Adversarial Attacks and Defenses for Automatic Speech Recognition Systems

Daniyal Parveez, Computer Science Department
Advisors: Dr. Zesheng Chen, Dr. Jack Li, and Dr. Chao Chen

PURDUE UNIVERSITY FORT WAYNE

Introduction

State-of-the-art **Automatic Speech Recognition (ASR)** systems, powered by AI technologies, convert spoken language into text (commonly referred to as transcription). Like other AI-driven systems, they are susceptible to various security threats. One particularly critical threat that has garnered significant attention in recent years is **adversarial attacks**.

Adversarial attacks involve subtly modifying inputs to deceive AI models. In many cases, a benign input that the model would typically process correctly can be carefully altered in a way that remains imperceptible to human observers yet causes the model to produce erroneous or even harmful outputs. The specific mechanics and objectives of adversarial attacks vary depending on the application.

For ASR systems, adversarial attacks typically involve embedding **imperceptible** perturbations into an audio signal. While these perturbations are nearly undetectable to human listeners, they can cause the model to generate a drastically different transcription from the intended speech.

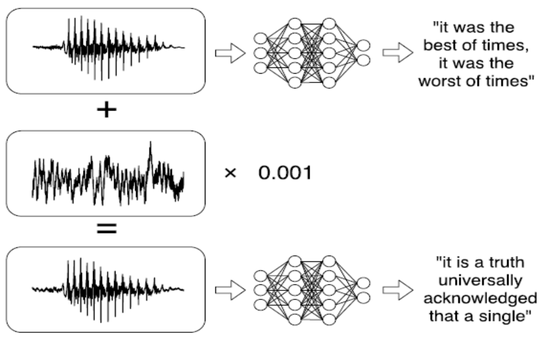

Figure 1: Example of an Adversarial Attack

Attack Methodology

We evaluate two popular ASR models—Wav2Vec2 and OpenAI Whisper (Base English variant). On the attack front, we consider three different adversarial attacks: CW [1], PGD [2], and AdvReverb [3]. Using a dataset of 113 clean audio samples from LibriSpeech, we apply each attack individually to both models. To assess the effectiveness of these attacks, we measure attack success rates & perception metrics.

Attack	Model	ASR
CW	Wav2Vec2	99.8736
	Whisper	100
PGD	Wav2Vec2	85.4614
	Whisper	99.115
AdvReverb	Wav2Vec2	96.1947
	Whisper	100

Table 1: Attack Success Rates

Model / Attack	MCD Mean	PESQ Mean	SNR Mean	L2 mean	L_inf mean
Wav2Vec2 / CW	8.221	1.639	18.881	1.783	0.046
Whisper / CW	6.59	2.161	23.839	1.2	0.027
Wav2Vec2/ PGD	6.659	2.066	24.214	0.833	0.005
Whisper / PGD	6.505	2.113	24.659	0.79	0.005
Wav2Vec2 / AdvReverb	5.502	1.862	15.539	3.483	0.132
Whisper / AdvReverb	4.303	2.13	18.686	2.412	0.136

Table 2: Perception Metrics

Defense Methodology

Defending against adversarial attacks is a critical area of research. In our study, we explore various detection methods to distinguish between benign and adversarial audio samples. Specifically, we evaluate the performance of four different defense systems:

- **MEH-FEST** (Minimum Energy in High FrEquencies for Short Time), which measures the energy in high frequencies of an audio when the speech is absent [4].
- **LFCE** (Low Frequency Cumulative Energy), which calculates the total energy in low frequencies of an audio.
- **SNA** (Simple Noise Adding), which adds random Gaussian noise to an input audio.
- **SNF** (Simple Noise Flooding), which finds the minimum noise value that can change the output of an ASR system.

Attack	Detection Method	RoC AUC Scores	
		Wav2Vec2	Whisper
CW	LFCE	0.540	0.516
	MEH-FEST	1	0.88
	SNA	0.998	0.993
	SNF	0.989	0.899
PGD	LFCE	0.506	0.502
	MEH-FEST	1	0.992
	SNA	1	1
	SNF	0.990	0.965
AdvReverb	LFCE	0.962	0.961
	MEH-FEST	0.480	0.534
	SNA	1	1
	SNF	0.990	0.933

Table 3: RoC AUC Scores for different detection methods

Conclusion

Through comprehensive experiments, we found that

- OpenAI Whisper is more vulnerable to adversarial attacks than Wav2Vec2.
- MEH-FEST effectively mitigates CW and PGD attacks but is ineffective against AdvReverb. Conversely, LFCE can defend against AdvReverb but fails to counter CW and PGD attacks.
- Both SNA and SNF provide defense against all three attacks.

References

[1] Nicholas Carlini and David A. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," CoRR, 2018.
[2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations, 2018.
[3] M. Chen, L. Lu, J. Yu, Z. Ba, F. Lin and K. Ren, "AdvReverb: Rethinking the Stealthiness of Audio Adversarial Examples to Human Perception," in *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1948-1962, 2024.
[4] Z. Chen, "On the Detection of Adaptive Adversarial Attacks in Speaker Verification Systems," in *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 16271-16283, Sept. 2023.

Acknowledgement

This research was funded by 2024-2025 PFW CS Graduate Research Assistantship and 2024-2025 ETCS Faculty Research Seed Grant.