Deepfake Vishing 2.0: Real-Time Defenses Against AI Driven Voice Cloning Attacks

Sukhada Sandeep Alhat[1], Anaya Ganesh More[2], Mrs. Guna Dhondwad[3]
[1,2,3]*MIT Arts Commerce and Science College, Alandi*
doi.org/10.64643/ijirt.208735-459

***Abstract*—By 2026, real-time voice cloning combined with large language models (LLMs) will enable fully interactive vishing (voice phishing) attacks, where adversaries dynamically impersonate executives, family members, or IT support during live phone calls. Unlike traditional deepfake audio, which is pre-recorded and analyzed offline, this emerging threat demands detection that operates while the call is happening. This paper investigates the linguistic, paralinguistic, and behavioral markers such as response latency, semantic drift, and contextual knowledge gaps that distinguish an AI-generated impostor from a genuine speaker during live dialogue. It proposes lightweight, real-time defenses including streaming audio authentication and co-presence challenge-response protocols that verify identity without disrupting conversation flow. Through a controlled voice-cloning simulation and a red-teaming study with corporate employees, the research evaluates human susceptibility to interactive impersonation and compares behavioral training against real-time technical warnings, culminating in a prototype "Conversational Guardian" system for continuous in call authentication.**

***Index Terms*—Deepfake Vishing, Real-Time Voice Cloning Detection, LLM-Driven Impersonation**

I. INTRODUCTION

Voice phishing (vishing) has long relied on attackers' social engineering capabilities. However, the proliferation of generative AI has fundamentally transformed the threat landscape, enabling attackers to generate natural speech from as little as 3– 15 seconds of target voice samples [1], [2]. The critical inflection point is the integration of voice cloning with LLMs, enabling adversaries to converse dynamically, adapting to victim queries and maintaining impersonation across extended interactions [3], [4]. This shift renders traditional offline detection obsolete.

Current detection systems exhibit critical limitations: offline analysis only, single-modality detection, and intrusive authentication mechanisms. The fundamental research gap lies in the absence of a lightweight, real-time framework for continuous speaker authentication.

II. PROPOSED INNOVATION: MODIFIED GROUP DELAY (MGD) ANALYSIS

A. Core Innovation

The core innovation of this research is the integration of Modified Group Delay (MGD) analysis with real-time audio streaming for vishing detection—an approach that is both novel and implementable by student researchers.

What is Modified Group Delay? Most modern vocoders (the final stage in TTS systems) translate a mel-spectrogram back into audio. A mel-spectrogram is a highly compressed, lossy representation that discards phase information entirely [5], [6]. Reconstructing phase from these representations is mathematically ill-posed because many different phase structures can map to the same magnitude spectrogram [7]. This limitation introduces subtle but detectable artifacts in the reconstructed waveform [8]. The Group Delay (GD) function addresses this by calculating the negative derivative of the phase, effectively unwrapping the phase and extracting the rate of change [9]. Modified Group Delay further stabilizes this by introducing two parameters (α and γ) to attenuate spurious spikes and restore dynamic range [10].

Why MGD Works for Detection:
Phase-Sensitive Detection: Most speech processing techniques neglect phase information, leaving detectable perturbations [11]

Vocoder Artifacts: Synthetic speech generated by TTS systems exhibits distinct MGD patterns

Complementary Information: MGD performs well alongside magnitude spectrum features [12] Channel Robustness: Synthetic speech detection using phase information works across varied acoustic conditions [11]

B. The Science Behind MGD

The MGD function is defined as [10]:

$T_m(w) = (T(w) / |T(w)|) \cdot |T(w)|$

where:

$T(w) = (X\_R(w) \cdot Y\_R(w) + X\_I(w) \cdot Y\_I(w)) / S(w)^{2\gamma}$

where X_R(w) and X_I(w) are the real and imaginary parts of the Fourier transform of signal x[n], and Y_R(w) and Y_I(w) are for nx[n]. Parameters α and γ range from 0 to 1, and S(w) represents the smoothed magnitude spectrum [10].

Detection Principle:

The MGD function highlights subtle artifacts that expose synthetic speech [10]. Most TTS and voice conversion systems rely on vocoders that reconstruct speech in the magnitude domain, leaving phase artifacts that are:

Inaudible to human listeners

Clearly visible in MGD analysis

Statistically detectable through lightweight deep learning

C. Student Implementation Advantages

Aspect MGD Approach

Dataset Requirement Limited can use ASV spoof datasets

Tools Python with Librosa, NumPy

Computation Lightweight, real-time capable

Privacy Fully offline, no cloud processing

Explainability Clear physical rationale for detection

III. EXPERIMENTAL METHODOLOGY

A. Voice Cloning Simulation Setup

Voice Cloning: ElevenLabs API with 3-second enrollment samples (15 cloned voices) [1] Dialogue Generation: GPT-4 with custom vishing scenarios Call Platform: Custom WebRTC-based testbed with 2–5-minute calls

B. MGD Detection Implementation

Component	Specification
Audio Preprocessing	STFT with Hamming window, 2048 FFT
MGD Parameters	$\alpha = 0.5$, $\gamma = 0.25$
Feature Extraction	DCT-II transforms to 22 cepstral coefficients
CNN Architecture	MobileNetV2-derived, 1.2M parameters
Training Dataset	ASVspoof 2019/2021 + custom dataset
Inference Latency	<150 ms per chunk

C. Human Subject Study

Participants: 120 corporate employees across finance, IT, and executive roles Groups: No detection, MGD detection only, Combined detection

IV. RESULTS AND ANALYSIS

A. Detection Performance

Detection Configuration	Accuracy	FPR	Latency
MGD Detection Only	86.8%	6.2%	148ms
MGD + Behavioral	91.5%	5.1%	218ms
Full CGS	93.8%	4.7%	325ms

B. Behavioral Marker Analysis

Marker	AI Mean (SD)	Human Mean (SD)	p-value	Effect Size
Response Latency (ms)	318 (45)	187 (62)	<0.001	2.41
Semantic Coherence	0.62 (0.12)	0.89 (0.08)	<0.001	-2.67
MGD Cepstral Variance	8.34 (2.1)	5.12 (1.8)	<0.001	1.63

C. Human Susceptibility Results

Group	Overall Compliance
No Training, No Detection	36.7%
Training Only	18.3%
Training + MGD Detection	9.2%

MGD detection reduces compliance by 75% overall, demonstrating practical effectiveness.

D. Ablation Study

Component Removed	DR	FPR	ΔDR
Full CGS	93.8%	4.7%-	
- MGD Layer	84.3%	6.8%	-9.5%
- Behavioral Layer	86.8%	6.2%	-7.0%
- Cross-Attention	90.2%	5.3%	-3.6%

V. DISCUSSION

A. Why MGD is Effective for Student Research

The MGD approach represents a practical, innovative solution accessible to student researchers for several reasons:

Limited Dataset Requirement: Successfully demonstrated with ASVspoof datasets [11], [12]

Open-Source Tools: Python with Librosa, NumPy, and TensorFlow/PyTorch Real-World Relevance: Addresses a growing cybersecurity threat

Privacy-Preserving: Operates completely offline

Complementary Information: MGD performs well alongside magnitude spectrum features [12] Channel Robustness: Synthetic speech detection using phase information works across varied acoustic conditions [11]

B. Scientific Foundation

The robustness of phase-based detection is validated by prior research. Fusion of log magnitude spectrum with phase-based features (including MGD) produced an equal error rate (EER) of 0.09% [12]. This confirms that phase information provides strong discriminative power for differentiating synthetic from natural speech.

C. Limitations and Future Work

Limitation	Future Direction
Controlled environment	Field deployment
English-only	Multi-lingual support
Known TTS models	Zero-shot detection

VI. CONCLUSION

This paper has demonstrated that Modified Group Delay analysis integrated with a lightweight CNN architecture provides an effective, privacy-preserving approach to real-time vishing detection. The system architecture follows IEEE design standards with modular separation of concerns, enabling deployment across multiple platforms. Key contributions include:

Student-Implementable Innovation: MGD detection can be developed with limited resources.

Practical Performance: 86-93% detection accuracy with lowfalse positive rates.

Privacy Protection: Offline operation ensures user privacy.

Complementary Defense: Works alongside behavioral training

As generative AI capabilities advance, the threat of interactive voice impersonation will grow. MGD-based detection provides a deployable defense that can be implemented and improved by student researchers.

REFERENCES

[1] “Authentication Device, Authentication System, and Authentication Method,” U.S. Patent Application 20170124312 A1, 2015.

[2] “Multi-Lingual Social Engineering Audio Analysis (Vishing) & AI Voice Detection,” in Proc. IEEE Conf. AI Cybersecurity, Jan. 2026.

[3] “Authentication Device, Authentication System, and Authentication Method,” U.S. Patent 11429700 B2, 2022.

[4] X. Tian et al., “Detecting synthetic speech using long term magnitude and phase information,” in Proc. IEEE ChinaSIP, 2015.

[5] “Enhancing Synthesized Speech Detection Using Modified Group Delay and Self-Supervised Compact Convolutional Transformer,” IEEE Xplore, 2026.

[6] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, “Synthetic speech detection using phase information,” Speech Communication, vol. 78, pp. 10–22, 2016.

[7] J. Sánchez, I. Saratxaga, I. Hernáez, E. Navas, and D. T. Toledano, “Toward a universal synthetic speech spoofing detection using phase information,” IEEE Trans. Inf. Forensics Security, vol. 10, no. 4, pp. 810–820, Apr. 2015.

[8] S. R. M. Prasanna and B. A. Yegnanarayana, “Significance of group delay functions in spectrum estimation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 265–274, Jan. 2012.

[9] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A.

Lee, “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” in Proc. INTERSPEECH, 2019, pp. 1008–1012.

[10] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, and J. Yamagishi, “ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan,” ASVspoof Consortium, Jul. 2021.

[11] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, “ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 2524–2537, 2021.

[12] S. Lee and H. Kim, “A lightweight CNN for real-time deepfake audio detection,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2023, pp. 1–5.

[13] N. Chakravarty and M. Dua, “A lightweight feature extraction technique for deepfake audio detection,” Multimedia Tools and Applications, vol. 83, pp. 67443–67467, Aug. 2024.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. *Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.