Laser Meager Listener: A Scientific Exploration of Laser-based Speech Eavesdropping in Commercial User Space

Payton Walker
Computer Science and Engineering Department
Texas A&M University
College Station, Texas, USA
prw0007@tamu.edu

Nitesh Saxena
Computer Science and Engineering Department
Texas A&M University
College Station, Texas, USA
nsaxena@tamu.edu

2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P) | 978-1-6654-1614-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/EUROSP53844.2022.00040

Abstract—**Human speech signals produce sound waves that induce vibrations on objects that they encounter. Such vibrations can be measured via *laser vibrometers* and possibly used in speech eavesdropping attacks. However, there is still much to learn about when this attack is feasible. In this paper, we aim to broaden our understanding of the viability of laser eavesdropping attacks to compromise speech in the commercial user space. In our study, we design experiments to measure the subtle vibrations induced on *commonly-available objects* by nearby speech, using commercially sold, high-precision laser vibrometers. To observe idealized success rates of the attack, we maintain certain physical parameters in favorable conditions that represent best case scenarios for an attacker. We test three primary attack scenarios considering different relative positions to the target object. Additionally, we consider many important experimental parameters to understand the generalizability of the attack, including: *speech sources*, *loudness levels*, vibration *propagation media*, and object *materials*.**

Our vibrometer recorded signals were analyzed via a two-pronged methodology including, (1) *time domain*, *frequency spectrum*, *cross correlation*, and *speech intelligibility metric* analyses and (2) an information extraction analysis using both *human listeners* and *automated recognition tools*. Our results suggest that eavesdropping attacks using a laser vibrometer may be practical in some situations and parameter settings (i.e., intelligence missions). However, we find that live aerial human speech and machine-rendered speech at a normal conversational loudness level does not show signs of significant leakage in our analysis.

1. Introduction

The idea of using lasers for spying over human speech conversations has existed since 1940, predating the actual invention of the laser in 1960. Mainly we hear about laser eavesdropping being used in intelligence operations by the military to spy on potential threats and protect national security. It may have been used in the past by the US against Russian embassies [1]. It is also believed that the CIA used a "laser microphone" to learn that Osama bin Laden was hiding in a building in Abbottabad [1]. Used for law enforcement, espionage or otherwise, the implications of such laser eavesdropping are clearly far-reaching and potentially devastating for the victims. Beyond the classical examples referenced above, numerous online articles, blogs and videos seem to suggest the possibility of real-life laser eavesdropping attacks (e.g., [2]–[6])[1]. In this work, we make new contributions exploring this attack methodology, in the commercial space, that furthers our understanding of its accuracy and feasibility. We ask ourselves the following research question: ***How much speech information can be learned using a publicly available laser vibrometer and in which scenarios?***

Speech and vibration are very closely related concepts. Vibrations in our vocal chords generate sounds that we formulate into speech in the form of sound waves. As these sound waves interact with objects, they induce vibrations proportional to the original speech signal [7]. Therefore, accurate measurements of the induced vibrations could potentially be used to reveal the actual speech. This functionality can be found in a technology called a "laser vibrometer". A laser vibrometer uses the concept of Laser Doppler vibrometry (LDV) to measure the subtle vibrations (i.e., displacement) of an object with very high precision. This raises concerns around the eavesdropping potential of such equipment. There is specific interest in investigating such an eavesdropping attack in the vibration domain for a few significant reasons. Unlike traditional microphone devices that require a close distance to the speech source to adequately record the speech and are easily hindered by a solid barrier between the microphone and the speech source, an attacker armed with a laser vibrometer could listen to private conversations without being near the speakers. Even transparent barriers such as windows may not be sufficient to hinder a laser vibrometer from eavesdropping on speech related information. As long as the attacker has a visual line of sight to the victim or an object near the victim such that the laser beam can travel to the target unobstructed, speech information could be learned using a laser vibrometer.

In this work, we study the practicality of using state-of-the-art, commercially-available, laser vibrometers to measure speech signals. We look to better understand the scope of such an attack in the commercial user space and provide insights into how such an eavesdropping attack

This work is partially supported by the National Science Foundation (NSF) under the grants: CNS-1714807, CNS-2030501, CNS-2139358.

[1] There are many "spy" microphones and kids toys being sold for under $100 in shops and online that claim to be capable of spying on speech. We are skeptical that such cheap gadgets could perform well at this challenging task and are motivated to use more sophisticated technology built by a large company.

DOI 10.1109/EuroSP53844.2022.00040

could be successful. While the discussion about this topic is present [2]–[6], we want to broaden our understanding of the attack using controlled experimentation and in-depth analysis.

Our study is three-fold. **First**, we design a specific speech eavesdropping attack using a laser vibrometer. Inspired by the real-world situation presented in [1], we chose a cup (filled with liquid) and a glass window as our objects of measurement. In the article, an intelligence expert describes that such an attack could be possible in a real office space by fixing a laser on a cup that was inside the room where the targeted conversation was being held. We define three measurement scenarios depicted in Figure 1. **Second**, we use different cups to establish what materials are more or less susceptible to such an eavesdropping attack and to generalize our results. **Third**, we consider multiple parameters including: *(1) speech source* (live human and machine-rendered via loudspeaker), *(2) sound pressure level (SPL)* or speech loudness (normal human conversation (40-60 dB) and loud (>70dB)), and *(3) propagation medium* (aerial, shared surface).

We chose to include the loudspeaker scenario because it allows us to test at greater loudness levels than normal human conversational loudness and it captures some real-world scenarios in the commercial space (i.e., business meeting with conference call). It also has a clear application for the gender recognition task because the speaker would not be visible in the room. That said, we still believe the scenarios involving live human and machine-rendered speech (at normal conversational loudness) traveling through the air are the most crucial to our research as they capture scenarios involving actual (potentially sensitive) conversations.

Our Contributions and Result Summary: We summarize our key contributions and results below:

- ***Design and Characterization of Attack Scenarios:*** We design a passive LDV attack to eavesdrop on live human and machine-rendered speech. We consider different positions of the malicious attacker, target speech source, and point of measurement and define three different attack scenarios; (1) the vibrometer has an unobstructed line of sight to the target cup (*Direct-Contact*), (2) a glass barrier sits between the vibrometer and the target cup (*Glass- Barrier*), and (3) the laser is focused on the surface of a nearby glass window (*Glass-Surface*). Each of these scenarios encompasses a different setup of the actors in the system. The attack scenarios are fully defined in Section 3.
- ***A Measurement Study Scoping Wide Parameterizations:*** We measure and evaluate the scenarios introduced in the first item across a wide set of parameters, and we acquired two state-of-the-art laser vibrometer devices as our measurement equipment. To our knowledge, our study has more controlled parameter settings than any other previous work on this attack, elaborated in Section 4.1.
- ***Multi-Pronged Analysis of Speech Information Leakage:*** Our research investigates potential speech information leakage from data collected via laser vibrometry. We first analyze the signal collected in the vibration domain by inspecting the time domain and frequency spectrum graphs of the signal to identify any indicators of information leakage. Then, we perform cross correlation between the measured signal (after noise reduction and speech enhancement) and the original speech signal. The raw vibration data collected in each scenario was converted to .wav sound files without compression. Additionally, we perform an attack on speech that considers two primary goals for an attacker; *speech recognition* and *(speaker) gender recognition*. We use both live humans and Automatic Speech Recognition (ASR) tools to achieve this goal. The sound files described for the cross correlation analysis above were re-used in Amazon Mechanical Turk studies to observe how well human listeners perform.

Implications and Scope of Our Work: Our works suggests that laser eavesdropping in the commercial space can potentially be successful when certain key conditions are met (i.e., loudness of speech, material of target object, etc.). However, this attack does begin to falter under certain parameter values (i.e., speech at normal conversational loudness, realistic distance between attacker and victim speech, etc.) This is especially the case for scenarios involving live human speech whereby we suspect that the produced sound waves are not strong enough to induce significant enough vibrations for compromising the speech. A caveat to these implications is that with highly advanced and high-cost laser technology [8] and the possibility of advanced speech extraction techniques, such laser eavesdropping attacks could have more positive results than what we observed. This is especially true in national intelligence scenarios, with their level of resources, which are outside the scope of this work.

We do recognize that certain settings can make this attack successful. Yet, we chose to reproduce *favorable* attack settings for an intentional bias in our experiments in order to test our negative hypothesis. By testing the attack's potential under ideal conditions, we can make new insights about the limiting conditions of this attack. Our work does not present a new attack, but rather focuses on exploring the application of laser vibrometry for speech eavesdropping.

2. Background: Laser Doppler Vibrometry

The change in a wave's frequency as it encounters an object in motion is called the Doppler effect [9]. In terms of Laser Doppler vibrometry (LDV), the frequency of the light beam (laser) shifts in proportion to its velocity as it is reflected off of the moving object that it is measuring. This effect is used by the vibrometer to measure vibrations. It can measure vibrational displacement, velocity, or acceleration of an object. Additionally, the data acquisition system processes the voltage signal generated by the interferometer and digital decoding electronics. The AC voltage signal is created by converting the frequency shifts recorded by the laser.

Converting Vibration to Speech: The AC signal acquired by the LDV is proportional to the instantaneous velocity and can be stored in digital format (.wav in our case). This is similar to how a microphone converts the signals

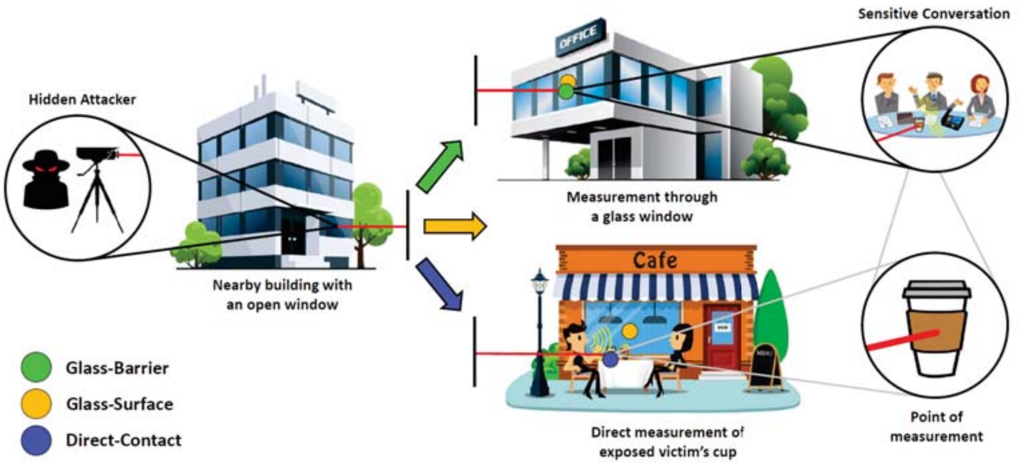

Hidden Attacker
Nearby building with an open window
OFFICE
Sensitive Conversation
Measurement through a glass window
Cafe
Direct measurement of exposed victim's cup
Point of measurement
Glass-Barrier
Glass-Surface
Direct-Contact

Figure 1: A malicious attacker can take advantage of three potential attack scenarios to eavesdrop on speech. Green: an attacker can direct their laser through the window of a building and focus on a cup that is near the speech source (*Glass-Barrier*). Yellow: an attacker can focus their laser on the surface of a nearby window (*Glass-Surface*). Blue: an attacker can focus their laser directly onto an exposed cup without passing the laser through a glass barrier (*Direct-Contact*).

produced by the vibrating diaphragm into audio signals. We are focused on speech reconstruction capabilities so we do not use the raw vibration data that is stored in ASCII format.

Commercial Availability: The company Polytec offers a wide array of vibrometer equipment that is available for purchase by the public including; single-point vibrometers [10], full-field vibrometers [11], microscope-based vibrometers [12], and special-application vibrometers [13]. Our study focuses on attacks using the portable, single-point vibrometers such as the PDV-100 and Vibroflex [14] laser vibrometers. Additionally, new LDV technologies have been developed that are capable of detecting speech by measuring window vibrations. Specifically, the Long-Range Laser Listening Device [8] reportedly can measure the vibrations of a window and eavesdrop on the speech inside a room from 500 meters away. Although this device is currently only available to law enforcement and government agencies, it could feasibly become available to the public in the future.

LDV Applications: Offering the highest resolutions for vibration measurements, Laser Doppler vibrometry is currently used in many applications of research and maintenance. Some examples of applications for LDV include turbine durability validation and development of nondestructive testing (Aerospace), recording physiological processes and investigating sound conduction in the eardrum (Medical), and measuring dynamic properties of Hard Disk Drive (HDD) systems and sub-components (Data Storage) [15]. Additionally, LDV can be used for audio and ultrasonic studies [15], which is the focus of our research. Specifically, we study the application of Laser Doppler vibrometry for eavesdropping on speech.We believe the viability and severity of a laser eavesdropping attack is very apparent and we look to investigate this attack further.

Capturing Live Speech: In order to capture a signal for reconstruction, the Nyquist-Shannon theorem tells us that we require a sampling rate that is twice the maximum frequency (Hz) of the signal. This indicates a minimum sampling requirement for any sensor looking to record and reconstruct speech, and that minimum can vary depending on the application. For example, almost all telephony systems have a cutoff frequency of 3.4 kHz which means that microphones with a sampling rate of 8 kHz are sufficient for capturing speech signals for telephone [16]. However, other applications such as automatic speech recognition require even greater sampling rates (12 kHz) [17]. Since LDV technology can capture vibrations with a very high sampling rate of 44.1 kHz, we are confident they can be used to reconstruct live human speech.

In term's of real life situations where speech may be captured using LDV, the challenges are introduced by the strength of vibrations induced by the target speech, and the limiting factors of the environment conditions. When speech must travel aerially to affect a nearby object, there is a necessary loudness of speech (SPL) and travel distance that must be met for strong enough vibrations to occur on the object being measured. If the speech is not loud enough, or it has to travel long distances, there may not be enough energy left by the time it interacts with an object, causing weak to no induced vibrations. Further, the objects material plays a role in whether the sound waves can effectively induce vibrations. Harder materials are not as easily affected by vibrations as softer and thinner materials [18]. In real life scenarios there are sensitive combinations of conditions that must be met for eavesdropping using LDV to be feasible, and this study seeks to better understand what those are.

3. Threat Model & Attack Scenarios

Our threat model was initially influenced by the scenario described in [1]. We do not consider the potentials of law enforcement or other intelligence agencies to perform this attack as the budget and equipment used in those situations are not commercially available. We recreate the nearby cup speech scenario and allow the attacker to position their equipment close to the target object for maximized attack success. In a real life attack, the LDV could be positioned at much farther distances and focus on an object through a window. Compared to standard eavesdropping attacks (i.e., wiretapping, bug), laser-based eavesdropping can be setup and performed from much greater distances and allows the attacker to go undetected. We chose a cup as our target object because they are common in both business and personal settings. We included water in the cup in order to foster any induced vibrations.

Threat Scenarios: As an illustrative example within our threat model, we consider a real-world situation where a small business meeting is happening in a standard conference room that has glass windows. There are two or more people physically in the meeting with some other attendees participating via a conference call. In the meeting, sensitive information pertaining to business operations is discussed and it is important that this information remains private. The physical attendees of the meeting may have a beverage that sits close to them on the conference table and is susceptible to induced vibrations from surrounding speech. Additionally, the remote attending person's speech is rendered by the speakerphone/conference phone that will also create vibrations which may be used to capture speech. Using vibration measurement technology an attacker can record those vibrations from some distance. Our hypothesis is that certain parameter settings may significantly limit attack success, and reveling those among our other *ideal* experimental settings will increase our understanding of the attack's practicality. Therefore, our results may act as a baseline for understanding the attack's potential.

The threat model considers the possibility of speech sound waves to induce a vibrational impact on the cup containing the beverage and a nearby window. If strong enough vibrations occurred and could be measured from the cup or window, there is a potential for speech information leakage. We also consider a more exposed scenario in which two people are sitting outside at a cafe having a private conversation (i.e., both the attacker and target speech are external). Here the attacker can focus the laser on the target object without passing through a window or at the surface of a nearby window. Figure 1 shows real-world scenarios that influence our experimental design. Notably, while the threat scenario we define considers a business setting and a public cafe, similar attack principles would also apply to eavesdropping speech in other specific scenarios (i.e., a person's home).

Attacker Specifications: The attacker in our threat model is equipped with a laser vibrometer that they can use to measure the minute vibrations of an object. They are able to achieve certain ideal conditions for their attack that are not likely in a realistic scenario, but will increase the feasibility and success of the attack. These settings include a close distance between the speech source and target object and a single speaker/volume per scenario. We also assume our attacker has access to off-the-shelf signal processing and speech recognition tools. Increasing the sophistication of the attacker and their skills will decrease the generalizability of our results as only a few specialized attackers would have such skills. Therefore, we design our attacker to better encompass a standard level of eavesdropping capabilities.

Attack Scenarios: We define three commercial attack scenarios that consider different relative positions between the (1) attacker, (2) speech source and (3) point of measurement, termed "internal" and "external". From a fixed location, the vibrometer focuses its laser on the surface of an object that is near a conversation. Speech from the conversation induces vibrations on the object which are measured by the laser. The three scenarios we consider are more or less practical and can be implemented in a live attack situation. To clarify, although the vibrometers in our setup use a visible laser that could reveal the attack, an attacker could also use an *infrared* sensor head to make the attack more surreptitious. While infrared lasers would have a different wavelength, the same concept of measuring fluctuations in that wavelength would be implemented to measure the vibrations.

In the *Direct-Contact* scenario, the laser has an unobstructed line of sight to the object that it measures. This represents a more favorable measurement setting in which this attack is most likely to be successful. The outdoor cafe scenario described earlier is an example of a setting where this type of attack could occur.

The *Glass-Barrier* scenario introduces a measurement obstacle between the vibrometer and the target object. This scenario represents a realistic situation in which an (external) attacker must direct the laser through a glass window to focus on an (internal) object. This scenario allows us to investigate the affect that measuring through glass has on the laser and whether it degrades measurement quality.

Lastly, the *Glass-Surface* scenario considers the potential for a nearby window to pick up speech vibrations. Here, we are able to test the situation where an (external) attacker tries to eavesdrop on (internal) speech by collecting data at an (external) point of measurement (i.e., outside surface of a window). So the attacker can potentially eavesdrop on speech on the other side of the window, while remaining completely hidden. Additionally, we are able to test the situation where an (external) attacker tries to eavesdrop on (external) speech (i.e., outdoor cafe) when there is no object like the cup, but there is a nearby window surface.

4. Our Methodology

We observe the impact of induced vibrations on a cup and a window, near a speech source, and investigate the potential speech leakage when those vibrations are measured using a laser vibrometer. Our methodology provides a broad insight into the viability of such an eavesdropping attack (or lack thereof) with high precision vibration measurements.

4.1. Experimental Parameters

Speech Source: Our methodology considers two speech sources: live human and machine-rendered (i.e., loudspeaker device). We use live human speech to determine if a conversation at a normal loudness level can cause vibrations on a nearby cup. The loudspeaker device allows us to replicate live human speech as machine-rendered speech and test its effect in a louder setting. The use of live human speech captures the most realistic scenario in which a live conversation may induce vibrations on nearby objects. Additionally, the use of machine-rendered speech from a loudspeaker device captures the scenarios in which a speakerphone/conference phone is used during a private meeting. As the phone renders speech of the remote attendee, vibrations produced from the phone propagate through the shared surface (i.e., conference table) to other objects on the surface, such as a cup.

Sound Pressure Level: Sound Pressure Level (SPL) refers, or “loudness” of a sound, is measured in decibels (dB) and it is estimated that the SPL of normal human conversations is between 40-60 dBs [19], [20]. As part of our methodology, we define two loudness levels; normal and loud. The Normal loudness setting refers to speech with an SPL of 40-60 dB. The Loud setting refers to speech with an SPL of >70 dB. Although it is unlikely that sensitive information is ever spoken or played at a loudness above 70 dB, we felt it was important to test in this decibel range as it allows us to gain a broader understanding of the threat and how increased loudness affects the propagation of vibrations. We use a digital sound level meter to verify the SPL (dB).

Vibration Propagation Medium: We test two different propagation mediums in which the vibrations from sound waves can travel. The *aerial* propagation medium refers to the air space. Sound waves that are projected in to the air, such as in live human speech, propagate through the air and impact the surrounding objects. Additionally, we define the *same surface* propagation medium for our machine-rendered speech scenario. Here, the speech source (i.e., loudspeaker device) shares a surface with the target cup, allowing any vibrations from the sound waves to propagate through the shared solid surface and directly into the cup. Again, an example of this scenario would be the vibrations from a conference phone propagating through the table that it is sitting on and into a cup that is sharing a surface.

Material: We also look to investigate how different materials react when exposed to vibrations. Our methodology uses three different cup materials, and a glass window, for our experiments. In the setting described in [1], the cup material is plastic. Therefore, we chose a standard 16oz plastic Solo cup [21] because of their popularity. Considering that coffee is a popular workplace drink that someone may take with them into a meeting, we also look at disposable cups used for coffee. We chose a standard 16oz Hefty paper coffee cup [22]. We also chose to use a 16oz Styrofoam cup [23] as they are extremely common in office settings. We suspect that some cup materials will be more susceptible to induced vibrations than others because of different physical properties. In each of our scenarios the cups are filled with the same amount of water for consistency. We also investigate how induced vibrations affect the glass of a nearby window. Our experiments use a standard double-pane window that was installed in the past year to represent windows that an attacker may encounter.

4.2. Favorable Conditions

We control a few experimental parameters in settings that will support speech eavesdropping success. Mainly, we look to increase the strength of induced vibrations on the target object, and reduce the difficulty of the signal processing tasks that the attacker must complete. We hypothesize that deviation towards more realistic values will only increase the difficulty of the attack and lower its potential for success.

Number of Speakers: Our experiments use single speaker audio in order to induce the strongest speech vibrations. Adding additional speakers at different loudness levels will increase the complexity of the speech recognition task. The attacker would have to perform additional signal processing to isolate individual speakers and their speech.

Speech Distance: In our experimental setups, we use a short distance (0.5 meters) between the speech source and the target object as another effort to capture higher quality vibration data. Sound waves from speech become weaker as they travel through the aerial medium, and induce the strongest vibrations on the nearest objects. In a real life situation where the speaker is likely farther from the target object, induced vibrations will be weaker and attack potential will decrease.

Vibrational Noise: Background vibrational noise in the environment can negatively affect the success of an eavesdropping attack. Since the target speech is extracted from the vibration domain data, any unrelated vibrational noise that is captured will make it more difficult to process the signal effectively. Depending on the level of background noise, the significant speech related data may be masked or obfuscated. We perform all of our experiments in scenarios that have no added vibrational noise in the environment so that signal processing does not become more complex.

Vibrometer Angle: In our experiments the laser beam is not angled so that it intersects the point of measurement perpendicularly. This position allows for the most accurate vibration measurements with the least error. The vibrometer device manual describes an error threshold for measurements that is determined by the angle of intersection from the laser to the target surface. As the laser angle deviates from the perpendicular position, the error margin increases.

Attacker Budget: The attacker in our model has a generous budget to purchase high-precision LDV technology. Although such a significant budget is not likely for a common day-to-day attacker, we want to demonstrate the eavesdropping potential of top-of-the-line, commercially available laser sensors. More realistic attackers would probably use lower fidelity sensors (laser or otherwise) which cannot capture the same quality of vibration information that we do in our experiments. A lower budget will certainly make the attack more difficult to achieve.

4.3. Equipment

We use high precision laser vibrometers, supplied by the company Polytec; the PDV- 100 portable laser vibrometer and the Vibroflex modular vibrometer [14]. The PDV-100 laser vibrometer can measure vibrational velocity with a frequency of up to 22kHz. The second vibrometer model that we used to collect data is the Vibroflex vibrometer equipped with the VFX-I-120 sensor head, VX-08 decoder, and VFX-O-SRI short distance lense. This setup has very high precision with a displacement resolution of 0.3 pm (i.e., can measure movements as small as 0.3 pm in distance). These devices represent LDV systems that are commercially available for purchase and within our attacker’s $50K budget. The PDV- 100 and Vibroflex units, along with the necessary software, cost around $28 K and $47 K, respectively.

For our machine-rendered speech scenarios, we use the Sony SRS-XB2 portable speaker as our loudspeaker device. The SRS-XB2 speaker device has a sampling rate of 44.1 kHz and a frequency transmission range of 20 Hz - 20 kHz. Using a loudspeaker allows us to increase the loudness of the speech past the normal range of human conversation. Additionally, the loudspeaker is a viable alternative to live human speech because of its ability to reproduce low frequencies. To ensure we met our loudness level requirements in each scenario, we used an SLM305 digital sound level meter. We found that the SLM305 sound level meter was adequate for our purposes because it samples twice per second, has a frequency response of 31.5 Hz to 85 kHz, and has a measuring range of 30 dB to 130 dB. For each experiment, the sound level meter was held at the point of measurement (i.e., beside the cup) to confirm the speech loudness (dB) was at the correct level at the moment the sound waves encountered the cup.

4.4. Experiment Setups

Each of our experiments used the same general setup. The laser vibrometer was attached to a tripod and positioned at a height equal to the height of the point of measurement (i.e., on cup or window surface). The vibrometer was located approximately 3 meters from the targeted object with the laser parallel to the floor and focused on the point of measurement. While the vibrometer has maximum measurement distances close to 30 meters, we chose a shorter distance to avoid complications and remain inline with our approach to test best case scenarios. Through initial testing with the cup, we found that focusing the laser on the surface of the liquid was not effective because the liquid surface is not reflective so we decided on measuring the side of the cup. We do believe measuring any point on the outside of the cup will yield similar results as vibrations propagating through the cup should disperse evenly throughout the entire object. For each scenario, the cup was filled with water and the amount remained consistent for every scenario. We chose water as our liquid for ease in conducting multiple experiments. As liquid conducts vibrations very efficiently, we suspect that it will enhance any induced vibrations.

For the Direct-Contact measurement scenario, there were no obstructions between the vibrometer and the cup. Measurements were taken in a closed/private lab space where all of our experimental parameters could be controlled. For the Glass-Barrier measurement scenario we initially setup outside of a conference room, framed by glass dividers, for measurements taken with the PDV-100 vibrometer. The glass divider acted as the "window" and the laser was pointed through the glass and focused on the cup inside the conference room. For the Vibroflex vibrometer measurements, we constructed our experimental setup in a computer lab space where we used an actual window for our glass barrier. Lastly, for the Glass-Surface measurement scenario we utilized the same computer lab setup as described above. However, instead of directing the laser through the glass at a cup on the other side; we focused the laser on the glass surface. Again, the speech source was located on the other side of the window with the played audio directed towards the window for maximum induced vibrations. Appendix Figures 5a and 6a in show images of our experimental setups for the PDV-100 and Vibroflex vibrometer collected data, respectively.

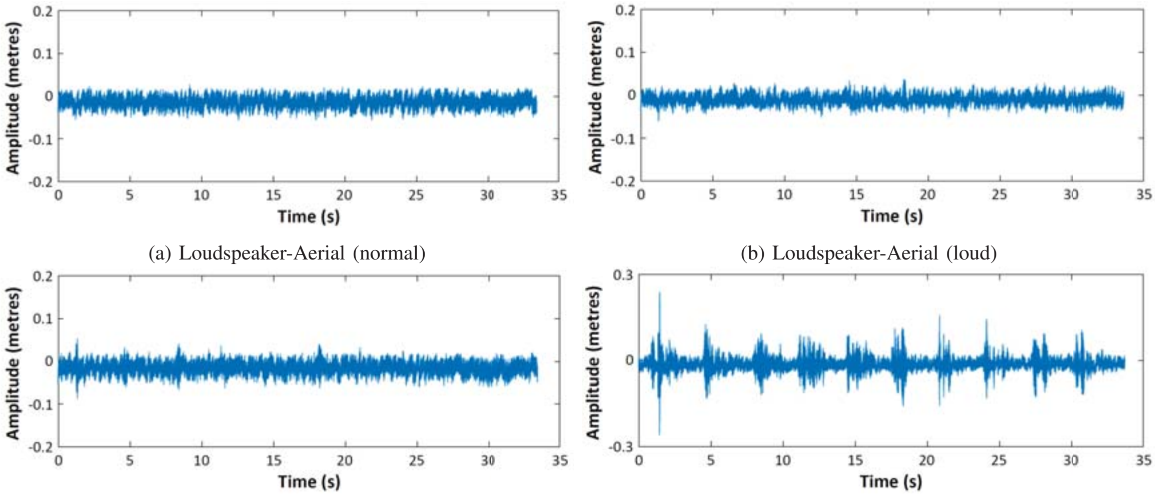
All experimental parameters were explored in the three scenarios described above. For the live human speech scenarios, a human participant stood inside the conference room, with their mouth approximately 0.5 meters from the cup, and read the transcribed speech samples. The human speaker used a normal conversational loudness when speaking and directed their speech towards the cup to maximize the effect of induced vibrations. For the machine-rendered speech scenarios, the loudspeaker device was either placed on a decoupled surface near the cup for the aerial propagation medium, or on the same surface as the cup for the same surface medium. Specifically, the loudspeaker device was placed on a chair for the aerial scenario. We ensured both the chair and table were mechanically decoupled so that the sound waves from the loudspeaker could only propagate aerially. Appendix Figures 5b, 5c, 6b and 6c show images of the loudspeaker arrangements for the aerial and same surface scenarios.

The vibrometers we use measure along the axis of the laser beam. If the laser beam intersects an object at a non-perpendicular angle, there is a margin of error that is elicited in the measurement. The error is equal to the cosine of the angular deviation of the laser from the perpendicular position. Therefore when the laser is perpendicular, the error margin is 0% (cos(0) = 1). We forcibly induce this condition of least margin of error in our measurements by aligning the vibrometer at the correct height and direction to intersect the target object at a perpendicular angle.

4.5. Data Collection

All data for this study was collected in a quiet office or lab space in order to reduce the effect of any unintended external noise. Data was collected in the absence of speech for each scenario to establish an initial control measurement for our signal analyses, before introducing the external speech.

Speech Samples: For our experiment scenarios that require machine-rendered speech played through a loudspeaker device, we selected prerecorded speech samples. The IEEE Recommended Practices for Speech Quality Measurements [24] provides a set of recorded speech samples from both male and female speakers in the Harvard sentences database. We chose to use speech samples from this database because they are approved and used for testing telephone and Voice over IP systems (inline with our threat scenarios involving remote callers). The sentences in each sample are phonetically balanced to use phonemes at the same frequency that they would appear in the spoken English language. Each speech sample recording contains a set of 10 different sentences spoken by a single male or female speaker. Both are native English speakers without any significant accents. We randomly chose the female-spoken sample labeled "List 1" and the male-spoken sample labeled "List 25" for our experiments. The speech sample from each speaker was replayed three times for a total of 60 sentence samples used per scenario. Additionally,we used the same sentences from the male-spoken samples (List 25) for our live human speaker scenario. For each measurement, the male human speaker reads all 10 sentences from the list. Additionally, for the purposes of speech extraction the most favorable setting



(a) Loudspeaker-Aerial (normal)

(b) Loudspeaker-Aerial (loud)

(c) Loudspeaker-Same Surface (normal)

(d) Loudspeaker-Same Surface (loud)

Figure 2: Time domain graphs of signals captured in the vibration domain in the Direct-Contact measurement scenario, from the plastic cup, using the Vibroflex LDV.

an attacker can target are those in which one person is speaking (i.e., allowing the attacker to avoid processing overlapped speech from multiple speakers). Therefore, we only explore attack scenarios involving a single speaker.

5. Signal Analysis

In this section, we describe our qualitative and quantitative analysis results for each experimental scenario. Due to space limitations, only the most significant graphs/figures supporting our analysis/results are presented in the paper and appendix as illustrative examples. For a full set of graphs, data, and materials we refer to our website: https://sites.google.com/view/laser-meager-listener/home. Throughout our study, our results have been consistently validated via both our quantitative (cross correlation, recognition, metrics) analysis and the qualitative (inspection-based in time and frequency domains) analysis.

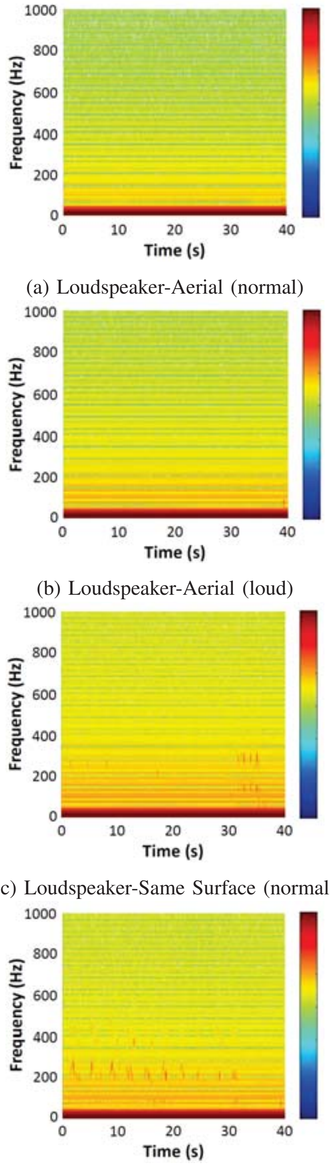
5.1. Analysis Methodology

We approach our analysis of the captured signals in four parts: (1) time domain analysis, (2) frequency spectrum analysis, (3) cross correlation analysis, and (4) a speech intelligibility metric analysis. We visually inspect the time domain graphs of the vibration signals. When the speech is strong enough, we expect to see 10 distinct peaks in the time domain graphs of the vibration signals. These peaks should align with the 10 sentences in the original audio file.

Similarly, we generate the power frequency spectrum from the original vibration signal and look for 10 distinct frequency markers that would correspond to the 10 sentence utterances in the audio. In order to convert a signal from the time domain to the frequency domain, transform functions are used to convert the sum or integral of sine waves in the signal to different frequencies that represent the frequency components of the original signal. We achieve this conversion using the Matlab tool and performing manual Fast Fourier Transform (FFT) operation on the raw data to produce the spectrogram. Through a bit of trial and error,we determined the set of parameters for our transform function that produced the best output (i.e., scale, detail, etc.). The *sampling rate* was set to 9600, the *window size* was set to 1200, the *noverlap* value was calculated as the floor of 2/3 the window size (or 800), and the chosen *nfft* value was 2048. Using the recommended *2ˆnextpow2(data size)* function to determine nfft, we found that our data size was too big and the function could not handle such a large nfft value. Therefore, we calculated the *2ˆnextpow2* of our window size to achieve the 2048 value. The frequency domain graphs show visual representations of vibrational impact and make it easier to identify the measurable effects.

We generate cross correlation graphs that compare the original audio file to the reconstructed audio files (from vibration signal) of each experimental scenario. In a correlation graph, a spike at lag=0 on the graph is an indicator of correlation between the two signals. The amplitude of the peak is equal to the correlation score. We classify the correlation score results from each scenario using the scale defined by [25] where a score of 0.00 - 0.3333 is considered *Weak*, a score of 0.3333 - 0.6666 is *Medium*, and a score of 0.6666 - 1.00 is *Strong*. Through observation we found that *Weak* scores refer to audio with no intelligible speech (0.0 score sounds like noise and 0.33 score sounds like noisy, unintelligible speech). *Medium* scores indicate some intelligible speech and clear speech presence, and *Strong* scores indicate the audio is mostly or completely intelligible to human listeners.

We performed some post-processing on the reconstructed audio files to enhance any underlying speech signal information that may have been captured. Specifically, we use the speech enhancement routines from the speech processing toolbox in Matlab, called VOICEBOX [26]. We initially tested out the routines specific to “speech enhancement”: *specsub*, *spendred*, *ssubmmse*, and *ssubmmsev*. We determined that the *ssubmmse* routine, which uses minimum-mean square error (MMSE) criteria



(a) Loudspeaker-Aerial (normal)

(b) Loudspeaker-Aerial (loud)

(c) Loudspeaker-Same Surface (normal)

(d) Loudspeaker-Same Surface (loud)

Figure 3: Frequency spectrum graphs of signals captured in the Direct-Contact measurement scenario, from the plastic cup, using the Vibroflex laser vibrometer.

for speech enhancement, had the best performance (i.e., speech enhancement and static noise reduction) on our samples. The *ssubmmse* routine was applied to all of our reconstructed audio files for use in the cross correlation analysis, and in the speech recognition studies (Section 6).

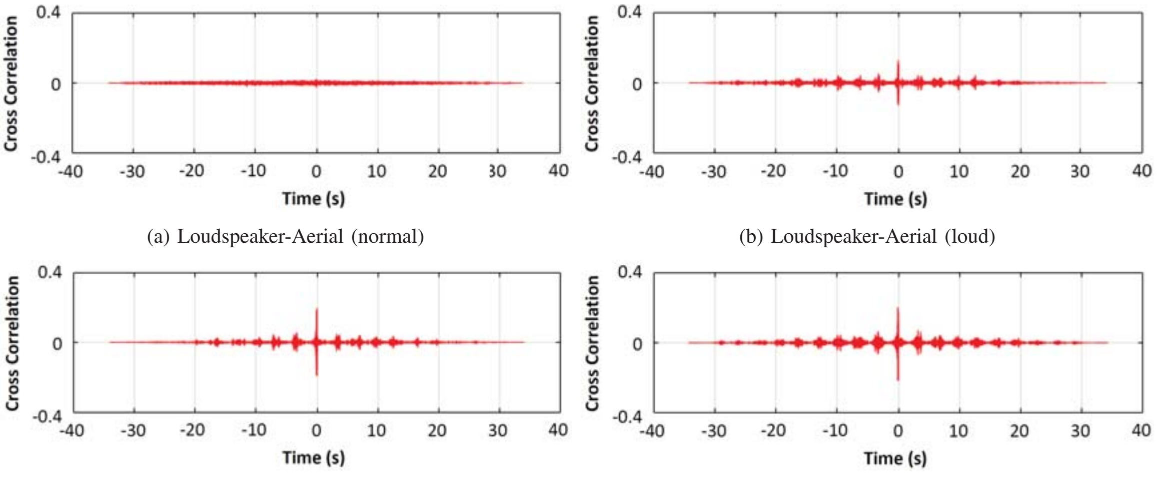
Lastly, we calculated speech metric values to determine presence and intelligibility for the highest quality recovered samples. We use the results from the initial analyses (1- 3) to select the loudspeaker scenarios that show the greatest presence of speech leakage, for each speech source and propagation medium that was tested. From the Direct-Contact scenario, we selected samples from the Loudspeaker-Aerial and Loudspeaker-Same Surface setups with the Loud SPL setting, and the Live Human-Aerial setup with the Normal SPL setting. We also included raw microphone recordings of the clean audio (no noise present) for a comparison of samples containing truly intelligible speech. Each of the recovered audio samples was manually processed to isolate the sentences, resulting in 10 individual samples containing one sentence each, for a total of 60 sentence samples per scenario. We calculate the Signal-to-Noise Ratio (SNR), Short-Term Objective Intelligibility (STOI) [27], and the Perceptual Evaluation of Speech Quality (PESQ) [28] metrics to further explore the quality of speech captured in our recovered samples. The SNR values indicate the overall presence of speech compared to the noise and the STOI and PESQ metrics score the intelligibility of the speech in the recovered sample. Comparing the metric values calculated for our recovered samples with those of the raw microphone recordings indicate the quality of speech (e.g., attack feasibility) that can be retrieved from vibration data.

5.2. Presence of Leakage

Through our initial time and frequency domain analysis, we observe certain parameter settings that consistently indicate the presence of speech information leakage throughout our different experimental scenarios. In our time domain analyses, we look for 10 distinct peaks in the time domain graphs generated from our enhanced samples. These peaks indicate that the speech was able to induce strong vibrations on the cup and may result in speech information leakage in the vibration domain. Additionally,we inspect the frequency spectrum graphs generated from our enhanced samples and look for 10 distinct frequency markers. This would indicate that frequencies unique to the speech were captured in the vibration data which could be used by an attacker to reconstruct that speech. To confirm this assumption we perform cross correlation analysis to compare the enhanced and original speech samples and identify when speech information is shared (e.g., leaked via vibration domain).

Loud SPL: Our time domain analysis revealed that most scenarios with peaks in the time domain graph, that correlate to induced vibrations from the speech, used speech at the Loud SPL level. The time domain graph for the Loud SPL, same surface setting in Figure 2d shows an example of the peaks we are looking for. In the Direct-Contact setting with Loud SPL speech, peaks were observed in the graphs generated from scenarios with the same surface propagation medium for all cup materials, and in the aerial propagation scenario for the Styrofoam cup. In the Glass-Barrier setting, we see peaks corresponding to the speech in both the aerial and same surface propagation mediums, for all cup materials, when the speech was played at a Loud SPL level. This suggests that the measured vibrations induced via the aerial propagation medium are stronger in the Glass-Barrier scenario due to vibrations that are also induced in the glass. Similarly, the aerial propagation medium scenario tested in the Glass-Surface setup also contained distinct peaks in the time domain graph when there was Loud speech.

Our frequency spectrum analysis had similar results to our time domain analysis. In the Direct-Contact, Loud SPL setup, we observed unique frequency markers in the spectrum graphs generated for scenarios with the same surface propagation medium for all cup materials, and



(a) Loudspeaker-Aerial (normal)

(b) Loudspeaker-Aerial (loud)

(c) Loudspeaker-Same Surface (normal)

(d) Loudspeaker-Same Surface (loud)

Figure 4: Cross correlation graphs comparing the original audio to the data collected in the Direct-Contact measurement scenarios, from the plastic cup, using the Vibroflex laser vibrometer.

the aerial propagation medium for the Styrofoam cup. Figure 3 depicts the frequency spectrum graphs generated from samples collected in the Direct-Contact scenario and Figure 3d shows the 10 speech related frequency markers. In the Glass-Barrier scenario with Loud speech, the frequency markers were present in the spectrum graphs for both the aerial and same surface propagation. We identify speech frequencies in all scenarios using the Loud SPL level. This is consistent with our observation that aerially induced vibrations are stronger in the Glass-Barrier scenario. Lastly, we find speech related frequencies in the Glass-Surface, aerial propagation, Loud SPL setup.

We further confirm our previous observations with cross correlation analysis for some settings. In the Direct-Contact scenario, the cross correlation graphs comparing the same surface propagation medium samples to the original speech showed some correlation (although *Weak*) for all cup materials and Loud SPL speech. Spikes in the graph indicating correlation were seen for the aerial propagation medium samples from the plastic and Styrofoam cups. Figure 4 shows the cross correlation graphs with spikes at lag=0. Additionally, our results for the Glass-Barrier and Glass-Surface attack settings were supported by our cross correlation analysis. Appendix Figures 8 & 9 show the cross correlation graphs for these scenarios.

Same Surface Propagation Medium: Another parameter that consistently produced results indicating speech leakage was the Same Surface propagation medium. Along with the results for the Same Surface - Loud SPL scenarios described above, our time domain analysis of the Direct-Contact samples also revealed that some settings Normal SPL speech levels produced graphs with distinct peaks (for plastic and Styrofoam cup materials). Figure 2c shows the few small peaks in the time domain graph of the Same Surface - Normal SPL scenario measured from the plastic cup. In the Glass-Barrier setup, we observed induced vibrations only when speech was played at the Loud SPL level with Shared Surface propagation.

Our frequency spectrum analysis of these samples produced similar results with frequencies unique to the original speech present in the spectrum graphs for scenarios using Loud SPL speech for all cup materials and attack setups. Additionally, these frequency markers were seen in the graphs for Direct-Contact scenarios using speech at the Normal SPL level, for the plastic and Styrofoam cups. Figure 3c shows the frequency spectrum graph for the plastic cup containing distinct markers.

Cross correlation analysis of the Same Surface propagation medium scenarios confirm our observations in prior analyses. We see some correlation (although *Weak*) between the original speech audio and the samples collected when speech was played at the Loud SPL level, for all cup materials and attack setups. Appendix Figure 9c shows the peak indicating some correlation at lag=0 for the plastic cup in the Glass-Barrier, Normal SPL scenario.

5.3. Absence of Leakage

We also identify parameter settings that consistently showed no indication of speech information leakage. The graphs generated from these scenarios contain no time domain peaks or frequency spectrum markers that are unique to the original speech audio. Our cross correlation analysis confirmed this with no spikes to indicate correlation with the original speech audio (e.g., no speech leakage).

Normal SPL - Aerial Propagation: In each of our analyses, there was one parameter combination that showed no sign of speech information leakage throughout each of the attack scenarios. When speech was propagated Aerially, at the Normal SPL level, we found no clear presence of speech information leakage in either our time domain (Figure 2a) or frequency spectrum (Figure 3a) analyses (e.g., no distinct peaks or frequencies specific to the original speech were seen in the generated graphs). The absence of speech information leakage in these scenarios was further supported by our cross correlation analysis. The cross correlation graphs have no spike at lag=0 to suggest a correlation with the original speech, seen in Figure 4a. Uniquely, the cross correlation graph for the Normal SPL - Aerial propagation scenario in the Glass-Surface setup, shown in Appendix Figure 8c, does indicate a *minor* correlation. Since the correlation is on the lower end of the *Weak* classification range, and there was no indication of leakage in the time/frequency domains, there is likely no chance for speech information leakage.

Live Human Speech: We also observe this lack of information leakage from our samples collected when speech came from a Live Human speaker (also Normal SPL level and Aerial propagation). The results of our time domain and frequency spectrum analyses were inline with our observation from the loudspeaker generated, Normal SPL, aerial propagation samples. Again, we further confirm the lack of speech information in these samples with cross correlation. Appendix Figure 7 shows the cross correlation graphs generated from the Live Human speech samples. These figures show no significant peaks to suggest correlation with the original speech.

5.4. Speech Metric Analysis

Our analysis using the different speech metrics demonstrates the lack of intelligible speech in the audio samples recovered from vibration data. Compared to the scores seen for the microphone recordings, we find that the recovered samples have much lower speech presence and intelligibility. Appendix Table 4 shows the average SNR, STOI, and PESQ scores for samples collected in the Direct-Contact, Styrofoam setup. We calculate scores for each Speech Source and Propagation Medium. We selected samples from the Loud SPL setting for the Loudspeaker scenarios because they produced stronger signal recovery. We compare these metric values to those calculated for raw microphone recordings of the original speech audio (at Normal SPL level).

The SNR results demonstrate the decreased presence of the original signal in our recovered samples compared to a plain audio recording. Additionally, we see that moving from the Same Surface to Aerial propagation medium, and then from the Loud to Normal SPL levels result in even lower SNR. Similarly, we can see the perceived intelligibility of our recovered samples is significantly lower than a normal microphone recording. The STOI metric results show a decrease from 0.91 (very intelligible) for the microphone recordings to 0.39 (poor intelligibility) and lower for our recovered samples. This decrease in intelligibility is further confirmed by our PESQ results where we see an immediate decrease by more than half (from 3.79 to 1.50) in the PESQ scores of our recovered samples compared to the microphone recordings. This indicates low intelligibility compared to the clear audio. And for both STOI and PESQ, we observe the same trends for Same Surface to Aerial and Loud to Normal SPL with each metric value decreasing.

6. Human/Automated Recognition Analysis

Having observed the present and absence of speech information leakage, we wanted to assess if this translates to speech and gender recognition. In a two-pronged approach, we used the reconstructed and enhanced audio samples previously described in Section 5. First, we use the power of live human listening and comprehension in a study with live human participants. Second, we utilize popular automatic speech recognition (ASR) tools such as Google and IBM Speech-to-Text services to understand how successful automated tools are when faced with our reconstructed samples. In recent years, speech recognition systems have been vastly improved with some systems achieving comparable performance (i.e., error rates) to actual human listeners [29]. However, the human ear remains superior in understanding speech in noisy environments [30], [31].

6.1. Live Human Study

Study Design: The goal of the study was to investigate the human perception of speech presence and intelligibility in the reconstructed audio samples. We conducted two separate Amazon Mechanical Turk studies to analyze the data collected from each of the two vibrometers. Our signal analysis of the vibration data collected with the PDV-100 identified 23/30 scenarios with the most potential for speech information leakage. Additionally, we included all 26 scenarios measured with the Vibroflex vibrometer.

We designed an online survey in Google Forms [32] to assess speech detection by live humans in the post-processed audio samples. We included the best sample (10 sentences) from each scenario and the samples on each page of the survey were anonymously labeled and appeared in a randomized order. Also, the participants were allowed to replay the samples as many times as they wanted. They were asked to listen to each sample and identify the speaker's gender and transcribe any words or phrases that they could understand. The participants were given a blank textbox where they could type their transcriptions. The participants were instructed to type "NA" in the textbox if they could not understand any words in the audio sample. Our survey had two qualification criteria; all respondents were native English speakers and did not have any known hearing impairments or loss. The survey took about 30 minutes and each participant was compensated $2.00 for completing it.

The survey included a "dummy" audio sample in order to test the participation of each respondent while taking the survey. The sample contained fully intelligible audio of the numeric digits "zero" thru "nine". We used this question to filter our responses and remove any that did not transcribe the dummy sample correctly, which we labeled "invalid" (e.g., did not make a good effort). Both studies were approved by our University's IRB and standard best practices were used to protect the participants' privacy.

Analysis: From the 126 total responses collected from the two studies, we identified 97 valid responses from 40 females, 56 males, and 1 non-binary person. The respondents represent a variety of ages (18-50+), educational levels (High School, Bachelors, Masters), and fields of study (Computer Science, Liberal Arts, etc..). The distribution of these demographics was similar between the two studies. We define five categories (Very Low, Low, Medium, High, Very High) to describe the success of speech recognition. Each category represents a range for speech transcription accuracy: Very Low = [0-10%], Low = (10-30%], Medium = (30-70%), High = [70-90%), and Very High = [90-100%].

Each of these categories define different levels of speech presence in the noisy audio; Very Low success: no speech detection and only noise is heard, Low success: the subtle presence of (unintelligible) speech may be detected

TABLE 1: Average and maximum speech decoding accuracies from our Mechanical Turk study and ASR analysis. Accuracies are reported for both vibrometers and spaces marked "–" indicate scenarios that were not tested for that vibrometer.

				PDV-100			Vibroflex		
				Mturk Study		Google Cloud STT	Mturk Study		Google Cloud STT
Attack Scenario	**Material**	**Speech Source – Propagation Medium**	**Loudness**	**Human Decoding Accuracy: Average (Maximum)**	**Gender Classification Accuracy: Average**	**ASR Decoding Accuracy: Maximum**	**Human Decoding Accuracy: Average (Maximum)**	**Gender Classification Accuracy: Average**	**ASR Decoding Accuracy: Maximum**
Direct-Contact	Paper	Human Speaker-Aerial	Normal	0% (0%)	32%	0%	--	--	--
		Loudspeaker-Aerial	Normal	0% (0%)	2%	0%	0% (0%)	4%	0%
			Loud	0.02% (1.23%)	42%	0%	0% (0%)	11%	0%
		Loudspeaker-Same Surface	Normal	0% (0%)	52%	0%	0% (0%)	4%	0%
			Loud	0% (0%)	0%	0%	0.03% (1.27%)	**85%**	1%
	Plastic	Human Speaker-Aerial	Normal	0% (0%)	0%	0%	--	--	--
		Loudspeaker-Aerial	Normal	0% (0%)	0%	0%	0% (0%)	4%	0%
			Loud	0.05% (1.27%)	8%	0%	0% (0%)	11%	0%
		Loudspeaker-Same Surface	Normal	0% (0%)	78%	0%	0% (0%)	47%	0%
			Loud	7.6% (**38.27%**)	**96%**	3%	4.78% (**29.63%**)	**100%**	2%
	Styrofoam	Human Speaker-Aerial	Normal	0.64% (4.94%)	**84%**	0%	--	--	--
		Loudspeaker-Aerial	Normal	1.58% (7.41%)	**100%**	0%	0% (0%)	40%	0%
			Loud	0.28% (3.8%)	**82%**	0%	0.11% (1.27%)	77%	0%
		Loudspeaker-Same Surface	Normal	0.13% (3.8%)	**94%**	0%	0.08% (1.27%)	**83%**	1%
			Loud	8.89% (**37.04%**)	**94%**	17%	17.81% (**65.43%**)	**98%**	10%
Glass-Barrier	Paper	Loudspeaker-Aerial	Normal	--	--	--	0% (0%)	23%	0%
			Loud	--	--	--	0.08% (1.27%)	**87%**	0%
		Loudspeaker-Same Surface	Normal	--	--	--	0% (0%)	4%	0%
			Loud	0% (0%)	38%	0%	0.03% (1.27%)	**83%**	0%
	Plastic	Loudspeaker-Aerial	Normal	--	--	--	0% (0%)	6%	0%
			Loud	--	--	--	0.13% (3.7%)	**91%**	0%
		Loudspeaker-Same Surface	Normal	0% (0%)	8%	0%	0% (0%)	17%	0%
			Loud	0% (0%)	**90%**	0%	0% (0%)	**87%**	0%
	Styrofoam	Human Speaker-Aerial	Normal	0% (0%)	22%	0%	--	--	--
		Loudspeaker-Aerial	Normal	0% (0%)	4%	0%	0% (0%)	6%	0%
			Loud	2.54% (17.28%)	2%	3%	0.16% (7.41%)	62%	0%
		Loudspeaker-Same Surface	Normal	0% (0%)	**80%**	0%	0% (0%)	0%	0%
			Loud	15.23% (**44.44%**)	18%	3%	0.66% (8.64%)	**83%**	0%
Glass-Surface	Glass	Loudspeaker-Aerial	Normal	--	--	--	0% (0%)	6%	0%
			Loud	--	--	--	0.08% (3.7%)	**83%**	0%

among the noisy audio, Medium success: an equal mix of noise and speech is perceived and can be partially transcribed, High success: almost all of the speech is understood and transcribed, and Very High success: all of the original speech is intelligible in the reconstructed audio and can be transcribed. We classify success using the maximum speech recognition accuracy observed from all responses in a scenario. Additionally, we define the same five categories (using different accuracy ranges) to describe gender recognition success. As a binary task, random guessing achieves 50% success. Therefore, the five categories we define encompass the accuracy range above 50% (below 50% is considered failed for performing worse than random guessing). We define our categories for gender recognition accuracy as: Very Low = [50-60%), Low = [60-70%), Medium = [70-80%), High = [80-90%), Very High = [90-100%].

Results: We calculated the speech decoding accuracy of each user and audio sample as: *# of correctly transcribed words / total # of words in the sample*. We observe the average and maximum speech decoding accuracy across all respondents in each of the tested scenarios. Table 1 displays the accuracies calculated from both MTurk studies. We found the scenarios that showed some presence of leakage in our previous analysis have some potential for speech comprehension by human listeners. In both studies, the samples reconstructed from the Direct-Contact, Same Surface, Loud SPL scenario for the plastic and Styrofoam cups had the greatest accuracies. The maximum decoding accuracies observed from the PDV-100 vibrometer samples were 38% and 37% and from the Vibroflex vibrometer samples were 39% and 65%. Additionally, the samples collected by the PDV-100 vibrometer in the Glass-Barrier scenario had a maximum decoding accuracy of 44% indicating a Medium level of transcription success. All other scenarios had Very Low success with maximum decoding accuracies below 10%. Among the samples collected in the Live Human and loudspeaker (Normal SPL) scenarios with aerial propagation, the maximum decoding accuracy was 7%. Our classifications for the speech recognition success are summarized in Table 2. The results for gender identification suggest a real potential for speaker information leakage. Many of the scenarios tested had 80% or greater gender identification (High or Very High success rate). In Table 2 we summarize the highest success rate observed for gender recognition. Our observations support previous claims made by other works involving eavesdropping on speech using LDV [33]–[35].

6.2. Automated Speech Recognition (ASR) Study

Study Design: For our ASR study we reused the post-processed audio samples from our live human study described above. Initially we tested a few ASR services including Google Cloud Speech Recognition [36], Microsoft BING Voice Recognition [37], and IBM Watson Speech to Text service [38]. We observed similar performances among the Google and IBM recognition tools and ultimately decided on the Google Cloud Speech Recognition system because of its mass popularity and because it has achieved comparable performances to humans in the task of speech recognition [39]. We fed each reconstructed and processed audio file into the Google Cloud Speech Recognition tool which automatically generates transcriptions using four different models; *Default*, *Command/Search*, *Phone call*, and *Video*. We recorded the maximum decoding accuracy achieved between the four models in Table 1. Decoding accuracy is *# of correctly transcribed words / total # of words in the sample*.

TABLE 2: We generalize the results of our analysis for each attack scenario, speech source - propagation medium, and speech loudness. Considering the accuracy results seen across all cup materials and vibrometer devices used, we describe the potential success for the speech and gender recognition tasks using a five value scale (Very Low, Low, Medium, High, Very High). "-": Failed gender recognition (accuracy <50%)

Attack Scenario	Speech Source – Propagation Medium	Speech Loudness	Speech Recognition (Human)	Gender Recognition (Human)	Speech Recognition (ASR)
Direct-Contact	Live Human – Aerial	Normal	Very Low	High	Very Low
	Loudspeaker – Aerial	Normal	Very Low	Very High	Very Low
		Loud	Very Low	High	Very Low
	Loudspeaker – Same Surface	Normal	Very Low	Very High	Very Low
		Loud	Medium	Very High	Low
Glass-Barrier	Live Human – Aerial	Normal	Very Low	-	Very Low
	Loudspeaker – Aerial	Normal	Very Low	-	Very Low
		Loud	Low	High	Very Low
	Loudspeaker – Same Surface	Normal	Very Low	-	Very Low
		Loud	Medium	Very High	Very Low
Glass-Surface	Loudspeaker – Aerial	Normal	Very Low	-	Very Low
		Loud	Very Low	High	Very Low

Analysis: We used the same five category definitions (Very Low, Low, Medium, High, Very High) that we created for human speech recognition to describe the success of ASR. Again, each category represents a range for speech transcription accuracy and are defined as: Very Low = [0-10%], Low = (10-30%], Medium = (30-70%), High = [70-90%), and Very High = [90-100%]. Table 2 shows our classification of speech recognition success by ASR for each attack scenario/ speech source/propagation medium/loudness setting. Our classification was based on the maximum observed accuracies for all cup materials and both vibrometers.

Results: The performance of ASR was not comparable to the human listeners. For the same scenarios, human listeners achieved decoding accuracies of 38%, 37%, and 65%; while the ASR tool achieved accuracies of 3%, 17%, and 9%. At most, the ASR tool was able to achieve Low success in only one of the experimental settings. For all other experimental settings, we find that ASR had Very Low success.

7. Summary & Discussion

Our analysis revealed certain vulnerability patterns between the different parameter settings that were tested. We also investigated speech information leakage via live human listeners and automated speech recognition tools. Overall the live human listeners performed better at decoding the audio samples (achieving up to Medium success for speech recognition) than the ASR tool (only achieving Low success for one setup). Significant decoding accuracies were observed in some limited scenarios suggesting eavesdropping attacks may be successful in optimal conditions. We found that these accuracies were only achieved for *Same Surface* propagation, *Loud* SPL settings. Gender recognition by human listeners was successful (High or Very High) for Direct-Contact, Styrofoam cup scenarios.

Parameter Setting Observations: We observed consistent trends among some of the controlled parameters and the time and frequency domain graphs revealed the subtle changes in induced vibrations across our parameter settings. First, we observed that lower speech volumes result in lower speech presence (Figures 2c & 2d). Additionally, we found that transitioning from the same surface to the aerial propagation medium lowered the speech presence (Figures 2b & 2d). Our speech recognition study results further confirm these findings with similar trends of improved/weakened speech.

Cup Material Observations: Our results across all scenarios indicate different materials do have an effect on induced vibrations and we observed that the Styrofoam material was the most sensitive. We consider the *density*, *elasticity*, *tension* and *temperature* to explain the effects of induced speech [40]. First, the density values for each material [41]–[43] correlate directly to our observation of vibrational impact. Next, we consider the elasticity, or an object's resistance to change [44]. Young's Modulus describes the stiffness of a material [45] and using these constants, we observed the least elastic material (Styrofoam) had the greatest vibrational propagation. The other physical properties, temperature and tension (amount of pulling or stretching force induced [46]) were consistent.

Limiting Factors and Potential Defenses: Since our experimental setup capitalizes on some idealistic settings for the attacker and the live human speech at typical conversational loudness still did not impact the target objects we studied, it seems that a laser eavesdropping attack may be faced with several challenges. Specifically, the loudness level of the target speech is the fundamental limitation of the attack's potential as it determines the strength of induced vibrations. Other factors like increased distance between speech source and target object, vibrational noise in the environment, and multiple speakers will make it more difficult to extract speech information. Therefore, certain defensive measures can be take that capitalize on these challenges. Sound *absorbing* materials placed inside a room can dampen the impact of speech waves on nearby objects and therefore reducing the level of induced vibrations. As these attack must target an object with a clear line-of-sight, blocking the ability to see any target objects would also mitigate the attack (e.g., closing the window blinds). And while not evaluated in the current work, our results suggest that any noise inside the room will only decrease the potential for attack success because of the conflicting sound waves that affect the target object. This can be accomplished using a simple white noise generator inside a room to essentially mask any speech related vibrations.

Speech Scenarios at Risk: From our experimental and

TABLE 3: Comparison of experimental parameter settings used in prior works vs in our current work. "✓": Condition satisfied, "✗": Condition not satisfied, "-": Information not available in literature.

Prior Work	*Year*	*Laser-based?*	*Speech from Live Human?*	*Conversational Loudness?*	*Aerial Propagation?*	*Multiple Materials?*
Li et al. [33]	2006	✓	-	-	-	✗
Shang et al. [35]	2009	✓	-	-	-	✗
Zalevsky et al. [55]	2009	✓	✓	-	✗	✗
Avargel et al. [56]	2011	✓	✓	-	✗	✗
Peng et al. [34]	2018	✓	✗	✗	-	✓
Lv et al. [61]	2019	✓	✗	-	✗	✗
Sami et al. [59]	2020	✓	✗	✗	✓	✗
Wang et al. [62]	2021	✓	✗	✗	✗	✗
Wang et al. [63]	2021	✓	✗	✗	✗	✗
Michalevsky et al. [51]	2014	✗	✗	✗	✗	✗
Davis et al. [54]	2014	✗	✗	✗	✓	✗
Han et al. [53]	2017	✗	✓	✗	✓	✗
Anand et al. [52]	2018	✗	✗	✗	✓	✗
Nassi et al. [58]	2020	✗	✗	✗	✓	✗
Our Work	2022	✓	✓	✓	✓	✓

speech recognition results we can conclude that certain speech scenarios are susceptible to laser-based eavesdropping. Live speech at louder volumes such as in business presentations or group announcements are more likely to result in successful speech recognition. Also, speech from a device (where the device itself can be targeted) would almost certainly be compromised by a laser eavesdropping attack. This is because of the vibrations that are directly proportional to the played speech.

8. Related Work

Previous works have investigated how induced vibrations, measured with low-fidelity sensors, can be used to reveal certain information. Marquardt et al. demonstrated how a mobile application with access to the phone's accelerometer data can determine what text was typed on a nearby keyboard [47]. Keystrokes on the soft keyboard of a mobile phone can also be determined using the onboard motion sensors [48]–[50]. Recent studies have also begun to focus on the eavesdropping implications of such vibration measurements. Michalevsky et. al found that vibration signals recorded by the gyroscope on a mobile phone can be used to perform speaker classification and speech parsing [51]. In [52], Anand et. al showed that the vibrations from a speaker device, propagated through a shared surface, can induce vibrations on the accelerometer that are capable of revealing speech information. Han et. al go even further to perform intelligible speech reconstruction using data from a set of non-acoustic sensors including a geophone, gyroscope, and accelerometer [53]. We use more precise vibration data and encompass a more general attack. A recent work by researchers at MIT investigated a new method for extracting vibration data by using a high-speed camera [54]. This work extracts vibration data from video recordings, relying on image processing techniques. Our work differs because vibrations are measured directly and can be immediately converted to audio data.

Other works have emerged that present unique laser applications for detecting or transmitting speech. Research conducted by Zalevsky et al. showed how a laser can be used to eavesdrop speech directly off of a human body and also detect heart beats [55]. A similar work by Avargel et al. [56] presented a remote speech measurement system that uses an LDV device pointed at a user's throat while they speak in order to detect speech. While detecting speech with an LDV by directly measuring the speech source may be viable, targeting a live human in a live attack scenario would be pose significant challenges. Our work focuses on the attack model that targets motionless objects near to the speech that can be affected by the speech sound waves. Another interesting work was conducted by Sullenberger et al. [57] where the authors define a method for targeted communication using messages encoded on modulated laser beams.

Most relevant to our work are the existing academic studies that investigate laser-based speech detection in the vibration domain by targeting a nearby object. Lamphone [58] is an attack that measures the vibrations of a light bulb that is near user speech in order to eavesdrop. Lidarphone [59] utilizes the lidar sensor of a robot vacuum cleaner to measure speech vibrations from a trash can. We notice however, that certain experimental parameters are minimally explored. In Lamphone [58] only loud machine-rendered speech is utilized which will significantly increase the strength of induced vibrations. In Lidarphone [59] there is a shared medium between the target object and the speaker system which will directly propagate stronger vibrations.

We have also seen academic studies that use the same technology as our work, Laser Doppler vibrometers, and appear to demonstrate speech detection capabilities. A project by Amendolare et al. looked to develop an audio surveillance device that measures the surface of a window to eavesdrop on internal speech [60]. In a work by Li et al. [33], the authors report that they can use an LDV device to remotely recover speech and enhance its intelligibility. Achieving a simpler task, Shang et al. [35] demonstrate real-time speech signal acquirement (i.e., speech detection). Peng et al. [34] designed a system that uses two LDV devices, measuring the same object, to recover speech. In more recent works, authors have implemented LDV speech capture to evaluate the effect of speckle noise [61] and develop new solutions to reduce speckle noise in laser-captured speech [62], [63].

While all of these existing works contribute to the study of laser-based speech capture, our work furthers the general understanding of the attack application's feasibility by contributing a broader set of parameters and settings for evaluation. Uniquely, we include certain parameter settings that have often been overlooked in the existing works. In Table 3 we compiled information about the experimental parameter settings of the most related prior works that explore speech eavesdropping/detection in the vibration domain. We can see that our work stands out by including experiments that represent realistic parameters settings for speech source, speech loudness, propagation media, along with testing multiple target materials.

9. Conclusion & Future Work

In this work, we empirically investigated the threat to speech privacy from vibrations induced by external speech when measured by a laser vibrometer. Our work demonstrated that in some unique cases such speech attacks may be successful. We noted however that under certain conditions that we explored, an eavesdropping attack using a commercially-available laser may not be adequate enough to compromise speech privacy *to the full extent of speech recognition*. We found that live human speech at normal conversational loudness (40-60 dB) was only able to induce vibrations on the nearby object in some limited scenarios. Overall, we believe this attack is much more challenging than has previously been suggested. We acknowledge that eavesdropping via laser can be effective in the context of national intelligence (i.e., specialized security operations with very high-intensity lasers only available to the military [8]). And it is possible such technology will eventually become commercially available. However, our analysis concludes that this attack may not be as viable in the commercial space due to limiting factors in the environment.

Although our work provides a comprehensive analysis of the attack's potential, there still remains many unexplored directions for studying laser-eavesdropping. Specifically, future work may investigate the leakage of other speech information including speaker recognition, emotion detection, and language detection. Additionally,we could explore new attack applications of laser vibrometers such as detecting the number of people in a room by measuring the exterior walls or monitoring when a person enters or leaves a space by measuring the entrance.

Acknowledgment

We would like to give special recognition to the company Polytec for lending us the laser vibrometer equipment that we used in this study. We would also like to thank the reviewers and our shepherd for guiding the final revisions of this paper.

References

[1] C. Arthur. (2013) Laser spying: is it really practical? [Online]. Available: https://www.theguardian.com/world/2013/aug/22/gchq-warned-laser-spying-guardian-offices

[2] Washington Post. (2013) British intelligence worried about lasers spying on the Guardian. That's not so crazy. [Online]. Available: https://wapo.st/2lUJKEZ

[3] Life Hacker. (2012) Build a Laser Microphone to Eavesdrop on Conversations Across the Street. [Online]. Available: https://lifehacker.com/build-a-laser-microphone-to-eavesdrop-on-conversations-5961503

[4] Stack Exchange. (2017) How practical is a laser microphone and how to protect against it? [Online]. Available: https://security.stackexchange.com/questions/151972/how-practical-is-a-laser-microphone-and-how-to-protect-against-it

[5] Youtube. (2011) Laser microphone for audio surveillance via window panes. [Online]. Available: https://www.youtube.com/watch?v=1MrudVza6mo

[6] ——. (2013) Fast Hacks #6 – Laser Spy Microphone. [Online]. Available: https://www.youtube.com/watch?v=K-96dX8ltO8

[7] Amrita Learning. (2019) Production and Propagation of Sound. [Online]. Available: https://tinyurl.com/ProductionAndPropOfSound

[8] Argo-A Security. (2019) Long-Range Laser Listening Device. [Online]. Available: https://tinyurl.com/ArgoASec

[9] Polytec. (2019) Laser Doppler Vibrometry. [Online]. Available: https://www.polytec.com/us/vibrometry/technology/laser-doppler-vibrometry/

[10] ——. (2019) Single-point Vibrometers. [Online]. Available: https://www.polytec.com/us/vibrometry/products/single-point-vibrometers/

[11] ——. (2019) Full-field Vibrometers. [Online]. Available: https://www.polytec.com/us/vibrometry/products/full-field-vibrometers/

[12] ——. (2019) Innovative Solutions for MEMS Characterization. [Online]. Available: https://www.polytec.com/us/vibrometry/products/microscope-based-vibrometers/

[13] ——. (2019) Special-application Vibrometers. [Online]. Available: https://www.polytec.com/us/vibrometry/products/special-application-vibrometers/

[14] ——. (2019) Vibroflex. [Online]. Available: https://www.polytec.com/eu/vibrometry/products/single-point-vibrometers/vibroflex/

[15] ——. (2019) Optical Vibration Measurement per Laser Vibrometry. [Online]. Available: https://www.polytec.com/us/vibrometry/areas-of-application/

[16] K. Wu, D. Zhang, G. Lu, and Z. Guo, "Influence of sampling rate on voice analysis for assessment of parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1416–1423, 2018.

[17] C. Ssnderson and K. Paliwal, "Effect of different sampling rates and feature vector sizes on speech recognition performance," in *TENCON '97 Brisbane - Australia. Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (Cat. No.97CH36162)*, vol. 1, 1997, pp. 161–164 vol.1.

[18] Sciencing. (2018) Which Materials Carry Sound Waves Best? [Online]. Available: https://sciencing.com/materials-carry-sound-waves-8342053.html

[19] Environmental Protection Department. (2017) Characteristics of Sound and the Decibel Scale. [Online]. Available: https://tinyurl.com/EPD-CharOfSound

[20] E. Sengpiel. (2011) Decibel Table–Loudness Comparison Chart. [Online]. Available: http://www.siue.edu/~gengel/ece476WebStuff/SPL.pdf

[21] Walmart. (2019) Red Solo Cup Cold Plastic Party Cups 16 Ounce Pack of 100. [Online]. Available: https://www.walmart.com/ip/Red-Solo-Cup-Cold-Plastic-Party-Cups-16-Ounce-Pack-of-100/264119372

[22] ——. (2019) Hefty Hot Disposable Cups with Lids, 16 Ounce, 20 Count. [Online]. Available: https://www.walmart.com/ip/Hefty-Hot-Disposable-Cups-with-Lids-16-Ounce-20-Count/521412668

[23] ——. (2019) Dart Drink Foam Cups, 16oz, White. [Online]. Available: https://www.walmart.com/ip/Dart-Drink-Foam-Cups-16oz-White-25-Bag-40-Bags-Carton/938096676

[24] IEEE Subcommittee on Subjective Measurements, ""harvard sentences", ieee recommended practice for speech quality measurements," *IEEE No 297-1969*, vol. 17, pp. 227–246, 1969. [Online]. Available: https://www.cs.columbia.edu/~hgs/audio/harvard.html

[25] IBM. (2019) Correlation Settings. [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=tab-correlation-settings

[26] M. Brookes. (2017) VOICEBOX: Speech Processing Toolbox for MATLAB. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752.

[29] C. Forrest. (2016) Microsoft's AI can now understand speech better than humans. [Online]. Available: https://www.techrepublic.com/article/microsofts-ai-can-now-understand-speech-better-than-humans/

[30] E. Myers. (2019) The Role of Artificial Intelligence and Machine Learning in Speech Recognition. [Online]. Available: https://www.rev.com/blog/artificial-intelligence-machine-learning-speech-recognition

[31] A. Nabeth-Halber. (2018) Are machines better than humans in speech transcription? Unravel the myth and reality. [Online]. Available: https://www.linkedin.com/pulse/machines-better-than-humans-speech-transcription-myth-nabeth-halber/

[32] Google. (2019) Forms. [Online]. Available: https://www.google.com/forms/about/

[33] W. Li, M. Liu, Z. Zhu, and T. Huang, "Ldv remote voice acquisition and enhancement," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, 2006, pp. 262–265.

[34] R. Peng, B. Xu, G. Li, C. Zheng, and X. Li, "Long-range speech acquirement and enhancement with dual-point laser doppler vibrometers," in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, 2018, pp. 1–5.

[35] J. Shang, Y. He, D. Liu, H. Zang, and W. Chen, "Laser doppler vibrometer for real-time speech-signal acquirement," *Chin. Opt. Lett.*, vol. 7, no. 8, pp. 732–733, Aug 2009.

[36] Google. (2019) Cloud Speech-to-Text. [Online]. Available: https://cloud.google.com/speech-to-text/

[37] Microsoft Azure. (2019) Speech Services. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/speech-services/

[38] IBM Watson. (2019) Speech to Text. [Online]. Available: https://www.ibm.com/watson/services/speech-to-text/

[39] A. Li. (2017) Google's speech recognition is now almost as accurate as humans. [Online]. Available: https://9to5google.com/2017/06/01/google-speech-recognition-humans/

[40] Study. (2019) How Wave Propagation Relates to the Properties of Materials. [Online]. Available: https://study.com/academy/lesson/how-wave-propagation-relates-to-the-properties-of-materials.html

[41] Aqua-Calc. (2019) Styrofoam. [Online]. Available: https://www.aqua-calc.com/page/density-table/substance/styrofoam

[42] ——. (2019) Polystyrene. [Online]. Available: https://www.aqua-calc.com/page/density-table/substance/polystyrene

[43] ——. (2019) Paper, standard. [Online]. Available: https://www.aqua-calc.com/page/density-table/substance/paper-coma-and-blank-standard

[44] Britannica. (2019) Elasticity. [Online]. Available: https://www.britannica.com/science/elasticity-physics

[45] . (2019) Young's Modulus. [Online]. Available: http://www-materials.eng.cam.ac.uk/mpsite/interactive_charts/stiffness-density/basic.html

[46] Wikipedia. (2019) Tension (physics). [Online]. Available: https://en.wikipedia.org/wiki/Tension_(physics)

[47] P. Marquardt, A. Verma, H. Carter, and P. Traynor, "(Sp)IPhone: Decoding Vibrations from Nearby Keyboards Using Mobile Phone Accelerometers," in *Proceedings of the 18th ACM Conference on Computer and Communications Security*, ser. CCS '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 551–562.

[48] L. Cai and H. Chen, "TouchLogger: Inferring Keystrokes on Touch Screen from Smartphone Motion," in *6th USENIX Workshop on Hot Topics in Security (HotSec 11)*. San Francisco, CA: USENIX Association, 2011.

[49] Z. Xu, K. Bai, and S. Zhu, "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*, ser. WISEC '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 113–124. [Online]. Available: https://doi.org/10.1145/2185448.2185465

[50] E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang, "ACCessory: Password Inference Using Accelerometers on Smartphones," in *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, ser. HotMobile '12. New York, NY, USA: Association for Computing Machinery, 2012.

[51] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing Speech from Gyroscope Signals," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 1053–1067.

[52] S. A. Anand and N. Saxena, "Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 1000–1017.

[53] J. Han, A. J. Chung, and P. Tague, "PitchIn: Eavesdropping via Intelligible Speech Reconstruction Using Non-acoustic Sensor Fusion," in *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2017, pp. 181–192.

[54] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video," *ACM Trans. Graph.*, vol. 33, no. 4, 2014.

[55] Z. Zalevsky, Y. Beiderman, I. Margalit, S. Gingold, M. Teicher, V. Mico Serrano, and J. Garcia-Monreal, "Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern," *Optics express*, vol. 17, pp. 21 566–80, 11 2009.

[56] Y. Avargel and I. Cohen, "Speech measurements using a laser doppler vibrometer sensor: Application to speech enhancement," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2011, pp. 109–114.

[57] R. M. Sullenberger, S. Kaushik, and C. M. Wynn, "Photoacoustic communications: delivering audible signals via absorption of light by atmospheric H2O," *Opt. Lett.*, vol. 44, no. 3, pp. 622–625, Feb 2019.

[58] B. Nassi, Y. Pirutin, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Real-Time Passive Sound Recovery from Light Bulb Vibrations," Cryptology ePrint Archive, Report 2020/708, 2020, https://ia.cr/2020/708.

[59] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, *Spying with Your Robot Vacuum Cleaner: Eavesdropping via Lidar Sensors*. New York, NY, USA: Association for Computing Machinery, 2020, p. 354–367.

[60] V. Amendolare and W. Sarraf. (2005) Laser Audio Surveillance Device. [Online]. Available: https://web.wpi.edu/Pubs/E-project/Available/E-project-010906-145352/unrestricted/ReportFinal.pdf

[61] T. Lv, X. Han, S. song Wu, and Y. Li, "The effect of speckles noise on the laser doppler vibrometry for remote speech detection," *Optics Communications*, 2019.

[62] Y. Wang, W. Zhang, X. Kong, Y. Wang, and H. Zhang, "Two-Sided LPC-Based Speckle Noise Removal for Laser Speech Detection Systems," *IEICE Transactions on Information and Systems*, vol. E104.D, pp. 850–862, 06 2021.

[63] Y. Wang, W. Zhang, Z. Wu, X. Kong, and H. Zhang, "Speckle noise detection and removal for laser speech measurement systems," *Applied Sciences*, vol. 11, no. 21, 2021.

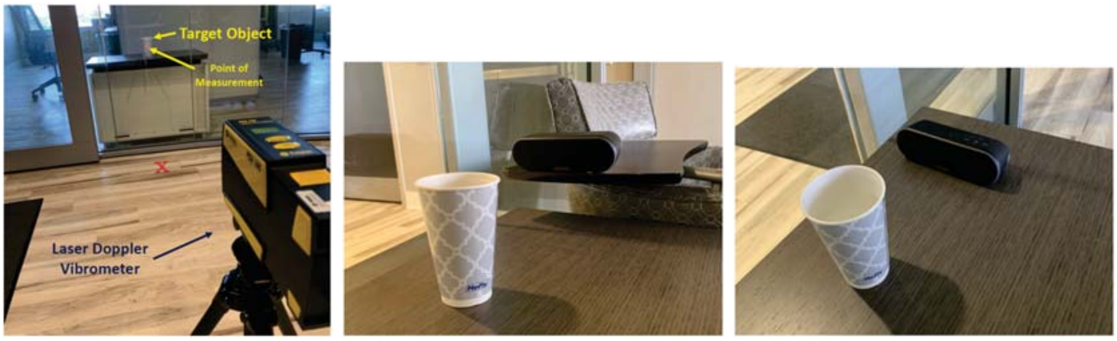
Appendix

1. Speech Metric Results

TABLE 4: Summary of the average speech metric values calculated for recovered samples from Direct-Contact scenarios and raw microphone recordings.

Speech Source – Propagation Medium	SPL	STNR	STOI	PESQ
Live Human – Aerial	Normal	1.21	0.20	1.37
Loudspeaker – Aerial	Normal	3.23	0.29	1.39
Loudspeaker – Same Surface	Loud	9.95	0.39	1.50
Microphone Recording	Normal	16.18	0.91	3.79

2. Images of Experiment Setups

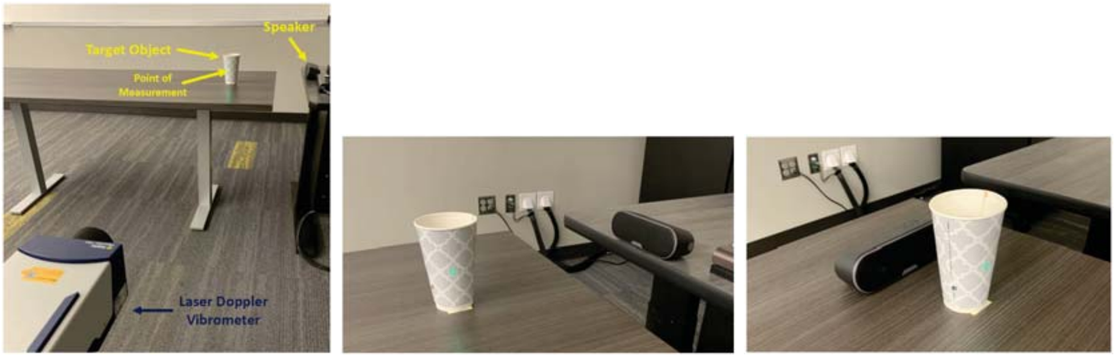


(a) Our experimental setup for the Glass-Barrier measurement scenario. The laser is passing through the transparent glass divider and focusing on the side of the cup. For Direct-Contact scenario, the table/cup moved to the spot marked by the red X.

(b) The cup and speaker arrangement for the aerial transfer medium scenarios. The speaker device rests on the writing board of a chair. The chair and writing board are decoupled from the table that the cup is resting on so the vibrations from the speech can only propagate through the air.

(c) The cup and speaker arrangement for the same surface transfer medium scenarios. The speaker device shares a surface with the cup. Any vibrations created from the speaker source can now propagate through the shared surface to induce vibrations in the cup.

Figure 5: Images of experiment setups for data collected with the PDV-100 vibrometer.



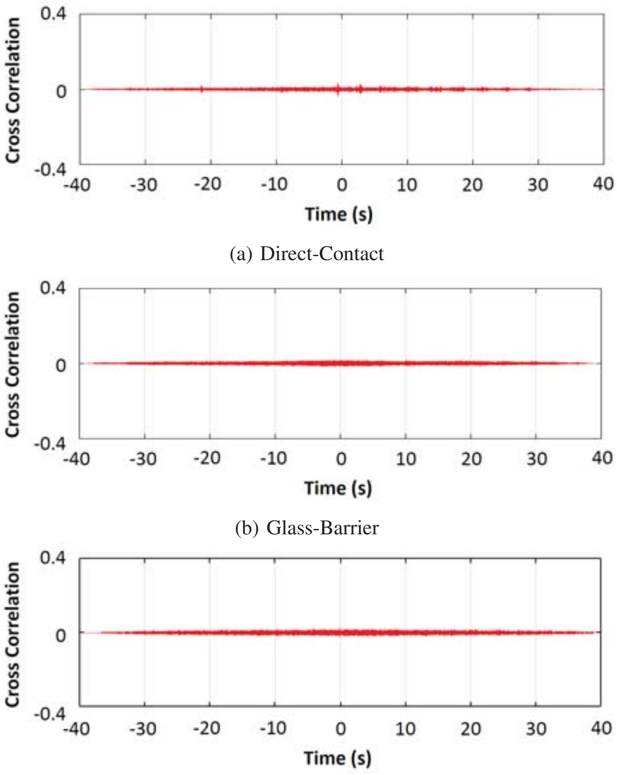
(a) Our experimental setup for the Direct-Contact (loudspeaker-aerial) measurement scenario. The laser is unobstructed and focused on the side of the cup.

(b) The cup and speaker arrangement for the aerial transfer medium scenarios. The speaker device rests on a separate table than the cup.

(c) The cup and speaker arrangement for the same surface transfer medium scenarios. The speaker device shares a surface with the cup.

Figure 6: Images of experiment setups for data collected with the Vibroflex vibrometer.

3. Cross Correlation Graphs

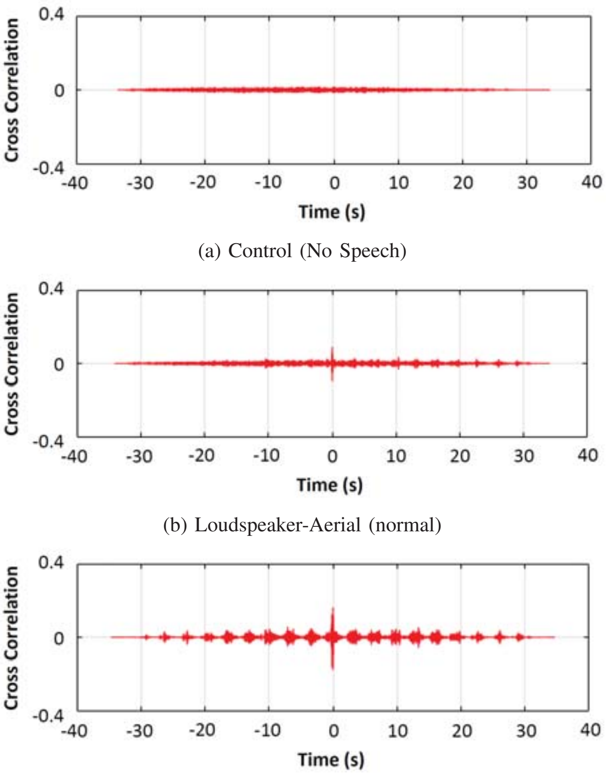


(a) Direct-Contact

(b) Glass-Barrier

(c) Glass-Surface

Figure 7: Cross correlation graphs comparing the original audio to the data collected in Live Human speech scenarios, in each attack setup, with the PDV-100 vibrometer.

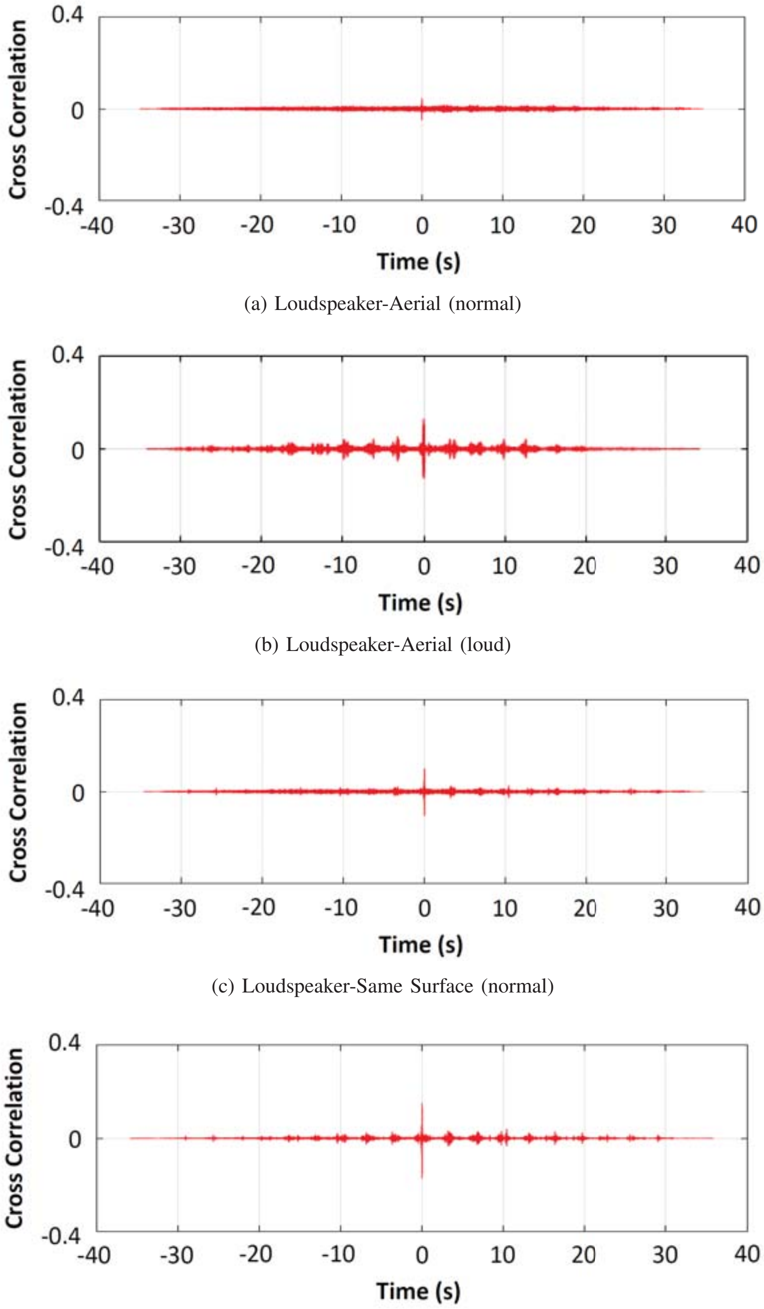


(a) Control (No Speech)

(b) Loudspeaker-Aerial (normal)

(c) Loudspeaker-Aerial (loud)

Figure 8: Cross correlation graphs comparing the original audio to the data collected in the *Glass-Surface* measurement scenario using the Vibroflex laser vibrometer.



(a) Loudspeaker-Aerial (normal)

(b) Loudspeaker-Aerial (loud)

(c) Loudspeaker-Same Surface (normal)

(d) Loudspeaker-Same Surface (loud)

Figure 9: Cross correlation graphs comparing the original audio to the data collected in the *Glass-Barrie*r measurement scenario, from the plastic cup, using the Vibroflex laser vibrometer.