dimensional world: Blessing of dimensionality. **Entropy**, MDPI, v. 22, n. 1, p. 82, 2020.

GUIDO, R. C. A tutorial on signal energy and its applications. **Neurocomputing**, Elsevier, v. 179, p. 264–282, 2016.

GUIDO, R. C. Zcr-aided neurocomputing: a study with applications. **Knowledge-based Systems**, Elsevier, v. 105, p. 248–269, 2016.

GUIDO, R. C. Enhancing teager energy operator based on a novel and appealing concept: signal mass. **Journal of the Franklin Institute**, Elsevier, v. 356(4), p. 1341–1354, 2018.

GUIDO, R. C. A tutorial-review on entropy-based handcrafted feature extraction for information fusion. **Information Fusion**, Elsevier, v. 41, p. 161–175, 2018.

GUPTA, P.; CHODINGALA, P. K.; PATIL, H. A. Energy separation based instantaneous frequency estimation from quadrature and in-phase components for replay spoof detection. In: IEEE. **2022 30th European Signal Processing Conference (EUSIPCO)**. [S.l.], 2022. p. 369–373.

GUPTA, P.; CHODINGALA, P. K.; PATIL, H. A. Replay spoof detection using energy separation based instantaneous frequency estimation from quadrature and in-phase components. **Computer Speech & Language**, Elsevier, v. 77, p. 101423, 2023.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157–1182, 2003.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Machine learning**, v. 46, n. 1, p. 389–422, 2002.

HANILÇI, C. Data selection for i-vector based automatic speaker verification anti-spoofing. **Digital Signal Processing**, Elsevier, v. 72, p. 171–180, 2018.

HERMANSKY, H. Perceptual linear predictive (plp) analysis of speech. **the Journal of the Acoustical Society of America**, Acoustical Society of America, v. 87, n. 4, p. 1738–1752, 1990.

HERRERA, A.; RIO, F. D. Frequency bark cepstral coefficients extraction for speech analysis by synthesis. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 128, n. 4, p. 2290–2290, 2010.

HIMAWAN, I. et al. Voice presentation attack detection using convolutional neural networks. In: **Handbook of Biometric Anti-Spoofing**. [S.l.]: Springer, 2019. p. 391–415.

HIMAWAN, I. et al. Deep domain adaptation for anti-spoofing in speaker verification systems. **Computer Speech & Language**, Elsevier, 2019.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, v. 9, n. 8, p. 1735–1780, 1997.

HUANG, L.; PUN, C.-M. Audio replay spoof attack detection using segment-based hybrid feature and densenet-lstm network. In: IEEE. **ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.], 2019. p. 2567–2571.

JADOUL, Y.; THOMPSON, B.; BOER, B. D. Introducing parselmouth: A python interface to praat. **Journal of Phonetics**, Elsevier, v. 71, p. 1–15, 2018.

JANA, S. et al. Replay attack detection for speaker verification using different features level fusion system. In: IEEE. **2021 Innovations in Power and Advanced Computing Technologies (i-PACT)**. [S.l.], 2021. p. 1–5.

JELIL, S.; SINHA, R.; PRASANNA, S. M. Spectro-temporally compressed source features for replay attack detection. **IEEE Signal Processing Letters**, IEEE, 2024.

JIA, W. et al. Feature dimensionality reduction: a review. **Complex & Intelligent Systems**, Springer, v. 8, n. 3, p. 2663–2693, 2022.

JOLLIFFE, I. T. **Principal component analysis**. [S.l.]: Springer, 2002.

KAMBLE, M. R.; PATIL, H. A. Combination of amplitude and frequency modulation features for presentation attack detection. **Journal of Signal Processing Systems**, Springer, v. 92, p. 777–791, 2020.

KAMBLE, M. R.; PATIL, H. A. Novel variable length teager energy profiles for replay spoof detection. **energy**, v. 32, p. 33, 2020.

KAMBLE, M. R.; PATIL, H. A. Detection of replay spoof speech using teager energy feature cues. **Computer Speech & Language**, Elsevier, v. 65, p. 101140, 2021.

KAMBLE, M. R.; TAK, H.; PATIL, H. A. Amplitude and frequency modulation-based features for detection of replay spoof speech. **Speech Communication**, Elsevier, v. 125, p. 114–127, 2020.

KHAN, A. et al. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. **Artificial Intelligence Review**, Springer, v. 56, n. Suppl 1, p. 513–566, 2023.

KIM, J. et al. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2021. p. 5530–5540.

KINNUNEN, T. et al. Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. **Training**, v. 10, n. 1508, p. 1508, 2017.

KINNUNEN, T. et al. t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. **arXiv preprint arXiv:1804.09618**, 2018.

KINNUNEN, T. et al. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. ISCA (the International Speech Communication Association), 2017.

KINNUNEN, T. et al. Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2195–2210, 2020.

KINNUNEN, T. et al. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In: **2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2012. p. 4401–4404.

KO, T. et al. Audio augmentation for speech recognition. In: **Interspeech**. [S.l.: s.n.], 2015. p. 3586–3589.

KO, T. et al. A study on data augmentation of reverberant speech for robust speech recognition. In: **2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 5220–5224.

LADEFOGED, P.; JOHNSON, K. **A course in phonetics**. [S.l.]: Cengage learning, 2014.

LAI, C.-L.; CHEN, K.-Y.; LEE, H.-y. Assert: Anti-spoofing with squeeze-excitation and residual networks. In: **Interspeech**. [S.l.: s.n.], 2019. p. 1018–1022.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, p. 436–444, 2015.

LI, J. et al. Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss. **IEEE Access**, IEEE, v. 8, p. 7907–7915, 2020.

LI, W.; XIA, M. Discriminative feature extraction based on swm for playback attack detection. In: IEEE. **2023 IEEE World Conference on Applied Intelligence and Computing (AIC)**. [S.l.], 2023. p. 763–767.

LIN, L. et al. A robust method for speech replay attack detection. **KSII Transactions on Internet and Information Systems (TIIS)**, Korean Society for Internet Information, v. 14, n. 1, p. 168–182, 2020.

LITTLE, M. et al. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. **Nature Precedings**, Nature Publishing Group UK London, p. 1–1, 2008.

LIU, M. et al. Replay attack detection using variable-frequency resolution phase and magnitude features. **Computer Speech & Language**, Elsevier, v. 66, p. 101161, 2021.

LIU, M. et al. Replay-attack detection using features with adaptive spectro-temporal resolution. In: IEEE. **ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.], 2021. p. 6374–6378.

LIU, X.; SAHIDULLAH, M.; KINNUNEN, T. Parameterized channel normalization for far-field deep speaker verification. **2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**, p. 1132–1138, 2021.

MAATEN, L. V. D. et al. Dimensionality reduction: A comparative review. **Journal of machine learning research**, v. 10, n. 66-71, p. 13, 2009.

MALEK, A. et al. **SuperKogito-spafe: v0.3.2**. Zenodo, 2023. Disponível em: https://doi.org/10.5281/zenodo.7686438.

MEMON, Q. et al. Audio-visual biometric authentication for secured access into personal devices. In: **Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science**. [S.l.: s.n.], 2017. p. 85–89.

MERIEM, F.; MESSAOUD, B.; BAHIA, Y.-z. Texture analysis of edge mapped audio spectrogram for spoofing attack detection. **Multimedia Tools and Applications**, Springer, v. 83, n. 6, p. 15915–15937, 2024.

MON, K. Z. et al. Spoof detection using voice contribution on lfcc features and resnet-34. In: IEEE. **2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)**. [S.l.], 2023. p. 1–6.

MUDA, L.; BEGAM, M.; ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. **Journal of Computing**, v. 2, n. 3, p. 138–143, 2010.

NEAMTU, C.-T.; MIHALACHE, S.; BURILEANU, D. Liveness detection–automatic classification of spontaneous and pre-recorded speech for biometric applications. In: IEEE. **2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)**. [S.l.], 2023. p. 1–5.

NOBLE, W. S. What is a support vector machine? **Nature Biotechnology**, Nature Publishing Group, v. 24, n. 12, p. 1565–1567, 2006.

PARK, D. S. et al. Specaugment: A simple data augmentation method for automatic speech recognition. In: **Interspeech**. [S.l.: s.n.], 2019. p. 2613–2617.

PATEL, T. B.; PATIL, H. A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In: **Sixteenth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2015.

PATIL, A. T.; PATIL, H. A. Significance of cmvn for replay spoof detection. In: IEEE. **2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)**. [S.l.], 2020. p. 532–537.

PATIL, H. A. et al. Non-cepstral uncertainty vector for replay spoofed speech detection. In: IEEE. **2022 30th European Signal Processing Conference (EUSIPCO)**. [S.l.], 2022. p. 374–378.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PELECANOS, J.; SRIDHARAN, S. Feature warping for robust speaker verification. In: **A Speaker Odyssey**. [S.l.: s.n.], 2001. p. 213–218.

PHAPATANABURI, K. et al. Exploiting magnitude and phase aware deep neural network for replay attack detection. **ECTI Transactions on Electrical Engineering, Electronics, and Communications**, v. 18, n. 2, p. 89–97, 2020.

PISANSKI, K.; RENDALL, D. The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. **The Journal of the Acoustical Society of America**, Acoustical Society of America, v. 129, n. 4, p. 2201–2212, 2011.

POINCARé, H. **La Science et l'Hypothèse**. [S.l.]: Flammarion, 1905.

PRABAKARAN, D.; SHYAMALA, R. A review on performance of voice feature extraction techniques. In: IEEE. **2019 3rd International Conference on Computing and Communications Technologies (ICCCT)**. [S.l.], 2019. p. 221–231.

PURWINS, H. et al. Deep learning for audio signal processing. **IEEE Journal of Selected Topics in Signal Processing**, v. 13, n. 2, p. 206–219, 2019.

PUTS, D. A.; APICELLA, C. L.; CÁRDENAS, R. A. Masculine voices signal men's threat potential in forager and industrial societies. **Proceedings of the Royal Society B: Biological Sciences**, The Royal Society, v. 279, n. 1728, p. 601–609, 2012.

QIAN, Y.; CHEN, N.; YU, K. Deep features for automatic spoofing detection. **Speech Communication**, Elsevier, v. 85, p. 43–52, 2016.

RAO, K. S.; REDDY, V. R.; MAITY, S. **Language identification using spectral and prosodic features**. [S.l.]: Springer, 2015.

REBY, D.; MCCOMB, K. Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. **Animal behaviour**, Elsevier, v. 65, n. 3, p. 519–530, 2003.

REYNOLDS, D. A. An overview of automatic speaker recognition technology. **IEEE International Conference on Acoustics, Speech, and Signal Processing**, IEEE, v. 4, p. 4072–4075, 2002.

SAHIDULLAH, M. et al. Integrated spoofing countermeasures and automatic speaker verification: An evaluation on asvspoof 2015. In: **Proceedings of the 17th Annual Conference ot the International Speech Communication Association**. [S.l.: s.n.], 2016. p. 1700–1704.

SAHIDULLAH, M.; KINNUNEN, T.; HANILÇI, C. A comparison of features for synthetic speech detection. In: **Sixteenth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2015.

SENK, C.; DOTZLER, F. Biometric authentication as a service for enterprise identity management deployment: a data protection perspective. In: IEEE. **2011 Sixth International Conference on Availability, Reliability and Security**. [S.l.], 2011. p. 43–50.

SHAHEED, K. et al. Deep learning techniques for biometric security: A systematic review of presentation attack detection systems. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 129, p. 107569, 2024.

SHEN, J. et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: **2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2018. p. 4779–4783.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.

SKOWRONSKI, M.; HARRIS, J. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. **The Journal of the Acoustical Society of America**, v. 116, n. 3, p. 1774–80, 2004.

SÜSLÜ, Ç.; EREN, E.; DEMIROĞLU, C. Uncertainty assessment for detection of spoofing attacks to speaker verification systems using a bayesian approach. **Speech Communication**, Elsevier, v. 137, p. 44–51, 2022.

TAK, H. et al. End-to-end anti-spoofing with rawnet2. In: **ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2021. p. 6369–6373.

TAPKIR, P. A. et al. Replay spoof detection using power function based features. In: IEEE. **2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)**. [S.l.], 2018. p. 1019–1023.

TEIXEIRA, J. P.; FERNANDES, P. O. Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices. **Procedia technology**, Elsevier, v. 16, p. 1228–1237, 2014.

TIAN, X. et al. Spoofing speech detection using temporal convolutional neural network. In: **2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)**. [S.l.: s.n.], 2016. p. 1–6.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

TODA, T.; BLACK, A. W.; TOKUDA, K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 15, n. 8, p. 2222–2235, 2007.

TODISCO, M.; DELGADO, H.; EVANS, N. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. **Computer Speech & Language**, Elsevier, v. 45, p. 516–535, 2017.

TODISCO, M.; DELGADO, H.; EVANS, N. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. **Computer Speech & Language**, Elsevier, v. 45, p. 516–535, 2017.

TODISCO, M.; DELGADO, H.-J.; EVANS, N. W. D. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In: **Odyssey**. [S.l.: s.n.], 2016. p. 283–290.

TODISCO, M. et al. Asvspoof 2019: Future horizons in spoofed and fake audio detection. **arXiv preprint arXiv:1904.05441**, 2019.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. **IEEE Transactions on Speech and Audio Processing**, v. 10, n. 5, p. 293–302, 2002.

ULYANOV, D.; VEDALDI, A.; LEMPITSKY, V. Instance normalization: The missing ingredient for fast stylization. **ArXiv**, abs/1607.08022, 2016. Disponível em: https://api.semanticscholar.org/CorpusID:16516553.

VALERO, X.; ALIAS, F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. **IEEE Transactions on Multimedia**, IEEE, v. 14, n. 6, p. 1684–1689, 2012.

VIIKKI, O.; LAURILA, K. Cepstral domain segmental feature vector normalization for noise robust speech recognition. **Speech Communication**, v. 25, n. 1, p. 133–147, 1998. ISSN 0167-6393.

VILLALBA, J.; LLEIDA, E. Speaker verification performance degradation against spoofing and tampering attacks. In: **FALA 10 Workshop, 2010**. [S.l.: s.n.], 2010. p. 131–134.

WALL, M. E.; RECHTSTEINER, A.; ROCHA, L. M. Singular value decomposition and principal component analysis. In: **A practical approach to microarray data analysis**. [S.l.]: Springer, 2003. p. 91–109.

WANG, X. et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. **Computer Speech & Language**, Elsevier, v. 64, p. 101114, 2020.

WOUBIE, A.; BÄCKSTRÖM, T. Voice quality features for replay attack detection. In: IEEE. **2022 30th European Signal Processing Conference (EUSIPCO)**. [S.l.], 2022. p. 384–388.

WU, Z. et al. Spoofing and countermeasures for speaker verification: A survey. **Speech Communication**, Elsevier, v. 66, p. 130–153, 2015.

WU, Z. et al. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: **Sixteenth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2015.

XU, L. et al. Device features based on linear transformation with parallel training data for replay speech detection. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, v. 31, p. 1574–1586, 2023.

YAMAGISHI, J. et al. Asvspoof: the automatic speaker verification spoofing and countermeasures challenge. **IEEE Journal of Selected Topics in Signal Processing**, IEEE, v. 11, n. 4, p. 588–604, 2017.

YAMAGISHI, J. et al. **ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan**. 2019. https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf. Acesso em: 15 jan. 2019.

YAMAGISHI, J. et al. Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. 2019.

YAMAGISHI, J. et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. **arXiv preprint arXiv:2109.00537**, 2021.

YANG, J.; DAS, R. K. Low frequency frame-wise normalization over constant-q transform for playback speech detection. **Digital Signal Processing**, Elsevier, v. 89, p. 30–39, 2019.

YANG, J.; DAS, R. K. Improving anti-spoofing with octave spectrum and short-term spectral statistics information. **Applied Acoustics**, Elsevier, v. 157, p. 107017, 2020.

YANG, J. et al. Discriminative features based on modified log magnitude spectrum for playback speech detection. **EURASIP Journal on Audio, Speech, and Music Processing**, Springer, v. 2020, p. 1–14, 2020.

YANG, S. et al. Effective dysphonia detection using feature dimension reduction and kernel density estimation for patients with parkinson's disease. **PloS one**, Public Library of Science San Francisco, USA, v. 9, n. 2, p. e88825, 2014.

YU, L. D. D. **Automatic Speech Recognition: A Deep Learning Approach**. [S.l.]: Springer, 2015.

YUDIN, O. et al. Speaker's voice recognition methods in high-level interference conditions. In: IEEE. **2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)**. [S.l.], 2019. p. 851–854.

ZEN, H.; TOKUDA, K.; BLACK, A. W. Statistical parametric speech synthesis. **Speech Communication**, v. 51, n. 11, p. 1039–1064, 2009. ISSN 0167-6393.

ZOUHIR, Y.; OUNI, K. Feature extraction method for improving speech recognition in noisy environments. **J. Comput. Sci.**, v. 12, n. 2, p. 56–61, 2016.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Leonardo Mendes de Souza Data nasc. 16/08/1984
Nacionalidade	Brasileira
Nome em citações bibliográficas	SOUZA, L. M. SOUZA, LEONARDO MENDES
Currículo Lattes	http://lattes.cnpq.br/5498910423038831
ORCID	https://orcid.org/0009-0006-3536-2177
Outro identificador	https://www.linkedin.com/in/leomendes84/
FORMAÇÃO ACADÊMICA	
2002/2005	Graduação em Ciência da Computação. Centro Universitário do Norte Paulista, UNORP, Brasil.
2009/2010	Mestrado em Física Computacional Universidade de São Paulo, USP, Brasil.
PRODUÇÃO BIBLIOGRÁFICA	
SOUZA, LEONARDO MENDES DE ; GUIDO, R.C. ; CONTRERAS, RODRIGO COLNAGO ; VIANA, MONIQUE SIMPLICIO ; BONGARTI, MARCELO ADRIANO DOS SANTOS. Improving Voice Spoofing Detection Through Extensive Analysis of Multicepstral Feature Reduction. Sensors, v. 25, p. 4821, 2025.	
LACERDA, MICHEL ALVES ; GUIDO, RODRIGO CAPOBIANCO ; DE SOUZA, LEONARDO MENDES ; ZULATO, PAULO RICARDO FRANCHI ; RIBEIRO, JUSSARA ; CHEN, SHI-HUANG . A WAVELET-BASED SPEAKER VERIFICATION ALGORITHM. International Journal of Wavelets Multiresolution and Information Processing, v. 08, p. 905-912, 2010.	
SOUZA, L. M.; FONSECA, E. S. ; GUIDO, R. C. . Predição Linear e Transformada Wavelet Discreta para Identificação de Patologias na Voz. INFOIP: Revista Online de Informática, Inovação e Pesquisa (FATEC, Rio Preto), v. 2, p. 3, 2010.	
SOUZA, L. M.; GUIDO, RODRIGO CAPOBIANCO ; CHEN, SHI-HUANG ; JUNIOR, SYLVIO BARBON ; SOUZA, LEONARDO MENDES ; VIEIRA, LUCIMAR SASSO ; RODRIGUES, LUCIENE CAVALCANTI ; ESCOLA, JOAO PAULO LEMOS ; ZULATO, PAULO RICARDO FRANCHI ; LACERDA, MICHEL ALVES ; RIBEIRO, JUSSARA . On the Determination of Epsilon during Discriminative GMM Training. In: IEEE International Symposium on Multimedia (IEEE ISM 2010), 2010, Taichung. Anais do IEEE ISM 2010, 2010. p. 362-364.	

GUIDO, R. C. ; BARBON JUNIOR, S. ; VIEIRA, L. S. ; ESCOLA, J. P. ; FANTINATO, P. C. ; RODRIGUES, L. C. ; ALMEIDA, L. O. B. ; SLAETS, J. F. W. . Autenticação biométrica por comando de voz com implementação em hardware dedicado. In: Workshop no Ambito do Projeto - Ambientes para co-projeto de hardware/software em plataformas FPGAS com aplicação em robótica móvel, 2008, São Carlos. Ambientes para co-projeto de hardware/software em plataformas FPGAS com aplicação em robótica móvel, 2008. v. 2.

SOUZA, L. M.; GUIDO, R. C. ; BARBON JUNIOR, S. ; VIEIRA, L. S. ; FANTINATO, P. C. ; SANCHEZ, F. L. ; ESCOLA, J. P. ; BOLOGNINI, E. J. ; ZULATO, P. R. F. ; RIBEIRO, J. ; LACERDA, M. A. . Speech compression based on the Discrete Shapelet Transform. In: CNMAC - Congresso nacional de matemática aplicada e computacional, 2008, Recife. CNMAC - Congresso nacional de matemática aplicada e computacional. Recife: CNMAC, 2008.

FANTINATO, PAULO CÉSAR ; GUIDO, RODRIGO CAPOBIANCO ; CHEN, SHI-HUANG ; SANTOS, BRUNO LEONARDO SILVEIRA ; VIEIRA, LUCIMAR SASSO ; BARBON JÚNIOR, SYLVIO ; RODRIGUES, LUCIENE CAVALCANTI ; SANCHEZ, FABRÍCIO LOPES ; ESCOLA, JO?AO PAULO LEMOS ; SOUZA, LEONARDO MENDES ; MACIEL, CARLOS DIAS ; SCALASSARA, PAULO ROGÉRIO ; PEREIRA, JOSÉ CARLOS . A Fractal-Based Approach for Speech Segmentation. In: 2008 Tenth IEEE International Symposium on Multimedia (ISM) (Formerly MSE), 2008, Berkeley. 2008 Tenth IEEE International Symposium on Multimedia, 2008. v. 1. p. 551-555.

PARTICIPAÇÃO EM BANCAS E ORIENTAÇÕES

Bancas de trabalhos de conclusão

SILVA, D. D.; DEZANI, R.; SOUZA, L. M.. Participação em banca de Valéria Teresinha Rodrigues. Sistema de gestão para o planejamento e controle financeiro e de estoque para organizações de pequeno porte. 2013. Monografia (Aperfeiçoamento/Especialização em Consultoria Web) - Faculdade de Tecnologia de São José do Rio Preto.

DEZANI, Henrique; SOUZA, L. M.; RODRIGUES, L. C.. Participação em banca de Guilherme Trovo de Souza. Desenvolvimento de um framework para geração de formulários JSP. 2011. Monografia (Aperfeiçoamento/Especialização em Consultoria Web) - Faculdade de Tecnologia de São José do Rio Preto.

CATELANI, M.; PARDO, M. H. S.; SOUZA, L. M.. Participação em banca de Igor Pavan de Souza.Prestação de serviços em contratos públicos: análise da melhoria do processo de contratação como forma de redução de custos. 2019. Trabalho de Conclusão de Curso (Graduação em Informática para Negócios) - Faculdade de Tecnologia de São José do Rio Preto.

CATELANI, M.; LUVIZARI-MURAD, L.; SOUZA, L. M.. Participação em banca de Henrique Augusto Trajano Oliveira.Conectando sudos e deficientes auditivos aos intérpretes. 2018. Trabalho de Conclusão de Curso (Graduação em Informática para Negócios) - Faculdade de Tecnologia de São José do Rio Preto.

CARDIA NETO, J. B.; MARTINS, R. R.; SOUZA, L. M.. Participação em banca de Heitor Soares Nogueira; Rafael de Freitas.O crescimento do mercado de e-sports. 2018. Trabalho de Conclusão de Curso (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva.

BARROS, E. S.; SOUZA, L. M.; SATORIO, K.. Participação em banca de José Ribamar Souza Neto; Fernando Vieira Fragoso.Contingência da Informação: Uma Análise Quantitativa e Qualitativa de Empresas da Cidade de Catanduva. 2017. Trabalho de Conclusão de Curso (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva.
BARROS, E. S.; SOUZA, L. M.; SATORIO, K.. Participação em banca de Dionatan Batista Brancalhone; Edinaldo Aparecido Ribeiro.Política de Segurança da Informação: Um Panorama Geral Sobre as Empresas da Cidade de Catanduva, Interior de São Paulo. 2017. Trabalho de Conclusão de Curso (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva.
MARTINS, R. R.; SOUZA, L. M.; CAMARGO, L. S. A.. Participação em banca de Renan Rocco; Valdeir Gomes Armario.Estudo de Solução de UMT em ambientes de TI para pequenas e médias empresas. 2017. Trabalho de Conclusão de Curso (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva.
CAMARGO, L. S. A.; SOUZA, L. M.; CARDIA NETO, J. B.. Participação em banca de Mateus Obozovsky Gregorato - Victor Torres de Souza.Acessibilidade e tecnologia assistiva: Uma abordagem de auxílio para usuários com necessidades especiais. 2017. Trabalho de Conclusão de Curso (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva.
SOUZA, L. M.; GRANDI, M. A.; CARDIA NETO, J. B.. Participação em banca de Carlos Ivo dos Reis.Análise do impacto das franquias de internet banda larga nas classes sociais de baixa renda: Estudo de caso na Fatec Catanduva. 2017. Trabalho de Conclusão de Curso (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva.
FACCIOLI, R. A.; SOUZA, L. M.; SANCHES, A. M.. Participação em banca de Léo Rodrigues Biscassi.Estudo de Caso Sobre Aplicação de Algoritmos Evolutivo para Predição de Est.. 2015. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Faculdade Barretos.
SANCHES, A. M.; SOUZA, L. M.; FACCIOLI, R. A.. Participação em banca de Elton Marques da Silva.JSF e Primefaces do Desenvolvimento Java Web: Estudo de Caso - Aplicação Web Vigilancia Social. 2015. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Faculdade Barretos.
BARBON JUNIOR, S.; SOUZA, L. M.; RODRIGUES, L. C.. Participação em banca de Marcela Ciani de Souza.Sistema de Informação baseado em Código de Resposta Rápida - QR CODE. 2011. Trabalho de Conclusão de Curso (Graduação em Informática para Negócios) - Faculdade de Tecnologia de São José do Rio Preto.
Orientações
REIS, C. I. P. Análise do impacto das franquias de internet banda larga nas classes sociais de baixa renda: Estudo de caso na Fatec Catanduva. 2017. Trabalho de Conclusão de Curso. (Graduação em Gestão da Tecnologia da Informação) - Faculdade de Tecnologia de Catanduva. Orientador: Leonardo Mendes de Souza.

PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS
SOUZA, L. M.; VIEIRA, L. S. ; GUIDO, R. C. ; BARBON JUNIOR, S. ; SANCHEZ, F. L. ; SERGIO, K. I. C. ; GUILHERME, M. A. B. ; FANTINATO, P. C. ; BASSI, R. D. S. ; FONSECA, E. S. . On the use of wavelets for voice morphing. In: CNMAC - Congresso nacional de matemática aplicada e computacional, 2007, Florianópolis. CNMAC, 2007.
Semana de Tecnologia - FATEC.Apoc - Sistema Programado para Celular. 2006. (Oficina).
VI Semana da Computação.Semana da Computação. 2004. (Outra).