

VoiceGuard – Deepfake Voice Cloning in Vishing Attacks: Threat Landscape, Detection Gaps, and Countermeasures in the Indian Context

Aishwarya Garje¹, Aditya Walunj², Anurag Prajapati³

^{1,2,3}Master of Computer Applications, D Y Patil Schools of Humanities & Sciences, Navi Mumbai, India

Abstract - AI-based speech synthesis has lowered the barrier to convincing voice impersonation. Modern zero-shot text-to-speech and voice-conversion systems can reproduce a speaker's timbre from very short audio prompts, and attackers can pair the result with public audio, social engineering, caller-ID manipulation and urgent payment requests. This paper examines deepfake vishing as a socio-technical threat, with emphasis on India. It reviews how voice cloning works, recent reported incidents, the limits of current detection, and practical countermeasures. Our central finding is that a voice-only trust decision is no longer safe: detection models generalise poorly across languages, codecs, noise conditions and unseen generators, while human listeners are easily swayed by context and urgency. We therefore propose VoiceGuard, a layered framework built on transaction-risk rules, independent call-back, voice biometrics as a supporting signal rather than sole proof, device and account context, awareness training for staff and families, rapid reporting and continuous testing. The paper is a literature and case-study review, not a prevalence study; the reported incidents should not be read as an estimate of national incidence.

Key Words: Deepfake, Voice Cloning, Vishing, Social Engineering, Synthetic Media, Voice Biometrics, Cybersecurity Awareness, AI Fraud.

1. INTRODUCTION

People routinely recognise relatives, managers and colleagues by their voice, which makes voice a natural, if informal, authentication cue. Conventional phishing defences concentrate on suspicious links, look-alike domains, spelling errors or visual design. A phone call removes most of those cues and puts the victim in a high-pressure, real-time conversation.

Deepfake vishing combines synthetic speech with social engineering. The attacker does not need a perfect clone; a plausible voice, a familiar identity, a believable scenario and a time-critical request are often enough to trigger action. The security problem therefore shifts from spotting a visibly fake artefact to checking whether the trusted person actually made the request.

India is a relevant setting because digital payments are widespread and vishing is already a major category of payment fraud [3], [4]. In the Reserve Bank of India's nationwide survey on awareness and use of digital

payments, vishing was the most prevalent method among the frauds or attempted frauds reported by respondents, at 54.2% [4].

Table -1: Summary of key observations and implications.

Area	Key observation	Implication
Attack enabler	Few-shot and zero-shot voice synthesis can imitate a speaker from short audio prompts [10], [11].	Public audio can become attacker input.
Human factor	Urgency, authority and familiarity can override normal scepticism [1].	Voice-only trust is unsafe for sensitive actions.
Detection	Performance can degrade under unseen generators, languages, codecs and noise [8], [9].	Laboratory accuracy is not a deployment guarantee.
Authentication	Caller ID and voice similarity are not sufficient proof of authority [1], [6], [7].	Use an independent channel and step-up approval.
India response	I4C runs the 1930 cyber financial fraud helpline [5]; CERT-In has issued deepfake guidance [1], [2].	Rapid reporting and verification belong in the control design.

2. OBJECTIVES AND RESEARCH QUESTIONS

The study has four objectives: (1) to explain how voice cloning enables deepfake vishing; (2) to map the threat using recent global and India-specific reported cases; (3) to identify technical and operational gaps in detection and authentication; and (4) to propose a practical, layered countermeasure framework for Indian consumers and organisations.

Four research questions guide the work.

- **RQ1:** How do modern voice-cloning systems support real-time or near-real-time impersonation?
- **RQ2:** Why do laboratory detection results often fail to carry over to live calls?

- **RQ3:** Which social-engineering patterns recur in reported Indian incidents?
- **RQ4:** Which combinations of technical, procedural and awareness controls reduce dependence on voice alone?

3. METHODOLOGY

We carried out a structured narrative review organised around the attack lifecycle rather than around any single detection model. Sources fall into three groups: Indian government and regulatory guidance from CERT-In, RBI and I4C [1]–[5]; technical research and benchmark programmes such as ASVspoo and NIST [6]–[12]; and incident reports from established news organisations [13]–[16].

Evidence was classified as primary or official guidance, peer-reviewed or preprint technical research, and incident reports. Incident reports are used only to illustrate attack patterns and not to estimate prevalence. Technical claims are stated with their original scope and limitations.

4. TECHNICAL MECHANISM OF DEEPPAKE VOICE VISHING

Voice cloning can be built on text-to-speech (TTS), voice conversion (VC) or a hybrid of the two. A speaker encoder, or a comparable representation, captures the characteristics of a target speaker, and a synthesis or decoder stack then produces speech with that timbre. Zero-shot systems reduce the amount of target-specific audio that is needed.

ControlSpeech, for example, describes zero-shot speaker cloning from only a few seconds of audio, with additional control over speaking style [10]. MobileSpeech points toward fast, lightweight zero-shot TTS suited to constrained devices [11]. These results do not mean every criminal tool matches their quality or latency, but they explain why a short public clip can be a valuable input for an attacker.

A typical attack chain runs from open-source intelligence (OSINT) collection, target selection and voice acquisition, through cloning and pretext construction, to call delivery, urgency and authority cues, a payment or credential request, money movement and, finally, delayed verification (Fig. 1).

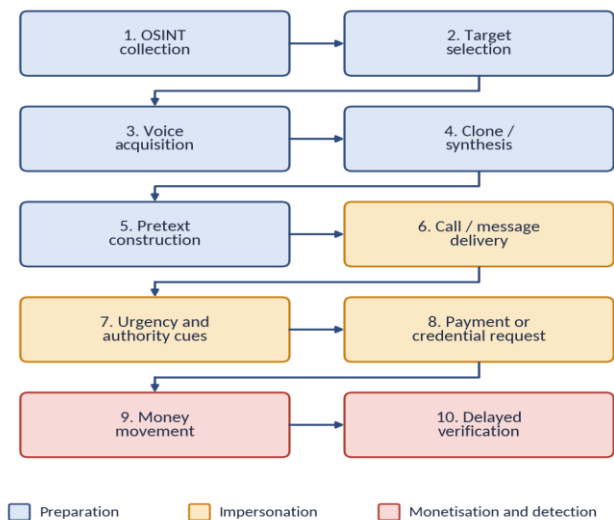


Fig -1: Lifecycle of a deepfake vishing attack, from preparation to monetisation.

Real-time voice conversion can blur the line between pre-recorded audio and a live conversation. In practice, quality still depends on the model, hardware, network, language, microphone, background noise and the attacker's operational setup.

5. THREAT LANDSCAPE

Deepfake vishing sits where synthetic media meets classic social engineering. Pretexts seen in reported cases include family emergencies, executive requests, law-enforcement or embassy claims, account-security alerts, investment offers and urgent vendor payments [13]–[16].

The core weakness is treating voice as evidence. A familiar voice creates a false sense of continuity: the victim may confirm who is speaking without asking whether the channel is authentic or whether the request itself is legitimate. Attackers can further combine voice cloning with spoofed caller IDs, WhatsApp or Teams accounts, leaked organisational information, public social-media audio and follow-up messages, which turns the attack into a multi-channel impersonation problem [1], [2].

6. INDIA-SPECIFIC INCIDENT EVIDENCE

Table 2 summarises four incidents reported in the Indian press between April 2024 and June 2026. All figures are as reported and, where cases were still under investigation, remain allegations.

Table -2: Reported deepfake-voice fraud incidents in India used as case illustrations.

Date and place	Reported scenario	Amount	Ref .
Apr 2024, Mumbai	Police arrested two people; an accountant's wife was allegedly called in a cloned version of her brother's voice and induced to transfer money in an emergency pretext.	₹3.8 lakh	[13]
Apr 2024, Mumbai	A 68-year-old businessman reportedly received a call from a caller posing as an Indian Embassy official, with a synthetic version of his son's voice claiming the son had been arrested in Dubai.	₹80,000	[14]
Dec 2024, Ranchi	An alleged AI-cloned voice of a senior railway official was used to request money from a friend under an emergency narrative.	₹2 lakh	[15]
Jun 2026, Delhi / Mumbai	A police case involving alleged voice cloning of a Mumbai company's managing director; a deputy general manager was instructed to make multiple transfers, reportedly routed through numerous accounts.	Alleged total above ₹10 crore	[16]

Across these cases the same themes recur: a familiar identity, urgency, apparent authority, a financial action, and verification that happens only outside the original call. The cases show patterns; they do not establish how common deepfake vishing is in India.

7. DETECTION GAPS

Four gaps limit how far detection can be trusted on live calls; Fig. 2 contrasts benchmark conditions with those of a real phone call.

Generalisation. A detector trained on known generators can fail when it meets a new synthesis system, different compression, re-recording, background noise or an unseen language. A 2026 in-the-wild study reported significant performance degradation across diverse languages and real-world acoustic conditions [8]; broader surveys reach similar conclusions about generalisation [9].

The telephone channel. A phone call is a hostile environment for detectors. Telephony codecs, packet loss, echo cancellation, microphones and speaker playback can

remove or distort the high-frequency artefacts that a detector may depend on [8], [9].

Probabilistic output. Even a strong classifier produces false positives and false negatives. A high laboratory accuracy figure is not a guaranteed decision on an unknown live call [9].

Socio-technical nature. Audio authenticity does not prove that a transaction is legitimate. A genuine executive can make an unusual request, and a synthetic voice can make a convincing one. Voice deepfake detection should therefore complement, and not replace, independent authorisation controls [1].

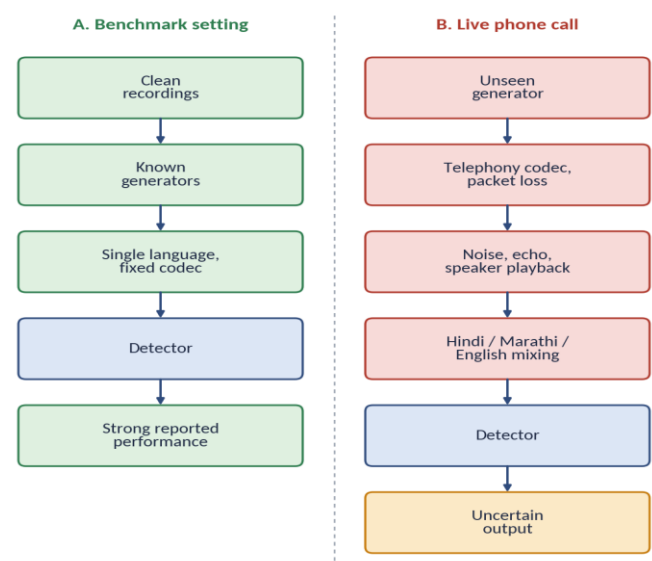


Fig -2: Conditions that separate benchmark evaluation (A) from live phone calls (B).

8. EVALUATION OF AUTHENTICATION MECHANISMS

Knowledge factors, such as questions about family or projects, are weak when the answers can be gathered through OSINT. Caller ID is also insufficient, because the displayed number is not a cryptographic identity [1].

Voice biometrics can add a useful signal, but a speaker-verification system must be spoof-resistant. ASVspoof 5 explicitly studies speech deepfakes and spoofing-robust automatic speaker verification, reflecting the need to test systems against adversarial synthetic speech and not only clean recordings [6], [7]. Evaluation programmes such as NIST SRE24 likewise provide benchmarks for speaker recognition [12].

Strong authentication should therefore be multi-factor and transaction-aware, combining an authenticated application identity, device and session context,

transaction limits, independent call-back, dual approval and a second trusted channel.

9. VOICE GUARD COUNTER MEASURE FRAMEWORK

VoiceGuard organises its controls into six layers (Fig. 3). A suspicious request has to pass through each layer before money or credentials change hands.

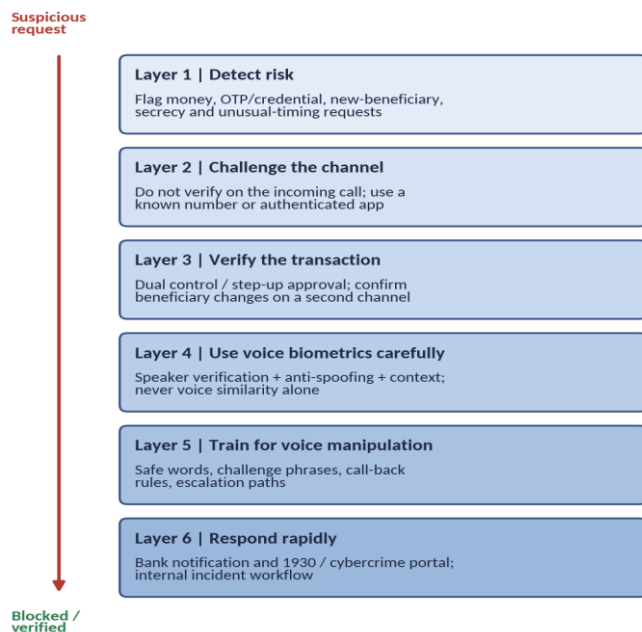


Fig -3: The six-layer VoiceGuard countermeasure framework.

Layer 1: Detect risk. Flag requests involving money, credentials, OTPs, new beneficiaries, urgent secrecy, unusual timing or changes in payment instructions.

Layer 2: Challenge the channel. Do not treat the incoming call as the verification channel. End the call and independently call a known number, or use the organisation's authenticated app or directory [1].

Layer 3: Verify the transaction. Require dual control or step-up authentication for high-value transfers, and confirm beneficiary changes through a separate channel.

Layer 4: Use voice biometrics carefully. Where available, combine speaker verification with anti-spoofing and contextual signals [6], [7]. Never make a high-value financial decision from voice similarity alone.

Layer 5: Train for voice-specific manipulation. Teach staff and families that "it sounds exactly like them" is not proof. Use rehearsed challenge phrases, safe words, call-back rules and escalation paths [1], [2].

Layer 6: Respond rapidly. Report suspected cyber financial fraud immediately through the national cybercrime system and the bank's channels. In India, the I4C cybercrime portal directs victims of cyber financial fraud to the 24×7 helpline 1930 [5].

10. PROPOSED OPERATIONAL PLAYBOOK

Individuals. Pause and do not transfer money. Ask for a call-back and use a saved or official number. Confirm the story with another person. Never share an OTP, PIN or password. If money has already moved, report it immediately (Fig. 4).

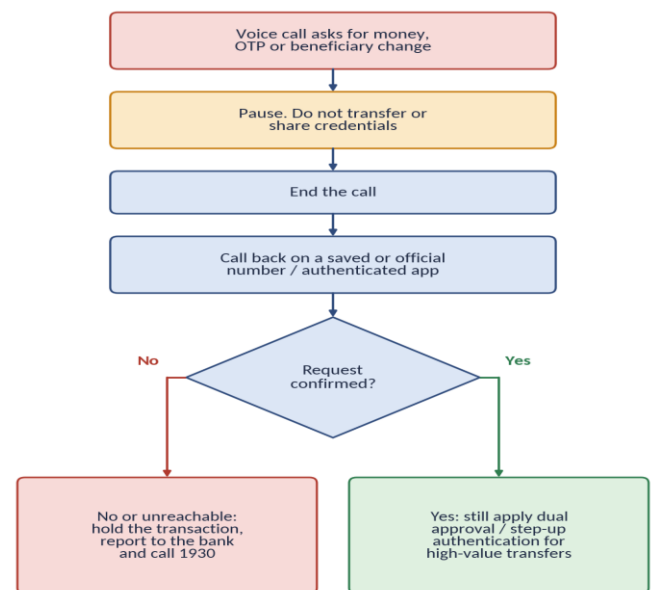


Fig -4: Independent call-back workflow for a suspicious voice request.

Organisations. Classify payment requests by risk and enforce maker-checker approval. Maintain verified contact directories, and disable beneficiary changes that rest on a phone call alone. Run deepfake-vishing simulations, and log and review every exception.

Banks and payment providers. Combine transaction analytics with account and device context, strengthen fraud warnings, and provide fast reporting and response channels. Voice-biometric systems should be evaluated under synthetic-speech attacks and realistic telephony conditions [7], [12].

11. DISCUSSION

The main design principle is to separate identity verification from transaction authorisation. A caller may be exactly who they claim to be and still ask for something they are not authorised to request. Conversely, a synthetic voice can impersonate a legitimate person but should fail

wherever the organisation demands an independent cryptographic or procedural proof.

The Indian context adds multilingual and code-switching complexity. Marathi, Hindi, English and other languages can occur within a single conversation, so detection systems need to be evaluated on representative Indian speech, accents, telephony conditions and noisy environments, and not only on clean benchmark clips [8].

Awareness programmes should also avoid a detector-only mindset. CERT-In's 2024 deepfake advisory and its 2026 guidance on AI-assisted exploitation both stress verifying voice and video messages and being cautious about AI-driven impersonation and social engineering [1], [2].

12. VOICEGUARD REFERENCE ARCHITECTURE

Fig. 5 and Table 3 show how the layers fit together as a decision pipeline, from the incoming call to the incident response.

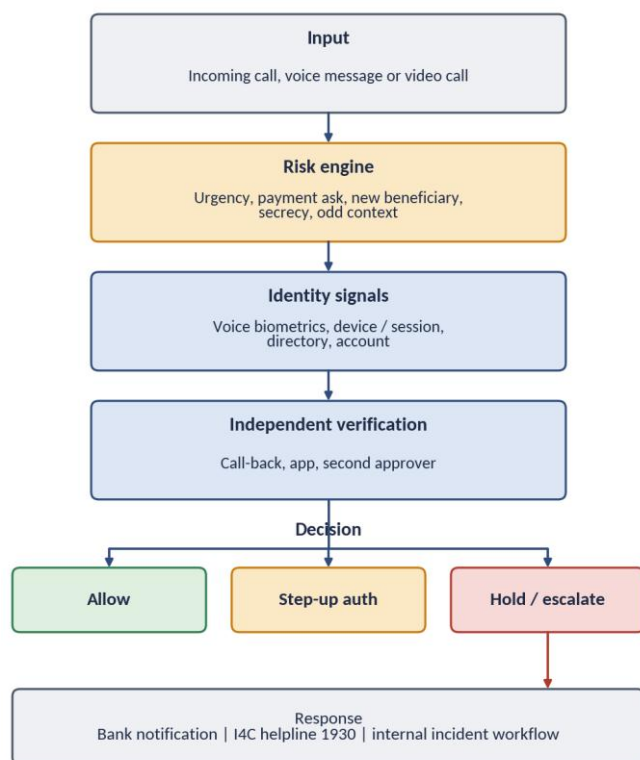


Fig -5: VoiceGuard reference architecture and decision flow.

Table -3: Components of the VoiceGuard reference architecture.

Component	Control
Input	Incoming call, voice message or video call
Risk engine	Urgency, payment request, new beneficiary, secrecy and unusual context
Identity signals	Voice biometrics, device and session context, known directory, account context
Independent verification	Call-back on a known number, authenticated app, or second-person approval
Decision	Allow, step-up authentication, hold, or escalate
Response	Bank notification, I4C helpline 1930 [5], internal incident workflow

13. LIMITATIONS AND FUTURE RESEARCH

This study relies on published research, official guidance and reported cases. It does not include original call recordings or a controlled user experiment. Media reports of incidents can contain incomplete or evolving investigative details and are therefore treated only as case illustrations.

Future work should build multilingual Indian datasets that cover Hindi-English code-switching and regional languages, telephony compression, replay, re-recording, noise and real-time voice conversion. Future evaluations should report false-accept and false-reject rates, equal error rate or calibrated risk metrics, latency, robustness to unseen generators, and performance on in-the-wild calls, and not only clean benchmark accuracy.

14. CONCLUSIONS

Deepfake vishing turns a familiar human signal, the voice, into an attack surface. The technology is moving toward lower data requirements, faster synthesis and finer control, while detection remains sensitive to distribution shift and real-world channels [8], [10], [11].

VoiceGuard therefore recommends layered defence: risk-based transaction controls, independent call-back, multi-factor authorisation, spoof-resistant voice biometrics where appropriate, verified contact directories, awareness training and rapid incident reporting. The aim is not to decide whether a voice sounds real, but to make sure that a financial or sensitive action can never be authorised by voice alone.

REFERENCES

- [1] CERT-In, "Deepfakes – Threats and Countermeasures," Advisory CIAD-2024-0060, Indian Computer Emergency Response Team, Nov. 27, 2024.
- [2] CERT-In, "Blueprint for Reducing Exposure and Responding to AI-Assisted Cyber Exploitation," Version 1.0, Indian Computer Emergency Response Team, May 25, 2026.
- [3] Reserve Bank of India, "Increasing Instances of Payment Frauds – Enhancing Public Awareness Campaigns Through Multiple Channels," Jun. 22, 2020.
- [4] Reserve Bank of India, Report on Currency and Finance 2023–24: India's Digital Revolution, Mumbai, India, 2024 (reporting the RBI survey on awareness and use of digital payments).
- [5] Indian Cyber Crime Coordination Centre (I4C), National Cyber Crime Reporting Portal and cyber financial fraud helpline 1930, Ministry of Home Affairs, Government of India.
- [6] ASVspoof Consortium, ASVspoof Workshop 2024: Speech Deepfake Detection and Spoofing-Robust Automatic Speaker Verification, 2024.
- [7] X. Wang et al., "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," 2024.
- [8] Li et al., "How well do current speech deepfake detection methods generalize to the real world?," 2026.
- [9] Pham et al., "A comprehensive survey with critical analysis for deepfake speech detection," 2024.
- [10] S. Ji et al., "ControlSpeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control," 2024.
- [11] S. Ji et al., "MobileSpeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech," 2024.
- [12] National Institute of Standards and Technology (NIST), 2024 Speaker Recognition Evaluation (SRE24), 2024.
- [13] "Two clone voice to defraud woman of 3.8L, arrested," The Times of India, Apr. 30, 2024.
- [14] "AI voice clone used to cheat Mumbai businessman of 80,000," The Times of India, Apr. 2024.
- [15] "AI voice cloning scam: Railway DG's friend loses 2 lakh," The Times of India, Dec. 2024.
- [16] "Delhi gang uses AI to clone Mumbai MD's voice, siphons 10 crore from firm," India Today, Jun. 17, 2026.