Week 2 Report

Project Information
Project Type: Individual
Student Name: Cat Lewin
Mentor Name: Dr. Rui Duan
Research Title: Perception-Aware Adversarial Attacks on Speech Audio

Problem Statement: This project develops a perception-aware adversarial attack framework for speech audio that uses AI-driven models of human hearing to create subtle changes—largely unnoticed by listeners—that can evade automated copyright detection systems, with the goal of identifying and studying weaknesses in these systems for research purposes.

Literature Review

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

- **Summary:** This paper came up with a way to attack speaker recognition systems without needing to know anything about how the system works internally (black-box). They used a technique that mimics a person's voice ("parrot-trained") to fool the system.
- **Contributions:** They introduced the idea of "parrot-trained" surrogate models, which can be built from just a short audio clip of someone's voice. These models make it possible to create adversarial audio that even works when played out loud into a microphone—no direct access to the target system needed.
- **Limitations:** This work focuses on attacking speaker recognition systems, so it doesn't really touch on general audio content or how people perceive the audio changes.
- **Expand/Improve:** I'm taking their idea further by applying it to broader detection systems like copyright enforcement, and adding models that better reflect how humans hear and interpret speech.

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

- **Summary:** This paper researched how to make audio that tricks detection systems but still sounds normal to people. They trained a model to understand what kinds of distortions humans can't easily notice, and used that to guide their attacks.
- **Contributions:** They built a regression model trained on real human feedback to measure how noticeable audio changes are. This model was then used to generate music adversarial examples that sneak past detection algorithms while sounding almost identical to the original.
- **Limitations:** The study did a great job with music, but it didn't explore how these techniques might work for speech, which is different in how it's processed and understood by listeners.
- **Expand/Improve:** I'm adapting their perception-aware framework to focus on spoken content, looking specifically at how intelligible speech remains when adversarial noise is added.

Title: Audio Adversarial Examples: Targeted Attacks on Speech-to-Text
Authors: Nicholas Carlini, David Wagner
Publication Year: 2018

- **Summary:** This paper showed that it's possible to subtly change an audio clip so that an automatic speech recognition (ASR) system hears something completely different—like hearing "okay google, search evil.com" instead of "play some music"—without humans noticing anything weird.
- **Contributions:** They created a white-box attack method that could reliably get ASR systems to output specific phrases, even with very small audio changes. It was one of the first to show how vulnerable these systems are to targeted manipulation.
- **Limitations:**Their approach is highly effective for speech-to-text, but it doesn't quantify how perceptible the sound disturbances are, and it doesn't look at other types of detection systems. Additionally, this is a white-box attack specific to Mozilla's DeepSpeech, and is not a particularly feasible attack in real time.
- **Expand/Improve:** I'm using models that try to align better with human perception and shifting the focus from changing transcriptions to evading systems like Content ID.

Title: Adversarial attacks on Copyright Detection Systems
Authors: Parsa Saadatpanah, Ali Shafahi, Tom Goldstein
Publication Year: 2019

- **Summary:** This study looked at how to evade copyright detection systems, like YouTube's Content ID, with slight changes to the audio. They found that you can often bypass detection without making the clip sound any different to a human.

- **Contributions:** They built and tested attacks on both open and closed systems, showing that even very small perturbations could evade detection. They also explored whether attacks that worked on one system could transfer to another.
- **Limitations:** The paper shows that these systems can be tricked, but mostly with music and without formally measuring how noticeable the changes are to listeners.
- **Expand/Improve:** I'm testing similar attacks on speech content, and working to make sure the changes are less noticeable using perceptual models.