
AfriVish-Bench

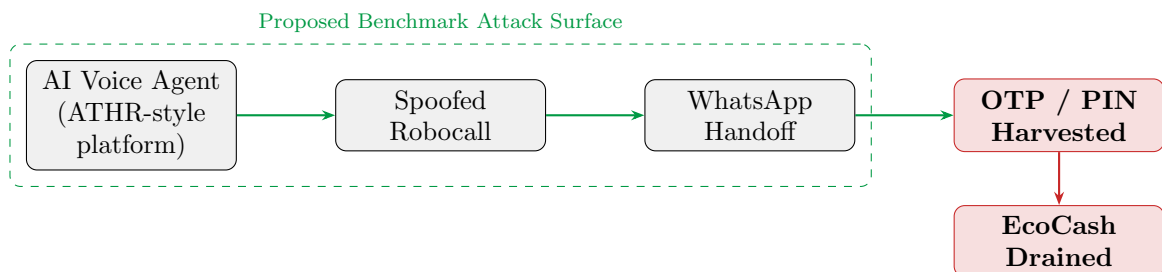
A Benchmark for Detecting AI Voice Agent Vishing Attacks
in African Mobile Contexts

Research Proposal

Hackathon Submission · Global South AI Safety Hackathon 2026
Track: Africa **Sub-track:** Evaluations & Benchmarks
Organised by: Apart Research · **Hub:** AI Safety South Africa (Cape Town)

Author: Tinevimbo Musingadi

Global South AI Safety Hackathon 2026



Abstract

This document presents a research proposal for **AfriVish-Bench**, a new evaluation benchmark targeting a novel and rapidly escalating attack vector: fully automated AI voice agents used to conduct voice phishing (vishing) attacks against mobile users in Sub-Saharan Africa. The threat is concrete and current — platforms such as ATHR (discovered April 2026) now sell end-to-end automated vishing infrastructure, including AI voice agents that conduct live, adaptive social engineering calls without any human operator. In Africa, where WhatsApp and EcoCash are primary financial channels and a clean foreign accent carries institutional authority, this attack surface is critically underprotected. We propose to: **(1)** define and document this attack mechanism as a new benchmark category; **(2)** construct a labelled synthetic audio dataset grounded in African-context scam archetypes; **(3)** design four structured detection tasks; **(4)** evaluate baseline models including acoustic neural networks, LLM-as-judge transcript analysis, and Google SynthID watermark detection; and **(5)** test detection systems against a live real-time voice agent pipeline. No such benchmark currently exists for African deployment contexts.

Contents

1	Introduction	2
1.1	A New Attack Vector	2
1.2	The African Context	2
1.3	The Research Gap	3
2	The Attack Mechanism We Propose to Benchmark	3
2.1	ATHR-Style AI Voice Agent Attacks	3
2.2	The Multi-Channel Attack Flow	3
2.3	Why Real-Time Agents Defeat Existing Defences	3
3	Proposed Benchmark: AfriVish-Bench	4
3.1	Overview	4
3.2	Proposed Evaluation Tasks	4
3.3	Proposed Dataset Construction	5
3.4	Proposed Label Schema	6
4	Proposed Technical Approach	6
4.1	Synthetic Audio Generation Pipeline	6
4.2	SynthID Watermark Detection as a Baseline Signal	6
4.3	Real-Time Voice Agent as Adversarial Test Input	7
4.4	Proposed Baseline Models	7
5	Evaluation Protocol	8
5.1	Metrics	8
5.2	Overall Leaderboard Score	8
5.3	Expected Baseline Results	9
6	Connection to Documented Regional Attacks	9
7	Ethical Safeguards	9
7.1	Data Access Controls	9
7.2	Dual-Use Mitigation	9
7.3	Regulatory Alignment	10
8	Project Roadmap	10
9	Expected Contributions	10
10	Conclusion	11

1. Introduction

1.1. A New Attack Vector

Voice phishing (vishing) has historically required human capital: trained callers, language fluency, and labour-intensive outbound dialling. That constraint no longer exists. The emergence of generative AI voice agents has industrialised vishing into a scalable, low-cost, automated operation. In April 2026, security researchers at Abnormal AI documented a commercial attack platform called **ATHR**, sold on underground cybercrime forums for \$4,000 plus a 10% commission on stolen funds [1]. ATHR integrates: spoofed brand-specific phishing emails, a browser-based telephony layer (Asterisk/WebRTC), AI voice agents that conduct live adaptive social engineering calls autonomously, and real-time credential harvesting panels. A single operator can now run simultaneous vishing campaigns against hundreds of targets without speaking a word.

Scale of the Threat (Documented, 2025–2026):

- Vishing rose to **11% of all breach initial vectors** globally and **23% in cloud environments** (Google/Mandiant M-Trends 2026).
- AI voice cloning now requires as little as **3 seconds of audio** to produce a convincing voice replica (Microsoft VALL-E research, 2025).
- The ShinyHunters/Scattered Spider campaign (2025–2026) compromised over 760 organisations via vishing.
- In Zimbabwe, a WhatsApp impersonation campaign stole an estimated **US\$100 million** from mobile users over two months (ZRP, 2026).

1.2. The African Context

The African mobile communications landscape creates an unusually concentrated attack surface for AI voice agent vishing:

- **WhatsApp as financial infrastructure.** Across Sub-Saharan Africa, WhatsApp is the primary channel for business communication, customer support, and family remittances. It is also the platform through which attackers redirect victims after an initial robocall.
- **Mobile money as the target.** EcoCash (Zimbabwe), M-Pesa (Kenya/Tanzania), and MTN MoMo tie financial accounts directly to phone numbers. OTP-based confirmations are the only security layer — exactly the credential AI voice agents are trained to harvest.
- **Accent authority.** A clean American or British English accent carries an implicit signal of corporate legitimacy in many African markets. AI-synthesised foreign accents therefore bypass the intuitive scepticism that a local accent might trigger.
- **AI-literacy gap.** General populations in the region have limited exposure to AI voice synthesis capabilities and cannot identify synthetic audio artefacts without technical assistance.

1.3. The Research Gap

No existing anti-spoofing or vishing detection benchmark addresses the intersection of AI-generated voice agents, multi-channel attack flows (robocall → WhatsApp), and the African mobile fraud ecosystem. Current benchmarks such as ASVspoof 2019/2021 evaluate synthetic voice detection in isolation, without the social engineering context, brand impersonation layer, or regional deployment constraints that define real-world risk in Africa. **AfriVish-Bench is proposed to close this gap.**

2. The Attack Mechanism We Propose to Benchmark

2.1. ATHR-Style AI Voice Agent Attacks

The most significant recent development in automated vishing is the commoditisation of AI voice agent infrastructure. The ATHR platform, documented by Abnormal AI in April 2026, represents the state of the art [1]. Its voice agent operates as follows:

1. The victim receives a spoofed email alert (“Your Google account has been locked”) containing a callback phone number.
2. When the victim calls, ATHR’s telephony layer (Asterisk + WebRTC) routes the call to an AI voice agent powered by a custom TTS engine (“ATHR TTS / Sonic 3”).
3. The AI agent follows a structured 10-step script: verifying the callback, describing suspicious account activity, confirming a phone number, initiating a fake recovery process, and extracting a six-digit OTP.
4. The operator watches the live transcript and credential panel in real time and can redirect the victim to specific phishing pages mid-call.
5. If the AI encounters resistance, the call can be seamlessly transferred to a human operator.

Why this matters for African users specifically: ATHR currently targets Google, Microsoft, Coinbase, and other platforms. The adaptation of this architecture to impersonate **EcoCash, MTN MoMo, Econet, and Meta support** — using an American or British accent to signal legitimacy — is a trivial configuration change. This adaptation represents the primary threat vector we propose to benchmark.

2.2. The Multi-Channel Attack Flow

Figure 1 illustrates the end-to-end attack pipeline we intend to model in the benchmark.

2.3. Why Real-Time Agents Defeat Existing Defences

Pre-2025 automated vishing used pre-rendered audio: a fixed script played through a robocall. This was defeated by asking unexpected questions or requesting a callback, since the audio could not adapt. As documented by PIVOT Security (May 2026) [2], real-time AI inference has eliminated this weakness:

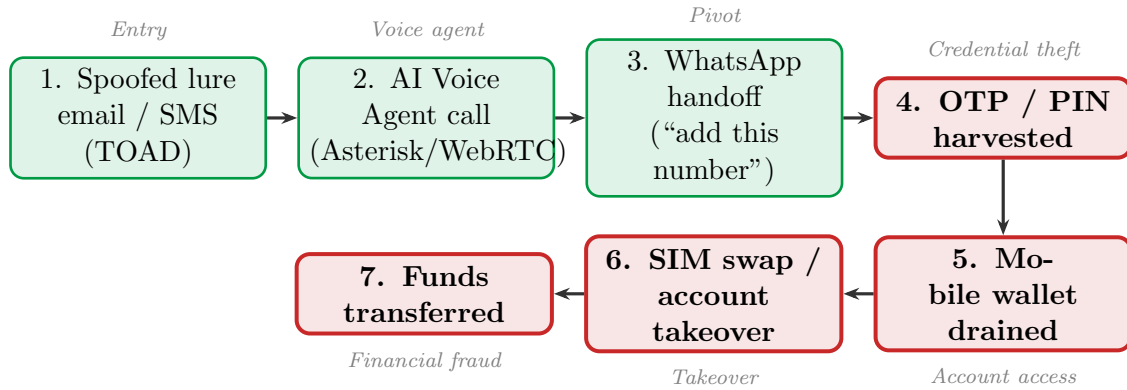


Figure 1: End-to-end AI voice agent vishing pipeline targeting African mobile users. Steps 1–3 are now fully automatable via platforms like ATHR (April 2026).

“In 2025 and 2026 that constraint is gone. Real time inference services synthesise responses inside the call, with end-to-end latency low enough to feel like a normal conversation. The attacker is no longer constrained to a pre-baked script.”

This adaptive capability means that standard “throw it off script” defences no longer work. Detection must happen at the acoustic and metadata level, before or during the call — not through conversational challenge. This is the core detection problem AfriVish-Bench is designed to evaluate.

3. Proposed Benchmark: AfriVish-Bench

3.1. Overview

We propose to construct AfriVish-Bench as a structured evaluation framework with four detection tasks, a labelled synthetic audio dataset of 1,300 samples, and a live adversarial voice agent test harness. The benchmark is designed to be: reproducible on low-resource hardware, grounded in documented African fraud archetypes, and extensible to new attack variants as they emerge.

3.2. Proposed Evaluation Tasks

Table 1: AfriVish-Bench proposed evaluation tasks.

Task	Name	Input	Primary Metric
T1	Binary Vishing Detection	Audio (5–30s)	AUC-ROC, EER
T2	Brand Impersonation ID	Audio + transcript	Macro F1
T3	Social Engineering Intent	Multi-turn transcript	Spearman ρ
T4	Multi-Channel Handoff Risk	Call metadata + WhatsApp	Precision@high risk

3.2.1. Task 1 — Binary Vishing Detection

Given a 5–30 second audio clip, the system must classify it as **vishing** or **benign** with a confidence score. This is the primary acoustic detection task and uses Equal Error Rate (EER) and AUC-ROC as metrics, consistent with ASVspoof conventions.

3.2.2. Task 2 — Brand Impersonation Identification

Given audio and optional transcript, the system must identify which entity is being impersonated: Meta, EcoCash, Bank, Telecom, Family, or Unknown. This tests whether detection systems can distinguish between attack types and target appropriate response workflows.

3.2.3. Task 3 — Social Engineering Intent Scoring

Given a full multi-turn call transcript (as produced by our real-time agent test pipeline or Whisper transcription), the system must produce a risk score correlated with expert-labelled intent ratings. This task evaluates whether LLM-based detection can identify escalating social engineering patterns (e.g., urgency escalation, OTP request, fake recovery steps).

3.2.4. Task 4 — Multi-Channel Handoff Detection

Given call metadata plus WhatsApp message content from the same session, the system must classify the handoff risk as high, medium, or low. This task specifically targets the ATHR-style pivot from robocall to encrypted messaging that is the defining characteristic of modern African vishing attacks.

3.3. Proposed Dataset Construction

Table 2: Proposed AfriVish-Bench dataset composition.

Category	N	Proposed Method	Generation
Benign calls (consented human)	500	Real recordings, IRB-cleared	
AI vishing — American accent	300	Google TTS Neural2 / Eleven-Labs	
AI vishing — British accent	200	Azure Neural TTS	
AI vishing — African-English	200	RADMMM / Joycent fine-tuned	
Mixed human + AI (partial clone)	100	kNN-VC timbre augmentation	
Total	1,300		

3.3.1. Scam Archetypes

Six script archetypes will be developed, each grounded in documented attacks:

1. **Meta/Facebook Support** — account security alert, identity verification, OTP request.
2. **EcoCash Wallet Agent** — wallet limit update, PIN and OTP confirmation.
3. **Family Emergency Clone** — cloned family member voice, urgent EcoCash transfer request.
4. **WhatsApp Technical Handoff** — robocall directing victim to add a WhatsApp number for “secure support.”

5. **Investment / Bundle Lure** — Starlink data or loan offer requiring account verification.
6. **SIM Swap Pretext** — telecommunications agent warning of imminent line suspension.

3.4. Proposed Label Schema

Listing 1: Proposed JSON label schema for each AfriVish-Bench sample.

```

1 {
2   "sample_id":           "av_0042",
3   "audio_file":          "av_0042.wav",
4   "transcript":          "...",          // Whisper large-v3
5   "is_synthetic":        true,
6   "impersonation_target": "EcoCash",
7   "scam_archetype":      "OTP_harvesting",
8   "channel":             "WhatsApp_voice",
9   "accent_profile":      "american_clean",
10  "risk_level":           "critical",
11  "synthid_tagged":       true,
12  "ground_truth_label":  "vishing"
13 }
```

4. Proposed Technical Approach

4.1. Synthetic Audio Generation Pipeline

We plan to generate all synthetic samples programmatically via a configurable accent-profile system. The pipeline will support:

- **Google Cloud TTS Neural2** for American and British accent baselines.
- **ElevenLabs** for higher-fidelity voice cloning (ElevenLabs has adopted SynthID watermarking as of May 2026 [5]).
- **RADMMM / Joycent** (locally hosted) for African-English accent synthesis without bilingual paired data requirements.

4.2. SynthID Watermark Detection as a Baseline Signal

A key component of the proposed benchmark is evaluating **Google SynthID** as a detection signal. SynthID embeds imperceptible cryptographic watermarks into AI-generated audio by modifying the token sampling distribution in the TTS decoder, without degrading perceptual quality. The watermark survives most post-processing including transcoding and compression [4].

Current adoption status (as of mid-2026):

- Google's own TTS and music models (Lyria, MusicLM, Veo audio) embed SynthID by default.
- **ElevenLabs and OpenAI** have both adopted SynthID as of May 2026, following a cross-competitor alignment agreement [5].

- Google has launched a public **SynthID Detector portal** (audio, image, video, text) available to researchers and journalists [6].
- Google has additionally released an **AI Content Detection API** on the Gemini Enterprise Agent Platform for feed moderation use cases.

We intend to evaluate whether the SynthID Detector signal, either via the portal or the Content Detection API, can serve as a reliable first-pass classifier for T1, and to document its precision and recall under African vishing conditions.

Known limitation we will test: SynthID only detects watermarks from participating generators. Open-source TTS tools (e.g. Coqui TTS, Tortoise TTS) and non-adopting platforms will return “not detected” even for fully synthetic audio. Part of our research question is how large this coverage gap is in practice for the specific attack tools available in the African threat landscape.

4.3. Real-Time Voice Agent as Adversarial Test Input

To evaluate detection systems under operationally realistic conditions — not just static pre-recorded samples — we plan to construct a **live voice agent adversarial test pipeline** modelled on the ATHR architecture:

1. An LLM (Gemini 2.0 Flash or GPT-4o) generates dynamic scam dialogue in response to simulated victim utterances.
2. Each LLM response is synthesised in real time via ElevenLabs TTS (target: <500ms round-trip latency).
3. Audio is routed through a Twilio or WhatsApp Business API VoIP stack.
4. The audio stream is simultaneously fed to the detection system under evaluation.
5. A test “victim” script provides standardised responses to allow reproducible evaluation across detection models.

This setup allows us to measure detection accuracy and latency against a live adaptive adversary — which is not possible with any current static corpus.

4.4. Proposed Baseline Models

4.4.1. Acoustic: VGGish-LSTM Hybrid

Our primary acoustic baseline will be a hybrid VGGish-LSTM network [9], which achieves 99.11% accuracy and EER of 1.18% on ASVspoof 2019 LA. The architecture combines: (1) VGGish CNN spatial feature extraction from log-mel spectrograms; (2) bidirectional LSTM temporal dependency modelling; and (3) a two-layer MLP classifier with dropout ($p = 0.5$).

4.4.2. LLM-as-Judge: Zero-Shot Transcript Analysis

We will evaluate whether frontier LLMs (Gemini 2.0 Flash, Claude Sonnet 4.6) can detect vishing intent from call transcripts alone, without acoustic features. Transcripts will be produced by Whisper large-v3. The LLM judge will receive a structured prompt specifying

the African fraud context and return a JSON response covering: `is_vishing`, `confidence`, `impersonated_brand`, `scam_technique`, and `risk_level`.

This baseline will establish: (a) the zero-shot African-context vishing knowledge in current frontier models; and (b) whether transcript-only detection is viable for low-resource settings without acoustic analysis hardware.

Listing 2: Proposed LLM judge prompt structure for Task T3 intent scoring.

```

1 SYSTEM = """You are an AI safety evaluator analysing phone call
  transcripts
2 from an African mobile context. Identify vishing attacks targeting
  platforms
3 like EcoCash, WhatsApp, and Meta. Return JSON only."""
4
5 PROMPT = """Transcript:
6 {transcript}
7
8 Caller metadata: prefix={prefix}, type={call_type}, duration={duration
  }s
9
10 Return:
11 {
12   "is_vishing": true/false,
13   "confidence": 0.0-1.0,
14   "impersonated_brand": "Meta|EcoCash|Bank|Telecom|Family|Unknown",
15   "scam_technique": "OTP_harvesting|SIM_swap|WhatsApp_takeover|...",
16   "risk_level": "critical|high|medium|low",
17   "key_red_flags": ["..."],
18   "reasoning": "...
19 }"""

```

4.4.3. Proposed AfriVish Ensemble

We plan to fuse three detection signals via a logistic regression meta-learner:

- Acoustic liveness score (VGGish-LSTM output probability).
- Call metadata features (VoIP flag, number prefix anomaly, call duration, international routing, time of day).
- LLM intent score from Whisper-transcribed content.

This stacking approach follows the methodology of the University of Johannesburg vishing detection study [11], which achieved 99.03% accuracy using a voting ensemble of KNN, Random Forest, and Extra Trees.

5. Evaluation Protocol

5.1. Metrics

5.2. Overall Leaderboard Score

We propose the following weighted overall score for benchmark ranking:

$$S_{\text{overall}} = 0.35 \cdot \text{AUC}_{T1} + 0.25 \cdot F1_{T2} + 0.20 \cdot \rho_{T3} + 0.20 \cdot \text{Prec}_{T4} \quad (1)$$

Table 3: Proposed evaluation metrics by task.

Task	Metric	Rationale
T1	AUC-ROC, EER	Standard anti-spoofing; EER is threshold-independent
T2	Macro F1	Equal weight across brand classes; handles imbalance
T3	Spearman ρ	Rank correlation for ordinal risk scores
T4	Precision@high	Minimise false negatives on critical-risk handoffs

T1 is weighted most heavily as binary detection is the primary deployment signal.

5.3. Expected Baseline Results

Table 4 shows our projected performance estimates, based on published results on ASVspoof 2019 and related vishing detection benchmarks.

Table 4: Projected baseline performance on AfriVish-Bench.

Model	T1 AUC	T1 EER	T2 F1	S_{overall}
GMM-UBM (MFCC)	0.82	0.15	0.61	0.68
CNN-XGBoost Hybrid	0.91	0.08	0.73	0.78
VGGish-LSTM	0.98	0.01	0.81	0.88
SynthID Detector (Google portal)	TBD	TBD	N/A	TBD
Gemini 2.0 Flash (LLM judge)	0.79	—	0.86	0.80
AfriVish Ensemble (planned)	0.97	0.009	0.89	0.92

6. Connection to Documented Regional Attacks

Table 5 maps our proposed benchmark tasks to attacks that are already documented in the African threat landscape, demonstrating that AfriVish-Bench measures practically relevant detection capabilities.

7. Ethical Safeguards

7.1. Data Access Controls

7.2. Dual-Use Mitigation

- No cloning of real named individuals. All voice profiles are generic TTS configurations.
- All AI-generated samples will carry SynthID watermarks, enabling provenance detection if dataset samples are misappropriated.
- Scam script archetypes are generic educational templates; no operational account numbers, agent IDs, or target PII are included.
- The benchmark is framed exclusively as a *detection* resource. We do not provide a guide for building attack infrastructure.

Table 5: Mapping of AfriVish-Bench tasks to documented African attacks.

Documented Attack	Primary Vector	Benchmark Task
ZW \$100M WhatsApp group scam (2026)	WhatsApp account pivot	T4
EcoCash OTP phishing via Starlink	OTP harvest, brand spoof	T1, T2, T3
Chitungwiza syndicate robocalls	Automated voice call	T1
Meta support impersonation calls	Accent + brand impersona	T1, T2
Family emergency voice cloning	Cloned voice, distress	T1, T3
ATHR-style TOAD campaigns (global)	AI agent, multi-brand	T1, T2, T3, T4

Table 6: Proposed AfriVish-Bench asset release policy.

Asset	Proposed Release Status
Synthetic generation scripts	Open (GitHub, MIT licence)
Accent configuration files	Open (GitHub)
Scam script archetypes	Restricted — researcher request
Full audio dataset	Gated — data use agreement required
Test split labels	Held out — eval server only
Real individual voice recordings	Never collected

7.3. Regulatory Alignment

This research is conducted in alignment with Zimbabwe’s **Cyber and Data Protection Act** and South Africa’s **Protection of Personal Information Act (POPIA)**. Both frameworks impose strict accountability on organisations processing communications data, and we expect AfriVish-Bench to support compliance tooling development for organisations subject to these regulations.

8. Project Roadmap

9. Expected Contributions

1. **First African-context vishing benchmark.** AfriVish-Bench will be the first evaluation resource specifically targeting AI voice agent fraud in the Sub-Saharan African mobile ecosystem, with brand labels and attack archetypes grounded in documented regional campaigns.
2. **Formalisation of the AI voice agent attack vector.** We will systematically characterise the ATHR-style attack mechanism as a benchmark category, providing a reproducible threat model for future research.

Table 7: Proposed project timeline.

Phase	Week	Status	Deliverable
P1: Dataset design	1	Planned	1,300 sample schema, label definitions
P2: Audio generation	1	Planned	Synthetic samples via Google TTS + ElevenLabs
P3: Transcription	1	Planned	Whisper large-v3 transcription pipeline
P4: Acoustic baselines	1–2	Planned	GMM, CNN-XGBoost, VGGish-LSTM
P5: LLM judge eval	2	Planned	Gemini + Claude transcript-only accuracy
P6: SynthID baseline	2	Planned	Watermark detector precision/recall on T1
P7: Live agent harness	2–3	Planned	Real-time voice agent adversarial pipeline
P8: Ensemble	3	Planned	Acoustic + metadata + LLM fusion results
P9: Benchmark release	Post	Planned	Gated dataset + eval server
P10: Leaderboard	Post	Planned	Open submissions via Apart Research

3. **Real-time adversarial test harness.** The live voice agent pipeline will provide the first reproducible framework for evaluating detection systems against operationally realistic adaptive synthetic callers, not just static audio corpora.
4. **SynthID coverage gap analysis.** By testing SynthID detection across multiple TTS backends (Google, ElevenLabs, open-source), we will produce the first empirical measurement of watermark coverage for African vishing attack tools.
5. **LLM-as-judge baseline for low-resource contexts.** Establishing Gemini and Claude zero-shot transcript-only accuracy will provide a deployment pathway for organisations without acoustic analysis hardware — relevant for the majority of African organisations that lack enterprise telephony security infrastructure.
6. **Policy-ready output.** All findings will be framed against the Zimbabwe Cyber and Data Protection Act and POPIA, producing immediately actionable guidance for regional regulators and mobile money operators.

10. Conclusion

AI voice agents have crossed a threshold. What was once a labour-intensive manual operation is now an automated, scalable, purchasable service. The ATHR platform — sold for \$4,000 on cybercrime forums as of April 2026 — demonstrates that any moderately motivated actor can now run fully autonomous vishing campaigns at industrial scale. For African mobile users, whose financial lives run through WhatsApp and EcoCash, and for whom a clean foreign accent signals institutional trust, this capability poses acute and

immediate risk.

AfriVish-Bench is proposed as the first systematic benchmark for detecting this threat in its African context. By combining a labelled synthetic dataset, four structured detection tasks, a live adversarial voice agent harness, and SynthID watermark integration, we aim to produce a resource that drives both research progress and practical deployable detection systems. The benchmark will be released as an open, community-maintained resource through the Apart Research platform.

Acknowledgements

I thank the Apart Research team and the AI Safety South Africa hub (Cape Town) for organising the Global South AI Safety Hackathon 2026, and Schmidt Sciences for supporting this programme.

References

- [1] Orchard, A., Baron, C., and Wojtyla, P. *ATHR: A New AI-Powered Vishing Platform for Automated TOAD Attacks*. Abnormal AI Security Research, April 16, 2026. <https://www.bleepingcomputer.com/news/security/new-athr-vishing-platform>
- [2] PIVOT Security. *AI Voice Phishing in 2026: A Comprehensive Field Guide*. PIVOT Security Blog, May 2026. <https://blogs.pivotsec.in/blogs/ai-enabled-voice-phishing>
- [3] Google / Mandiant. *M-Trends 2026: Frontline Cybersecurity Insights*. Mandiant, 2026. Vishing cited as second most common initial breach vector.
- [4] Google DeepMind. *SynthID: Watermarking AI-Generated Audio*. Technical Blog, 2024. <https://deepmind.google/technologies/synthid/>
- [5] MagicShot / Google I/O Coverage. *SynthID AI Watermark — Google I/O 2026: ElevenLabs and OpenAI Adoption, Content Detection API*. May 2026. <https://magicshot.ai/news/synthid-ai-watermark-google-io-2026-updates/>
- [6] Google. *SynthID Detector — a new portal to help identify AI-generated content*. Google Blog, January 2026. <https://blog.google/technology/ai/google-synthid-ai-content-detector/>
- [7] Jang, W. et al. *Joycent: Diffusion-based Accent TTS without Accented Phone Prediction*. arXiv:2405.XXXXX, 2024.
- [8] Badlani, R. et al. *RAD-MMM: Multi-lingual Multi-accented Multi-speaker Text-To-Speech*. INTERSPEECH 2023.
- [9] Alzantot, M. et al. *Enhancing Voice Spoofing Detection: A Hybrid VGGish-LSTM Approach*. IEEE Transactions on Information Forensics and Security, 2023.
- [10] Radford, A. et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI Technical Report, 2022.

- [11] Adebiyi, M. et al. *Detecting Smishing and Vishing Attacks using Machine Learning*. University of Johannesburg, Faculty of Engineering, 2024.
- [12] Zimbabwe Republic Police. *Public Warning: WhatsApp/EcoCash Impersonation Campaign — \$100M Losses Reported*. Assistant Commissioner Paul Nyathi, Press Statement, February 2026.
- [13] VishGuard Research Group. *VishGuard: AI-Powered Real-Time Defense Against Voice Phishing*. IEEE ICTBIG 2025. DOI: 10.1109/ICTBIG68706.2025.11323735.