



AI-Generated Voice Impersonation in Financial Fraud: A Narrative Review and Proposed Payment Verification Protocol

¹ Mary A. Olugbemi-Bello

²Ubakaeze Victor Chiagozie

³Fatimah Yesuf

⁴Babajide J. Sunmonu

¹University Of The Potomac, U.S.A

Bello.adewumi@gmail.com

<https://orcid.org/0009-0007-1765-2143>

²Institution: Southern Utah University, U.S.A.

ubakaezevc@suu.edu

<https://orcid.org/0009-0005-7719-723X>

³Emerson College Engagement Lab, Boston, Massachusetts

⁴Mddus Limited, Glasgow, United Kingdom

fz.yesuf@gmail.com

Abstract

Objective: This narrative review examines how organisations can verify urgent payment requests when a familiar-sounding voice can no longer be treated as reliable evidence of the caller's identity. **Methods:** This is a structured narrative review. The literature search was run on 28 December 2025 and included only sources published on or before 30 November 2025, a predefined evidence cut-off (Section 2.1). Scholarly databases, technical repositories and primary guidance from law-enforcement and financial-regulatory agencies were consulted. **Results:** Two frequently cited incidents (the 2019 UK energy-firm case and the 2024 Arup case) are reported in the media as voice- or deepfake-enabled fraud, but neither has public forensic confirmation of synthetic voice; the Arup case additionally involved a video call rather than a voice-only channel. In laboratory experiments on human detection, Mai et al. (2023) found single-clip accuracy of about 73% for deepfake clips and 67.78% for genuine clips, above chance but well below the level required to authenticate a payment request; the clips came from one older text-to-speech system, and brief training raised confidence-weighted accuracy by only 3.84%. Barrington et al. (2025), using commercial ElevenLabs clones of 220 speakers, found that listeners judged a clone to be the same identity as the real speaker about 80% of the time and identified AI-generated voices correctly only about 60% of the time. Müller, Pizzi, et al. (2022) found that players and a detection model struggled with the same attacks, and Warren et al. (2024), with 1,212 participants, reported average accuracy of about 73%, with deepfakes accepted as human more often than genuine clips were rejected. None of these experiments tested listeners known to be familiar with the speaker. Automated detection systems achieve EER below 2% on the ASVspoof 2019 LA benchmark, but Müller, Czempin, et al. (2022) summarise the deterioration on in-the-wild audio as about 200–1,000%, with EERs of 34% to 91% across the configurations they tested (50% is chance). Models trained on clean ASVspoof 2019 data lost 10.7

to 20.5 EER points on telephony-encoded versions of the same attacks (Müller et al., 2024), although the best ASVspoof 2021 system, trained without matched data, still reached an EER of 1.32% on its telephony track (Yamagishi et al., 2021). These findings do not prove that every detector fails in every payment call, but they do argue against relying on detection or human listening alone. Conclusion: On the available evidence, voice similarity is not a reliable authentication factor for high-risk payment requests. The author proposes a four-step procedural protocol (pause the payment, call back on a pre-existing verified number, confirm any new bank details through a genuinely independent channel, and require a second approver above a risk-based threshold) as guidance drawn mainly from established anti-fraud principles and law-enforcement advice, informed by the detection evidence reviewed. The protocol has not been empirically tested as a combined intervention and no specific reduction in fraud is claimed.

Keywords: voice cloning, deep fake audio, payment fraud, business email compromise, verification, synthetic speech detection

Introduction

AI voice-cloning technology has developed within a few years into a tool that can be misused for fraud. Contemporary systems can generate realistic clones of a person's voice from short audio samples; the resulting voice can be used in real time during a telephone conversation or recorded in advance and sent as a voice message (Khanjani et al., 2023). This creates a specific risk for staff who process payments, because a convincing impersonation of a chief executive, finance director or supplier requesting a transfer or a change of bank details may be taken seriously on the strength of the voice alone.

The scale of business email compromise (BEC) as a broader category is substantial. According to the U.S. Federal Bureau of Investigation's Internet Crime Complaint Center (IC3), BEC produced adjusted losses of USD 2.77 billion across 21,442 complaints in 2024, within total reported cybercrime losses of USD 16.6 billion (FBI IC3, 2025a). These figures are not an estimate of voice-cloning losses: IC3 does not currently disaggregate BEC losses by the impersonation technique used, so the share attributable specifically to synthetic voice is unknown. What is documented is that BEC has historically relied on email spoofing and social engineering, and that in December 2024 the FBI warned that criminals are now using generative AI including voice cloning to impersonate relatives and well-known public figures in order to elicit payments (FBI IC3, 2024). A further advisory on 15 May 2025 reported the use of AI-generated voice messages to impersonate senior U.S. government officials in an ongoing campaign aimed at gaining access to personal accounts (FBI IC3, 2025b).

Regulatory interest in voice cloning is increasing, but the evidence base has clear gaps. Voice-cloning fraud stories are common in the business press, yet very few have been supported by published forensic analysis, and the academic literature on voice-cloning fraud specifically remains limited. At the same time, work on

the detection of synthetic speech has developed rapidly and now faces well-documented difficulties with audio from generators the detector has not seen and with variable transmission channels.

This article deals with a practical issue that comes with these two developments: *How should an organisation verify an urgent payment request when a familiar-sounding voice cannot be trusted on its own?* The emphasis is on calls and voice messages asking for a fund transfer or payment change. The article presents the research available on voice impersonation and its detection, outlines a verification process that is based on the research findings and on the principles of anti-fraud, and makes explicit the distinction between research findings and the author's recommendations.

Research Question

What does the available evidence show about the reliability of voice as a means of authenticating a payment request, and which verification controls do law-enforcement and anti-fraud guidance recommend in response?

Review Method

This article is a structured narrative review and the selection of the approach was due to the purpose of the article being to assemble evidence from academic research, technical literature, and primary regulatory or law-enforcement guidance on each of the following: voice impersonation using AI technology, detection of synthetic speech and payment verification. Moving through a structured narrative approach enables evidence from a variety of sources to be taken into account and therefore explicit criteria for search, selection and interpretation of material could still be maintained. The method is described transparently, giving the reader a clear idea of the extent of the evidence base, the rationale for the selection of sources and the limitations that may impact the conclusions.

Search Date and Evidence Cut-Off

The literature search was executed on 28 December 2025. The evidence cut-off was set at 30 November 2025. The search was not limited to a single year: it used a publication-date filter of 2018 to 2025 where the interface allowed, and material published after 30 November 2025 was set aside so that the evidence base did not shift during writing and revision. Seed references were used for citation chasing (Section 2.2). Inclusion decisions followed the criteria in Section 2.4. The consequence for the reader is that any claim about the "current" state of synthetic-speech generation or detection in this article refers to the published literature as of 30 November 2025; work released after that date, including any new detection architectures or forensic analyses of reported cases, is not represented. Because generation and detection both change

quickly, readers assessing operational implications should re-check the primary literature before selecting any specific detection technology.

Sources Searched

The search combined academic databases and research repositories, technical literature and primary institutional guidance. Academic and technical sources were searched through Google Scholar, PubMed Central, IEEE Xplore, the ISCA Archive, arXiv and Semantic Scholar. For Google Scholar the first 100 results per query were screened by title. For the other databases the queries were run against title, abstract and keyword fields with the publication-date filter set to 2018–2025 where the interface allowed. The retained search record logs the following numbers of records by source: Google Scholar 22, IEEE Xplore 12, ISCA Archive 8, PubMed Central 4, arXiv 5, and Semantic Scholar 2 (53 in total). These are the counts held in the search record; they should not be read as the total number of hits each database returned. The search log does not retain a mapping between each query string and its yield, so Table 1 records only what the retained documentation supports and does not reconstruct missing query-level figures. These sources were chosen to capture interdisciplinary evidence on fraud, cybersecurity, speech technology, artificial intelligence, human perception and automated detection.

Table 1: Search Record by Source

Database / repository	Query record retained	Search field / filter	Search date	Records logged
Google Scholar	Search strings listed in Section 2.3; exact query-to-yield mapping was not retained	First 100 results per query screened by title	28 December 2025	22
IEEE Xplore	Search strings listed in Section 2.3; exact query-to-yield mapping was not retained	Title, abstract and keyword fields; 2018–2025 publication filter where available	28 December 2025	12
ISCA Archive	Search strings listed in Section 2.3; exact query-to-yield mapping was not retained	Title, abstract and keyword fields; 2018–2025 publication filter where available	28 December 2025	8
PubMed Central	Search strings listed in Section 2.3; exact query-to-yield mapping was not retained	Title, abstract and keyword fields; 2018–2025 publication filter where available	28 December 2025	4
arXiv	Search strings listed in Section	Title, abstract and keyword fields;	28 December 2025	5

	2.3; exact query-to-yield mapping was not retained	2018–2025 publication filter where available		
Semantic Scholar	Search strings listed in Section 2.3; exact query-to-yield mapping was not retained	Title, abstract and keyword fields; 2018–2025 publication filter where available	28 December 2025	2
Citation chasing	Backward reference-list checking and forward “Cited by” checking from seed references; exact per-seed query/count record was not retained	Reference lists and Google Scholar “Cited by” feature	28 December 2025	No reliable numerical total retained

Primary guidance and evidence were also gathered directly from the official websites of the FBI IC3, the U.S. Financial Crimes Enforcement Network (FinCEN), the U.S. Federal Trade Commission (FTC) and the Association of Certified Fraud Examiners (ACFE). Each site was searched using the terms "deepfake", "voice cloning", "synthetic media" and "business email compromise"; results were screened for advisories, alerts and reports directly relevant to voice impersonation or payment-verification controls. These institutional sources were cited especially where information was needed about patterns of fraud, regulatory warnings, institutional controls or documented fraud situations.

In addition, seed references were used to identify further studies through backward citation chasing (references cited by the seed items) and forward citation chasing (later work citing the seed items, via Google Scholar's "Cited by" feature). Records added through citation chasing are described in the Appendix rather than being merged into the database yields above.

Search Strategy and Search Terms

The search strategy combined keywords relating to synthetic speech, voice impersonation, fraud, detection and payment controls. Terms were combined with Boolean AND within each line and, where the interface supported it, phrases were quoted so that only exact matches were returned. The principal search strings, run against title, abstract and keyword fields with the 2018–2025 filter, were:

- "voice cloning" AND (fraud OR scam)
- "audio deepfake" AND detection
- ("synthetic speech" OR "speech deepfake") AND detection AND (generalisation OR generalization)

- "CEO fraud" AND (voice OR audio)
- "business email compromise" AND (callback OR verification)
- ("payment verification" OR "payment authorisation") AND ("dual authorisation" OR "dual control")
- ASVspoof
- deepfake AND ("financial institution" OR bank OR payment)

Variants of these terms (for example, "voice conversion", "text-to-speech", "replay attack") were used when scanning results to catch studies that describe the same technology or fraud mechanism under different labels. Retrieved records were exported to a spreadsheet with fields for source, query, title, authors, year, venue and reason-for-inclusion-or-exclusion; screening decisions were recorded against the criteria in Section 2.4 by the sole author, without independent duplication.

Inclusion and Exclusion Criteria

Explicit inclusion and exclusion criteria were set to keep the evidence base focused on voice-based impersonation, synthetic speech, AI-enabled fraud, detection performance and payment verification. Visual deepfakes were excluded as a separate field. Table 2 lists the criteria.

Table 2: Inclusion and Exclusion Criteria

Criterion	Inclusion Criteria	Exclusion Criteria
Topic relevance	Sources directly addressing voice cloning, synthetic speech, audio deepfakes, AI-enabled impersonation, fraud, or payment verification controls were considered eligible.	Sources unrelated to voice-based impersonation, synthetic speech, AI-enabled fraud, or payment verification were excluded.
Publication date	Material published on or before 30 November 2025 was eligible.	Material published after 30 November 2025 was excluded from the substantive evidence base.
Research focus	Studies examining human detection, automated detection, synthetic-speech generation, generalisation, fraud mechanisms, or organisational verification practices were included.	Studies focused exclusively on unrelated AI applications or detection problems without relevance to voice or payment fraud were excluded.
Media type	Research concerning audio, speech, voice conversion, text-to-speech, or related multimodal evidence where audio was relevant was considered.	Studies limited exclusively to video deepfakes without an audio or voice component were excluded.
Source quality	Peer-reviewed academic studies, recognised technical research, and primary guidance from relevant government, law-enforcement, regulatory, or professional organisations were included.	Unverified online material, promotional content, and sources lacking sufficient evidence or identifiable provenance were excluded.

Case evidence	Reported fraud cases were considered where they provided relevant evidence about voice impersonation, payment fraud, or associated verification weaknesses, with the evidentiary status explicitly identified.	Blog posts or news reports used as the sole basis for claiming that a particular incident was a confirmed voice-cloning case were excluded.
Language	Sources available in English were included.	Non-English sources were excluded because the review was restricted to English-language evidence.
Accessibility and relevance	Sources containing sufficient information to evaluate their relevance to the research question were considered.	Sources for which the available information was insufficient to assess their relevance or evidentiary basis were excluded.

These criteria mattered most for reported fraud incidents. If a media report or secondary source described an incident as an AI or deepfake fraud, it was not treated as a confirmed voice-cloning case unless forensic evidence of synthetic speech was publicly available. Without such evidence, the incident is described as a reported or alleged case. This separation helps avoid overstating the empirical evidence and keeps facts apart from interpretations and media characterisations.

Source Selection and Evidence Assessment

Sources were evaluated on relevance to the research question, methodological quality, publication date and the authority of the publisher or issuing body. Quality was assessed differently by evidence type, because the same criteria cannot fairly be applied to a controlled experiment and a press report. Peer-reviewed experimental and benchmark studies were judged on sample size, task design, generator diversity and the match between test conditions and the payment-call setting. These criteria were turned into a four-item checklist that was applied to each research source at revision (Table 3): (1) sample or data adequacy; (2) diversity and currency of the generators or attacks tested; (3) match between test conditions and a payment call (a speaker the listener knows, a telephone channel, a realistic share of fraudulent calls); and (4) availability of data or code. Each item is rated Met, Partly met or Not met. No validated instrument exists for this mixed body of evidence, so the checklist adapts the criteria above and the ratings reflect the author's reading of each paper. Preprints were eligible where authorship, methodology and reproducibility (code and data availability) allowed an informed judgement; none of the research sources cited in this article is an unreviewed preprint. Institutional guidance from the FBI IC3, FinCEN, the FTC and the ACFE was treated as authoritative for its issuing body's own operational recommendations and reported statistics, but not as independent evidence of the effectiveness of any specific control. News reports were used only for background on incidents and were never treated as evidence that a particular fraud was voice-cloning-enabled. The synthesis distinguishes throughout between (i) research findings, (ii) reported or alleged

incidents, (iii) institutional guidance and (iv) the author's own recommendations. The design and main limits of each research source are summarised in Table 4 (Section 3). Three research sources (Barrington et al., 2025; Müller, Pizzi, et al., 2022; Warren et al., 2024) were added at revision outside the search protocol (Section 5).

Table 3: Appraisal of the Research Sources

Source	1. Sample or data	2. Generators or attacks	3. Match to a payment call	4. Data or code
Mai et al. (2023)	Partly met: 529 listeners, but 100 clips from one speaker per language	Not met: one pre-trained text-to-speech system, older than the state of the art	Not met: clean clips, unfamiliar speaker, listeners warned, about half the clips fake	Met: data on OSF
Barrington et al. (2025)	Met: 220 speakers; 604 listeners	Partly met: one commercial system (ElevenLabs)	Partly met: commercial clones, but strangers to listeners, full-band audio, no telephone channel	Met: audio on Hugging Face; listener data on Zenodo
Müller, Pizzi, et al. (2022)	Met: 410 users, 13,229 rounds (published abstract)	Partly met: 13 attack algorithms, all from the 2019 benchmark	Not met: clean English read speech, unfamiliar speakers, feedback after every round, self-selected online players	Not stated in the version reviewed
Warren et al. (2024)	Met: 1,212 participants; three datasets	Partly met: over 100 algorithms in one dataset, one cloning tool in another; none a current commercial system	Partly met: no feedback, headphones, three listeners per clip; older datasets, no telephone channel, celebrity voices not	Met: coded data, models and code on Zenodo

			shown to be known to listeners	
Müller, Czempin, et al. (2022)	Met: 37.9 hours, 58 speakers, twelve architectures	Partly met: generators in the wild not identified	Not met: public-figure audio, clips of about 4 seconds, no telephone channel; detectors trained on ASVspoof 2019 only	Met: dataset published
Müller et al. (2024)	Partly met: four models; ASVspoof and in-the-wild data	Partly met: ASVspoof attacks plus unidentified in-the-wild generators	Partly met: includes telephony-encoded data; models trained on clean 2019 data	Not stated in the paper
Yamagishi et al. (2021)	Met: 48 speakers; 42 submissions to the logical-access task	Partly met: 13 attack algorithms in logical access; over 100 in the deepfake task	Partly met: real VoIP and PSTN transmission, but attacks drawn from the 2019 evaluation set	Met: databases published on Zenodo
Khanjani et al. (2023); Almutairi & Elgibreen (2022)	Surveys; not rated	-	-	-

Limitations of the Review Method

The present review has some methodological limitations. Firstly, it is not a formal systematic review based on a framework like PRISMA. Search was therefore not conducted to ensure complete coverage of all publications relevant to the search. Secondly, the search and screening process were not repeated independently by another reviewer, so possibly relevant studies or other interpretations were not identified. Third, only English language materials have been reviewed, so relevant evidence in other language(s) may not have been included.

The rapidly evolving technology of voice cloning also has an impact on the evidence base. Detection methods and methods of fraud are constantly changing, and it is not always possible to get a satisfactory

amount of forensic information from publicly reported frauds to determine exactly how an impersonation was created. Moreover, the review extends to academic research, technical publications and institutional guidance with varying methodologies and evidentiary value. These should not therefore be considered equal to other sources of evidence listed below.

Lastly, instead of trying to tackle every type of deepfake-related fraud, the review concentrates just on payment impersonation and current verification procedures. The results should therefore be applied to this limited context and other cybersecurity, regulatory, professional and organisational advice should be reviewed and applied when creating payment controls.

Evidence on Voice Impersonation and Detection

Synthetic-speech systems can now generate the characteristics of a target voice from short audio samples. Khanjani et al. (2023) describe two main approaches: text-to-speech synthesis, which produces speech from text, and voice conversion, which re-renders spoken input in another speaker's voice. Either output can be used in social-engineering attacks in which a familiar-sounding voice, sometimes accompanied by a recognisable email address, is used to request a change of bank details or an urgent transfer. Reported cases of such fraud should be treated with care, because most high-profile incidents described as "voice deepfakes" have not been supported by published forensic analysis; the 2019 UK energy-firm case and the 2024 Arup case are examined in Section 3.1.

The evidence reviewed below indicates that neither human listening nor current automated detection can, on its own, reliably distinguish synthetic from genuine speech in the conditions typical of a payment call. Sections 3.2 and 3.3 report those findings in detail; Section 4 sets out the procedural verification response that follows from them.

Table 4 summarises the research sources behind Sections 3.2 and 3.3 and the features of each that limit how far it transfers to a payment call.

Table 4: *Characteristics of the Research Sources Reviewed in Sections 3.2 and 3.3*

Source	Design and sample	Conditions and main results	Limits for payment calls
Mai et al. (2023)	Online experiment; 529 Prolific participants; English and Mandarin; 20 clips (single-clip task) or 20 clip pairs (paired task) each; one female speaker per language; clips from one pre-trained text-to-speech system	Listeners knew deepfakes were present; about half the clips were fake; deepfakes identified 73% of the time, genuine clips 67.78%; training raised	Clean clips, no telephone channel; speakers unlikely to be known to listeners; older synthesiser; equal base rate; speaker familiarity not tested

		confidence-weighted accuracy by 3.84%	
Barrington et al. (2025)	Two online perceptual studies; 220 US speakers; clones from ElevenLabs Instant Voice Cloning; 304 listeners (identity) and 300 (naturalness)	Clone judged the same identity as the real speaker about 80% of the time; AI-generated voices identified correctly about 60% of the time	Listeners were strangers to the speakers; clips of about 5 seconds, full-band audio, no telephone channel; one commercial system; senior author has an interest in a detection company
Müller, Pizzi, et al. (2022)	Online game; 410 users and 13,229 rounds (published abstract); clips from the ASVspoof 2019 evaluation set; a detection model played alongside	Players and the detector struggled with the same attacks; no advantage for IT professionals; native English speakers did better	Older generators; clean English read speech; feedback after each round; harder attacks over-sampled; unfamiliar speakers
Warren et al. (2024)	1,212 US participants (Prolific); 20 clips each from WaveFake, ASVspoof 2021 deepfake track or FakeAVCeleb; three listeners per clip; no feedback	Individual accuracy 65%, 71% and 81% by dataset (average about 73%); genuine audio classified better than deepfakes by 15 to 21 points; more deepfakes accepted as human than genuine clips rejected	Older datasets; one cloning tool in FakeAVCeleb; no telephone channel; celebrity voices not shown to be known to listeners
Müller, Czempin, et al. (2022)	Twelve detection architectures re-implemented and trained on ASVspoof 2019 only; tested on ASVspoof 2019 and a new in-the-wild set (37.9 hours, 58 celebrities and politicians, 17.2 hours deepfake)	Authors summarise deterioration as about 200–1,000%; in-the-wild EERs 33.9–91.1% across configurations; for example RawGAT-ST 1.2% to 37.2% (full-length input)	Public-figure audio, not telephone calls; academic models, not commercial products
Müller et al. (2024)	Four detection models; ASVspoof 2019 and 2021 data; performance gap split into “hardness” and “difference”	Gaps of 30–78 EER points in the wild and 11–21 points on telephony-encoded logical access, mostly “difference”	Explains why detectors fail; does not estimate how often they would fail in practice
Yamagishi et al. (2021)	ASVspoof 2021 challenge report: logical-access track with VoIP and PSTN transmission and several codecs; deepfake track with media compression	Best logical-access EER 1.32%; best deepfake-track EER 15.64%; baseline systems about 9–19% (logical access)	No codec-by-codec comparison with clean audio; logical-access attacks from the 2019 evaluation set (13 algorithms)
Khanjani et al. (2023)	Survey of audio deepfake generation and detection methods	Secondary source; no new experiment	Used for background on methods

Almutairi & Elgibreen (2022)	Review of audio deepfake detection methods	Secondary source; no new experiment	Used for background on method choice, accents and noise
---	--	-------------------------------------	---

Voice-Cloning Technology and Its Application in Fraud

The voice-cloning systems exist in two general classes: Text-to-Speech (TTS) synthesis and voice conversion (Khanjani et al., 2023). Speaking from a text input (TTS systems) or changing the voice of a spoken input (voice conversion systems) is done via the voice model of a target speaker. Both methods have gained increasing popularity and are becoming increasingly data efficient. Khanjani et al. (2023) point out that there are a number of openly available tools that can generate plausible speech from a few seconds of reference audio, reducing the technical hurdles for malicious applications. There are numerous high-profile cases reported in which this technology has been used in the context of payment fraud, but the evidentiary status of each needs to be stated precisely.

Case 1: UK energy firm (2019). The CEO of an energy firm in the UK was directed on the phone to wire €220,000 (about USD 243,000) into a supplier's account; the case was reported at the time as the first known fraud of this kind to involve artificial intelligence. Later, the CEO reported that the voice of the caller was a “subtle German accent” and a speech “melody” of his German parent-company boss. The insurer, Euler Hermes Group SA, claimed that the attack was probably carried out with commercially available voice-generation software. However, no names have emerged, no forensic analysis of the audio has been released and the technology employed has not been independently verified. Evidence status: alleged, not certified. The case was initially introduced by Stupp (2019) in *The Wall Street Journal*, and has been referenced in secondary sources (Damiani, 2019).

Case 2: Hong Kong multinational (2024). An employee of Arup, a multinational engineering company, transferred approximately USD 25.6 million (HKD 200 million) after joining a video call in which multiple apparent colleagues, including a person presented as the company's CFO, are reported to have been rendered using deepfake video and synthetic voice. The case was reported by the Hong Kong police in February 2024 and publicly confirmed by Arup in May 2024. Because Arup was a video call rather than a voice-only contact, it is included here as evidence of payment-control failure under a media-rich impersonation attack, and as evidence that combined visual and audio cues can be socially convincing. It is not evidence about the reliability of voice cloning in a voice-only channel, and no independent forensic analysis of the audio or video has been published. Evidence status: alleged; reported by law enforcement but not forensically established in the public record (CNN, 2024).

In both cases the attacker is reported to have impersonated a senior figure and asked for a transfer. The public record does not isolate the contribution of the voice in either case: the energy-firm CEO described the caller's accent and speech melody, and the Arup employee is reported to have been deceived in a video call. Neither case has been confirmed by published forensic evidence as involving AI-generated voice as distinct from other impersonation techniques or a combination of techniques. The two incidents therefore illustrate the payment-control failures that this article addresses without, on their own, quantifying the prevalence of voice-cloning fraud.

Regulatory warnings. There are clear warnings from police and regulators. In December 2024, the FBI issued a public service announcement that criminals use generative AI to create “short audio clips containing a loved one's voice” to commit family-emergency scams, as well as to “obtain access to bank accounts” via voice impersonation (FBI IC3, 2024). In May 2025, the FBI issued an advisory about the use of AI-generated voice messages in an operation targeting senior U.S. government officials (FBI IC3, 2025b). In November 2024, the U.S. Financial Crimes Enforcement Network (FinCEN) released Alert FIN-2024-Alert004, which describes deepfake media used against financial institutions and their customers, chiefly to circumvent identity verification with fabricated identity documents, and asks institutions to reference the key term FIN-2024-DEEPFAKEFRAUD in Suspicious Activity Reports (SARs) where they identify related activity (Synthetic Media refers to any content that appears genuine but is actually fake) (FinCEN, 2024). Since 2023, the U.S. Federal Trade Commission (FTC) has cautioned consumers that voice cloning needs only “a short audio clip” and “a voice-cloning program” to impersonate a family member and advises contacting the supposed relative on a phone number the consumer knows to be theirs (FTC, 2023).

Human Detection of Synthetic Speech

A central question for payment verification is whether a human listener can distinguish a genuine voice from a synthetic one when no other cues are available. Four experiments are directly relevant. The first, by Mai et al. (2023), used clean recorded clips rather than telephone audio, recruited 529 participants through Prolific and tested their ability to discriminate genuine from deepfake speech in English and Mandarin under two task conditions: a unary task (a single clip judged as “real” or “fake”) and a binary task (a genuine clip and a deepfake clip presented together for comparison). The 100 clips came from one female speaker per language, synthesised with one pre-trained text-to-speech system that the authors describe as older than the state of the art; listeners were told that deepfakes were present, and about half the clips were fake.

In the unary task, participants correctly classified deepfake clips on approximately 73% of trials and genuine clips on 67.78% of trials. Overall accuracy in the unary task was 70.35%, which was above chance but well below the level required to authenticate a payment request based on a voice alone. Approximately one in three genuine clips was judged to be fake, which the authors attribute partly to listeners' awareness that deepfakes were present. In the binary task, in which listeners could compare a genuine clip with a deepfake clip side by side, accuracy increased to 85.59%. However, this comparative condition does not reflect the operational situation of receiving an inbound call, where no reference clip is available. Performance did not differ significantly between English and Mandarin (Mai et al., 2023).

A brief familiarization intervention, in which participants were exposed to examples of deepfake speech before the test, increased confidence weighted accuracy by an average of 3.84% ($p = 0.001$). The authors describe this improvement as slight, conclude that detection capability remains unreliable, and consider efforts to substantially improve human detection to be unrealistic (Mai et al., 2023). The practical implication for organizations is not that human listeners are incapable of detecting deepfakes. A 73% single clip detection rate is meaningfully above chance. Rather, the evidence indicates that a short awareness raising session is unlikely to increase human detection accuracy to a level sufficient for authenticating a payment request.

The second experiment, by Barrington et al. (2025), used a current commercial system. The authors took recordings of 220 US speakers collected through Prolific, built a clone of each speaker with the ElevenLabs Instant Voice Cloning service from two recorded prompts, and generated cloned versions of the speakers' other clips (mean clip length 5.37 seconds). In an identity study, 304 listeners heard two clips back to back and said whether they came from the same person; they were not told that some voices were AI-generated. In a naturalness study, 300 listeners judged single clips as real or AI-generated; half the clips were AI-generated, and listeners first heard two real and two AI examples. Listeners judged a clone and the real counterpart to be the same identity about 80% of the time, and identified AI-generated voices correctly only about 60% of the time. Longer clips and unscripted speech were associated with better naturalness judgments, and the authors suggest that engaging a caller in a longer, open-ended conversation may help listeners (Barrington et al., 2025).

This study improves on Mai et al. in the number of speakers and the quality of the generator, but two limits remain for payment calls. The speakers were strangers to the listeners, so the identity task measures how well listeners match two unfamiliar voices, not whether they recognise a voice they know, and the clips were full-band recordings rather than telephone audio. The authors also note that listeners' attention was

fully on the voice and that people may be less attentive in real calls. The senior author declares a financial interest in a company that develops techniques to detect AI-generated content (Barrington et al., 2025).

The third, by Müller, Pizzi, et al. (2022), was an online game. Players decided one clip at a time whether English read speech from the ASVspoof 2019 evaluation set was authentic or fake, and were then shown the true label and the prediction of a detection model; the game over-sampled attacks that earlier players had found harder. The published abstract reports 410 users and 13,229 rounds. Players and the detector struggled with the same types of attack, IT professionals had no advantage over other players, and native English speakers did better than non-native speakers (Müller, Pizzi, et al., 2022). The clips came from older generators and clean studio speech, players received feedback after every round, and the speakers were unfamiliar, so the accuracy figures do not transfer directly to a payment call.

The fourth, by Warren et al. (2024), is the largest study to date. The authors recruited 1,212 US participants through Prolific (native English speakers with no hearing difficulty, using headphones). Each participant judged 20 clips from one of three widely used datasets (WaveFake, the ASVspoof 2021 deepfake track or FakeAVCeleb) as human or computer-generated, without feedback, and three participants judged each clip. Mean individual accuracy was 65% on WaveFake, 71% on ASVspoof 2021 and 81% on FakeAVCeleb, and the authors summarise average human accuracy as 73%. By majority vote, listeners classified genuine audio better than deepfake audio on every dataset, by 15, 18 and 21 percentage points, and they accepted more deepfakes as human than they rejected genuine clips as fake. Participants relied on prosody, accent, breathing and background noise, and some held preconceptions, such as that computers cannot produce accents, that led them to wrong answers. Detectors retrained on each dataset did not simply outperform people: listeners and detectors made different errors, with listeners prone to false negatives and detectors to false positives (Warren et al., 2024). The authors separate deepfake detection from speaker verification and note that a listener who knows the target's voice may use that knowledge; their FakeAVCeleb clips use celebrity voices, but the study did not test whether participants knew them.

Taken together, the four experiments show that listeners who are not known to be familiar with the speaker detect synthetic speech above chance, with reported accuracies from about 60% to 86% depending on the generator, dataset and task, but not reliably. The studies differ on the direction of error: Mai et al. found that about a third of genuine clips were judged fake, whereas Warren et al. found that listeners accepted more deepfakes as human than they rejected genuine clips. Barrington et al. found that listeners often accepted a commercial clone as the same speaker. None of the experiments tests the situation that matters

most for payment fraud, a listener known to be familiar with the caller's voice. The conclusion that voice similarity is unreliable for authentication is therefore an inference from related tasks, not a direct finding.

Automated Detection Systems and Their Limitations

There has been active research into automated deep fake speech detection, in part due to the ASVspoof challenge series. Several surveys offer detailed information on the advancement of the field and its ongoing challenges.

Detection architectures: Khanjani et al. (2023) review the various detection methods from deep convolutional networks, ResNet architectures, recurrent networks (biLSTM), and temporal convolutional networks (TCN). Almutairi and Elgibreen (2022) report that the type of method affects detection more than the acoustic features used, and they note difficulties with accented speech and environmental noise.

Benchmark performance vs. real-world generalization: On the ASVspoof 2019 Logical Access (LA) benchmark, the standard evaluation set for text-to-speech and voice-conversion spoofing attacks, top systems achieve an equal error rate (EER) below 2%. Müller, Czempin, et al. (2022) re-implemented twelve detection architectures, trained them on ASVspoof 2019 data only, and evaluated them on a new “in-the-wild” set of 37.9 hours of recordings of 58 celebrities and politicians, 17.2 hours of which were deepfakes (average clip length 4.3 seconds). The authors summarise the deterioration relative to the benchmark as roughly 200 to 1,000 percent. The underlying values are in Table 1 of Müller, Czempin, et al. (2022). With full-length input, RawGAT-ST had an EER of 1.2% on the ASVspoof 2019 evaluation data and 37.2% on the in-the-wild set; RawNet2, 3.2% and 37.8%; an LCNN with constant-Q features, 6.4% and 65.6%; and a Transformer with constant-Q features, 7.5% and 43.8%. Across all the configurations they tested, in-the-wild EERs ranged from 33.9% to 91.1%, where 50% is chance. Because relative changes depend on a low benchmark baseline, the absolute rates are the more informative figures. Müller et al. (2024), using four detection models and ASVspoof 2019 and 2021 data, decomposed the degradation into two components: “hardness” (new deepfakes being intrinsically harder to detect) and “difference” (models being tested against generators they were not trained on). For the 2021 deepfake set and the in-the-wild set, the gap (14 to 27 and 30 to 78 EER points across the four models) came almost entirely from “difference”. The practical reading is not that every detector will fail on every payment call. In-distribution performance can be excellent, and vendor systems that are continually updated with new generators may hold up better than the frozen research systems tested in these studies. The reading is that a detector's benchmark EER is a poor guide to its accuracy on audio produced by a generator, or through a channel, that it has not encountered.

Telephony conditions: The ASVspoof 2021 challenge added a Logical Access track in which bona fide and spoofed utterances were transmitted over real voice-over-IP and public switched telephone network (PSTN) systems with several codecs, including a-law, μ -law and G.722. The organisers describe the data as intentionally more challenging than the 2019 edition. Even so, the best submitted system reached an EER of 1.32% on this track, and the organisers report that logical-access results were close to those of earlier editions despite the added channel variability; the four baseline systems reached pooled EERs of about 9% to 19% (Yamagishi et al., 2021). The report does not compare each codec with clean audio, so it does not show how much telephony encoding alone cost. On the deepfake track, which used compressed media audio and more than 100 spoofing algorithms, the best evaluation EER was 15.64%, and the organisers attribute the gap from progress-phase results to overfitting. In a separate analysis, four models trained on clean ASVspoof 2019 data lost 10.7 to 20.5 EER points on the telephony-encoded version of the same attacks, mostly because of “difference”, so the effect of the channel depends on how a detector was trained (Müller et al., 2024). Because almost all payment calls traverse a telephony network of one kind or another, this is the setting in which any deployed detector would have to operate. The ASVspoof 2021 results therefore show that detectors can be built to cope with telephony channels, while the generalisation problem described above remains.

Emerging countermeasures: The FTC Voice Cloning Challenge was announced in November 2023, and four winning submissions were named in April 2024: an algorithm that distinguishes genuine from synthetic voice patterns (AI Detect), a watermarking technique that distorts voice samples so that they are harder to clone (DeFake), a method that authenticates the human origin of a recording at the point of creation (OriginStory), and a real-time voice-clone detector from Pindrop Security, which received a recognition award (FTC, 2024). These approaches address different stages of the problem, from prevention to detection. The sources reviewed contain no independent validation of any of them at scale in live telephony.

Summary of detection evidence: On the evidence reviewed, neither human listeners nor current publicly reported automated detectors are, on their own, adequate to authenticate a voice on an inbound payment call. Human listeners perform above chance at detecting clones but far below what an authentication decision requires, and in one study often accepted a commercial clone as the same speaker (Barrington et al., 2025), and in another study listeners and trained detectors made different errors, so neither is a clean fallback for the other (Warren et al., 2024); automated detectors generalise poorly across unseen generators, and telephony transmission adds a further source of variability. This does not mean that detection has nothing to offer — vendor products, watermarking and liveness scoring may prove useful as they mature

— but it does argue against treating voice similarity, or any single-layer detection tool, as sufficient evidence of caller identity. The procedural response is set out in the next section.

Proposed Payment Verification Approach

Proposed Payment Verification Process

The following four-step process is offered as guidance, not as a validated fraud-reduction protocol. It draws mainly on established anti-fraud principles (dual authorisation, out-of-band verification, segregation of duties) and on law-enforcement and regulator advice; the evidence in Section 3 explains why voice alone cannot be relied on, but it does not test these controls. No claim is made that the combined protocol has been empirically tested or that it reduces voice-cloning fraud by any specific percentage. Each step is labelled with its evidence status (established practice, regulatory guidance, or the author's recommendation informed by the evidence) and identifies the person or role expected to act.

Step 1: Pause the payment (Author's recommendation, informed by evidence). Any staff member who receives an urgent call, voice message or written instruction requesting a payment or a change of bank details (most often the finance officer, accounts-payable clerk or executive assistant handling the request) should suspend action on that request pending verification, regardless of the seniority of the apparent requester. Time pressure is a recurring feature of impersonation-based payment fraud. The UK energy-firm caller is reported to have pressed for an urgent transfer (Stupp, 2019), and the FBI's 2024 advisory describes AI-generated voice scams built around a crisis and a request for immediate financial assistance (FBI IC3, 2024). Payment policy should give the person handling the request explicit authority to hold the payment without seeking further approval; release should require positive confirmation from the designated authoriser after Steps 2–4 have been completed.

Step 2: Call back on a pre-existing verified number (Established practice, supported by law-enforcement and consumer guidance). The recipient or, for higher-risk payments, a designated verifier separate from the person who received the request should call the requester back on a number obtained from a pre-existing record that predates and is independent of the current request. Acceptable pre-existing records include: the organisation's own HR or supplier master data; a signed contract or purchase order on file; a previous invoice paid to the same payee; or a directory entry that has been in place for at least one prior verified interaction. Numbers supplied in the request itself, in the email accompanying it, in a signature block on a new document, or on a website whose URL was provided in the request, are not pre-existing records for this purpose. Independent callback is advised in FBI guidance. The 2024 advisory tells recipients to hang

up and call the number of the organisation concerned directly, and the 2025 advisory tells recipients to identify a phone number for the purported caller independently and call to verify (FBI IC3, 2024; FBI IC3, 2025b). FTC consumer guidance gives the same advice for family-emergency calls (FTC, 2023). These sources are addressed to consumers and, in the 2025 case, to government officials and their contacts, not to accounts-payable staff, so they support the principle rather than the specific procedure set out here. FinCEN's alert (FinCEN, 2024) is addressed to financial institutions and names multifactor authentication and live verification checks as mitigations; it does not set out a payment callback procedure. Callback addresses the core weakness of voice-cloning attacks: if the attacker controls only the inbound channel, a call to a number from a pre-existing record reaches the genuine person and does not depend on how the inbound voice sounds. Organisations should maintain, and periodically re-verify, a directory of authorised requesters and their contact details.

Step 3: Confirm any new or changed bank details through a genuinely independent channel (Author's recommendation, informed by established practice). Any new bank account, or any change to existing bank details, should be confirmed through a channel that the attacker is unlikely to control at the same time as the requesting channel. A reply to the email address that accompanied the request is not automatically independent (that mailbox may be exactly what the attacker has compromised), and neither is a letter on company letterhead, which can be forged or sent from a compromised office. Genuinely independent channels include: a callback on the pre-existing verified number from Step 2; a face-to-face confirmation with the named requester or, for a supplier, with a previously known contact at the supplier organisation; or a message on an organisation-managed secure messaging platform where account provisioning is controlled by IT, not by the requester. Where email is the only feasible channel, the reply should go to an address taken from the pre-existing supplier record rather than from any address supplied with the request, and a second confirmation should be sought on a different medium. Every change of bank details should be recorded with the date, the verifier and the channel used.

Step 4: Require a second approver for higher-risk payments (Established practice). Payments meeting the risk criteria set out below should require positive approval from a second authorised individual who is organisationally independent of the requester and of the first approver, before the payment is released. The second approver should act after Steps 1–3 have been completed and documented, and should be expected to review the callback record and the bank-details confirmation, not merely to sign off on the amount. Dual authorisation is a long-standing anti-fraud control in financial operations; the ACFE's Occupational Fraud 2024 report found that a lack of internal controls (32% of cases) or an override of existing controls (19%) contributed to more than half of the occupational fraud cases studied (ACFE, 2024). Those cases involve insiders, not external impersonators, so the finding supports the general value of internal controls but does

not measure the effect of dual approval against voice-cloning fraud, and no such effect has been quantified in the reviewed literature.

Defining "higher-risk" payments (Author's recommendation; illustrative thresholds). Each organisation should set its own thresholds, calibrated against its own transaction profile, risk appetite and regulatory obligations. A workable starting point is to review the distribution of the organisation's own outbound payments over the previous 12 months and to set the monetary threshold near a percentile (for example, the top 5–10%) at which additional verification is operationally tolerable. The USD 10,000 figure below is used purely to illustrate the shape of such a threshold and is not a recommended universal value. Criteria to consider include:

- Any single payment above a defined monetary threshold (for example, USD 10,000 used here only as an illustration; the appropriate figure depends on the organisation).
- Any payment to a new payee, or to a payee whose bank details have changed within a defined recent window (for example, the previous 30 days).
- Any payment requested outside normal business hours or through a channel that bypasses the standard approval workflow.
- Any payment in which the request emphasises urgency, secrecy or confidentiality, or discourages independent verification.

These criteria and any monetary thresholds should be documented in the organisation's payment policy, reviewed at least annually, and updated when the organisation's payment profile changes materially. All four steps above are proposed as guidance rather than as measured fraud-reduction controls: their empirical effect as a combined protocol has not been tested in the reviewed literature. Table 5 summarises what each safeguard protects against, its limitations and its evidence status.

Table 5: *Safeguards for Verifying Payment Requests When Voice Identity Is Uncertain*

Safeguard	What It Protects Against	Limitations	Evidence Status
Pause the payment	Prevents impulsive compliance with urgent, fraudulent requests	Requires organisational culture that tolerates delay; does not itself verify identity	Supported by documented fraud patterns (FBI IC3, 2024); no study reviewed here tests a pause as an intervention
Callback on a previously verified number	Bypasses voice impersonation by initiating contact on a trusted channel	Requires a maintained, up-to-date verified contact directory; does not protect if the attacker also controls the callback number (e.g., SIM swap,	Advised in FBI IC3 (2024, 2025b) and FTC (2023) guidance; not empirically quantified

		call forwarding) or has altered the contact record	
Independent confirmation of new bank details	Prevents fraudulent redirection of payments to attacker-controlled accounts	Adds processing time; requires a second verified contact channel; may be bypassed if the attacker compromises email	Author's recommendation based on out-of-band verification principles; not tested as an intervention
Dual authorisation for higher-risk payments	Adds a second checkpoint; a single compromised individual cannot authorise payment alone	Thresholds must be calibrated; adds delay; ineffective if both approvers are compromised or collude	Established anti-fraud control; ACFE (2024) links weak or overridden controls to occupational fraud, but the effect of dual approval against voice-cloning fraud is not quantified
Human listening for voice artifacts	May detect low-quality synthetic speech with audible artifacts	Listeners not known to be familiar with the speaker were correct about 60% to 86% of the time, depending on study, generator and task; training improved confidence-weighted accuracy by about 4% (Mai et al., 2023); a commercial clone was judged the same identity as the real speaker about 80% of the time (Barrington et al., 2025)	Tested experimentally; found unreliable as a standalone safeguard (Mai et al., 2023; Müller, Pizzi, et al., 2022; Barrington et al., 2025; Warren et al., 2024)
Automated deepfake detection software	Can identify synthetic speech patterns invisible to human listeners	Severe generalisation failures on unseen generators (Müller, Czempin, et al., 2022; Müller et al., 2024); telephony adds channel variability (Yamagishi et al., 2021); not independently validated at scale in live telephony	Tested on benchmarks; real-world performance substantially lower; prototype solutions emerging (FTC, 2024)

Discussion

The evidence reviewed does not support treating voice similarity as an authentication factor for high-risk payment requests. Across four laboratory studies, listeners not known to be familiar with the speaker performed above chance at detection but well below the level that authentication requires (Mai et al., 2023;

Müller, Pizzi, et al., 2022; Warren et al., 2024), brief training produced only marginal gains (Mai et al., 2023), and listeners often accepted a commercial clone as the same speaker (Barrington et al., 2025). None of these studies tested a listener known to be familiar with the speaker, so the conclusion for familiar voices is an inference. Automated detectors achieve strong in-distribution performance but generalise poorly to unseen generators (Müller, Czempin, et al., 2022; Müller et al., 2024), and telephony transmission adds further channel variability (Yamagishi et al., 2021).

The incident evidence should be handled with corresponding care. Many widely cited cases have been described in the press as AI-enabled deepfake fraud, yet in the two cases discussed here no public forensic analysis of the audio has been published. These cases illustrate that impersonation-driven payment fraud is a real operational problem; they do not, on their own, quantify how much of that problem is attributable specifically to synthetic voice.

In this setting, a procedural response is more tractable than a detection-based response. The four-step protocol proposed in Section 4.1 (pause, callback on a pre-existing verified number, independent bank-detail confirmation and a second approver for higher-risk payments) is grounded in established anti-fraud principles and is designed to remain useful even if the underlying detection landscape shifts. It has not been tested as a combined intervention, and Section 6.1 sets out how it might be evaluated.

The protocol also has practical limits. It presupposes at least two authorised approvers and a maintained directory of verified contacts, which very small organisations may lack. It adds delay, which matters for legitimate urgent payments such as payroll or tax deadlines, so organisations should decide in advance who may authorise an exception and how it is recorded. Callback and dual approval both fail if the attacker controls the callback number or deceives both approvers (Table 5). The guidance sources cited here are mostly US-based, so organisations elsewhere should check their own regulator's advice. These points are the author's reasoning, not findings from the reviewed literature.

Limitations

The results and recommendations of this review have several important limitations that should be taken into account when interpreting them.

First, the review is not systematic and did not follow a systematic review protocol. One reviewer carried out the literature search and screening without duplication. Relevant studies might have been overlooked, such as non-English literature, conference proceedings not captured in the searched databases, and grey

literature in the private sector. The research evidence is also small: nine of the 19 cited sources are research studies or reviews, four of them test human listeners (Mai et al., 2023; Müller, Pizzi, et al., 2022; Barrington et al., 2025; Warren et al., 2024), and none tests listeners known to be familiar with the speaker. None of the original search strings targeted studies of human listeners, so three of these studies (Barrington et al., 2025; Müller, Pizzi, et al., 2022; Warren et al., 2024), all published inside the evidence window, were added at revision outside the search protocol; other relevant studies of human listeners may be missing. The search record is incomplete (Section 2.2), and the evidence cut-off and seed references were predetermined (Section 2.1).

Secondly, there is limited evidence of voice cloning fraud. In the two most frequently cited examples, involving a UK energy company and Arup, both incidents have been reported as involving deepfake technology. However, neither case has been substantiated through published forensic audio analysis. The absence of confirmed cases does not mean that voice cloning fraud is not being used. It may instead reflect victims' reluctance to disclose information, underreporting of incidents, or the difficulty of detecting and verifying such attacks. Nevertheless, readers should be cautious about treating reported incidents as established facts, particularly when claims of voice cloning are not supported by forensic evidence.

Third, the suggested verification process has yet to be tested as a combined protocol. The individual elements are based on proven anti-fraud practice (dual authorisation, callback verification), as well as on evidence on the weakness of voice-based identification (Mai et al., 2023), however, the effectiveness of the combined process in reducing payment fraud has not been measured. The illustrative risk thresholds are provided as a basis for the calibration process for the organisation, but not as validated thresholds. Much of the supporting guidance is addressed to consumers rather than to payment staff, and the evidence on dual approval comes from research on insider fraud (Section 4.1).

Fourth, voice cloning and detection is a very new field. This detection landscape is based on evidence up to November 2025. The quality of synthetic speech and the detection systems' capabilities have probably changed since then. While the procedures outlined in this article may be technology independent, readers are urged to keep abreast of changes in detection technologies and vendor assessments in the literature before using any particular detection technology.

Fifth, this review is for telephone calls and voice messages calling for payments. It does not mention voice cloning in other situations (such as identity theft, misinformation, personal harassment), nor multimodal deepfakes (combined video and voice), except in the context of the Arup case.

Conclusion

AI-generated voice impersonation is treated by regulators and law-enforcement bodies as a live and growing risk to organisational payment processes, as reflected in advisories issued by the FBI, FinCEN and the FTC and in reports of high-value payment fraud. Independent forensic confirmation that any of the most-cited cases involved synthetic voice is, however, not yet available in the public record, and this article treats the risk as real and plausibly rising without asserting a specific trend.

Two conclusions follow from the evidence reviewed. First, a familiar-sounding voice on the telephone is no longer sufficient evidence of the caller's identity for the purposes of authorising a payment: human listeners are meaningfully but insufficiently accurate, brief training does not close the gap, and listeners in one large study often accepted a commercial clone as the same speaker (Barrington et al., 2025). Second, current publicly reported automated detection systems perform strongly on in-distribution benchmarks but degrade on unseen generators and face added channel variability on telephony audio, so they should not, at present, be relied on as the primary safeguard against voice-cloning fraud in a payment call.

In the absence of a reliable detection based defence, procedural verification is the most readily available line of protection. The four-step approach proposed in Section 4.1, which involves pausing the payment, calling back using a preexisting verified number, confirming any new bank details through a genuinely independent channel, and requiring a second approver for higher risk payments, draws on long standing anti-fraud principles and addresses the specific weakness exploited by voice cloning attacks. This approach is presented as practical guidance rather than as a validated intervention, and no specific reduction in fraud is claimed.

For the protocol to have any effect it needs to be embedded in payment policy, supported by training that expects staff to complete the verification steps even when the request appears to come from a senior authoriser, and reviewed periodically as voice-generation and detection technology evolve. Future research directions are set out in Section 6.1.

Future Empirical Research

Future empirical work should test whether the verification protocol proposed here actually reduces the acceptance of fraudulent payment requests in realistic organisational settings. A useful design would place employees or teams in simulated payment-fraud scenarios (urgent inbound calls, AI-generated voice messages, requests to change banking details, and layered social-engineering pressure) under two or more verification regimes and would compare their behaviour.

Key outcome measures include the rate at which fraudulent requests are accepted, the rate at which legitimate requests are incorrectly rejected, compliance with each step of the protocol, time to complete verification and the frequency with which controls are bypassed. Studies should also vary the attacker's capabilities, in particular whether the attacker controls only the inbound channel or multiple channels (email, phone number, calendar) simultaneously, to test whether the protocol still holds under more sophisticated attacks and to provide empirical grounding for what is, at present, a reasoned proposal rather than a measured intervention.

REFERENCES

- ACFE. (2024). *Occupational fraud 2024: A report to the nations*. Association of Certified Fraud Examiners. <https://www.acfe.com/en-us/-/media/files/acfe/pdfs/rtn/2024/2024-report-to-the-nations.pdf>
- Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5), 155. <https://doi.org/10.3390/a15050155>
- Barrington, S., Cooper, E. A., & Farid, H. (2025). People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15, 11004. <https://doi.org/10.1038/s41598-025-94170-3>
- CNN. (2024, May 16). Arup revealed as victim of \$25 million deepfake scam involving Hong Kong employee. *CNN Business*. <https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk>
- Damiani, J. (2019, September 3). A voice deepfake was used to scam a CEO out of \$243,000. *Forbes*. <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>
- FBI IC3. (2024, December 3). Criminals use generative artificial intelligence to facilitate financial fraud (PSA I-120324-PSA). Federal Bureau of Investigation, Internet Crime Complaint Center. <https://www.ic3.gov/PSA/2024/PSA241203>
- FBI IC3. (2025a). 2024 Internet crime report. Federal Bureau of Investigation, Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf
- FBI IC3. (2025b, May 15). Senior US officials impersonated in malicious messaging campaign (PSA I-051525-PSA). Federal Bureau of Investigation, Internet Crime Complaint Center. <https://www.ic3.gov/PSA/2025/PSA250515>
- FinCEN. (2024, November 13). FinCEN alert on fraud schemes involving deepfake media targeting financial institutions (FIN-2024-Alert004). U.S. Department of the Treasury, Financial Crimes

- Enforcement Network. <https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf>
- FTC. (2023, March 20). Scammers use AI to enhance their family emergency schemes. Federal Trade Commission, Consumer Advice. <https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes>
- FTC. (2024, April 8). FTC announces winners of Voice Cloning Challenge. Federal Trade Commission. <https://www.ftc.gov/news-events/news/press-releases/2024/04/ftc-announces-winners-voice-cloning-challenge>
- Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in Big Data*, 5, 1001063. <https://doi.org/10.3389/fdata.2022.1001063>
- Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*, 18(8), e0285333. <https://doi.org/10.1371/journal.pone.0285333>
- Müller, N. M., Czempin, P., Diekmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize? In *Proceedings of Interspeech 2022* (pp. 2783–2787). ISCA. <https://doi.org/10.21437/Interspeech.2022-108>
- Müller, N. M., Evans, N., Tak, H., Sperl, P., & Böttinger, K. (2024). Harder or different? Understanding generalization of audio deepfake detection. In *Proceedings of Interspeech 2024* (pp. 2705–2709). ISCA. <https://doi.org/10.21437/Interspeech.2024-247>
- Müller, N. M., Pizzi, K., & Williams, J. (2022). Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22)* (pp. 85–91). ACM. <https://doi.org/10.1145/3552466.3556531>
- Stupp, C. (2019, August 30). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *The Wall Street Journal*. <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>
- Warren, K., Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedeles, C., Pasternak, M., Layton, S., Butler, K., Gates, C., & Traynor, P. (2024). “Better be computer or I’m dumb”: A large-scale evaluation of humans as audio deepfake detectors. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)* (pp. 2696–2710). ACM. <https://doi.org/10.1145/3658644.3670325>
- Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (pp. 47–54). ISCA. <https://doi.org/10.21437/ASVSPOOF.2021-8>