Acoustic Side Channel Attack
Samuel Levesque, Ryan Martineau, Ryan Reid, Prof. Aaron Carpenter
Department of Electrical and Computer Engineering, Wentworth Institute of Technology

Abstract
This presentation describes an acoustic keyboard side channel attack on a variety of smartphones using acoustic and timing analysis in conjunction with a natural language model to be able to accurately predict what a user is typing. This white-hat hack has been explored previously on physical keyboards, but soft smartphone keyboards pose several new challenges. This article describes how sounds emitted from smartphone speakers can be captured and analyzed with digital signal processing (DSP), fed into a natural language model, and result in an accurate prediction of what the user is typing on their smartphone.

Digital Signal Processing
The Digital Signal Processing System analyzes the recorded data by graphing it as a spectrogram. A spectrogram is ideal for this scenario because it is able to graph frequency, time, and amplitude. Frequency and amplitude help to determine what character is being pressed. It can determine if the character was a letter, space, backspace, or a send. When a word is being typed, the DSP is able to determine how many letters are contained in the word by counting the number of peaks in-between spaces. When a space occurs, the DSP is able to tell a word has ended. When a backspace occurs, the letter count for the current word is decremented. When the Send button is hit, it will break out of the analyzing the data and send the information to the natural language model.

Introduction
Users touch their smartphones an average of 2,617 times per day to store, send, and receive data. [1] People feel confident using these devices to interact with the world and share their most private and personal information, whether it be to the masses on apps like Facebook and Twitter or to a single person with iMessage or WhatsApp. There is an expectation of privacy with these devices. Yet, many hackers have found ways to gain access to protected data without the user’s knowledge by exposing certain physical vulnerabilities of the phone. One exploit that has been explored on desktop keyboards is an attack that reconstructs the user’s message by analyzing the acoustics of keyboard typing and the timing between these sounds. This vulnerability manifests itself differently on mobile phones because the sounds are emitted directly from the speaker of the phone itself and do not usually differ between characters.

This acoustic side channel attack against smartphones utilizes a dual microphone system to eliminate background noise to capture acoustic emanations from smartphone keyboards. The recordings are analyzed by a DSP system (MATLAB) using a spectrogram to graph the frequencies, amplitude, and time on a single graph. From here, the DSP is able to tell what type of character is being pressed using the data from the spectrogram. The DSP is then able to tell the natural language model how many words are in the phrase and how many letters are contained in every word. The natural language model then finds words that fit the character length given by the DSP and matches it with words that are often used together. A basic diagram of this system can be seen below:

Recording
In order to capture these acoustic emanations from a smartphone, a person typing on a smartphone keyboard needs to be recorded. To accomplish this task, the institute’s radio station studio was utilized to get a relatively noiseless environment. To record these sounds, an iPhone SE was placed directly in front of the microphone array of a laptop, providing clean recordings with relatively low noise. This also kept the recordings of these sound relatively consistent. Next, a student typed a short message which was recorded using a free audio recording program called Audacity.

Natural Language Model
Using the open-source web scraping tool twiterscraper, text from discussions on Twitter are collected and used to construct a small corpus for the language model. Once the corpus is constructed, the text is preprocessed, removing twitter handles, hashtags, links, whitespace, and other punctuation from the corpus. Using several basic language modeling techniques to analyze word and word sequence probabilities, such as Markov chaining, text is generated from the available corpus. The DSP analysis of the sounds of a user typing on their keyboard provides additional data to the language model which narrows its search to words and word sequences of a specific length. With additional timing data/analysis, the construction of a more permanent database solution for the corpus of text, and the use of machine learning all in progress, the expectation is that model will improve significantly over the next several iterations.

\[
P(X_1, \ldots, X_n) = \sum_{i=1}^{n} P(X_i | X_{i-1})\]

Figure 6 – Equation demonstrating the theory behind Markov chaining. Essentially, it is the probability of a word occurring given a word or set of words that occurred previously. [2]

Future Work
• Introduce timing analysis – time between each key press could reveal more information about which letter was pressed
• Begin testing of a more complex recording system which can handle various real-world scenarios
• Design a more permanent method of storing the text corpus
• With timing data, incorporate machine learning to create a more complex, but accurate language model
• Perfect recording method, implementing a dual microphone system

References