



Detecting Synthetic Voice Clones in Recorded Evidence: A Workflow for Audio Examiners

Emeka Chandrasekar^{1*}, Aditya Subramaniam², Jorge Lindo³

¹Department of Science, Niger Graduate Institute of Science, Ghana

²Department of Science, Siberian Institute for Advanced Science, Italy

³Department of Science, Pampas Academy of Science, Nepal

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijfs.2026.v1.i1.pp14-19](https://doi.org/10.62715/ijfs.2026.v1.i1.pp14-19)

Received: 10 October 2025

Accepted: 05 December 2025

Published: 19 January 2026

Abstract

Neural speech synthesis can now imitate a named individual from a short sample of that person's voice, and recordings whose authenticity would once have been accepted without comment are being challenged as possible clones. Forensic audio examiners have long-standing procedures for authenticating recordings, but those procedures were written to detect edits, deletions and re-recording, not speech that was never spoken. Research on automatic spoofing detection has advanced quickly, yet its detectors are trained and tested on laboratory corpora and are known to perform poorly on material unlike their training data. This article proposes an eight-stage workflow that places the new detectors inside the established framework of forensic authentication. The stages run from framing the question and preserving the file, through examination of the container, suitability assessment, structured listening and signal analysis, to the use of automated detectors under case-matched validation and a reasoned, reviewed conclusion. For each stage the article states what is examined, why, and what the result can and cannot support. The workflow is a proposal built from published standards and literature. It has not been validated on casework or test material here, and the article identifies what such validation would need to show.

Keywords: *Audio forensics, Voice cloning, Speech synthesis, Deepfake detection, Recording authentication, Forensic validation*

© 2026 Emeka Chandrasekar et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A recorded voice has traditionally been challenged in two ways: the recording has been tampered with, or the speaker is someone else. A third challenge is now common in principle and increasingly in practice: the speech was generated by a machine trained to sound like the person. Generative models that produce raw audio, beginning with WaveNet (van den Oord et al., 2016), and sequence models that map text to spectrograms, such as Tacotron (Wang et al., 2017), made synthetic speech natural enough to pass casual listening. Speaker-conditioned systems then showed that a new voice could be imitated from a few seconds of reference audio without retraining (Jia et al., 2018). Tools built on these ideas are now widely available.

* Corresponding author: Emeka Chandrasekar, Department of Science, Niger Graduate Institute of Science, Ghana. E-mail: emeka.chandrasekar@niger.edu

© 2026 Emeka Chandrasekar et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For the examiner the consequence runs in both directions. A fabricated recording may be offered as genuine, and a genuine recording may be disowned as fabricated. Either way a court will want to know whether there are signs of synthesis and how much weight their presence or absence deserves. The established literature on audio authentication addresses edits and discontinuities (Koenig and Lacey, 2009; Maher, 2009) and predates neural cloning.

This article sets out a workflow intended to fill that gap. It is a methodological proposal. The stages are drawn from existing authentication practice, from the research literature on spoofing countermeasures and from general principles of forensic evaluation. The workflow has not been applied to casework or to a test collection for the purposes of this article, and no performance figures are claimed for it.

2. Background

Two bodies of knowledge inform the design. One describes how cloned speech is made and what traces the process tends to leave. The other describes how well those traces can be detected, and under what conditions detection fails.

2.1. How cloned speech is produced

Most current systems work in two parts. An acoustic model converts text, or a source utterance in the case of voice conversion, into an intermediate representation such as a mel spectrogram, conditioned on an embedding that captures the target speaker's vocal characteristics. A neural vocoder then turns that representation into a waveform. Generative adversarial vocoders such as HiFi-GAN produce high-quality output much faster than real time (Kong et al., 2020). Traces of synthesis can arise at either stage: prosody and timing that are too regular or oddly placed come from the acoustic model, while fine spectral and phase irregularities, especially at high frequencies, come from the vocoder.

2.2. What detection research has shown

Spoofing countermeasures were first developed to protect automatic speaker verification, and the survey by Wu et al. (2015) describes the early field. The ASVspoof challenge series has supplied common data and metrics (Todisco et al., 2019; Yamagishi et al., 2021). Detectors range from classifiers on hand-designed features, such as constant Q cepstral coefficients (Todisco et al., 2017), to networks that learn directly from the waveform (Tak et al., 2021). On the challenge data the best systems make few errors.

The difficulty is generalisation. Müller et al. (2022) tested detectors trained on challenge data against real-world recordings and found that performance deteriorated sharply. Lossy compression, telephone transmission and re-recording through a loudspeaker all remove or mask the fine detail on which detectors rely, a problem that the 2021 challenge was designed to expose (Yamagishi et al., 2021). New synthesis methods appear faster than detectors are retrained, and data sets built from a range of vocoders show that a detector may recognise some and miss others (Frank and Schönherr, 2021). An examiner cannot therefore treat a detector's output as a finding in itself.

3. Design Principles

Four principles shaped the workflow. First, no single method is sufficient, so several independent lines of examination are combined and their agreement or disagreement is itself informative. Second, the examination proceeds from the least to the most interpretive step, and from the file as an object to the speech it contains, so that early findings are not coloured by later ones. Third, every automated tool is used only within conditions for which its performance has been measured, in line with the consensus on validation in forensic voice comparison (Morrison et al., 2021). Fourth, the examiner's exposure to case information is managed, because knowledge of what the instructing party hopes to hear is a recognised source of bias in expert judgement (Dror, 2020).

Figure 1 shows the eight stages in order, and Table 1 summarises what each examines and the main limit on what it can show. The stages are described below.

4. The Proposed Workflow

The description assumes a digital audio or audio-visual file submitted with a question about whether the speech in it was synthesised. Where only an analogue or re-recorded copy exists, the earlier stages yield less and the conclusion must be correspondingly cautious.

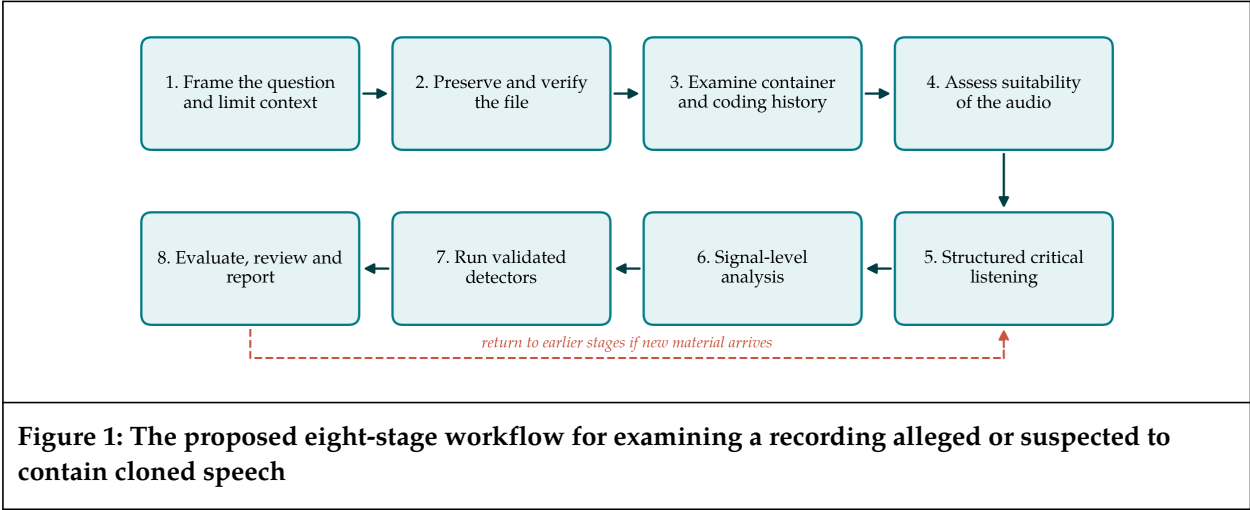


Table 1: Stages of the proposed workflow, what each examines and its principal limitation		
Stage	What is examined	Principal limitation
1. Frame the question	Competing propositions; context needed	Relies on the instructing party
2. Preserve and verify	Original file, hash values, custody	Says nothing about content
3. Container and coding	Format, metadata, encoder traces	Metadata can be stripped or forged
4. Suitability	Duration, bandwidth, noise, codec	Poor material may bar any conclusion
5. Critical listening	Prosody, breathing, room acoustics	Subjective; good clones pass
6. Signal analysis	Spectrum, phase, noise floor, hum	Artefacts masked by lossy coding
7. Automated detectors	Scores from validated detectors	Weak on unseen synthesis methods
8. Evaluate and report	All findings, both propositions	Conclusion is graded, never categorical

4.1. Stage 1: Frame the question and limit context

The examiner records the question as a pair of propositions, for example that the speech was uttered by a human speaker and captured by the stated device, or that it was generated by a synthesis system. Information needed for the technical work is requested: the claimed recording device and application, the route by which the file reached the submitter, and any intermediate platforms. Information about the wider case that has no technical bearing is kept from the examiner where the organisation allows a case manager to filter it. The examiner also notes at this point which recordings of undisputed origin are available, such as other files from the same device or account, because these will be needed as reference material at Stage 3 and again when automated tools are tested at Stage 7.

4.2. Stage 2: Preserve and verify the file

The file is acquired and handled as digital evidence under the principles of [ISO/IEC 27037 \(2012\)](#): a forensic copy is made, cryptographic hash values are calculated and recorded, and all work is done on the copy. Where the original device is available it should be obtained, since a recording found in place on a telephone, with its application database and system logs, carries far more provenance information than the same audio forwarded as a message attachment.

4.3. Stage 3: Examine the container and coding history

Before the audio is played, the file structure is examined. Container format, header fields, timestamps, encoder identifiers and the arrangement of data blocks are compared with what the claimed device and application are known to produce, ideally using reference recordings made on the same model and software version. [Koenig and Lacey \(2009\)](#) describe this kind of structural analysis for edited recordings, and it applies equally here. A file whose structure matches an editing program or a synthesis toolkit and not the claimed recorder calls for explanation. Signs of double compression indicate that the audio has been decoded and re-encoded at least once. Consistency with the claimed device does not exclude synthesis, since generated audio can be played into or imported onto a real device.

4.4. Stage 4: Assess suitability

The examiner documents the properties that determine what later stages can deliver: net duration of the questioned speech, sampling rate and effective bandwidth, codec and bit rate, signal-to-noise ratio, reverberation, overlapping talkers and background sound. Narrowband telephone audio at a low bit rate retains little of the high-frequency region where vocoder traces are usually sought. If the material falls outside the conditions for which the laboratory's tools have been tested, that is recorded now, before any result is seen, and it limits the strength of conclusion available at Stage 8.

4.5. Stage 5: Structured critical listening

Listening is carried out on calibrated equipment against a written checklist. Points attended to include the placement and sound of breaths, hesitations and fillers, the naturalness of pitch movement across phrase boundaries, the handling of unusual names and numbers, coarticulation between words, and whether the voice sits in a plausible acoustic space that changes as a real speaker moves. Background sound and reverberation should be continuous with the speech and not laid beneath it. Observations are written down as descriptions. Listening rarely settles the matter, since well-made clones sound natural and poor recordings of real speech sound odd, but it directs the signal analysis to particular passages.

4.6. Stage 6: Signal-level analysis

Wideband and narrowband spectrograms, long-term average spectra and phase displays are examined for features associated with synthesis: an abrupt upper band limit unrelated to the codec, periodic patterning or smeared harmonics at high frequencies, unnaturally uniform noise between words, and discontinuities where separately generated segments were joined. The acoustic environment is analysed for consistency of noise floor and reverberation time between speech and pauses. If mains hum is present, the electric network frequency criterion may show whether the recording has a continuous time base and, where reference data exist, whether it fits the claimed date ([Grigoras, 2007](#)). A cloned voice inserted into a genuine ambient recording can show a mismatch between the two layers. Each observation is recorded with the passage in which it occurs.

4.7. Stage 7: Automated detectors under case-matched validation

One or more automated detectors are then applied. The condition for using them is that the laboratory holds validation results for material resembling the questioned recording in codec, bit rate, bandwidth, duration, language and noise. Where it does not, a case-specific test is run: genuine speech and synthetic speech from several current systems are passed through the same coding and transmission chain as the questioned file, and the detector's scores for the two classes are compared. Only if the classes separate under those conditions is the score for the questioned recording given weight. Scores are reported with the validation results that give them meaning and are never converted into a bare statement that the recording is or is not synthetic.

4.8. Stage 8: Evaluate, review and report

The findings are brought together and assessed under each proposition in turn, following the logic of evaluative reporting set out by the European Network of Forensic Science Institutes ([ENFSI, 2015](#)). The examiner asks how probable the whole set of observations would be if the speech were human and

directly recorded, and how probable if it were synthesised. The conclusion is expressed as a graded level of support, with the reasons stated, or as inconclusive where the material or the validation does not allow more. A second examiner reviews the work, preferably having formed a view on the listening and signal stages before seeing the first examiner's conclusion. The report states the limits of the methods in plain terms.

5. Fitting the Workflow to Practice

Laboratories differ in resources, and the workflow can be scaled. A unit without its own detectors can still perform Stages 1 to 6 and report on file structure, coding history and acoustic consistency, stating that automated detection was not attempted. A unit that does run detectors needs a maintained library of genuine and synthetic reference speech and a routine for regenerating the synthetic part as new systems appear. Reference recordings from common telephones and messaging applications support Stage 3 and are worth collecting in any case.

The workflow also interacts with speaker comparison. A request to say whether a cloned voice "is" a named person is misconceived, since a successful clone is by design similar to its target, and automatic speaker recognition systems can be deceived by it (Wu *et al.*, 2015). Authenticity should therefore be examined before any comparison of voices is undertaken, and a comparison should not proceed on material whose authenticity is in real doubt.

6. Limitations and What Remains to Be Shown

The proposal has not been validated. Before it could be recommended for routine use, its error rates would need to be measured on collections of genuine and cloned recordings that have passed through realistic channels, with examiners blind to ground truth. Such a study should report the two kinds of error separately, since calling a real recording synthetic and calling a clone real have different consequences. It should also test whether the combination of stages outperforms the detectors alone, and how much examiners agree with one another at the listening stage.

Some limits are inherent. Detection is an adversarial problem: a forger can degrade a clone by compression or re-recording precisely to hide its traces, and can test the output against public detectors before using it. Absence of detectable artefacts is therefore weak evidence of authenticity, and provenance evidence from the device, the network and witnesses will often carry more weight than signal analysis. The workflow is offered as a disciplined way of saying what the audio itself can contribute and of declining to say more.

7. Conclusion

Cloned speech does not make audio evidence unusable, but it changes what authentication has to address. The workflow proposed here orders the examination from file to signal to automated score, ties every tool to validation under case-like conditions, and ends in a graded and reviewed opinion. Its components are already familiar to audio examiners. What is new is their arrangement around a question the older procedures did not ask, and the insistence that a detector's output be read only alongside evidence of how that detector behaves on comparable material. Empirical testing of the workflow as a whole is the necessary next step.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Dror, I.E. (2020). Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Analytical Chemistry*, 92(12): 7998-8004. <https://doi.org/10.1021/acs.analchem.0c00704>
- ENFSI (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. European Network of Forensic Science Institutes.
- Frank, J. and Schönherr, L. (2021). WaveFake: A data set to facilitate audio deepfake detection. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
- Grigoros, C. (2007). Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. *Forensic Science International*, 167(2-3): 136-145. <https://doi.org/10.1016/j.forsciint.2006.06.033>
- ISO/IEC (2012). *ISO/IEC 27037:2012 Information technology. Security techniques. Guidelines for identification, collection, acquisition and preservation of digital evidence*. International Organization for Standardization, Geneva.
- Jia, Y., Zhang, Y., Weiss, R.J., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 31.
- Koenig, B.E. and Lacey, D.S. (2009). Forensic authentication of digital audio recordings. *Journal of the Audio Engineering Society*, 57(9): 662-695.
- Kong, J., Kim, J. and Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33.
- Maher, R.C. (2009). Audio forensic examination. *IEEE Signal Processing Magazine*, 26(2): 84-94. <https://doi.org/10.1109/MSP.2008.931080>
- Morrison, G.S., Enzinger, E., Hughes, V., et al. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, 61(3): 299-309. <https://doi.org/10.1016/j.scijus.2021.02.002>
- Müller, N.M., Czempin, P., Dieckmann, F., Froghyar, A. and Böttinger, K. (2022). Does audio deepfake detection generalize? *Proceedings of Interspeech 2022*, 2783-2787.
- Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N. and Larcher, A. (2021). End-to-end anti-spoofing with RawNet2. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6369-6373.
- Todisco, M., Delgado, H. and Evans, N. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45: 516-535. <https://doi.org/10.1016/j.csl.2017.01.001>
- Todisco, M., Wang, X., Vestman, V., et al. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *Proceedings of Interspeech 2019*, 1008-1012.
- van den Oord, A., Dieleman, S., Zen, H., et al. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
- Wang, Y., Skerry-Ryan, R.J., Stanton, D., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *Proceedings of Interspeech 2017*, 4006-4010.
- Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F. and Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66: 130-153. <https://doi.org/10.1016/j.specom.2014.10.005>
- Yamagishi, J., Wang, X., Todisco, M., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 47-54.

Cite this article as: Chandrasekar E., Subramaniam A. and Lindo J. (2026). Detecting Synthetic Voice Clones in Recorded Evidence: A Workflow for Audio Examiners. *International Journal of Forensic Science* . 1(1), 14-19. doi: 10.62715/ijfs.2026.v1.i1.pp14-19