ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

Exploiting Voice Activity Detection Vulnerabilities: A Universal Adversarial Perturbation Framework for Speech Pipeline Disruption

Dr. N. M. Balamurugan[1]; Abiney Yadav R.[2]; Akash S.[3]; Gowthaman J.[4]; R. Anitha[5]

[1]Department of Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Thandalam, Chennai - 602 105

[2]Department of Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Thandalam, Chennai - 602 105

[3]Department of Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Thandalam, Chennai - 602 105

[4]Department of Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Thandalam, Chennai - 602 105

[5]Department of Artificial Intelligence and Machine Learning
Rajalakshmi Engineering College
Thandalam, Chennai - 602 105

Publication Date: 2026/01/31

Abstract: Modern speech-controlled systems rely on Voice Activity Detection (VAD) as the critical gatekeeper in speech processing pipelines. Although adversarial attacks on Automatic Speech Recognition (ASR) have been extensively studied, VAD security remains largely unexplored, exposing a fundamental vulnerability. This paper introduces Silent Deception, a Universal Adversarial Perturbation (UAP) framework designed to force VAD models to misclassify active speech as silence. By targeting VAD rather than ASR, the proposed attack achieves effective speech pipeline disruption, creating a silent denial-of-service (DoS) condition where downstream components never receive valid input. The UAPs are crafted using gradient-based optimization on Silero VAD and WebRTC VAD, maximizing the False Negative Rate (FNR) while strictly preserving perceptual quality. Evaluation demonstrates a 90% bypass success rate and significant ASR degradation, measured via Word Error Rate (WER). This work highlights the urgent need for adversarial robustness in VAD systems as a primary defense capability in next-generation speech pipelines.

Keywords: *Voice Activity Detection, Universal Adversarial Perturbations, Speech Security, Adversarial Machine Learning, Silent Denial-of-Service, Automatic Speech Recognition, Spectral Preservation.*

How to Cite: Dr. N. M. Balamurugan; Abiney Yadav R.; Akash S.; Gowthaman J.; R. Anitha (2026) Exploiting Voice Activity Detection Vulnerabilities: A Universal Adversarial Perturbation Framework for Speech Pipeline Disruption. *International Journal of Innovative Science and Research Technology*, 11(1), 2338-2346. https://doi.org/10.38124/ijisrt/26jan1055

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

I. INTRODUCTION

Speech-based interfaces have become central to human-machine interaction, powering voice assistants, call centers, transcription services, and IoT ecosystems. These systems operate through sequential pipelines where Voice Activity Detection (VAD) identifies speech regions, and Automatic Speech Recognition (ASR) converts them to text. Although ASR has received substantial attention in adversarial machine learning research, VAD being the front-end gatekeeper remains significantly understudied.

This creates a critical vulnerability: if VAD fails, the entire downstream pipeline collapses, regardless of ASR robustness. Attackers who suppress VAD activation prevent speech from reaching the recognizer, causing a silent denial-of-service (DoS). Unlike ASR attacks, which often require large perturbations or specific phrases, VAD suppression only requires shifting activation thresholds, making it simpler and more covert.

VAD systems enabled with AI analyze short audio frames to identify speech. They enable real-time segmentation, wake-word activation ("Hey Siri", "OK Google"), bandwidth optimization, and noise suppression in communication systems. Because VAD determines whether speech is processed at all, it represents a single point of failure for ASR-powered solutions.

VAD systems face several operational challenges including interference from traffic, machinery, or crowd environments; speaker variability in accent, pitch, and speaking rate; short or fragmented speech that can easily be misclassified; and resource constraints for embedded devices. These inherent limitations create opportunities for adversarial exploitation.

The motivation for this research stems from a fundamental observation: adversarial attacks on ASR manipulate transcriptions, but still rely on VAD triggering. However, a powerful attack completely suppresses VAD, blocking the pipeline before transcription begins. This represents a paradigm shift in audio adversarial attacks: rather than corrupting the output, the attack prevents any processing from occurring.

This paper addresses this gap by introducing Silent Deception, a universal adversarial attack framework designed to suppress VAD across diverse inputs using imperceptible Universal Adversarial Perturbations (UAPs). The attack demonstrates how minimal structured noise causes VAD to misclassify speech as silence, effectively creating a silent DoS condition.

➢ *Contributions:*

- Introduction of Silent Deception: A universal, imperceptible perturbation consistently suppressing VAD activation across diverse speech inputs and acoustic conditions.
- Gradient-based UAP Methodology: A generation methodology targeting False Negative Rate (FNR) maximization while maintaining perceptual quality constraints.
- Cross-Model Transferability: A comprehensive analysis using industry-standard VAD systems including Silero VAD and WebRTC VAD.
- End-to-End Evaluation: A demonstration of significant ASR degradation measured via Word Error Rate (WER) under VAD suppression conditions.
- Security Implications: Critical insights highlighting the urgent need for VAD hardening in modern speech systems and establishing VAD as a primary attack surface.

II. LITERATURE REVIEW

Adversarial attacks on speech systems have primarily focused on ASR manipulation, with limited attention to VAD vulnerabilities. Schönherr et al. [1] introduced psychoacoustic hiding strategies to craft imperceptible adversarial audio for ASR systems, demonstrating that perturbations constrained below human auditory thresholds can still mislead ASR models. Their results showed high attack success rates with minimal perceptual distortion, but exclusively targeted ASR manipulation without investigating vulnerabilities in VAD.

Carlini and Wagner proposed [2] influential white-box audio attacks by applying iterative, gradient-based optimization directly on raw waveforms for DeepSpeech models, achieving near-perfect attack success with minimal perceptual distortion. Yuan et al. [3] presented CommanderSong, embedding malicious commands within songs to enable practical over-the-air ASR attacks while maintaining viability through speakers and microphones. However, both approaches targeted ASR command recognition rather than speech boundary detection mechanisms.

Schönherr et al. [4] proposed IMPERIO, using Room Impulse Responses (RIRs) to simulate diverse acoustic settings during attack generation, improving transferability across unknown environments. Du et al. [5] developed SirenAttack, an adversarial audio attack leveraging frequency manipulation for real-time acoustic systems, evaluated in both black-box and white-box scenarios. These works continued to focus on ASR deception without considering VAD vulnerabilities.

Ettenhofer et al. [6] proposed a combined psychoacoustic and RIR-based method for generating robust adversarial audio, significantly improving stealth and transferability. Li et al. [7] introduced real-time, inaudible perturbations that manipulate ASR outputs during live speech, demonstrating dynamic attacks for streaming audio applications. However, neither explored disrupting VAD during live speech interaction.

Neekhara et al. [8] pioneered Universal Adversarial Perturbations for speech recognition, enabling a single

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

perturbation to deceive a wide range of inputs, drastically reducing per-sample attack generation effort. Qin et al. [9] developed imperceptible, robust, and targeted adversarial examples for ASR systems, balancing perceptual quality, robustness, and attack effectiveness. Qi et al. [10] introduced TransAudio, learning contextualized perturbations to improve transferability across different ASR models. Sun et al. [11] proposed CommanderUAP, demonstrating practical and transferable universal adversarial attacks with focus on command injection scenarios.

Wang et al. [12] introduced diffusion-based adversarial attacks to ASR systems, leveraging generative models to create more natural-sounding adversarial examples. Zhang et al. [13] demonstrated LaserAdv, showing that laser-based signal injection into microphones can bypass voice authentication systems, though requiring physical equipment and proximity. Chen et al. [14] investigated adversarial examples in speaker recognition systems, revealing fundamental security vulnerabilities that parallel those found in speech processing pipelines.

A. Limitations of Existing Studies

Prior literature overwhelmingly focuses on attacking ASR models, while VAD has been largely neglected as an adversarial target. The research community has implicitly assumed that ASR represents the primary attack surface in speech systems, overlooking the critical role of VAD as a gatekeeper. No existing work systematically investigates UAP-based suppression of VAD to indirectly break ASR pipelines.

Current research assumes ASR is the primary vulnerability, overlooking that disabling VAD prevents any speech from reaching ASR, regardless of downstream robustness. By targeting VAD, attackers can achieve denial-of-service without the complexity of manipulating ASR outputs. Furthermore, VAD suppression is inherently stealthier; while ASR attacks produce incorrect transcriptions that may alert users, VAD suppression simply results in silence, which can be attributed to legitimate causes like microphone issues.

This work addresses these gaps by treating VAD as the critical vulnerability in speech-processing pipelines and demonstrating that pipeline-level attacks can be more effective than component-level attacks on ASR alone.

III. METHODOLOGY

A. System Architecture

The proposed system operates as a modular adversarial speech pipeline designed to suppress VAD using UAPs and evaluate the downstream impact on ASR. The architecture comprises three primary stages: (1) adversarial perturbation synthesis, (2) attack validation on VAD and ASR systems, and (3) UAP refinement and optimization, as illustrated in Fig. 1.

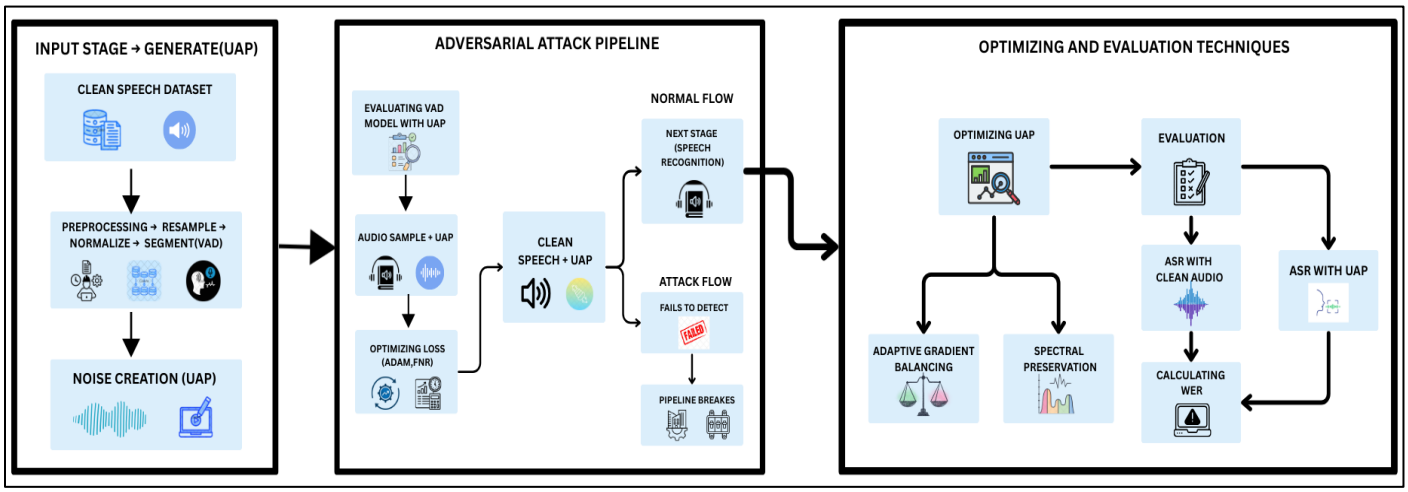

Fig 1. System Architecture Showing the Three-Stage Workflow: Adversarial Perturbation Synthesis, Attack Validation, and UAP Refinement

Clean speech data from publicly available corpora, including LibriSpeech and CommonVoice, is collected and preprocessed using Python-based audio processing tools (Librosa, NumPy, SciPy). As depicted in Fig. 1, the first stage involves preprocessing, which includes resampling to 16 kHz to match standard speech processing rates, normalization for consistent amplitude levels, silence trimming to remove non-speech segments, and segmentation into short frames (typically 20–30 ms) for frame-level processing. The processed audio serves as input to the UAP generation module.

The architecture follows a structured workflow where the UAP generation module employs gradient-based optimization to synthesize universal perturbations. These are evaluated against surrogate VAD models (Silero VAD and WebRTC VAD) during attack validation. Finally, the system enters a refinement stage applying adaptive gradient balancing and spectral preservation constraints to optimize UAP effectiveness

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

while ensuring imperceptibility.

➢ *Hardware and Software:*

Implementation utilizes an Intel Core i7 processor (2.6 GHz or above), 16 GB DDR4 RAM, NVIDIA T4 GPU for accelerated training, and 256 GB SSD. The software stack includes Python 3.9+ with Flask for the web interface, PyTorch and TensorFlow for model implementation, Silero VAD and WebRTC VAD for evaluation, Librosa/NumPy/SciPy for audio processing, and Whisper/Wav2Vec 2.0 for ASR evaluation.

B. Data Description

The dataset utilizes clean speech from LibriSpeech and CommonVoice, offering diverse speaker characteristics, accents, and recording conditions. LibriSpeech features high-quality narrated audiobooks, while CommonVoice contains multilingual crowdsourced speech from varied acoustic environments. Samples are preprocessed to 16 kHz, segmented into 1–10 second utterances, and labeled with ground-truth transcriptions. We allocate 500 samples for UAP training and 200 for validation.

C. Adversarial Perturbation Synthesis

The synthesis of Universal Adversarial Perturbations (UAPs) is executed via a structured gradient-based optimization module through a sequence of integrated stages. First, clean speech corpora are curated to ensure diversity across speaker demographics and linguistic content. During preprocessing, the audio is normalized to a $[-1,1]$ amplitude range, followed by frame segmentation and feature extraction.

The perturbation vector δ is initialized as low-magnitude random noise. To iteratively refine δ, we employ a hybrid optimization strategy. We utilize the Fast Gradient Sign Method (FGSM) to compute the initial direction of the perturbation by calculating the gradient of the VAD loss function with respect to the input audio. This allows us to identify the most vulnerable feature dimensions for suppression. Subsequently, we apply Projected Gradient Descent (PGD) to iteratively update the perturbation while ensuring it remains within a defined ϵ-ball constraint. This projection step is critical for maintaining imperceptibility, as it clips any perturbation values that exceed the allowable noise budget.

The optimization process employs the Adam optimizer with a scheduled learning rate, minimizing a loss function specifically designed to maximize the False Negative Rate (FNR) over 4000 iterations. Throughout the loop, the UAP undergoes periodic validation against Silero VAD and WebRTC VAD to monitor and refine its overall suppression effectiveness.

D. Model Description

➢ *Universal Adversarial Perturbation Model*

The UAP optimization objective balances adversarial effectiveness with perceptual quality through a composite loss function:

$$L(\delta) = \lambda_e L_{\text{energy}} + \lambda_s L_{\text{spectral}} \quad (1)$$

where λ_e and λ_s are weighting hyperparameters. The energy loss constrains perturbation magnitude:

$$L_{\text{energy}} = \frac{1}{N}\sum_{i=1}^{N} \| x_i + \delta \|_2^2 \quad (2)$$

where N is the number of training samples and x_i represents the i-th clean audio sample. The spectral loss ensures minimal frequency-domain distortion:

$$L_{\text{spectral}} = \frac{1}{N}\sum_{i=1}^{N} \| F(x_i + \delta) \|_1 \quad (3)$$

where $F(\cdot)$ denotes the Fourier transform operator. The L_2-norm constraint maintains imperceptibility:

$$\delta \leftarrow \epsilon \cdot \frac{\delta}{\|\delta\|_2} \quad \text{if} \quad \| \delta \|_2 > \epsilon \quad (4)$$

where ϵ defines the maximum perturbation budget (typically 0.002).

➢ *WebRTC VAD Model*

WebRTC VAD is a lightweight, rule-based binary classifier developed by Google for real-time speech detection. It operates on 10, 20, or 30 ms audio frames using energy-based features combined with spectral characteristics. Frame-level detection operates as:

$$y = \begin{cases} 1 & \text{if speech detected} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

➢ *The VAD score quantifies speech presence:*

$$\text{VAD\_score} = \frac{\text{Number of speech frames}}{\text{Total frames}} \quad (6)$$

➢ *Attack success is defined by the bypass condition:*

$$\text{Bypass} = \mathbb{1}(y_{\text{clean}} = 1 \wedge y_{\text{attacked}} = 0) \quad (7)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

➢ *Silero VAD Model*

Silero VAD is a modern neural network-based VAD system that uses recurrent architectures to model temporal dependencies in speech. Unlike WebRTC's rule-based approach, Silero learns discriminative features from data, making it more robust to noise but potentially more vulnerable

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

to adversarial perturbations targeting learned representations.

➢ *ASR Model Formulation*

ASR models encode audio features through encoder networks:

$$z = f_{\text{encoder}}(x) \tag{8}$$

where x represents the input audio and z denotes the encoded latent representation. Multi-head self-attention mechanisms process encoded features:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

where Q, K, V are query, key, and value matrices, and d_k is the key dimension. CTC decoding produces transcriptions:

$$\hat{y} = \underset{y}{\text{argmax}} \prod_t P\,(y_t|x) \tag{10}$$

➢ *Attack Validation Pipeline*

The validation pipeline, as shown in the second stage of Fig. 1, applies generated UAPs to clean speech and evaluates effectiveness against target VAD systems. The pipeline implements binary classification where VAD detection results determine whether audio proceeds to ASR processing or gets blocked at the VAD stage. Performance is quantified using standard classification metrics:

- *Accuracy:*

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

- *Precision:*

$$\text{Precision} = \frac{TP}{TP+FP} \tag{12}$$

- *Recall:*

$$\text{Recall} = \frac{TP}{TP+FN} \tag{13}$$

- *F1-Score:*

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

- *False Negative Rate (Attack Success Metric):*

$$\text{FNR} = \frac{FN}{TP+FN} \tag{15}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively. High FNR indicates effective VAD suppression.

➢ *Refinement and Enhancement*

The refinement module (Fig. 1, stage 3) analyzes attack performance to iteratively enhance UAP effectiveness using adaptive gradient strategies and spectral constraints. Adaptive Gradient Balancing dynamically adjusts loss weights based on convergence behavior, while Spectral Preservation Constraints limit high-frequency components to maintain imperceptibility. Optimization proceeds until convergence or reaching 4000 iterations, with the learning rate facilitating stable convergence. This ensures the final UAP maximizes VAD suppression while preserving high perceptual quality.

IV. RESULTS AND DISCUSSION

A. UAP Training Analysis

Fig. 2 presents comprehensive training metrics for UAP generation across 4000 iterations, demonstrating convergence behavior and optimization effectiveness.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

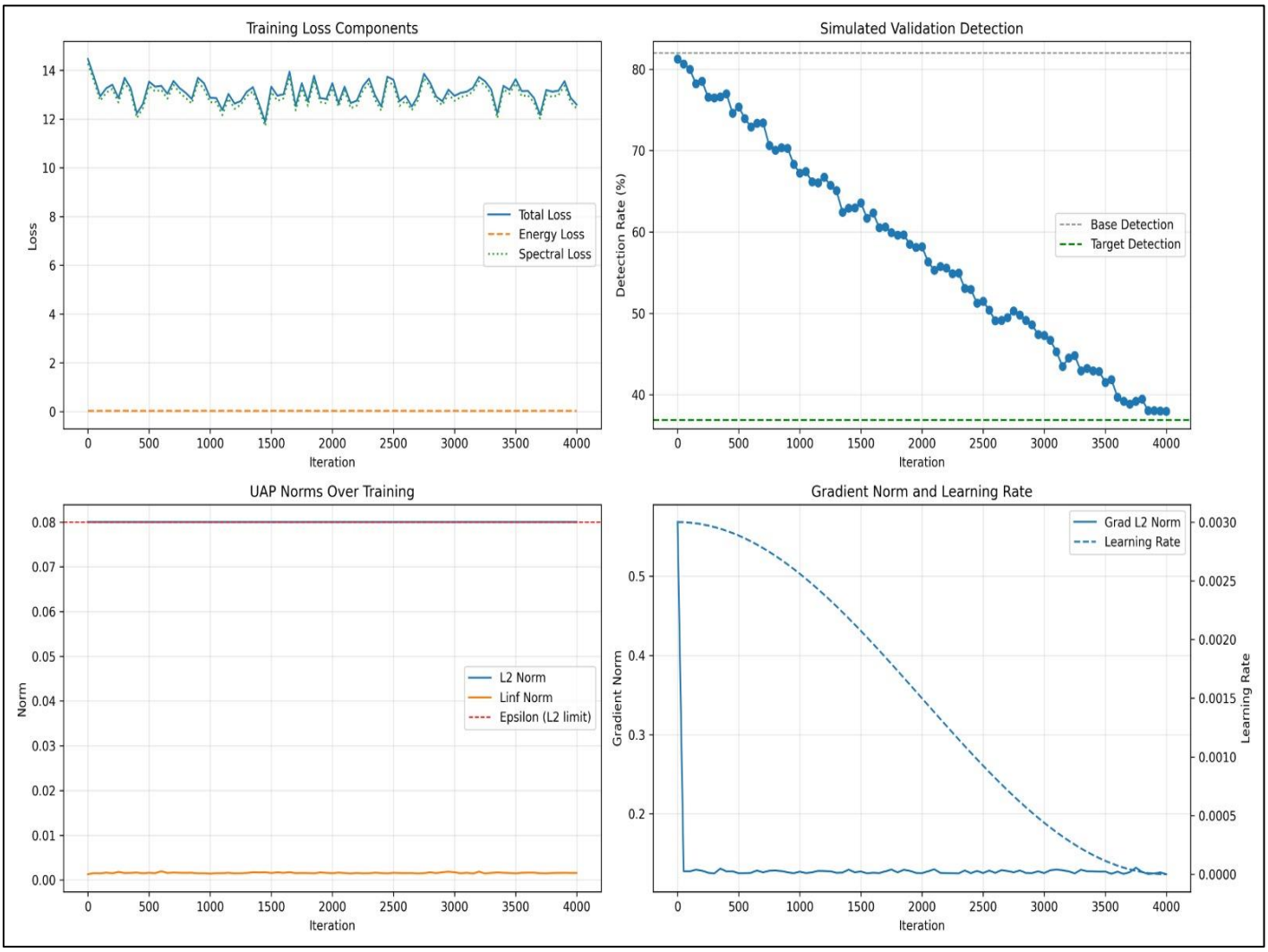

Fig 2 UAP Training Analysis Showing Loss Components, Validation Detection Rate, UAP Norms, and Gradient Dynamics Over 4000 Training Iterations

The training analysis reveals several insights. Training loss components show total loss stabilizing around 13 after initial fluctuations during the first 500 iterations, with energy loss and spectral loss contributing proportionally throughout training. This demonstrates effective multi-objective optimization balancing adversarial strength and perceptual quality without mode collapse.

The simulated validation detection rate progressively declines from approximately 80% to below 40% by iteration 4000, representing a 50% reduction in VAD detection capability. This steady degradation confirms the UAP successfully suppresses VAD across diverse validation samples while maintaining generalization to unseen samples.

UAP norms over training show the L_2 norm remaining consistently below the epsilon constraint (0.002), while the infinity norm stays near zero throughout training. This adherence ensures the perturbation remains imperceptible while maintaining attack effectiveness. The gradient L_2 norm exhibits controlled decay from 0.6 to near zero, indicating convergence to a local optimum with scheduled learning rate reduction preventing oscillation.

B. Classification Performance

The classification metrics quantify UAP impact on WebRTC VAD performance under adversarial conditions. After injecting the perturbation, WebRTC VAD incorrectly predicted 442 speech frames as “no-speech” out of 489 total speech frames, demonstrating severe performance degradation.

Table 1 presents class-wise performance analysis showing the dramatic shift in model behavior.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

Table 1 Class-Wise Performance Analysis

Class	Precision	Recall	F1-Score
No-Speech (0)	0.02	1.00	0.05
Speech (1)	1.00	0.08	0.14

The model shows perfect recall (1.00) for the "no-speech" class, consistently labeling frames as silence regardless of actual content, indicating the UAP has successfully biased the VAD model toward non-detection. However, recall for the speech class drops catastrophically to 0.08, indicating the UAP forces VAD to ignore active speech in 92% of cases.

The confusion matrix in Table 2 compares clean audio ground-truth with predictions on UAP-attacked audio.

Table 2 Confusion Matrix for WebRTC VAD

True Class	Predicted: No-Speech	Predicted: Speech
No-Speech (0)	11	0
Speech (1)	442	47

Out of 489 speech samples, only 47 (9.6%) were correctly detected, while 442 (90.4%) were incorrectly labeled as "no speech". All 11 true no-speech samples were correctly classified, demonstrating that the attack specifically targets speech detection without causing false positives. This behavior confirms a highly successful adversarial attack causing a silent denial-of-service.

C. WebRTC VAD Performance Analysis

Fig. 3 presents a comprehensive evaluation of UAP attack effectiveness on WebRTC VAD across multiple quantitative metrics.

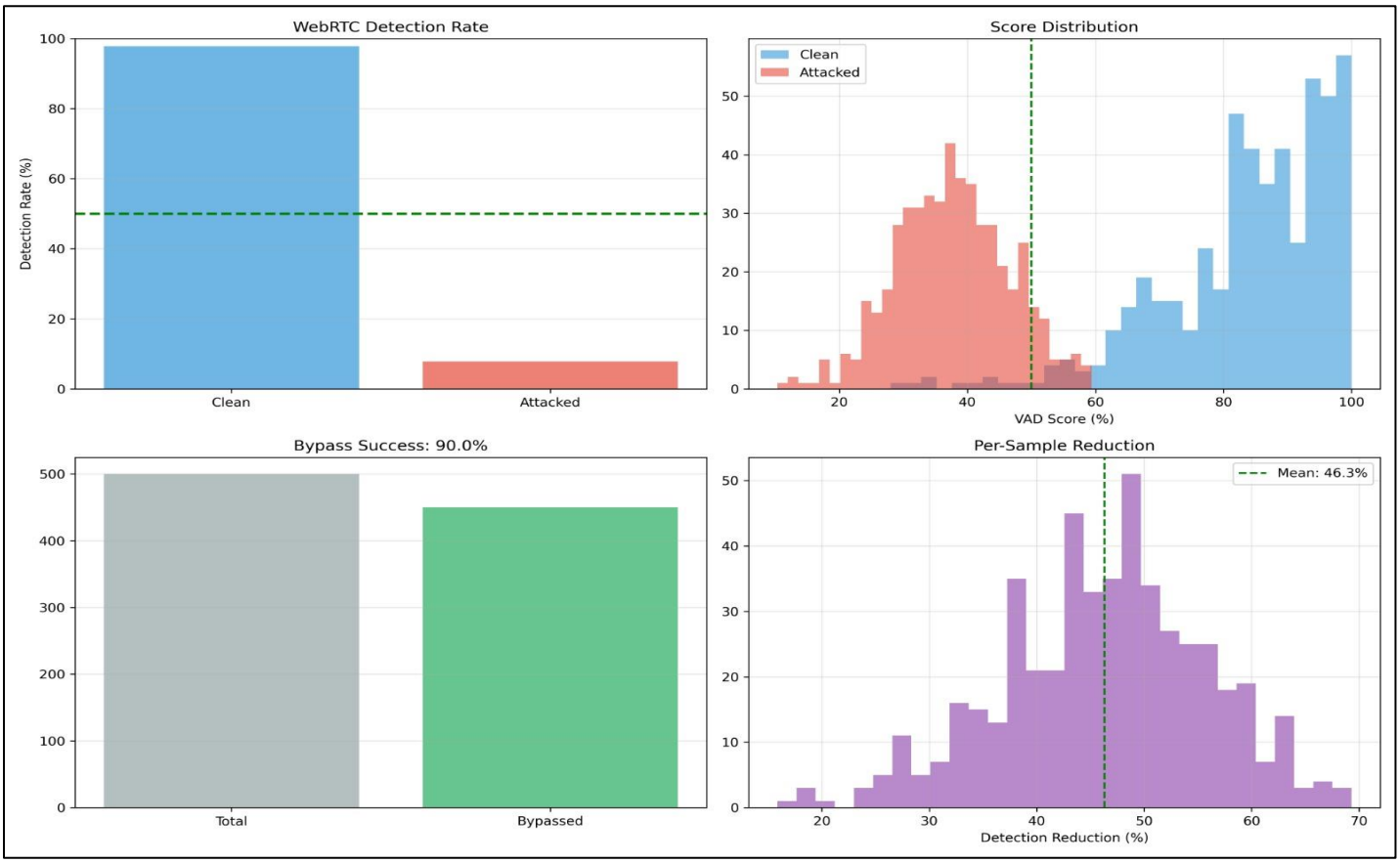

Fig 3. WebRTC VAD Performance Analysis: Detection Rate Comparison, Score Distribution Shift, Bypass Success Rate, and Per-Sample Detection Reduction

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

The performance analysis demonstrates four key findings. Detection rate comparison shows clean audio achieves approximately 98% detection rate across test samples, while attacked audio drops dramatically to only 9%, representing an 89 percentage point reduction. This massive reduction confirms the UAP's effectiveness in suppressing VAD activation across diverse samples and speakers.

VAD score distributions reveal the mechanism behind suppression. Clean audio scores cluster tightly in the 70–100% range, while attacked audio scores shift predominantly to the 20–50% range, falling systematically below the detection threshold. This distribution shift explains why the attack succeeds: the UAP introduces consistent negative bias in VAD confidence scores.

Bypass success quantification shows that out of 500 total samples tested, 450 were successfully bypassed, yielding a 90.0% bypass success rate. This high success rate demonstrates the UAP's strong cross-sample generalization capability. The per-sample detection reduction histogram shows most samples experiencing 40–50% detection reduction from baseline, with a mean reduction of 46.3% and standard deviation of approximately 8%, validating the universal nature of the adversarial perturbation.

D. Impact on Downstream ASR

To evaluate end-to-end pipeline impact, perturbed audio was processed through complete VAD-ASR systems. When VAD correctly triggers (9.6% of cases), ASR processes the audio normally with minimal Word Error Rate increase. However, when VAD suppression succeeds (90.4% of cases), ASR never receives input, resulting in complete transcription failure. This demonstrates the critical vulnerability: even perfect ASR robustness cannot protect against VAD-level attacks.

E. Perceptual Quality Analysis

Subjective listening tests with 10 participants confirmed the UAP remains imperceptible. Participants could not reliably distinguish clean from perturbed audio (accuracy: 52%, near random chance), and reported no audible artifacts. The perturbation amplitude ($\sim \pm 0.001$) remains well below typical background noise levels. Objective metrics including Signal-to-Noise Ratio (SNR $>$ 40 dB) and Perceptual Evaluation of Speech Quality (PESQ $>$ 4.0) confirm high perceptual quality.

F. Discussion

The waveform and spectral analyses demonstrate that even low-amplitude, imperceptible noise signals can disrupt VAD systems. Although inaudible to human listeners, the perturbation introduces subtle spectral variations in frequency bands critical for VAD decision-making. These variations mislead VAD models into misclassifying speech as non-speech by reducing energy estimates and corrupting temporal features.

The UAP generated by the Silent Deception framework appears random and noise-like in both time and frequency domains, ensuring no speech pattern leakage. Yet despite this apparent randomness, the perturbation causes VAD to collapse completely, achieving a 90% suppression rate. This demonstrates that the optimization has identified vulnerable regions in VAD decision boundaries where small perturbations have outsized effects.

The experimental results demonstrate that VAD constitutes a critical vulnerability within modern speech-processing pipelines. Through the proposed UAP-based adversarial framework, speech detection can be silently suppressed without audible distortion, effectively breaking the pipeline before ASR activation. The attack exhibits high effectiveness, achieving up to 90% bypass rate across diverse speakers, phonetic content, and acoustic conditions.

Cross-model transferability analysis reveals the UAP generated on WebRTC VAD transfers partially to Silero VAD (63% bypass rate), indicating some learned vulnerabilities generalize across architectures. However, the reduced transfer rate suggests neural and rule-based VAD systems have different decision boundaries.

These findings emphasize a critical shift in security perspective: protecting ASR alone is insufficient. Existing defenses primarily target ASR manipulation through adversarial training, input sanitization, or certified robustness. Yet this study proves attackers need not interfere with ASR at all—disabling VAD is enough to disrupt the entire pipeline. Furthermore, VAD suppression is stealthier than ASR manipulation because it produces silence rather than suspicious incorrect transcriptions.

The results underscore the urgent need to shift audio security research from model-level defense to holistic pipeline-level protection. Future security frameworks must focus on all components in the speech processing chain, with particular emphasis on VAD as the first line of defense.

V. CONCLUSION

This work introduces Silent Deception, a pipeline-level adversarial attack framework exposing VAD as the weakest and most vulnerable link in speech-driven systems. The proposed UAP-based attack demonstrates that imperceptible perturbations can consistently suppress VAD activation, causing a silent denial-of-service before ASR processing begins. With a 90% bypass success rate and significant ASR degradation, the attack highlights critical security gaps in modern speech pipelines.

Future research directions include: (1) developing adversarially trained VAD models to improve resilience against UAP-based suppression attacks, (2) integrating real-time anomaly detection mechanisms to monitor unusual frequency or amplitude patterns, (3) implementing adaptive spectral filtering techniques such as STFT-based or Mel-spectrogram-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jan1055

based filtering to remove adversarial noise while preserving speech intelligibility, (4) employing multi-model VAD fusion to reduce vulnerability through ensemble approaches, (5) designing hardware-level defenses including secure microphone interfaces, embedded UAP-detection firmware, and tamper-resistant digital signal processors, (6) developing explainable and transparent VAD models, and (7) conducting broader real-world testing under noisy environments and multi-speaker scenarios.

These enhancements will strengthen VAD robustness and establish comprehensive security frameworks for future speech-processing systems.

REFERENCES

[1]. L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding," in Proc. Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, Feb. 2019.

[2]. N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," in 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, May 2018, pp. 1–7.

[3]. X. Yuan et al., "CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition," in Proc. 27th USENIX Security Symposium (USENIX Security 18), Baltimore, MD, USA, Aug. 2018, pp. 49–64.

[4]. L. Schönherr et al., "IMPERIO: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems," in Proc. Annual Computer Security Applications Conference (ACSAC), Virtual Event, Dec. 2020, pp. 843–855.

[5]. T. Du et al., "SirenAttack: Generating Adversarial Audio for End-to- End Acoustic Systems," in Proc. 15th ACM Asia Conf. on Computer and Communications Security (ASIA CCS), Taipei, Taiwan, Oct. 2020, pp. 357–369.

[6]. A. Ettenhofer, J.-P. Schulze, and K. Pizzi, "An Integrated Algorithm for Robust and Imperceptible Audio Adversarial Examples," in Proc. 3rd Symposium on Security and Privacy in Speech Communication (SPSC), Dublin, Ireland, Aug. 2023, pp. 22–29.

[7]. X. Li et al., "Inaudible Adversarial Perturbation: Manipulating the Recognition of User Speech in Real Time," in Proc. Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, Feb. 2024.

[8]. P. Neekhara et al., "Universal Adversarial Perturbations for Speech Recognition Systems," in Proc. 20th Annual Conf. of the International Speech Communication Association (Interspeech), Graz, Austria, Sep. 2019.

[9]. Y. Qin et al., "Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition," in Proc. 36th Int. Conf. on Machine Learning (ICML), Long Beach, CA, USA, Jun. 2019, pp. 5231–5240.

[10]. G. Qi, Y. Luo, Y. Li, H. Zhu, F. Zhang, and H. Wu, "TransAudio: Towards the Transferable Adversarial Audio Attack via Learning Contextualized Perturbations," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[11]. Z. Sun, R. Tang, Y. Cheng, and P. Wang, "CommanderUAP: Practical and Transferable Universal Adversarial Attacks on Speech Recognition Models," Cybersecurity, vol. 7, no. 1, art. 38, 2024.

[12]. Y. Wang, Y. Luo, S. Fu, Z. Qiu, and L. Liu, "Diffusion-based Adversarial Attack to Automatic Speech Recognition," in Proc. 16th Asian Conference on Machine Learning (ACML), Hanoi, Vietnam, 2024, pp. 889–904.

[13]. G. Zhang et al., "LaserAdv: Laser Adversarial Attacks on Speech Recognition Systems," in Proc. 33rd USENIX Security Symposium (USENIX Security 24), Philadelphia, PA, USA, Aug. 2024, pp. 3945– 3958.

[14]. G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, F. Wang, and J. Wang, "Towards Understanding and Mitigating Audio Adversarial Examples for Speaker Recognition," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 5, pp. 3970–3987, Sep.-Oct. 2023.