Demo: Accelerometer-based Smartphone Eavesdropping

Zhongjie Ba*, Tianhang Zheng†, Zhan Qin*, Hanlin Yu*, Liu Liu*, Baochun Li†, Xue Liu‡, Kui Ren*
*Zhejiang University, †University of Toronto, ‡McGill University
{zhongjieba,qinzhan,yuhanlin,evence,kuiren}@zju.edu.cn
th.zheng@mail.utoronto.ca,bli@ece.toronto.edu,xueliu@cs.mcgill.ca

ABSTRACT

In this demonstration, we show that audio signals emitted by a smartphone speaker can be captured by the accelerometer on the same smartphone, and accelerometers on recently released smartphones can cover most of the fundamental frequency band of adult speech. Based on these pivotal observations, we present AccelEve, a new side channel attack that allows smartphone applications to eavesdrop on the smartphone speaker without the requirement of sensitive system permissions. Through analyzing the accelerometer measurements of a smartphone, AccelEve is able to: 1) recognize the speech information (text) carried by the acceleration signal; 2) reconstruct the audio signal played by the smartphone speaker. This demo will present experimental validations for our observations and the proposed system.

CCS CONCEPTS

• **Security and privacy → Side-channel analysis and countermeasures**.

KEYWORDS

motion sensor, smartphone eavesdropping

ACM Reference Format:
Zhongjie Ba*, Tianhang Zheng†, Zhan Qin*, Hanlin Yu*, Liu Liu*, Baochun Li†, Xue Liu‡, Kui Ren*. 2020. Demo: Accelerometer-based Smartphone Eavesdropping. In *MobiCom 2020 (MobiCom '20), September 21–25, 2020, London, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3372224.3417323

1 INTRODUCTION

Sound is a vibration that can be transmitted through gas, liquid or solid. Motion sensors can response to external vibrations and thus have been exploited to eavesdrop on audio devices. In USENIX Security'14, Michalevsky *et al.* [4] proposed to use a smartphone gyroscope to eavesdrop on a loudspeaker placed on the same table. They were the first to show that sound signals traveling through a solid surface can affect smartphone gyroscopes. Later, Zhang *et al.* [5] studied the possibility of utilizing a smartphone accelerometer to capture voice signals traveling through the air. However, experimental results [1] show that airborne speech signals might not have sufficient power for generating and transmitting vibrations to motion sensors. In S&P'18, Anand *et al.* [1] systematically studied the threat of motion sensors to speech privacy and reported that only speech signals propagating through solid medium can have a noticeable impact on the motion sensors of a smartphone. *The only feasible setup reported previously is to use a gyroscope to eavesdrop on an independent loudspeaker, in which the adversary can only get speech signals with severe attenuation and distortion.*

In this work, we revisit the threat of motion sensors to speech privacy and report two fundamental observations. First, all previous works failed to cover the most adverse setup where the motion sensor and the target speaker are on the same smartphone. In this setup, speech signals emitted by the speaker always produce strong responses in motion sensor measurements due to the shared motherboard. Second, motion sensors on recently released smartphones can cover most of the fundamental frequency band of adult speech (85-255Hz). Contrary to the widely-held belief that the sampling rate of the motion sensor on a smartphone cannot exceed 200Hz, we observed sampling rates up to 500Hz in recently released smartphones. This indicates that motion sensors on smartphones are able to capture a significant amount of human speech.

On top of the above observations, we propose AccelEve, a new side channel attack that enables smartphone applications to eavesdrop on the smartphone speaker through analyzing the measurements of the built-in accelerometer. In this attack, the adversary is a spy app whose objective is to capture the speech information played by the smartphone speaker. It can be disguised as any kind of applications since the accelerometer is classified as a zero-permission sensor in most mobile systems. During the attack, the spy app continuously collects the accelerometer's readings in the background and extracts speech information using two deep learning models: 1) speech recognition: this model mainly learns the semantic information captured by the accelerometer. It can recognize pre-trained numbers, letters, and hot words from acceleration signals. 2) speech reconstruction: this model mainly learns the mapping between the acceleration signal and the audio signal. It can convert acceleration signals into audio signals and enable the adversary to double-check the results of the recognition model with human ears.

2 THE DESIGN OF ACCELEVE

The workflow of the AccelEve is shown in Figure 1. The main intent of this system is to extract speech information from accelerometer measurements. For conventional speech recognition tasks with audio signals, the raw signal is normally transformed to its mel-frequency cepstrum (MFC) representation in order to discard redundant and superfluous information in the high-frequency band. For acceleration signals, however, the MFC-based approaches are not applicable due to the low sampling rates of accelerometers.

MobiCom '20, September 21–25, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7085-1/20/09.
https://doi.org/10.1145/3372224.3417323

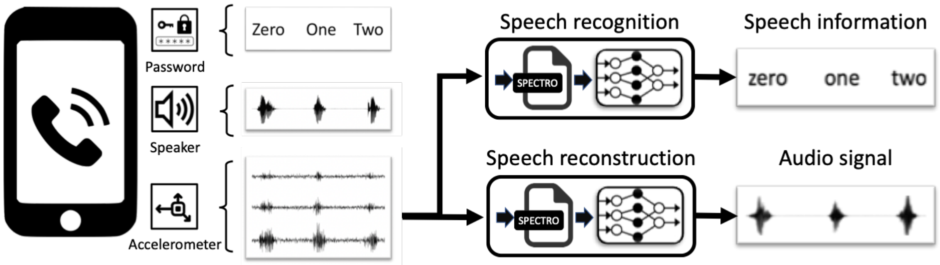

Figure 1: The workflow of accelerometer-based smartphone eavesdropping

Therefore, we employ the spectrogram representation to preserve the speech information to the utmost extent and use deep learning techniques for recognition and reconstruction. Our proposed system is mainly composed of three modules: preprocessing, recognition, and reconstruction.

Preprocessing: This module converts raw acceleration signals into single word spectrogram-images as input to deep neural networks. In particular, it converts unequally-spaced raw samples into a uniformly sampled signal with timestamp-based interpolation, eliminates significant distortions, e.g., gravity and human activities, via high-pass filtering, cuts long acceleration signals into single word segments, and converts the three axes of each segment into spectrograms through Short-Time Fourier Transform (STFT). The output of this module is a spectrogram-image whose red, green, and blue channels correspond to the x-axis, y-axis, and z-axis of the acceleration signal, respectively.

Speech Recognition: In this module, we adopt DenseNet [3] as the base network to recognize the spectrogram-images processed by the prepocessing module. Direct connections between any two layers in DenseNet extremely encourage the reuse of features and improve the flow of information. Therefore, using DenseNet, the recognition module can achieve the best accuracy with fewer parameters compared to other deep networks like VGG and ResNet. Moreover, we set the dropout rate as 0.3 in DenseNet and also apply an adaptive optimization using cross-entropy loss with weight decay to enhance generalizability.

Speech Reconstruction: This module targets to convert an acceleration signal into an audio (speech) signal with enhanced sampling rates (1500Hz). The intuition behind is that most of the speech information in the high frequency band is composed of the harmonics of the fundamental frequency. In this module, we first reconstruct speech spectrograms from the acceleration spectrograms by a reconstruction network composed of an encoder, residual blocks, and a decoder. Robust L_1 loss is applied to train the reconstruction network, with weight decay to enhance generalizability. Speech signals are then recovered from the reconstructed speech spectrograms by the Griffin-Lim algorithm [2].

3 THE DEMONSTRATION

For our demo, we first present experimental validations to demonstrate our pivotal observations and the feasibility of the proposed attack. After that, we conduct an end-to-end attack with a third party Android application AccDataRec in order to illustrate the effectiveness of the three modules of AccelEve. Finally, we briefly go through some evaluation results on the benchmark datasets, which demonstrate the effectiveness and high accuracy of the proposed attack under various settings.

3.1 Feasibility Study

This part demonstrates the capability of accelerometers in capturing audio signals emitted by the smartphone speaker. The feasibility study is conducted from three aspects:

Significance: stimulate the accelerometer of a smartphone with audio signals played at different volume levels.

Effectiveness: test smartphones released in different years and report the actual sampling rates of their accelerometers.

Robustness: evaluate and analyze the impact of acoustic noise and human activities on accelerometer measurements.

3.2 End-to-end Case Study

This part demonstrates the entire process of the attack, from collecting the accelerometer measurements to extracting the speech information. Specifically, we first play a series of speech signals on a Samsung S8 and collect accelerometer measurements through a third party Android application AccDataRec running in the background. Then, we explain the problems of raw accelerometer measurements and use the pre-processing module to address them. After obtained the spectrogram-images from the pre-processing module, we use a recognition network to search and identify pre-trained hot words. Finally, we feed the spectrogram-images into the reconstruction module and use the GL algorithm to recover the audio signal. The recovered audio shows the effectiveness of the attack, which will be played in our demo.

REFERENCES

[1] S Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1000–1017.

[2] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 4700–4708.

[4] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: recognizing speech from gyroscope signals. In *Proceedings of the 23rd USENIX conference on Security Symposium*. USENIX Association, 1053–1067.

[5] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 301–315.