Fast Detection of Handheld Phone-Distracted Driving by Sensing the Driver's Hand-Grip

Ruxin Wang, *Student Member, IEEE*, Long Huang, and Chen Wang

***Abstract*—Handheld phone distraction is the leading cause of traffic accidents. However, few efforts have been devoted to detecting when phone distraction happens, which is a critical input for taking immediate safety measures, not only related to the driver but also important to surrounding vehicles, pedestrians, and vehicle networks. This work proposes a fine-grained handheld phone-use monitoring system, which detects the start of phone-distracted driving activities and further distinguishes different handheld phone-use scenarios, which enables estimating the impacts on traffic safety to take measures accordingly. Specifically, the proposed system emits periodic ultrasonic pulses to sense if the phone is being held in hand or placed on support surfaces (e.g., seat and cup holder), and the unique signal interference resulted from the contact object's damping, reflection, and refraction is analyzed based on the sounds that return to the microphone. We derive the short-time Fourier transform to describe such impacts and develop a CNN-based binary classifier to distinguish phone use between handheld and handsfree. Moreover, the system leverages the embedded inertial sensors to capture the phone's motion dynamics and recognize specific handheld phone distractions (e.g., holding the phone for calling). The system periodically samples the driver's phone-use status and use an error correction window to correct misclassified samples. As a result, the start, end, and duration of each handheld phone distraction activity can be obtained. Extensive experiments show that our system achieves 99.5% accuracy in recognizing handheld phone-use instances and a 0.76-second median error in detecting the start of a handheld phone distraction.**

***Index Terms*—Distracted driving, driver behavior, handheld device, vehicle safety.**

I. Introduction

USING a handheld device while driving is a dangerous behavior. The driver can be impacted by all three types of distractions from the phone (i.e., visual, manual, and cognitive), which increases the risk of crashing by up to 23 times [1]. Every year, around 660,000 drivers attempt to use phones while driving, and 14% of fatal crashes involve phones [2]. Though law enforcement and insurance penalty policies help raise public awareness and lower car accidents, they achieve limited effects. Reports show that handheld device distractions cause 1.6 million crashes annually in the U.S. In 2020 alone, over 50,098 people were injured or killed in car accidents related to cell phone use [3], [4]. Since the COVID-19 pandemic, a 17% increase in driver phone use has been found, because more people attempt to take Zoom calls, read Instagram messages, or text while driving [5]. More efforts are still urgently needed to reduce the driver's handheld phone use, and incorporating human factors into Vehicle-to-Vehicle (V2V) communications is key to enhancing traffic safety.

There has been active work on using the phone itself to prevent distractions. By recognizing when the phone user is driving, the phone could automatically turn on the do-not-disturb mode and prohibit phone use (e.g., delaying messages and routing calls to voice mail). For example, cellphone handovers and signal strength variations can be used to recognize a phone in a moving car [6], [7]. To further distinguish whether the phone user is a driver or a passenger, researchers have developed in-vehicle localization methods, which estimate whether the phone is closer to the driver or passenger seat [8], [9]. However, most users refuse to disable phone services completely while driving though acknowledging the dangers [10]. They may have concerns about missing important notifications and calls during long-distance driving. They may also prefer to use the less distracting and legally allowable handsfree phone operation, ask the passenger to read/reply or pull over to a safe area to cope with emergencies. Thus, preventing a driver from reaching out to the phone is more practical and effective than disabling all phone services for an entire trip. More specifically, we need to know when a driver holds the phone to immediately address the handheld phone distraction.

This work aims to capture the precise timing (e.g., start, end, and duration) of each distracted driving activity, which is a critical input to numerous safety systems for taking immediate safety measures. For example, knowing when the driver picks up the phone, all Apps could be shut down by the phone at once except for emergency calls. And the nearby automobiles (especially self-driving cars) could be notified to take precautionary measures. Additionally, such information could be used to determine who is at fault in a car accident or personalize insurance rates. The prior work to monitor the driver's phone use mainly relies on monitoring the display on/off, the phone

Manuscript received 31 March 2023; revised 7 January 2024 and 29 February 2024; accepted 1 March 2024. Date of publication 12 March 2024; date of current version 15 August 2024. This work was supported in part by LA-BoR under Grant LEQSF(2020-23)-RD-A-11 and in part by NSF under Grant CNS-2155131. An earlier version of this paper was presented at the IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems, 2021 [DOI: 10.1109/MASS52906.2021.00058], which achieved the initial success of using active acoustic sensing to detect handheld phone-use. The review of this article was coordinated by Dr. Tao Dusit Niyato. *(Corresponding author: Chen Wang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by LSU Institutional Review Board (IRB) under Application No. IRBAM-21-0756, and performed in line with the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects.

The authors are with the School of Electrical Engineering, Computer Science, Louisiana State University, Baton Rouge, LA 70803 USA (e-mail: rwang31@lsu.edu; lhuan45@lsu.edu; chenwang1@lsu.edu).

Digital Object Identifier 10.1109/TVT.2024.3374589

lock status, the phone lifting action [11], [12], and the phone dynamics related to distracting phone activities (e.g., calling and texting) [13]. But based on such indirect phone-use indicators, these methods are hard to determine the detailed timing of each distracted driving instance. Moreover, they have limited abilities to cover the diverse phone distraction scenarios and are not sufficiently reliable in the practical in-vehicle environment, which is noisy.

It is noted that our system starts to work when the phone user has been identified as the driver by existing methods [8], [9]. The phone's speaker actively plays ultrasonic pulses to sample the phone-use status periodically. The acoustic signals traveling on the device surface could be uniquely damped, reflected, and refracted by a gripping hand. Due to the palm skin and unique contact areas, the resulting signals reaching the phone microphones are different from the scenarios when the phone is placed on a seat, cup holder, pocket, or phone mount. Based on that, our system accurately detects when the driver grabs/holds/drops the phone. To further recognize the fine-grained handheld device use, we leverage the embedded inertial sensors to study the phone's motions resulted from specific phone uses, such as reading, calling, scrolling, and texting. Because no additional hardware is required, users of our system can continue to use their existing cars without technological restrictions.

In particular, we develop a learning-based distracted driving monitoring system, which continuously monitors the phone-use status and captures the fine-grained distracted driving activities related to handheld device use. We utilize active acoustic sensing to recognize the surface of the object that contacts the phone and inertial sensing to recognize the phone-use motion dynamics. The microphone data is used to derive the short-time Fourier transform, which describes the unique time-frequency characteristics of the signal interference caused by the gripping hand or a support surface in the vehicle. Based on that, we develop a CNN-based binary classifier to distinguish whether the phone is handheld or handsfree. To recognize the specific handheld phone use, we derive statistic motion features from the accelerometer and gyroscope data and develop an SVM-based classifier for analysis. After getting a sequence of phone-use status samples, we further utilize an error correction filter to address the misclassified samples and output the start, end, and duration of each specific handheld phone distraction activity.

Our contributions can be summarized as follows:

- This work proposes a continuous phone-use monitoring system to detect the driver's handheld device distractions, which enables many prompt safety measures and is a significant addition to vehicular networks.
- We develop an active acoustic sensing method to recognize the phone's contact surface. Specifically, we derive the short-time Fourier transform from the sensing sound to describe the characteristics of the phone's contact surface and use a CNN-based binary classifier to distinguish handheld phone use from various handsfree scenarios.
- We utilize inertial sensors to capture the unique phone motion dynamics associated with different phone-use scenarios and develop an SVM-based multi-class algorithm for fine-grained handheld phone distraction recognition.
- We design error correction schemes to process the most recent phone-use estimation samples, which facilitates detecting the start of a handheld phone-use activity and its end in noisy in-vehicle environments.
- Extensive experiments with different phone/car models and participants show that our system accurately recognizes the fine-grained phone distraction scenarios and estimates their timing information.

II. Related Work

There has been a rising interest in monitoring unsafe driving behaviors. The vehicle's speed, acceleration, and deflection angle can be estimated from the phone sensor data to recognize the dangerous driving behaviors [14], [15]. To improve the drivers' awareness of their driving habits, Chen et al. further classify the abnormal driving behaviors among different vehicle maneuver types by using phone sensors [16]. Xu et al. focus more on the driver's attention and use Doppler shifts of the phone audio signals to sense the inattentive driving events, such as eating, drinking, and turning back [17]. But none of these works could effectively address the handheld device distraction, one of the leading causes of traffic accidents.

The existing research efforts to prevent handheld phone distraction are on differentiating the phone user to be the driver or the passenger based on its in-vehicle location. Yang et al. propose a relative-ranging system, which sends acoustic signals in a programmed sequence from the stereo car speakers and captures the time differences of their arrivals at the phone to determine whether it is closer to the driver seat or the passenger seat [8]. Wang et al. use the phone's inertial sensors to measure its centripetal acceleration when the vehicle makes turns. By comparing to a reference point, they estimate whether the phone is on the right or left side of the car [9]. Chu et al. release the requirement of additional infrastructure and rely entirely on the phone sensors to differentiate the micro-activities between the driver and the passenger, such as with which foot to enter the car first and along which direction to fasten the seat belt [18]. There are also infrastructure-free methods to recognize the phone user during driving, which localize the phone based on its motion dynamics or camera views [19], [20]. However, these methods are far from satisfactory to address the handheld device distraction, as they cannot detect when the distracted driving happens to take proper safety measures right away, which requires capturing the interaction between the phone and the driver.

There are several solutions to capture phone-driver interactions based on cameras. For example, Chuang et al. monitor the driver's gaze direction using the phone front camera [21]. A recent work installs multiple cameras in the car to capture the interaction between the driver and the phone, which complements the blind spots of each single camera [22]. However, these vision-based methods are limited by light conditions, camera view angles, or high installation overhead.

We propose to monitor phone-driver interactions based on sensing the gripping hand. There have been several studies on detecting the grips of mobile devices. For example, the phone's rotations, vibrations, and touch events can be measured by

inertial sensors and the touchscreen to infer the user's phone-use postures, such as with which hand (or both hands) to hold the device and which finger (e.g., index finger and left/right thumb) to operate on the screen [23], [24]. These are motion-driven approaches. Ono et al. attach a pair of vibration motor and receiver on the phone case to recognize the user's hand postures [25], and Kim et al. achieve similar functions based on acoustic sensing [26]. Both methods use a support vector machine as the classifier. However, the above studies all assume the phone is already in the user's hand and then recognize the type of phone grip. Few of them investigate distinguishing a handheld phone from that placed on many other surfaces such as a table, seat, and phone mount. Furthermore, it is unknown whether they could work in the in-vehicle environment, which suffers from complex acoustic noises and vibration noises related to the engine, road conditions, and the wind. More importantly, none of them is able to demonstrate the user-phone interaction monitoring and capture the phone-grip start, end, and duration.

III. Background and System Architecture

A. Distracted Driving Instance

This work aims to reduce the impacts of distracted driving caused by handheld phone use. We define a *distracted driving instance* as the handheld phone-use activity, which begins from the driver's hand reaching the phone and ends until the phone is dropped off. This entire period is subject to the combination of all three types of distractions (i.e., visual, manual, and cognitive). Compared to single-distraction-type activities, such as checking the navigation system (visual), making a handsfree phone call (cognitive), and eating/driving (manual), handheld phone use is the most dangerous and is prohibited by law. Therefore, one efficient and direct way to prevent handheld phone use by a driver is to detect when and how long the driver holds the phone and then disable or restore the phone services accordingly. It also facilitates sending early warnings, notifying nearby automobiles, assisting law enforcement, and personalizing insurance rates.

B. Sensing the Gripping Hand Acoustically

We leverage the acoustic signals that propagate on or near the phone surface to sense the gripping hand or other objects that come in contact with the phone. In particular, we use the phone speaker to send ultrasonic signals for sensing periodically. The signal traveling on the phone case would be interfered with by the driver's gripping hand or the support surfaces on which the phone is placed, such as the seat and center console. The resulted sound reaching the phone microphone contains useful information that could describe how differently the original signal is damped, reflected, and refracted by the gripping hand and the support surfaces. Fig. 1 illustrates how the acoustic signal interacts with the driver's hand, where the sound recorded by the microphone includes the damped direct-path signal, the reflected signals, and the air-borne refracted signals (near-surface). These signal components are mainly determined by the material, area, and pressure of the contact surface. Because the hand's skin, geometry, and gripping strength are distinct from any support surface in the car, the gripping hand can be distinguished by acoustic sensing.

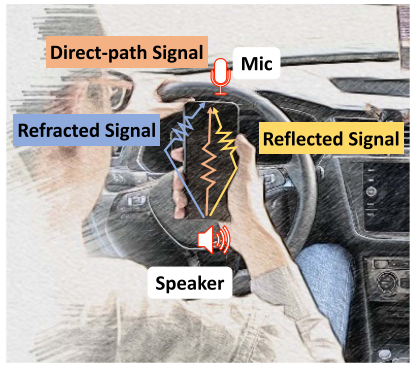


Fig. 1. Illustration of acoustic signal interaction with driver's hand.

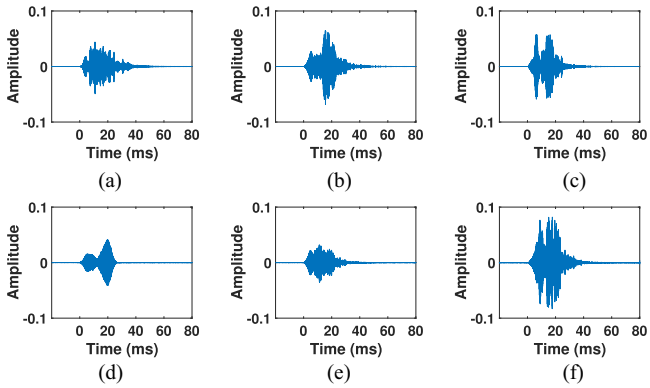


Fig. 2. Acoustic response of different phone placement. (a) Handheld (b) Center console (c) Cup holder (d) Pocket (e) Seat (f) Phone mount.

To show the feasibility, we play an ultrasonic chirp sound using the phone's speaker, which sweeps from 18 kHz to 22 kHz in 25 ms. Fig. 2 shows the waveforms of the recorded sounds when the phone is on six different support surfaces in a car, including a hand. We observe that the microphone-recorded sound is distinguishable in the waveforms among all six phone placement scenarios, which shows the potential of differentiating the gripping hand from the other phone placement scenarios. Moreover, while the sensing signal sweeps along the frequency, its amplitude is reinforced or suppressed with different scales, and at the same frequency, the amplitude change is also unique for each support surface. This phenomenon reflects the frequency diversity of the sound to sense the various support surfaces, which motivates us to use the sound with rich frequencies rather than a single frequency to achieve robust sensing. It is noted that we use the 2D time-frequency images shown in Fig. 5 instead of directly utilizing the captured audio data in the time domain to differentiate between handheld and handsfree phone use status. The reason is that our system identifies phone-use status by sensing the smartphone's contact surfaces' materials, and different materials impact the signal's frequencies differently and presenting distinct patterns at time-frequency domain.

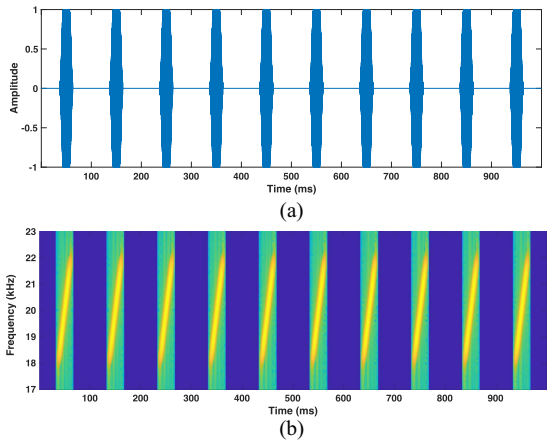


Fig. 3. Periodic ultrasonic pulse signals for sensing. (a) Waveform of the periodic ultrasonic pulse signal. (a) Spectrogram of the periodic ultrasonic pulse signal.

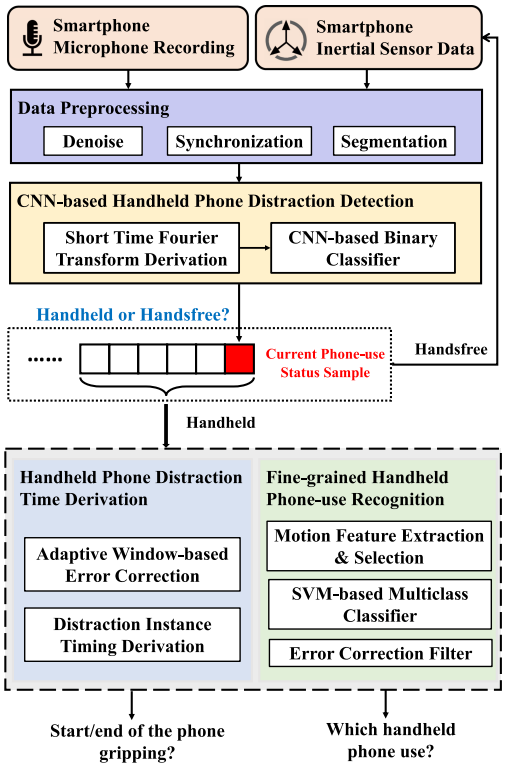


Fig. 4. Architecture of our system.

C. Challenges

We also face some challenges when using acoustic signals to sense the gripping hand. Specifically, we find that the microphone keeps receiving sounds for a long time after the sensing signal stops at 25 ms, as shown in Fig. 2. These sounds are mainly environmental reflections, which are much stronger in the vehicle's confined space than indoor or outdoor scenarios. They also heavily rely on the in-vehicle phone locations and should not be used for analysis. One exception is the in-pocket scenario, because the fabric of the pocket is a good sound-absorbing material, which significantly damps the outward sounds and reduces the

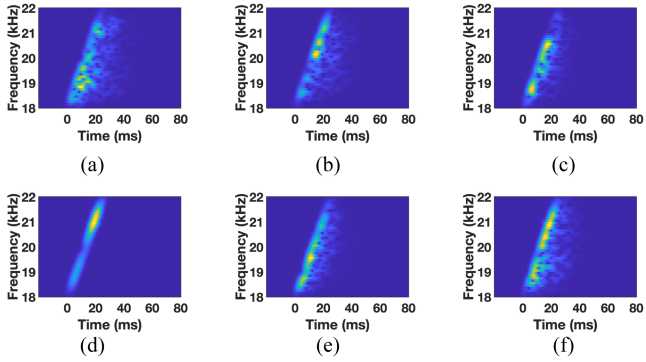


Fig. 5. Short-time Fourier transform of different phone use statuses. (a) In hand (b) On center console (c) On cup holder (d) In pocket (e) On seat (f) On phone mount.

echoed back sounds. As a result, the in-pocket waveform is more different from that of the other five scenarios. In comparison, the handheld scenario is harder to be differentiated from the center console, cup holder, and phone mount scenarios. We thus rely on deep learning to recognize the handheld scenario.

Furthermore, different people's hands may exhibit slight differences when holding the phone, due to their individually unique hand shapes and gripping strengths. Even the same person may hold the device slightly differently when texting, scrolling, and calling. These variances need to be considered and addressed. Furthermore, our acoustic system must work under a noisy in-vehicle environment, where the background noises result from the different road conditions, driving speeds, and car audio sounds. Additionally, to accurately estimate the start and end of a distracted driving instance, our system needs to handle the classification errors and the noisy transient states when the phone is being grabbed or dropped off.

D. System Design

The goal of our work is to eliminate the handheld phone-use distraction based on detecting the gripping hand. To achieve the goal and address the above challenges, we develop a phone-use monitoring system, which sends unique signals for sensing and uses a deep learning-based algorithm to recognize the various in-vehicle phone-use statuses. Our system can work with existing phone localization methods [8], [9] to effectively eliminate handheld phone distraction. For example, our system could start after the phone user is identified as the driver. Alternatively, our system could continuously sense the phone use status, and once it is in hand, the phone localization method further confirms if this is the driver's hand.

1) Sensing Signal Design: The sensing signal is used to interact with the object that is in contact with the phone and capture its characteristics in the acoustic domain to differentiate whether the phone is in the driver's hand or on a support surface of the vehicle. Based on our feasibility study and challenge analysis in Sections III-B and III-C, we design the sensing signal with repetitive ultrasonic chirps. Fig. 3(a) and (b) illustrate the original waveform of the sensing signal and its spectrogram. In particular, each pulse signal lasts for a short period (i.e., 25 ms),

and every two pulses are separated by a stop period (i.e., 75 ms). The short pulses suffer less from the echo sounds which usually last for much longer, and the stop time reduces the interference between adjacent pulses. Only the 25 ms pulse sound is used for analysis.

Moreover, each pulse signal is designed to sweep from 18 kHz to 22 kHz to leverage the rich frequency information, which facilitates capturing more characteristics of the object in contact with the phone. Besides, this high-frequency range is not impacted much by the in-vehicle noises, which are mainly on lower frequencies. The sounds in these frequencies are also demonstrated to be hardly audible and not invasive [27]. Furthermore, we apply a Hamming window to smooth the two ends of each pulse to reduce the spectral leakages and the speaker hardware noises caused by the sudden frequency jumps at the start and the end of each pulse. As a result, the pulse signals could sample the phone-use status ten times per second to detect the hand-grip.

2) System Flow: The architecture of our system is shown in Fig. 4, which takes the phone's microphone recording and inertial sensor data (e.g., accelerator and gyroscope) as input. *Data Pre-processing* is performed first to calibrate the data for analysis. It applies a bandpass filter to remove the noises outside the sensing signal's frequency range and synchronizes the microphone's data by referring to the original audio. Based on that, we can find the start and end of the pulse signal to obtain the pulse segment, which is one sample of the phone-use status. We apply similar denoising schemes to the inertial sensing data and further segment it based on the timing information of the obtained audio segments.

The core of our system consists of three components. The *CNN-based Phone-use Status Recognition* processes the pulse sound and recognizes the phone-use status at the current sampling point. Specifically, we derive the short-time Fourier transform (STFT) from the pulse sound to describe the time-frequency characteristics of the contact surface in the acoustic domain. The 2D STFT is input to the CNN-based binary classifier to differentiate the phone-use status between handheld and handsfree.

Based on a series of the most recent phone-use status samples, the *Handheld Phone Distraction Timing Derivation* further finds the start and end of each complete distracted driving activity. In particular, we develop an adaptive window-based error correction method to examine the current status sample and correct the classification errors based on the results of the recent samples. We further use a threshold-based method to determine whether a distracted driving activity occurs. If the phone-use status toggles back and forth too quickly, it is unlikely to be from human action and is corrected. Once a distracted driving instance is confirmed, the system would take safety measures immediately, such as sending early warnings to the driver and notifying nearby self-driving vehicles to take precautions via V2V networks. If the phone is detected to be dropped off, the phone services can be restored and the vigilance levels of surrounding vehicles can be reduced.

After identifying a handheld phone distraction activity, the *Fine-grained Handheld Phone-use Recognition* further enables recognizing detailed handheld phone use, such as reading, calling, scrolling, and texting. This is important to quantify the impact of distracted driving. For example, texting is shown to be more dangerous than calling [28]. In particular, we extract the phone's motion dynamics features from both the microphone and the inertial sensor data segments. The two domain features, after Principal Component Analysis (PCA), are feed into an SVM-based multi-class classifier for handheld scenario determination.

IV. Method Design

A. Data Pre-Processing

After obtaining the data from the microphone buffer and inertial sensors (e.g., accelerator and gyroscope), we first pre-process it for denoising, synchronization, and segmentation. In particular, for the data from the microphone, we design a bandpass filter with the 18 kHz-to-22 kHz passband to reduce the noises outside of the sensing signal's frequency range. For example, the engine, road, and wind noises can be removed, which are mainly on frequencies below 6 kHz [29], and the car audio sound impact could be reduced. After denoising, we can focus better on the sensing signal changes caused by different contact objects.

Next, we run a synchronization scheme to locate the pulse signal in the microphone data precisely. Specifically, we iteratively shift the microphone data $\hat{x}$ and compute its cross-correlation with the original pulse signal x. The shift length leading to the maximum cross-correlation coefficient indicates the time delay between the two signals as expressed by

$$delay = \underset{m}{argmax} \sum_{n=0}^{N-m-1} \hat{x}(n+m)x(n), \tag{1}$$

where m is the number of samples to shift. After subtracting this delay, we can find the start and end of the sensing sound by referring to the original pulse signal. The resulted 25 ms pulse segment is used for further analysis. We further normalize the amplitude of the pulse segment to be within the range $[-1, 1]$. It is important to note that the pulse signal is generated every 100 ms, and the 75 ms microphone audio that comes after the pulse is mainly the echo sounds. This audio part is heavily affected by the phone's in-vehicle location and is discarded.

Furthermore, most phones are embedded with two microphones for noise cancellation (e.g., one at the top and one at the bottom). By using the two acoustic channels, we can leverage the spatial diversity to capture more characteristics of the contact object. Therefore, we use the two mics to independently sense the contact object and integrate their results to make a decision, which reduces the errors of each single mic and is robust.

We apply similar denoising schemes, including a highpass filter (3 Hz), to the accelerometer and gyroscope data along three axes. This helps filter out low-frequency noise readings from inertial sensors caused by sudden jerks and turns while driving. This ensures minimal impact on the accuracy of our fine-grained handheld use recognition. And further segment them with the microphone data segments' timing into chunks of the same time length (e.g., 0.1 s). These inertial sensor segments are used

with the microphone data segment together for distinguishing different handheld phone-use scenarios.

B. Short-Time Fourier Transform

We derive the STFT from the microphone data segment to describe the characteristics of the contact object in the acoustic domain. STFT presents the frequency spectrum along time, which captures how each spectral point of the signal is interfered with by the hand or a support surface in the vehicle. In particular, we use a sliding window with the length 480 samples to compute the Discrete-Time STFT (DT-STFT) of the pulse signal, which results in a 2D image. The value of each image pixel at sample m and frequency f is expressed by (2), where $w(n)$ is a window function.

$$DT\text{-}STFT(m, f) = \sum_{n=-\infty}^{\infty} \hat{x}(n)w(n-m)e^{-j2\pi fn} \quad (2)$$

Though the derived DT-STFT covers the microphone's all frequencies, which span from 0 to 24 kHz, we crop the image to only focus on the pulse signal's frequency range from 18 kHz to 22 kHz. Fig. 5 shows the feasibility of using the DT-STFT image of the pulse signal to differentiate six different scenarios. We can observe that the DT-STFT images show distinct pixel patterns among all the contact objects. For example, the in-hand scenario presents several strong spectral points around 19 kHz, while the center console, pocket, and seat show lower amplitudes around this frequency. When the phone is on the center console, cup holder, pocket, and phone mount, the received pulse signal has great amplitudes between 20 kHz and 22 kHz. In comparison, the gripping hand suppresses the pulse signal significantly on these frequencies. The reason is that the impact on the pulse signal depends on the contact object's material, contact area, and pressure, which may reinforce the signal at some frequencies but suppress it at others. Our next step is to utilize a deep learning algorithm to discriminate the handheld phone use from most handsfree scenarios using the DT-STFT images.

C. CNN-Based Binary Classifier

We resize all the DT-STFT images into a fixed size and process them using a binary classifier based on Convolutional Neural Network (CNN). CNN is widely used to analyze images by learning their patterns. To recognize the gripping hand based on DT-STFT images, we develop a CNN-based binary classifier with three convolution layers and one fully connected layer, which is a CNN structure widely used on mobile devices [30]. The output dimensions in each layer are tuned to reduce the processing time while ensuring accuracy. Specifically, the dimensions of the output can be calculated as

$$dimensions = \left(\tfrac{m-k+2d}{l}+1\right) \times \left(\tfrac{m-k+2d}{l}+1\right) \times t \quad (3)$$

where m, k, l, d and t are the input image size, the kernel size, the step length, the number of padding applied and the number of filters.

The detailed structure of our CNN classifier is shown in Table I. In particular, the dimensions of the normalized input image is 150×150. The convolutional kernel size is 3×3 and the pooling kernel size is 2×2. The step length is set as 1, the number of padding applied is set as 0, and the number of filters is 32. The dimensions after the first convolution operation is $148 \times 148 \times 32$ as computed by the above equation. Since the kernel size of the pooling layer is 2, the dimension after the first pooling operation is $74 \times 74 \times 32$. We keep the same configuration for the rest of the convolution and pooling layers. At the end of the model, we utilize the softmax function to normalize the network output and obtain a probability for each class as the decision confidence or CNN score. Since our system's inputs are images, and the computing resources needed for learning of image features is a crucial factor to consider. We thus choose the efficient and low-cost optimizer - Adam [31], [32]. We use sparse categorical cross-entropy as the model's loss function since we expect class labels to be provided as integers instead of one-hot encoding ones.

TABLE I
STRUCTURE OF OUR CNN-BASED BINARY CLASSIFIER

Layer	Output Shape	Param #
Input: short-time Fourier transform	(150, 150, 3)	0
Conv2D + RecLineU	(148, 148, 32)	896
Max Pooling 2D	(74, 74, 32)	0
Conv2D + RecLineU	(72, 72, 32)	9248
Max Pooling 2D	(36, 36, 32)	0
Conv2D + RecLineU	34, 34, 32)	9248
Max Pooling 2D	(17, 17, 32)	0
Flatten	(9248)	0
Dropout	(9248)	0
Dense	(128)	1183872
Dense_1	(60)	7740
Dense_2	(2)	122
Output: Probability in [0, 1]	(1)	0

Our CNN-based algorithm performs the binary classification to discriminate the handheld and handsfree phone uses, which consists of two phases. During the training phase, we involve a number of people in collecting the handheld and handsfree phone-use instances. Moreover, the various handheld phone-use activities are considered to cover various scenarios when the user holds the phone still, taps/swipes on the phone screen, and hold the phone close to face (e.g., making phone calls). It is important to note that these phone-use activities generate sounds and cause the handheld status to be unstable. Our system does not rely on these sounds to recognize handheld phone use, because they differ significantly among people and activities. These acoustic noises mainly reside at low frequencies and are suppressed by our bandpass filter. Though the phone can be used differently in the driver's hand, our CNN algorithm can still distinguish them from the handsfree scenarios, as the phone is consistently in the user's hand, which is discernible from other contact objects. Additionally, we train two CNN models for Mic 1 and Mic 2 of the phone, for analyzing the contact object from two acoustic channels.

During the testing phase, the DT-STFT images of the testing pulse sound are input to the two CNN models to process independently. The CNN scores of the two models are integrated to make the classification decision. This result is the phone-use status sampled by one sensing pulse. We compare the binary classification performance when using SVM and CNN models.

We find that using the CNN model outperforms SVM. We use the term FLOPS to measure the operational requirements of a network model and to indicate the computing power of hardware like GPUs, providing an estimate of a model's training time on such hardware. The 150 MFLOPS is calculated using a *FLOPs calculator with tf.profiler for neural network architecture written in TensorFlow 2.2+* [33], when the model is trained on a MacBook Pro (13-inch, 2017, Two Thunderbolt 3 ports). The memory usage during training usually amounts to 5 MB on the CPU. Generally, there are no specific training requirements; most commercial laptops are capable of training our model.

D. Handheld Phone-Use Monitoring

The accurate classification obtained with each sensing pulse is the basis for monitoring phone use and detecting distracted driving instances. But monitoring phone use in practical in-vehicle scenarios is more challenging. Even the classification error of a single sample could come at a tremendous cost. We continue to study practical phone-use monitoring and correct sample errors to cope with the false positives and false negatives in classification results.

Our system is designed to sample the phone-use status ten times per second. The phone-status monitoring result is a sequence of labels between *handheld* and *handsfree*, based on which the system decides when the user grabs or drops off the phone. We design an adaptive window-based error correction filter to process the label sequence based on the *flip-and-merge* rule. The adaptive window starts from the first sample of the current instance and compares it with its adjacent next sample. If their labels are the same, the window grows its size by one and examines the next consecutive sample. This recursion continues until the sample status changes. The current window extracts a sample chunk, and its size W is recorded. Then, the above process repeats to find the next chunk.

The *flip-and-merge* rule further determines each chunk to be an error or a valid chunk with two thresholds th_1, th_2 ($th_1 < th_2$), where a valid chunk represents a complete or a partial instance. The intuition is that when a driver uses a phone, the duration can not be too short (even for checking time). If $W \geq th_2$, the chunk is determined to be a valid chunk. If $W < th_1$, the entire chunk is considered to be misclassified because the phone status toggles back and forth too fast, and the labels of its all samples are flipped. This chunk after correction is merged to its closest valid chunk. If $th_1 < W < th_2$, we need to examine the labels of its two valid neighbor chunks, v_{pre} and v_{next}. If $v_{pre} = v_{next}$, we consider this current chunk to be erroneous, so it is flipped and merged with its neighbor chunks. If $v_{pre} \neq v_{next}$, we keep the label of the current chunk and merge it with the valid neighbor chunk that has the same label. As a result, the handsfree and handheld instances are obtained. Especially, the handheld instance is detected, if the prior chunk is a handsfree instance and the current chunk size grows larger than th_2 (it is not necessary to wait to obtain an entire chunk). The first sample of the current chunk then captures the handheld instance start, and its end is determined when the next chunk is confirmed to be a handsfree instance. Empirically, we use 0.5 s and 0.8 s for th_1 and th_2.

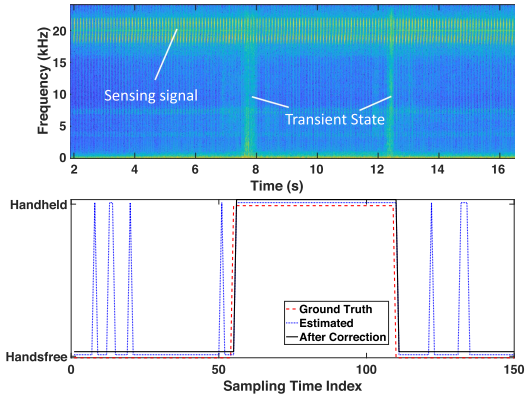


Fig. 6. Phone-use monitoring and classification error correction.

Fig. 6 illustrates the phone-use monitoring when a driver grabs the phone for 5 seconds and then drops it off. The top figure presents the spectrogram of this process, where the ultrasonic pulses periodically sense the phone-use status, and the transient state sounds (i.e., phone-grab and drop-off actions) show the main signal powers at lower frequencies. The bottom figure illustrates the phone-use status monitoring results. We observe that though some samples are mistakenly classified, they can be corrected by our adaptive window-based filter. The resulted phone-use status sequence is close to the ground truth curve. From this monitoring result, we can detect the complete distracted driving instance as well as determine its start, end, and duration.

E. Fine-Grained Handheld Phone-Use Recognition

After the phone is recognized to be in the driver's hand, it is important to further know how the phone is specifically used, such as calling, reading, texting, and scrolling. These different user-phone interactions present different distraction levels and exert different impacts on the future vehicle status. For example, texting distracts the driver much more than handheld-phone calling and causes the driver to respond to traffic even slower [2], [34], [35]. Knowing such detailed handheld phone use is critical to estimate/quantify the distraction impacts and manage the traffic accordingly.

While acoustic sensing is efficient in differentiating handheld status from handsfree, we find it has limited capability to further distinguish the various handheld phone-use scenarios, such as reading, scrolling, and calling. Fig. 8(a) illustrates the difficulty of using acoustic sensing alone to analyze handheld scenarios. The overlapped clusters of different handheld scenarios, though showing some feasibility, are not sufficiently reliable. This is because the contacting surface keeps moving when the user interacts with the phone, which cause significant noises to our acoustic method. Inertial sensors are good at capturing the phone motions caused the user's handheld phone-use. Fig. 7 presents the unique phone motions captured by the accelerometer and the gyroscope when the user is reading, calling, scrolling or texting on the phone. We observe that the inertial sensor data exhibit

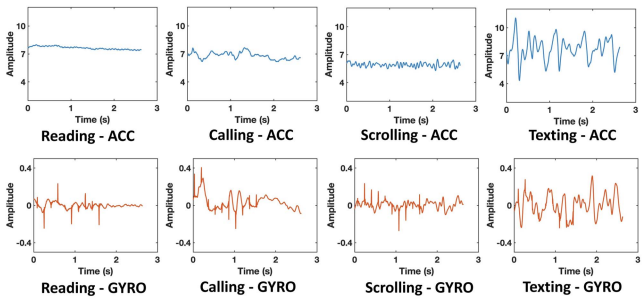


Fig. 7. Motion dynamics captured by accelerometer and gyroscope of different handheld scenarios.

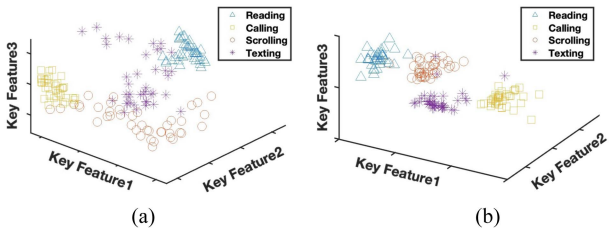


Fig. 8. Illustration of using audio and inertial sensors' features for fine-grained handheld recognition. (a) Audio Features. (b) Inertial Sensors' Features.

distinct patterns among different handheld scenarios, demonstrating the unique motion dynamics of the device during phone use. Fig. 8(b) further presents the effectiveness of using inertial sensors to distinguish different handheld scenarios, which are separated in isolated clusters. Therefore, we propose to enhance our system by incorporating inertial sensing and developing a sensor fusion method. It is important to note that fine-grained handheld phone-use recognition is performed after a handheld distraction activity has been detected. We still use the active acoustic sensing data to examine how the gripping hand interferes with the periodically emitted pulse signal. Differently, we use the accelerometer and gyroscope data to passively monitor the phone's motions resulted from human-phone interactions. We observe that the ultrasonic sounds emitted by the phone are not strong enough to affect the data from inertial sensors. Therefore, we derive unique features to describe detailed interactions between the driver and phone across two domains and use a Support Vector Machine (SVM) classifier for fine-grained handheld phone-use recognition.

Specifically, we derive the Mel-frequency Cepstral Coefficients (MFCC) from each microphone data segment and the statistical features from each inertial sensor data segment. The statistic features include Max, Min, Variance, Standard Derivation, Range, Skewness, Kurtosis, and Quartiles. We further use Principal Component Analysis (PCA) for feature selection and find the key features that better capture the unique phone motion dynamics. In particular, we recursively eliminate one feature from the feature set and use PCA to compute the sum of weights/coefficients/loading scores for the remaining features. The feature sets achieving the highest weight sums are selected, whose clustering performances are further compared to determine the key feature set. We then feed the key features into the

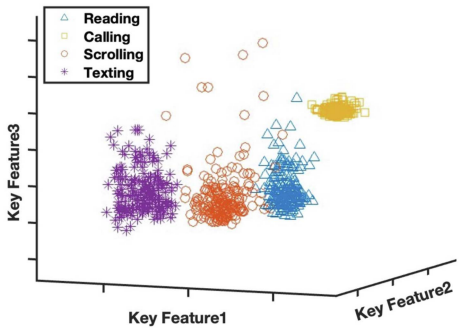


Fig. 9. Using cross-domain features to recognize fine-grained handheld phone use.

SVM multi-class algorithm for handheld phone-use recognition, and the output is each phone-use status sample (e.g., for 1 ms).

Furthermore, we design an error correction filter to address the mis-classified phone-use status samples. The intuition is that the shifting between different phone uses can not be too quick. We use the majority vote to process the phone-use status samples in a sliding window and correct the minority samples. We find that a window of a short time length (e.g., 1 s) is sufficient for the error correction, given that it takes a user 1.67 s to type a word seconds [36]. Fig. 9 illustrates using three derived key features to distinguish four handheld phone-use scenarios (before error correction). We observe that these different handheld phone uses are separated in different clusters. While calling is the more distinctive from other phone uses, reading, calling and scrolling, are close to each other. The reason is that the three phone uses are associated with the same hand-grip pose. But they can still be differentiated based on the minute motion dynamic differences, and our error-correction filter can further improve the fine-grained handheld phone-use recognition performance.

V. Performance Evaluation

A. Experimental Setup

To evaluate our system, we develop an experimental platform based on Android, which periodically sends ultrasonic pulse signals and records the stereo sounds and inertial sensor data simultaneously. The sensing signal is programmed to play for 25 ms, followed by a 75 ms pause, and this pattern repeats periodically. We use this platform to collect data from four phone models, Samsung Galaxy S20, Samsung Galaxy S8, Motorola Moto G8, and Google Pixel2, and the data is processed offline. Samsung Galaxy S8, Motorola Moto G8, and Google Pixel2 run Android 9.0, Samsung Galaxy S20 runs Android 12.0, and the microphone sampling rate is set to 48 kHz. We also test two vehicle models, Nissan Rogue (Car A) and Volkswagen Tiguan (Car B). We recruited eighteen participants (6 females and 12 males, age 21∼33) for data collection. The authors were trained to act as drivers, while the participants sat on the passenger side. Tests were conducted to assess hand shape and size, body fat ratio, and phone use behaviors, including grip strength, typing, swiping behaviors, and grip pose. As shown in Fig. 10, each participant was asked to use the phone in eleven

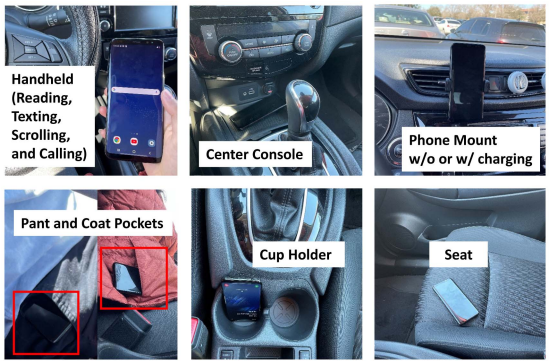


Fig. 10. Eleven experimental scenarios in the vehicle.

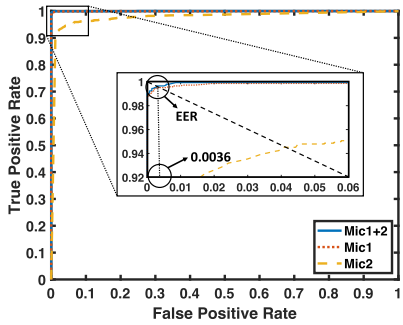


Fig. 11. Distracted detection performance of our system.

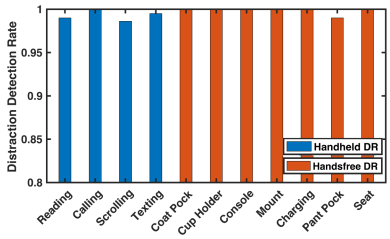


Fig. 12. Distraction detection under different phone-use contexts.

scenarios, including four handheld phone uses (i.e., holding the phone still or reading, texting, scrolling, and calling) and seven handsfree scenarios (i.e., in a coat pocket, pant pocket, cup holder, center console, phone mount, phone charging on phone mount and seat). For each scenario, the participant was asked to re-grab or reposition the phone 40 times for two main reasons: 1) to enlarge dataset, and 2) to include behavioral inconsistency, unfixed phone orientation and phone location differences.

The overall performance is evaluated based on eighteen participants, eleven scenarios, car A and Samsung S8. We apply half of the data for training and the rest for testing. We also investigate the various impact factors based on four participants and eleven scenarios. In particular, the device model and the car model impacts are studied. Moreover, different in-vehicle environments where the practical in-vehicle noises caused by the engine, road conditions, and traffic are involved. Additionally, the impact of the car audio (e.g., radio sounds) is studied. Furthermore, we monitored four participants' phone-distracted driving activities and fine-grained handheld phone use, in which each participant was asked to use the phone by grabbing it 40 times from the seat, center console, cup holder, phone mount, and pocket for an hour of monitoring. Due to safety reasons, the front passenger performed the experiments.

Evaluation Metrics: We use True Positive Rate (TPR) to measure the proportion of actual positives (e.g., handheld distraction instances) that are correctly identified by our system. It is calculated as the number of successful recognitions divided by the total number of tests. False Positive Rate (FPR) is used to assess the proportion of actual negatives (e.g., handsfree instances) that are incorrectly classified as positives by our system. It is calculated as the number of unsuccessful recognitions divided by the sum of tests. Equal Error Rate (EER) is where TPR equals FPR. Additionally, we use Distraction Detection Rate (DR) to represent the percentage of instances (e.g., handheld distraction instances) detected by the system.

B. Phone Distraction Detection Performance

1) Handheld Vs. Handsfree: The ROC curves of our system to detect phone distraction are presented in Fig. 11. We find the system achieves a high TPR and low FPR to distinguish *handheld* from *handsfree*. In particular, when integrating the two microphones, our system achieves 99.7% TPR and 0.5% FPR, and the EER is 0.36%. The results are very promising as the system correctly differentiates the handheld and handsfree scenarios, regardless of how the driver uses the phone and who holds the phone. The results also indicate that our system is effective in practical usage. Furthermore, we find Mic 1 performs better than Mic 2. The reason is that Mic 1 is at the top of the phone, far from the bottom speaker. Compared to Mic 2, which is close to the speaker, Mic 1 receives sounds that travel across the phone case and interact better with the contact object to capture its characteristics.

2) Phone-Use Contexts: Next, we investigate how the system distinguishes eleven phone statuses between *handheld* and *handsfree*. Fig. 12 presents the DR in four handheld and seven handsfree scenarios. We observe that our system performs well for all eleven scenarios, obtaining a mean 99.6% DR. For example, calling performs the best among the four handheld scenarios with a 100% DR. The DRs of reading, texting, and scrolling are slightly lower, which are 99.0%, 99.5%, and 98.61%, respectively. "The reason is that hand movements in these three scenarios cause noise and slightly unstable contact between the phone and hand. Moreover, reading and scrolling are coexisting behaviors that are hard to differentiate. For the seven handsfree scenarios, except for the pant pocket, which performs with a 99.0% DR, the other six scenarios achieve a 100% DR and are recognized as handsfree. These results indicate that our system successfully detects handheld phone distractions based on their contact with phone.

3) Individual Difference: We also study how the system performs across different users. Fig. 13 presents the DR for four types of phone use (i.e., handheld and handsfree) across eighteen users. We observe that the system accurately detects phone distractions for all participants, with an average DR of 99.7%. Moreover, more than half of the users achieve a DR of

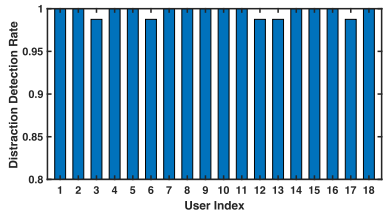


Fig. 13. Distraction detection performance for different users.

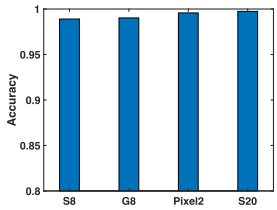


Fig. 14. Impact of different devices for phone distraction detection.

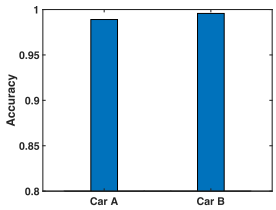


Fig. 15. Impact of car models for phone distraction detection.

100%, with the lowest DR being 98.8%. The results show that our system can work for different users regardless of their unique hand geometry and gripping strengths.

C. Impact Factor Study

1) Device Models: We now investigate the impacts of device models. Our participants were asked to use four different phones in Car A, and the above eleven types of phone statuses were collected. Fig. 14 shows the classification accuracy for each device. We observe that all four devices accurately distinguish handheld phone use from handsfree. In particular, Samsung Galaxy S20 performs the best with 99.7% accuracy. The performances of Google Pixel 2, Samsung Galaxy S8, and Motorola G8 are slightly lower, which are at 99.6%, 98.9%, and 99.0%, respectively. The results indicate that our system performs well with a range of Android phone models. Considering the adaptability of Android phones to different types, we believe that our system is compatible with Apple phones as well, and therefore our system can be broadly deployed on different devices [37].

2) Car Models: Similarly, the shells and interiors of different car models may affect the performance of our system. Therefore, we repeat the above experiments in Car B using Samsung Galaxy S8. Fig. 15 shows the performance of each car model. It can be observed that both car models achieve good performance. In particular, Car B achieves an accuracy of 99.6%, which is slightly higher than Car A. The reason may be that Car B has a thick shell, which suffers less from wind, road, and engine noises. The results show our system is able to detect distracted driving with different car models.

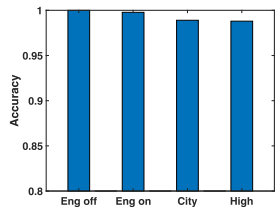


Fig. 16. Impact of engine status for phone distraction detection.

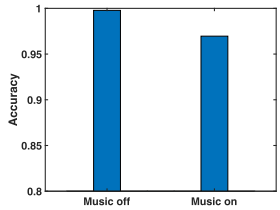


Fig. 17. Impact of in-car music for phone distraction detection.

3) Vehicle Engine Status: The car engine at different statuses or speeds generates different noise levels, including increased or decreased road and wind noises. We thus evaluate our system under different engine statuses, including *city driving*, *highway driving*, *engine on*, and *engine off*. We use Car A and Samsung Galaxy S8 for this impact study. Fig. 16 presents the classification results under the four different engine statuses. Not surprisingly, *engine off* performs the best with 100% accuracy, as this is a quiet in-vehicle environment. *Engine on* also performs well with 99.8% accuracy. *City driving* and *highway driving* achieve a slightly lower accuracy, which are 98.9% and 98.8%, respectively, though they suffer from different types of noises. In particular, *city driving* mostly involves the noise from frequent accelerations and braking in the traffic, while *highway driving* experiences more engine and wind noises. However, our system is robust enough to detect phone use distraction in both driving environments.

4) Car Audios: When driving, the drivers may turn the radio or music on. The car audio sounds may interfere with our sensing signal and affect the system's performance. We, therefore, evaluate our system with the car music on. It is noted that car audios primarily operate in the audible frequency range [38], [39], while our system works in the ultrosonic frequency range. Moreover, our sounds are internal, which are generated and recorded by the same device. As shown by prior work [40], the smartphone's own speaker sounds leave much higher Signal-to-Noise Ratio (SNR) to its microphone data compared to external sounds when they are at the same frequency. A prior study also demonstrates the feasibility of using ultrasonic sounds to record breath sounds in vehicles [41]. Our experiments in a car audio environment further confirm the limited impact of car audios on ultrasonic signals. The experiment was done with Car A and Samsung Galaxy S8, under the *engine on* status. The music sounds were between 56∼60 dB. Fig. 17 compares the performances of our system when the music is on or off. We observe that the music sounds do have a slight impact on our system performance. The classification accuracy degrades to 97.0% when the music is on,

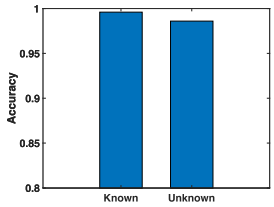


Fig. 18. Distraction detection for unenrolled users (not in the training model).

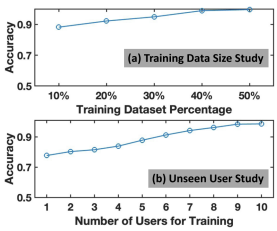


Fig. 19. Training data size and unseen user study performance.

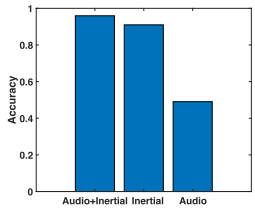


Fig. 20. Using different feature sets to recognize fine-grained handheld phone use.

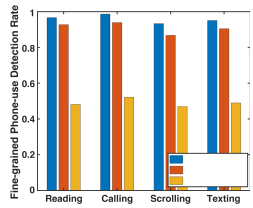


Fig. 21. Detailed performance of recognizing fine-grained handheld phone use.

which is still high. The result confirms the robustness of our system to work under car audios.

D. Training Data Size Study

To balance the effort required from users in data collection for model training with the goal of achieving good performance, we evaluate our system with different training data size splits. Specifically, we select five different percentages for the training dataset split, 10%, 20%, 30%, 40%, and 50%, for each of the eighteen participants for training, and used a fixed 50% of the dataset for testing. As illustrated in Fig. 19(a), we find that when we use more than 20% of the collected data for training, the achieved accuracy is over 90%. Particularly, our system achieves 92.2% accuracy when using 20% of the collected data for training. When we use 30% and 40% of the data for training, the accuracy increases to 94.9% and 98.9%, respectively. Our system performs the best when 50% of the collected data is used for training, achieving 99.6% accuracy. This study demonstrates that our system can minimize user effort by reducing the training data size while maintaining good handheld phone detection accuracy.

E. Unseen Users

To investigate whether each individual user's training data is required, we conduct a study with participants who are not included in the training data set or unknown to our system. Specifically, we divide the participants into two groups, with 10 and 8, respectively. We use the 10 participants' data to train the handheld phone distraction detection system, while the 8 unknown participants' data are used for testing. As illustrated in Fig. 18, our system achieves 98.6% accuracy in detecting handheld phone distractions for unknown users, which is just slightly lower than 99.6% achieved by the users included in the training set. The results indicate that our system has the potential to exempt a new user from training. The reason is that our handheld phone distraction detection is based on detecting the object's surface that contacts the phone, and the differences between the human skins are much smaller than the other in-vehicle surfaces. Furthermore, we reduce the number of users included in the training model from 10 to 1 and still use the 8 unknown users' data for testing. The results are presented in Fig. 19(b). Our system gets over 90% accuracy in phone distraction detection when we use more than 5 enrolled users for training. Specifically, our system obtains 91.3%, 94.2%, 96.2%, 98.4% and 98.6% classification accuracy in detecting phone distraction when we use 6, 7, 8, 9 and 10 enrolled users for training, respectively. The results indicate that it is possible to pre-train our system with a data set and exempt the new users from collecting new training data.

F. Fine-Grained Handheld Phone-Use Recognition

We evaluate the performance of our system in recognizing four fine-grained handheld phone use scenarios (i.e., reading, calling, scrolling, and texting) with eighteen participants. Fig. 20 presents the classification performance when using acoustic sensing, inertial sensing, and their fusion, respectively. We find our system achieves the highest performance with the fusion of acoustic and inertial sensing. The accuracy is 95.9% in recognizing the different handheld phone use scenarios. The inertial sensing plays a dominant role, which alone achieves 91.0% accuracy. This is much higher than using acoustic signals alone, whose accuracy is 49.0%, while the random guest rate is 25%. We further study the recognition performance regarding each type of handheld phone-use scenario. Fig. 21 shows that our system achieves a high TPR after for all the four handheld phone uses. In particular, with the fusion of acoustic and inertial sensors, our system achieves 96.7% TPR in reading, 98.7% TPR in calling, 93.3% TPR in scrolling, and 95.1% TPR in texting. Moreover, when we only use inertial sensors for recognition, the performance is slightly lower, and the TPRs are 92.7%, 93.9%, 86.8%, and 90.5%, respectively. But if only using the acoustic signals, the TPRs are 48.1%, 52.7%, and 48.9%,

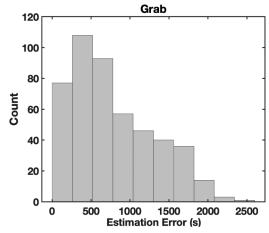


Fig. 22. Distraction start estimation.

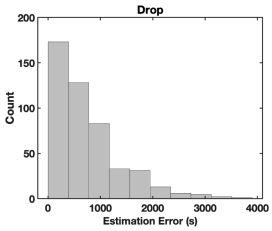


Fig. 23. Distraction end estimation.

respectively. The results confirm that only using the acoustic signals is hard to distinguish the different handheld device use scenarios and that the fusion of acoustic and inertial sensors enables the fine-grained handheld device distraction recognition. More specifically, while the acoustic sensing is used to recognize the contract surface of the phone, the inertial sensing is able to recognize the phone's motions resulted from specific handheld phone uses.

G. Practical Phone-Use Case Study

Lastly, we conducted eight case studies to monitor four participants' phone use. In each case, the assumed driver (front passenger) grabs the phone from one place in the vehicle, uses it, and then drops it at one place in the vehicle. Each case is repeated 20 times. The eight different cases are as follows:

Case 1: Seat - reading - seat.
Case 2: Center console - reading - center console.
Case 3: Cup holder - reading - cup holder.
Case 4: Phone mount - reading - phone mount.
Case 5: Pocket - reading - pocket.
Case 6: Pocket - calling - seat.
Case 7: Phone mount - texting - phone mount.
Case 8: Center console - scrolling - cup holder.

Our system achieves a 99.6% DR to capture the distracted driving instances with all of these cases, and the FPR is 0.6%. This performance is the combined result of distraction detection and the status sample error correction. We then evaluate the performance of our system to detect when the driver grabs and drops the phone. Figs. 22 and 23 present the distributions of the absolute time errors to detect the start and the end of each distraction instance. Our system achieves a median error of 0.67 seconds to determine the start of the distraction instance, and a median error of 0.56 seconds to determine the end time. These time errors are mainly associated with the complex transient states when the user grabs and drops the phone. We also find that most larger errors (e.g., between 1 s and 2 s) occur when the user grabs the

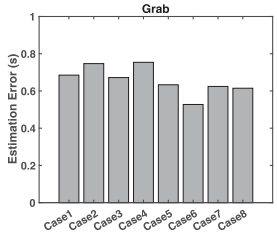


Fig. 24. Distraction start estimation for different study cases.

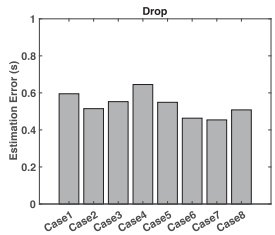


Fig. 25. Distraction end estimation for different study cases.

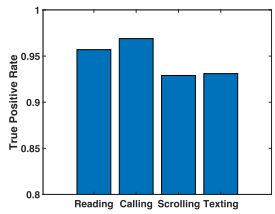


Fig. 26. Fine-grained phone-use recognition monitoring.

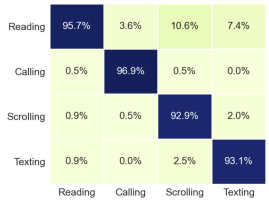


Fig. 27. Confusion matrix of recognition monitoring.

phone from or drops it to a phone mount. The reason is that fetching a phone from or putting it on the phone mount is a less smooth process and takes a longer time compared to the pocket, cup holder, center console, and seat. Figs. 24 and 25 further present the distracted instance start/end detection for each individual study case. The results confirm that our system can effectively capture the start and the end of a distraction instance in different cases.

We further evaluate our system performance in detecting the fine-grained handheld phone-use recognition in these case studies. Fig. 26 presents the performance of our system in recognizing the fine-grained phone uses (i.e., reading, calling, scrolling, and texting) in these more practical scenarios, where the classification accuracy is calculated after applying a 1-second error correction window. We observe that our system achieves 95.7%, 96.9%, 92.9%, and 93.1% TPR in recognizing reading, calling, scrolling, and texting, respectively. The confusion matrix of recognizing the four phone uses accuracy is presented in Fig. 27, we find our system achieves an overall accuracy of

95.2%. The results show that our system can effectively detect each distracted driving instance and recognize the fine-grained handheld phone-use scenarios. Besides, we observe a relatively lower true positive rate in recognizing scrolling and texting, compared to calling and reading. The reason is related to the fact that in practical scenarios, a user may not consistently text or scroll but has these phone-use actions accompanied by reading. For example, the user may read a message, text back, and then reads the returned message. Or the user may scroll the web page, read the content, and scroll the page again. The mixed texting-reading and scrolling-reading explain why some segments during the participants' texting and scrolling are classified as reading. The results motivate us to further compute the texting-reading and scrolling-reading ratios to estimate the impact of these fine-grained phone uses. We leave this to future work.

VI. Discussion & Future Work

Our system enhances traffic safety from three perspectives. For the driver, it can reduce handheld phone distractions by blocking or postponing non-emergency functions when the phone is in the driver's hand. Additionally, it can enable semi-autonomous driving modes to assist with traffic response and maintaining safe distances, similar to cruise control [42]. For surrounding vehicles, particularly self-driving cars, it alerts their systems to potentially unsafe nearby vehicles, allowing them to take precautionary measures like maintaining a longer distance. This alert includes accounting for possibly slower reactions from distracted drivers. For transportation management, the system transmits data on human factors to vehicular networks, aiding in traffic analysis and planning to reduce accidents and traffic congestion. Our proposed system may have some limitations. Its performance may fluctuate under certain conditions that are not covered by this work, such as heavy traffic, poor weather and extreme temperatures. The system's effectiveness also depends on the smartphone's hardware, especially the ultrasonic and inertial sensors' sensitivity, fidelity and range. Additionally, the variability in user behaviors and phone handling may result in occasional misclassifications. We will further study these in future work.

There are multiple topics we plan to explore in future work. 1) Our system currently focuses on law enforcement, which typically prohibits only handheld phone use. We believe our system can be extended to include handsfree phone use detection. For example, it could detect a phone in a phone mount when it makes a phone call. 2) We will explore the use of different time-frequency characteristic representation methods, such as the Discrete Wavelet Transform (DWT) to distinguish between handheld and handsfree phone use. 3) As shorter signals can also be used for sensing as demonstrated by prior work [43], we will study the balance between performance and computational overhead incurred by using different lengths of sensing signal. 4) We will investigate more potential impacts in practical driving scenarios, including frequent stops, various invasive maneuvers, and evaluate our system under such impacts. 5) The BFGS optimizer may enhance parameter learning rates and model performance, offering greater precision in parameter updates than Adam and we plan to explore this optimizer comparison in future work. We will adopt state-of-the-art methods to perform data augmentation and enlarge the dataset.

VII. Conclusion

This work proposes a learning-based phone-use monitoring system to address handheld phone distractions by sensing the driver's gripping hand. First, the system actively emits periodic ultrasonic pulse signals to continuously sense the material of the object in contact with the phone. Then, it determines whether the phone is being held by hand or placed on a surface within the vehicle, such as the seat, center console, pocket, or phone mount. After identifying a handheld phone distraction, the system employs a combination of acoustic and inertial sensing to recognize specific handheld phone usage activities, such as texting or calling. We develop an error correction window to correct misclassified phone-use status samples and to detect the start and end of each distracted driving activity related to handheld device use. Through comprehensive experiments involving various phone/car models and participants, the system is proven effective in providing fine-grained monitoring of driver phone use.

Acknowledgment

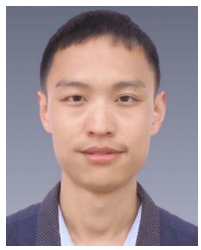
This paper conducts a more comprehensive study and takes one step further to achieve fine-grained handheld phone distraction detection by leveraging cross-domain sensing.

References

[1] V. K. Lee, C. R. Champagne, and L. H. Francescutti, "Fatal distraction: Cell phone use while driving," *Can. Fam. Physician*, vol. 59, no. 7, pp. 723–725, 2013.
[2] T. Zebra, "Texting and driving statistics," 2023. Accessed: Jan. 7, 2024. [Online]. Available: https://www.thezebra.com/resources/research/texting-and-driving-statistics/
[3] T. Covington, "Texting and driving statistics," 2021. Accessed: Jan. 7, 2024. [Online]. Available: https://www.thezebra.com/resources/research/texting-and-driving-statistics
[4] A. Leicht and K. A. Smith, "Texting and driving statistics 2023," 2024. Accessed: Jan. 7, 2024. [Online]. Available: https://www.forbes.com/advisor/car-insurance/texting-driving-statistics/
[5] Zendrive, "Zendrive collision report," 2021. Accessed: Jan. 7, 2024. [Online]. Available: https://live.zendrive.com/collision-report
[6] D. Gundlegård and J. M. Karlsson, "Handover location accuracy for travel time estimation in GSM and UMTs," *IET Intell. Transport Syst.*, vol. 3, no. 1, pp. 87–94, 2009.
[7] G. Chandrasekaran et al., "Tracking vehicular speed variations by warping mobile phone signal strengths," in *Proc. Int. Conf. Pervasive Comput. Commun.*, 2011, pp. 213–221.
[8] J. Yang et al., "Detecting driver phone use leveraging car speakers," in *Proc. ACM Int. Conf. Mobile Comput. Netw.*, 2011, pp. 97–108.
[9] Y. Wang et al., "Determining driver phone use by exploiting smartphone integrated sensors," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 1965–1981, Aug. 2016.
[10] C. News, "Why so many people text and drive, knowing dangers," 2014. Accessed: Jan. 7, 2024. [Online]. Available: https://www.cbsnews.com/news/why-so-many-people-text-and-drive-knowing-dangers/
[11] F. Li, H. Zhang, H. Che, and X. Qiu, "Dangerous driving behavior detection using smartphone sensors," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst.*, 2016, pp. 1902–1907.

[12] W.-H. Lee, X. Liu, Y. Shen, H. Jin, and R. B. Lee, "Secure pick up: Implicit authentication when you start using the smartphone," in *Proc. 22nd ACM Symp. Access Control Models Technol.*, 2017, pp. 67–78.
[13] K. B. Ahmed, B. Goel, P. Bharti, S. Chellappan, and M. Bouhorma, "Leveraging smartphone sensors to detect distracted driving activities," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3303–3312, Sep. 2019.
[14] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *Proc. IEEE Intell. Veh. Symp.*, 2012, pp. 234–239.
[15] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya, and M. C. Gonzalez, "Safe driving using mobile phones," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1462–1468, Sep. 2012.
[16] Z. Chen, J. Yu, Y. Zhu, Y. Chen, and M. Li, "D 3: Abnormal driving behaviors detection and identification using smartphone sensors," in *Proc. 12th Annu. IEEE Int. Conf. Sens., Commun., Netw.*, 2015, pp. 524–532.
[17] X. Xu, J. Yu, Y. Chen, Y. Zhu, S. Qian, and M. Li, "Leveraging audio signals for early recognition of inattentive driving with smartphones," *IEEE Trans. Mobile Comput.*, vol. 17, no. 7, pp. 1553–1567, Jul. 2018.
[18] H. Chu, V. Raman, J. Shen, A. Kansal, V. Bahl, and R. R. Choudhury, "I am a smartphone and i know my user is driving," in *Proc. IEEE 6th Int. Conf. Commun. Syst. Netw.*, 2014, pp. 1–8.
[19] V. Paruchuri and A. Kumar, "Detecting driver distraction using smartphones," in *Proc. IEEE 29th Int. Conf. Adv. Inf. Netw. Appl.*, 2015, pp. 468–475.
[20] J. Wahlström, I. Skog, P. Händel, B. Bradley, S. Madden, and H. Balakrishnan, "Smartphone placement within vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 669–679, Feb. 2020.
[21] M.-C. Chuang, R. Bala, E. A. Bernal, P. Paul, and A. Burry, "Estimating gaze direction of vehicle drivers using a smartphone camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 165–170.
[22] C. Jin, Z. Zhu, Y. Bai, G. Jiang, and A. He, "A deep-learning-based scheme for detecting driver cell-phone use," *IEEE Access*, vol. 8, pp. 18580–18589, 2020.
[23] M. Goel, J. Wobbrock, and S. Patel, "Gripsense: Using built-in sensors to detect hand posture and pressure on commodity mobile phones," in *Proc. 25th Annu. ACM Symp. User Interface Softw. Technol.*, 2012, pp. 545–554.
[24] C. Park and T. Ogawa, "A study on grasp recognition independent of users' situations using built-in sensors of smartphones," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol.*, 2015, pp. 69–70.
[25] M. Ono, B. Shizuki, and J. Tanaka, "Touch & activate: Adding interactivity to existing objects using active acoustic sensing," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, 2013, pp. 31–40.
[26] N. Kim, J. Lee, J. J. Whang, and J. Lee, "SmartGrip: Grip sensing system for commodity mobile devices through sound signals," *Pers. Ubiquitous Comput.*, vol. 24, pp. 643–654, 2020.
[27] K. Ashihara, "Hearing thresholds for pure tones above 16 kHz," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. EL52–EL57, 2007.
[28] EndDD, "Distracted driving research and statistics," 2023. Accessed: Jan. 7, 2024. [Online]. Available: https://www.enddd.org/research-stats/
[29] G. Cerrato, "Automotive sound quality–powertrain, road and wind noise," *Sound Vib.*, vol. 43, no. 4, pp. 16–24, 2009.
[30] M. Xu, J. Liu, Y. Liu, F. X. Lin, Y. Liu, and X. Liu, "A first look at deep learning apps on smartphones," in *Proc. World Wide Web Conf.*, 2019, pp. 2125–2136.
[31] S. Exchange, "When training a neural network, why choose adam over l-BGFs for the optimizer," Stack Exchange, 2020. Accessed: Jan. 7, 2024. [Online]. Available: https://scicomp.stackexchange.com/questions/34172/when-training-a-neural-network-why-choose-adam-over-l-bgfs-for-the-optimizer
[32] S. Ivanov, "Picking an optimizer for style transfer," 2017. Accessed: Jan. 7, 2024. [Online]. Available: https://blog.slavv.com/picking-an-optimizer-for-style-transfer-86e7b8cba84b
[33] PyPI, "Keras-flops 0.1.2," 2020. Accessed: Jan. 7, 2024. [Online]. Available: https://pypi.org/project/keras-flops/
[34] EndDD. ORG, "Distracted driving research and statistics - text messaging," 2023. Accessed: Jan. 7, 2024. [Online]. Available: https://www.enddd.org/research-stats/
[35] W. K. Winingham, "The risks of texting and driving," 2023. Accessed: Jan. 7, 2024. [Online]. Available: https://www.wkw.com/auto-accidents/blog/risks-texting-driving/
[36] abc7news, "Study finds people are texting as fast as they type on a keyboard," 2019. Accessed: Jan. 7, 2024. [Online]. Available: https://abc7news.com/type-texting-speed-typing/5587092/
[37] Y. Finance, "IOS vs Android market share by country: Top 30 countries using iphones," 2023. Accessed: Jan. 7, 2024. [Online]. Available: https://finance.yahoo.com/news/ios-vs-android-market-share-135251641.html
[38] D. Hawley, "How to set crossover frequency for a car audio system," 2021. Accessed: Jan. 7, 2024. [Online]. Available: https://www.jdpower.com/cars/shopping-guides/how-to-set-crossover-frequency-for-a-car-audio-system
[39] D. Walden, "How to set crossover frequency for car audio system," 2023. Accessed: Jan. 7, 2024. [Online]. Available: https://bellengineering.net/car-audio/how-to-set-crossover-frequency-for-car-audio-system/
[40] L. Huang and C. Wang, "Notification privacy protection via unobtrusive gripping hand verification using media sounds," in *Proc. 27th Annu. Int. Conf. Mobile Comput. Netw.*, 2021, pp. 491–504.
[41] Z. Wang, S. Li, Z. Li, S. Wang, and J. Cui, "Ultrasonic-based submillimeter ranging system for contactless respiration monitoring," *AIP Adv.*, vol. 13, no. 8, 2023, Art. no. 085117.
[42] S. Edelstein, "What is adaptive cruise control?," 2021. [Online]. Available: https://medium.com/axinc-ai/yolov5-the-latest-model-for-object-detection-b13320ec516b
[43] L. Huang and C. Wang, "PCR-Auth: Solving authentication puzzle challenge with encoded palm contact response," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 1034–1048.
[44] R. Wang, L. Huang, and C. Wang, "Preventing handheld phone distraction for drivers by sensing the gripping hand," in *Proc. IEEE 18th Int. Conf. Mobile Ad Hoc Smart Syst.*, 2021, pp. 410–418.

Ruxin Wang (Student Member, IEEE) received the B.S. degree in electrical and engineering from Northeastern Forestry University, Harbin, China, in 2017, and the M.S. degree in computer engineering from Syracuse University, Syracuse, NY, USA, in 2019. She is currently working toward the Ph.D. degree with Louisiana State University, Baton Rouge, LA, USA. Her research interests include mobile sensing and computing, cybersecurity, human-computer interaction, deep learning and AI in healthcare. She was the recipient of the N2Women fellowship at the 27th Annual International Conference On Mobile Computing And Networking.



Long Huang received the B.Eng. degree in electrical and electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2015, and the M.Sc. degree in electrical and electronic engineering from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2019. He is currently working toward the Ph.D. degree with Louisiana State University, Baton Rouge, LA, USA. His research interests include mobile computing, signal processing, and Internet of Things. He was the recipient of the Best Paper Award from the 14th International Workshop on Wireless Sensor, Robot and UAV Networks, IEEE WISARN 2021.

Chen Wang received the Ph.D. degree from Rutgers University, Piscataway, NJ, USA, in 2019. He is currently an Assistant Professor of computer science with Louisiana State University, Baton Rouge, LA, USA, and also Leads the Mobile and Internet SecuriTy (MIST) Laboratory. His research interests include cyber security and privacy, sensing, cyber-physical systems and smart healthcare. He has authored or coauthored a number of papers at high-impact conferences, including IEEE S&P, ACM CCS, ACM Mobicom and IEEE Infocom. He was the recipient of the five Best Paper awards from IEEE CNS 2018, IEEE CNS 2014, ASIACCS 2016, EAI HealthyIoT 2019, and IEEE INFOCOM WKSHPS 2021. From 2014 to 2023, his research studies have been reported by more than 170 media outlets, including IEEE Spectrum, NSF Science 360, CBS TV, BBC News, NBC, IEEE Engineering 360, Fortune, ABC News, and MIT Technology Review.