Audio and Voice Deepfake Detection: a Survey on Cross-Dataset Generalization Bottlenecks

João Victor F. de Castro[1,5], **Pedro Luiz Seabra Rocha**[1,5], **Jean P. de O. Lima**[4,5], **Antônio João G. de Azambuja**[2,5], **Leonardo B. Oliveira**[3,5], **Anderson da Silva Soares**[1,5]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – Goiás – Brazil

[2]Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – São Paulo – Brazil

[3]Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – Minas Gerais – Brazil

[4]Universidade Federal do Mato Grosso (UFMT)
Várzea Grande – Mato Grosso – Brazil

[5]Centro de Excelência em Inteligência Artificial (CEIA) – UFG
Goiânia – Goiás – Brazil

joao_castro@discente.ufg.br, pedro_rocha@discente.ufg.br,
jeanphelipedeoliveira@gmail.com, azambuja@ita.br,
leob@dcc.ufmg.br, anderson@inf.ufg.br

Abstract. *Synthetic speech generation has become accessible enough to enable fraud and disinformation, raising the need for robust audio deepfake detectors. Following PRISMA 2020, this paper surveys the cross-dataset generalization bottlenecks of audio and voice deepfake detectors, with 37 included studies. The studies that decompose the performance drop attribute it mainly to distribution difference between corpora rather than to intrinsic target difficulty: in-domain Equal Error Rate near 1% degrades to 27–34% out-of-distribution. Self-supervised front-ends and data diversification are the strategies most consistently associated with improvements across the included studies, though none resolves the bottleneck alone. We further show that comparability is undermined by protocol and metric inconsistencies, and outline an agenda toward standardized cross-corpus protocols.*

Resumo. *A geração de fala sintética tornou-se acessível o suficiente para viabilizar fraudes e desinformação, intensificando a necessidade de detectores robustos de deepfake de áudio. Este artigo apresenta uma revisão conduzida segundo o protocolo PRISMA 2020 sobre as limitações de generalização cross-dataset em detectores de deepfake de áudio e voz, com 37 estudos incluídos. Os estudos que decompõem a queda de desempenho atribuem-na sobretudo à diferença de distribuição entre corpora, e não à dificuldade intrínseca do alvo: a Taxa de Erro de Igualdade (in-domain) próxima de 1% degrada para 27%–34% em conjuntos out-of-distribution. Os front-ends auto-supervisionados e a diversificação de dados são as estratégias mais consistentemente associadas a*

ganhos entre os estudos incluídos, embora nenhuma resolva o gargalo isoladamente. Mostra-se ainda que a comparabilidade é comprometida por inconsistências de protocolos e métricas, e propõe-se uma agenda para protocolos cross-corpus padronizados.

1. Introdução

O avanço da síntese e da conversão de voz tornou a fala sintética difícil de distinguir da fala humana em determinadas condições de escuta, canal e duração do sinal. Modelos de *text-to-speech* (TTS), *voice conversion* (VC) e, mais recentemente, *audio language models* (ALMs) baseados em *codecs* neurais ampliam a superfície de ataque para fraudes, falsificação de identidade e desinformação [Yi et al. 2023]. Em resposta, a comunidade desenvolveu contramedidas (*countermeasures*, CMs), impulsionadas pelo desafio ASVspoof [Yamagishi et al. 2021] e por arquiteturas como AASIST [Jung et al. 2022] e *front-ends* auto-supervisionados como o wav2vec 2.0 [Tak et al. 2022].

Apesar do desempenho expressivo *in-dataset*, acumula-se evidência de que esses detectores generalizam mal: modelos com *Equal Error Rate* (EER) inferior a 2% no treino sofrem degradação por vezes superior a 1000% relativos quando avaliados em gravações *in-the-wild*, ataques não vistos ou áudio submetido a outros canais e *codecs* [Müller et al. 2022]. Esse gargalo *cross-dataset* (ou *out-of-distribution*, OOD) é hoje um dos principais obstáculos à implantação confiável de detectores [Müller et al. 2022, Müller et al. 2024, Khan et al. 2023], e o tema é fragmentado: as contribuições dispersam-se entre fatores causais, estratégias de mitigação e protocolos, e a avaliação *cross-corpus* sistemática permanece rara e não unificada [Khan et al. 2023, Ranjan et al. 2023, Sali et al. 2023].

Revisões anteriores já trataram o campo, e deste trabalho as separam quatro aspectos. [Yi et al. 2023] oferecem panorama abrangente, mas tratam a generalização como um tópico entre vários; [Khan et al. 2023] realizam revisão e comparação empírica com avaliação de generalizabilidade — o trabalho mais próximo —, porém com recorte anterior à consolidação dos *front-ends* SSL e dos ataques por *codecs* neurais; [Ranjan et al. 2023] sintetizam o campo em *survey track*; e [Sahidullah et al. 2023] cobrem *presentation attack detection* em capítulo de manual. Frente a eles, esta revisão toma o gargalo *cross-dataset* como *objeto* e não como capítulo; adota protocolo PRISMA documentado, onde as demais são narrativas; estende a janela até 2026, incorporando *codecs* neurais, ALMs, *partial fake* e voz cantada; e dedica um eixo analítico próprio às práticas de avaliação.

A revisão organiza-se em quatro perguntas: **RQ1** — quais fatores explicam a queda *cross-dataset*/OOD; **RQ2** — quais abordagens foram propostas e qual a evidência de sua eficácia; **RQ3** — como a generalização é avaliada e quais inconsistências comprometem a comparabilidade; **RQ4** — quais lacunas e direções emergem. As contribuições são: (i) protocolo de busca e seleção rastreável, de 995 registros a 37 estudos (Seção 3); (ii) mapeamento dos fatores de degradação e das estratégias de mitigação, por eixo mutuamente exclusivo (Seção 4); (iii) análise crítica das práticas de avaliação (Seção 5); e (iv) agenda de direções futuras (Seção 6).

2. Fundamentação Teórica

Por uniformidade, empregam-se em português *contramedida* (*countermeasure*) e *características extraídas manualmente* (*handcrafted features*); mantêm-se em itálico os termos sem tradução consolidada (*bona fide*, *cross-dataset*, *cross-corpus*, *out-of-distribution*, *in-the-wild*, *front-end*, *back-end*, *spoofing*); e usa-se *corpus*/*corpora* para as coleções de fala, reservando *dataset* aos compostos consagrados e aos nomes próprios dos conjuntos.

Entende-se por *deepfake* de áudio a fala sintética ou manipulada gerada por aprendizado profundo com o intuito de passar por fala autêntica (*bona fide*). Os ataques agrupam-se em *Logical Access* (LA — TTS e VC) e *Physical Access* (PA — *replay*) [Yamagishi et al. 2021], a que se somam o *partially fake audio* e os ataques por *codecs* neurais e ALMs. A detecção é formulada como classificação binária por uma contramedida independente do ASV, com um *front-end* de extração de características — de LFCC e CQCC à forma de onda bruta ou a modelos auto-supervisionados (SSL) como wav2vec 2.0, WavLM e XLS-R [Tak et al. 2022] — seguido de um *back-end* (LCNN, ResNet/Res2Net, RawNet2, AASIST [Jung et al. 2022]), sendo o EER, por vezes com a *tandem Detection Cost Function* (t-DCF), a métrica predominante.

Convém distinguir dois casos frequentemente confundidos: a generalização a *ataques não vistos* — algoritmos ausentes do treino, mas de mesma origem *bona fide*, como no ASVspoof 2019 LA — e a generalização *cross-dataset*, foco desta revisão, em que o detector é treinado em um *corpus* e avaliado em outro, distinto na origem, nas condições acústicas e de canal e nos ataques (p. ex. treinar no ASVspoof e avaliar no In-the-Wild). Trabalhos que decompõem essa queda mostram que ela decorre predominantemente da diferença de distribuição entre os *corpora*, e não de maior dificuldade intrínseca do alvo [Müller et al. 2022, Müller et al. 2024].

3. Metodologia

Adota-se o PRISMA 2020, cujo fluxograma de quatro fases e *checklist* de 27 itens [Page et al. 2021a, Page et al. 2021b] foram seguidos. A escolha justifica-se por transparência e reprodutibilidade, mitigação de viés de seleção por critérios definidos *a priori*, e comparabilidade pela extração padronizada; o PRISMA é hoje referência também em Computação [Kitchenham and Charters 2007].

3.1. Bases, busca e janela temporal

As buscas foram executadas em 12 de junho de 2026 em Scopus e Web of Science (Core Collection), com filtros de idioma (inglês e português) e tipo de documento, mantendo-se log de base, data, *string* e retornos. A *string*, idêntica em ambas as bases salvo pela sintaxe de campo, foi:

```
("audio deepfake*"OR "speech deepfake*"OR "synthetic
speech"OR "spoofed speech"OR "voice spoofing"OR
"anti-spoofing"OR "presentation attack") AND (detect*
OR countermeasure* OR classif*) AND (generaliz* OR
"cross-dataset"OR "cross-corpus"OR "cross-domain"OR
"out-of-distribution"OR "in-the-wild"OR unseen OR robust*)
```

Não se aplicou filtro de ano na consulta às bases: os 995 registros recuperados abrangem publicações de 2008 a 2026, e a delimitação temporal operou exclusivamente

como critério de elegibilidade na triagem (Seção 3.2), excluindo trabalhos anteriores a 2018 salvo marcos fundacionais. O conjunto final distribui-se entre 2019 e 2026, com dois estudos de 2019, dois de 2021 e os demais a partir de 2022 — concentração que reflete o adensamento da discussão de generalização *cross-dataset* na esteira do ASVspoof 2021 e do *corpus* In-the-Wild. Não houve procedimento complementar de identificação: sem *snowballing*, busca manual ou inclusão de referências conhecidas, a busca nas bases é a única via de identificação.

3.2. Critérios de elegibilidade

Os critérios foram definidos *a priori*; para cada um indica-se a regra de decisão aplicada. Incluídos os estudos que (1) tratam de detecção de *deepfake*/*spoofing* em áudio ou voz — regra: a detecção precisa ser o objeto principal, excluindo-se trabalhos em que ela é subproduto de outra tarefa; (2) abordam generalização ou avaliação *cross-dataset* — regra estrita: o estudo precisa reportar ao menos uma avaliação em que o *corpus* de teste difira do de treino, ou analisar criticamente essa condição, não bastando variar o algoritmo de ataque dentro do mesmo *corpus*; (3) são revisados por pares ou *preprints* indexados nas bases, sem versão revisada localizada; e (4) apresentam método, avaliação empírica ou *survey* em inglês ou português.

Excluídos os que focam apenas em vídeo ou imagem; tratam só de síntese sem detecção; são anteriores a 2018, salvo marco fundacional; são resumo, pôster sem método, duplicata, editorial ou *keynote*; ou não têm texto completo disponível. O critério (2) foi aplicado de forma estrita: alegações de generalização sustentadas apenas por avaliação *in-dataset* levaram à exclusão, categoria que responde isoladamente pela maior parte das exclusões (Figura 1).

3.3. Seleção, extração e qualidade

O processo seguiu as quatro fases do PRISMA 2020, com deduplicação e triagem por título e resumo no Rayyan e registro do motivo de exclusão de cada estudo avaliado em texto completo (Figura 1). Para cada estudo incluído extraiu-se uma linha em planilha estruturada com os campos: referência, ano, *venue*, *corpora*, características/*front-end*, arquitetura, estratégia de generalização, protocolo (*in-*/*cross-dataset*), métrica e resultado, limitações e RQ(s) atendida(s). A planilha e o log não acompanham esta versão, mas estão disponíveis mediante solicitação.

A triagem, a elegibilidade e a extração foram conduzidas por um único revisor, sem dupla avaliação independente nem procedimento formal de resolução de discordâncias — limitação registrada na Seção 6.3. Como mitigação: critérios operacionalizados em regras explícitas, reduzindo o julgamento caso a caso; registro auditável do motivo de exclusão no Rayyan; e formulário de extração fixo, aplicado uniformemente.

Dos 37 incluídos, 27 conduzem eles próprios avaliação com *corpus* de teste distinto do de treino, e 10 satisfazem o critério (2) pela segunda via — analisam criticamente a condição *cross-dataset* sem propor detector novo, correspondendo ao eixo de avaliação/análise crítica (Seção 4.3). A triagem interceptou ainda cinco trabalhos de *anti-spoofing* facial recuperados por compartilharem o vocabulário de generalização *cross-domain*; nenhum alcançou o conjunto final.

Não se aplicou instrumento formal de avaliação de risco de viés (tipo AMSTAR ou ROBIS) nem escore de qualidade: os estudos foram tratados com peso semelhante na síntese narrativa, limitação registrada na Seção 6.3. O que se garantiu foi o rigor da extração, não a ponderação das evidências: os 37 estudos foram integralmente lidos em texto completo — nenhum caracterizado apenas por resumo ou metadados, e os candidatos sem texto completo foram excluídos e contabilizados como tal — com formulário preenchido sem lacunas nos campos substantivos. Em compensação, as afirmações sintéticas da Seção 6.1 vêm acompanhadas do número de estudos que as sustentam, e a linguagem sobre eficácia relativa é deliberadamente cautelosa: fala-se em estratégias *mais consistentemente associadas a ganhos*, não em estratégias *mais eficazes*. A síntese é narrativa e tabular, não meta-analítica, dada a heterogeneidade de protocolos e métricas.

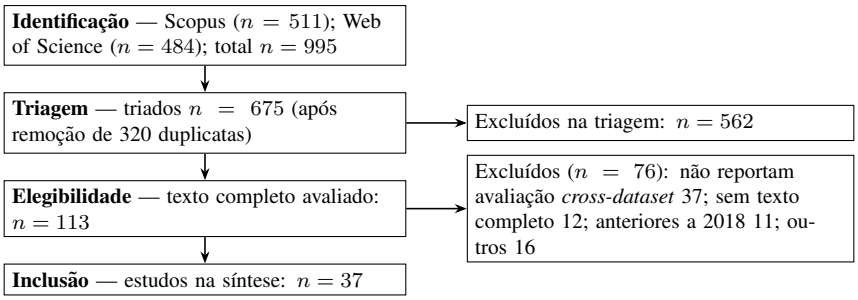


Figura 1. Fluxograma PRISMA 2020. A busca nas bases é a única via de identificação.

4. Resultados

4.1. *Corpora* e *benchmarks*

O ASVspoof domina a literatura: a edição de 2019 (LA) aparece em 24 dos 37 estudos e a de 2021 (LA/DF) em 7, firmando-se como *benchmark* de fato para treino e avaliação *in-dataset*. Como alvos OOD recorrentes despontam o In-the-Wild [Müller et al. 2022], em 13 estudos, e o WaveFake, em 7. A Tabela 1 resume os *corpora* mais empregados; as contagens *não são mutuamente exclusivas*, pois um estudo costuma usar vários. Observa-se diversificação recente com *corpora* específicos: *partially fake audio* (PartialSpoof, Half-Truth [Luong et al. 2025]), edição neural de fala (SINE [Huang et al. 2024]), voz cantada (SingFake [Zang et al. 2024]), *codecs* neurais (Codecfake [Xie et al. 2025]) e *replay* (VSDC, BTAS). Essa concentração de treino, contrastada com a heterogeneidade da avaliação, é um primeiro indício das inconsistências da Seção 5.

4.2. *Front-ends* e arquiteturas

Observa-se transição clara das características extraídas manualmente para representações aprendidas. As cepstrais e espectrais permanecem frequentes — LFCC (6), CQT/CQCC (4), *mel-spectrogram* (8), MFCC (4) —, mas a forma de onda bruta processada por camadas *sinc* (10) e, sobretudo, os modelos SSL consolidaram-se como abordagem dominante: wav2vec 2.0/XLS-R em 8 estudos, WavLM em 3 e Whisper em 3, o maior bloco quando agregados. As contagens também se sobrepõem. Diversos estudos relatam que *front-ends* SSL pré-treinados em larga escala melhoram a generalização [Kawa et al. 2023, Wang and Yamagishi 2024, Müller et al. 2024]. No

Tabela 1. *Corpora* mais empregados (n = número de estudos; as contagens se sobrepõem). LA: *Logical Access*; PA: *Physical Access*; DF: *Deepfake*; ITW: *in-the-wild*.

Corpus	Ataque	n	Papel típico
ASVspoof 2019 (LA)	TTS, VC	24	Treino / *in-dataset*
In-the-Wild	TTS, VC (ITW)	13	Avaliação OOD
WaveFake	TTS (*vocoders*)	7	Avaliação OOD
ASVspoof 2021 (LA/DF)	TTS, VC, *codec*	7	Avaliação OOD / canal
ASVspoof 2015	TTS, VC	3	Treino / histórico
VSDC	*replay* (PA)	3	Avaliação multi-ataque
PartialSpoof	*partial fake*	2	Treino / avaliação
FoR	TTS, VC	2	Avaliação OOD
Codecfake	*codec* neural / ALM	1	Treino / avaliação
BTAS 2016	*replay* (PA)	1	*Cross-database*

back-end recorrem AASIST (8), ResNet/Res2Net (7), LCNN (7), RawNet2 (6) e *Transformer*/Conformer (5); a combinação SSL + AASIST recorre como linha de base forte [Zang et al. 2024, Wang and Yamagishi 2024]. Ressalve-se que melhor desempenho *in-dataset* não implica melhor generalização: um estudo de sistematização reimplementou doze arquiteturas sob avaliação uniforme e constatou degradação de até $\sim$1000% ao migrar do ASVspoof para o In-the-Wild [Müller et al. 2022].

4.3. Eixos das estratégias de generalização

As estratégias organizam-se em cinco eixos de atribuição *mutuamente exclusiva*: cada estudo foi alocado ao eixo de sua contribuição *principal* — prevalecendo o que o próprio texto reivindica como novidade —, daí a soma reproduzir o total de 37 estudos ($14 + 7 + 3 + 3 + 10$).

O eixo de arquitetura/característica (14) propõe representações ou módulos discriminativos: Res2Net [Li et al. 2021], redes de valor complexo que exploram a fase [Müller et al. 2023], prosódia/emoção [Conti et al. 2022], fusão de características acústicas [Gong and Li 2025], localização por *self-attention* [Wu et al. 2022] e bandas de frequência interpretáveis [Salvi et al. 2023]. O de dados/*augmentation* (7) aproxima as distribuições por geração de dados — múltiplos *vocoders* [Müller et al. 2023, Park and Kwak 2025, Wang and Yamagishi 2024], nível do algoritmo de ataque [Ge et al. 2024], simulação de canal/*replay* [Zhang et al. 2021], equidade [Singh Yadav et al. 2024a] e seleção ativa [Wang and Yamagishi 2023]. O de *front-end* SSL (3) explora representações autosupervisionadas [Wang et al. 2026, Kawa et al. 2023, Wang and Yamagishi 2024], e o de treino/otimização (3) reúne adaptação de domínio [Wang et al. 2019], minimização *sharpness-aware* [Shim et al. 2023] e aprendizado contínuo [Zhang et al. 2023]. Por fim, avaliação/análise crítica (10) não propõe detector novo, mas diagnostica fatores e limitações [Shih et al. 2024, Salvi et al. 2023, Kang et al. 2025, Müller et al. 2022, Müller et al. 2024, Sali et al. 2023], incluindo revisões [Sahidullah et al. 2023, Badugu et al. 2025, Ranjan et al. 2023].

5. Avaliação *Cross-Dataset*

5.1. Protocolos

Os estudos distribuem-se em três regimes cuja confusão é, em si, fonte de inconsistência. No primeiro, *in-dataset a ataques não vistos*, treino e teste partilham a origem *bona fide*, variando apenas os algoritmos de *spoof* (p. ex. ASVspoof 2019 LA, A07–A19); EER baixo nesse regime não evidencia generalização [Li et al. 2021, Ranjan et al. 2022]. No segundo, multi-*dataset* intra-base, avalia-se em vários *corpora* isoladamente, sem cruzar fronteiras [Khan et al. 2024]. O terceiro, foco desta revisão, é a avaliação *cross-dataset*/OOD: treina-se em um *corpus* e avalia-se em outro, distinto na origem, no canal e nos ataques [Müller et al. 2022, Müller et al. 2023, Kawa et al. 2023, Müller et al. 2024].

A discrepância entre os regimes é acentuada: detectores com EER *in-domain* próximo de 1% sobem para 27%–34% no In-the-Wild [Müller et al. 2022, Müller et al. 2023, Singh Yadav et al. 2023]. Um trabalho subsequente decompôs essa degradação e mostrou que ela é dominada pela *diferença de distribuição* entre *corpora*, e não pela *dificuldade intrínseca* do alvo, quantificando que, para o In-the-Wild, a componente de dificuldade é praticamente nula [Müller et al. 2024].

A Tabela 2 sintetiza resultados sob protocolo *cross-dataset* explícito, selecionados por três critérios: (i) *corpora* de treino e avaliação distintos e declarados; (ii) EER no conjunto OOD e, quando disponível, o *in-domain* do mesmo modelo, permitindo contraste interno; e (iii) cobertura de alvos OOD e eixos distintos, em vez de listagem exaustiva. Os valores são transcritos como reportados, sem recomputação.

Tabela 2. Resultados sob avaliação *cross-dataset* explícita (EER, %; menor é melhor). Não constituem comparação direta entre métodos: os estudos diferem em *corpus* de treino, conjunto OOD, partição e pré-processamento, e cada linha só é interpretável no contraste interno entre suas colunas *in-dom.* e OOD.

Estudo	*Front-end*	*In-dom.*	OOD	Alvo OOD
RawGAT-ST [Müller et al. 2022]	Forma de onda (*sinc*)	1.23	33.94	In-the-Wild
Complex-valued [Müller et al. 2023]	Espectral complexo	—	26.95	In-the-Wild
PS3DT [Singh Yadav et al. 2023]	*Mel-spectrogram*	4.54	29.72	In-the-Wild
Whisper+MFCC [Kawa et al. 2023]	SSL (Whisper) + MFCC	—	26.72	In-the-Wild
FADEL [Kang et al. 2025]	*a preencher*	1.21	5.60	ASVspoof21 LA
SAM co-train [Shim et al. 2023]	*a preencher*	0.99	18.20	ASVspoof21 DF
CFPRF [Luong et al. 2025]	*a preencher*	7.61	43.25	LlamaPartialSpoof

Entre as linhas avaliadas no In-the-Wild, a configuração com *front-end* SSL apresenta o menor EER OOD (26,72%) frente às três não-SSL (26,95%, 29,72% e 33,94%). A leitura exige cautela: o sistema SSL foi treinado no ASVspoof 2021 DF e os demais no 2019 LA, de modo que a comparação não isola o efeito do *front-end*. O padrão é consistente com os estudos que manipulam o *front-end* sob treino fixo [Wang and Yamagishi 2024, Wang et al. 2026], mas a tabela, por si só, não o demonstra.

5.2. Métricas e inconsistências

A métrica predominante é o EER, por vezes com t-DCF [Li et al. 2021, Luong et al. 2025]. Sua adequação é questionada: o limiar ótimo é calculado *a*

posteriori sobre o teste, mascarando o desempenho sob limiar fixo — a condição real de implantação [Luong et al. 2025] —, daí a defesa de métricas dependentes de limiar e da análise de falsos positivos [Luong et al. 2025, Singh Yadav et al. 2024b]. Soma-se a divergência de *dimensão* reportada: EER na maioria, mas também AUC e *balanced accuracy* [Conti et al. 2022, Kang et al. 2025] e F1 de *frame*/locução em *partial fake* [Huang et al. 2024, Luong et al. 2025], grandezas não conversíveis. Ilustra a fragilidade a avaliação de ALMs quantizados em regime *zero-shot* [Dutta et al. 2025]: F1 elevado (até 0,946) convivendo com *balanced accuracy* próxima de 0,5 revela viés de classe que o F1 oculta.

Emergem daí quatro inconsistências recorrentes. Primeiro, a *ambiguidade terminológica*: "generalização" é aplicada indistintamente a ataques não vistos *in-dataset* e a cenários *cross-dataset*, levando a alegações de robustez sustentadas por avaliação intra-base [Khan et al. 2024, Li et al. 2021]. Segundo, a *concentração de treino*, sobretudo no ASVspoof 2019 LA (24 dos 37), denunciada como fonte de resultados enganosos por seu caráter *single-sourced* [Sali et al. 2023, Khan et al. 2023]. Terceiro, a *ausência de partições padronizadas* em conjuntos OOD: WaveFake e In-the-Wild carecem de protocolo único, e cada trabalho define a sua partição [Shim et al. 2023, Wang and Yamagishi 2023]. Quarto, a *escassez de avaliação cross-corpus sistemática* [Khan et al. 2023, Ranjan et al. 2023], somada à sensibilidade à *seed* de inicialização [Kang et al. 2025, Müller et al. 2022, Müller et al. 2024]. Em conjunto, essas inconsistências explicam por que a síntese deve ser narrativa, e não meta-analítica.

6. Discussão

6.1. Síntese por pergunta de pesquisa

Cada afirmação vem acompanhada do número de estudos que a sustenta diretamente, tornando visível a extensão da evidência.

Quanto à **RQ1**, o diagnóstico de que a queda decorre predominantemente da diferença de distribuição entre *corpora* repousa de forma direta sobre *dois* estudos que decompõem experimentalmente a degradação [Müller et al. 2024, Müller et al. 2022]; os demais são consistentes com essa leitura, mas não a testam isoladamente — conclusão bem fundamentada quanto ao método, porém de base empírica estreita, que convém tratar como hipótese de trabalho robusta e não como fato estabelecido. Os fatores específicos incluem descasamento de canal e *codec* ($n = 2$) [Zhang et al. 2021, Singh Yadav et al. 2023], dependência de artefatos não robustos do *benchmark* de treino ($n = 2$) [Shih et al. 2024, Sali et al. 2023], sensibilidade à concatenação *long-form* ($n = 1$) [Liu et al. 2024] e vieses demográficos ($n = 2$) [Singh Yadav et al. 2024b, Singh Yadav et al. 2024a].

Quanto à **RQ2**, as abordagens mais consistentemente associadas a ganhos são os *front-ends* SSL combinados a *back-ends* AASIST ($n = 4$) [Kawa et al. 2023, Wang and Yamagishi 2024, Wang et al. 2026, Xie et al. 2025] e a diversificação de dados ($n = 3$) [Wang and Yamagishi 2023, Wang and Yamagishi 2024, Park and Kwak 2025]; as de otimização mostram ganhos reais, porém modestos ($n = 4$) [Wang et al. 2019, Shim et al. 2023, Zhang et al. 2023, Zeng et al. 2023]. Essa ordenação reflete a *consistência* com que ganhos são relatados, e não uma comparação controlada — que a heterogeneidade de protocolos não permite realizar. A evidência é parcial: vários estudos

relatam EER OOD ainda entre 26% e 47%, indicando que nenhuma estratégia isolada resolve a limitação.

Quanto à **RQ3**, a constatação é metodológica e sustenta-se sobre o conjunto amplo dos estudos: ambiguidade entre regimes, concentração de treino, ausência de partições OOD padronizadas e heterogeneidade de métricas inviabilizam comparação quantitativa direta e meta-análise, e a própria adequação do EER é contestada em favor de métricas sensíveis a limiar fixo ($n = 2$) [Luong et al. 2025, Dutta et al. 2025].

6.2. Lacunas e direções futuras (RQ4)

Emergem seis direções prioritárias. A primeira é um protocolo *cross-corpus* padronizado, com *splits* oficiais e conjuntos OOD de referência, condição necessária para superar a avaliação *single-sourced* dominante [Khan et al. 2023, Sali et al. 2023]. A segunda é a adoção de métricas operacionais — limiar fixo, *balanced accuracy* e falsos positivos por subgrupo —, diante da evidência de viés de classe oculto por métricas agregadas [Dutta et al. 2025, Luong et al. 2025, Singh Yadav et al. 2024b].

A terceira é a robustez a *codecs* neurais e ALMs, deslocamento qualitativamente distinto dos anteriores. Os *vocoders* clássicos reconstroem a forma de onda a partir de representações contínuas tempo-frequência, deixando artefatos localizados e espectralmente estáveis — resíduos de banda alta, inconsistências de fase e padrões periódicos de *upsampling* —, precisamente as pistas que detectores treinados em ASVspoof aprendem a explorar [Shih et al. 2024, Salvi et al. 2023]. Os *codecs* neurais projetam o sinal em espaço latente discreto por quantização vetorial residual e reconstroem com decodificador treinado sob perdas adversariais e perceptuais: a distorção deixa de ser resíduo espectral concentrado e torna-se erro de quantização distribuído pelo espectro, moldado ao envelope da fala natural. Nos ALMs, a geração autorregressiva sobre tokens desloca a assinatura do domínio espectral local para regularidades temporais de mais longo alcance. A consequência é dupla: detectores que dependem das pistas clássicas generalizam mal porque tais pistas não são produzidas [Xie et al. 2025, Dutta et al. 2025]; e áudio autêntico recodificado por *codec* neural pode ser classificado como falso, elevando os falsos positivos. Somam-se os cenários de voz cantada [Zang et al. 2024], *partial fake* [Luong et al. 2025] e edição neural [Huang et al. 2024].

A quarta é o tratamento explícito de equidade, integrando a mitigação de viés demográfico ao projeto do detector e não como verificação posterior [Singh Yadav et al. 2024a, Singh Yadav et al. 2024b]. A quinta é a explicabilidade e análise causal, avançando do diagnóstico de *que* os detectores dependem de artefatos [Shih et al. 2024, Salvi et al. 2023] para representações comprovadamente robustas à mudança de distribuição. A sexta, sugerida pela organização por eixos aqui adotada, é uma taxonomia formal das metodologias de mitigação, que explicite composições e sobreposições entre abordagens — passo necessário para que ablações comparem estratégias equivalentes.

6.3. Limitações

Primeiro e mais relevante, a triagem e a extração foram conduzidas por um único revisor, sem dupla avaliação independente; ainda que mitigado pelos critérios *a priori*, pelo registro dos motivos de exclusão e pelo formulário fixo, o desenho não elimina o risco de viés

de seleção, e recomenda cautela nas decisões de fronteira do critério (2). Segundo, não se aplicou instrumento formal de risco de viés: os estudos foram tratados com peso semelhante apesar de níveis de rigor distintos, razão pela qual as afirmações sobre eficácia relativa foram formuladas em termos de consistência de ganhos, não de superioridade demonstrada. Terceiro, a busca restringiu-se a duas bases e a inglês e português, sem procedimento complementar de identificação, o que pode ter omitido literatura indexada apenas em outras bases ou disponível só como *preprint*. Quarto, restringiu-se o conjunto final aos estudos com texto completo efetivamente lido, o que reforça a confiabilidade da extração mas exclui trabalhos acessíveis apenas em resumo. Quinto, não acompanha esta versão material suplementar com a lista dos estudos e os campos extraídos, o que limita a verificação independente — compensado parcialmente pela *string* integral (Seção 3.1), pelos critérios operacionalizados (Seção 3.2) e pelos campos de extração reportados no corpo do artigo. Por fim, a síntese é narrativa e tabular, não meta-analítica, decisão imposta pela heterogeneidade discutida na Seção 5.

7. Conclusão

Esta revisão mapeou, sob o PRISMA 2020, as limitações de generalização *cross-dataset* em detectores de *deepfake* de áudio e voz, partindo de 995 registros e incluindo 37 estudos publicados entre 2019 e 2026. Os achados indicam que a queda *cross-dataset* é atribuída, pelos estudos que a decompõem experimentalmente, sobretudo à diferença de distribuição entre *corpora*; que os *front-ends* auto-supervisionados e a diversificação de dados são as abordagens mais consistentemente associadas a ganhos, ainda que nenhuma resolva isoladamente o gargalo e que a heterogeneidade de protocolos impeça comparação controlada; e que a comparabilidade é seriamente comprometida pela ambiguidade entre regimes, pela concentração de treino e pela heterogeneidade de métricas. Como contribuições, o trabalho oferece um processo de seleção rastreável, um mapeamento por eixos mutuamente exclusivos, uma análise crítica das práticas de avaliação e uma agenda centrada na padronização de protocolos *cross-corpus*, em métricas operacionais, na robustez a *codecs* neurais e ALMs, em equidade e em explicabilidade. Reitera-se que a síntese é narrativa e conduzida por um único revisor, com as demais limitações declaradas na Seção 6.3.

Referências

Badugu, V. K., Veesa, S., and Singh, M. (2025). The Kolmogorov-Arnold networks: A new foundation paradigm for hierarchical function learning in ASV anti-spoofing research. In *2025 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pages 359–366. IEEE.

Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M. C., and Tubaro, S. (2022). Deepfake speech detection through emotion recognition: A semantic approach. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8962–8966. IEEE.

Dutta, B., Ranjan, R., Sathvik, S., Vatsa, M., and Singh, R. (2025). Can quantized audio language models perform zero-shot spoofing detection? In *Interspeech 2025*, pages 4578–4582. ISCA.

Ge, W., Wang, X., Yamagishi, J., Todisco, M., and Evans, N. (2024). Spoofing attack augmentation: Can differently-trained attack models improve generalisation? In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12531–12535. IEEE.

Gong, L. Y. and Li, X. J. (2025). Deepfake voice detection: An approach using end-to-end transformer with acoustic feature fusion by cross-attention. *Electronics*, 14(10):2040.

Huang, S.-F., Kuo, H.-C., Chen, Z., Yang, X., Yang, C.-H. H., Tsao, Y., Wang, Y.-C. F., Lee, H.-Y., and Fu, S.-W. (2024). Detecting the undetectable: Assessing the efficacy of current spoof detection methods against seamless speech edits. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–659. IEEE.

Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J. S., Lee, B.-J., Yu, H.-J., and Evans, N. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Kang, J. Y., Yoon, J. W., Kim, S., Han, M. H., and Kim, N. S. (2025). FADEL: Uncertainty-aware fake audio detection with evidential deep learning. In *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Kawa, P., Plata, M., Czuba, M., Szymański, P., and Syga, P. (2023). Improved deepfake detection using Whisper features. In *Interspeech 2023*, pages 4009–4013. ISCA.

Khan, A., Malik, K. M., and Nawaz, S. (2024). Frame-to-utterance convergence: A spectra-temporal approach for unified spoofing detection. In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 10761–10765. IEEE.

Khan, A., Malik, K. M., Ryan, J., and Saravanan, M. (2023). Battling voice spoofing: A review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*, 56(S1):513–566.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University.

Li, X., Li, N., Weng, C., Liu, X., Su, D., Yu, D., and Meng, H. (2021). Replay and synthetic speech detection with Res2Net architecture. In *ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6354–6358. IEEE.

Liu, X., Wang, X., and Yamagishi, J. (2024). A preliminary case study on long-form in-the-wild audio spoofing detection. In *2024 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE.

Luong, H.-T., Rimon, I., Permuter, H., Lee, K. A., and Chng, E. S. (2025). Robust localization of partially fake speech: Metrics and out-of-domain evaluation. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 2205–2210. IEEE.

Müller, N., Czempin, P., Diekmann, F., Froghyar, A., and Böttinger, K. (2022). Does audio deepfake detection generalize? In *Interspeech 2022*, pages 2783–2787. ISCA.

Müller, N. M., Evans, N., Tak, H., Sperl, P., and Böttinger, K. (2024). Harder or different? understanding generalization of audio deepfake detection. In *Interspeech 2024*, pages 2705–2709. ISCA.

Müller, N. M., Sperl, P., and Böttinger, K. (2023). Complex-valued neural networks for voice anti-spoofing. In *Interspeech 2023*, pages 3814–3818. ISCA.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021b). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372:n160.

Park, S. and Kwak, I.-Y. (2025). Impact of neural-vocoder-generated spoofed speech on spoofing-detection performance. *Korean Journal of Applied Statistics*, 38(5):679–692.

Ranjan, R., Vatsa, M., and Singh, R. (2022). STATNet: Spectral and temporal features based multi-task network for audio spoofing detection. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE.

Ranjan, R., Vatsa, M., and Singh, R. (2023). Uncovering the deceptions: An analysis on audio spoofing detection and future prospects. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, pages 6750–6758.

Sahidullah, M., Delgado, H., Todisco, M., Nautsch, A., Wang, X., Kinnunen, T., Evans, N., Yamagishi, J., and Lee, K.-A. (2023). Introduction to voice presentation attack detection and recent advances. In *Handbook of Biometric Anti-Spoofing*, Advances in Computer Vision and Pattern Recognition, pages 339–385. Springer.

Sali, Y., Erden, M., and Haznedaroğlu, A. (2023). Investigation of approaches in the literature and potential problems for audio spoofing detection. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Salvi, D., Bestagini, P., and Tubaro, S. (2023). Towards frequency band explainability in synthetic speech detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 620–624. IEEE.

Shih, T.-H., Yeh, C.-Y., and Chen, M.-S. (2024). Does audio deepfake detection rely on artifacts? In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12446–12450. IEEE.

Shim, H.-j., Jung, J.-w., and Kinnunen, T. (2023). Multi-dataset co-training with sharpness-aware optimization for audio anti-spoofing. In *Interspeech 2023*, pages 3804–3808. ISCA.

Singh Yadav, A. K., Bhagtani, K., Bestagini, P., and Delp, E. J. (2024a). Improving fairness in synthetic speech detectors. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pages 362–366. IEEE.

Singh Yadav, A. K., Bhagtani, K., Salvi, D., Bestagini, P., and Delp, E. J. (2024b). FairSSD: Understanding bias in synthetic speech detectors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4418–4428. IEEE.

Singh Yadav, A. K., Xiang, Z., Bhagtani, K., Bestagini, P., Tubaro, S., and Delp, E. J. (2023). PS3DT: Synthetic speech detection using patched spectrogram transformer. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 496–503. IEEE.

Tak, H., Todisco, M., Wang, X., Jung, J.-w., Yamagishi, J., and Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*.

Wang, H., Dinkel, H., Wang, S., Qian, Y., and Yu, K. (2019). Cross-domain replay spoofing attack detection using domain adversarial training. In *Interspeech 2019*, pages 2938–2942. ISCA.

Wang, X. and Yamagishi, J. (2023). Investigating active-learning-based training data selection for speech spoofing countermeasure. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 585–592. IEEE.

Wang, X. and Yamagishi, J. (2024). Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 10311–10315. IEEE.

Wang, Y., Lin, X., Qiu, Y., and Tan, G. (2026). Leveraging Gaussian attention for adapter-based fine-tuning of WavLM in audio deepfake detection. *International Journal of Intelligent Systems*, 2026(1):6786731.

Wu, H., Kuo, H.-C., Zheng, N., Hung, K.-H., Lee, H.-Y., Tsao, Y., Wang, H.-M., and Meng, H. (2022). Partially fake audio detection by self-attention-based fake span discovery. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 9236–9240. IEEE.

Xie, Y., Lu, Y., Fu, R., Wen, Z., Wang, Z., Tao, J., Qi, X., Wang, X., Liu, Y., Cheng, H., Ye, L., and Sun, Y. (2025). The Codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*, 33:386–400.

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop*.

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., and Zhao, Y. (2023). Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

Zang, Y., Zhang, Y., Heydari, M., and Duan, Z. (2024). SingFake: Singing voice deepfake detection. In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12156–12160. IEEE.

Zeng, C., Wang, X., Miao, X., Cooper, E., and Yamagishi, J. (2023). Improving generalization ability of countermeasures for new mismatch scenario by combining multiple advanced regularization terms. In *Interspeech 2023*, pages 1998–2002. ISCA.

Zhang, X., Yi, J., Tao, J., Wang, C., and Zhang, C. Y. (2023). Do you remember? overcoming catastrophic forgetting for fake audio detection. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 41819–41831.

Zhang, Y., Zhu, G., Jiang, F., and Duan, Z. (2021). An empirical study on channel effects for synthetic voice spoofing countermeasure systems. In *Interspeech 2021*, pages 4309–4313. ISCA.