Perception-Aware Adversarial Attacks on Speech Audio

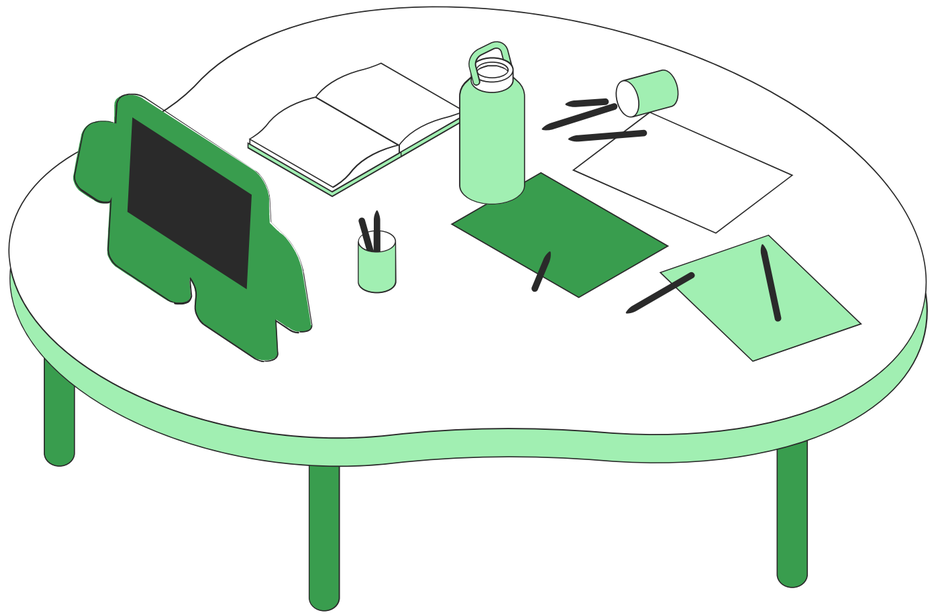
Student: Cat Lewin
Mentor: Dr. Rui Duan
Monday, June 9th 2025

Problem Statement

Goal: This project develops a framework that uses AI-based models of human auditory perception to generate subtle modifications to speech audio—largely unnoticed by listeners—that can evade automated copyright detection systems.

Purpose: To investigate and identify potential vulnerabilities in these systems in order to inform the design of more robust and resilient detection technologies.

Related Work



Duan et al., 2024: Explores black-box adversarial attacks on speaker recognition using surrogate models trained on short voice samples.

Duan et al., 2022: Introduces a perception-aware framework for generating adversarial music examples guided by human perceptual ratings.

Saadatpanah et al., 2019: Demonstrates that automated copyright systems like YouTube’s Content ID can be deceived using minor audio perturbations.

Carlini & Wagner, 2018: Proposes targeted attacks on speech-to-text systems that produce specific transcriptions with high perceptual similarity.

AI Methods Being Used

Adversarial Machine Learning

Crafting small, strategic changes to audio signals using gradient-based methods.

Perceptual Modeling

Applying AI techniques trained on human feedback to estimate how noticeable perturbations are to listeners.

Surrogate Modeling

Exploring the use of approximate models that replicate the behavior of detection systems in a black-box setting.

Next Steps & Timeline

Week 3-4

Implement perceptual loss functions and baseline attacks.

Week 5

Apply attacks to speech audio and analyze the effectiveness against simulated detection systems.

Week 6

Begin evaluating audio intelligibility and prepare preliminary findings.

Current Challenges / Questions

How can we effectively model human perception in the optimization process?

What evaluation metrics best balance audio intelligibility with detection evasion?

How accurately can we approximate the behavior of real-world detection systems like Content ID?

How to appropriately & ethically test detection evasion of copyrighted speech?

References

Rui Duan, Yao Liu, Zhe Qu, Leah Ding, and Zhou Lu. 2024. Parrot-Trained Adversarial Examples: Pushing the Practicality of Black-Box Audio Attacks against Speaker Recognition Models.

Rui Duan, Leah Ding, Zhe Qu, Yao Liu, Shangqing Zhao, and Zhou Lu. 2022. Perception-Aware Attack: Creating Adversarial Music via Reverse-Engineering Human Perception.

Nicholas Carlini and David Wagner. 2018. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text.

Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. 2019. Adversarial Attacks on Copyright Detection Systems.