Rajat Sethi

CPSC 8580

**Paper Summary: Devil’s Whisper**

In Automated Speech Recognition (ASR) and Intelligent Voice Control (IVC) systems, sound serves as the primary (possible sole) method of communication between the device and its environment. With these devices, humans can issue commands (play a song, navigate to my house, open the door) using their voice instead of typing it out. As these systems are controlled by large companies, their language processing models are kept a secret to the public. As such, testing them becomes a “black-box” challenge, in which only the input and output can be studied. This paper discusses their black-box approach to generating Adversarial Examples (AE) that can enter malicious commands to an ASR/IVC.

The key novelty in the Devil’s Whisper project is their simplified model. Since the white-box model is unavailable, the research team created their own simple model by taking the results from their black-box testing. This way, their attacks can be generalized to work on multiple ASR/IVC systems. Their other novelty was their usage of successful attacks from other papers as their testing method. For example, Devil’s Whisper uses the same attack as “CommanderSong”, another research paper that placed quiet AEs in popular songs.

The obvious limitation of this paper is the black-box nature of the devices they’re exploiting. While analyzing the devices through black-box testing is the whole purpose of this paper, they cannot check with the actual models since they’re guarded under company secrets. As such, it is quite possible that the team missed some malicious commands or some form of attack. Having that validation from the manufacturers themselves could open several doors for future attacks/defenses.

While I don’t have any ideas for the paper itself, I would like to take some of the methodologies from this paper for my own research. My project for this class is to break into the Google Nest Mini in some capacity. I originally tried to use high-pitch frequencies to mask commands, but to no avail. However, this paper showed me that high-pitch attacks have been done in the past and could still be possible. I will have to look at adversarial examples referenced in this paper and see if I can use any of their techniques.