REPURPOSING OPTICAL MICE FOR ACOUSTIC EAVESDROPPING

ZHIMIN MEI

MPhil

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Computing

Repurposing Optical Mice for Acoustic Eavesdropping

Zhimin MEI

A thesis submitted in partial fulfillment of the requirements for
the degree of Master of Philosophy
September 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Zhimin MEI

Abstract

Acoustic eavesdropping presents a longstanding challenge in the realm of personal information security and privacy preservation. In this work, we introduce a novel eavesdropping method called JerryAttack, which repurposes an optical mouse as a covert eavesdropping device. Specifically, we transform the mouse's integrated low-resolution but high-frame-rate image sensor into a high-speed camera for visual vibrometry, capable of capturing acoustic vibrations from nearby loudspeakers. Our contributions are threefold: First, we utilize the 'pixel grabber' register as a backdoor to extract the pixel stream from the image sensor. Second, we establish an acoustic-optical side channel that enables effective acoustic eavesdropping. Third, we thoroughly explore two attack scenarios: voice profiling and speech reconstruction. Our findings reveal that the sound recovered through our side channel achieves a mean SNR of 7.3 dB, comparable to standard microphone recordings in noisy environments like cafes. Additionally, when combined with a classification neural network, JerryAttack identifies individuals with an overall accuracy of 83.27% across six languages. Moreover, when cooperated with joint channel information, JerryAttack consistently achieves good intelligibility, with a median STOI score exceeding 0.7 in reconstructed results.

Publications Arising from the Thesis

1. Zhimin Mei, Donghui Dai, Jingyu Tong, Zheng Gong, and Lei Yang "Repurposing Optical Mice for Acoustic Eavesdropping", in Proc. of **IEEE INFOCOM**, 2025.

2. Sicong Liao, Jingyu Tong, Zhimin Mei, Donghui Dai, Yuanhao Feng, Qiongzheng Lin, and Lei Yang "POSTER: A One-size-fits-all Solution for Cross-Technology Communication via Transformer", in Proc. of **ACM MobiSys**, 2024.

Acknowledgments

It is with profound pleasure that I acknowledge the significant contributions of those who have played pivotal roles in the successful completion of my MPhil dissertation. My deepest gratitude extends to a multitude of individuals whose diverse contributions have made this journey possible.

Firstly, I wish to express my deepest appreciation to my supervisor, Prof. Lei Yang. His invaluable guidance throughout my academic journey has been of paramount importance. His innovative academic insights have not only enriched my learning but also shaped my scholarly trajectory. His steadfast commitment to scholarship has continually served as an inspiration to me.

Special mention must be made of Prof. Xiaoming Wu and Prof. Kai Zhou, whose insightful advice and constructive feedback during my confirmation have significantly enhanced the quality of the dissertation. Concurrently, I wish to express my gratitude to the BoE chair, Prof. Yuanqing Zheng, and external examiners, Prof. Xiaoyu Ji and Prof. Zhetao Li. Their meticulous attention and patience during the examination of the initial version of the dissertation and the oral examination have further refined my work.

I would like to extend my sincere thanks to the partners in our research group. I am particularly grateful to my co-authors, Mr. Donghui Dai, Mr. Jingyu Tong, and Mr. Zheng Gong. Their relentless efforts in our shared projects and their companionship have been sources of both strength and joy. Additionally, I must also express special

thanks to Dr. Qingrui Pan, Mr. Xiaopeng Zhao, and Ms. Xuanzhi Wang. Dr. Pan guided me in reading literature efficiently and identifying research challenges, which was very helpful for my subsequent personal research. Mr. Zhao often discussed specific ideas and programming details with me, practically advancing my experiment progress. Ms. Wang provided me with a great deal of encouragement and thoroughly explained her research methods, which significantly boosted my confidence. I am also thankful to Dr. Zhenlin An, Dr. Yuanhao Feng, Ms. Xueyuan Yang, Mr. Sicong Liao, Mr. Shen Wang, Ms. Zhuhang Li, Mr. Zhicheng Wang, Mr. Fengrui Zhang, Mr. Guosheng Wang, and all collaborators for the shared experiences and friendship during this journey.

I express my heartfelt gratitude to my parents, Mr. Yonghong Mei and Ms. Xiaohong Li, and my sister Ms. Xinyu Mei for their boundless love, patience, and unwavering support. Their strength and love have been my pillar of support over the years. I am also thankful to my good friends at PolyU, Ms. Zhu Wang, Mr. Qianyue Chen, Mr. Jiahang Li and Mr. Ruibin Li. Their constant friendship and moral support have been a great comfort.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Acoustic eavesdropping remains a significant security threat, employing stealthy techniques to intercept private conversations. Various methods leverage diverse channels to achieve this, such as motion sensors [1, 2, 3, 4, 5, 6, 7], wireless signals [8, 9, 10, 11, 12, 13], and camera inputs [14], each exploiting different vulnerabilities to access sensitive audio data. Particularly, utilizing a camera to capture mechanical vibrations (e.g., acoustic signals), known as visual vibrometry, has undergone extensive research over the years [15, 16, 17, 18, 19, 20, 21, 22, 23]. Yet, not every camera is suitable for acoustic eavesdropping. The Nyquist sampling theorem stipulates that the sampling rate should be double that of the signal's highest frequency. Furthermore, phonetic research indicates that vowels contain the primary energy in speech, and distinguishing different vowels often requires comparing their first two formants, which are distinct frequency components of the sound. The average second formant frequency for approximately 81.25% of vowels is around 1.5 kHz or lower [24]. Consequently, to perform accurate acoustic eavesdropping, a camera capable of achieving frame rates of 3 kHz or higher is necessary.

In this work, we revisit acoustic eavesdropping through visual vibrometry by exploiting an underutilized acoustic-optical side channel: the optical mouse. Common in

computing setups, the optical mouse incorporates a low-resolution digital camera to capture images of the surface below it. As the mouse is moved, it sequentially records images, which are then processed by an onboard digital signal processor (DSP). The DSP analyzes these images to detect patterns or shifts in position over time, allowing it to calculate the direction and speed of the mouse's movement. Driven by the demands of gaming applications, modern optical mice are designed to support exceptionally high frame rates, often exceeding 3.7 kHz. Furthermore, the optical sensor typically focuses on a small area (e.g., 3.1 mm^2) within its pixel array (e.g., 26×26), allowing the mouse to detect subtle variations at the sub-millimeter level induced by structure-borne soundwaves. The high frame rate and a finely focused sensing area characteristics highlight the significant potential of optical mice as tools for visual-vibrometry-enabled acoustic eavesdropping, a capability that has remained largely untapped until now.

To address this, we introduce a novel eavesdropping framework called JerryAttack, named in tribute to 'Jerry' Mouse from the 'Tom and Jerry' cartoons. Fig. 1.1 illustrates the attack scenario. We envision a typical setup where a target's workspace includes an optical mouse adjacent to a loudspeaker device, either integrated or external, both placed on a shared surface like a desk. Sound from the speakers causes vibrations that travel across the desk to the mouse, subtly altering the images captured by the mouses built-in camera. By analyzing the pixel stream from these images, an attacker can potentially detect the target's activities or reconstruct audible conversations during video conferences.

JerryAttack stands in stark contrast to traditional eavesdropping methods that often necessitate noticeable alterations to the environment or the introduction of obvious eavesdropping devices. By utilizing a commonplace computer accessory that is widely used in settings ranging from personal workspaces to corporate offices, JerryAttack can be integrated into daily routines without attracting attention. A demo video can be found at https://youtu.be/zlsrEucXh9U.

Fig. 1.1: Illustration of JerryAttack. The attacker can reconstruct the sound broadcasted from the loudspeaker via the acoustic-optical side channel.

However, effectively implementing this acoustic eavesdropping framework faces two main challenges:

• *How to extract pixel values from optical mice?* An optical mouse typically incorporates an image sensor paired with a DSP that preprocesses images to detect movement. This system traditionally outputs coarse-grained movement data, such as displacements along the X and Y axes measured in pixel units. This standard setup generally lacks the sensitivity required to detect the fine, subtle vibrations caused by sound from speakers. Visual vibrometry necessitates access to raw texture images captured by the sensor, but typical optical mice do not provide a high-level interface for accessing these images directly. This limitation poses a significant challenge for recovering sound from acoustic vibrations.

Upon reviewing technical datasheets, we made an unexpected discovery: mainstream optical sensors used in the mice, such as the ADNS-3050 [25], include a special register called the 'pixel grabber'. Originally intended for debugging or testing, this register captures and retains the value of a single pixel from each image frame captured by the sensor. By accessing this register, we can tap into a continuous and stable stream of pixel values at a high frame rate of 3.7 kHz. This access effectively turns the register into a backdoor for data extraction. Leveraging this capability allows us to overcome significant obstacles in accessing detailed image data, essential for sound recovery via visual vibrometry. For more technical information, please see § 4.

• *How to recover sound from the pixel stream?* The economical, low-resolution image sensors (e.g., 26×26 pixels) integrated into these mice are prone to thermal noise and inherent harmonics, potentially masking the subtle vibrations that are indicative of sound within the pixel data. Additionally, pixel values are quantized to 7 bits within a grayscale limit of 127, in stark contrast to the 16-bit quantization utilized in audio encoding. Consequently, the audio information encapsulated within the pixel stream undergoes significant compression and loss. These aspects increase the complexities in recovering audible signals from the acoustic-optical side channel.

To navigate this challenge, our initial approach involves leveraging traditional signal processing methods to accurately model the channel, followed by the implementation of a basic recovery algorithm. This foundational step sets the stage for further refinement. To enhance the initial, raw audio recovery, we utilize a band-split recurrent neural network (BSRNN) [26]. This advanced neural network architecture is specifically tailored to address the unique constraints posed by the pixel stream, such as its limited resolution and the presence of noise. The BSRNN methodically processes the audio data across different frequency bands, allowing for a more nuanced restoration of the sound. Both measures greatly mitigate the losses inherent in the pixel stream and improve the fidelity of the recovered audio. The technical details refer to § 5.

Contribution. While the diverse applications of optical mice are well-documented, their use for acoustic eavesdropping presents a novel innovation. Considering their ubiquity, the potential security implications of such an attack are significant and could be widespread. Our work introduces innovative methods for voice profiling and speech reconstruction via the developed acoustic-optical channel. Additionally, we have successfully validated JerryAttack on 9 optical mouse models across 7 manufacturers including giant Logitech and Razer. We tested the system across multiple corpora in six languages, achieving an average SNR of 7.3 dB in sound recovery micro-benchmark, an 83.27% overall accuracy for 48 individuals identification in voice profiling, and a median STOI of around 0.7 in joint-speech reconstruction, demonstrating the efficacy and potential impact of our approach in real-world settings.

Chapter 2

Related Work

The related work can be grouped into four categories:

(1) Motion Sensor Based Acoustic Eavesdropping: Motion sensor-based eavesdropping has emerged as a significant concern in mobile security, with several pivotal studies shedding light on the potential for motion sensors to capture speech signals. Michalevsky et al. [1] pioneered this field with their work on 'Gyrophone,' demonstrating that smartphone gyroscopes could recognize speech from acoustic vibrations despite the limited sampling rate of 200 Hz. Building upon this, Zhang et al. [2] introduced 'AccelWord,' which utilized accelerometers to detect voice commands, highlighting the sensitivity of these sensors to speech vibrations. Anand et al. [3] systematically analyzed the impact of speech on smartphone motion sensors, concluding that only loudspeaker-generated speech signals transmitted through a solid surface could significantly affect motion sensors and raising questions about the threat posed by everyday speech scenarios.

Further expanding the scope, Hu et al. [4] presented AccEar, an attack that employs a conditional Generative Adversarial Network (cGAN) to reconstruct high-fidelity audio from low-frequency accelerometer signals, marking a significant advancement in eavesdropping capabilities by overcoming hardware limitations. In a related vein,

Ba et al. [5] proposed AccelEve, a deep learning-based system that recognizes and reconstructs speech from accelerometer measurements, challenging the common belief about the narrow band of speech signals that motion sensors can capture. Lastly, Kwong et al. [6] showcased that mechanical components in magnetic hard disk drives could act as unintended microphones, extracting and parsing human speech, which introduces novel defense mechanisms against such cyberphysical attacks. Also worth mentioning, Yao et al. [7] propose an on-board eavesdropping method using a smartphone accelerometer at an extremely low 5 Hz sampling rate, exploiting stable rhythm features for classification tasks including scene, digit, city, and place recognition, posing a significant privacy threat despite the low sampling rate.

However, a common limitation across these studies is that the motion sensors used, such as accelerometers and gyroscopes, typically operate within a sampling rate range of 100 Hz to 500 Hz. In contrast, our JerryAttack leverages the mouse optical sensor, which can easily exceed 3 kHz, providing a significantly higher sampling rate. This enhanced capability allows for the capture of high-fidelity acoustic data, making it better suited for precise eavesdropping in real-time scenarios.

(2) Wireless Signal Based Acoustic Eavesdropping: The field of wireless signal-based eavesdropping has seen significant advancements, with research exploring various wireless technologies to intercept acoustic communications. Millimeter-wave (mmWave) technology, for instance, has proven effective for high-resolution eavesdropping. Hu et al. [11] introduced mmEcho, a system that uses mmWave to measure micrometer-level vibrations induced by sound, enabling eavesdropping without line-of-sight or prior knowledge of the target's vocabulary. Also, Hu et al. [10] developed MILLIEAR, a system capable of reconstructing audio from vibrations using generative machine learning models. In the realm of WiFi signals, Wei et al. [9] demonstrated how acoustic eavesdropping can be achieved by extracting speaker vibrations through WiFi. Similarly, Wang et al. [12] expanded eavesdropping capabilities through Impulse Radio Ultra-Wideband (IR-UWB) technology, with their

UWHear system capable of sensing audio through walls. RFID technology has also been explored, with Wang et al. [13] showing how audio can be intercepted via RFID by capturing sub-mm level vibrations.

However, wireless signal-based eavesdropping systems also face some challenges that limit their practicality. First, their performance is heavily influenced by environmental factors such as ambient noise, interference, and signal attenuation. Second, these systems often require specialized signal transceiver equipment, particularly high-precision mmWave devices, which are very expensive.

(3) Visual Vibration Based Acoustic Eavesdropping: The field of visual vibrometry, which focuses on recovering sound from silent video footage, has evolved significantly through the integration of signal processing, machine learning, and computer vision techniques. Initial research by Akutsu et al. [15] explored the potential of visual data for audio recovery, with subsequent studies like those by Fuse et al. [16] demonstrating how vibrations captured in the video can reconstruct sound. Innovations continued with Mim et al. [17] applying optical flow techniques to detect minute vibrations in video for sound extraction. A pivotal advancement was the 'Visual Microphone' concept by Davis et al. [27], showcasing sound recovery by analyzing object vibrations within video frames. This was further enhanced by high-speed video analysis [20] and the application of machine learning, particularly cross-modal generative adversarial networks [21], which have bridged visual and auditory data, enhancing multimedia processing. Some newer work uses high-speed cameras to observe lamp lights [22] or small shiny objects [23] to recover sound signals. These developments underscore a multidisciplinary approach combining traditional and modern audio-visual data analysis and reconstruction techniques.

Although these visual vibration-based systems can achieve high-quality sound restoration, they require expensive high-speed cameras and precise calibration before attacking to accurately observe the target, which limits its application in practical scenarios.

(4) Repurposing of Optical Mice: Recent studies have demonstrated innovative uses of optical mice in various fields of research and technology development. Ng et al. [28] explored the potential of optical mice in harmonic oscillator experimentation, showcasing their utility in physics education. Similarly, Ng et al. [29] investigated the application of optical mice as two-dimensional displacement sensors, offering a cost-effective solution for precise measurements. Tresanchez et al. [30] repurposed optical mouse sensors as incremental rotary encoders, highlighting their accuracy and efficiency. Palacin et al. [31] demonstrated the use of optical mice for indoor mobile robot odometry measurement, contributing to advancements in robotics navigation. Lastly, Ullrich et al. [32] explored the utilization of optical sensors from mice to create new input devices, emphasizing the versatility and adaptability of these components. These studies collectively underline the significant potential of repurposing optical mice for innovative applications.

Comparision with Existing Acoustic Eavesdropping Systems: We present a detailed comparison between JerryAttack and existing systems in Table. 2.1: First, JerryAttack outperforms motion sensor-based systems by leveraging the naturally high sampling rate (abbreviation SR) of the optical sensor in the mouse, providing superior accuracy and efficiency. Second, unlike many existing systems, JerryAttack does not require expensive professional equipment, such as millimeter-wave transceivers or high-speed cameras, making it more accessible and cost-effective. Third, JerryAttack utilizes a unique acoustic-optical channel to filter out ambient airborne sound noise, a feature that several existing systems lack, ensuring more re-

Table 2.1: JerryAttack VS Existing Systems

Eavesdropping System	High SR?	Cheap?	Noise-resistant?	Easy-to-deploy?
Motion sensor-based	✗	✓	✗	✓
Wireless signal-based	✓	✗	✗	✓
Visual vibration-based	✓	✗	✗	✗
JerryAttack (Our system)	✓	✓	✓	✓

liable signal detection in noisy environments. Finally, JerryAttack does not require precise calibration before the attack, unlike visual vibration-based systems, making it significantly easier to deploy in real-world scenarios.

Chapter 3

Threat Model

3.1 Attack Scenarios

We envisage a scenario where the audio output from the loudspeakers creates vibrations that travel through the desk to the mouse, leading to minute alterations in the imagery recorded by the mouse's onboard camera. The objective of the attacker in this scenario is to clandestinely capture confidential personal information emanating from the loudspeaker. This could encompass a variety of sensitive data as follows:

- **Patterns of Media Consumption:** Identifying which movies or music the user is consuming and discerning specific preferences in content, habitual viewing, or listening patterns.
- **Response to Voice Commands**: Capturing the response to spoken instructions from voice-activated computers can reveal the user's personal preferences, specific commands for controlling smart home devices, or sensitive inquiries made to these assistants.
- **Insights from Virtual Meetings:** Acquiring detailed content of discussions, such as corporate strategies or private personal matters shared during online meetings

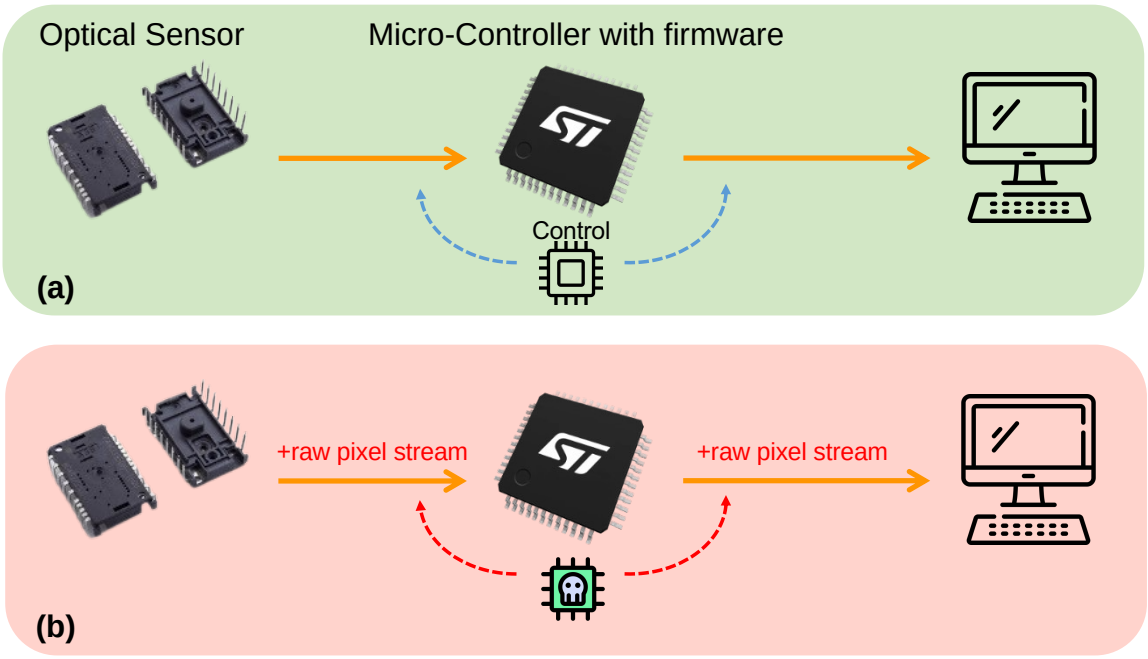


Fig. 3.1: Illustration of firmware update.

on platforms like Zoom or Google Meet.

- **Sensitive Online Calls:** Intercepting private communications on digital platforms, such as Skype and WhatsApp calls, to obtain confidential information.

In short, we can identify the victim's behaviors or directly reconstruct human speech from the acoustic vibrations.

3.2 Attacker's Capabilities

One might wonder why the attacker does not hack the microphone for acoustic eavesdropping directly. The answer lies in the stringent permissions required to access microphones or other acoustic sensors. Even when permissions are obtained, operating systems usually alert the user to the activation of sensors, such as displaying microphone icons, which could warn the victim of a breach. Additionally, some anti-recording hardware [33] can emit ultrasonic waves and interfere with the normal

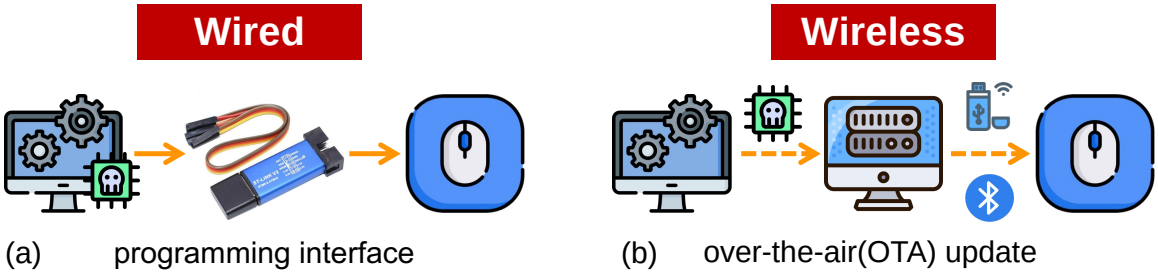


Fig. 3.2: Two approaches for firmware modification.

recording of the microphone based on intermodulation distortion [34]. In contrast, JerryAttack operates with greater discretion. For the effective activation of JerryAttack, the attacker is presumed to have the following two key capabilities:

- **Firmware Updates:** The attacker can compromise the optical mouse firmware, allowing access to the 'pixel grabber' register to capture the real-time pixel stream. To explain the fundamentals of mouse firmware, lets refer to Fig. 3.1. As depicted in part (a), under normal conditions, the optical sensor transmits basic data such as x and y direction displacements and motion status to the microcontroller, which then relays this information to the computer. However, in part (b), the attacker seeks to modify the firmware, enabling the optical sensor to transmit the additional raw pixel stream externally.

  Firmware modification can be carried out through both wired and wireless methods, as shown in Fig. 3.2. The wired method involves using a programming interface, like ST-Link [35], to directly inject malicious firmware into the mouses microcontroller. The wireless method utilizes over-the-air (OTA) updates [36], where the malicious firmware is first downloaded to the host computer via the HTTP/HTTPS protocol. It is then injected into the mouse using either 2.4 GHz RF or Bluetooth, depending on the mouse type.

  In practice, the attacker may deceive the user into downloading the compromised firmware by presenting a fake update prompt from a fraudulent website, which overwrites the original firmware. Once the firmware is modified, the captured pixel stream is then transmitted to the host computer via USB or Bluetooth.

It is important to highlight that this capability does not rely on the attacker needing significant preparation time before launching the attack. Instead, it emphasizes the security risk posed by the 'pixel grabber,' which enables covert firmware modification, even at the manufacturing stage.

- **Data Exfiltration:** The adversary possesses the ability to transmit the acquired pixel stream over the Internet, which is roughly 7kbps. In the context of broader Internet bandwidths, often measured in megabits per second (Mbps), the covert transmission of such modest amounts of data is unlikely to capture the user's attention. The attacker then reconstructs the sound signal on a separate machine with enough computing resources. In addition, using the mouse's wireless channel, such as Bluetooth or 2.4 GHz RF, to directly send the pixel stream to the attacker without affecting the victim's normal use is also an area worth exploring.

  For a Bluetooth optical mouse, implementing dual Bluetooth roles communication in the updated firmware could be a suitable solution. This would involve configuring the mouse to act as both a **peripheral** (connecting to the victims computer) and a **central** (connecting to the attackers computer), similar to how certain devices, such as the Sony WH-1000XM4 headphones [37], can simultaneously connect to multiple devices and switch between communication channels. In this setup, the mouse maintains its primary HID connection with the victims computer as a peripheral, while in the background, it can also initiate a Bluetooth Low Energy (BLE) connection to the attackers computer as a central, enabling it to transmit the pixel stream. Communication between devices is managed through dynamic channel switching, where the mouse can switch between channels based on priority, depending on activity. For example, when the victim is actively using the mouse, the mouse prioritizes the HID communication. During idle periods or low-activity moments, it can switch to the secondary communication channel to transmit data without interrupting the victims experience. The switching mechanism must be seamless and fast to avoid alerting the victim.

  For a USB wireless optical mouse, a similar dual communication approach could

be implemented, but with the added complexity of handling RF (Radio Frequency) communication in addition to Bluetooth. Commercial mouse hardware is typically optimized for a one-to-one connection, lacking the necessary protocols and memory to manage multiple 2.4 GHz RF connections concurrently. Modifying a USB wireless mouse to establish dual connections would require extensive hardware alterations, making it less feasible. In this case, the firmware could be updated to allow the mouse to communicate with both the victim's computer and the attacker's computer by utilizing two separate communication protocols: USB wireless (2.4 GHz RF) for the victims device and BLE for the attackers device. Many commercial mice, such as the Logitech MX Anywhere 3 [38], already support both Bluetooth and USB wireless, which could make this approach more practical.

In our work, we operate under the assumption that the victim positions the optical mouse close to the computer (i.e., < 20cm). This assumption is reasonable in practicality, as users typically maintain the mouse near the computer for easy use.

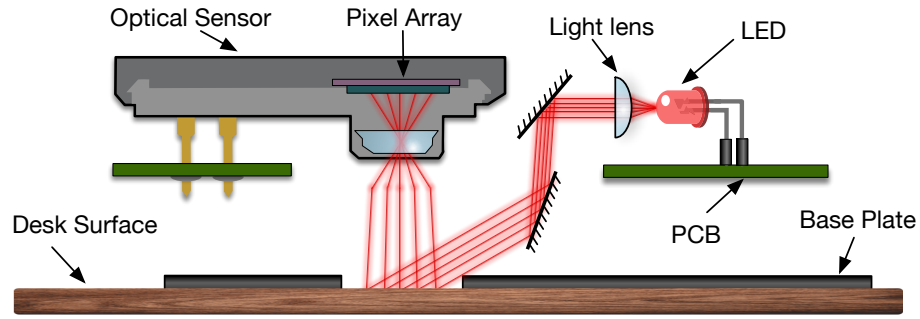
Chapter 4

Transforming Mice into High-Speed Cameras

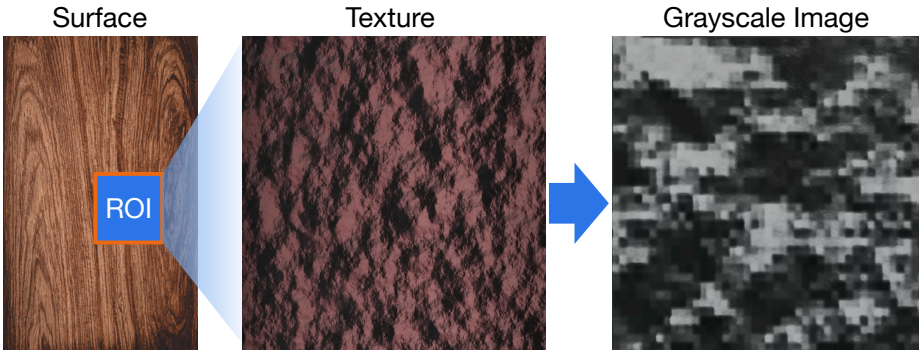
In this chapter, we detail the operational mechanism of optical mice and subsequently describe how to repurpose these devices into high-speed, single-pixel cameras, serving as side channels for acoustic eavesdropping.

4.1 Background of Optical Mice

Mice, crucial peripherals for computer interaction, are available in various forms such as mechanical (ball) and optical mice. Among these, optical mice have gained market dominance due to their enhanced durability and reliability compared to mechanical mice. Fig.4.1(a) illustrates the internal structure of an optical mouse, featuring components such as an optical sensor with pixel array, a small LED, some light lens, and a single printed circuit board (PCB) that integrates a DSP, a microcontroller, and a communication module (PS/2, USB, or wireless). Fig.4.1(b) explains the working principle of an optical mouse. Specifically, the LED projects light through the optical lens onto the desk surface. The reflected light, carrying the texture of the surface,



(a) Schematic Diagram



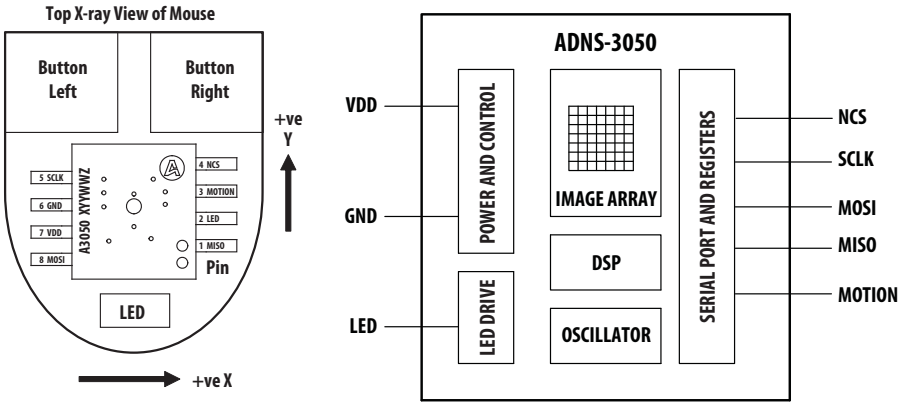
(b) Working Princicple

Fig. 4.1: Illustration of Optical Mouse

is captured by the image sensor's pixel array, which converts these optical signals into grayscale values that represent light intensity. The onboard DSP processes these images to detect changes between consecutive frames, thus determining the mouses movement direction and speed. This process ultimately translates the physical movement of the mouse into cursor navigation on the computer screen, allowing users to interact with their digital environment.

4.2 Acquiring Pixel Stream

Optical mice are equipped with a compact image sensor that comprises an array of pixel diodes. Unlike regular cameras that boast millions of pixels, the image sensors in optical mice have only a few hundred pixels which is enough to be used as a



(a) Structure (b) Image Sensor

Fig. 4.2: Illustration of the ADNS image sensor. (a) shows the schematic structure of an optical mouse equipped with an ADNS-3050 image sensor. (b) shows the primary components of the image sensor.

reference for calculating displacement. However, what these mouse cameras lack in pixel count compared with common cameras in smartphones, they compensate with an exceptionally high frame rate, far surpassing regular cameras. Optical mice typically support frame rates of 3 kHz or higher, which may reach up to 20 kHz in models designed for high-end gaming.

Optical mice come with various camera configurations. For illustrative purposes, we will use the Logitech G402 optical mouse as an example, however, the datasheet for its optical sensor AM-010 is not public. Fortunately, the AM-010 can be considered a variant of the ADNS-3050 (since the AM-010 is a variant of the PMW-3320 [39], which

Table 4.1: ADNS-3050 Registers

Address	Register Name	Read/Write	Default Value
0x00	PRODUCT ID	R	0x09
0x01	REVISION ID	R	0x00
0x02	MOTION STATUS	R/W	0x00
0x03	DELTA X	R	0x00
0x04	DELTA Y	R	0x00
...	...	...	...
0x0b	PIX GRABBER	R/W	0x00
...	...	...	...

Algorithm 1 Acquiring pixel stream from the pixel array via the grabber register

```
Initialization: N ← 26, ImgBuff[N × N]
Initialize: SPI CONFIG, CLOCK CONFIG
repeat
    SPI_WRITE(0x0b, 0x01)
    index ← 0
    Clean ImgBuff
    repeat
        temp ← SPI_READ(0x0b)
        ImgBuff[index] ← temp
        index ← index + 1
    until index == N × N -1
    Send ImgBuff to PC
until Power Shut Down
```

in turn is a variant of the ADNS-3050 [40, 41]), while the biggest difference is that the pixel array size increases from 19×19 of ADNS-3050 to 26×26 of AM-010. Therefore, this paper will refer to the datasheet of the ADNS-3050 for explanations [25], which integrates more than 30 registers. Detailed specifications of these registers are provided in Table 4.1. For example, the registers at $0x00$ to $0x01$ hold the product ID and revision number. The 'MOTION_ST' register at address $0x02$ shows motion status, where zero indicates stillness and a non-zero value signals movement. Motion data (Δx and Δy) is stored from $0x02$ to $0x04$. A microcontroller reads these values via the SPI bus and transmits them to the host computer.

Pixel Grabbing. During our examination of the mouse's internal registers, we identified a notable register known as the 'pixel grabber,' located at address $0x0b$. This register enables the extraction of a single-pixel value from each frame captured by the image sensor. Each time the register is accessed, its location pointer advances to the next pixel, moving sequentially across every location. Using this feature, an entire frame can be constructed by continuously reading pixel values (26×26 reads in this case) from the $0x0b$ by the SPI bus. Algorithm 1 shows the pseudo-code of the whole process. In our experiment, we can achieve a stable readout rate of about 3.7 kHz (i.e., the operational speed of the AM-010 sensor), effectively turning the

mouse into a high-speed, single-pixel camera that outputs a continuous pixel stream. The continuous pixel stream is packaged and sent to the host computer. Our custom software monitors the serial or wireless connection to capture the complete pixel stream and transmits it to the attacker's server for further processing.

Ubiquity of the 'Pixel Grabber' register. The most crucial step in the previous pixel-grabbing process is extracting pixel values from the 'pixel grabber' register(Abbreviation PG). This naturally raises the question: Do mainstream mice include such a register? To answer this, we surveyed 11 popular mice from leading manufacturers such as Logitech and Razer. The results, presented in Table 4.2, reveal that 9 of these 11 mice are equipped with one of five different image sensors, each with publicly available datasheets and PG. Furthermore, a review of PixArt's official website [42], a major optical sensor manufacturer, shows that all PMW and ADNS series sensors come with public datasheets and PG. This indicates that all mice using these sensors are susceptible to the JerryAttack, demonstrating our attack's wide applicability. The remaining two models, the high-end Logitech G903 and Razer Viper V3 Pro feature the HERO-25K and Focus Pro-35K sensors, respectively. These sensors are custom-developed and lack publicly available datasheets, so it is unclear whether they include PG. However, these mice are priced above $150 and are

Table 4.2: Survey of the 'pixel grabber' register on 11 mainstream mice

Manufacturer	Model	Image Sensor	Public datasheet?	With PG?
Logitech	G903	HERO-25K	No	Unknown
Razer	VIPER V3 PRO	Focus Pro-35K	No	Unknown
Logitech	G402	AM-010	Yes	Yes
Logitech	G100S	AM-010	Yes	Yes
Logitech	G302	AM-010	Yes	Yes
Razer	DeathAdder Elite	PMW-3389	Yes	Yes
Madcatz	RAT3	ADNS-3090	Yes	Yes
Asus	GT200	ADNS-3050	Yes	Yes
Zowie	EC1	ADNS-3060	Yes	Yes
Tucano	GAMEZONE Toros	ADNS-3050	Yes	Yes
Cool Master	Lite L Combo	ADNS-3050	Yes	Yes

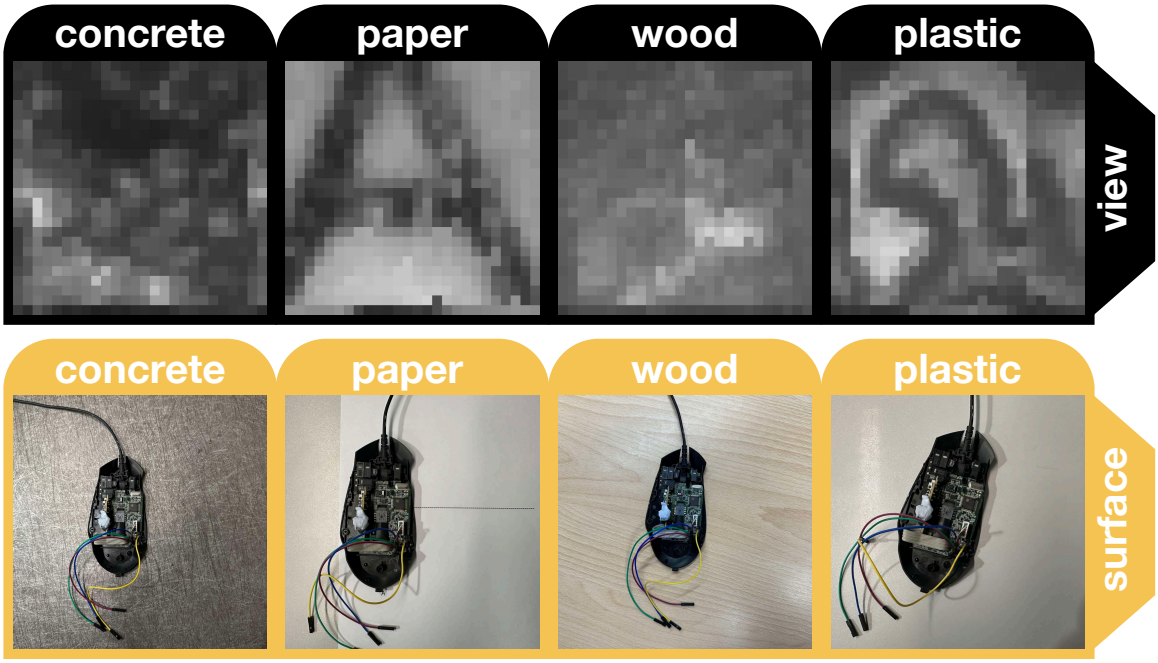


Fig. 4.3: Illustration of images captured by the pixel sensor across different types of textures.

typically targeted at professional e-sports players, making them uncommon in standard office settings. Despite this, if we really want to attack a sensor without a public datasheet, perhaps reverse engineering [43] could offer a possible solution, which may capture the data exchanged between the sensor and the host using a protocol analyzer and test different register settings to observe how they affect the sensors output and functionality.

4.3 Feasibility Verification

To demonstrate the capability of transforming an optical mouse into an ultra-fast camera, we present some initial experimental results in Fig. 4.3. In these experiments, the mouse was placed on various materials including concrete, paper, wood, and plastic. Each image was constructed using 26×26 pixel values collected via the pixel grabber. These images distinctly capture the unique textures of each surface. A

notable feature in the captured images is the letter 'A', measuring 3.1mm^2, clearly visible on the surfaces. These results convincingly validate the feasibility of extracting pixel values from an optical mouse using the pixel grabber register.

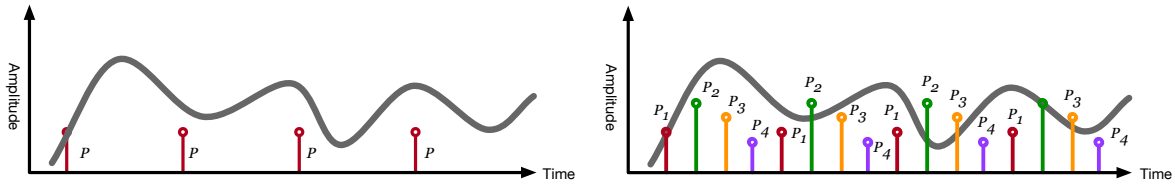
Chapter 5

Developing the Acoustic-Optical Side Channel

After converting the optical mouse into an ultra-fast camera, we leverage its inherent sensitivity to vibrational disturbances caused by sound waves. In this chapter, we detail the development of the acoustic-optical side channel.

5.1 Channel Model

Speaker-generated vibrations manifest as structure-borne waves, a type of mechanical waves that carry energy through solid materials by causing particle interactions within the medium. There are three main propagation modes: Rayleigh waves, longitudinal waves, and transverse waves [44]. Each displays unique traits affecting their interaction with surroundings and detection mechanisms. In our scenario, the mouse detects surface texture changes primarily induced by Rayleigh waves, which causes the solid surface to float up and down [45] and, in turn, makes the distance between the image sensor and desktop fluctuate accordingly. Combined with Fig. 4.1, this interaction between the loudspeaker and the mouse forms a classic backscatter sys-



(a) Single-Pixel Sampling

(b) Multi-Pixel Sampling

Fig. 5.1: Illustration of the sampling process. The sampling process involves multiplying the constant pixel offsets by the signal over time.

tem where the loudspeaker modulates the light with varying amplitudes due to the Rayleigh waves, and the image sensor captures this amplitude-modulated (AM) data. Next, we formally model this channel.

Single-Pixel Sampling: We first examine a simplified scenario where the acoustic-optical channel comprises just a single pixel. In the absence of an acoustic signal, this pixel maintains a steady light intensity, denoted as p. However, the introduction of an acoustic signal $S(t)$ alters the gap between the surface and the pixel, akin to how a backscatter system operates. In this context, the acoustic signal modifies the surface's reflective properties, resulting in a fluctuating light intensity captured by the pixel. As Fig. 5.1(a) shows, the sampling results are formalized as $p \cdot S(t)$ at the moment t.

Muti-Pixel Sampling: We proceed to explore the scenario of multi-pixel sampling where the image sensor contains an array of $\sqrt{N} \times \sqrt{N}$ pixels. Given that the area covered by the image sensor is quite small (approximately $3.1mm^2$) relative to the wavelength of sound (around 5m), we can assume that the acoustic signal impacts all pixels uniformly.

As previously mentioned, the pixel grabber sequentially selects a single pixel value from each frame, moving from left to right and from top to bottom. Fig. 5.2 illustrates the procedure of acquiring a pixel stream. Consequently, the sampling results are formalized as follows:

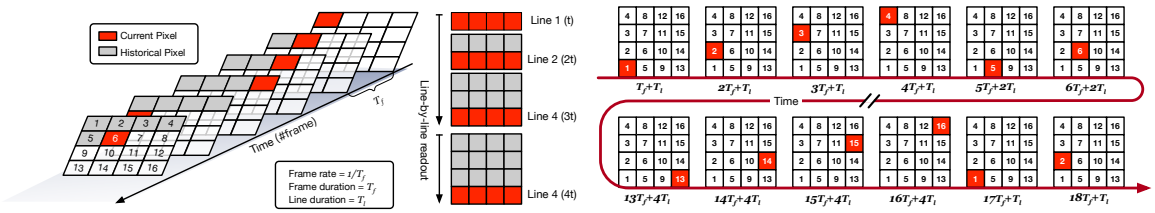


Fig. 5.2: Pixel Grabbing Process with a 4×4 Image Sensor. For every frame, the pixel grabber extracts a single pixel value, progressing systematically from top to bottom and left to right, at a frame rate of f Hz. The image sensor implements a line-by-line sequential readout strategy, facilitating the capture of an entire row of pixels in one reading cycle. As a result, pixel values are sampled at non-uniform intervals $T_f + T_l, 2T_f + T_l, 3T_f + T_l, 4T_f + T_l, 5T_f + 2T_l \ldots$, where T_f represents the time to read a frame, and T_l denotes the time to read a line.

$$\begin{bmatrix} \tilde{S}(t_1) \\ \tilde{S}(t_2) \\ \tilde{S}(t_3) \\ \vdots \\ \tilde{S}(t_N) \end{bmatrix} = \begin{bmatrix} p_1 & 0 & 0 & \cdots & 0 \\ 0 & p_2 & 0 & \cdots & 0 \\ 0 & 0 & p_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & p_N \end{bmatrix} \begin{bmatrix} S(t_1) \\ S(t_2) \\ S(t_3) \\ \vdots \\ S(t_N) \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_N \end{bmatrix} \tag{5.1}$$

$$\text{or} \quad \tilde{S} = MS + n \tag{5.2}$$

where $\tilde{S} = [\tilde{S}(t_1), \tilde{S}(t_2), \cdots, \tilde{S}(t_N)]^T$ corresponds to the sampled outcomes of the acoustic signal, the set $\{p_1, p_2, \cdots, p_N\}$ represents the baseline intensity offsets for each pixel, and $n = [n_1, n_2, \cdots, n_N]^T$ denotes the thermal noise affecting each pixel. Particularly, $\tilde{S}(t_i) = p_i \cdot S(t_i)$. The above equation models the sampling process over a period that includes N samples. This pattern of sampling the acoustic signal is consistently repeated. Fig. 5.1(b) visualizes the whole sampling procedure. It is worth noting that multi-pixel sampling captures the acoustic signal at different scales caused by the inherent intensity offsets.

Let T_f represent the time interval between two consecutive frames. Ideally, sampling would occur at each interval T_f. However, due to the architecture of CMOS sensors,

which typically process pixel values line by line to optimize memory addressing, the timing for capturing pixel values varies. As Fig. 5.2 shows, acquiring a pixel from the j^{th} line takes approximately jT_l seconds, where T_l is the duration required to read a line of pixels. Therefore, the timing for the i^{th} sample is given by $iT_f+((i//\sqrt{N})+1)T_l$, where i ranges from 1 to N, indicating that the sampling across the image sensor is not uniform. This variation in timing can affect the precision of capturing the acoustic signal due to slight discrepancies in the sampling intervals.

5.2 Naive Sound Recovery

Building upon Eqn. 5.2, the acoustic signal can be recovered by the following straightforward way:

$$S = M^{-1}(\tilde{S} - n) \approx M^{-1}\tilde{S} \tag{5.3}$$

This approximation is valid under two ideal conditions. First, the magnitude of each component of $\tilde{S}(t_i)$ must significantly exceed that of n_i, i.e., $\tilde{S}(t_i) \gg n_i$. This is typically achieved when the loudspeaker is close to the mouse, thereby providing a high SNR. Second, the matrix M must be known and invertible. We can determine M when the mouse is stationary before any eavesdropping begins. The idle signal received, $\tilde{S}_0$, can be modeled as:

$$\tilde{S}_0 = M \cdot \mathbf{1}_N + n \approx [p_1, p_2, p_3, \ldots, p_N]^T \tag{5.4}$$

where $\mathbf{1}_N$ is a vector of N ones, indicating no acoustic signals are present, and T represents the transpose. Besides, M is a full-ranked matrix because all p_i are not 0. The inverse of M (i.e., M^{-1}) is defined as:

$$M^{-1} = \begin{bmatrix} \frac{1}{p_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{p_2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{p_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{p_N} \end{bmatrix} \tag{5.5}$$

The above framework describes how we deduce each pixels baseline intensity from the quiescent signal $\tilde{S}_0$, when the mouse remains stationary.

Algorithm. The collected sequence of pixel values is first segmented into K blocks, each denoted as $\{\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_K\}$. Each block contains N samples that correspond to the N pixels, structured as follows:

$$\tilde{S}_k = [\tilde{S}(t_{kN+1}), \tilde{S}(t_{kN+2}), \ldots, \tilde{S}(t_{kN+N})]^T \tag{5.6}$$

where k ranges from 0 to $K-1$. The recovery of the acoustic signal from these segments is then performed using the following steps:

$$\begin{aligned} S_k &\approx M^{-1}\tilde{S}_k \\ &= [\frac{\tilde{S}(t_{kN+1})}{p_1}, \frac{\tilde{S}(t_{kN+2})}{p_2}, \ldots, \frac{\tilde{S}(t_{kN+N})}{p_N}]^T \\ &= \tilde{S}_k \oslash \tilde{S}_0 \end{aligned} \tag{5.7}$$

where $\oslash$ represents element-wise division. Concatenating these K reconstructed segments, i.e., $\{S_1, S_2, S_3, \ldots, S_K\}$, allows for the comprehensive recovery of the entire acoustic signal, ensuring a thorough and precise restoration process.

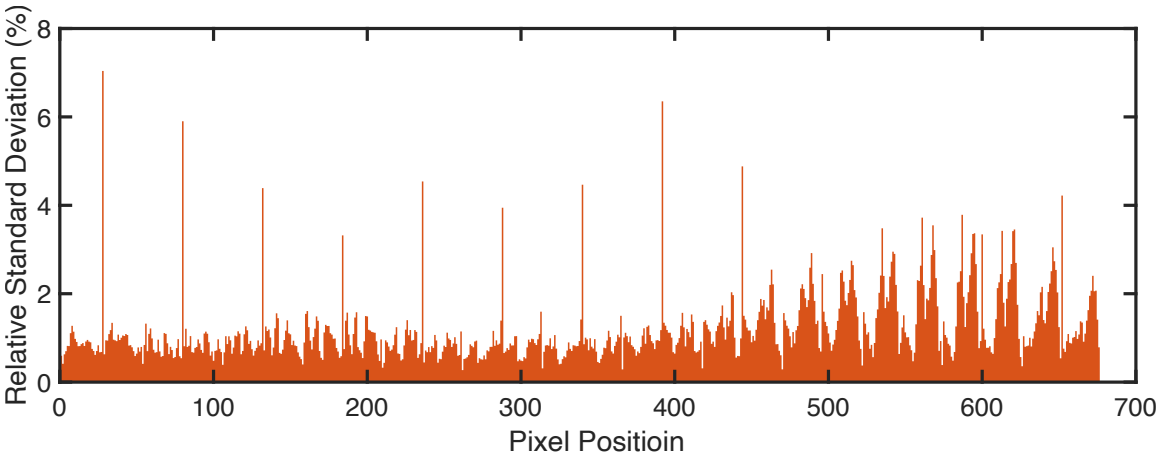


Fig. 5.3: Pixel value distributions. This illustrates the thermal noise in a stationary optical mouse, captured across all the 676 pixels. The y-axis represents the standard deviation of the pixel value relative to the maximum intensity value.

5.3 Channel Characteristics

Next, we further investigate the characteristics of the acoustic-optical channel via empirical experiments.

Characteristic I: Thermal Noise. Economical image sensors, such as those found in optical mice, often struggle with noise management, a challenge evident in our experimental observations. We immobilized the mouse and analyzed the values of a 26×26 pixel grid. The distributions of their relative standard deviations (RSTD), which we define as the standard deviation normalized by the maximum intensity value (e.g., 127), are illustrated in Fig. 5.3. Among these 676 pixels, we noted a maximum RSTD of 7% and an average RSTD of 1.19%. This is in stark contrast to the RSTD of 0.02% typically observed in standard CMOS sensors within smartphones [46], indicating that thermal noise-induced fluctuations in an optical mouse's pixel sensor are significantly higher, about 60 times greater than those in regular sensors. Such pronounced noise levels lead to inaccuracies in signal recovery as per Eqn. 5.3 due to the approximation assumption.

Characteristic II: Inherent Harmonics. We analyze the time-frequency spec-

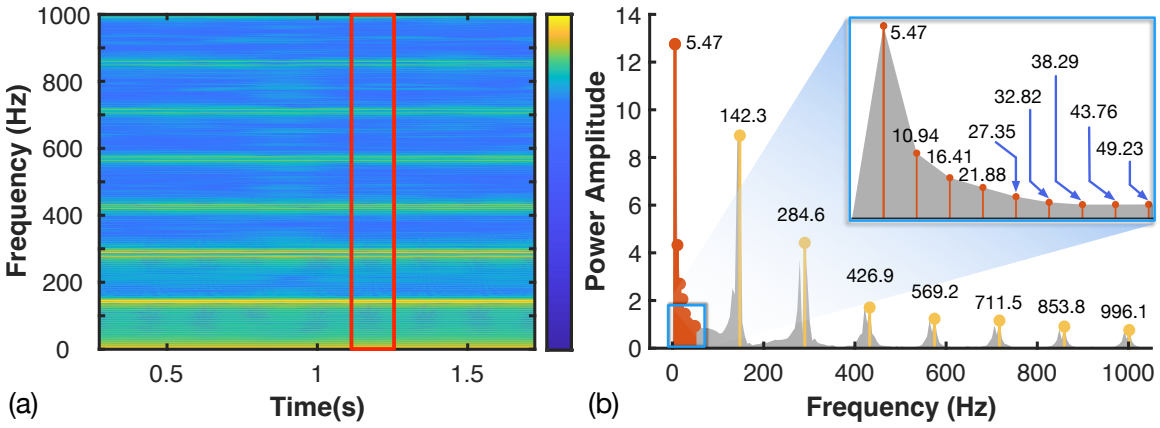


Fig. 5.4: STFT and FFT result of the pixel stream captured without external sound stimulus

trum of the pixel stream captured from an optical mouse in the absence of any external sound stimulus. We configured the Short-Time Fourier Transform (STFT) with a window size of 2048 samples and an overlap of 2032 samples. Under no excitation conditions, we expect a uniform spectrum due to stable pixel values. However, the spectrum reveals the presence of many inherent harmonics, as illustrated in Fig. 5.4(a). To delve deeper, we show the spectrum across a single FFT window in Fig. 5.4(b). We identified two distinct sets of harmonics, with their fundamental frequencies being 5.47 Hz and 142.31 Hz, respectively. These harmonics stem from residual intensity offsets and non-uniform sampling within the device.

One might be wondering how the inherent harmonics are generated. As shown in Fig. 5.2, the periodic reading of the same pixel every $16T_f$ second, for instance, $(17T_f + T_l) - (T_f + T_l) = 16T_f$, generates a hidden periodic signal at the frequency of $1/16T_f$. On the other hand, pixel sampling should occur uniformly at a frequency of $1/T_f$. However, due to the sequential line-by-line readout process, the sampling of pixels on the i^{th} line experiences a delay of T_l seconds relative to the $(i-1)^{\text{th}}$ line. This delay introduces a new periodic signal in the spectrum with an interval of $4T_f + T_l$. Extending this principle to an image sensor with a $\sqrt{N} \times \sqrt{N}$ pixel array, the inherent frequencies are derived as $1/(NT_f)$ and $1/(\sqrt{N}T_f + T_l)$. In our case,

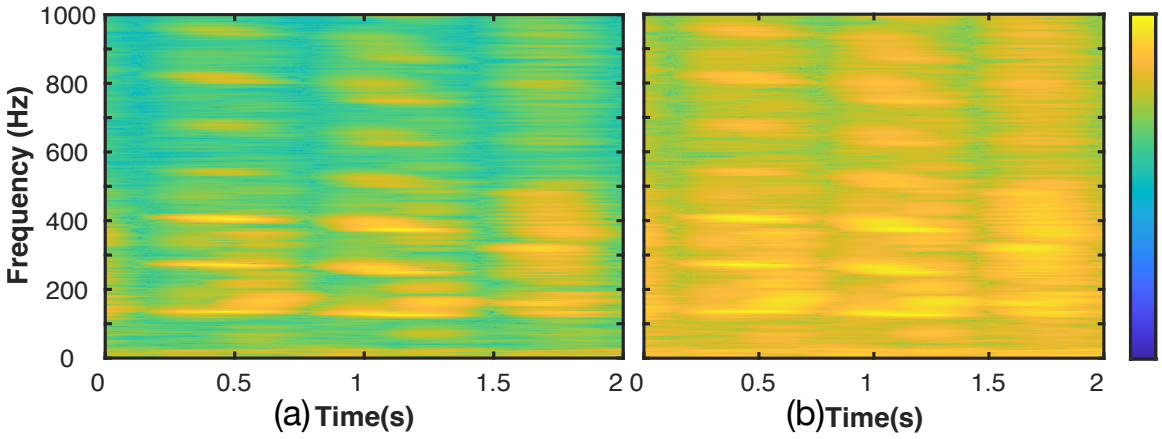


Fig. 5.5: Spectrum comparison between coarse-grained and fine-grained quantification. (a) shows the STFT result of 16-bit quantized human speech while (b) illustrates that of 7-bit.

$N = 676$, $T_f = 1/3700 \approx 270\mu s$ and $T_l \approx 7\mu s$, we precisely calculate the inherent frequencies to be 5.47 Hz and 142.31 Hz, which perfectly clarifies the origin of the observed harmonics. To summarize, by considering each pixel as a separate sensor, the pixel stream is effectively generated by multiple sensors operating at varying scales. This results in the production of inherent harmonics within the data.

Characteristic III: Lossy Encoding. The image sensor in an optical mouse operates using 7-bit encoding for pixel intensity, which limits the quantization to 128 discrete levels. This is a stark contrast to the 16-bit quantization employed in standard audio processing, where 65,536 levels are available. The 16-bit standard is specifically designed to accommodate the broad dynamic range and sensitivity of human hearing, ensuring high-fidelity audio reproduction. When the quantization depth is reduced to just 7 bits, as in the optical mouse's image sensor, there is a significant decrease in the ability to accurately capture and reproduce the fine details of audio signals. This reduction in resolution has a direct impact on the quality of audio that can be transmitted through this side channel.

The diminished quantization depth introduces considerable noise and distortion, making it much more difficult to recover the original sound with clarity. This effect is

clearly demonstrated in Fig. 5.5, where a comparison is made between speech samples quantized at 7-bit and 16-bit levels. The 7-bit quantization introduces noticeable artifacts and a substantial increase in background noise, which degrades the overall quality of the recovered audio. Consequently, the lower resolution not only affects the fidelity of the sound but also complicates the process of sound recovery, making it a more challenging task to extract intelligible audio from the quantized data.

5.4 Sound Enhancer

The acoustic-optical side channel presents significant challenges as an acoustic medium due to the aforementioned characteristics. These obstacles make the task of sound recovery through traditional signal processing techniques almost impractical. Inspired by the recent work in band-split recurrent neural networks [47, 48, 26], we employ a specialized neural network to enhance the recovered sound. This network is designed to refine the quality of the acoustic signal recovered through the naive algorithm. The neural network architecture of the sound enhancer is shown in Fig. 5.6(a), with its key components described below:

(1) Preprocessing. Initially, the pixel stream is processed through the naive sound recovery algorithm at a frequency of 3.7 kHz. Then the time-domain acoustic signal is transformed into a complex-valued spectrogram via the Short-Time Fourier Transform (STFT). This spectrogram serves as the input for the sound enhancer. This step is crucial for converting the raw pixel stream into a format that saves both amplitude and phase information in the frequency domain, ultimately aiming to produce a clearer and more accurate representation of the original acoustic environment.

(2) Band Splitter. The initial step involves segmenting the input spectrogram into various spectral bands. Given a complex-valued spectrogram $\mathbf{X} \in \mathbb{C}^{F \times T}$, it is segmented into K distinct frequency bands, denoted as $\{W_1, W_2, \ldots, W_K\}$, with each

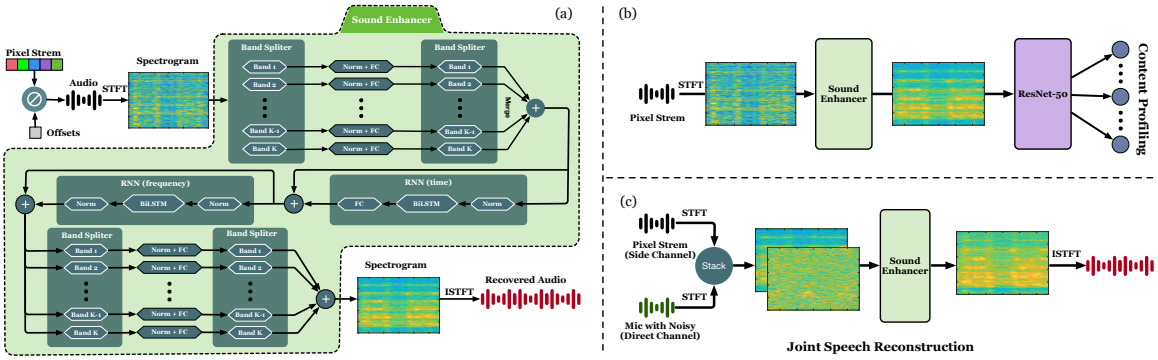


Fig. 5.6: Neural Network Architectures for Audio Processing. (a) illustrates the neural network architecture designed for sound enhancement, detailing the layers and connections utilized. (b) depicts the voice content profiling architecture, which integrates a sound enhancer followed by a Resnet-50 to categorize sound content. (c) shows the joint speech reconstruction architecture, which utilizes dual-channel information of pixel streams and low-SNR microphone recordings to recover high-quality speech.

$W_k \in \mathbb{C}^{F_k \times T}$. Here, F_k denotes the specific frequency range of each band, and F is the total frequency span, i.e., $F = \sum_{k=1}^{K} F_k$, while T signifies the temporal axis. To focus on the lower spectrums by improving their resolutions, the segmentation into bands is executed using a non-linear approach: the frequency range 0∼1 kHz is segmented into 10 segments of 100 Hz each, the frequency range 1 kHz∼2 kHz is divided into 5 segments of 200 Hz each, and the remaining range is divided into 400 Hz sub-bands. Subsequently, all the frequencies sub-band W_k are transformed into real-valued features $Z_k \in \mathbb{R}^{N \times T}$ via layer normalization module (Norm) and fully connected (FC) layer, where N donates the feature dimension. The spectral features from all K bands are then integrated into a composite feature tensor $\mathbf{Z} \in \mathbb{R}^{N \times K \times T}$, paving the way for enhanced signal representation.

(3) Correlation via RNNs. To analyze correlations both within the temporal and spectral dimensions of the signal, we employ two distinct Recurrent Neural Network (RNN) modules. Initially, for temporal correlation, an RNN module processes the feature tensor $\mathbf{Z}$ along the time axis. To optimize the model's efficiency, all K sub-spectral features are inputted through a single RNN layer, reducing the overall

model size. Following this, for frequency correlation, a second RNN module does the same operation along the frequency axis of the tensor. Despite their different operational domains, both RNN modules share a similar structure, beginning with a batch normalization layer applied to their inputs, followed by a BiLSTM layer [49] and a fully connected layer. A residual connection facilitates the integration of the original input with the FC layer's output, enhancing learning by allowing the flow of gradients and reducing the risk of vanishing gradients in deeper architectures. By stacking multiple such (e.g., 8) RNN modules, a more profound network capable of capturing complex temporal and spectral dependencies is constructed, with the final layer's output represented as $\mathbf{Q} \in \mathbb{R}^{N \times K \times T}$.

(4) Signal Reconstruction. The processed tensor $\mathbf{Q}$, rich with interleaved temporal and spectral information, is then reintroduced to the band-splitting process for reconstruction. At this stage, each spectral band is independently normalized and processed through an FC layer, akin to the preprocessing stage, ensuring that the unique characteristics of each band are maintained and enhanced. The spectral features from all K bands are amalgamated into a unified feature set, which undergoes a final transformation to reconstruct the acoustic signal in the time domain. This reconstruction phase is critical for translating the multidimensional spectral-temporal features back into an audible signal, effectively completing the sound enhancement and recovery process. This method leverages the strengths of RNNs in capturing sequential data patterns, offering a sophisticated approach to restoring audio signals from noisy inputs.

(5) Loss. The network is optimized by minimizing a combined loss that incorporates both frequency-domain and time-domain mean absolute error (MAE) as follows:

$$\mathcal{L}_{\text{loss}} = \|\mathbf{X} - \overline{\mathbf{X}}\|_1 + \|\text{ISTFT}(\mathbf{X}) - \text{ISTFT}(\overline{\mathbf{X}})\|_1 \tag{5.8}$$

where $\mathbf{X}$ and $\overline{\mathbf{X}}$ represent the spectra of the predicted and the ground truth signals,

respectively. This dual-component loss function ensures that the recovered signal closely matches the ground truth in the frequency spectrum and the time waveform, enhancing both spectral fidelity and temporal accuracy.

5.5 Summary

The acoustic-optical side channel, challenged by factors like thermal noise, inherent harmonics, and lossy encoding, struggles as an acoustic medium. Traditional signal processing proves nearly ineffective for sound recovery due to these complexities. Deep learning emerges as a potent solution, adept at tackling the unique challenges of using optical mice for acoustic sensing. To demonstrate the efficacy of our signal enhancer, Fig. 7.3 shows the original, recovered, and enhanced spectrograms of three short speeches, where the enhanced versions closely resemble the original except for slightly attenuated high-frequency components.

Chapter 6

Micro-Benchmark

The experiment design for evaluating JerryAttack is structured into three key components: a micro-benchmark and two distinct attack scenarios, which are introduced in Chapter 7. (1) The micro-benchmark focuses on assessing the feasibility and versatility of JerryAttack while also investigating the influence of various environmental factors, including distance, material, ambient noise, and orientation, on the system's performance. (2) The first attack scenario, voice profiling, aims to test JerryAttacks effectiveness in performing relatively simple tasks, such as gender identification, individual recognition, and digit classification, providing insights into its capabilities in basic voice analysis. (3) The second attack scenario, speech reconstruction, evaluates JerryAttack's performance on a more complex and challenging task, reconstructing speech when cooperated with joint channel information, offering a deeper understanding of the system's potential vulnerabilities. Together, these components form the basis for a comprehensive evaluation of JerryAttack's overall performance and resilience. In this chapter, we first introduce the implementation of JerryAttack and then present the details of the micro-benchmark.

6.1 Implementation

JerryAttack has been developed with a comprehensive suite of 2.1k lines, encompassing both C++ and Python code. A visual depiction of our experimental setup is presented in Fig. 6.1. Functionally, JerryAttack operates across both the client (optical mouse) and server environments.

Attack Devices and Setup. For our attack methodology, we selected the Logitech G402 optical mouse [50] as our primary device. This technique is adaptable to any optical mouse equipped with the 'pixel grabber' register, such as some models within the Logitech G-series range [51]. On the client end, the G402 model integrates an STM32 microcontroller unit (MCU) with its mouse optical sensor. To facilitate the extraction of pixel values from the sensor, we implemented modifications to the firmware within the MCU. The core segment of this customized firmware is illustrated in Algorithm 1. In practice, we used an ST-Link for direct hardware programming through the connection pins. Notably, over-the-air (OTA) updates, as discussed in 3.2, are also feasible for our attack, allowing for wireless firmware updates similar to those used by Razer mouse [52]. However, we utilized ST-Link to ensure the stability and convenience of the experiment.

On the software and hardware integration front, our setup included a Lenovo R7000P running Ubuntu 20.04 as the host system. Audio output was managed through an Adin speaker [53], which connects to the host via USB or Bluetooth. By default, the mouse was positioned in close proximity to the speaker, typically within a 20cm radius, to ensure optimal capture of audio-induced vibrations.

Server Configuration: Our neural networks are trained on a high-performance server equipped with an Intel(R) Xeon(R) Gold 6348 CPU@2.60 GHz, 256 GB of RAM, and three NVIDIA 4090 GPUs, tailored for demanding tasks such as signal enhancement, classification, and speech reconstruction. Audio data is segmented into 6-second clips with an initial rate of 3.7 kHz, upscaled to 8 kHz for processing.

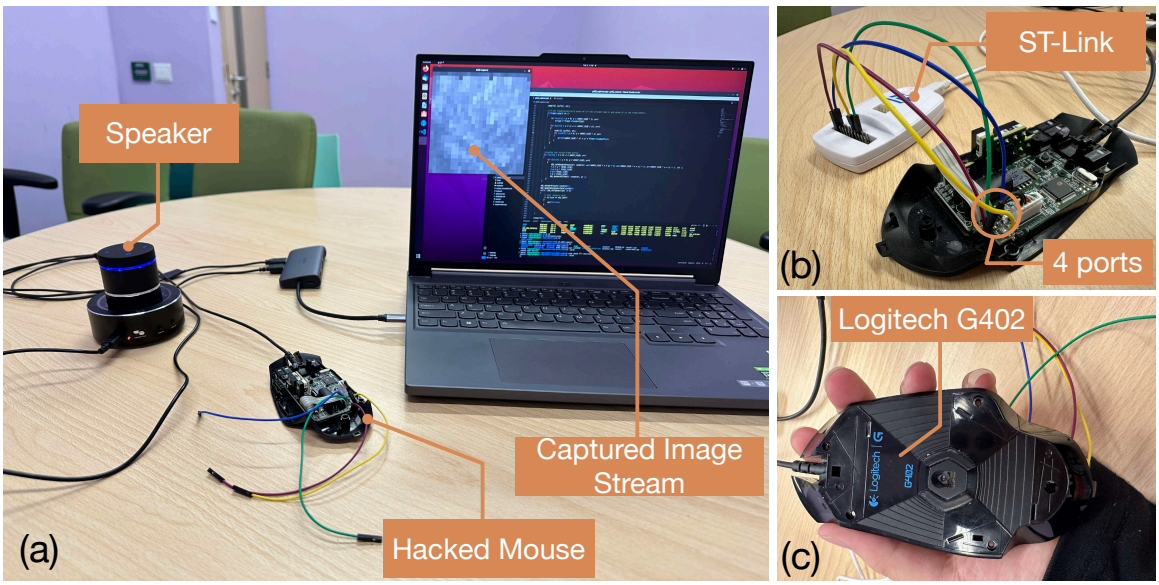


Fig. 6.1: Experimental Setup

We allocate 80% of these segments for training and 20% for validation and testing. Training employs the Adam optimizer and a cosine learning rate scheduler adjusting from 10^{-5} to 10^{-3}. We use a batch size of 28, running our dataset through 500 epochs, each lasting about 8 hours, to ensure both efficiency and accuracy.

Experimental Setup We use the AudioMNIST dataset [54], a publicly available resource that includes 30,000 audio samples of spoken digits (0-9) by 60 different speakers, to train and validate the model. The SNR is the primary criterion for sound quality, defined by the equation:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{|S_{\mathrm{gt}}|^2}{|S_{\mathrm{pt}} - S_{\mathrm{gt}}|^2} \right) \tag{6.1}$$

where S_{pt} and S_{gt} represent the predicted and ground truth time-domain acoustic signals, respectively. This metric helps quantify the effectiveness of our sound enhancement and behavior recognition system, highlighting improvements in clarity and accuracy of the audio output used for inference.

6.2 Overall Sound Quality

In this experiment, we captured sound across three distinct environments: an office, a home, and a cafe, with respective ambient noise levels of 35, 65, and 85 dBA. By default, the optical mouse was positioned 10 cm away from the speaker. For benchmarking purposes, we also recorded the data using a conventional microphone to establish a baseline comparison.

(1) Sound Quality across Scenes. Fig. 6.2 compares the sound quality outcomes from Naive Sound Recovery (NSR), Network Enhanced Recovery (NER), and Microphone-Based Recovery (MBR). Our analysis yields several key insights. Firstly, MBR achieved significantly higher SNRs in quieter environments like the home and office, registering mean SNRs of 13.41 dB and 12.09 dB, respectively. Conversely, in the noisier caf setting, the SNR noticeably decreased to about 7.72 dB. Secondly, the SNRs for NSR and NER remained more consistent (3.43 dB and 7.35 dB), showing resilience to ambient noise. This stability is attributed to the acoustic-optical channels unique ability to bypass background noise. Remarkably, the performance of NER in noisy conditions nearly matched that of the conventional microphone, validating the efficiency of JerryAttack. Third, incorporating a neural network led to an average SNR improvement of approximately 3.92 dB, increasing from 3.43 dB to 7.35 dB. This demonstrates the power of neural networks to identify the patterns caused by the inherent harmonics and other flaws of the side channel.

(2) Sound Quality across Samples. Further detailed in Fig. 6.3, the SNR distributions for NER in the three environments demonstrate a consistent pattern. More than 50% of the cases achieved an SNR of at least 6.69 dB, with the lowest recorded SNR being 4.13 dB. The 90th percentile reached up to 8.62 dB. These findings emphasize the acoustic-optical channels stability and reliability.

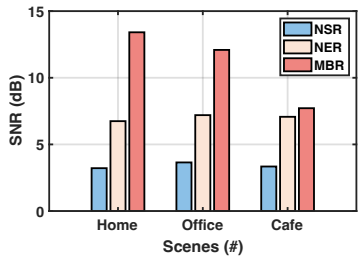


Fig. 6.2: SNR vs Scenes

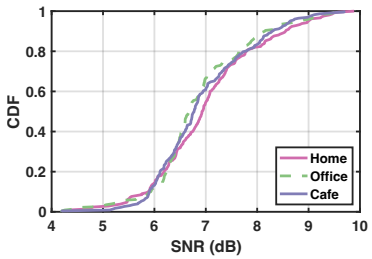


Fig. 6.3: SNR vs Samples

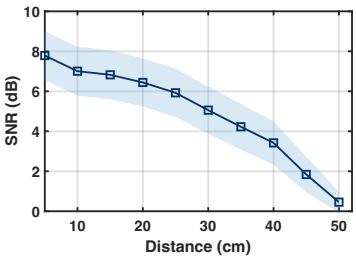


Fig. 6.4: Distance Impact

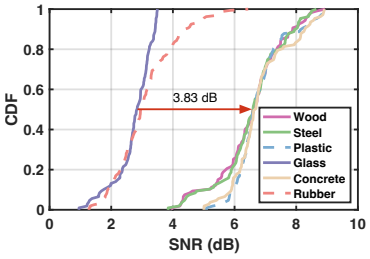


Fig. 6.5: Material Impact

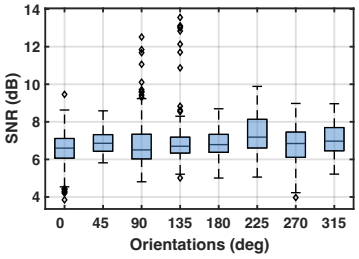


Fig. 6.6: Orient. Impact

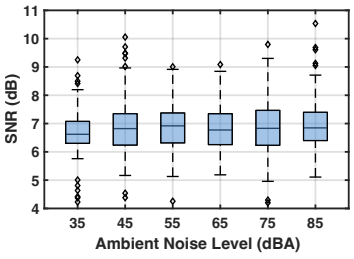


Fig. 6.7: Noise Impact

6.3 Versatility

Table 6.1 details the specifications of optical mice examined in our study, featuring 9 models across 7 manufacturers. The optical sensors used in these models include the AM010, PMW3389, and various ADNS-series sensors. Notably, the 'pixel grabber' register address may vary across different optical sensors. This highlights that varying parameters are required to execute the attack across different models successfully. Our experimental results confirm the effectiveness of JerryAttack on all tested devices,

Table 6.1: Experiment Devices and Results

Manufacturer	Model	Image Sensor	Addr.	FPS	SNR (dB)
Logitech	G402	AM010	$0x0b$	3700	7.32
Logitech	G100S	AM010	$0x0b$	3700	7.21
Logitech	G302	AM010	$0x0b$	3700	7.29
Razer	DeathAdder Elite	PMW3389	$0x64$	3600	7.31
MADCATZ	RAT3	ADNS3090	$0x40$	3500	7.19
ASUS	GT200	ADNS3050	$0x0b$	3000	7.03
ZOWIE	EC1	ADNS3060	$0x40$	3200	7.08
Tucano	Gamezone Toros	ADNS3050	$0x0b$	3000	6.95
Cool Master	Lite L Combo-Mouse	ADNS3050	$0x0b$	3000	6.98

with frame rates ranging from 3 kHz to 3.7 kHz, and the performance disparity remaining below 0.4 dB. This consistency underlines the adaptability and robustness of JerryAttack across a diverse array of optical mice.

6.4 Impact Analysis

(1) Impact of Distance: To assess the effectiveness of JerryAttack in relation to varying distances, we positioned the optical mouse at different proximities to the speaker, ranging from 5 cm to 50 cm in 5 cm increments. For each distance setting, 20 measurements were recorded. The resulting SNRs are depicted in Fig. 6.4. Notably, for distances less than 20cm, the SNR remains at a relatively high level (>6 dB). But when the distance exceeds 20cm, the SNR drops rapidly until it reaches 0.44 dB at 50cm. This trend suggests that the success of such an eavesdropping attack is heavily dependent on the optical mouse being placed near the speaker.

(2) Impact of Materials: In our exploration of the acoustic properties of various desktop materials using JerryAttack, we assessed the performance across six different surfaces: wood, steel, plastic, concrete, glass, and rubber. The results, displayed in Fig. 6.5, reveal significant variations in performance. Glass and rubber surfaces showed notably poor results, with a median SNR of only 2.56 and 2.81 dB respectively. This inefficiency of glass is attributed to its transparency, which allows light to pass through rather than reflect, effectively rendering optical mice non-functional on such surfaces. Rubber, known for its energy-absorbing properties, similarly impedes efficient sound recovery by dampening vibrations. Conversely, the remaining materials demonstrated comparably higher and consistent SNR values around 6.39 dB.

(3) Impact of Orientation: Next, we assess the impact of the orientation of the optical mouse. The mouse was rotated from 0° to 315° in increments of 45°, relative

to the loudspeaker. The resulting SNRs are shown in Fig. 6.6. Notably, there was no significant difference observed among these orientations. This lack of variability is attributed to the wavelength of the structure-borne waves being considerably larger than the dimensions of the image sensor. Consequently, it can be inferred that all pixels within the sensor receive nearly uniform influence from the acoustic vibrations, resulting in minimal differences in performance across various orientations. This finding suggests that orientation is not a critical factor that could influence performance.

(4) Impact of Background Noise: Finally, we assessed the impact of background noise on the effectiveness of the attack. The results are shown in Fig. 6.7. Consistent with previous findings, background noise had no discernible effect on the attack's efficacy. This is because the vibrations detected by the mouse are transmitted through solid materials rather than air, rendering airborne noise irrelevant. This characteristic distinguishes JerryAttack from microphone-based eavesdropping, which is susceptible to interference from sound masking devices.

Chapter 7

Attack

7.1 Attack Scenario I: Voice Profiling

In this section, we detail an attack scenario where voice profiling is used to determine a victim's identity or habits through recovered audio. To this end, high-fidelity audio recordings are not necessarily required. Instead, we deploy a ResNet [55] in conjunction with the sound enhancer for classification, as depicted in Fig.5.6(b). This setup enables us to extract valuable information from the speech.

(1) Speech Classification. We trained the classification network to classify digitized speech from the AudioMNIST dataset [54], which consists of audio samples of single-digit numbers. This model was tested to determine if the recovered sounds could be accurately classified. The results are depicted in Fig. 7.1, which includes a confusion matrix illustrating the network's performance. Overall, we achieved an accuracy of 82.2% across the ten digits. The highest accuracy, 97%, was observed for the digit one, whereas the lowest, 66%, was for the digit three. Notably, the digit three often got misclassified as four, accounting for a 20% error rate, likely due to their similar pronunciations.

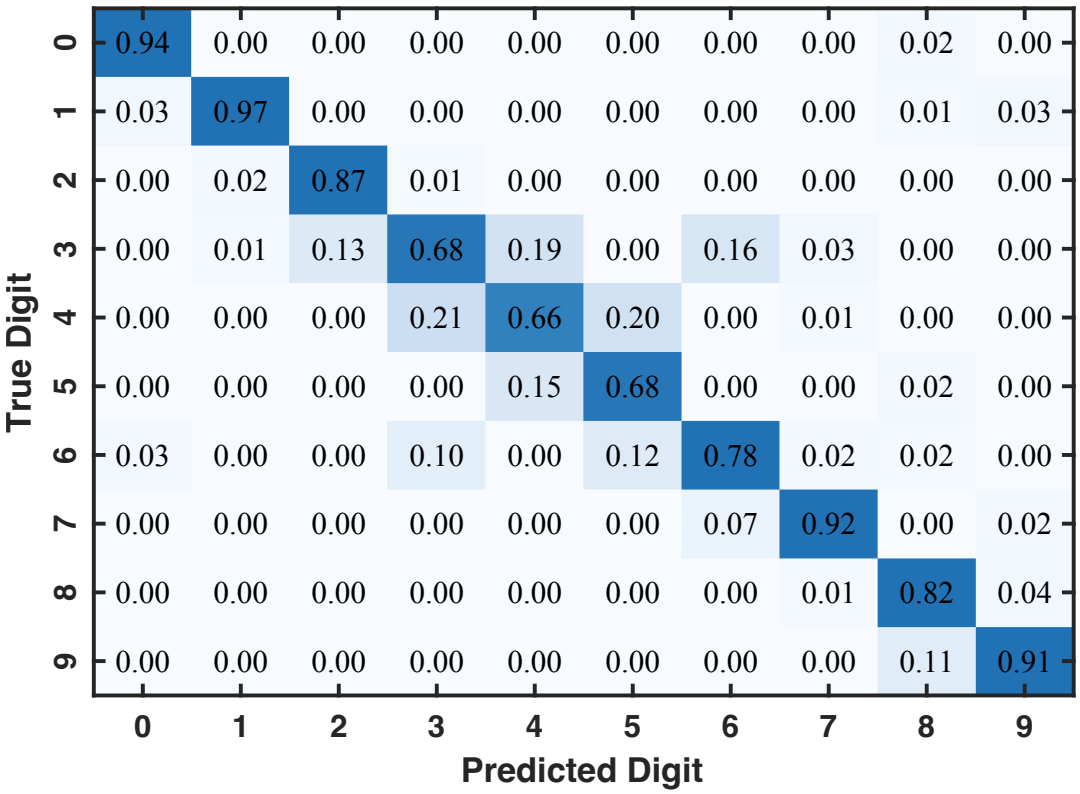


Fig. 7.1: Digit Classification Results

(2) Gender Classification. We trained our network to identify the gender of speakers across multiple languages, including English, Spanish, French, Chinese, German, and Italian. The used speech data was selected from three public datasets, including ASR corpus [56] for English speech, THCHS-30 [57] for Chinese speech, and Multilingual LibriSpeech(MLS) [58] for another 4 languages. Results are presented in Fig 7.2. With only two possible outcomesmale or femalethis task is inherently simpler than multi-category classifications. The overall accuracy achieved 99.13%, with the highest recorded accuracy reaching 100%.

(3) Human Identification. We further leveraged the same dataset for individual identification, encompassing a cohort of 48 distinct speakers. Each speaker contributed an average of 379 training speech samples and 162 test speech samples. The specific data distribution is detailed in Fig 7.2. Our network was trained to discern the identity of speakers from these samples. Consequently, we achieved an

	#	Gender	Length (s)	Train Sample	Test Sample	SNR (dB)	Gender (%)	Individual (%)
LibriSpeech-English	1	M	1505	351	150	6.54	98.07	68.67
	2		1208	281	120	8.27		90
	3		1503	350	150	7.05		75.33
	4		1503	350	150	8.14		88.67
	5	F	1511	352	151	7.31	98.99	82.12
	6		1441	336	144	8.54		89.58
	7		1503	350	150	7.3		79.33
	8		1502	350	150	7.29		79.33
MLS-Spanish	9	M	1255	292	125	7.59	98.81	87.2
	10		2225	519	222	7.15		86.49
	11		1585	369	158	7.57		82.91
	12		1658	386	165	7.41		84.24
	13	F	1470	343	147	6.95	98.57	78.91
	14		1806	421	180	6.74		73.89
	15		1585	369	158	8.45		89.87
	16		1430	333	143	8.11		86.01
MLS-French	17	M	2238	522	223	7.9	99.74	88.34
	18		1940	452	194	8.08		90.21
	19		1246	290	124	8.4		87.9
	20		2393	558	239	7.32		81.17
	21	F	2395	558	239	6.97	99.86	81.17
	22		1251	291	125	8.09		88
	23		1509	352	150	8.37		90
	24		1794	418	179	7.62		87.15

	#	Gender	Length (s)	Train Sample	Test Sample	SNR (dB)	Gender (%)	Individual (%)
THCHS30-Chinese	25	M	1271	296	127	7.79	98.11	87.4
	26		1940	452	194	7.43		82.99
	27		1310	305	131	7.45		87.02
	28		1296	302	129	7.94		88.37
	29	F	1381	322	138	7.88	100	86.96
	30		1563	364	156	7.82		85.26
	31		1330	310	133	8.32		87.97
	32		1464	341	146	8.05		87.67
MLS-German	33	M	1512	352	151	7.12	98.75	82.12
	34		1793	418	179	7.6		87.15
	35		1327	309	132	6.84		74.24
	36		1795	418	179	8.16		86.03
	37	F	1788	417	178	7.21	98.64	76.4
	38		1791	417	179	7.53		84.92
	39		1791	417	179	7.51		88.27
	40		1240	289	124	7.4		82.26
MLS-Italian	41	M	1365	318	136	7.77	99.72	82.35
	42		1568	365	156	7.29		79.49
	43		1998	466	199	5.92		67.84
	44		2323	542	232	7.5		81.03
	45	F	2001	466	200	6.95	99.85	70
	46		1572	366	157	8.25		85.99
	47		1800	420	180	8.16		87.78
	48		1449	338	144	7.63		83.33

Fig. 7.2: Gender and Human Classification Result on Multilingual Datasets

overall accuracy of 83.27%. The mean accuracies for identification across the six languagesEnglish, Spanish, French, Chinese, German, and Italianare reported as 81.37%, 83.59%, 86.22%, 86.48%, 83.01%, and 79.2% respectively.

Summary. The success of these classification tasks highlights that the sound captured and reconstructed through the acoustic-optical side channel preserves the vital attributes of the original audio. This confirms the practicality of JerryAttack as a viable tool for acoustic eavesdropping, capable of effectively extracting key information from audio data.

7.2 Attack Scenario II: Speech Reconstion

While the recovered sound maintains a relatively high SNR and retains features recognizable by neural networks, it may not always be comprehensible to humans. To evaluate speech intelligibility, we use the Short-Time Objective Intelligibility (STOI) measure, which correlates the amplitude envelopes of clean and processed speech across various frequency bands to assess intelligibility.

7.2.1 Speech Intelligibility

To examine the relationship between SNR and STOI, experiments were conducted by introducing various levels of noise to a pristine voice sample. Fig. 7.4 displays the resulting STOI scores as a function of SNR, with values ranging from 0 (completely unintelligible) to 1 (perfect intelligibility). Speech quality is categorized as 'poor' (below 0.5), 'fair' (0.5-0.6), 'good' (0.6-0.8), and 'excellent' (above 0.8) based on the STOI scores. The findings indicate that speech achieves 'good' intelligibility only when the SNR is above 7.21 dB. Additionally, the figure shows the SNR distribution (depicted with a red line) of speeches recovered from the Multilingual dataset, predominantly ranging between 6 and 9 dB, aligning with 'fair' and 'good' quality. This limited intelligibility is largely attributed to inherent harmonics in the system that replicate voice frequencies, complicating the clarity of the speech.

To improve intelligibility, we propose a joint attack that combines the side channel

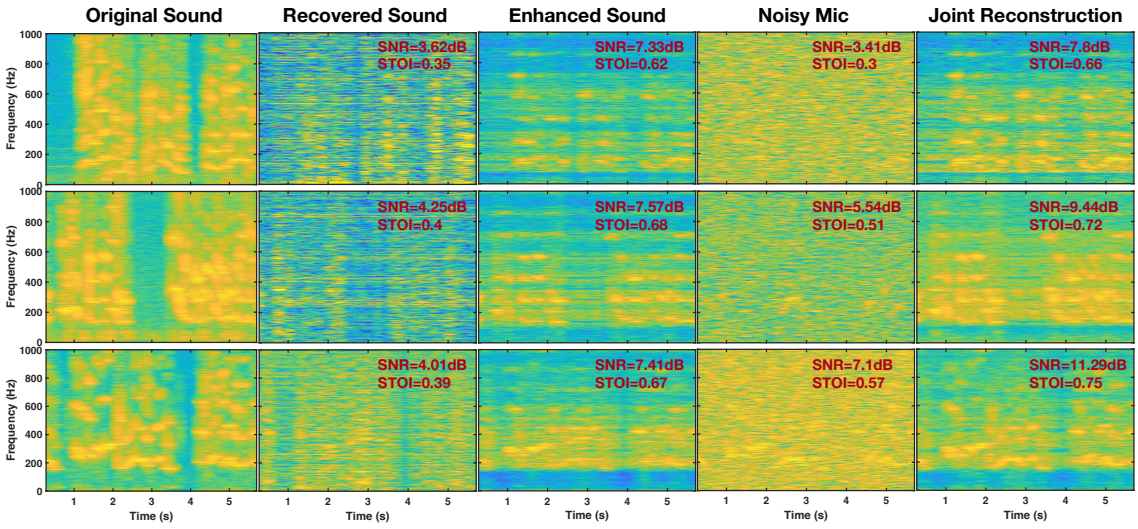


Fig. 7.3: Spectrogram Comparison of Speech Processing. Each row represents a different speech sample. The first column displays the original sound's spectrogram. The second and third columns illustrate the spectrograms after processing by the naive recovery algorithm and the sound enhancer, respectively. The fourth column shows the sound recovered using a microphone at a distance. The final column presents the results from the joint reconstruction. All samples have been well-marked the SNR and STOI with respect to original sound.

with a direct eavesdropping channel, such as a microphone. In this scenario, the attacker strategically positions a microphone at a significant distance from the victim, potentially attaching it to an exterior wall or using a directional microphone array to capture sound from afar. Due to this distant placement, the SNR of the direct channel is often too low for effective speech reconstruction on its own. We concurrently capture the same speech through both the side and direct channels and then reconstruct the speech using a joint neural network, as depicted in Fig. 5.6(c). Each channel processes a spectrogram and outputs an enhanced version. The goal is to merge the enhanced outputs from both channels, aiming for their combined result to closely align with the ground truth spectrum. This strategy leverages the strengths of both audio sources, potentially overcoming the limitations of each channel when used independently and achieving a more precise reconstruction of the speech.

7.2.2 Results

We assess the reconstruction efficacy from three aspects:

(1) Visual Representation. To clarify the mechanics of the attack, Fig. 7.3 shows spectrograms of speeches captured by the side and direct channels in the third and fourth columns, respectively. Notably, the direct channel's spectrogram is heavily obscured by noise, which nearly overwhelms the entire spectrum. The joint reconstruction results, shown in the fifth column, demonstrate a significant improvement in SNR compared to the side channel alone. This improvement is due to the side channel's ability to provide a complementary spectrum that enhances the signal components and suppresses the noise elements from the direct channel.

(2) Speech Quality. To assess the reconstructed speech quality, we evaluated 300 speech samples from the previously mentioned multilingual dataset using both the direct channel (D) and the side channel (S). The STOI distribution of the reconstructed results is illustrated in Fig. 7.5. Key insights include: Firstly, using the side

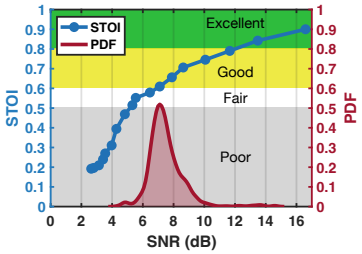


Fig. 7.4: STOI vs SNR

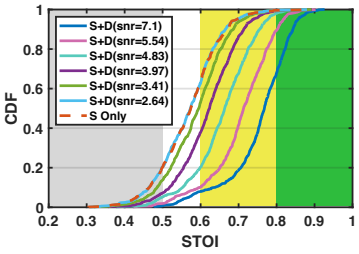


Fig. 7.5: Intelligibility

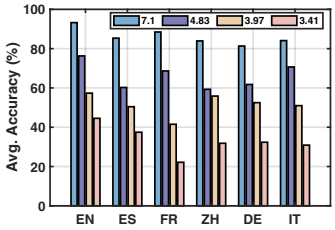


Fig. 7.6: Speech2Text

channel alone, we achieved a median STOI of approximately 0.58, indicating 'fair' intelligibility. Secondly, the direct channel provides an additional boost in intelligibility when the SNR exceeds 2.64 dB, improving speech quality from 'fair' to 'good,' even with a weak direct channel SNR of 3.97 dB. Generally, the direct channel increases STOI by 0.04 per 1 dB increment in SNR. Finally, when the SNR of the direct channel surpasses 5.54 dB, speech intelligibility reaches 'excellent' levels, marking a great enhancement in speech clarity.

(3) Speech-to-Text Accuracy. We also used the AssemblyAI Speech-to-Text Recognition API [59] to gauge the intelligibility of the reconstructed speech indirectly. As depicted in Fig. 7.6, a similar trend emerges across different languages where a direct channel SNR of 7.1 dB helps achieve an average accuracy of up to 90%. Conversely, a 4.83 dB SNR yields an accuracy slightly above 60%, showcasing the positive impact of higher SNR levels on speech recognition accuracy.

Summary. The speech reconstructed solely from the side channel reaches a 'fair' level of intelligibility. However, integrating a direct channel significantly enhances intelligibility, as it helps verify and reduce the impact of inherent harmonics, thereby improving the clarity and accuracy of the reconstructed speech.

Chapter 8

Conclusion and Future Works

In this chapter, we reflect on the key findings of this thesis and explore potential future directions for expanding and improving the work. The conclusion will summarize the contributions of JerryAttack, an innovative acoustic eavesdropping technique that repurposes optical mice, and highlight the significance of this research in uncovering security vulnerabilities in common devices. Following this, the future works section will present two promising research directions: first, expanding the attack scenarios to accommodate dynamic mouse movement, and second, exploring robust countermeasures to mitigate the security risks posed by such attacks. These discussions aim to provide a clear roadmap for further research in enhancing mobile security practices and developing more resilient systems against unconventional threats.

8.1 Conclusion

This thesis introduces a novel eavesdropping technique that repurposes optical mice as covert acoustic surveillance devices. By utilizing the optical sensor embedded in a standard optical mouse, we demonstrated how this widely available device could be adapted to capture vibrations and convert them into audio signals, a technique

we have termed 'JerryAttack'. This approach leverages the high frame rates and sensitivity of optical mouse sensors, which are typically designed for precise movement tracking, to detect subtle vibrations caused by sound waves. The findings from this research reveal the potential security risks posed by everyday computing peripherals when repurposed for malicious purposes, emphasizing the need for heightened awareness and preventive measures.

The idea of using non-traditional devices for acoustic eavesdropping has been explored in various forms in prior research. Techniques such as exploiting smartphone gyroscopes, Wi-Fi signals, and high-speed cameras to capture sound by analyzing mechanical vibrations have been studied extensively. However, these methods often require specialized equipment or access to devices that are closely monitored for security breaches. For example, visual vibrometry using high-speed cameras has shown promise in recovering sound from video footage of vibrating objects, but these approaches typically involve expensive equipment and are limited by the visibility of the target object.

In contrast, JerryAttack offers a novel, cost-effective solution by repurposing ubiquitous optical mice as acoustic sensors. This method presents several key advantages over existing techniques as follows:

Stealth and Ubiquity: Optical mice are widely used in both personal and corporate environments, making them ideal for covert operations. Their common presence in everyday settings means they are unlikely to arouse suspicion, unlike more specialized surveillance equipment.

High Frame Rate for Accurate Sound Recovery: The optical sensors in mice, with frame rates often exceeding 3 kHz, can capture fine surface vibrations caused by sound waves. This high frame rate allows JerryAttack to recover sound with a mean SNR comparable to standard microphones, especially in noisy environments.

Robustness to Environmental Noise: Unlike traditional acoustic sensors that

capture air-borne sound waves, JerryAttack focuses on structure-borne sound waves transmitted through solid surfaces. This unique characteristic makes it highly robust to environmental noise, effectively filtering out background noise typically carried by air, and ensuring clearer sound recovery in diverse settings.

Low-Cost Implementation: JerryAttack requires minimal hardware, only simple firmware modifications of the mouse's microcontroller that could be implemented via the OTA method, making it a cost-effective alternative to more specialized and expensive surveillance technologies.

Enhanced Performance Through Machine Learning: The integration of advanced machine learning techniques for signal enhancement and classification, significantly improves the intelligibility of reconstructed speech. This capability makes JerryAttack highly effective for practical eavesdropping applications.

The broader implications of these findings underscore a significant privacy risk, as commonly used devices like optical mice are not typically scrutinized for security vulnerabilities. This work not only introduces a new attack vector but also highlights the urgent need for robust countermeasures. By advancing our understanding of how everyday devices can be repurposed for surveillance, this thesis makes a significant contribution to the field of mobile security. It emphasizes the importance of vigilance and innovation in developing protective measures against emerging threats. The insights gained from this research open new avenues for securing devices that are not traditionally viewed as security risks, thereby helping to protect personal and corporate privacy in an increasingly interconnected world.

8.2 Future Works

8.2.1 Expanding Potential Attack Scenarios: Transitioning the Mouse from Static to Dynamic

In this thesis, JerryAttack requires the optical mouse to remain static during the eavesdropping process for effective sound recovery. However, expanding the attack scenario to allow the mouse to move dynamically, as it would during normal usage, presents an intriguing challenge. The main difficulty arises from the fact that the optical sensor in the mouse, modeled as a multi-sampler system, has samplers that correspond to fixed physical positions. When the mouse moves, the baseline light intensity bias p_i for each sampler changes continuously, which can significantly impact the performance of sound recovery. To address these challenges and make JerryAttack feasible in dynamic scenarios, we propose the following potential solutions:

Leveraging Mouse Displacement Registers: Modern optical mice store displacement data in two registers, Δx and Δy, which track movement along the x and y axes, respectively, over short intervals (e.g., 1ms, based on the mouse's polling rate). By utilizing this real-time displacement data, it may be possible to infer the current positions of the samplers and dynamically adjust the p_i array. This method would allow us to compensate for the shifting baseline intensity biases caused by mouse movement, ensuring more accurate sound recovery. Future work could focus on developing algorithms to incorporate displacement data into the sound recovery model, enabling continuous recalibration of the sensor positions in real time.

Separating High-Frequency Vibrations from Low-Frequency Movements: In dynamic scenarios, distinguishing between high-frequency, small-amplitude vibrations caused by structure-borne sound waves (the target of JerryAttack) and low-frequency, large-amplitude vibrations from normal mouse movement is crucial. The key challenge is isolating these subtle sound-induced vibrations from the noise created

by mouse movement. Future research could explore advanced signal processing techniques or machine learning models capable of filtering out low-frequency movement noise while preserving the high-frequency vibrations containing acoustic information. Successfully extracting these high-frequency vibrations would allow sound recovery even when the mouse is in active use.

These approaches, leveraging real-time displacement data to infer the p_i array and separating vibrations based on frequency and amplitude, offer promising solutions for extending JerryAttack to dynamic usage scenarios. Future research in these areas could greatly improve the flexibility and effectiveness of the attack, enabling it to function even when the mouse is being used normally. This would significantly expand the range of potential attack vectors and increase the real-world applicability of the method.

8.2.2 Countermeasures Exploration

To address the security risks associated with optical mice being repurposed for acoustic eavesdropping, one of the future research topics should focus on developing and implementing a range of hardware-based and software-based countermeasures. These countermeasures aim to mitigate the vulnerabilities identified in this thesis and enhance overall mobile security practices.

Hardware Modifications: The 'pixel grabber' register within optical mice, originally intended for debugging and sensor integration testing, remains inactive during standard operations. This feature, although seemingly harmless, provides a potential entry point for unauthorized access, acting as a 'backdoor' for eavesdropping. To mitigate this vulnerability, manufacturers should consider disabling or securely encapsulating this testing interface in commercial products. This could involve hardware-level changes that render the register inaccessible or reprogramming the firmware to prevent unauthorized access to pixel data. By eliminating this backdoor, the risk of

data breaches through optical mice can be significantly reduced.

Operating System Enhancements: Operating systems (OS) have the potential to monitor and flag unusual device behaviors, such as atypical data transmissions between the optical mouse and the host computer. Future work should focus on enhancing the OS's ability to recognize these anomalies and alert users to potential security threats. Developing advanced monitoring algorithms that can detect unusual data patterns indicative of eavesdropping attempts would provide an additional layer of security. These algorithms could employ machine learning techniques to continuously learn and adapt to new forms of eavesdropping, improving detection accuracy and reducing false positives.

Physical Countermeasures: Employing anti-vibration materials on mouse pads or other surfaces can significantly reduce the transmission of sound vibrations to optical mice. This measure is particularly crucial in high-security environments, where sensitive conversations or activities occur frequently. By absorbing vibrations that could otherwise be detected and analyzed, these materials create a physical barrier to potential eavesdropping attempts. Future research should investigate the effectiveness of various materials and designs to enhance their protective capabilities against acoustic surveillance.

Awareness and Training: Raising awareness about the security risks posed by seemingly innocuous devices like optical mice is essential to strengthening overall mobile security practices. Training programs and informational campaigns should be developed to educate users, IT professionals, and manufacturers about the potential for optical mice to be exploited for eavesdropping. Emphasizing the importance of regular device audits, firmware updates, and cautious use of peripheral devices can help mitigate risks and foster a more security-conscious environment.

By combining these specific countermeasures, a comprehensive defense strategy against acoustic eavesdropping via optical mice can be developed. This multi-layered ap-

proach addresses both the hardware and software vulnerabilities identified in this study, providing robust protection against this unconventional but significant security threat in the context of mobile security.

References

[1] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. Gyrophone: Recognizing speech from gyroscope signals. In *Proc. of USENIX Security*, 2014.

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

[3] S Abhishek Anand and Nitesh Saxena. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *Proc. of IEEE S&P*, pages 1000–1017, 2018.

[4] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth Pathak, Guoming Zhang, and Xiuzhen Cheng. Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1757–1773. IEEE, 2022.

[5] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *NDSS*, volume 2020, pages 1–18, 2020.

[6] Andrew Kwong, Wenyuan Xu, and Kevin Fu. Hard drive of hearing: Disks that eavesdrop with a synthesized microphone. In *2019 IEEE symposium on security and privacy (SP)*, pages 905–919. IEEE, 2019.

[7] Qingsong Yao, Yuming Liu, Xiongjia Sun, Xuewen Dong, Xiaoyu Ji, and Jianfeng Ma. Watch the rhythm: Breaking privacy with accelerometer at the extremely-low sampling rate of 5hz. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1776–1790, 2024.

[8] Tao Ni, Guohao Lan, Jia Wang, Qingchuan Zhao, and Weitao Xu. Eavesdropping mobile app activity via radio-frequency energy harvesting. In *Proc. of USENIX Security*, 2022.

[9] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 130–141, 2015.

[10] Pengfei Hu, Yifan Ma, Panneer Selvam Santhalingam, Parth H Pathak, and Xiuzhen Cheng. Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 11–20. IEEE, 2022.

[11] Pengfei Hu, Wenhao Li, Riccardo Spolaor, and Xiuzhen Cheng. mmecho: A mmwave-based acoustic eavesdropping method. In *Proc. of the ACM Turing Award Celebration Conference-China*, pages 138–140, 2023.

[12] Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani B Srivastava. Uwhear: Through-wall extraction and separation of audio vibrations using wireless signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 1–14, 2020.

[13] Chuyu Wang, Lei Xie, Yuancan Lin, Wei Wang, Yingying Chen, Yanling Bu, Kai Zhang, and Sanglu Lu. Thru-the-wall eavesdropping on loudspeakers via rfid by capturing sub-mm level vibration. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–25, 2021.

[14] Yan Long, Pirouz Naghavi, Blas Kojusner, Kevin Butler, Sara Rampazzi, and Kevin Fu. Side eye: Characterizing the limits of pov acoustic eavesdropping from smartphone cameras with rolling shutters and movable lenses. In *Proc. of IEEE S&P*, pages 1857–1874, 2023.

[15] Mariko Akutsu, Yasuhiro Oikawa, and Yoshio Yamasaki. Extract voice information using high-speed camera. In *Proc. of Meetings on Acoustics*, volume 19, 2013.

[16] Yohei Fuse, Yusuke Yasumi, and Tetsuya Takiguchi. Sound recovery using vibration mode of an object in video. *Open Access*, 2018.

[17] Khatuna Zannat Mim, Abdullah Arafat Miah, and Mohiuddin Ahmad. Extraction of sound signal from tiny vibrations in motion magnified video using optical flow. In *Proc.of IEEE IC4ME2*, pages 1–4, 2019.

[18] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Fredo Durand, and William T Freeman. The visual microphone: Passive recovery of sound from video. 2014.

[19] Mark Sheinin, Dorian Chan, Matthew O'Toole, and Srinivasa G Narasimhan. Dual-shutter optical vibration sensing. In *Proc. of IEEE/CVF CVPR*, pages 16324–16333, 2022.

[20] Robert Lea, Samuel Reiter, and Uriel Tar. Auditory recovery from silent high-speed video using optical flow. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(10):1622–1631, 2015.

[21] Author Placeholder. Visual to auditory: Sound recovery from visual stimuli using cross-modal generative adversarial networks. In *Proc. of IEEE ICASSP*, pages 1234–1238, 2018.

[22] Ben Nassi, Yaron Pirutin, Raz Swisa, Adi Shamir, Yuval Elovici, and Boris Zadov. Lamphone: Passive sound recovery from a desk lamp's light bulb vibrations. In *Proc. of USENIX Security*, 2022.

[23] Ben Nassi, Raz Swissa, Jacob Shams, Boris Zadov, and Yuval Elovici. The little seal bug: Optical sound recovery from lightweight reflective objects. In *IEEE Security and Privacy Workshops (SPW)*, pages 298–310, 2023.

[24] John Cunnison Catford. *A practical introduction to phonetics*. Oxford University Press, 2001.

[25] Avago Technologies. ADNS-3050 Optical Mouse Sensor Datasheet. `https://www.mouser.com/datasheet/2/678/avagotechnologies_ADNS-3050-1217285.pdf`.

[26] Jianwei Yu and Yi Luo. Efficient monaural speech enhancement with universal sample rate band-split rnn. In *Proc. of IEEE ICASSP*, pages 1–5, 2023.

[27] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics*, 33(4):79:1–79:10, 2014.

[28] Tuck Wah Ng and Kar Tien Ang. The optical mouse for harmonic oscillator experimentation. *American Journal of Physics*, 2005.

[29] Tuck Wah Ng. The optical mouse as a two-dimensional displacement sensor. *Sensors and Actuators A-physical*, 2003.

[30] Marcel Tresanchez, Tomas Palleja, Merce Teixido, and Jordi Palacin. The optical mouse sensor as an incremental rotary encoder. *Sensors and Actuators A-physical*, 2009.

[31] Jordi Palacin, I. Valganon, and R. Pernia. The optical mouse for indoor mobile robot odometry measurement. *Sensors and Actuators A-physical*, 2006.

[32] Sebastian Ullrich, Jakob T. Valvoda, and Torsten Kuhlen. Utilizing optical sensors from mice for new input devices. http://ftp.informatik.rwth-aachen.de/Publications/AIB/2006/2006-15.pdf, 2006. RWTH Aachen University.

[33] Xun Yi, Zhuoling Mo, Zhili Zhao, Haixin Zou, and Yousheng Chen. Research on anti-eavesdropping system of multi-source ultrasonic in space. In *Proc. of ICEACE*, 2023.

[34] José Carlos Pedro and Nuno Borges Carvalho. *Intermodulation distortion in microwave and wireless circuits*. Artech House, 2002.

[35] STMicroelectronics. St-link/v2 in-circuit debugger/programmer for stm8 and stm32. https://www.st.com/en/development-tools/st-link-v2.html, 2024.

[36] Jan Bauwens, Peter Ruckebusch, Spilios Giannoulis, Ingrid Moerman, and Eli De Poorter. Over-the-air software updates in the internet of things: An overview of key principles. *IEEE Communications Magazine*, 58(2):35–41, 2020.

[37] Connect the headphones to two devices simultaneously (multi-point connection). https://www.sony.com/electronics/support/wireless-headphones-bluetooth-headphones/wh-1000xm5/articles/00249936, 2024.

[38] Logitech mx anywhere 3s. https://www.logitech.com/en-us/products/mice/mx-anywhere-3s.html?srsltid=AfmBOop5uTbOqCn_gSl_FoXl3siKNgzwUe_FYhG4XlmN8hvq7I6pCFW_, 2024.

[39] sensors. https://sensor.fyi/sensors/, 2024.

[40] pmw3320-am010-hacking. https://www.overclock.net/threads/pmw3320-am010-hacking.1587477/page-2, 2024.

[41] Mouse Specs. Flawless mouse sensors. https://mousespecs.org/sensors/, 2024.

[42] Pixart imaging. https://www.pixart.com/, 2024.

[43] Pramod Subramanyan, Nestan Tsiskaridze, Kanika Pasricha, Dillon Reisman, Adriana Susnea, and Sharad Malik. Reverse engineering digital circuits using functional analysis. In *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1277–1280, 2013.

[44] Jean Mandel and Louis Brun. *Mechanical waves in solids*, volume 222. Springer, 2014.

[45] WOLFRAM. Built to last: Understanding earthquake engineering. https://blog.wolfram.com/2011/03/18/built-to-last-understanding-earthquake-engineering/, 2011.

[46] Alessandro Foi, Sakari Alenius, Vladimir Katkovnik, and Karen Egiazarian. Noise measurement for raw-data of digital imaging sensors by automatic segmentation of nonuniform targets. *IEEE Sensors Journal*, 7(10):1456–1461, 2007.

[47] Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[48] Jianwei Yu, Yi Luo, Hangting Chen, Rongzhi Gu, and Chao Weng. High fidelity speech enhancement with band-split rnn. *arXiv preprint arXiv:2212.00406*, 2022.

[49] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[50] Logitech G402 Hyperion Fury FPS Gaming Mouse. https://www.logitechg.com/en-hk/products/gaming-mice/g402-hyperion-fury-fps-gaming-mouse.910-004070.html, 2014.

[51] Logitech G Gaming Mice. https://www.logitechg.com/en-hk/products/gaming-mice.html, 2023.

[52] Razer atheris firmware updater — rz01-02170. https://mysupport.razer.com/app/answers/detail/a_id/13048/~/razer-atheris-firmware-updater-%7C-rz01-02170, 2023.

[53] Adin Speaker. https://www.aliexpress.com/item/1005002224994045.html, 2023.

[54] Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE/CVF CVPR*, pages 770–778, 2016.

[56] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Proc. of IEEE ICASSP*, pages 5206–5210, 2015.

[57] Dong Wang and Xuewei Zhang. Thchs-30: A free chinese speech corpus. *arXiv:1512.01882*, 2015.

[58] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020.

[59] Speech-to-text. https://www.assemblyai.com/products/speech-to-text, 2023.