Project Proposal:

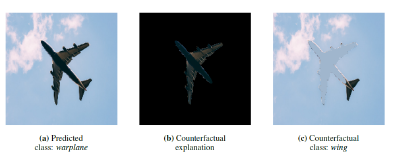
Trustworthy Machine Learning

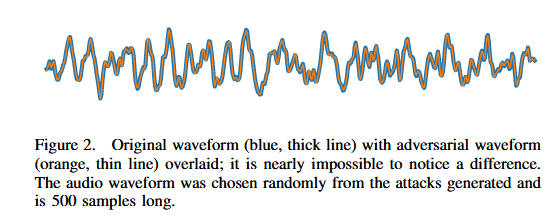
Andrew Nguyen

The project seeks to explore the field of adversarial attacks on automatic speech recognition systems(ASR). An ASR has the goal to “convert raw human audio to text” through two main steps: signal processing step to extract features that are fed into the machine learning step to infer a speech recognition model [1]. According to Abdullah’s overview [2], there are three main attacks on these systems. Firstly, optimization attacks query or use information on weights to computer gradients that can craft an input to flip the decision boundary. Secondly, signal processing attacks effect the feature engineering portion by crafting designing an audio sample that leverages differences between human listening and feature extractions. Thirdly, miscellaneous attack such as exploiting hardware limitations, replaying sound bite of individual or homophones in language. Due to prominence of ASR in popular applications like Alexa and Siri, it is to provide users with trust for the systems. According to [4], “much work is yet to be done [in xAI and audio]”. Work so far has only focused on generating a visual agent to understand voice classification that gained more trust compared with simply text or voice-only interactions.

The project proposes to apply evasive attacks in signal processing and/or indirect optimization attacks on a pretrained ASR model to generate successful attack audio samples to then develop counterfactual explanations for the complex ASR system. These attacks seek to differentiate human-interpretation and ASR-interpretation of audio.

The network model(popular model for ASR) from Mozilla DeepSpeech [3] was chosen due to time and resource constraints that can take many days to train. Then the project seeks to craft a signal processing attack through some of the four methodologies outlined in [1]: Time-Domain inversion, random phase generation, high frequency addition, time scaling. on this pretrained model as it is known to be very efficient (“require less than 15 queries and a few seconds to generate attack with black-box access”) that attacks the preprocessing and feature pipeline of the system. In addition, indirect optimization attacks may be explored that require thousands of queries to the black-box system to generate an adversarial audio sample that attacks the neural network itself. Carlini [5] showed an adaption of optimization of adversarial examples through additive perturbations from image domain to audio domain that worked with great success that are nearly imperceptible to the human ear, but it took roughly one hour to generate one adversarial example on a singular GPU which limits this approach for my project. Finally, the project seeks to provide expand the presence of explainable AI(xAI) in this field. Much work in xAI exists in the computer vision domain with one such idea called counterfactual explanations. The example utilized in [4] on image classification had three images(shown below) a) Warplane in the sky, b) Warplane without a tail wing on black background, and c) the wing part from the warplane isolated. Image (a) was predicted as a warplane whereas removing (b) from (a) that produced (c) causes the system to predict the tail portion as a wing rather than plane. This means (b) is a “critically minimum portion” to get classified as a warplane. The input to computer vision inference systems include an array of pixel values. ASR systems take an array of amplitude from an audio reading but **first preprocessed and features extracted** entering the inference system **[2]**. This additional step prevents a direct translation from image domain to audio domain of counterfactuals. The project also seeks to adapt this work of counterfactual explanation to automatic speech recognition attacks by finding the critical feature in feature extraction that allows the attack to be successful.





References:

[1] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. Butler, and J. Wilson, “Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems,” Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS), 2019.

[2] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, “ SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems”, 2020.

[3] “Deep Speech 0.9.3,” Last accessed in 2022, available at <https://github.com/mozilla/DeepSpeech/releases/tag/v0.9.3>

[4] P. Gohel, P. Singh, M. Mohanty, “ Explainable AI: current status and future directions”, 2021

[5] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018, pp. 1–7.