Vulnerability Analysis

Automatic Speech Recognition

Attack Type

Evasion

Date

2025-01-11

Author

AIShield

Job Id

gAAAAABn...qdL-KBOw==

Executive Summary:

Adversarial Efficacy (max): 68.75

Alert: Critical

Actions Needed: yes

We identified a Critical severity issue, indicating that an attacker can evade the original model with an attack efficacy of 68.75% . This vulnerability introduces substantial risks, as evasion attacks can expose sensitive data, leading to potential financial, legal, and reputational repercussions.
Note: ASR uses the Word Error Rate (WER) to measure accuracy. WER calculates the percentage of incorrectly transcribed words out of the total words in a transcription. A lower WER indicates superior ASR performance, capturing speech with minimal errors.

This assessment aligns with the following OWASP Machine Learning Security Top-10 Vulnerabilities(v0.3 Draft):

- ML01:2023 Input Manipulation Attack
- ML07:2023 Transfer Learning Attack
- ML09:2023 Output Integrity Attack

The proposed measures align with these CVE/CWE Entries:

- CWE-20: Improper Input Validation
- CWE-707: Improper Neutralization
- CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
- CWE-1288: Improper Validation of Consistency within Input

For more details about OWASP Vulnerabilities, please refer to Section 4.1 of the report.
For more details about CVE/CWE Entries, please refer to Section 4.2 of the report

Copyright Bosch Global Software Technologies Private Limited. All rights reserved, also regarding any disposal, exploration, reproduction, editing, distribution, as well as in the event of applications for industrial property rights.

1. Security:

1.1 Gradient based Testing: Strength Vs Efficacy on Original Model:

Evasion technique for ASR models involve evaluate the durability of models and their capacity to maintain accuracy in adverse conditions by introducing noise during speech recognition. The table display the relationship of SNR and WER for each epsilon of the ASR model, with base WER of 9.56% . The table provides a range of SNR values, which represent varying degrees of background noise. As the SNR decreases, the WER increases, indicating that the ASR model's accuracy deteriorates in noisier environments.

Epsilon	Avg WER (%)
0.006	46.83
0.008	43.38
0.01	41.01
0.02	55.35
0.04	77.27
0.06	100.29

Note: Table-1 shows the relation between epsilon and average WER.

This average WER is calculated based on WER scores that meet both of the following conditions: WER > threshold (i.e., 2*(original model WER)+1)% and SNR > 1dB = (2*9.56+1)% and SNR > 1dB = 20.12% and SNR >1dB.

Epsilon	Attack Efficacy (%)
0.006	26.5
0.008	33.17
0.01	38.33
0.02	67.92
0.04	68.75
0.06*	30.08

Note: Table-2 show the percentage of audio samples that are succesfully evaded by adding noise corresponding to each epsilon.

Note: * In Table 2, for certain epsilon values, the added perturbation significantly decreases the SNR, leading to a complete loss of information in the signal. Therefore, only samples with an SNR greater than 1 dB are considered when calculating attack efficacy.

Copyright Bosch Global Software Technologies Private Limited. All rights reserved, also regarding any disposal, exploration, reproduction, editing, distribution, as well as in the event of applications for industrial property rights.

2. Performance:

2.1 Inference Time of model:

Inference time of a model refers to the amount of time it takes for the model to generate predictions on new input data. It is the time taken by the model to process input data and produce output. inference time is influenced by factors such as the model's complexity, the input data's size, and the computational resources available for inference. It is an important metric to consider when deploying machine learning models in production environments where fast and efficient processing is necessary.

In the context of this scenario, Original Model Inference Time in ms is 1177.42.

More details on the hardware infrastructure, and inference time distributions will be populated in the upcoming reports.

2.2 Original Model Metric:

Word Error Rate (WER) on original data is crucial because it serves as a benchmark for the model's initial performance under ideal conditions. It provides a reference point for assessing improvements or degradation in accuracy when the model is subjected to various challenges, such as noisy environments or accent variations. By knowing the base WER, developers can measure the model's adaptability and identify areas for enhancement, ultimately ensuring better real-world usability and reliability.

Model	Number of Samples	Word Error Rate(%)
Original Model	1200	9.56

Copyright Bosch Global Software Technologies Private Limited. All rights reserved, also regarding any disposal, exploration, reproduction, editing, distribution, as well as in the event of applications for industrial property rights.

3. Mitigation Strategies:

3.1 Adversarial and Diverse Training Data (Training Phase):

Enhance the robustness of the ASR model against evasion attacks by ensuring the training data is both diverse and incorporates adversarial examples [1]. A diverse dataset includes various accents, languages, dialects, and speech patterns, enabling the model to better adapt to real-world variations.
Additionally, employing adversarial training techniques by integrating adversarial examples into the dataset further strengthens the model's resilience to evasion attacks. This combined approach ensures a more comprehensive and robust defense mechanism for the ASR system.

3.2 Adaptive Noise Reduction (Inference Phase):

Implement noise reduction techniques as a preprocessing step to filter out background noise from audio input. This can significantly enhance the model's ability to perform accurately in noisy environments and make it more resilient to adversarial attempts [2].

3.3 Language Model Integration:

Integrate Natural Language Processing (NLP) models with your ASR system to improve contextual understanding of spoken language. By considering the context, the ASR model can become more resistant to attacks that rely solely on manipulating individual words [3].

References

1. Ian J. Goodfellow and Jonathon Shlens and Christian Szegedy, "Explaining and Harnessing Adversarial Examples", 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings 2015

2. Olivier, Raphael, and Bhiksha Raj. "There is more than one kind of robustness: Fooling whisper with adversarial examples." arXiv preprint arXiv:2210.17316 (2022).

Copyright Bosch Global Software Technologies Private Limited. All rights reserved, also regarding any disposal, exploration, reproduction, editing, distribution, as well as in the event of applications for industrial property rights.

5. Appendix:

Appendix section offers supplementary data on our methodology.

5.1 OWASP Vulnerabilities(v0.3 Draft):

ML01:2023 Input Manipulation Attack
This weakness directly relates to adversarial attacks where an automated mechanism, such as machine learning, doesn't properly detect or handle inputs that have been modified to mislead the model.

ML07:2023 Transfer Learning Attack
If the machine learning model does not ensure that structured messages or data are well-formed, it can be susceptible to attacks like the Transfer Learning Attack.

ML09:2023 Output Integrity Attack
The description of the adversarial attack mentions the importance of input validation to detect and prevent adversarial attacks.

For an expanded understanding of the OWASP Top - 10 ML Vulnerabilities, please visit
https://owasp.org/www-project-machine-learning-security-top-10/

5.2 CVE(Common Vulnerabilities and Exposures)/CWE(Common Weakness Enumeration):

CWE-20: Improper Input Validation
A weakness where the product receives input or data but does not validate or incorrectly validates that the input has the properties required to process the data safely and correctly. This can lead to various security issues, such as altered control flow, arbitrary control of a resource, or arbitrary code execution.

CWE-707: Improper Neutralization
The product does not ensure or incorrectly ensures that structured messages or data are well-formed and that certain security properties are met before being read from an upstream component or sent to af downstream component.

CWE-1039: Automated Recognition Mechanism with Inaccurate Detection
This weakness occurs when a product uses an automated mechanism, to recognize complex data inputs, but it does not properly detect or handle inputs that have been modified or constructed such that it causes the mechanism to detect an incorrect concept.

CWE-1288: Improper Validation of Consistency within Input
The product receives a complex input with multiple elements or fields that must be consistent with each other, but it does not validate or incorrectly validates that the input is actually consistent.This can allow attackers to trigger unexpected errors, cause incorrect actions to take place, or exploit latent vulnerabilities.

For comprehensive data related to CVE/CWE list,please visit
https://cwe.mitre.org/

The digest value of the pdf file is 49295410ff761ccc1155fce2426fc660d7916b9ba7256e81f79b2e432e2393c7

Copyright Bosch Global Software Technologies Private Limited. All rights reserved, also regarding any disposal, exploration, reproduction, editing, distribution, as well as in the event of applications for industrial property rights.