SoK: Assessing the Threat Potential of Vibration-based Attacks against Live Speech using Mobile Sensors

Payton Walker
University of Alabama at Birmingham
Birmingham, Alabama, USA
prw0007@uab.edu

Nitesh Saxena
University of Alabama at Birmingham
Birmingham, Alabama, USA
saxena@uab.edu

ABSTRACT

Existing academic research on vibration-based speech attacks has introduced interesting and intellectually appealing threat vectors with proof-of-concept demonstrations in controlled environments. The attacks presented in these studies exploit different types of sensors such as MEMS motion sensors, laser-based sensors, and some other sensors (camera, position error signal, piezo-disc) to measure the vibrations induced on an object by nearby sensitive speech. Such sensors are commonly found on mobile devices like smartphones and tablets that can be exposed to sensitive speech, revealing the significance of this potential threat. These studies have amassed significant attention in news and media and introduced concern to people about the safety of their day-to-day speech and around their personal, wireless and IoT devices. However, we hypothesize that the controlled experiments in the prior research maintain critical parameter values that are favorable to attack success (deviating from the limiting settings in a real-world scenario) and produce results that suggest a greater real-life threat level than actually exists.

The contributions made in this paper are as follows; First, we provide a detailed review of 10 existing academic research works related to vibration-based eavesdropping attacks. Second, we identify key experimental parameters that can impact the success of eavesdropping in the vibration domain. Third, we build a framework to evaluate the existing literature based on the Percent Parameters in Favored Settings (PPFS) Score metric that we define. Lastly, we use our defined framework to evaluate the feasibility of the existing vibration-based speech attacks to compromise *live human speech* to the extent of *full speech recognition*. The results of our evaluation suggest that none of the existing vibration-based eavesdropping attacks have a high likelihood of successfully compromising live human speech in a real-world scenario.

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

side-channel, vibration, speech eavesdropping, SoK

WiSec '21, June 28-July 2, 2021, Abu Dhabi, UAE

ACM ISBN 978-1-4503-8349-3/21/06…$15.00
https://doi.org/10.1145/3448300.3467825

ACM Reference Format:
Payton Walker and Nitesh Saxena. 2021. SoK: Assessing the Threat Potential of Vibration-based Attacks against Live Speech using Mobile Sensors. In *14th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '21), June 28-July 2, 2021, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3448300.3467825

1 INTRODUCTION

The threat of sensitive speech eavesdropping is becoming a great concern for many people as news and media, covering academic studies on the topic, report that these eavesdropping attacks can compromise speech in day-to-day scenarios (i.e., full speech recognition). Specifically, recent works have explored the *vibration domain* side-channel to attack sensitive speech. The attacks presented in the existing literature exploit a multitude of different sensors for collecting vibration data (MEMS motion sensors [13, 31, 41, 42, 68], lidar sensor [54], electro-optical sensor [45], high speed camera [21], position error signal [36], and piezo-electrical disks [40]). These sensors can collect data from an object without physically interfering with it (i.e., passively) and are ubiquitous on wireless devices such as smartphones and tablets, which increases the severity of the potential threat as the popularity of personal wireless and IoT devices (that are exposed to sensitive user speech) grows. The majority of these attacks attempt speech recognition and to achieve this goal attackers will use signal processing techniques and can implement basic or sophisticated machine learning (i.e., deep learning). If truly successful, these attacks could have devastating impacts in real-life situations by allowing an attacker to compromise speech in situations where speech privacy is otherwise protected in the audio domain. For example, we assume speech privacy in situations such as; talking about our medical history in a doctor's office, discussing confidential information in a business meeting, or even speaking sensitive information to a voice assistant device in our homes. In these scenarios, many of the limiting conditions that make eavesdropping in the audio domain a difficult task do not apply to the alternative vibration domain attack. Additionally, the constant presence of vibration sensors in devices like our smartphones makes such vibration domain attacks more likely, even in highly private situations.

In a vibration-based eavesdropping attack, the attacker exploits the vibrations induced by the speech source as the sound waves propagate through the space. When eavesdropping on live human speech, an attacker can exploit the vibrations induced on objects that are close to the speaker (i.e., a nearby cup, chip bag, etc.). For the purposes of this study, and as it should be considered throughout this paper, we define "live speech" as speech that is spoken by a live human (e.g., sound waves that induce vibrations are produced by

vocal chords). Conversely, we define "machine-rendered speech" as any speech audio (whether a recording of live speech or synthetically produced) that is generated by a mechanical speaker device. If the speech audio is machine-rendered, an attacker can exploit the vibrations of the speaker device itself (i.e., the vibrations of a smartphone when it plays audio in speaker mode).

These attacks could reveal significant information and therefore have the potential for some significant real-world applications. For example, these speech tasks can be used for sensitive jobs including national intelligence operations that look to spy on a person of interest that may be hiding out. Further, compromising speech via these attacks can be used in the commercial sector for modeling user behavior and generating targeted advertisements. Although likely intended for improving user experience, this potential of specialized ads has actually made general users more concerned that apps on their smartphones may be eavesdropping on all of their private conversations in order to learn what ads to target with [8, 18, 27, 43]. The idea that vibration-based eavesdropping is a real threat only enhances this concern as the permission-less motion sensors ubiquitous on smartphones can be easily compromised by an attacker.

While the existing studies introduce very interesting and intellectually appealing threat vectors, *can these threats actually be deemed concerning in real-life* is what we seek to explore. Or, are these threats merely "textbook threats"? This paper will examine a broad gamut of previous works on speech leakage in the vibration domain [13, 21, 31, 36, 40–42, 45, 54, 68]. The 10 works that we evaluate were selected because they best encompass the basic elements of vibration based speech attacks that we are interested in investigating. Specifically, we looked for studies that report on the physical parameters of their experiments, claim to demonstrate a successful attack against speech, or amassed significant attention in top academic conferences or the media and influenced the popular belief that such vibration-based attacks may be viable in a real-world situation.

We examine these works in part by identifying important experimental parameters that can affect the success of vibration-based speech leakage. If we consider a real-world situation where sensitive speech may be targeted by an attacker, there are many factors of the physical environment that can limit the potential success. The parameters we identify largely focus on these physical limitations so that we can assess how the experimental setups of the existing research compare to the real-world settings an attacker would encounter. ***We hypothesize that each parameter in the controlled experiments (of the existing academic research) that deviates from the limiting settings of a real-world scenario will improve the observed success of the attack.*** We believe inadequate representation of certain real-world limiting factors may attribute to the positive results reported in existing works. These studies have consequently introduced concern that people's day-to-day speech is vulnerable to such attacks.

The parameters we consider include: *speech source* (Live Human speaker, Machine-rendered), *speech loudness* (Normal conversational loudness (40-60 dB), "Loud" speech (>70 dB)), the *propagation medium* by which speech signals travel (through the air (Aerial), over a shared solid surface (Shared Surface), or via direct contact (Touching)), the vibration *sensor fidelity* (Low-fidelity (0-5 kHz), High-fidelity (>5 kHz)), the level of *background noise* in the attack environment (High (noisy), Low (ambient)), the *distance* between speech source and point of measurement (Very Close [0-0.5 m], Close (0.5 - 2 m], Far (>2 m)), and whether the attack attempted Speech Recognition (Yes/No). We suspect that most attacks that were previously perceived at a high threat level will show to be less of a threat in real-world settings.

Contributions: The contributions made in this work are as follows:

(1) We provide a detailed review and systematization of 10 existing research works on vibration-based speech attacks including experiment details and attack accuracies (Sec. 3).
(2) We identify key experimental parameterizations that must be understood in order to determine the viability of vibration-based speech leakage in real-world scenarios (Sec. 4).
(3) We define a simple framework for evaluating the existing literature based on an evaluation metric, Percent Parameters in Favored Settings (PPFS) Score, that we created; as well as some experimental laser vibrometry data to reinforce our qualitative framework (Sec. 5).
(4) Lastly, we evaluate a set of current research works on vibration-based eavesdropping, using the frameworks we defined, to assess the feasibility for these prior studies to be successful against live human speech (Sec. 6). This framework can be used to evaluate other attacks as they are identified in future works by other researchers.

The current research works do explore some aspects of vibration-based speech attacks; but we are missing a broad and comprehensive review of the current literature that systematically identifies the areas of this topic that need new or further understanding. While this study is largely applied work, it uniquely builds on the basic elements of security and privacy in speech environments and provides a comprehensive analysis of the realistic parameters that affect the success of vibration-based speech attacks. The results of our evaluation suggest that the existing attack methodologies are less likely to successfully compromise *live human speech* than has previously been perceived. This analysis could be invaluable, we believe, to the community as it can inform future academic research on evaluating this threat in the face of growing wireless device popularity.

2 BACKGROUND

2.1 Side-Channel Attacks

Side-channel attacks uniquely exploit data leakage that occurs naturally (or that the attacker does not cause directly) in a system in order to compromise user information. These attacks can be described as "passive" attacks and therefore have a few defining characteristics that potentially make them more dangerous [38]. Side-channel attacks will not impact the system or environment because data is only recorded, not changed. This also means that any resources of the target system are not changed and the victim cannot be alerted to the attack. Although many forms of side-channel attacks exist, we focus on evaluating *vibration-based* side-channel attacks against *speech* (i.e., eavesdropping). The "data leakage" that occurs is in the form of acoustic sound waves or vibration emissions from the speech source that allows the attacker to potentially learn the target speech. Figure 5 in the Appendix shows the typical flow of a vibration-based, side-channel speech attack.

2.2 Vibration Domain

We often think of the audio and vibration domains as being separate. However, these domains work closely with each other in many common scenarios. The best example of this is human speech where the vibrations of our vocal chords produce sound waves that we formulate into speech. These sound waves travel through the air and induce minute vibrations in our eardrums that cause us to hear it. Similarly, microphone devices use an inbuilt diaphragm, that vibrates in response to sound waves, to record audio. In this process, the vibrations are converted to an analog signal and output as audio. Hence, the vibration domain is a viable avenue for exploiting sensitive speech *when high quality vibration data can be collected*.

Some of the first studies that looked at side-channel attacks via the vibration domain targeted the touchscreen inputs of smartphones [17, 39, 48, 67]. These works exploited the vibration data recorded by the inbuilt MEMS motion sensors when a user typed using the soft keyboard on their smartphone. It was demonstrated that the MEMS motion sensor data could be used to infer the touch input (i.e., passwords) from the user. Exploiting the vibration side-channel can also be applied to compromise sensitive (audible) speech. An attacker must identify a target object that is subsequently affected by the speech in the area (e.g., by vibrations induced from sound waves). For example, an attacker executing a vibration-based speech attack may target the vibrations induced on a cup sitting on someones desk as they speak to another person or take a phone call. Additionally, if the speech source is machine-rendered, an attacker may be able to exploit the vibrations of the body of the speaker device as it plays the target audio and the vibrations of the internal speakers are propagated throughout the entire device.

2.3 Public Perception

There has been a lot of attention in news and media and among the general public on the potential for attackers to use vibrations sensors to eavesdrop on our speech. Many of the existing research works on the subject have been individually recognized. A prior work claimed to exploit the gyroscope on smartphones to eavesdrop speech and it was featured in many online articles [25, 28, 30]; even leading to a YouTube tutorial published by the authors [44]. Eavesdropping via laser is another popular attack that has received online attention [12], including at-home tutorials to build a laser microphone device yourself [24, 59]. The authors of these tutorials claim that the devices can be used to eavesdrop speech inside a target room or from across the street, without providing sufficient scientific evidence.

Another work reported that even an HDD could be converted to a microphone. This naturally sparked a lot of curiosity and fear (as HDDs are widely used), and resulted in significant media coverage [19, 22, 64]. Similarly, a more recent study reported that they could eavesdrop speech by exploiting the vibrations of a light bulb in the room. Within a short time, the claims made from this research were published in multiple articles [29, 46, 66]. Often times the authors of these articles will use verbiage to describe the conclusions from these academic studies that unintentionally overstates the threat potential actually realized in the work. And this can convince people that their day-to-day speech is at a higher risk of being eavesdropped than it likely is. Outside of controlled academic settings, an attacker must overcome many limiting factors present in real-world environments.

3 REVIEW OF EXISTING LITERATURE

In this section, we will discuss recent works that explore speech recognition (or related task) via the vibration domain. Most of the papers we evaluate specifically present side-channel speech attacks using vibrations ([13, 21, 31, 36, 41, 42, 45, 54]). We also include two studies ([40, 68]) in our evaluation that do not specifically present their work as an attack, but have clear potential for similar speech attack applications. Table 1 shows the experimental parameter values for each existing work (collected from the literature). Certain parameter settings represent more limiting and realistic attack scenarios, while others are favorable and will increase attack success.

3.1 Attacks using MEMS Motion Sensors

PitchIn: In the eavesdropping attack PitchIn [31], multiple non-acoustic sensors (e.g., geophone, accelerometer, gyroscope) were used to reconstruct nearby speech signals. The non-acoustic sensors were networked and the fusion of their measurements was used for speech reconstruction. The combination of sensors achieved a sampling frequency of 1 kHz. In their experiments, the speech source was a live human speaking at 85 dB loudness. To evaluate the quality of the reconstructed speech samples, Han et al. conducted a live study involving human listeners. The participants were asked to listen to the reconstructed speech signals and attempt to transcribe what was said. The results show that there was some word recognition among the human listeners which hints at the potential threat of this attack.

Gyrophone: Gyrophone [42] is a study that explored the use of an MEMS gyroscope sensor (200 Hz sampling frequency) on a smartphone to measure acoustic signals that are nearby. In the experimental setup, the speech audio was played from a loudspeaker device at 75 dB. Figure 3a in the Appendix depicts the experimental setup used in their work. With the reconstructed audio from their gyroscope data, Michalevsky et al. evaluated the intelligibility of the speech using the Sphinx speech recognizer. They also trained different machine learning models (SVM, GMM, DTW) for speaker recognition. Lastly, they performed isolated word recognition with leave-one-out cross validation. In general, the authors found that the reconstructed speech audio was partially intelligible; but more significantly they could successfully perform speaker recognition with their reconstructed audio (compromising the speech to some extent). However, inspection of their experimental design reveals that the loudspeaker system and the smartphone were placed on the same surface which can propagate stronger vibrations. Appendix Figure 4a shows the actual experimental setup from the literature.

AccelEve: AccelEve [13] is an eavesdropping attack presented by Ba et al. This attack exploits the speech reverberations (vibrations) that are generated from a smartphone's internal speakers and captured by the accelerometer motion sensor (200 Hz sampling frequency). The authors use the accelerometer of the same smartphone to measure the induced vibrations. Appendix Figure 3b is a general depiction of the experimental setup. In this work the authors implement deep learning-based speech recognition. Further, the authors of this work report greater sampling rates observed in the smartphones that they used (up to 500 Hz). Specifically, the authors report 70% speaker identification accuracy, 78% digit recognition accuracy, and 55% digit+letter recognition accuracy. A more recently published

Table 1: Experimental parameters used in each existing work on exploiting speech via the vibration domain. Bold text in the Attack Goal column indicate prior works that specifically attempt the highly threatening, speech recognition task. "*": Previous works marked with an asterisk were not specifically presented as speech *attacks*, but still have clear applications for potential attackers.

Previous Work	Sensor	Attack Description	Attack Goals	Sensor Resolution	Speech Source	Speech Loudness	Propagation Medium	Background Noise	Speech Distance
Pitchin [30]	MEMS Motion Sensors	Eavesdrop speech via sensor fusion (geophone, gyroscope, accelerometer)	**Speech Recognition**	1 kHz	Live Human	85 dB	Aerial	Ambient (~50 dB)	1 m
Gyrophone [41]		Measure acoustic signals using MEMS motion sensors on a smartphone	**Speech Recognition**, Speaker & Gender Identification	200 Hz	Machine-rendered	75 dB	Same Surface	Ambient (~50 dB)	0.15 m
AccelEve [9]		Accelerometer-based smartphone eavesdropping using speech reverberations	Word Recognition, Speaker Identification	500 Hz	Machine-rendered	70+ dB	Same Surface	70+ dB	< 0.1 m
Kinetic Song Comprehension [40]		Music identification using audio reverberations measured by smartphone motion sensors	Song Recognition	200 Hz	Machine-rendered	70+ dB	Same Surface	Ambient (~50 dB)	< 0.1 m
AccelWord* [71]		Hotword detection using accelerometer readings	Word Recognition	200 Hz	Machine-rendered	20-70 dB	Aerial	25 dB	0.3 m
Lidarphone [55]	Lidar	Acoustic side-channel attack using lidar sensors in robot vacuum cleaners	Word Recognition, Speaker & Gender Identification	1.8 kHz	Machine-rendered	70, 75 dB	Aerial/ Same Surface	70, 75, 77 dB	1.5 m
Lamphone [44]	Laser	Recover sound from lightbulb vibrations via laser	**Speech Recognition**	2 kHz	Machine-rendered	70, 90 dB	Aerial	Ambient (~50 dB)	0.01 m
The Visual Microphone [20]	Camera	Speech recovery using video recorded by a high speed camera	**Speech Recognition**	2 kHz	Machine-rendered	80-110 dB	Aerial	Ambient (~50 dB)	0.5-2 m
Hard Drive of Hearing [35]	PES	Eavesdrop speech via PES signal of HDD	**Speech Recognition**, Song Recognition	34.5 kHz	Machine-rendered	75, 85, 90 dB	Aerial	Ambient (~50 dB)	0.5 m
V-Speech* [39]	Piezo	Reconstruct speech via nasal vibrations measured using piezo-electrical discs	**Speech Recognition**	16 kHz	Live Human	40-60 dB	Touching	80 dB	0 m

work, Spearphone [11], similarly exploited smartphone speech reverberations and achieved even greater recognition accuracies while using off-the-shelf classifiers. This work has similar implications as our analysis of AccelEve demonstrates.

Kinetic Song Recognition: Another recent work by Matovu et al. [41] presents an attack that identifies popular songs played from a victim's smartphone by exploiting vibration data collected by the accelerometer of the smartphone. The model of this attack is very similar to that of the AccelEve attack described previously, using the same setup shown in Figure 3b in the Appendix. The attack uses a malicious app installed on the victim's phone that records speech reverberations via the accelerometer. In their experiments, the authors collect vibration data using this app while the phone plays different songs from its internal speakers. The classifier built by the authors using this data is able to achieve song recognition accuracy ≥80%. Another part of their study was exploring different surface materials that the phone is placed on during the experiments, as well as the volume level of the phone's speakers. The authors report that the surface type that the phone is placed on had little to no impact on the success of the attack.

Accelword: In an academic paper by Zhang et al., the authors present AccelWord [68], a speech recognition framework that was developed to detect when a user says a "hotword" by analyzing the MEMS accelerometer data of a nearby smartphone. Although Accelword is a benign application, its methodology could easily be adapted for a speech attack. The experiments that were conducted used a smartphone speaker to play the source audio at volumes of 20-70 dB. To evaluate their framework, the authors performed hotword detection using 10 live human speakers. They measure the true positive (TP), false positive (FP), and accuracy of the hotword classifier. Lastly, they compared their AccelWord framework to Google Now[65] and S Voice[55] for both hotword detection and energy consumption. Training their model on known hotwords, the authors found that AccelWord could successfully detect spoken hotwords (achieving an accuracy of 86%) and can be used for speaker recognition. Although AccelWord did not perform quite as well as Google Now and S Voice at hotword detection, it only consumes half as much energy comparably.

3.2 Attacks using Other Low Fidelity Sensors

LidarPhone: LidarPhone is an acoustic side-channel attack presented by Sami et al. [54] that exploits vibration data collected by the lidar sensors found on common wireless robot vacuum cleaners. The lidar sensors are laser-based sensors with a sampling resolution of 1.8 kHz that can be used to measure vibrations induced on objects that are near a target speech source. In this work, the authors implement their attack using a Xiaomi Roborock vacuum cleaning robot [53] to compromise speech (digits) and music played from a victim's computer speakers. In their setup, the victims computer speakers were placed on a desk with a subwoofer placed on the ground. A trash can was placed on the ground near the desk and acted as the target object that was measured by the lidar sensor. Figure 3c in Appendix shows the experimental setup used in their work. Notably, in this setup we see that there is a shared surface between the subwoofer (significant source of vibrations) and trash can which would induce most of the vibrations measured by the lidar sensor. Additionally, the target speech is played at volumes ≥70 dB which will further strengthen the induced vibrations. With their collected data, the authors built classifiers for gender, digit, and music recognition and reported recognition accuracies over 90% for all tasks, as well as 67.5% speaker recognition accuracy.

Lamphone: A recent study demonstrated how to exploit the vibrations of a hanging light bulb inside a room to eavesdrop on speech [45]. Nassi et al. utilized a Thorlabs PDA100A2 electro-optical

sensor [2] and a telescope to measure the light bulb from 75 meters away. The experimental setup used in this study is depicted in Appendix Figure 3d. In the target room the sensitive speech was played from a loudspeaker device at 70 and 90 dBs, at a distance of 1 cm from the light bulb. To evaluate the quality of the audio recovered from their sensor data, the authors first listened to the audio samples personally. Next, they tested the samples against an ASR system (Google Speech-to-Text [26]) and the samples containing songs against Shazam [5] and SoundHound [58]. The results of their evaluation revealed the reconstructed speech samples could be understood by human listeners and ASR, and the reconstructed song samples could be recognized by Shazam and SoundHound. Images of their experiments are shown in Appendix Figure 4b.

The Visual Microphone: In the work titled The Visual Microphone [21], Davis et al. utilized a Phantom V10 high speed camera and image processing techniques to extract vibration data from an object that they recorded. In their experiments the authors played the source audio (samples from the TIMIT dataset [23]) from a loud speaker at volumes of 80-110 dB. To evaluate the quality of the reconstructed signals, the authors began by using segmental STNR to measure accuracy. Additionally, they used a perceptually-based metric for speech intelligibility from a separate academic publication [60]. Lastly, the authors investigated the similarity between the spectral shapes of the reconstructed and original signals using the Log Likelihood Ratio (LLR). The results of their evaluation revealed that gender recognition could be achieved using the recovered audio, some music in the reconstructed audio can be understood by human listeners, and sound can be reproduced using the vibration data.

3.3 Prior Work using High Fidelity Sensors

Hard Drive of Hearing: In [36], Kwong et al. demonstrated an eavesdropping attack that exploits the mechanical components of a hard disk drive (HDD) to recover nearby speech. In their experiments, speech was played from a loudspeaker device at volumes louder than the range for normal human conversation. The Position Error Signal (PES) data, measured by the HDD, was used to reconstruct the audio signal. The HDD inadvertently acts as a microphone and can allow an attacker to eavesdrop on nearby speech. To evaluate the quality of the reconstructed samples, the authors performed a side-by-side visual comparison of the time-domain graphs. They calculated the similarity metric using discrete cross-correlation between the time series data. The speech intelligibility was determined using STNR and PESQ. Lastly, the Shazam[5] song identification tool was used on the reconstructed samples that contained song audio. When the source speech was 75 dB there was clear information leakage. At 85 dB intelligible speech was reconstructed. And at 90 dB, Shazam correctly identified a song played near the HDD.

V-Speech: The research titled V-Speech [40] is an interesting work by Maruri et al. that presented a novel speech sensing and processing solution that allows for speech recognition and even human-to-human communication in noisy settings. This work was not presented in the context of an attack, but we include it in our evaluation to determine its potential for real-world attack application. This solution, implemented in a pair of sensor-equipped eyeglasses, measured the vibrations of a user's nasal bone as they spoke. As the vibrations were captured by the piezoelectric discs (attached to the nose pads of the eyeglasses), a signal transformation was performed that utilizes both machine learning and signal processing techniques. V-Speech was tested in both quiet and noisy environments and the quality of the reconstructed audio samples was evaluated using the PESQ and Word Error Rate (WER) metrics. The authors also subjectively listened to the audio files to gauge the speech quality and varying STNR values were considered to maximize the intelligibility of the reconstructed speech. Unique to most other academic studies on vibration-based speech attacks, the experimental design to test V-Speech used *live human speakers* that spoke within an appropriate loudness range for human conversation (40-60 dB). The results of their evaluation was positive as they were able to reconstruct intelligible speech from the vibration data that was rated fair to good on how natural it sounded. This work demonstrated the vulnerability of speech when an attacker can acquire direct vibration data from the speech source. Similar to the nasal bone, any loudspeaker device that plays audio will incur vibrations proportional to the speech audio.

4 PARAMETERIZATIONS

Through our study of the existing literature on speech leakage in the vibration domain, we identify some key physical parameters that can affect the success of these attacks. Figure 1 depicts the experimental parameters that we consider in our evaluation of the previous works. Controlling these parameters and testing them in various experimental settings is crucial to understanding the feasibility of these attacks against sensitive speech in real-world scenarios. Additionally, evaluating these parameters will best inform what defensive strategies can work and should be explored in future research.

1. Speech Source: Speech source is a variable that can change between different attack settings. Often the source of sensitive speech that an attacker may target will be from live human speakers. However, current speaker technology is capable of projecting very clear audio which introduces new scenarios where sensitive speech may be gleaned. If we consider a conference call during a business meeting, the conference phone device will play the speech of the remote participants that may contain confidential information. Through experimental analysis we can compare the vulnerability of *live human speech* and *machine rendered speech* in vibration-based eavesdropping attacks to determine how speech source affects an attacker's ability to compromise the speech. If we consider the potential for attacking live human speech vs machine-rendered speech, we find certain properties that suggest machine-rendered speech may be generally easier to attack. Speech played from a speaker system will have consistent audible properties throughout, while live speech can be variable. Also, machine-rendered speech produces highly directed audio towards one direction, while live human speech emanates in a broader area. Therefore, we consider machine-rendered speech to be a favorable setting when evaluating an attack's feasibility against *live human speech*, like in this study.

2. Sound Pressure Level: Probably the most important parameter to consider when researching speech attacks is the Sound Pressure Level (SPL), or loudness, of the speech source. In regards to the vibration-based attack, the SPL of the target speech is significant because it directly determines the strength of the vibrations induced by the sound waves. It is reasonable to assume that there is a threshold

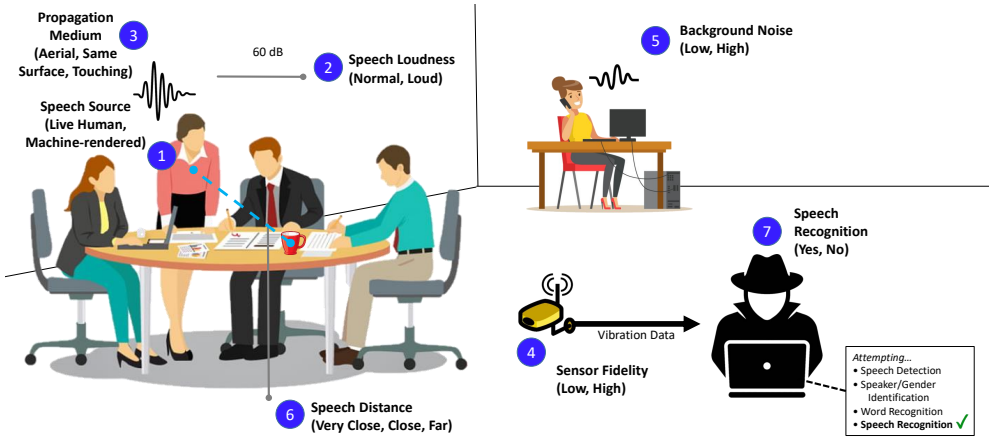


Figure 1: Figure depicting the seven parameters relevant to a real-world vibration-based side-channel attack scenario that we consider for our evaluation. The "realistic" parameter settings that an attacker would likely encounter are: 1) Speech Source (Live Human), 2) Speech Loudness (Normal), 3) Propagation Medium (Aerial), 4) Sensor Fidelity (Low), 5) Background Noise (High), 6) Speech Distance (Far), and 7) Speech Recognition (Yes).

of speech loudness that can ensure the success or failure of eavesdropping in the vibration domain. To evaluate the speech loudness used in the prior research experiments, we define two categories of speech loudness representing the *Normal* SPL range for human conversations (40-60 dB), as well as *Loud* speech (≥70 dB). In a real-world scenario, sensitive speech that would be targeted by an eavesdropping attack is not likely to be in the Loud SPL setting (i.e., speaking in a doctor's office or bank). Truly sensitive speech is going to be spoken within the Normal loudness range for human conversation, or possibly even quieter.

3. Propagation Medium: Propagation medium refers to the means by which sound waves from audio travel to the destination. Live human speech always travels *aerially*, simply traveling outward in the open air space. However, since the speech source does not have to be a live human, we also consider the possible propagation mediums for speech originating from a speaker device. When a speaker device plays some audio, sound waves are projected outward similar to live human speech. However, a speaker device will also generate internal vibrations that are proportional to the sound waves being played. This means that audio from a speaker device can also propagate through a *shared surface* when the vibrations from a speaker device travel along that surface and affect a different object. Along those lines, a specific scenario in which a speaker device is *touching* another object allows the vibrations from the speaker to propagate directly into the object it is touching. Understanding these different mediums, and how they can be utilized in different attack scenarios, is crucial for understanding the practicality of these attacks.

4. Vibration Sensor Fidelity: Exploiting the vibrations induced by sound waves can be achieved using any equipment that is capable of capturing vibration data. Existing research has demonstrated how to achieve this using different types of sensors. Each method provides a way to capture minute vibrations, and potentially speech information. However, some of these sensors have greater sampling rates which collect finer-grained vibration data. *Low-fidelity* sensors (≤5 kHz sampling frequency) such as MEMS motion sensors are cheap and ubiquitous and are a viable and likely option for a real-world attacker. However, the lower sampling rates of these sensors will result in lower Signal-to-noise Ratio (SNR) because environmental noise will be more present in the reduced data sampling. Also, the Nyquist Sampling Theorem says a sampling frequency of at least 5 kHz is required to obtain intelligible speech [49]. On the other hand, *High-fidelity* sensors (>5 kHz sampling frequency) such as the Laser Doppler vibrometers are able to collect high quality measurements of even the smallest vibrations and are more likely usable for full speech recognition. Because these sensors have greater sampling rates, they will achieve higher SNR for their recordings as additional noise will have a lower impact. These complex devices are only available to the public in a limited capacity and are significantly more expensive (tens of thousands of dollars) than the low-fidelity sensing equipment. Other high-fidelity sensors that we have seen include piezoelectrical discs and the PES signal of HDDs. These methods are not expensive to fund but are challenging to implement in a real-world attack because they require direct contact (piezo) or specialized firmware hacking (PES).

5. Background Noise: Another common factor that can influence the success rate of speech-based attacks is the level of background noise in the attack environment. If there are other audio signals traveling through the air, it can affect the quality of the content of the sound waves before they reach the sensing equipment. We consider two levels of background noise in this study: Low and High. We say there is *Low* background noise when the experimental setting only contains ambient noise aside from the source speech. All works that tested with this level of background noise were setup in an office (or similar) space where the ambient noise is ≤60 dB. Additionally, we say there is *High* background noise when the experimental setting injects noise in the test space to mimic a more realistic setting (i.e., work or public space with multiple speakers or music in the background). Here, the background noise is >60 dB and has the potential to compromise the source speech signals as they travel.

6. Speech Distance: We consider the distance between the speech source and point of measurement as another important physical parameter to consider when evaluating the attack's real-world feasibility. In some attacks, such as those that utilize MEMS motion sensors, the distance between the speech source and the point of vibration is equal to the distance between the speech source and sensor device (vibrations induced on sensor). Conversely, some attacks can

utilize long-range measurement devices such as the Laser Doppler vibrometers (LDV) where the point of measurement is an object near the target speech. Meanwhile, the actual LDV device can be a further away so the speech distance is still fairly short. To evaluate the existing work, we define three distance range classifications: *Very Close* distance is [0 - 0.5] meters, *Close* distance is (0.5 - 2] meters, and *Far* distance is >2 meters, between the speech source and the point of vibration measurement. In a real-world attack, it is far more likely that the point of measurement will be a *Far* distance from the speech source.

7. Attempted Speech Recognition: The main concern about speech privacy is the threat of *full speech recognition* by an attacker. Of all speech tasks, full speech recognition is the most difficult, but also the most severe because all speech information is compromised. Aside from speech recognition, there are other less complex speech tasks that are achievable with lower quality data. In the literature that we evaluate, we find studies that explore Speaker/Gender Identification, Word Recognition, and Song Recognition. To clarify, although Song Recognition can essentially reveal the same information as full speech recognition (e.g., lyrics of a song), we consider it a lower complexity speech task because in terms of signal processing and ML modeling, less information is required to achieve song recognition (i.e., short samples of the music, even without lyrics). Therefore, to consider an eavesdropping attack feasible in a real-world scenario against live human speech, the attack should be confirmed via experimentation to achieve *full speech recognition*. We consider whether the attacks in the previous literature achieved full speech recognition in their experimentation (Yes/No). If an attack is confirmed experimentally to achieve full speech recognition, it may be feasible that the attack can achieve full speech recognition in a real world attack scenario. But if an attack is not confirmed initially, the feasibility of successful speech recognition in the real world decreases.

5 EVALUATION FRAMEWORK

Although the works described above do lend some knowledge about passively eavesdropping speech via the vibration domain, there remains a lack of understanding about how feasible executing these attacks may be in real-world situations where certain parameters may not be in the favored setting. We evaluate a set of previous works on different parameters that determine how realistic their experimental attack scenarios are. Our evaluation is summarized in Table 3. From this evaluation we determine a potential risk level for each of the prior works (i.e., how likely is it that the work presented in the literature can be applied to a real-world attack scenario *against live human speech*). We estimate potential risk by weighing the positive attack results from these prior works against their representations of realistic, limiting parameter values in their experimental designs.

The seven parameters we consider are 1) the *speech source*, 2) the *speech loudness* (SPL), 3) the *propagation medium* of the speech, 4) the *fidelity* of vibration sensor used, 5) the amount of *background noise* in the environment, 6) the *distance* between the speech source and target object, and 7) if *speech recognition* was attempted. We selected physical parameters (1-6) that represent the attack environment, and are commonly reported in published papers, because we recognize in real-world scenarios there are certain physical factors that can be very limiting to a vibration-based attack. Additionally, we found that these parameters have been controlled in unrealistic, favorable settings in the experiments of existing works which claim potential for real life attack success. Because of this, we also consider whether the existing work attempted full speech recognition. In order to best replicate a realistic scenario that an attacker would face, the experimental setup would use Live Speech as the speech source, at a Normal loudness for human conversations, at a Far distance from the target object, where the speech travels Aerially in an environment with High background noise, and some cheap, ubiquitous, Low-fidelity sensor (i.e., MEMS) is used to record vibration data. Additionally, the attacker would likely try to accomplish full speech recognition. Table 2 lists the different parameters that we consider in our evaluation and their potential values. The table also depicts which parameter settings are most likely in a real-world scenario.

Table 2: List of experimental parameters that we consider in our evaluation and their potential settings, as well as their value for calculating the PPFS Score. *Realistic, non-favorable settings.

Parameter	Settings	PPFS Value (*favorable = 1*) (*non-favorable = 0*)	*Realistic Scenario
Speech Source	1. Live Human 2. Machine-rendered	0 1	Live Human
Speech Loudness	1. Normal (40-60 dB) 2. Loud (70+ dB)	0 1	Normal
Propagation Medium	1. Aerial 2. Same Surface 3. Touching	0 1 1	Aerial
Sensor Fidelity	1. Low (0-5 kHz) 2. High (>5 kHz)	0 1	Low
Background Noise	1. Low (~50 dB) 2. High (60+ dB)	1 0	High
Speech Distance	1. Very Close (0 - 0.5 m) 2. Close [0.5 - 2 m] 3. Far (>2 m)	1 1 0	Far
Speech Recognition	1. Yes 2. No	0 1	Yes

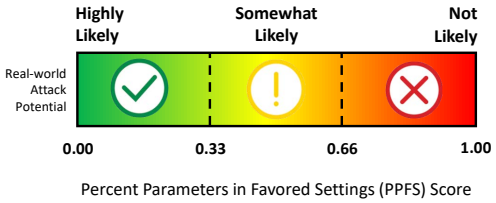


Figure 2: We define three categories to describe the potential success of a real-world vibration-based speech attack based on the Percent Parameters in Favored Setting (PPFS) metric that we defined. PPFS scores of [0.00-0.33) are labeled as Highly Likely; scores of [0.33-0.66] are labeled as Somewhat Likely; and scores of (0.66-1.00] are labeled as Not Likely.

PPFS Score: To assess the potential risk of each of the previous works (for use in a successful real-world attack), we define the metric Percent Parameters in Favored Settings (PPFS) Score. This metric expresses the portion of experimental parameters (from the set of seven that we consider) that were tested in favorable settings. It is important to note this observation because studies that maintain a lot of these parameters in favorable settings during their experiments may produce more positive speech leakage results than an attacker would likely achieve in a real-world attack situation. We calculate the PPFS Score as;

$$PPFS = \frac{params\_favored}{params\_total}$$

where *params_favored* equals the number of parameters that were maintained in favorable settings during experimentation and that

could have attributed to the reported success. It can be calculated using the equation;

$$params\_favored = \sum_{i=0}^{k} p\_val_i = \begin{cases} 0, & if\ nonfavorable\ setting \\ 1, & otherwise \end{cases}$$

where k is the set of all parameters, and p_val is a value of 0 or 1 that is assigned to each parameter based on whether it represents a favored or non-favored setting. Summing these values and dividing by *params_total*, the total number of parameters (7 in our case), results in a PPFS Score for each work. We consider one setting of each parameter as "non-favorable", and classify any remaining settings as "favorable". For example, Far distance is considered non-favorable, so both the Close and Very Close distances are considered favorable. We did not independently recreate each experiment, but rather determined the value of each parameter from the information provided in the literature of each work. We define three ranges for PPFS score values to categorize the potential risk (i.e., potential for success) of each prior work. PPFS scores of [0.00-0.33) are labeled as Highly Likely; scores of [0.33-0.66] are labeled as Somewhat Likely; and scores of (0.66-1.00] are labeled as Not Likely. Figure 2 depicts our defined PPFS value scale.

As this is an initial study to evaluate vibration-based speech attacks, we use a simple model for comparing our evaluation parameters. In our PPFS score calculation, all 7 parameters are considered equally valuable. Although certain parameters are likely more significant for attack success (i.e., speech loudness), we reserve more extensive evaluation of parameter correlations and adjusted weighting for future research. Our work focuses on the major parameters that will affect attack success, so we believe for an initial evaluation metric, comparing each parameter equally is still very revealing and speaks to the true feasibility of attacks presented in the prior literature. The attack potential classifications for each of the evaluated works is shown in Table 3 (using Figure 2 icons).

6 EVALUATION AND CHARACTERIZATION OF VIBRATION-BASED SPEECH ATTACKS

Here we will present our evaluation of the existing literature based on the framework described in Section 5. We discuss the parameterizations present in each work and determine their PPFS Scores to categorize their potential for success in a real-world attack against live human speech. We seek to learn the true feasibility of the current "known" attack methodologies in order to validate or refute some of the public confusion and paranoia around whether or not our everyday speech is being eavesdropped, and how easy those type of attacks could be. As a founding example, a recent work by Anand et al. [9] evaluated a previously confirmed attack on speech using MEMS motion sensors [42]. The authors determined the threat to be lower than previously perceived because certain experimental parameters (i.e., sound wave propagation) had not been considered. They concluded that the threat posed by the attack in a real-world scenario was lower than originally suggested because of a certain limiting factor that is required for the attack to be successful - propagating the sound waves via a shared surface. Appendix Figure 6 depicts frequency spectrum graphs presented in Speechless [9] that demonstrate the significance of the Same Surface propagation medium, in comparison to Aerial propagation, for inducing strong vibrations. Additionally, the work conducted in [10] explored a common mitigation technique for speech eavesdropping and demonstrated a correlation between physical barriers around a space and the success of eavesdropping attacks. Along a similar line, we consider a set of parameters, like the propagation medium of sound waves, and their representation in the experimental setups of the existing literature. Table 3 summarizes our evaluation of each work including the PPFS Score calculations and attack potential classifications.

6.1 Evaluating MEMS Motion Sensor Attacks

In our evaluation we looked at five different papers that explored the use of MEMS motion sensors to eavesdrop speech via the vibration side-channel. The AccelEve [13] and Kinetic Song Comprehension [41] attacks look to compromise speech audio played from the onboard smartphone loudspeakers by capturing the speech reverberations induced in the smartphone using the MEMS motion sensors (i.e., accelerometer, gyroscope, etc.). Although these works are designed to compromise machine-rendered audio played from a victim's smartphone, we chose to evaluate their potential to compromise live human speech because of the generalized fear of eavesdropping that has resulted from coverage of these works in the media [6, 20, 61], and the ever growing presence of wireless/mobile devices that are equipped with these sensors. These articles claim the research has shown user's "calls" can be eavesdropped in attacks when the research has actually only demonstrated eavesdropping on audio output from the phone. We acknowledge that these attacks do have merit in their own domains, but in terms of the threat to *live human speech*, to the extent of full speech recognition, we need a clearer understanding of their feasibility.

Looking at the experimental setups used in these papers we do find that some of the parameters were kept in the more realistic settings. Both works used Low-fidelity sensors in their attacks which would produce the lower quality of data a real-world attacker would likely have to work with. Additionally, the AccelEve attack was tested in environments with High background noise. However, we also observe that the majority of parameters in each of these studies was maintained in favorable settings. Specifically, we see that these attacks were tested with Machine-rendered speech, in the Loud SPL range, propagated through a Shared Surface, and with a Very Close distance between the speech source and point of measurement (all contained within the same smartphone housing). Additionally, neither work actually attempted full speech recognition, and the Kinetic Song Comprehension [41] attack was only tested in environments with Low background noise. Considering all of these evaluation parameters, we calculate the PPFS Scores 0.71 and 0.86 for the AccelEve and Kinetic Song Recognition works, respectively, classifying their potential for real-world attack success as **Not Likely**.

The other studies that use MEMS sensors look to compromise speech that comes from a source separate from the vibration sensors. The Pitchin [31] attack was tested with the realistic parameter settings of Low sensor fidelity, Live Human speech, Aerial propagation of the sound waves, and attempted full speech recognition. However, the experiments used speech in Loud SPL setting, at a Close distance to the point of measurement, and in an environment with Low background noise. The PPFS Score calculated for the Pitchin attack is 0.43 and it receives an attack potential classification of

Table 3: Evaluation summary of the existing literature detailing the experimental parameter settings that they tested. The cells shaded in green indicate *non-favorable, realistic* settings that receive a value of 0. Cells shaded with red or yellow indicate *favorable* settings with a value of 1. "PPFS": Percent Parameters in Favored Settings; "*": Previous works marked with an asterisk were not specifically presented as speech attacks by the authors, but still have clear applications for potential attackers.

	Sensor	Sensor Fidelity	Speech Recognition	Speech Source	Speech Loudness	Propagation Medium	Background Noise	Speech Distance	PPFS	Attack Potential
Expected Realistic Attack		**Low**	**Yes**	**Live Human**	**Normal**	**Aerial**	**High**	**Far**	**0.00**	✓
Pitchin [30]	MEMS Motion Sensors	Low	Yes	Live Human	Loud	Aerial	Low	Close	0.43	!
Gyrophone [41]		Low	Yes	Machine-rendered	Loud	Same Surface	Low	Very Close	0.71	✗
AccelEve [9]		Low	No	Machine-rendered	Loud	Same Surface	Low, High	Very Close	0.71	✗
Kinetic Song Comprehension [40]		Low	No	Machine-rendered	Loud	Same Surface	Low	Very Close	0.86	✗
AccelWord* [71]		Low	No	Machine-rendered	Normal, Loud	Aerial	Low	Very Close	0.57	!
Lidarphone [55]	Lidar	Low	No	Machine-rendered	Loud	Aerial, Same Surface	High	Close	0.71	✗
Lamphone [44]	Laser	Low	Yes	Machine-rendered	Loud	Aerial	Low	Very Close	0.57	!
The Visual Microphone [20]	Camera	Low	Yes	Machine-rendered	Loud	Aerial	Low	Close	0.57	!
Hard Drive of Hearing [35]	PES	High	Yes	Machine-rendered	Loud	Aerial	Low	Very Close	0.71	✗
V-Speech* [39]	Piezo	High	Yes	Live Human	Normal	Touching	High	Very Close	0.43	!

Somewhat Likely. The Gyrophone [42] attack also had the realistic settings of attempted speech recognition and Low sensor fidelity, but all other parameters were kept in favorable settings. The PPFS Score of Gyrophone is calculated as 0.71 which indicates the majority of parameters were in favorable settings, classifying the attack as **Not Likely** for potential success against live speech. Lastly, the AccelWord [68] application, which actually targets the live speech of a user, uses the realistic parameter settings of Low sensor fidelity and speech at a Normal loudness that propagates Aerially (representing most sensitive speech scenarios). However, other parameters such as Low background noise and Very Close distance between the speech source and vibration sensor can deviate the results from what would be observed by a real-world attacker. We calculate the PPFS Score for AccelWord and get 0.57 which classifies the application as having **Somewhat Likely** potential for attack success.

6.2 Evaluating Other Low-fidelity Sensor Attacks

Aside from MEMS motion sensors, a few of the previous works that we evaluated explored the use of other low-fidelity sensors for capturing vibration data. First, the Lidarphone [54] attack uses the low-fidelity lidar sensor (laser-based sensor) found on robot vacuum cleaners to compromise speech. In terms of our evaluation, the Lidarphone attack maintained three parameters in realistic settings. Along with the Low fidelity sensor, speech was propagated in the Aerial medium and in environments with High background noise. If we look at the favorable parameter settings, we see that Lidarphone used Machine-rendered speech in the Loud SPL setting, with a Close distance between the speech source and target object, and did not attempt speech recognition. This attack scenario involves speech audio that was propagated both Aerially (through tweeter speakers) and through a Same Surface (subwoofer placed on the ground). Although we do not consider this parameter in the favored setting (such that it affects the PPFS Score calculation), it is important to note that without the portion of the sound waves that propagate through the Same Surface medium in their attack scenario, it is likely the authors would have observed lower success rates. The final PPFS Score calculation for Lidarphone is 0.71 which classifies this attack as **Not Likely** for success against live human speech.

Next, in the Lamphone [45] experiments the parameters that were maintained in realistic settings were Low sensor fidelity, Aerial propagation of the sound waves, and attempted full speech recognition. However, we see that the experiments used Machine-rendered speech in the Loud SPL setting, in an environment with Low background noise, and with a Very Close distance between the speech source and light bulb. Considering these parameter values we calculate the PPFS Score for Lamphone and get 0.57, classifying the attack as **Somewhat Likely** to successfully compromise live speech spoken by a human. Lastly, The Visual Microphone [21] study developed a methodology to extract vibration data of an object from the video recordings of a low-fidelity camera. Similar to Lamphone, the Visual Microphone experiments also attempted speech recognition targeting Aerially propagated speech. However, this attack was also tested with Machine-rendered speech at a Loud SPL setting, with Low background noise, and Close distance between the speech source and target object. The PPFS Score calculated for The Visual Microphone attack is 0.57 so it is also classified as **Somewhat Likely**.

6.3 Evaluating High-fidelity Sensor Attacks

Among the previous research works that we evaluated, two of them investigated the use of high-fidelity vibration sensors to reconstruct speech. The Hard Drive of Hearing study [36] explored an attack that exploits the high resolution Position Error Signal (PES) sensor that is used to monitor the offset of the read/write head in HDDs. Even though this sensor provides certain benefits such as finer quality data, it is important to consider the challenges that are incurred when an attacker chooses a high fidelity sensor. For example, the PES methodology used in this study requires firmware hacking of the target HDD which significantly increases the difficulty for the

attacker. Through our evaluation of the experiments used to test this attack, we find that only two parameters were kept in realistic settings; speech was propagated Aerially and the attack attempted full speech recognition. And aside from the High fidelity sensor, the experiments in this study also maintained the favorable parameter settings of Machine-rendered speech in the Loud SPL setting, Low background noise, and Very Close distance between the speech source and the HDD. The results of our PPFS Score calculation for this attack is 0.71 which classifies the potential threat to live human speech in real-world scenarios as **Not Likely**.

The last work that we evaluated is the benign application V-Speech [40] which demonstrates the potential use of high-fidelity piezo-electrical discs in speech reconstruction (from nasal bone vibrations). The clear limitation to using the piezo-disc methodology in the attack domain is that physical contact between the discs and the speech source is required to obtain the high resolution vibration measurements needed for full speech reconstruction. In terms of the parameters we consider in our evaluation, the V-Speech application tested over half of our parameter set in realistic settings. This includes using speech from Live Human speaker that is spoken at a Normal loudness level, in environments with High background noise, and successfully achieving full speech recognition. Weighing these against the parameters that were tested in favorable conditions, we calculate the PPFS Score to get 0.43. Therefore, V-Speech is classified as **Somewhat Likely** to compromise live human speech (e.g., not spoken by the user of V-Speech).

6.4 Data-Driven Evaluation using LDV

Existing Literature: A few research works have been released that specifically explore the use of high-fidelity laser vibrometers for eavesdropping speech in the vibration domain [37, 50, 56]. In controlled experiments, the authors evaluate the potential for these vibrometer devices to record fine enough vibration data such that quality speech can be recovered. These works report some success in reconstructing speech from the vibrometer measurement, but further inspection of the provided literature reveals unreported parameter settings that could have significantly improved the success that they observed. Specifically, the work by Shang et al. [56] did not report the loudness of their speech source, the propagation medium of the sound waves, or the amount of background noise in their experiments. The work by Li et al. [37] did not report their speech source, speech loudness, propagation medium, background noise level, or speech distance parameter values. Without knowing the values of these parameters in their experiment designs, there is little evidence that such methods would work against live human speech in a real-world scenario. Lastly, the work by Peng et al. [50] did report on most of the parameters that we consider and we find that many were in fact kept in favored settings which would improve the observed success. In their experiments, the speech was machine-rendered, played at a Loud volume (75 dB), and was located at a Very Close distance to the target object (0.5 meters). Further, the authors used two high-fidelity vibrometer sensors to obtain an unrealistic sampling rate for the vibration data. Therefore, the results from these laser-based evaluations further support that certain favored parameter settings are required to fully recover speech from vibrations.

7 SUMMARY

We have compiled a set of existing literature that explore vibration-based side-channel attacks against speech [13, 21, 31, 36, 41, 42, 45, 54], and two other related studies [40, 68]. These works have presented positive results on the potential success of their attacks in the experimental scenarios that were tested. However, whether or not these results adequately translate to feasibility in a real-world setting was unknown. We evaluated the prior works across seven experimental parameters that describe conditions of the attack environment. We consider the vibration sensor fidelity, speech source, speech loudness, speech distance, propagation medium, background noise level, and whether speech recognition was attempted. In a real-world situation, these parameters would have values that limit the potential for attack success. However, in controlled experiments these parameters can be set to more *favored* values that are less likely to occur in the real-world and can improve attack success.

Along these lines, we reason that prior works with many favored experimental settings are not likely to be successful in a real life attack situation. On the other hand, if there is an attack that has confirmed success with realistic and non-favorable experimental parameter settings, then there is high likelihood of real-world success. We defined the Percent Parameters in Favored Settings (PPFS) Score metric to express how favored each of the prior experiments were and defined a scale for the PPFS scores to classify the likelihood for real-world attack success. We determined that half of the works evaluated are Somewhat Likely to be successful in a real attack, while the other works are Not likely. All works we evaluated had over 1/3 of their parameters in favored settings, meaning none can be classified as Highly Likely for real-life feasibility.

The research community has attempted to develop some defenses to relevant attacks in both audio domains [7, 33, 34, 47, 51, 52] and vibration domains [57, 62]. However, it seems certain physical parameters can be exploited to easily thwart such attacks. In future work we plan to establish a correlation between the evaluation parameters and attack success, and develop mitigation strategies focused around the most significant parameters. Further, because these attacks do not seem very effective in real-world settings, a determined attacker might actually deploy traditional mechanisms of eavesdropping such as implanting insiders or "bugs" [1, 3, 4, 63], or exploiting the always listening voice assistant devices [14–16, 32, 35].

8 CONCLUSION

In this work, we evaluated current studies on vibration-based speech attacks. We explore these eavesdropping attacks in terms of their potential for application in real-world situations against live human speech. Our observations from the academic literature are summarized in Table 3. We determined that much of the current research does not precisely recreate the limiting parameter settings of a real-world scenario. This has led to a misunderstanding among a lot of people about the true feasibility of these attacks to compromise our day-to-day speech. We do not believe, at the present time, that the vibration-based speech attacks presented in the prior literature are likely to be successful against sensitive, *live human speech* in real-world situations. Moving forward, our evaluation framework may become a useful tool in evaluating wireless and IoT device vulnerability to eavesdropping attacks via induced speech vibrations.

ACKNOWLEDGEMENT

We would like to give special thanks to our shepherd Dr. Shivam Bhasin, as well as the set of anonymous reviewers for their valuable feedback on this paper. This work is partially supported by the National Science Foundation (NSF) under the grants: CNS-1714807, CNS-1526524, CNS-1547350.

REFERENCES

[1] AV Security [n.d.]. *Bugging Threats*. AV Security. https://avsecurity.com/bugging-threats/

[2] ThorLabs [n.d.]. *PDA1000A2 - Si Switchable Gain Detector*. ThorLabs. https://www.thorlabs.com/thorproduct.cfm?partnumber=PDA100A2

[3] U.S.Department of Commerce Western Region Security Office 2001. *Bugs and Other Eavesdropping Devices*. U.S.Department of Commerce Western Region Security Office. https://www.wrc.noaa.gov/wrso/security_guide/intro-17.htm

[4] Daily Mail 2016. *Voice recorders in binders and secret meetings with undercover agents: How FBI penetrated Cold War-style Russian spy ring in New York City*. Daily Mail. https://www.dailymail.co.uk/news/article-3484620/FBI-penetrated-New-York-based-Russian-spy-ring-using-hidden-recorders.html

[5] 2020. *Shazam*. https://www.shazam.com/

[6] Abeerah Hashim. 2019. *Spearphone Attack Allows Android Apps to Listen To Your Loudspeaker Conversations*. https://latesthackingnews.com/2019/07/20/spearphone-attack-allows-android-apps-to-listen-to-your-loudspeaker-conversations/

[7] A.B.M. Alim Al Islam, Tusher Chakraborty, Taslim Arefin Khan, Mahabub Zoraf, and Chowdhury Sayeed Hyder. 2017. Towards Defending Eavesdropping on NFC. *Journal of Network and Computing Applications* (2017), 11–23. https://doi.org/10.1016/j.jnca.2017.10.013

[8] Abdullah AlShawaf. 2019. *Why Do We See Ads for Things We Have Just Talked About*. https://medium.com/swlh/why-do-we-see-ads-for-things-we-have-just-talked-about-ba0924f9abee

[9] S. Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors. In *Proceedings of the IEEE Symposium on Security and Privacy (SP '18)*. 1000–1017.

[10] S. Abhishek Anand, Payton Walker, and Nitesh Saxena. 2019. Compromising Speech Privacy under Continuous Masking in Personal Spaces. *Proceedings of the 17th International Conference on Privacy, Security and Trust*, 1–10.

[11] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2021. Spearphone: A Lightweight Speech Privacy Exploit via Accelerometer-Sensed Reverberations from Smartphone Loudspeakers. In *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec'21)*.

[12] Charles Arthur. 2013. *Laser spying: is it really practical?* The Guardian. https://www.theguardian.com/world/2013/aug/22/gchq-warned-laser-spying-guardian-offices

[13] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Z. Qin, B. Li, X. Liu, and K. Ren. 2020. Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer. In *Network and Distributed System Security Symposium (NDSS) (NDSS '20)*.

[14] Joshua Bote. 2019. *Google workers are eavesdropping on your private conversations via its smart speakers*. USA Today. https://www.usatoday.com/story/tech/2019/07/11/google-home-smart-speakers-employees-listen-conversations/1702205001/

[15] Matthew Broersma. 2019. *Smart Speakers Hacked To Listen In, Steal Passwords*. https://www.silicon.co.uk/workspace/smart-speakers-hijacked-297599?cmpredirect

[16] Fabian Bräunlein and Luise Frerichs. 2019. *Smart Spies: Alexa and Google Home expose users to vishing and eavesdropping*. USA Today. https://srlabs.de/bites/smart-spies/

[17] Liang Cai and Hao Chen. 2011. TouchLogger: Inferring Keystrokes On Touch Screen From Smartphone Motion. In *Proceedings of the 6th USENIX Workshop on Hot Topics in Security*.

[18] Christian Cawley. 2019. *Does Your Phone Listen to You for Ads? (Or Is It Just Coincidence?)*. https://www.makeuseof.com/tag/your-smartphone-listening-or-coincidence/

[19] Thomas Claburn. 2019. *From hard drive to over-heard drive: Boffins convert spinning rust into eavesdropping mic*. The Register. https://www.theregister.com/2019/03/07/hard_drive_eavesdropping/

[20] Danny Bradbury. 2019. *Your Android's accelerometer could be used to eavesdrop on your calls*. https://nakedsecurity.sophos.com/2019/07/23/spearphone-researchers-eavesdrop-on-phone-loudspeakers/

[21] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. 2014. The Visual Microphone: Passive Recovery of Sound from Video. *ACM Transaction on Graphics* (2014), 79:1–79:10. http://doi.acm.org/10.1145/2601097.2601119

[22] David Ehrenstein. 2019. *Your Hard Drive May Be Listening*. Physics. https://physics.aps.org/articles/v12/24

[23] W. M. Fisher, G. R. Doddington, and K. M. Goudiemarshall. 1986. The darpa speech recognition research database: specifications and status. *Proc. DARPA Workshop on speech recognition* (1986), 93–99.

[24] David Galloway. 2012. *Build a Laser Microphone to Eavesdrop on Conversations Across the Street*. Life Hacker. https://lifehacker.com/build-a-laser-microphone-to-eavesdrop-on-conversations-5961503

[25] Phil Goldstein. 2014. *Researchers show how to turn a phone's gyroscope into a crude microphone for eavesdropping*. Fierce Wireless. https://www.fiercewireless.com/wireless/researchers-show-how-to-turn-a-phone-s-gyroscope-into-a-crude-microphone-for-eavesdropping

[26] Google Cloud. 2020. *Speech-to-Text*. https://cloud.google.com/speech-to-text

[27] Jefferson Graham. 2019. *Is Facebook listening to me? Why those ads appear after you talk about things*. https://phys.org/news/2019-06-facebook-ads.html

[28] Andy Greenberg. 2014. *The Gyroscopes in Your Phone Could Let Apps Eavesdrop on Conversations*. Wired. https://www.wired.com/2014/08/gyroscope-listening-hack/

[29] Andy Greenberg. 2020. *Spies Can Eavesdrop by Watching a Light Bulb's Vibrations*. Wired. https://www.wired.com/story/lamphone-light-bulb-vibration-spying/

[30] Jason Hahn. 2014. *Your smartphone's gyroscope can be turned into an eavesdropping hacker's microphone*. Digital Trends. https://www.digitaltrends.com/mobile/your-smartphones-gyroscope-can-be-turned-into-an-eavesdropping-hackers-microphone-privacy/

[31] Jun Han, Albert Jin Chung, and Patrick Tague. 2017. PitchIn: Eavesdropping via Intelligible Speech Reconstruction Using Non-acoustic Sensor Fusion. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN '17)*. 181–192.

[32] Charlotte Jee. 2019. *Smart speakers can be hijacked by apps that spy on users*. MIT Technology Review. https://www.technologyreview.com/2019/10/21/330/smart-speakers-can-be-hijacked-by-apps-that-spy-on-users/

[33] Yeongseok Kim and Youngjin Park. 2017. Effect of active noise control and masking sound on speech intelligibility. *Applied Acoustics* 123 (2017), 152–157. https://doi.org/10.1016/j.apacoust.2017.02.021

[34] Alexey Krasnov, Edward R Green, Bret Engels, and Barry Corden. 2019. Enhanced Speech Privacy in Office Spaces. *Building Acoustics* 26, 1 (2019), 57–66. https://doi.org/10.1177/1351010X18798105

[35] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill Squatting Attacks on Amazon Alexa. In *Proceedings of the 27th USENIX Conference on Security Symposium (SEC'18)*. USENIX Association, 33–47.

[36] Andrew Kwong, Wenyuan Xu, and Kevin Fu. 2019. Hard Drive of Hearing: Disks that Eavesdrop with a Synthesized Microphone. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (SP '19)*.

[37] Weihong Li, Ming Liu, Zhigang Zhu, and Thomas S. Huang. 2006. LDV Remote Voice Acquisition and Enhancement. *18th International Conference on Pattern Recognition* 4, 262–265.

[38] Mahesh Parahar. 2019. *Difference between Active Attack and Passive Attack*. https://www.tutorialspoint.com/difference-between-active-attack-and-passive-attack

[39] Philip Marquardt, Arunabh Verma, H. Carter, and P. Traynor. 2011. (sp)iPhone: Decoding Vibrations From Nearby Keyboards Using Mobile Phone Accelerometers. In *Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS '11)*. Association for Computing Machinery.

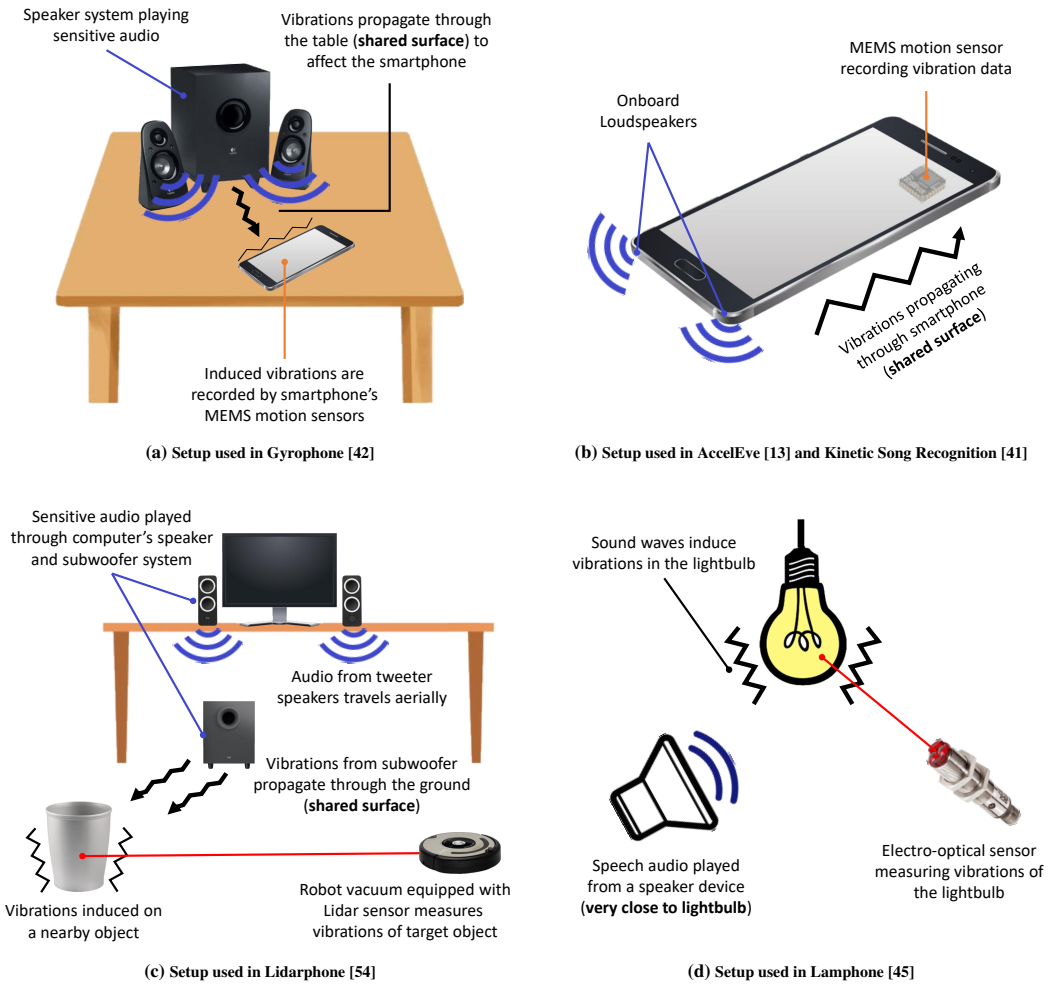
[40] Héctor A. Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Marco Beltman, Lama Nachman, and Hong Lu. [n.d.]. V-Speech: Noise-Robust Speech Capturing Glasses Using Vibration Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* ([n. d.]).

[41] Richard Matovu, Isaac Griswold-Steiner, and Abdul Serwadda. 2019. Kinetic Song Comprehension: Deciphering Personal Listening Habits via Phone Vibrations. *ArXiv* abs/1909.09123 (2019).

[42] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *Proceedings of the 23rd USENIX Security*

Symposium (USENIX Security '14). 1053–1067.
[43] Katie Moore. 2019. *Is my phone listening to me? We tested it, this is what happened*. https://www.wwltv.com/article/news/investigations/katie-moore/when-we-talk-our-phones-are-listening-how-else-could-this-happen/289-04f2c434-4ca3-49a5-abff-b9d86f21f19d
[44] Gabi Nakibly and Yan Michalevsky. 2015. *Gyrophone: Eavesdropping Using a Gyroscope*. YouTube. https://www.youtube.com/watch?v=PVpBrum-H0Y
[45] Ben Nassi, Yaron Pirutin, Adi Shamir, Yuval Elovici, and Boris Zadov. 2020. Lamphone: Real-Time Passive Sound Recovery from Light Bulb Vibrations. Cryptology ePrint Archive, Report 2020/708. https://eprint.iacr.org/2020/708
[46] Lindsey O'Donnell. 2020. *'Lamphone' Hack Uses Lightbulb Vibrations to Eavesdrop on Homes*. Threat Post. https://threatpost.com/lamphone-hack-lightbulb-vibrations-eavesdrop/156551/
[47] Y. Ohshio, H. Adachi, K. Iwai, T. Nishiura, and Y. Yamashita. 2018. Active Speech Obscuration with Speaker-dependent Human Speech-like Noise for Speech Privacy. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC '18)*. 1252–1255.
[48] Emmanuel Owusu, Jun Han, Sauvik Das, Adrian Perrig, and Joy Zhang. 2012. ACCessory: Password Inference Using Accelerometers on Smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems and Applications (HotMobile '12)*. Association for Computing Machinery, 6 pages.
[49] Pery Pearson. [n.d.]. *Sound Sampling*. Human Interface Technology Laboratory. http://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/I.B.3.a.SoundSampling.html
[50] Renhua Peng, Binbin Xu, Guoteng Li, Chengshi Zheng, and Xiaodong Li. 2018. Long-range speech acquirement and enhancement with dual-point laser Doppler vibrometers. *2018 IEEE 23rd International Conference on Digital Signal Processing*, 1–5.
[51] C. Phunruangsakao, P. Kraikhun, S. Duangpummet, J. Karnjana, M. Unoki, and W. Kongprawechnon. 2020. Speech Privacy Protection based on Controlling Estimated Speech Transmission Index. In *Proceedings of the 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON '20)*. 628–631.
[52] Ding Qi, Nan Longmei, and Xu jinfu. 2018. A Speech Privacy Protection Method Based on Sound Masking and Speech Corpus. *Procedia Computer Science* 131 (2018), 1269–1274. https://doi.org/10.1016/j.procs.2018.04.342
[53] Roborock Technology. 2020. *Roborock S5 Max Robot Vacuum & Mop Cleaner*. https://us.roborock.com/pages/roborock-s5-max
[54] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with Your Robot Vacuum Cleaner: Eavesdropping via Lidar Sensors. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys '20)*. Association for Computing Machinery, 354–367.
[55] Samsung. 2020. *What is S Voice?* https://www.samsung.com/global/galaxy/what-is/s-voice/
[56] Jianhua Shang, Weibiao Chen, Huaguo Zang, Yan He, and Dan Liu. 2009. Laser Doppler Vibrometer for Real-time Speech-Signal Acquirement. *Chinese Optics Letters* 7, 11 (2009), 732.
[57] Prakash Shrestha, Manar Mohamed, and Nitesh Saxena. 2016. Slogger: Smashing Motion-Based Touchstroke Logging with Transparent System Noise. In *Proceedings of the 9th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '16)*. Association for Computing Machinery, 67–77.
[58] SoundHound. 2020. *Discover, search, and play any song – featuring voice control.* https://www.soundhound.com/soundhound
[59] Mike Szczys. 2010. *Laser Mic Makes Eavesdropping Remarkably Simple*. HACKADAY. https://hackaday.com/2010/09/25/laser-mic-makes-eavesdropping-remarkably-simple/
[60] Cees Taal, Richard Hendriks, R. Heusdens, and Jesper Jensen. 2011. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *Audio, Speech, and Language Processing, IEEE Transactions on* 19 (10 2011), 2125 – 2136. https://doi.org/10.1109/TASL.2011.2114881
[61] Tara Seals. 2019. *Popular Samsung, LG Android Phones Open to 'Spearphone' Eavesdropping*. https://threatpost.com/samsung-lg-android-spearphone-eavesdropping/146625/
[62] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. 2017. WALNUT: Waging Doubt on the Integrity of MEMS Accelerometers with Acoustic Injection Attacks. *Proceedings of the IEEE European Symposium on Security and Privacy*, 3–18.
[63] Jackie Ward. 2016. *Bugs and Other Eavesdropping Devices*. CBS San Fransisco Bay Area. https://sanfrancisco.cbslocal.com/2016/05/13/hidden-microphones-exposed-as-part-of-government-surveillance-program-in-the-bay-area/
[64] Ryan Whitwam. 2019. *Researchers Turn Hard Drives Into Covert Listening Devices*. ExtremeTech. https://www.extremetech.com/electronics/287324-researchers-turn-hard-drives-into-covert-listening-devices
[65] Wikipedia. 2019. *Google Now*. https://en.wikipedia.org/wiki/Google_Now
[66] Davey Winder. 2020. *How Hackers Use An Ordinary Light Bulb To Spy On Conversations 80 Feet Away*. Forbes. https://www.forbes.com/sites/daveywinder/2020/06/14/how-to-use-an-ordinary-light-bulb-to-spy-on-conversation-80-feet-away-security-research-lamphone-hack/#a309f705be1d
[67] Zhi Xu, Kun Bai, and Sencun Zhu. 2012. TapLogger: Inferring User Inputs on Smartphone Touchscreens Using on-Board Motion Sensors. In *Proceedings of the Fifth ACM Conference on Security and Privacy in Wireless and Mobile Networks (WISEC '12)*. Association for Computing Machinery, 113–124.
[68] Li Zhang, Parth H. Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. AccelWord: Energy Efficient Hotword Detection through Accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '15)*.

A APPENDIX



(a) Setup used in Gyrophone [42]

(b) Setup used in AccelEve [13] and Kinetic Song Recognition [41]

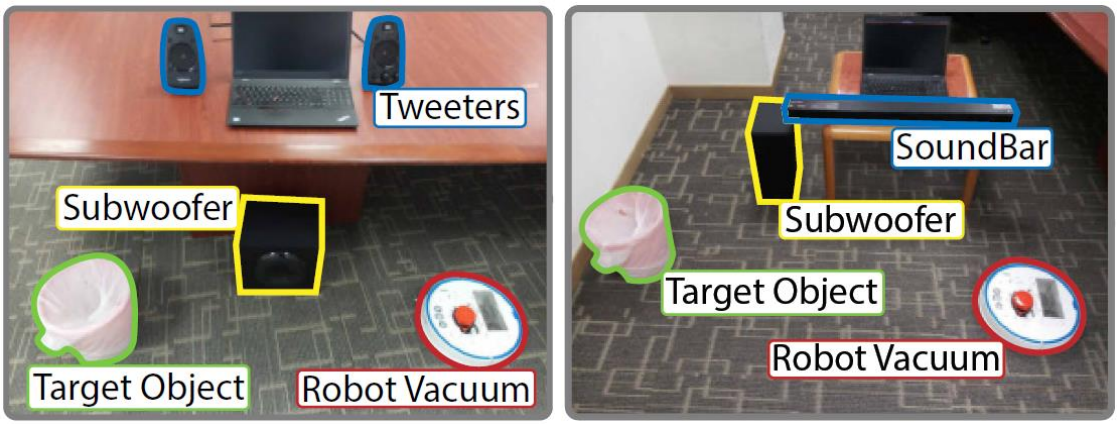
(c) Setup used in Lidarphone [54]

(d) Setup used in Lamphone [45]

Figure 3: Figures depicting different experiment setups used in some of the prior works we evaluated. We can see favored conditions in the experimental setups (e.g., shared surface propagation, close distance between speech source and target object) that would not occur in a real life attack against *live human speech*.

(a) Sensing via MEMS in Gyrophone attack [42]

(b) Sensing via Electro-Optical Sensor in Lamphone attack [45]



(c) Sensing via Lidar in Lidarphone attack [54]

Figure 4: Images of experiments conducted in three vibration-based side-channel attack papers that eavesdrop speech via different sensor methodologies.

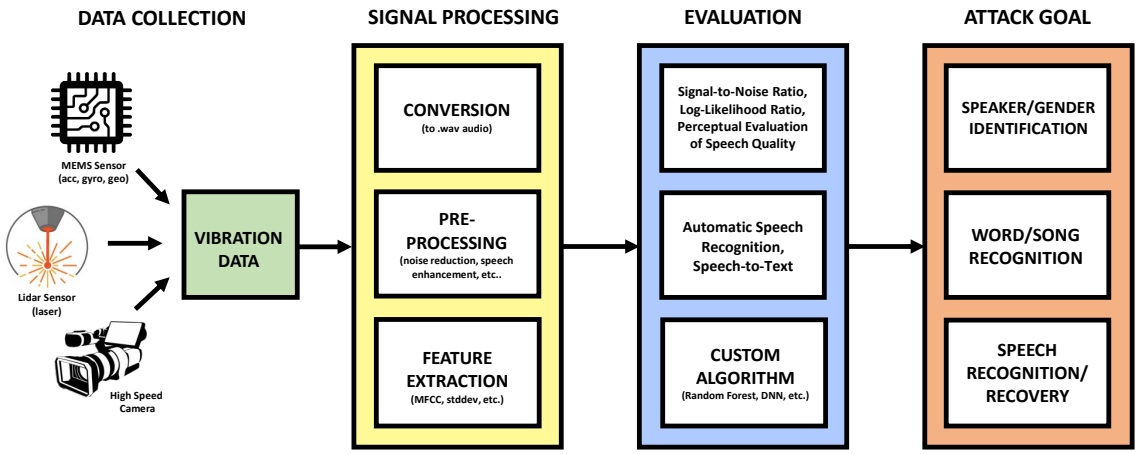
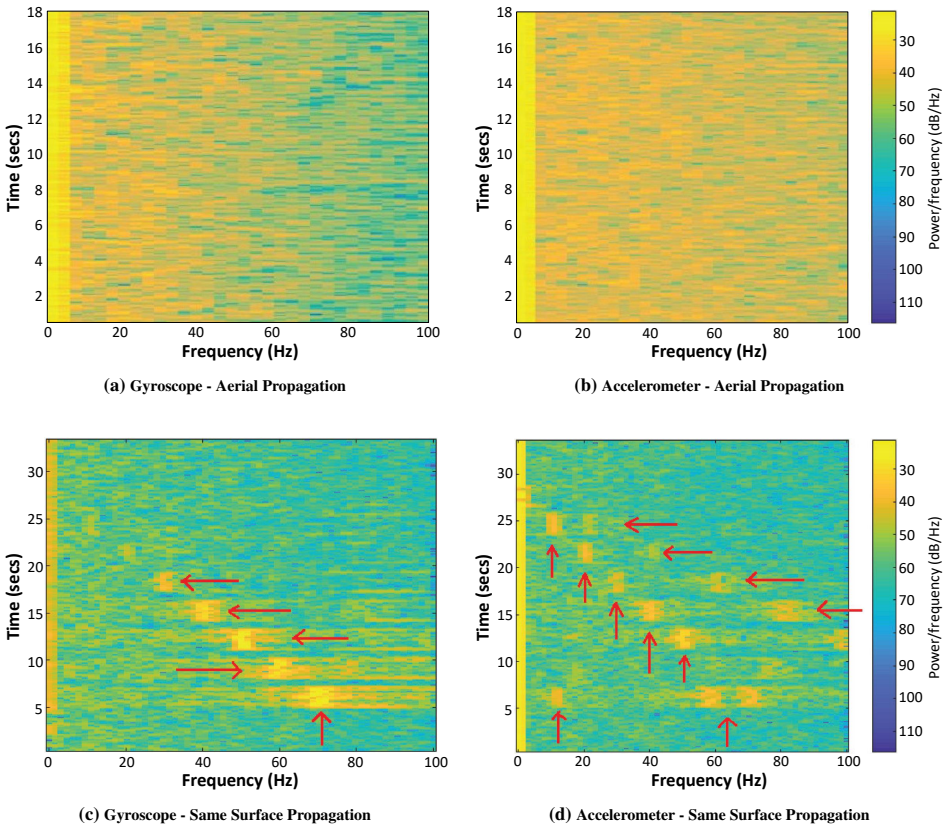


Figure 5: In a vibration-based speech attack the attacker discretely collects information from the system without affecting it. The attacker can use the data it collects, via some sensory equipment, to execute additional speech-based attacks.



(a) Gyroscope - Aerial Propagation

(b) Accelerometer - Aerial Propagation

(c) Gyroscope - Same Surface Propagation

(d) Accelerometer - Same Surface Propagation

Figure 6: Frequency spectrum graphs generated from gyroscope and accelerometer data in the presence of speech propagating aerially vs over a shared surface - from the work Speechless [9]. We can clearly see that speech leakage is significantly reduced when the propagation medium changes from the (favored) same surface setting to the (realistic) aerial propagation setting.