Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating Audio Deepfake Detection Across Languages and Mobile Communication Channels

Abdullahi Salad Abdi

deerow99@simad.edu.so

SIMAD University

Abdikarim Abdi Abdulle

SIMAD University

Maryan Mohamud Hambali

SIMAD University

Omar Osman Haji Abdi

SIMAD University

Systematic Review

Keywords: audio deepfakes, speech anti-spoofing, low-resource languages, cross-lingual transfer, speaker independence, domain shift, mobile codecs, evaluation protocols

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11250417/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Audio deepfake detectors are frequently evaluated using curated speech corpora whose language, speaker, synthesis method and recording channel differ from operational audio. This narrative review critically examines how these differences affect the interpretation of detector performance, with particular attention to low-resource languages and mobile-mediated recordings. A purposive synthesis of 2024–2026 peer-reviewed journal literature links benchmark design, speaker and source separation, low-resource-language evidence, cross-dataset generalization, compression and acoustic stress testing. The evidence indicates that high discrimination on matched tests cannot establish reliability against unfamiliar generators, speakers, languages or transmission chains. We propose an evidence-informed evaluation workflow comprising documented data provenance; speaker-, source- and generator-disjoint partitions; matched baselines; isolated and combined distribution shifts; subgroup metrics with uncertainty; calibration and application-specific error analysis; and reproducibility reporting. Somali is discussed as an illustrative under-resourced case, not as a language for which the reviewed studies establish detection performance. Three original conceptual figures and three comparative tables translate the synthesis into a practical research and reporting framework. This framework is a methodological proposal, not a validated detector, standard or meta-analytic finding. Its use could make future evaluation of underrepresented languages more transparent and better aligned with real recording conditions.

1 Introduction

Neural text-to-speech (TTS) and voice-conversion (VC) systems support accessibility and language technologies but also enable synthetic or manipulated audio that can be used for impersonation and the falsification of recorded speech. Recent reviews document progress in deepfake-detection architectures and available datasets, while identifying persistent challenges in generalization and trustworthy evaluation [1–4, 10]. The main question is therefore not simply which classifier reports the lowest error on a familiar corpus, but what evidence is required to claim that a detector will behave reliably beyond its development conditions.

Published results are shaped by data collection, speaker overlap, generator families, encoding and processing. The ASVspoof 5 journal paper explicitly expands speaker, recording and attack diversity, whereas cross-dataset analyses show that detector behavior can change substantially under domain shift and codec degradation [5, 7]. A matched-domain equal error rate (EER) is informative for a specified benchmark, but is not a universal measure of reliability on unseen speech.

The evaluation challenge is intensified for low-resource languages, which may have limited consented recordings, fewer synthesis systems, uncertain accent coverage and little independently collected test audio. An IEEE Access study of Urdu demonstrates the feasibility of dedicated evaluation, but also illustrates sensitivity to acoustic and signal-level perturbations [6]. One language-specific result cannot

establish transfer to Somali or other African languages with different linguistic and recording characteristics.

This review connects linguistic representation, partition integrity, attack novelty and mobile-channel robustness in one practical evaluation protocol. Figure 1 summarizes the mismatch between typical matched benchmarks and plausible deployment conditions. Somali is used only as an illustrative setting for future study; this review does not claim Somali-specific empirical performance, nor does it report a new dataset or trained detector [5–8].

2 Review approach and relationship to existing literature

This paper uses a focused narrative-review approach. Peer-reviewed journal literature published from 1 January 2023 to 3 October 2026 was purposively identified through targeted searches of publisher and bibliographic pages for combinations of “audio deepfake detection”, “speech anti-spoofing”, “low-resource language”, “cross-lingual”, “domain shift”, “codec”, “robustness” and “evaluation”. Reference chaining was used to locate directly relevant journal work. The final cited evidence is limited to articles with verifiable DOIs in Scopus-indexed journals reported in Q1 or Q2 in at least one relevant 2025 SCImago subject category; quartiles are journal- and category-specific and should be reconfirmed at submission. Conference papers, preprints and datasets without an eligible journal article are not included as primary citations.

Selection prioritized work that informs data provenance, speaker and generator separation, underrepresented-language evaluation, compression, acoustic stress, reporting or the interpretation of detector scores. This is a selective thematic synthesis, not a comprehensive or reproducible systematic search. No protocol registration, PRISMA study counts, pooled effects or formal risk-of-bias estimates are claimed. The illustrative studies in Table 3 do not represent an exhaustive inventory of all work published during the period.

Broad reviews already cover model families, feature representations, available corpora, spoofing attacks and the standard assessment metrics [1–4, 10]. Maltby et al. instead investigate robustness failure mechanisms through a combined survey and cross-dataset case study [7], while Firc et al. examine the comparability of evaluation procedures and detectors [8]. Table 1 maps these contributions against the narrower, protocol-centered contribution of this review. The article does not present existing recommendations as a universally adopted consensus standard.

Table 1
Relationship between eligible journal literature and the focus of this narrative review.

Literature strand	Illustrative journal evidence	Contribution to this review
Broad speech and multimedia surveys	Li et al. [1]; Pham et al. [2]; Alnaqbi and Ikuesan [3]; Khan et al. [4]; Zhang et al. [10]	Focus the synthesis on evaluation under language, attack and channel shift rather than cataloguing architectures.
Benchmark construction	ASVspoof 5 journal description [5]	Translate documented speaker and attack diversity into separation and provenance checks.
Low-resource and accented speech	Urdu benchmark [6]; Irish-accented resource [11]	Distinguish corpus availability from representativeness and cross-language validity.
Robustness and comparative testing	Maltby et al. [7]; Firc et al. [8]; Wani and Amerini [9]; Kim et al. [15]	Link codec and adversarial stress to interpretable error reporting and practical testing.

3 Linguistic coverage and partition integrity

3.1 Language resources are not equivalent to language representativeness

A dataset may contain recordings labeled with many languages while providing weak coverage of within-language accent, age, speech style, device and channel variation. Recent surveys discuss multilingual datasets, including MLAAD, but the original MLAAD conference publication is intentionally excluded from this journal-only reference list [1, 10]. Any use of such a resource in a future experiment should identify the exact corpus version, generator and bona fide source distributions, licensing and limitations. A language label or large sample count alone is not evidence of generalizable authenticity cues.

Authentic and synthesized recordings are often acquired through different pipelines. If nearly all synthetic recordings use one bitrate, recording environment or source pool while bona fide speech comes from another, a classifier may learn collection artifacts rather than synthesis features. The ASVspoof 5 resource demonstrates the importance of broader speaker and attack coverage and separate speaker partitions, but its specific design should not be treated as proof of representativeness for every application or language [5].

3.2 Evidence from low-resource and accented speech

The Urdu study by Owais et al. evaluates multiple feature and model combinations under baseline and augmented corpus configurations and tests sensitivity to noise, pitch and tempo perturbations [6]. It supplies direct language-specific evidence, but comparisons beyond its data, speech varieties and attack systems require independent tests. A 2026 Scientific Data resource focused on Irish-accented English

audio and audiovisual deepfakes illustrates the separate importance of accent-aware collection and source-level integrity checks [11]. Neither study establishes detector performance for Somali.

For future Somali research, the minimum descriptive unit should include the speaker and, where reliably documented, the spoken variety, speaking context and acquisition device. A held-out Somali test partition should not share speakers or derivative source files with training. Because no Somali-specific results are synthesized here, claims about Somali detector accuracy or relative linguistic difficulty would be unwarranted [5, 6, 11].

3.3 Speaker, source and generator independence

Speaker-disjoint splits test generalization to new voices, source-disjoint splits prevent a single utterance and its transformed variants from appearing on opposite sides of the train–test boundary, and generator-disjoint splits test unfamiliar synthesis or VC methods. Each is a separate requirement; satisfying one does not imply the others. The ASVspoof 5 journal paper documents multiple speaker-disjoint partitions and diverse attacks, providing a concrete model for transparent partition descriptions [5]. For small corpora, publication of auditable split manifests is especially important.

4 Robustness under mobile and out-of-domain conditions

4.1 Acoustic and device shift

Mobile audio can be altered by microphone response, reverberation, background noise, automatic gain control, denoising and changes in sample rate. Urdu experiments include noise, pitch and tempo perturbations, while broader robustness research finds that domain-dependent features may not transfer reliably [6, 7]. Acoustic changes can either mask synthesis cues or create new artifacts shared by bona fide and generated speech. Results must therefore be disaggregated by transformation instead of reported only as one averaged score.

Controlled stress tests are useful because the same held-out recording can be evaluated at named signal-to-noise ratios (SNRs) and noise types. Such transformations do not reproduce every characteristic of actual handset and transmission recordings. Independently collected field audio is necessary to evaluate device, room, human behavior and end-to-end transmission jointly [6–9].

4.2 Codec, transcoding and communication-channel shift

Compression may alter the spectral or temporal details used by a detector. Maltby et al. identify codec masking and artifact overfitting as important robustness concerns, and Wani and Amerini test a compression-aware teacher–student architecture on named benchmark and codec conditions [7, 9]. These findings warrant explicit testing of audio-file compression, narrowband speech where relevant and sequential transcoding. A codec observed in training is not an adequate substitute for a genuinely unseen codec chain.

Future reports should state codec implementation, nominal bitrate, sampling rate, transcoding order, loudness normalization and whether processing was applied symmetrically to bona fide and fake speech. Class-dependent processing can itself create a label shortcut. Figure 3 separates controlled signal-level stress testing from independent field testing, which must not be conflated [7, 9, 10].

4.3 Novel attacks and combined distribution shifts

A system trained on known TTS or VC methods may depend on generator-specific artifacts. ASVspoof 5 broadens the attack inventory, while the robustness analysis of Maltby et al. describes degradation under unseen generation and data conditions [5, 7]. An unseen generator in a familiar language on clean audio is not the same challenge as an unseen generator producing an underrepresented language and transmitted through a previously unseen channel.

The evaluation design should therefore include both isolated shifts and compound shifts, with the latter reserved for independently held-out combinations. Robustness claims should identify exactly which generator families, languages, speaker populations and channels were unseen at test time. Aggregate performance can be reported only alongside these disaggregated results [5, 7–9].

5 An integrated evaluation and reporting framework

The central contribution is an evidence-informed minimum framework for future studies, not a newly validated performance measure. Figure 2 provides the proposed workflow, from provenance to reproducibility, and Table 2 specifies what an author should test and report at each stage. The protocol integrates benchmark partitioning, low-resource-language evidence, comparative evaluation and codec-robustness findings [5–10].

Table 2
Proposed minimum evaluation and reporting matrix.

Dimension	Minimum evaluation protocol	Required reporting
Provenance	Record authentic sources, generation tools, processing, permission and corpus version.	Corpus and version, language/speaker/utterance counts, licenses and generator list.
Partition integrity	Audit speaker- and source-disjoint partitions; add generator-disjoint tests.	Split manifests, overlap checks, adaptation and augmentation policy.
Matched baseline	Use a genuinely held-out evaluation partition from the training-domain distribution.	Confusion matrix, precision, recall, F1, ROC-AUC, EER and confidence intervals.
Cross-language transfer	Evaluate each held-out language; distinguish zero-shot from adaptation.	Per-language and subgroup results, test sizes and annotation uncertainty.
Attack novelty	Exclude named generator or VC families from training and validation.	Per-generator and pooled-with-breakdown performance; attack inventory.
Acoustic / mobile shift	Apply named noise, SNR, codec and transcoding chains to held-out audio.	Metric changes by noise type, SNR, bitrate, sampling rate and channel.
Combined shift	Hold out combinations of language, generator and communication channel.	Separate compound-shift scores with the exact shift configuration.
Calibration / costs	Validate threshold on a separate representative set; examine confidence and errors.	Threshold-selection protocol, FPR/FNR, calibration, use-case error consequences.
Resource profile	Test latency and memory on stated hardware and software.	Inference time, model size, device, energy where available.
Reproducibility	Release lawful metadata, splits and code or justify restrictions.	Seeds, versions, preprocessing, access terms and sensitivity analyses.

5.1 Discrimination, calibration and decision consequences

Accuracy can conceal serious minority-class errors. Precision, recall, F1 and class-specific confusion matrices show the relationship between false alarms and missed synthetic speech, while ROC-AUC and EER characterize discrimination from different perspectives. The comparative study by Firc et al. cautions against comparing detectors tested under incompatible datasets and protocols [8]. Reports should indicate whether the positive class is synthetic speech and provide uncertainty estimates with resampling units that respect source or speaker clustering.

Calibration is distinct from discrimination: a useful ranking of utterances does not guarantee that output scores represent probabilities suitable for a particular workflow. Any threshold should be selected

without access to the final test partition and reviewed in light of expected error consequences. Voice-authentication, content moderation and forensic triage may require different operating points, so a single recommended threshold is inappropriate [1, 8, 10].

5.2 Resource constraints and defense claims

Transformer representations, handcrafted spectral features and compact distilled systems have different computational and domain-shift trade-offs. Studies using pathological speech features and compression-aware distillation illustrate complementary approaches, but their measured improvements apply to the tested datasets and processing conditions [9, 14]. For low-resource deployment, evaluations should also report latency, memory, file size and feasible offline operation.

Robustness to ordinary acoustic noise should not be conflated with adversarial resistance. Kim et al. assess inference-time noise addition as an adversarial robustness strategy; their results motivate dedicated threat-model reporting rather than broad claims that signal augmentation defeats all attacks [15]. The final paper should state whether its intended use is benchmark research, authentication support or investigative triage, and limit conclusions accordingly.

6 Evidence from selected journal studies

The evidence is heterogeneous in task, corpus, generator family and test design. Table 3 provides a concise, study-level interpretation of selected empirical publications, including what each contributes to the proposed framework and what remains untested. Its descriptions are qualitative and should not be interpreted as a cross-study ranking or pooled effect estimate [5–9, 11–15].

Table 3
Selected empirical journal evidence and implications for future evaluation.

Study / year	Setting or resource	Evaluation emphasis	Relevance and limits
Wang et al. (2026) [5]	ASVspoof 5; diverse speech speakers and attack algorithms	Speaker-disjoint benchmark design and attack diversity	Informs partitioning; does not establish transfer to all languages or mobile channels.
Owais et al. (2026) [6]	Urdu deepfake speech configurations	Low-resource benchmarking; noise, pitch and tempo perturbations	Direct low-resource evidence, not a Somali-language validation.
Maltby et al. (2026) [7]	Cross-dataset speech detection case study	Dataset bias, codec masking, artifact overfitting	Explains shift-related failure; tests remain dataset specific.
Firc et al. (2025) [8]	Comparative detector evaluation	Metrics, detector comparison and evaluation differences	Supports standardized reporting; no universal detector ranking follows.
Wani and Amerini (2025) [9]	Compression-aware deepfake detection	Codec robustness and teacher–student learning	Illustrates engineering trade-offs; no guarantee on unseen codec chains.
Kumar et al. (2026) [11]	Irish-accented English audiovisual/audio resource	Accent-aware collection and file-level integrity checks	Useful dataset-design example; not a broad African-language study.
Muruganandham et al. (2025) [12]	FoR benchmark variants	Within-dataset and rerecorded-condition performance	Shows why acquisition variants should be reported separately.
Al-Shamayleh et al. (2025) [13]	Scene-manipulated fake audio	Transfer-oriented acoustic feature design	Adjacent forensic evidence; task differs from speech TTS/VC detection.
Speech-pathology feature study (2024) [14]	ASVspoof 2019 and 2021	Complementary acoustic features and cross-benchmark tests	Reported gains are bound to the evaluated attack/test regime.
Kim et al. (2025) [15]	Audio deepfake detector under adversarial tests	Inference-time noise and adversarial robustness	Different threat model from routine mobile background noise.

7 Research priorities for underrepresented languages

First, research programs need independently collected, consented authentic speech and carefully documented synthetic speech for languages underrepresented in major benchmarks. For a future Somali

corpus, researchers should define the sampled dialects or varieties, speaker characteristics, recording contexts, devices and generator coverage. Native-speaker review and data-governance procedures should be reported rather than inferred from a language label [5, 6, 10, 11].

Second, future experiments should predefine leakage checks and publish per-language, per-speaker-group, per-generator and per-channel errors with uncertainty. Negative findings and subgroup sample sizes are part of the result, not optional supplementary details. Adaptation effects should be measured separately from true zero-shot transfer, and all augmented derivatives should remain in the same partition as their source recordings [5, 6, 8].

Third, practical studies should test compound shifts and compare simulations with independently acquired audio. Mobile-network compression, denoising, reverberation and unseen generation can interact; isolated results cannot establish that a detector is resilient when they coincide. Figure 3 and Table 2 provide an explicit starting plan, whose added value should itself be measured in prospective multicorpus studies [7, 9, 12, 15].

Fourth, operational use should distinguish supportive forensic signals from conclusive attribution. A deepfake detector may assist triage or authentication review, but high-stakes decisions require contextual verification, carefully selected thresholds, access controls, retention policies and routes for contesting false alerts. These governance considerations follow from uncertainty in generalization and should not be mistaken for demonstrated deployment safety [3, 8, 10].

8 Limitations

This is a purposive narrative review restricted to selected 2023–2026 Q1/Q2 Scopus-indexed journal literature rather than all influential research. The deliberate exclusion of conference papers means original MLAAD, EMNLP cross-domain and workshop proceedings are not primary references; eligible journal surveys may discuss those resources as secondary evidence. Study designs, linguistic populations, attack inventories, processing pipelines and metrics differ, precluding credible pooled error rates or a universal model ranking. The proposed figures and matrix are original analytical syntheses, not independently validated standards. Somali is an illustrative target for future research, and no empirical claim about Somali-language detector performance can be inferred from Urdu, Irish-accented English or general benchmarks. Journal quartiles are category- and year-dependent and require checking again at submission.

9 Conclusion

A detector’s matched-benchmark performance is a limited statement about a defined data distribution, not evidence of general reliability across languages, voices, synthesis methods and mobile communication channels. The reviewed journal literature motivates a more transparent protocol: documented provenance; independent speaker, source and generator partitions; isolated and combined shifts; subgroup and uncertainty reporting; calibration; and resource-aware field validation [5–10]. The

three figures and three tables translate this argument into practical steps for future investigators of underrepresented languages. Prospective studies with independently acquired speech are needed to test whether the framework improves the reproducibility and relevance of audio deepfake evaluation.

Declarations

Funding. [Confirm with all authors whether external funding was received and provide the accurate statement before submission.]

Competing interests. [Obtain and insert the confirmed conflict-of-interest declaration from every author before submission.]

Author contributions. [Insert verified CRediT contributions for each author before submission.]

Ethics approval and consent to participate. Not applicable: this manuscript is a literature-only narrative review and reports no newly recruited human participants, new recordings, or analysis of individual-level participant data.

Consent for publication. Not applicable; no identifiable participant data or images are presented.

Data availability. Not Applicable

Use of generative AI and AI-assisted technologies. OpenAI ChatGPT was used to assist with manuscript drafting, organization, language editing and preparation of original conceptual figures. The authors take responsibility for verifying the final content, including references, scientific claims and illustrations.

Acknowledgements. None are reported; add any approved acknowledgements before submission if applicable.

References

1. Li, M., Ahmadiadli, Y., & Zhang, X-P. (2025). A survey on speech deepfake detection. *ACM Computing Surveys*, *57*(7), 165. https://doi.org/10.1145/3714458
2. Pham, L., Lam, P., Tran, D., Tang, H., Nguyen, T., Schindler, A., et al. (2025). A comprehensive survey with critical analysis for deepfake speech detection. *Computer Science Review*, *57*, 100757. https://doi.org/10.1016/j.cosrev.2025.100757
3. Alnaqbi, M., & Ikuesan, R. A. (2026). A systematic review of audio deepfake detection techniques for digital investigation. *Discover Computing*, *29*, 202. https://doi.org/10.1007/s10791-026-10077-1
4. Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing*, *28*, 48. https://doi.org/10.1007/s10791-025-09550-0

5. Wang, X., Delgado, H., Tak, H., et al. (2026). ASVspoof 5: design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, *95*, 101825. https://doi.org/10.1016/j.csl.2025.101825
6. Owais, M., Jadoon, K. K., Sandhu, A. I., Ali, Z., Mahmood, Z., Yahya, M., & Wahid, A. (2026). Deepfake audio detection in low-resource languages: a case study of Urdu. *Ieee Access : Practical Innovations, Open Solutions*, *14*, 12407–12421. https://doi.org/10.1109/ACCESS.2026.3654621
7. Maltby, H., Wall, J., Glackin, C., Moniri, M., Salami, I., Li, L., & Cannings, N. Robustness in deepfake speech detection: a survey of failure mechanisms including an experimental case study. Computer Speech & Language. Published online 28 July 2026. https://doi.org/10.1016/j.csl.2026.102037
8. Firc, A., Malinka, K., & Hanáček, P. (2025). Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors. *Cybersecurity*, *8*, 50. https://doi.org/10.1186/s42400-024-00346-1
9. Wani, T. M., & Amerini, I. (2025). Dynamic knowledge condensation with audio-selective transformer for audio deepfake detection. *Discover Computing*, *28*, 231. https://doi.org/10.1007/s10791-025-09746-4
10. Zhang, B., Cui, H., Nguyen, V., & Whitty, M. (2025). Audio deepfake detection: what has been achieved and what lies ahead. *Sensors (Basel, Switzerland)*, *25*(7), 1989. https://doi.org/10.3390/s25071989
11. Kumar, T., Umair, M., Leavy, S., Eustace, P., Curry, E., & Asghar, M. N. Irish-accented English audio-visual deepfake datasets with deep packet inspection-inspired media integrity validation. Scientific Data. Published online 30 July 2026. https://doi.org/10.1038/s41597-026-07940-7
12. Muruganandham, P., Thangasamy, G. R., Jayaraman, S., & Dharmarajan, R. (2025). LSTM autoencoder based parallel architecture for deepfake audio detection with dynamic residual encoding and feature fusion. *Scientific Reports*, *15*, 23514. https://doi.org/10.1038/s41598-025-08198-6
13. Al-Shamayleh, A. S., Riasat, H., Alluhaidan, A. S., Raza, A., El-Rahman, S. A., & AbdElminaam, D. S. (2025). Novel transfer learning based acoustic feature engineering for scene fake audio detection. *Scientific Reports*, *15*, 8066. https://doi.org/10.1038/s41598-025-93032-2
14. Chaiwongyen, A., Duangpummet, S., Karnjana, J., Kongprawechnon, W., & Unoki, M. (2024). Potential of speech-pathological features for deepfake speech detection. *Ieee Access : Practical Innovations, Open Solutions*, *12*, 121958–121970. https://doi.org/10.1109/ACCESS.2024.3447582
15. Kim, I., Doan, T-P., Hong, S., & Jung, S. (2025). Inference-time noise addition for improving adversarial robustness of audio deepfake detection system. *Ieee Access : Practical Innovations, Open Solutions*, *13*, 200669–200682. https://doi.org/10.1109/ACCESS.2025.3635633

Figures

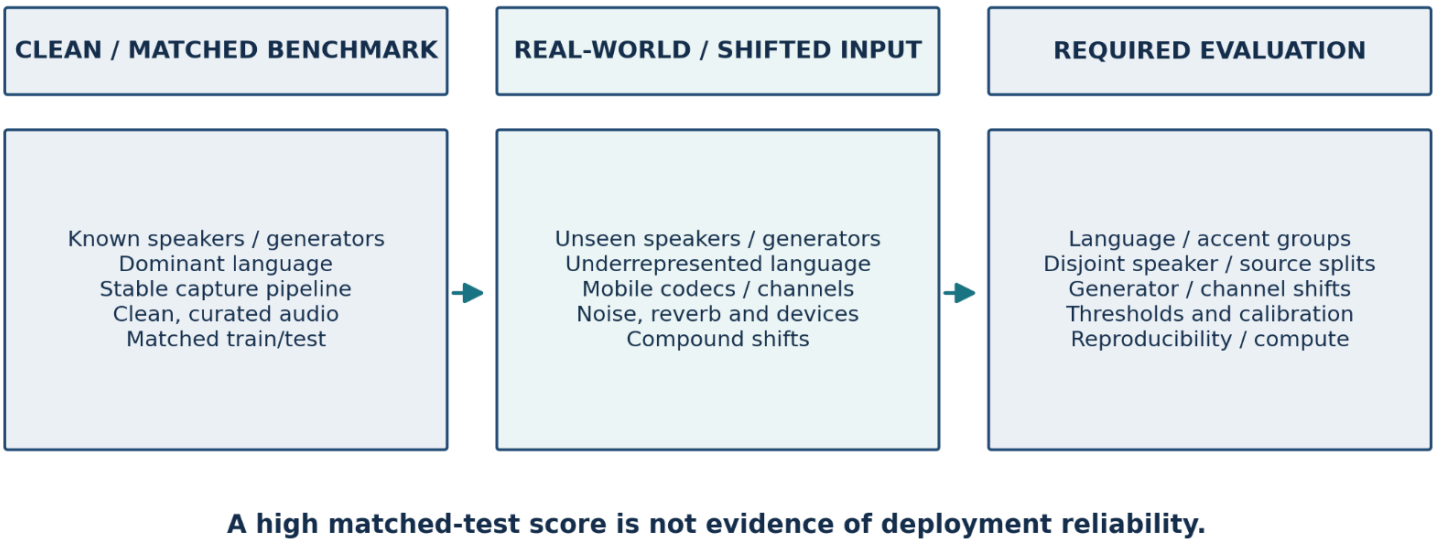


Figure 1

Conceptual benchmark-to-deployment gap. The diagram is an original synthesis of evaluation considerations and does not depict measured effect sizes.

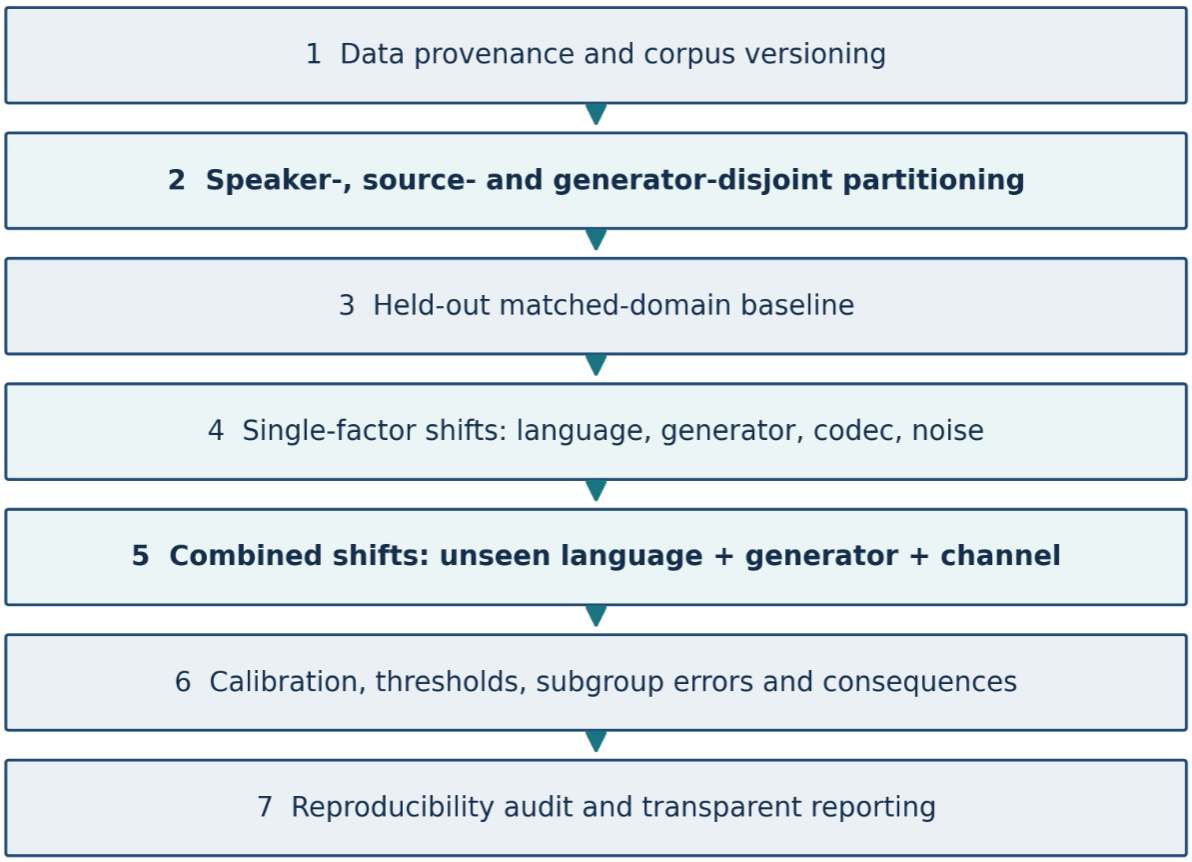


Figure 2

Proposed evaluation workflow for low-resource-language and mobile-channel audio deepfake detection. Steps are methodological recommendations requiring prospective validation.

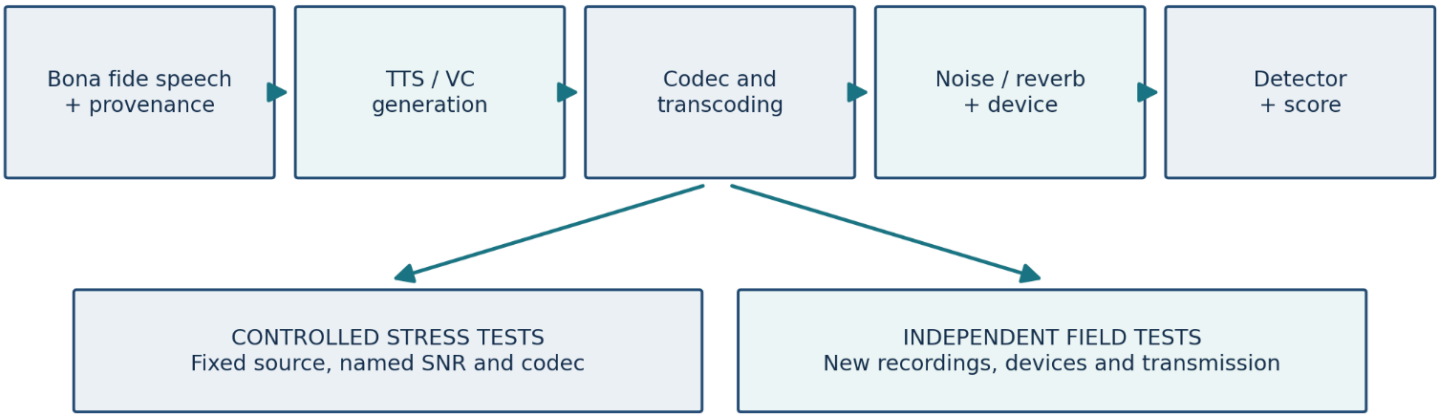


Figure 3

Proposed mobile-channel stress-test pipeline. Controlled transformations and independently collected field recordings supply different forms of evidence.