BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

A Review of Anti-Spoofing and Speaker Verification Techniques for Privacy-First, On-Device Detection of AI-Cloned Voice Scams and Family Voice Verification

[1]Shravani Suryawanshi, [2]Prof. Varsha Babar

[1]Student, [2]Assistant Professor
[1,2]Department of Artificial Intelligence and Data Science,
[1,2]Ajeenkya DY Patil School of Engineering, Pune, India
[1]shravani.suryawanshi@dypic.in, [2]varsha.babar@dypic.in

Abstract

Cloned voices are now cheap to make and easy to misuse in phone scams that imitate a relative or a colleague. This review paper examines what the research literature offers for defending ordinary users against such calls and uses it to propose Voice-Clone Guardian, a mobile application that runs entirely on the phone. The application combines two checks that are usually studied separately: an anti-spoofing model that looks for signs of synthetic speech, and a speaker-verification model that compares the voice with a profile enrolled for a family member. The paper reviews benchmark, model and survey studies, justifies the choice of a lightweight graph-attention countermeasure (AASIST-L class) and an ECAPA-TDNN speaker encoder, and describes the audio front-end, enrolment procedure, scoring pipeline and decision rules, with a diagram for each. The literature also shows clear limits: detectors trained on benchmark data degrade sharply on real-world audio, short clips give weak evidence, and mobile operating systems block direct access to call audio. The work is at the review and design stage; no prototype has been built and no performance figures are claimed for the proposed system. An evaluation plan is given for the next phase.

Keywords: anti-spoofing, voice cloning, speaker verification, on-device AI, audio deepfake detection, mobile security, privacy.

I. Introduction

Voice cloning no longer needs a studio or hours of recordings. Zero-shot text-to-speech systems such as VALL-E condition on a three-second recording of an unseen speaker [1]. A short voice note, a social-media video or a single phone call can be enough source material. The same capability that helps with dubbing and assistive speech can be used in a call that says, in the voice of a son, a daughter or a manager, that money is needed urgently.

Researchers have worked on a closely related problem for years under the name of spoofing countermeasures for automatic speaker verification (ASV) [2]. The ASVspoof challenge series has supplied shared datasets and metrics since 2015 [3], [4]. Progress on these benchmarks is real: a lightweight graph-

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

attention model with only 85 thousand parameters reaches an equal error rate (EER) of 0.99% on ASVspoof 2019 [5]. Yet the same literature warns that such numbers do not carry over to ordinary audio. Detectors that look excellent on ASVspoof 2019 lose a great deal when tested on deepfakes collected from the internet [6], and a speaker-verification system that scores 1.63% EER on genuine trials degrades to 23.83% once spoofed trials are included [7].

What is missing is not another benchmark model but a clear picture of how these pieces could fit into something an ordinary person can use during a suspicious call. Such a tool has to run on a phone, cope with short and noisy audio, respect the privacy of the people whose voices it stores, and tell a worried user what to do next rather than only output a probability. This paper reviews the relevant work with that target in mind and proposes Voice-Clone Guardian, a privacy-first mobile application that combines synthetic-speech detection with verification against a voice profile enrolled for a family member.

The contributions are: (i) a review of anti-spoofing, speaker-verification and spoofing-aware work, organised around what it means for a phone-based consumer tool; (ii) a justified selection of models, datasets, audio processing and mobile technologies; (iii) a system design, with diagrams of the architecture, enrolment, audio analysis and decision logic; and (iv) a candid account of limitations together with an evaluation plan. Nothing has been implemented and no results for the proposed system are reported. Section II reviews related work, Section III states the problem, Section IV justifies the technology choices, Section V describes the architecture, Section VI walks through an example, Section VII compares the design with existing approaches, Section VIII discusses privacy, security and limitations, and Section IX gives the evaluation plan before Section X concludes.

II. Related Work

A. Voice Cloning and Spoofing Attacks

Wu et al. divide attacks on speaker verification into impersonation, replay, speech synthesis and voice conversion, and review the early countermeasures for each [2]. Modern zero-shot cloning [1] belongs to the synthesis and conversion group, which the ASVspoof challenges call logical access (LA), while replaying a recording through a loudspeaker is physical access (PA) . For this project the distinction matters because a cloned voice reaches the user either through the telephone network, which adds codecs, or through the speakerphone and microphone, which adds a replay-like acoustic path.

B. Benchmarks and Datasets

The ASVspoof series began with spoofing detection for speaker verification and has grown with each edition. The 2019 edition added logical-access attacks created with a range of text-to-speech and voice-conversion systems and evaluated countermeasures with the tandem detection cost function (t-DCF), which measures the effect of a countermeasure on the verification system it protects [8]. All partitions of the 2019 data derive from the VCTK corpus [6]. ASVspoof 2021 moved closer to practical conditions: its LA task added encoding and transmission effects, and a new deepfake (DF) task used compressed speech of the kind posted online. The organisers report that countermeasures were robust to the new encoding and transmission effects, whereas DF systems did not generalise well across source datasets [3]. ASVspoof 5 goes further, with crowdsourced data from many more speakers in varied acoustic conditions and, for the first time, adversarial attacks; its organisers report that the attacks significantly compromise the baseline systems [4].

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

C. Countermeasure Models

Early countermeasures relied on hand-crafted features and Gaussian mixture models, and later systems moved to neural networks and ensembles [2], [8]. End-to-end models take the raw waveform. RawNet2 does this with a convolutional encoder [9], and AASIST adds heterogeneous spectro-temporal graph attention on top of a RawNet2-style encoder [5]. AASIST reports an EER of 0.83% on the ASVspoof 2019 LA evaluation set and, notably for this project, a lightweight variant, AASIST-L, with 85 thousand parameters and 0.99% EER [5]. A different line of work puts a self-supervised wav2vec 2.0 front-end before the classifier. Tak et al. paired a wav2vec 2.0 XLS-R front-end with an AASIST back-end and data augmentation, motivated by the limited generalisation of conventional countermeasures, and reported the lowest EERs known at the time on the ASVspoof 2021 LA and DF databases [10].

D. Speaker Verification

Speaker verification asks whether a test utterance comes from the claimed person, normally by comparing fixed-length embeddings. ECAPA-TDNN improved TDNN-based embeddings with channel attention and multi-layer feature aggregation [11] and is trained on large-scale speaker-recognition data. These systems are trained to separate real people, not to question whether a voice is real: when spoofed trials were added to an ASVspoof 2019 protocol, the EER of a state-of-the-art ASV system rose from 1.63% to 23.83% [7].

E. Combining Verification and Anti-Spoofing

The spoofing-aware speaker verification (SASV) challenge was created to encourage joint treatment of the two tasks [12]. Simple fusion already helps: in the baseline study, score-sum fusion brought the EER back to 1.71% and a neural fusion to 6.37% [7], and the best of 23 challenge submissions reduced the EER of a conventional verification system from 23.83% to 0.13% [12]. A related idea comes from Pianese et al., who detect deepfakes by comparing the audio under test with reference recordings of the same person in the embedding space of off-the-shelf speaker-verification models, either against the centroid of the references or against the most similar one [13]. Because nothing is trained on fake speech, the method does not depend on having seen the generator, and it showed good generalisation on three test sets [13]. This is close to what a family-voice profile offers, with the difference that the reference recordings here would be few and short.

F. Generalisation Outside the Lab

Müller et al. re-implemented twelve architectures under one protocol and found that constant-Q and linear-frequency spectrograms gave on average 37% lower EER than mel spectrograms, all else equal [6]. They also collected 37.9 hours of audio of celebrities and politicians, 17.2 hours of it deepfakes, and observed EER degradation of up to one thousand percent; a RawNet2 model retrained on every ASVspoof 2019 split still showed about 33% EER on this set [6]. They suggest that the community may have tailored its solutions too closely to the ASVspoof benchmark [6].

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE I

Summary of the Reviewed Literature and the Gap Each Study Leaves for This Project

Ref.	Study and method	Key finding as reported	Limitation here
[2]	Wu et al. (2015): Survey of attacks and countermeasures for speaker verification	Called for standard datasets and warned that countermeasures over-fit to known attacks	Predates zero-shot voice cloning
[8]	Wang et al. (2020): ASVspoof 2019 database (LA and PA) and t-DCF evaluation	Standard benchmark linking countermeasure and ASV performance	All partitions derive from VCTK [6]
[11]	Desplanques et al. (2020): ECAPA-TDNN speaker embeddings	Improved verification accuracy on speaker-recognition benchmarks over earlier TDNN systems	Not built to question whether speech is real
[9]	Tak et al. (2021): RawNet2, end-to-end raw-waveform countermeasure	Countermeasure without hand-crafted features; encoder family reused by AASIST	Weak on in-the-wild data when trained on ASVspoof 2019 only [6]
[5]	Jung et al. (2022): AASIST graph-attention countermeasure; AASIST-L variant	EER 0.83% (about 297k parameters) and 0.99% (85k parameters) on ASVspoof 2019 LA	Single benchmark; no on-phone latency reported
[10]	Tak et al. (2022): wav2vec 2.0 (XLS-R) front-end with AASIST back-end and augmentation	Lowest EERs reported at the time on ASVspoof 2021 LA and DF	About 300M-parameter front-end
[6]	Müller et al. (2022): Twelve architectures re-implemented; new 37.9 h in-the-wild set	cqtspec/logspec beat melspec by 37% EER on average; EER degraded by up to 1000% on real-world audio	Public-figure audio, not phone calls
[3]	Liu et al. (2023): ASVspoof 2021: LA, PA and DF tasks; 54 teams	LA countermeasures robust to new encoding and transmission effects; DF lacks cross-dataset generalisation	Challenge data; no on-device constraint

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Ref.	Study and method	Key finding as reported	Limitation here
[7], [12]	Shim et al. (2022); Jung et al. (2022): SASV: spoofed trials added to verification; score and embedding fusion	ASV EER 1.63% to 23.83% with spoofed trials; score-sum 1.71%; best submission 0.13%	ASVspoof-based data; no family enrolment scenario
[13]	Pianese et al. (2022): Deepfake detection by comparing speaker-verification embeddings with reference audio of the same person	Good generalisation on three test sets; trained on real speech only	Needs reference recordings of the speaker
[4]	Wang et al. (2024): ASVspoof 5: crowdsourced data, diverse conditions, adversarial attacks	Attacks significantly compromise baseline systems	Challenge setting; no deployment constraints

G. Research Gaps

Reading the studies together, five gaps stand out, and each shapes the design that follows. First, benchmark results do not transfer reliably to real audio [6], [3], so any claim for a consumer tool has to be tested on out-of-domain and channel-degraded recordings. Second, compact models are the exception rather than the rule; AASIST-L is one of the few reviewed works that reports a parameter count and file size, and none of the reviewed papers reports latency, memory or energy on a phone. Third, verification and anti-spoofing are combined mostly in challenge settings built from ASVspoof data [12], [7], not in a setting where a person enrols a few short clips of relatives. Fourth, short and noisy clips are under-studied: the countermeasures above are typically trained on windows of about four seconds [5], [6], and their behaviour on phone-quality fragments of that length is rarely reported. Fifth, none of the reviewed studies addresses how a result should be presented to a non-expert who is deciding, under pressure, whether to send money.

III. Problem Statement and Design Objectives

A person receives a call or voice message that sounds like someone they know and asks for money, a code or another urgent action. The person cannot easily tell whether the voice is genuine, existing detectors were built for benchmark evaluation rather than for use on a phone, and time is short. The application should help them judge the call and then steer them to a verification step that does not depend on the voice itself. The following objectives guide the design.

1) *On-device processing:* all analysis, models and voice profiles stay on the phone.
2) *Two complementary checks:* a synthetic-speech score and a similarity score against an enrolled family member, kept separate.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

3) *Short, noisy audio:* tolerate calls of a few seconds and decline to guess when the audio is insufficient.

4) *Honest uncertainty:* an explicit "inconclusive" outcome, and no message that promises safety.

5) *Consent and minimal data:* store only embeddings of people who agreed, and delete raw audio.

6) *Work within OS limits:* rely only on audio paths that mobile systems actually permit.

IV. Technology Selection and Justification

A. Anti-Spoofing Model

Table II lists the candidates considered. We propose AASIST-L as the starting point for the on-device countermeasure for three reasons. It works directly on the raw waveform, which removes a feature-engineering choice, including the mel-versus-log-spectrogram question raised in [6]. Its published size, 85,306 parameters and a model file of about 332 KB [5], makes deployment on a phone plausible, whereas the XLS-R front-end used in [10] has about 300 million parameters, more than three orders of magnitude more. And the accuracy given up is small on the benchmark where both were reported, 0.99% against 0.83% EER [5]. That gap should not be over-read: the AASIST authors note that EER can vary noticeably with the random seed, citing a baseline whose EER ranged from 1.19% to 2.06% across seeds [5].

Two cautions apply. The reported figures come from ASVspoof 2019 only, which [6] shows to be a poor predictor of performance on real data, so the model would have to be fine-tuned on additional data with augmentation such as codec simulation and noise, in the spirit of the augmentation used in [10]. And graph-attention operations may not convert cleanly to every mobile runtime, which has to be checked early. If AASIST-L proves too weak or too awkward to deploy, a small convolutional model on log or constant-Q spectrograms is the fallback, in line with the feature findings of [6]; larger self-supervised models remain a research reference rather than a shipping option.

TABLE II

Candidate Models Considered and Their Fit for On-Device Use

Component	Candidate	Size or accuracy as reported	Suitability for the phone
Anti-spoofing	AASIST-L [5]	85,306 parameters, about 332 KB; EER 0.99% on ASVspoof 2019 LA evaluation set	Primary candidate; check mobile-runtime operator support
Anti-spoofing	AASIST [5]	About 297k parameters; EER 0.83% on the same set	Fallback if latency allows
Anti-spoofing	RawNet2 [9]	End-to-end raw-waveform network; best in-the-wild model in [6] was a RawNet2 with 4-s input	Baseline for comparison and ablations

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Component	Candidate	Size or accuracy as reported	Suitability for the phone
Anti-spoofing	wav2vec 2.0 XLS-R + AASIST [10]	XLS-R front-end of about 300M parameters; lowest EERs reported at the time on ASVspoof 2021 LA and DF	Too large to ship; research reference
Speaker verification	ECAPA-TDNN [11]	Trained on speaker-recognition data; published in smaller and larger channel configurations	Primary candidate; smaller width, quantised
Scoring	Centroid and best-reference comparison [13]	Both strategies used with off-the-shelf speaker-verification models	Adopted for the family profile
Fusion	Score-sum and embedding fusion [7]	Score-sum fusion gave EER 1.71% in the SASV baseline study	Learned fusion kept for later; rules first

B. Speaker Encoder

For speaker verification we propose ECAPA-TDNN [11], again for maturity and evidence: it is trained on large-scale speaker-recognition data, published in smaller and larger channel configurations, and available as pretrained models in open toolkits. The SASV challenge also released pretrained spoofing-detection and speaker-verification models as open source [12], which gives a ready starting point and a reference to compare against. Scoring follows the person-of-interest idea of [13]: the test embedding is compared with the family member's profile by cosine similarity, using the centroid, the best individual reference, or both (Section V-D).

C. Audio Processing

Audio is converted to 16 kHz mono, a rate commonly used by speech models. Telephone speech is band-limited, so training data must include narrowband and codec-processed versions of the signals, as ASVspoof 2021 did for its LA task [3]. Speech detection removes silence and non-speech before scoring and before training, so that the model cannot lean on silence length as a shortcut. Analysis windows of four seconds with a hop of one to two seconds match the input length used by RawNet2 and AASIST [5], [6] and let the result update as the call continues. For the speaker encoder, 80-band log-Mel features are used, following the usual ECAPA-TDNN set-up.

D. Datasets

Table III lists the datasets and the role of each. The in-the-wild set should be used for testing only, since tuning on it would remove its value as an out-of-domain check. A family-style evaluation set still has to be collected: consenting volunteers record short clips on handsets and on speakerphone, and cloned versions are produced with open voice-cloning models. This needs ethics approval and careful handling, because the recordings are themselves biometric data.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE III

Datasets and Their Intended Role

Dataset	Reported content	Role here
ASVspoof 2019 LA [8]	TTS and VC attacks; derived from VCTK [6]	First training and sanity check
ASVspoof 2021 LA/DF [3]	Adds encoding and transmission effects (LA) and compression (DF)	Channel-robustness tests
ASVspoof 5 [4]	Crowdsourced, diverse acoustics, adversarial attacks	Harder evaluation
In-the-wild [6]	37.9 h of public-figure audio, 17.2 h fake	Out-of-domain testing only
Own recordings (to be collected)	Consenting volunteers; handset and speakerphone; clones from open TTS	Family-style evaluation

E. Mobile Deployment

Models would be exported to a mobile runtime such as TensorFlow Lite, ONNX Runtime Mobile or PyTorch's on-device runtime and quantised to 8-bit integers where accuracy allows. Integer quantisation usually improves latency on mobile CPUs, but its effect on a countermeasure has to be measured, because quantisation noise could blur exactly the faint artefacts that detectors look for. This is a hypothesis to test, not a result. The model files are bundled with the application, so no network access is needed at run time.

V. Proposed System Architecture

A. Overview

Fig. 1 shows the structure. Audio enters by one of the two paths available on a phone (Section V-B), passes through a front-end that cleans and windows it, and is scored by two branches in parallel. Their outputs are calibrated and smoothed before a rule-based decision engine maps them to a message.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

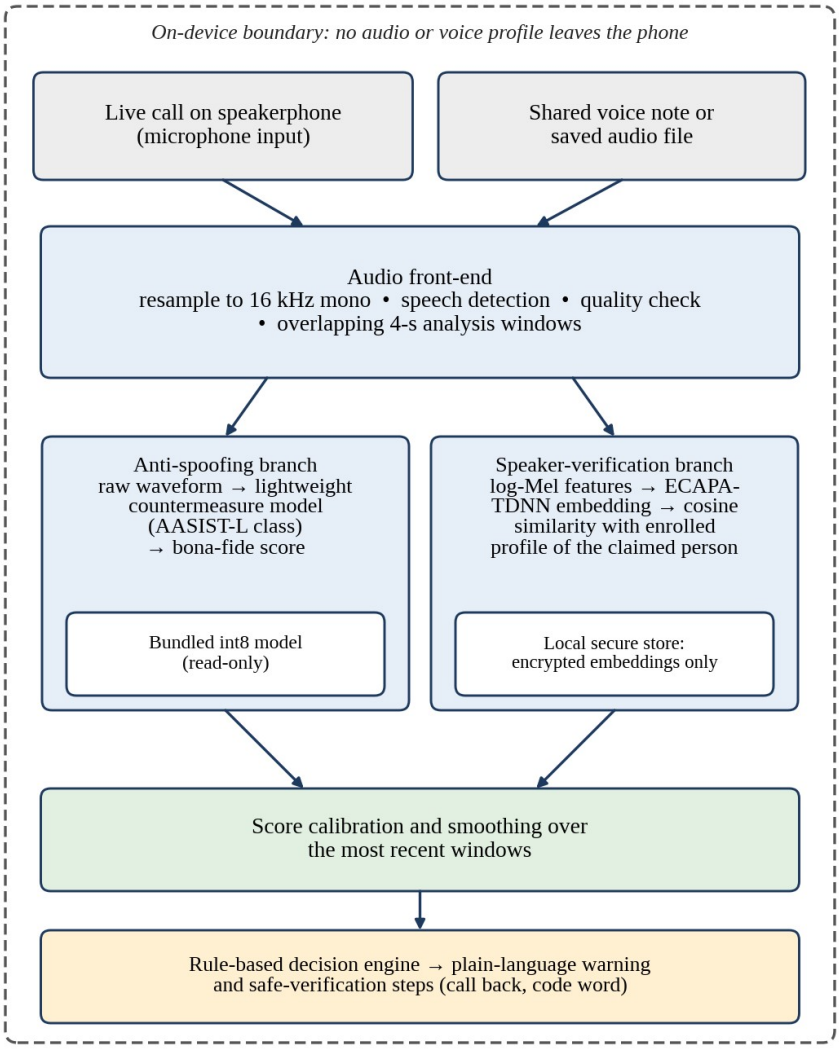


Fig. 1. Proposed architecture. Everything inside the dashed boundary runs on the phone; only the user's screen shows results.

B. Getting Audio Under Mobile OS Restrictions

The most natural design, listening to the call itself, is blocked on common platforms. Android's documentation states that capturing from the VOICE_CALL source requires the CAPTURE_AUDIO_OUTPUT permission, which is not available to third-party applications [14]. We are not aware of a public iOS interface that gives third-party apps access to cellular call audio; this should be checked against current Apple documentation. Fig. 2 summarises the three paths considered. Path A uses the speakerphone and microphone. It needs only the normal microphone permission, but the audio has passed through the loudspeaker and the room, so it behaves like a replay of the call and also contains the user's own voice; the user is asked to stay quiet while the check runs. Path B accepts voice notes and audio files shared into the app, which is cleaner and fully offline, although messaging apps compress the audio and the check is not live. Path C, direct capture of call audio, is not planned for the first version. Recent Android versions also restrict background microphone use, so analysis should run only while the app is

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

visible or in a foreground service with a notification; the exact rules depend on the Android version targeted and must be confirmed. Recording calls can also raise consent questions in some jurisdictions, so the app keeps audio in memory only and says so.

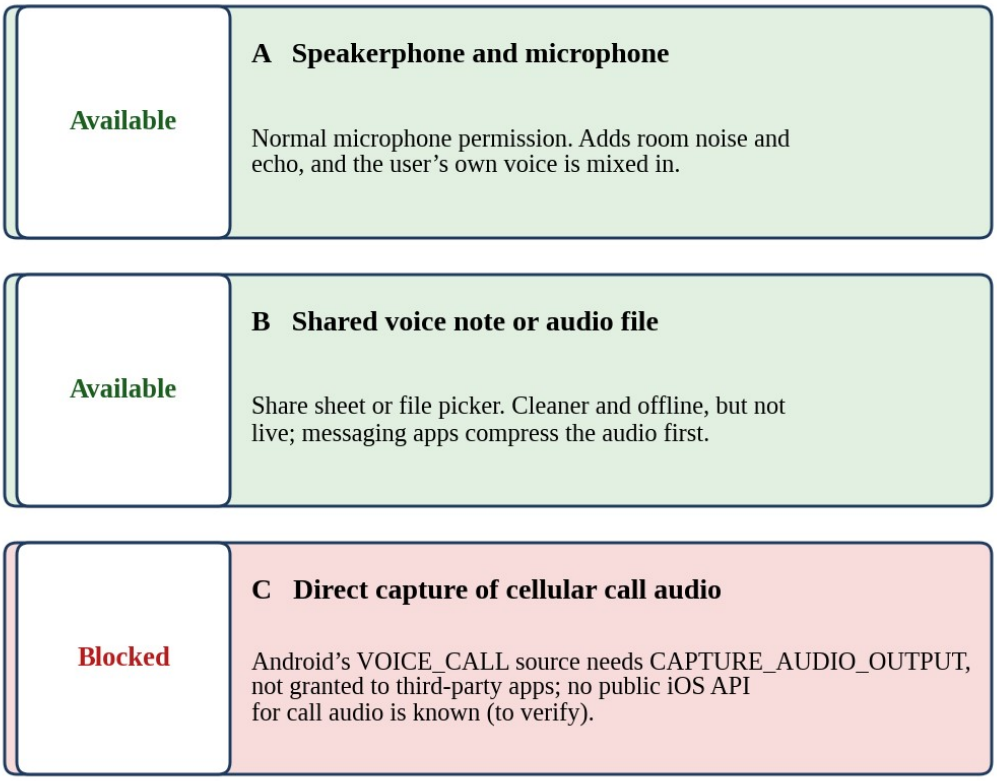


Fig. 2. Audio acquisition paths and what mobile operating systems allow.

C. Voice Enrolment

Enrolment builds the profile against which a claimed voice is compared (Fig. 3). The owner adds a family member, and that person gives consent on screen before anything is recorded. They then read several short prompts, ideally in more than one condition, since a profile built only from quiet-room recordings may not match a voice heard over a speakerphone. Each clip passes a quality check for speech length, noise and clipping, and is then screened by the anti-spoofing model, so that a synthetic clip cannot be enrolled as a relative. This step matters because enrolment is itself an attack surface: if an attacker could send a cloned voice note and have it stored as "Mum", the profile would work against the user. For the same reason enrolment should happen with the person present or over a channel the owner trusts. For every accepted clip an ECAPA-TDNN embedding is computed. The profile keeps the centroid and the individual reference embeddings, following the two comparison strategies of [13], encrypted with a key held in Android Keystore or the iOS Keychain. Raw audio is deleted immediately, and the owner can remove a profile at any time. How many clips and how many seconds of net speech are enough is an empirical question for the evaluation; no number is fixed here.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

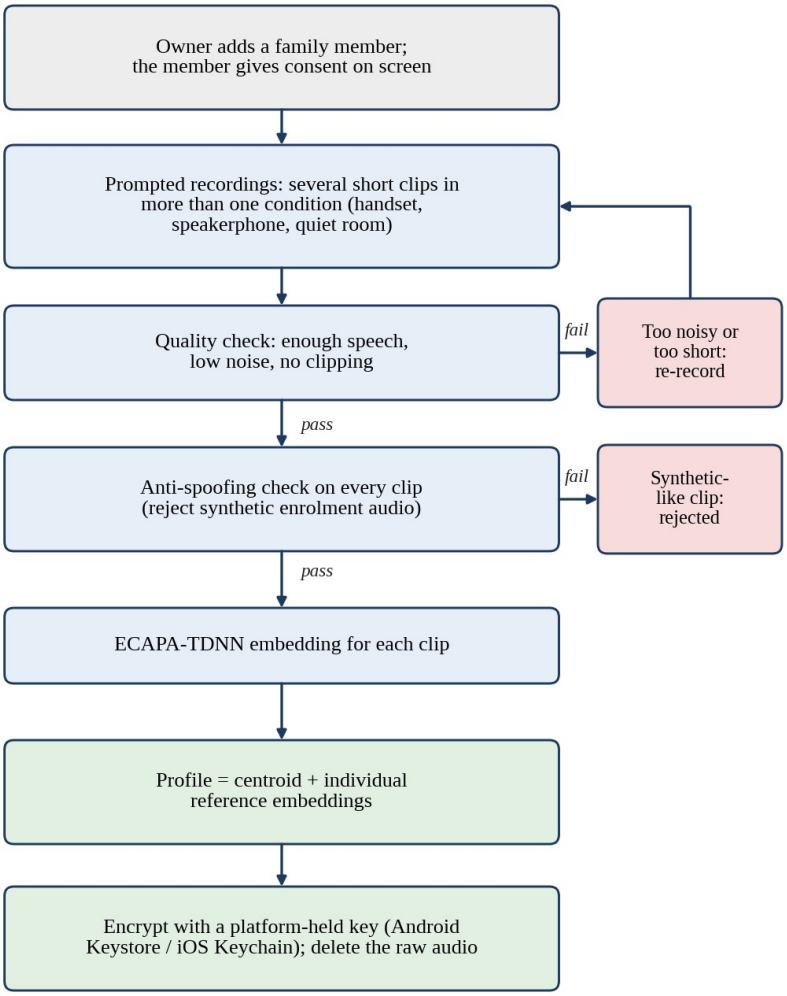


Fig. 3. Voice enrolment flow for one family member.

D. Audio Analysis

During a check (Fig. 4), the front-end converts the signal to 16 kHz mono, discards non-speech and cuts overlapping four-second windows once enough net speech has accumulated. Each window goes down both branches. The countermeasure produces a bona-fide score s_{cm}. The speaker branch produces an embedding e and compares it with the claimed person's profile. With centroid μ and reference embeddings r_k, the two scores are

$$s_c = (e \cdot \mu) / (||e||\,||\mu||), \quad s_m = \max_k (e \cdot r_k) / (||e||\,||r_k||) \tag{1}$$

and the speaker score s_{sv} is taken as one of them or a combination chosen during evaluation. Individual windows are noisy, so each score stream is smoothed with a running median over the last N windows,

$$\tilde{s}_t = \text{median}\{\, s_{t-N+1}, \ldots, s_t \,\} \tag{2}$$

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

and a warning is shown only when the same verdict persists over several consecutive windows. Raw scores are calibrated on a development set before thresholds are applied, because raw outputs of different models are not directly comparable.

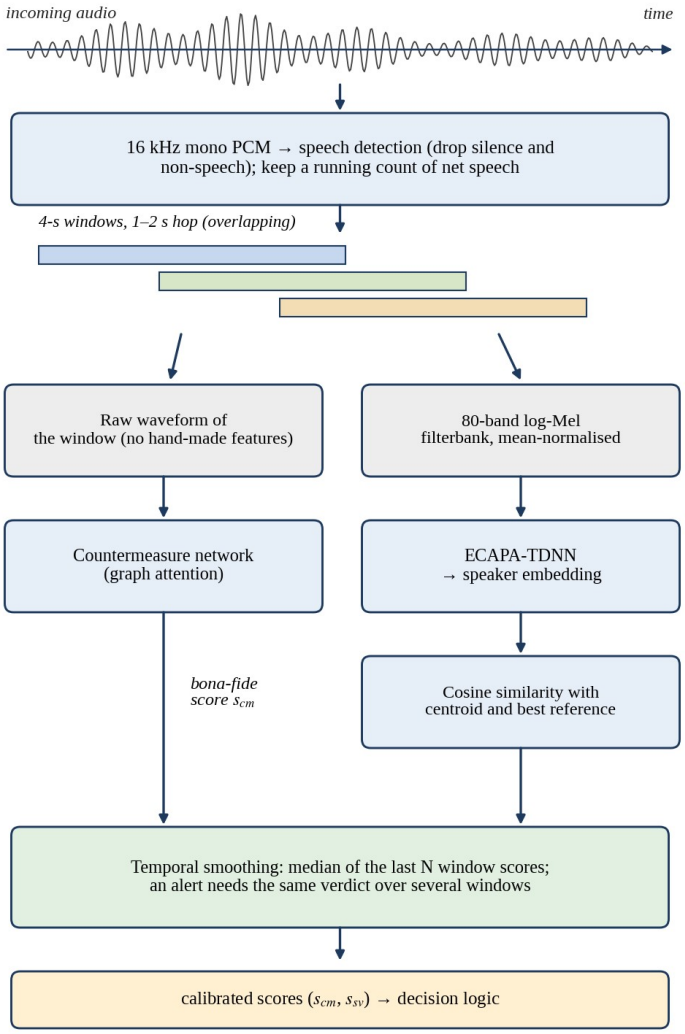


Fig. 4. Audio analysis pipeline: windowing, the two scoring branches, and smoothing.

E. Decision Logic

The engine applies two thresholds, τ_{low} and τ_{high}, to the smoothed countermeasure score, which leaves an uncertainty band between them (Fig. 5). Below τ_{low} the voice looks synthetic. If the claimed person also matches, the result is the most serious one, since a good clone is expected to resemble the enrolled voice. Above τ_{high} no synthetic traces were found: a match then gives the lowest level of concern, while a mismatch gives a caution, because the caller may be someone else or the audio may simply be poor. Between the thresholds the app reports that the result is inconclusive. A speaker match alone never produces a reassuring result, since speaker-verification scores were shown to be vulnerable to spoofed trials [7]. If the user does not name a person, the speaker branch is skipped and only the countermeasure outcomes are available.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

The thresholds are not set in this paper. They must be fixed on development data at a conservative operating point, because the two kinds of error differ in cost. Draft messages for each outcome are in Table IV. They describe what was found rather than certify the call, and each points to a verification step that does not rely on the voice: calling back on a saved number or asking for a code word agreed in advance by the family.

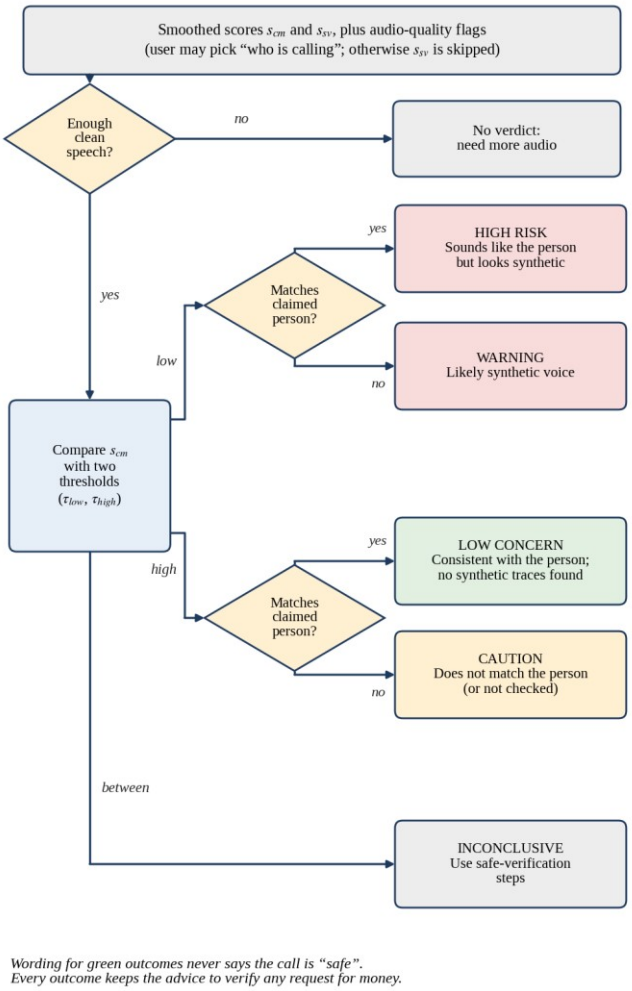


Fig. 5. Decision logic. Colour is a reinforcement only; every outcome also carries a text label.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE IV

Draft Messages and Suggested Next Steps

Outcome	Draft message	Suggested next step
High risk	"This voice sounds like [Name], but it shows signs of being computer-generated. Do not send money or share codes."	Hang up; call back on the saved number; ask for the family code word
Warning	"This voice shows signs of being computer-generated."	Same as above
Low concern	"This sounds like [Name] and no signs of synthetic speech were found. This check is not proof."	Still confirm any request for money or codes by call-back
Caution	"This voice does not match [Name] well, or no person was selected."	Poor audio or another person: confirm identity another way
Inconclusive / no verdict	"Not enough clear speech to decide" or "The result is unclear."	Retry in a quieter place; use call-back or code word

VI. Illustrative Scenario

The following example describes intended behaviour; it is not an experiment, and the system has not been run. A parent receives a call from an unknown number. The caller, who sounds like the parent's daughter, says her phone is broken and asks for an urgent bank transfer. The parent switches to speakerphone, opens the app, selects the daughter's profile as the claimed person and starts a check. After a few seconds of net speech the front-end has its first window. If the countermeasure score is low and the speaker score is high, the app shows the high-risk message of Table IV and recommends hanging up and calling the daughter on her saved number. If instead the audio is short or the scores fall between the thresholds, the app says the result is inconclusive and suggests the same call-back step, because the safe action does not depend on what the models concluded. That is also why the inconclusive outcome exists: on poor lines the honest answer will often be that the evidence is weak.

VII. Comparison with Existing Approaches

Table V compares the proposed design with the families of approaches reviewed in Section II. The comparison is qualitative and limited to what the reviewed papers report; for the proposed system the entries describe design intent, not measured behaviour.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE V

Qualitative Comparison of Approaches (Entries for the Proposed System Are Design Intent)

Approach	Spoof check / identity check / enrolled voice	Phone-side use	Evidence in reviewed work
Standalone countermeasures, e.g. AASIST, RawNet2 [5], [9]	Yes / No / No	Small variant (AASIST-L); not tested on phones	Strong on benchmarks; large drop on in-the-wild audio [6]
Standalone speaker verification, e.g. ECAPA-TDNN [11]	No / Yes / Yes	Not addressed	EER rises from 1.63% to 23.83% with spoofed trials [7]
Speaker-verification-based deepfake detection [13]	Indirect / Yes / Yes (reference audio)	Not addressed	Three test sets; trained on real speech only
SASV systems [12], [7]	Yes / Yes / Yes	Not the focus	ASVspoof-based challenge data
Proposed Voice-Clone Guardian	Yes / Yes / Yes (consenting family)	Design goal	None yet; evaluation planned (Section IX)

The main difference from the literature is not any single component, all of which exist, but their combination around the practical limits of a phone. The family-profile comparison borrows from person-of-interest detection [13], the joint use of two scores from the SASV work [12], [7], and the small model from lightweight countermeasures [5]. Compared with a standalone countermeasure, the design adds an identity check, which gives the user something to act on ("this sounds like your daughter but looks synthetic"). Compared with standalone speaker verification, it avoids the failure that spoofed trials cause [7]. Compared with SASV systems built on challenge data, it assumes few, short enrolment clips from relatives, handset and speakerphone conditions, and an interface for non-experts. We did not find in the reviewed literature a study that tests this combination on a phone, so any advantage over the alternatives is a hypothesis.

VIII. Privacy, Security and Limitations

A. Privacy by Design

Voice is biometric data, and an embedding is a template for recognising a person, so it deserves the same care as the audio. The design therefore processes audio in memory, never writes call audio to storage, and

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

keeps only encrypted embeddings of people who agreed to be enrolled. On Android the application can be built without the network permission, so that the claim of no upload can be checked by anyone who inspects the app's permissions; model updates would then arrive through normal app-store releases. A local event log, if kept, would record outcomes and times, not audio. In India, processing of personal data on a consent basis falls under the Digital Personal Data Protection Act, 2023; legal review is needed before any release, and no legal analysis is attempted here.

B. Short Samples and Channel Effects

Both models see less evidence from a short clip than from a long one, and phone audio is compressed and band-limited. ASVspoof 2021 found countermeasures robust to its added encoding and transmission effects but not to unseen source data [3], and [6] shows how far real-world audio can be from benchmark audio. The enrolment profile itself is built from only a few clips, so it may not capture how a relative sounds when ill, excited or frightened; this last point is an expectation, not a finding from the reviewed studies. The app therefore refuses to give a verdict until enough clean speech has been heard, and reports "inconclusive" rather than forcing a decision.

C. False Detections

A genuine relative on a bad line may be flagged as synthetic, and a clone from an unseen generator may pass. The first error erodes trust and the second gives false reassurance; DF results in [3] and the in-the-wild results in [6] suggest the second is a real risk. Thresholds should therefore be conservative, the uncertainty band should be kept, and the wording of green outcomes must never imply safety. Every outcome ends with advice to verify by another channel.

D. Adversarial Pressure and the Cost of On-Device Models

An on-device model can be extracted from the application package, which lets an attacker test clones against the same detector offline. This is the price of not sending audio to a server, and it is one reason the app should not claim to be unbeatable. ASVspoof 5 introduced adversarial attacks and found that attacks significantly compromised the baselines [4], and poisoning of voice-authentication systems is a further concern. Replay of a genuine recording of the relative is a further blind spot: a detector trained against synthetic speech has nothing to flag and the speaker check will match, which is another reason the advice always ends with out-of-band verification.

E. Coverage and Fairness

The benchmarks reviewed are largely English-centred; ASVspoof 2019, for example, derives from VCTK [6]. Nothing in the reviewed material reports performance on Indian languages, code-mixed speech, children or elderly speakers. Both detectors and speaker models need testing across these groups before any claim of reliability is made.

IX. Evaluation Plan and Future Work

The next phase is to build a prototype and evaluate it in stages. (i) *Countermeasure:* fine-tune and test on ASVspoof 2019, 2021 and 5 using EER and per-attack results, then on the in-the-wild set as a purely out-of-domain test [6], including versions passed through telephone codecs, added noise and speakerphone replay. (ii) *Speaker verification:* EER on standard speaker-recognition test sets, then on family-style recordings, varying the number and length of enrolment clips. (iii) *Joint system:* the SASV-EER metric [12]

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

and a per-outcome confusion analysis on spoofed and genuine trials from enrolled speakers. (iv) *Resources:* time to first verdict, memory, battery use and heating on mid-range Android phones, with and without 8-bit quantisation. (v) *Usability:* a small user study on whether the wording in Table IV changes what people would do, and whether the inconclusive state is understood. (vi) *Ethics:* consent and approval for all human recordings. Beyond this, compressing a self-supervised model to phone size is a possible route to better generalisation, and an iOS version depends on what Apple's interfaces allow; both are open questions rather than plans with known outcomes.

X. Conclusion

This paper reviewed anti-spoofing, speaker-verification and spoofing-aware verification research from the viewpoint of a consumer protecting a family from cloned-voice calls, and proposed Voice-Clone Guardian, an on-device application that combines the two kinds of check. The literature supports the main ingredients: lightweight raw-waveform countermeasures, ECAPA-TDNN speaker embeddings, person-of-interest comparison and score fusion. It also shows that benchmark results transfer poorly to real audio, that short and compressed clips are hard, and that mobile systems limit access to call audio. The design responds with conservative decision rules and an explicit inconclusive state, protected enrolment, data kept on the device, and advice that always points to verification not based on the voice. Because nothing has yet been built or measured, the claims here are limited to the design and to what the cited studies report; the evaluation plan sets out how they will be tested.

Acknowledgment

The author thanks Prof. Varsha Babar for guidance during the literature survey and design stages of this work, and the Department of Artificial Intelligence and Data Science, Ajeenkya DY Patil School of Engineering, Pune, for institutional support.

References

[1] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," arXiv:2301.02111, 2023.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[3] X. Liu et al., "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2507–2522, 2023.

[4] X. Wang et al., "ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *Proc. ASVspoof Workshop*, 2024.

[5] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 6367–6371.

[6] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?," in *Proc. Interspeech*, 2022, pp. 2783–2787.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

[7] H.-j. Shim et al., "Baseline systems for the first spoofing-aware speaker verification challenge: Score and embedding fusion," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, 2022.

[8] X. Wang et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114.

[9] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 6369–6373.

[10] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, 2022.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[12] J.-w. Jung et al., "SASV 2022: The first spoofing-aware speaker verification challenge," in *Proc. Interspeech*, 2022.

[13] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, "Deepfake audio detection by speaker verification," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2022.

[14] Android Developers, "MediaRecorder.AudioSource," *Android API Reference*. [Online]. Available: https://developer.android.com/reference/android/media/MediaRecorder.AudioSource (accessed Oct. 2026).