Adversarial Attacks on Language Models

Ahmed Tamer Samir Mohamed, Salma Mohamed Hamed Mostafa, Saad Ahmed Saad Ali
Omar Khaled Abdel Aleem Ali, Fatma Hussein Abdel Wahed Zaher, Youssef Hesham Abdel Fattah Mohamed
Department of Electronics and Electrical Communications Engineering
Cairo University
Supervised by: Prof. Dr. Hanan Ahmed Kamal and Dr. Mohamed Abdo Tolba

***Abstract*—This paper presents a comprehensive investigation into the adversarial vulnerabilities of deep learning models deployed in natural language processing (NLP), large language models (LLMs), and automatic speech recognition (ASR). We systematically evaluate the impact of state-of-the-art adversarial attacks on representative architectures such as BiLSTM-GloVe for sentiment analysis, Qwen-1.5 for LLMs, and Wav2Vec 2.0 for ASR. Our study implements and benchmarks a diverse suite of attack algorithms. In addition, we design and implement a range of defense mechanisms. Experimental results reveal the nuanced interaction between attack granularity and defense effectiveness, demonstrating that hybrid and layered defenses can significantly reduce attack success. Our findings highlight critical trade-offs between robustness, clean-data performance, and computational overhead, and provide insights for deploying resilient, trustworthy language AI systems in safety-critical and real-world environments. This work establishes unified benchmarks and defense blueprints, advancing the secure adoption of language models across various domains.**

***Index Terms*—Adversarial Attacks, Deep Learning, Large Language Models, Natural Language Processing, Automatic Speech Recognition, Model Robustness, Defense Mechanisms, White-Box Attacks, Black-Box Attacks, Security of AI Systems**

I. Introduction

Language models have become a cornerstone of artificial intelligence, enabling machines to process, understand, and generate human language. Evolving from simple statistical models to advanced deep learning architectures, these systems now underpin a wide range of applications, from speech recognition to text analysis and generative AI.

This paper investigates the adversarial vulnerabilities of language models across three major domains:

- **Traditional Natural Language Processing (NLP):** Models focused on tasks like sentiment analysis and text classification, often using architectures such as RNNs or early Transformers.
- **Large Language Models (LLMs):** Powerful, pre-trained models capable of advanced reasoning and generation, but with an expanded attack surface.
- **Automatic Speech Recognition (ASR):** Systems that convert spoken language into text, serving as the first step for many language-based applications.

Across these domains, we demonstrate that language models—regardless of their complexity—are susceptible to adversarial attacks. Small, carefully crafted perturbations in input data can cause significant failures, highlighting critical security challenges for real-world AI deployments. This work provides a concise overview of these vulnerabilities and sets the stage for a deeper exploration of attack strategies and defense mechanisms in subsequent sections.

II. Related Work

Adversarial attacks on language models have become a major concern as these models are increasingly integrated into real-world systems. While early research centered Aynı zamanda adversarial vulnerabilities in computer vision, recent studies have systematically exposed similar threats in natural language processing (NLP), large language models (LLMs), and automatic speech recognition (ASR) [1]–[5].

A. Adversarial Attacks on NLP Models

Text-based adversarial attacks exploit the discrete and semantic structure of language. Word-level attacks such as TextFooler [2] replace important words with semantically similar alternatives, maintaining grammaticality while causing model misclassification. Character-level attacks like TextBugger [3] and DeepWordBug [6] introduce small character manipulations (insertions, deletions, swaps), which are often imperceptible to humans but highly effective against NLP models. These attacks are commonly evaluated in black-box settings, using frameworks like TextAttack [7] for standardized benchmarking.

B. Adversarial Attacks on Large Language Models

LLMs present a broader attack surface, with jailbreak and prompt injection attacks posing new risks. Jailbreak attacks craft adversarial prompts to bypass safety and alignment mechanisms, eliciting undesired or harmful outputs [5], [8]. Prompt injection attacks exploit retrieval-augmented or tool-integrated systems by embedding malicious instructions in external data, hijacking model behavior during inference [**?**]. Automated and transferable adversarial suffix generation [5].

C. Adversarial Attacks on Audio and ASR Models

Adversarial attacks on ASR systems have revealed that even state-of-the-art speech recognition models are susceptible to carefully crafted audio perturbations. Early work demonstrated that adding imperceptible noise can cause ASR systems to transcribe incorrect or attacker-specified text [4]. Attacks such as the Fast Gradient Sign Method (FGSM) and Projected

Gradient Descent (PGD), originally developed for vision, have been adapted to the audio domain, generating adversarial waveforms that remain intelligible to humans but deceive ASR models [4], [9].

More advanced methods leverage psychoacoustic principles, hiding perturbations beneath the auditory masking threshold to ensure imperceptibility [4]. Optimization-based attacks, such as those using the Cramér Integral Probability Metric (Cramér-IPM), jointly optimize for transcription loss and perceptual similarity, achieving high attack success rates while preserving audio quality [9]. Transfer-based black-box attacks exploit the transferability of adversarial examples between different ASR architectures, but recent work has shown that architectural and training objective mismatches can significantly reduce transfer success [10].

D. Defense Mechanisms

Defenses against adversarial attacks on language and ASR models combine proactive and reactive techniques. Virtual Adversarial Training (VAT) [11] regularizes models by smoothing output distributions around perturbed inputs, improving robustness without requiring labeled adversarial examples. Character-level purification heuristics correct common adversarial text manipulations [7], [12]. For LLMs, multi-stage input filtering systems, such as Adversarial Prompt Shield (APS), detect and block toxic or adversarial prompts [5]. In ASR, signal processing defenses like spectral gating, MP3 compression, and input quantization can remove or disrupt adversarial noise [13]–[15].

III. Natural Language Processing

This section will analyze adversarial attacks on traditional NLP models, such as BiLSTM-GloVe, focusing on sentiment analysis and text classification tasks. Detailed experiments and results will be provided in the final manuscript.

IV. Large Language Models

This section will explore adversarial vulnerabilities in large language models, such as Qwen-1.5, including jailbreak and prompt injection attacks. Experimental results and defense strategies will be detailed in the final manuscript.

V. Automatic Speech Recognition

A. Experimental Setup

1) Model: Pre-trained Wav2Vec 2.0, a state-of-the-art self-supervised Transformer-based ASR model.

2) Dataset: LibriSpeech test-clean subset (100 samples). LibriSpeech is a standard benchmark for ASR, ensuring reproducibility and comparability.

3) Metrics:

- **Character Error Rate (CER):** Measures the percentage of characters incorrectly predicted.
- **Word Error Rate (WER):** Measures the percentage of words incorrectly predicted.
- **Signal-to-Noise Ratio (SNR):** Ensures adversarial audio remains intelligible to humans, with SNR thresholds referenced from auditory science [16] (greater than or equal to 7 dB for 100% intelligibility in complex noise).

B. Attacks

1) Tested Attacks:

1) **Fast Gradient Sign Method (FGSM):** The Fast Gradient Sign Method (FGSM) is a seminal single-step adversarial attack technique designed to generate adversarial examples by introducing imperceptible perturbations to input data, thereby misleading machine learning models [1]. Specifically, FGSM crafts adversarial examples using the equation $x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y_t))$, where x_{adv} is the adversarial input, x is the original input, ϵ controls the perturbation magnitude, and $\nabla_x L(x, y_t)$ is the gradient of the loss function with respect to the input for a target transcription y_t. In the context of automatic speech recognition (ASR), this attack crafts adversarial audio samples that force the model to produce a specified target transcription [17].
2) **Projected Gradient Descent (PGD):** The Projected Gradient Descent (PGD) attack is a powerful iterative adversarial attack method that builds upon the principles of the Fast Gradient Sign Method (FGSM) by applying multiple small perturbations to optimize the adversarial example within an ℓ_∞ norm constraint [18]. PGD iteratively updates the perturbation using $\eta = \eta + \alpha \cdot \text{sign}(\nabla_\eta L(x+\eta, y_t))$, where η is the perturbation, α is the step size, and $\nabla_\eta L(x+\eta, y_t)$ is the gradient of the loss function with respect to the perturbation for a target transcription y_t. The adversarial example is then computed as $x_{adv} = x + \eta$, where x_{adv} is the adversarial input and x is the original input. In automatic speech recognition (ASR) systems, PGD generates adversarial audio samples by iteratively adjusting the input waveform to maximize the loss function, aiming to force the model to produce a specified target transcription [17].
3) **Cramér Integral Probability Metric (IPM):** The Cramér Integral Probability Metric (IPM) attack is an advanced adversarial method designed to generate robust adversarial audio signals for automatic speech recognition (ASR) systems [9]. The method employs an iterative update process defined by the equations:

$$L = L_{CTC} + \lambda \cdot \text{mean}(\ell_{\text{Cramér}})$$

$$\delta \leftarrow \text{AdamUpdate}(\delta, \nabla_\delta L)$$

where $\ell_{\text{Cramér}} = \sum_{j=1}^{m} (F_{\text{org}}(t_j) - F_{\text{adv}}(t_j))^2 \cdot \Delta t$, with L_{CTC} being the Connectionist Temporal Classification loss, λ a weighting factor, $\ell_{\text{Cramér}}$ the Cramér loss measuring distribution discrepancy, δ the perturbation, and AdamUpdate the optimization step using the gradient $\nabla_\delta L$. By leveraging the Cramér-IPM, this attack minimizes the discrepancy between the distributions of original and adversarial audio samples, ensuring the adversarial signals remain close to the legitimate speech manifold. Unlike traditional attacks like FGSM or PGD,

the Cramér-IPM approach combines the CTC loss with a statistical distance metric, enhancing robustness against over-the-air playback without relying on costly Expectation Over Transformation (EOT) operations [9]. This method optimizes perturbations to maximize transcription errors while maintaining perceptual similarity, making it effective for both targeted and non-targeted attacks.

4) **Imperceptible, Robust, and Targeted Attack:** The Imperceptible, Robust, and Targeted Attack algorithm [4] is designed to generate adversarial audio examples that force automatic speech recognition (ASR) systems to transcribe a specific target phrase, while ensuring the perturbations remain inaudible to human listeners using psychoacoustic principles. The algorithm operates through two stages with key equations:

$$x_{adv} = x + \delta$$

$$l = l_{Net} + \alpha \cdot l_{Psd}$$

where x_{adv} is the adversarial input, x is the original audio, δ is the perturbation updated iteratively in Stage 1 using CTC loss optimization and clipped within an adaptive ϵ constraint, l_{Net} is the CTC loss, l_{Psd} is the psychoacoustic loss based on the masking threshold, and α is a weighting factor. It operates in two stages: Stage 1 generates an initial perturbation using CTC loss optimization with an adaptive ϵ constraint, and Stage 2 refines this perturbation to stay below the psychoacoustic masking threshold. This ensures the adversarial audio is imperceptible while achieving targeted transcription errors, leveraging psychoacoustic principles for stealth.

2) Results:

Original CER: 0.9% — Original WER: 2.84%.

TABLE I
ATTACK EFFECTIVENESS OF FGSM

ϵ	A-CER (%)	A-WER (%)	SNR (dB)	ΔCER	ΔWER
0.001	1.72	5.07	57.56	0.82	3.87
0.005	2.83	7.16	43.58	1.93	4.33
0.01	1.73	5.89	37.56	0.83	3.05
0.02	2.42	7.60	31.54	1.52	4.76
0.05	2.85	7.83	23.58	1.95	4.99
0.10	3.57	9.68	17.56	2.67	6.84
0.20	7.60	15.16	11.54	6.70	12.32

TABLE II
ATTACK EFFECTIVENESS OF PGD

ϵ, α	A-CER (%)	A-WER (%)	SNR (dB)	ΔCER	ΔWER
0.001, 0.0001	4.00	9.00	62.48	2.90	6.76
0.005, 0.0005	4.50	11.00	50.05	3.40	8.76
0.01, 0.001	5.34	14.70	37.56	4.44	11.86
0.02, 0.002	9.80	22.72	31.54	8.90	19.88
0.05, 0.005	24.21	44.36	23.58	23.31	41.52
0.10, 0.01	37.90	61.98	17.56	37.00	59.14
0.20, 0.02	58.39	81.44	11.54	57.49	78.60

TABLE III
ATTACK EFFECTIVENESS OF CRAMÉR-IPM

ϵ	A-CER (%)	A-WER (%)	SNR (dB)	ΔCER (%)	ΔWER (%)
0.001	20.5	26.1	18.39	19.6	23.3
0.005	18.8	32.1	16.17	17.9	29.3
0.010	40.8	59.6	14.27	39.9	56.8
0.020	61.8	81.1	12.08	60.9	78.3
0.050	93.4	99.0	9.42	92.5	96.2
0.100	95.5	98.9	8.19	94.6	96.1
0.200	92.9	98.4	8.05	92.0	95.6

TABLE IV
ATTACK EFFECTIVENESS OF IMPERCEPTIBLE ATTACK

Stage	WER (%)	CER (%)	SNR (dB)	ΔWER (%)	ΔCER (%)
Stage 1	75.88	75.88	40.48	73.04	74.98
Stage 2	100.00	100.00	30.39	97.16	99.10

TABLE V
COMPARISON OF ATTACK EFFECTIVENESS: WAV2VEC2 VS. ORIGINAL PAPERS

Attack	**Model**	**Condition**	**WER (%)**
FGSM	Wav2vec2 (This Work)	$\epsilon = 0.01$	5.89
	Wav2vec2 (This Work)	$\epsilon = 0.2$	15.16
	DeepSpeech 2 [17]	$\epsilon = 0.01$	77.9
	DeepSpeech 2 [17]	$\epsilon = 0.2$	100
	Espresso [17]	$\epsilon = 0.01$	65.3
	Espresso [17]	$\epsilon = 0.2$	108
PGD	Wav2vec2 (This Work)	$\epsilon = 0.01$	14.70
	Wav2vec2 (This Work)	$\epsilon = 0.2$	81.44
	DeepSpeech [17]	$\epsilon = 0.01$	97.5
	Espresso [17]	$\epsilon = 0.2$	139.3
Cramér-IPM	Wav2vec2 (This Work) III	$\epsilon = 0.05$	99.0
	DeepSpeech [9]	Average	88.19
Two-Stage (Stage 1)	Wav2vec2 (This Work)	-	75.88
	Lingvo [4]	-	100 (100% success)
	Wav2vec2 (This Work) (Stage 2)	-	100
	Lingvo [4] (Stage 2)	-	100 (100% success)

These results highlight the robustness of wav2vec2 against FGSM and PGD at lower ϵ values, with WER significantly increasing only at higher perturbations. The Cramér-IPM attack proves highly effective, achieving near-perfect error rates, while the two-stage attack successfully refines perturbations for imperceptibility. Future work could explore optimizing SNR at higher ϵ values to balance attack success and audio quality, especially for real-world applications.

C. Defenses

1) Tested Defenses:

1) **Randomized smoothing (RS):** Randomized smoothing [19] creates noise-invariant predictions. And it is an effective defense with speech to text models [17]. We implemented it with:
   - **Adaptive variance scaling [20]:** adjusting the noise level (variance) added to an input signal based on its local characteristics, it dynamically scales the noise to be lower in high-variance regions (to preserve speech content) and higher in low-variance regions (to mask potential adversarial perturbations).

- **ROVER voting [21]:** ROVER (Recognizer Output Voting Error Reduction) is a method for combining multiple transcription outputs from a speech recognition system. It aggregates transcriptions by voting on words at each position across noisy samples, selecting the most frequent word to produce a final, more robust transcription, reducing errors caused by noise or adversarial attacks.

2) **MP3 compression:** MP3 compression, a lossy audio encoding scheme based on psychoacoustic models, serves as an effective defense against Audio Adversarial Examples (AAEs) in Automatic Speech Recognition (ASR) systems [14]. By discarding audio information below human hearing thresholds, MP3 compression can reduce adversarial noise (AN) embedded in AAEs, which is often imperceptible to humans but misleads ASR systems into incorrect transcriptions [14].
3) **Spectral gating (SG):** Spectral gating is a defense strategy that applies noise reduction to audio signals to remove adversarial perturbations [13]. The spectral gating defense works follows:
   a) **Voice Activity Detection**: Apply VAD to generate a boolean mask identifying speech segments.
   b) **Noise Profile Extraction**: Extract non-speech segments based on the VAD mask. If no non-speech segments detected, use portions from the audio's beginning and end (0.5 seconds each).
   c) **Noise Reduction**: Perform spectral subtraction, configured with parameters for stationary noise, proportion of noise to remove, and frequency/time smoothing.
4) **Quantization (Q):** Input quantization reduces the precision of audio signals by mapping continuous values to discrete levels, effectively attenuating adversarial perturbations [15]. This process can disrupt adversarial perturbations while preserving speech intelligibility.

2) Comprehensive Pipeline Evaluation:

Table VI summarizes the performance of various pipeline configurations for defending against PGD adversarial attack with $\epsilon = 0.1$ on the target model, evaluated using Character Error Rate (CER), Word Error Rate (WER), change in CER relative to the baseline (ΔCER), and average processing time per sample.

TABLE VI
COMPREHENSIVE PIPELINE EVALUATION SUMMARY

Pipeline Configuration	CER (%)	WER (%)	ΔCER (%)	ΔWER (%)	Clean Data WER (%)	Avg Time (s)
Baseline (No Defense)	37.90	61.98	0.00	0.00	2.83	0.0663
SG	2.84	7.01	35.06	54.97	5.67	0.2268
MP3	7.08	14.94	30.82	47.04	4.49	0.4710
Q	14.70	29.96	23.20	32.02	4.28	0.0547
RS	11.46	22.95	26.44	39.03	3.21	1.1155
SG → MP3	2.79	6.95	35.11	55.03	7.06	0.5753
SG → Q	3.35	8.50	34.55	53.48	6.47	0.1425
MP3 → Q	7.53	17.43	30.37	44.55	4.01	0.4888
SG → MP3 → Q	3.52	8.66	34.38	53.32	7.54	0.5816
SG → RS	16.87	27.27	21.03	34.71	24.65	1.1237
MP3 → RS	5.28	11.12	32.62	50.86	5.03	1.4871
Q → RS	6.56	14.22	31.34	47.76	4.71	1.0738
SG → MP3 → RS	20.20	30.59	17.70	31.39	28.98	1.5576
SG → Q → RS	17.25	27.22	20.65	34.76	24.81	1.1653
MP3 → Q → RS	5.11	11.12	32.79	50.86	5.29	1.4704
SG → MP3 → Q → RS	19.76	30.11	18.14	31.87	29.09	1.5762

TABLE VII
COMPARISON OF WER/CER REDUCTION: THIS WORK VS. ORIGINAL PAPERS

Defense	Metric	Reduction (This Work)	Reduction (Original)
RS	WER	39.03%	∼30–40% [17]
MP3	CER	30.82%	18–21% [14]
SG	WER	54.97%	Not reported WER but 50–60% in speaker ID (strong ASR) [13]
Q	WER	32.02%	10–30% [15]

3) Proposed Adaptive Defense Pipeline: The adaptive defense pipeline combines lightweight detection with robust mitigation:

1) **Detection Stage:** 8-bit quantization applied to input audio
2) **Decision Logic:** Classifies as adversarial if transcription dissimilarity (WER) larger than 12%
3) **Mitigation Stage:** For detected attacks, applies Spectral Gating (SG) + MP3 compression (64kbps)

Evaluation on 100 samples (50 adversarial from all 4 tested attacks, 50 clean):

TABLE VIII
ADAPTIVE DEFENSE PERFORMANCE

Metric	Value
Accuracy	0.990

TABLE IX
CONFUSION MATRIX

	Predicted Adversarial	Predicted Clean
Actual Adversarial	50 (TP)	0 (FN)
Actual Clean	1 (FP)	49 (TN)

Timing Analysis:

- **Average total time/sample:** 0.375s
- **Breakdown:**
  - Original transcription: 0.044s (11.7%)
  - Quantization + transcription: 0.045s (12.1%)
  - SG→MP3 defense: 0.534s

Adversarial Samples (50):

- **Detection rate:** 100.0% (50/50)
- **WER reduction:** 57.49% (65.91% → 8.42%)

Clean Samples (50):

- **Correct classification:** 98.0% (49/50)
- **WER impact:** Minimal degradation (4.21% → 3.93%)

The adaptive pipeline achieves near-perfect attack detection (100% recall) while maintaining low false positives (2%). The defense reduces adversarial WER by 57.49 percentage points with minimal computational overhead, demonstrating an optimal balance between robustness and efficiency. The single false positive case confirms the pipeline's conservative design, prioritizing security over minimal clean-sample degradation.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[2] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? a strong baseline for natural language attack on text classification and entailment," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8018–8025, 2020.

[3] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," in *Proceedings of the 2019 Network and Distributed System Security Symposium*, 2019.

[4] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Long Beach, California, USA: PMLR, 2019, pp. 5231–5240. [Online]. Available: https://proceedings.mlr.press/v97/qin19a.html

[5] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.15043

[6] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops*. IEEE, 2018, pp. 50–56.

[7] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020.

[8] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," 2022. [Online]. Available: https://arxiv.org/abs/2211.09527

[9] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Towards robust speech-to-text adversarial attack," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montréal, QC, Canada, 2021, École de Technologie Supérieure (ÉTS), Département de Génie Logiciel et des TI. [Online]. Available: https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/attachments/71497788/90c7c857-4259-4819-818e-cede46ed36cd/robust.pdf

[10] X. Gao, Z. Li, Y. Chen, C. Liu, and H. Li, "Transferable adversarial attacks against asr," *arXiv preprint arXiv:2411.09220*, 2024, institute for Infocomm Research, A*STAR, Singapore; National University of Singapore; University of California Riverside; Shenzhen Research Institute of Big Data, CUHK, Shenzhen. [Online]. Available: https://arxiv.org/abs/2411.09220

[11] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.

[12] L. Li, D. Song, and X. Qiu, "Text adversarial purification as defense against adversarial attacks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 338–350.

[13] P. O'Reilly, A. Bugler, K. Bhandari, M. Morrison, and B. Pardo, "Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models," in *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022, pp. 25 274–25 289. [Online]. Available: https://papers.neurips.cc/paper_files/paper/2022/file/c204d12afa0175285e5aac65188808b4-Paper-Conference.pdf

[14] I. Andronic, L. Kürzinger, E. R. C. Rosas, G. Rigoll, and B. U. Seeber, "Mp3 compression to diminish adversarial noise in end-to-end speech recognition," in *Speech and Computer – 22nd International Conference, SPECOM 2020, Proceedings*, A. Karpov and R. Potapova, Eds. St. Petersburg, Russia: Springer Science and Business Media Deutschland GmbH, October 2020, pp. 22–34, arXiv:2007.12892.

[15] Q. Li, Y. Meng, C. Tang, J. Jiang, and Z. Wang, "Investigating the impact of quantization on adversarial robustness," in *ICLR 2024 Workshop on Practical Machine Learning for Low-Resource Settings (PML4LRS)*, 2024, arXiv:2404.05639. [Online]. Available: https://arxiv.org/abs/2404.05639

[16] J. Ma and P. C. Loizou, "Snr loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, no. 3, pp. 340–354, 2011, available: https://ecs.utdallas.edu/loizou/speech/snrloss_spcom.pdf.

[17] P. Żelasko, S. Joshi, Y. Shao, J. Villalba, J. Trmal, N. Dehak, and S. Khudanpur, "Adversarial attacks and defenses for speech recognition systems," *IEEE*, 2021, all authors were with the Center of Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, USA. Manuscript received March 4, 2021. Supported by DARPA Award HR001119S0026-GARD-FP-052. E-mail: piotr.andrzej.zelasko@gmail.com.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1706.06083

[19] J. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 1310–1320.

[20] S. Lyu, S. Shaikh, F. Shpilevskiy, E. Shelhamer, and M. Lécuyer, "Adversarial robustness for multi-step defences," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024, university of British Columbia; Google DeepMind.

[21] R. Olivier and B. Raj, "Sequential randomized smoothing for adversarially robust speech recognition," *arXiv preprint arXiv:2112.03000*, 2022, language Technologies Institute, Carnegie Mellon University.