EchoLight: Sound Eavesdropping based on Ambient Light Reflection

Guoming Zhang, Zhijie Xiang, Heqiang Fu, Yanni Yang, and Pengfei Hu*
School of Computer Science and Technology, Shandong University, China

***Abstract*—Sound eavesdropping using light has been an area of considerable interest and concern, as it can be achieved over long distances. However, previous work has often lacked stealth (e.g., active emission of laser beams) or been limited in the range of realistic applications (e.g., using direct light from a device's indicator LED or a hanging light bulb). In this paper, we present EchoLight, a non-intrusive, passive and long-range sound eavesdropping method that utilizes the extensive reflection of ambient light from vibrating objects to reconstruct sound. We analyze the relationship between reflection light signals and sound signals, particularly in situations where the frequency response of reflective objects and the efficiency of diffuse reflection are suboptimal. Based on this analysis, we have introduced an algorithm based on cGAN to address the issues of nonlinear distortion and spectral absence in the frequency domain of sound. We extensively evaluate EchoLight's performance in a variety of real-world scenarios. It demonstrates the ability to accurately reconstruct audio from a variety of source distances, attack distances, sound levels, light intensity, light sources, and reflective materials. Our results reveal that the reconstructed audio exhibits a high degree of similarity to the original audio over 40 meters of attack distance.**

***Index Terms*—Optical eavesdropping, Ambient light, Generative adversarial network**

I. Introduction

Sound eavesdropping can secretly steal confidential information without the victim's consent during phone calls, online meetings, or even face-to-face conversations, etc., thus posing a great threat to privacy. In recent years, various new eavesdropping technologies have emerged, allowing attackers to easily intercept sounds near the victim using millimeter-wave radar [1]–[7], WiFi signal [8], [9], motion sensors [10]–[13], electromagnetic radiation [14], and more. However, these methods suffer from one or more of the following limitations: 1) require intrusion into the victim's device; 2) rely on active sensors; 3) have limited attack range.

Optical eavesdropping attacks [15]–[19] have emerged as an intriguing avenue for extracting information. Glowworm [15], which recovers sound from a loudspeaker's power indicator LED. However, it may not be applicable to most existing loudspeakers lacking an indicator LED, and it cannot recover sound from a live human. Lamphone [16] uses the acoustic response of a suspended light bulb on the roof to recover sound which requires that the loudspeaker be placed at a close distance (1 centimeter), which severely limits its applicability. Laser-based eavesdroppings [20]–[23] are active methods that require the eavesdropper to direct a laser beam directed at a vibrating object and recover sound from the reflected laser beam. However, it can be easily detected by the victim due to the laser spot.

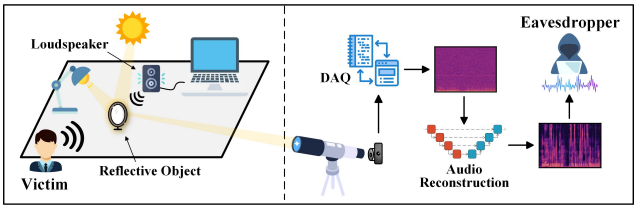


Fig. 1. EchoLight eavesdropping attack. The light source can be either natural sunlight or artificial indoor lighting, while the sound to be recovered can be human speech or played through a loudspeaker.

In this paper, we ask a fundamental question: Is it possible to passively recover sound over long distances using ambient light? To answer this question, we designed EchoLight, an approach to recover sound (including both loudspeakers and live human voices) by capturing ambient light (such as sunlight or indoor lighting) reflected or even diffusely reflected from nearby reflective objects. As shown in Fig. 1, these reflective objects, situated close to the victim, experience slight vibrations induced by sound waves. Consequently, the reflection of ambient light is deflected when the object's surface is displaced due to the vibrations caused by the audio signal. We employ a photodetector to capture the deflection of ambient light resulting from the acoustic wave vibrations. The optical measurements obtained by the photodetector carry the vibration information of the acoustic wave within the room. By converting these optical measurements into an audio signal, we can directly reconstruct the sound.

To realize EchoLight, however, several challenges need to be overcome. First, the limited acoustic energy hitting the object, coupled with the object's passive vibration, leads to poor frequency response in the reflective material. Second, to avoid detection and to be applicable to more scenarios, we utilize ambient light that is reflected or even diffusely reflected to recover the sound. Consequently, the reflection efficiency and intensity of the light we ultimately capture is relatively poor. These factors result in a lower signal-to-noise ratio in the sound signals, and the sound also lacks high-frequency components. To address this challenge, we first use signal processing and denoising techniques to enhance the

*Pengfei Hu is the corresponding author.

quality of the signal and then employ a bandwidth extension method based on a conditional Generative Adversarial Network (cGAN) architecture to reconstruct the sound.

We implement and evaluate `EchoLight` in various scenarios and settings. The evaluation results demonstrate that the audio reconstructed by `EchoLight` exhibits a significant resemblance to the original audio. Our contributions can be summarized as follows:

- We introduce `EchoLight` a non-invasive, passive eavesdropping method that utilizes ambient light to directly reconstructed audio from the subtle vibrations of everyday reflective objects.
- We employ a cGAN model to recover the high-frequency component of the captured light signals. This generative model is capable of reconstructing high-quality audio from preprocessed light signals.
- We validate the effectiveness of `EchoLight` by conducting a comprehensive exploration of various settings, including sound source distance, attack distance, reflective material, sound level, and moving audio source.

The remaining paper is organized as follows. Section II discusses the related work. In Section III, we present our attack scenario. Section IV covers the preliminaries, encompassing feasibility analysis and an introduction to the cGAN. In Section V, we outline our audio reconstruction method, which is employed to reconstruct audio from optical signals. In Section VI, we analyze and explore the results under various experimental settings. In Section VII, we engage in discussions, and finally, we conclude in Section VIII.

II. Related work

In this section, we extensively review state-of-art eavesdropping techniques. Some methods exploit smartphone motion sensors (gyroscope and accelerometer) to gather private user information [10], [12], [24], [25]. In [13], multiple non-acoustic sensors are fused for intelligible signal reconstruction. [26] explores eavesdropping using vibration sensors in glasses' nose pads. [27] employs a vibra-motor, vibrating due to the movement of magnetic substances, as an acoustic sensor. [28] shows how earphones can become eavesdropping microphones through malware, and [29] treats mechanical hard drives as synthetic microphones. Additionally, [22], [23] transforms a robotic vacuum cleaner's lidar radar into a laser microphone, and RFID [30] is also utilized for eavesdropping.

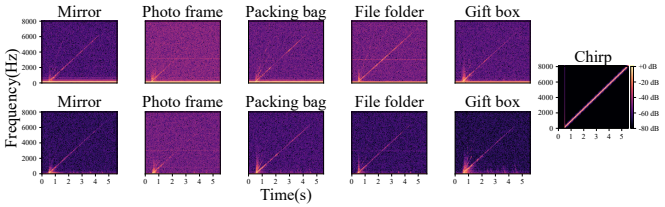
The aforementioned eavesdropping attacks are invasive, requiring physical tampering or malware installation. On the other hand, some methods [10], [12], [13], [24], [25] rely on hot word vocabularies for tasks like word classification or speaker identification. Unlike these, our non-intrusive approach doesn't depend on pre-established vocabulary, enhancing practicality in real-world scenarios.

Ongoing development in non-intrusive eavesdropping attacks includes techniques like [8], [9] using WiFi Received Signal Strength (RSS) and Channel State Information (CSI) analysis for word identification. Another method [31], based on Impulse Radio Ultra-Wideband, separates sound from multiple sources. However, these methods face challenges like limited vibration resolution due to longer wavelengths and the need for large antenna setups, thereby increasing operational difficulty. Moreover, mmWave-based eavesdropping attacks have been explored. For instance, mmEve [4] captures mmWave signals from smartphones in earphone mode to recover speech. Wavesdropper [32] employs millimeter waves for through-wall eavesdropping on human speech. Another method in [5] employs piezoelectric film and millimeter wave signals for voice eavesdropping. mmSpy [6] also utilizes mmWave for eavesdropping on phone conversations but treats it as a classification problem, limiting its ability to recognize specific keywords or numbers. Furthermore, [7] captures quality sound from the user's throat with millimeter-wave radar, but the subject must remain still and close to the radar probe.

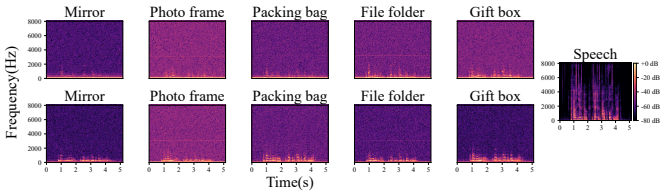
Optical eavesdropping is relevant to this article. The visual microphone [33] records subtle surface vibrations with a high-speed camera (2200 FPS). However, it requires substantial computing resources and high hardware specifications, leading to lengthy processing times for recovering brief audio snippets. Laser microphones [20]–[23], [34] use lidar sensors to illuminate vibrating objects and collect reflected signals for audio reconstruction. However, these methods have several drawbacks: 1) They are active, relying on active sensors that need signal transmission and reception. 2) They lack sufficient concealment, as specialized optical sensors can detect directed laser beams since the eavesdropper needs to inject the laser beam into the victim's room. 3) They are position-sensitive, as laser transceivers require precise calibration during signal reception, and even slight positional deviations can affect the signal quality. 4) Certain methods [22], [23] are considered intrusive attacks, requiring malware installation in the victim's device to extract laser data. 5) They have equipment restrictions, as specialized equipment with limited availability (such as specialized laser transceivers) may be required, and these may only be accessible to specific national departments.

Lamphone [16], on the other hand, captures indoor light bulb vibrations using a telescope and photoelectric sensor at a distance. However, this method is limited by the loudspeaker's proximity to the light bulb, requiring a distance of 1 centimeter, which may not be practical in daily offices or meeting rooms. Additionally, [16], [33] require very high sound levels (over 100 dB). Glowworm [15] collects light signals from a power supply device's LED indicator using a telescope and photoelectric sensor for audio reconstruction. However, this method depends on hardware vulnerabilities (power consumption fluctuations affecting the LED), which manufacturers can counteract easily.

There is also a category of methods that exploit optical side effects to retrieve information from optical emissions of victim devices. For example, some methods aim at recovering content from monitors [17], [35]–[37], or retrieving keystrokes from physical and virtual keyboards [18], [19], [38], [39]. However, it is worth noting that these attacks do not specifically focus on speech eavesdropping and audio reconstruction.



(a) Frequency response of chirp signal using indoor lighting (top) and sunlight (bottom)



(b) Frequency response of speech audio using indoor lighting (top) and sunlight (bottom)

Fig. 2. Feasibility analysis results.

III. Threat Model

The scenario we consider is that the victims are in an office or conference room discussing confidential business matters, and the meeting may or may not use loudspeakers. Sunlight or indoor light sources illuminate everyday objects (reflective objects) in the room, with the light being reflected out through the window.

The eavesdropper captures reflected ambient light (sunlight or indoor lighting) with a telescope, received by a photodetector. The photodetector is connected to a Data Acquisition Card (DAQ) to convert the analog signal to a digital format. After the digital signal is processed, the high-frequency information is recovered using a bandwidth extension algorithm. Finally, the eavesdropper receives a reconstructed, high-quality sound.

In this attack scenario, we assume the eavesdropper has the following conditions and capabilities for launching an `EchoLight` attack:

- The attacker can't access or damage the victim's device, nor install any sensors in the victim's room.
- The victim's room has common items, like packing bags, vanity mirrors, file folders, etc., where sound waves induce minute vibrations in reflective objects.
- There must be an unobstructed line of sight between the telescope and the reflective object in the victim's room, with the reflective object receiving light.
- The attacker doesn't need knowledge of specific words or prior information spoken by the victim; they should be able to reconstruct audio from any complete sentence.
- The attacker doesn't need expensive equipment; passive sensors, telescopes, and signal processing tools, not categorized as spy equipment, are sufficient.

IV. Preliminaries

In this section, we explore the feasibility of our eavesdropping attack and introduce the cGAN for signal bandwidth extension.

A. Feasibility Analysis

For a successful eavesdropping attack, we explored if the surface vibrations from the acoustic wave are sufficient to recover sound and if the ambient light (such as sunlight or indoor desk lamp light) reflected from the surface carries enough information about the acoustic vibration. We selected reflective objects like mirrors, photo frames, packing bags, file folders, and gift boxes. We placed the loudspeakers close to reflective objects on separate tables to eliminate any potential effects of table vibrations.

During the experiment, we set the sound level of the loudspeaker to 90 dB and played a chirp signal with a frequency range from 100 Hz to 8 kHz, as well as speech audio. To capture the reflected ambient light, we placed a telescope at a distance of 5 meters from the reflective object. A photodetector was placed behind the eyepiece of the telescope to detect the changing light signal received by the telescope. Subsequently, we used a DAQ to convert the optical analog signal detected by the photodetector into a digital signal for further analysis.

Fig. 2 illustrates the results of the feasibility analysis using both sunlight and indoor desk lamp light sources. It is evident that each material exhibits a high maximum response frequency (around 4-7 kHz) to the chirp signal. The signal strength decreases with increasing frequency, resulting in no response at certain high frequencies. Conversely, each material's maximum response frequency to speech audio is relatively low (about 1000 Hz) due to the weak high-frequency component, unlike the chirp signal with high amplitude across all frequency bands. Additionally, the response of reflective objects to high frequencies is weaker than that to low frequencies, making it challenging for them to generate any substantial response at high frequencies. We found that audio recovered under sunlight has a higher response frequency and less noise compared to desk lamp light.

The results of our feasibility analysis show that there is a certain correlation between light signals and sound signals, both in the case of specular reflection (e.g., mirror) and diffuse reflection (other materials). The vibration of the object's surface is capable of generating a signal that can be used for audio recovery, with a frequency response of around 1000 Hz. In addition, ambient light carries the information of the vibrations caused by the sound when reflected from the reflective object. Therefore, it is possible to use ambient light to recover sound from the vibrations of reflective objects.

B. Conditional Generative Adversarial Network

The conditional Generative Adversarial Network (cGAN) [40] consists of a generator and a discriminator, and its architecture is shown in Fig. 3. The generator produces images that satisfy given conditions by utilizing input conditions and random noise sequences. These images, combined with the

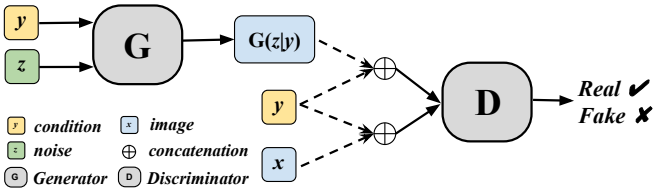


Fig. 3. Architecture of conditional Generative Adversarial Network.

specified conditions, are then assessed by the discriminator to be either true or false. The discriminator and the generator undergo alternating training throughout the entire training process, gradually enhancing the discriminator's ability to discern authenticity and the generator's capacity to forge images. This process resembles a two-player minimax game, ultimately reaching a Nash equilibrium state in which the generator and the discriminator no longer strive to enhance their respective capabilities. Through adversarial training, the generated images achieve a high level of realism, leading the discriminator to classify images produced by the generator as real with a probability of 0.5.

The intersection of environmental audio and speech reconstruction is an emerging field, introducing new challenges in reconstructing speech from environmental audio, with cGAN models serving as innovative solutions to address this issue.

V. EchoLight Design

EchoLight is composed of two modules: Speech Enhancement and Bandwidth Extension. The Speech Enhancement module eliminates the intense noise present in the captured optical signal, achieving a flawless noise reduction. On the other hand, the Bandwidth Extension module dramatically enhances the overall audio quality by recovering the lacking high-frequency components in the optical signal. In this section, we describe the specific design details of each component in EchoLight.

A. Speech Enhancement

A direct transformation of the optical signal into audio introduces considerable noise, making it unrecognizable. Thus, we first apply speech enhancement techniques to the optical signal. Specifically, our speech enhancement process consists of the following steps:

- **Filtering**. Since the light bulb operates at 100 Hz, the light signal captured by the photodetector contains 100 Hz and its harmonics (200 Hz, 300 Hz, etc.). We use a high-pass filter with a cutoff frequency of 100 Hz to remove components below 100 Hz, and band-stop filters to remove the light frequency and its harmonics.
- **Denoising**. The captured optical signal exhibits pronounced noise, similar to white noise, which is not a result of the original audio. To deal with this, we leverage Sox - Sound eXchange [41], a widely used command-line tool for audio processing, to effectively diminish this type of interference. Specifically, we select a short segment of silent audio, extract its noise features, and then denoise the entire audio segment based on the noise features.
- **Peak Normalization**. To adjust the dynamic range of the denoised signal, we normalize the audio peak to 0 dB and adjust other parts of the audio accordingly.

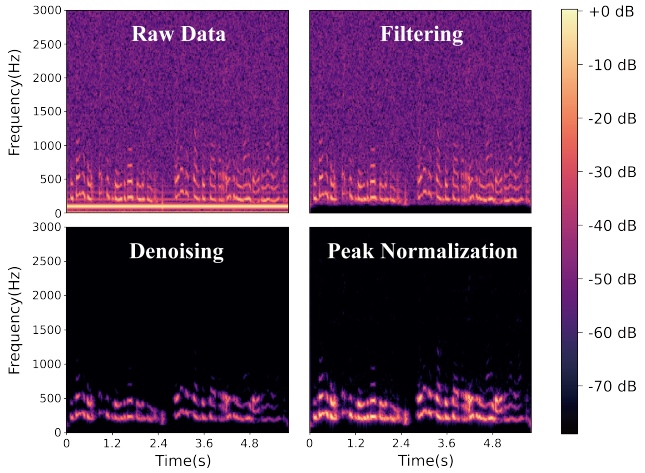


Fig. 4. The influence of the three stages of speech enhancement.

The implementation results of each stage are illustrated in Fig. 4. The speech enhancement technology used in our study is lightweight and straightforward in nature. This approach allows us to obtain a relatively clean optical signal, which can then be fed into the trained model for rapid reconstruction of high-quality audio. This ensures a certain level of real-time capability in the eavesdropping process, which is a critical requirement for eavesdroppers.

B. Bandwidth Extension

Due to the poor frequency response of the reflective object and the reflection efficiency, the collected optical signal always lacks these high-frequency components. The goal of bandwidth extension is to recreate the high-frequency components based on the low-frequency components of the audio signal. This is achieved by learning the mapping relationship between low-frequency and high-frequency components in the sound. In our method, we design the bandwidth extension method based on cGAN.

1) Model input and output: The application field of cGAN primarily focuses on image translation and related tasks, where both the input and output are in the form of images. In our case, we apply the Short-Time Fourier Transform (STFT) to convert the input audio into a spectrogram, which is then used as the input and output form for the model. The audio has a sampling frequency of 16 kHz, and all audios are divided into segments with a length of 2.0240625 s (32385 sampling points). Each segment is transformed into a 256×256 spectrogram using STFT with specific parameters (the number of FFT points and window length are set to 510, and the hop length is set to 127). Subsequently, the generator's output is also a 256×256 spectrogram. We utilize the Griffin-Lim (GL) algorithm and inverse Short-Time Fourier Transform (iSTFT)

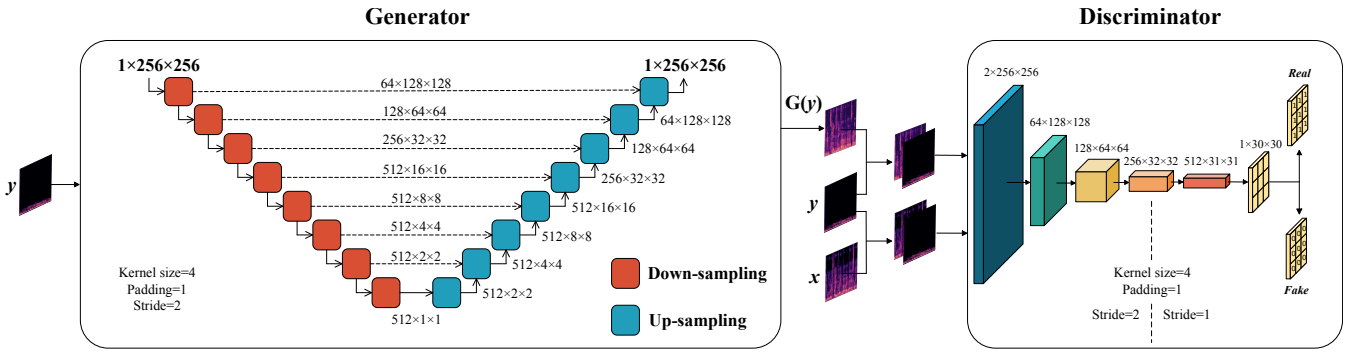


Fig. 5. Model Architecture for bandwidth extension.

to reconstruct the sound signal. It's worth noting that the time cost for the integration process of audio reconstruction is relatively minimal and can be practically negligible.

2) Model architecture: Fig. 5 illustrates our model architecture, where the spectrograms of the original audio x and the speech-enhanced light signal y are used as inputs to the model. The generator adopts a Unet architecture, consisting of 8 Down-sampling and 8 Up-sampling processes. The Down-sampling introduces feature information into the corresponding Up-sampling process through skip connections. The output of the final Up-sampling layer has the same dimension as the input spectrogram. On the other hand, the discriminator consists of 5 layers of 2D convolutional layers. The last layer of the general discriminator is designed to output a scalar value representing probability. However, in our discriminator, the last layer employs only 1 convolution kernel and outputs 30×30 patches. Subsequently, each patch is discriminated as either real or fake.

3) Objective function: In our model, the generator's objective is to minimize the L1 distance between the generated spectrogram and the spectrogram of the original audio, while simultaneously making the discriminator output "True." On the other hand, the discriminator aims to accurately distinguish between real and fake inputs. As a result, our objective function can be formulated as follows:

$$G_{loss} = -\mathbb{E}_y[\log D(G(y))] + \mathbb{E}_{x,y}[\lambda \parallel G(y) - x \parallel_1] \quad (1)$$

$$D_{loss} = -\mathbb{E}_x[\log D(x)] - \mathbb{E}_y[\log(1 - D(G(y)))] \quad (2)$$

where G_{loss} and D_{loss} represent the losses of the generator G and the discriminator D, respectively. x denotes the original real spectrogram, while y represents the spectrogram containing only low-frequency components. The parameter $\lambda = 100$ is the weight of the $L1$ loss. Both the discriminator and the generator aim to minimize the objective function.

4) Training parameters: The training process involves 200 epochs. During the initial 100 epochs, the learning rate is set to 0.0002, and then it gradually decays linearly to 0 over the last 100 epochs. We utilize the Adam optimizer with a first-order momentum of 0.5 and a second-order momentum of 0.999 to update the parameters. Throughout the training process, the discriminator and generator are trained alternately.

5) Spectrogram to audio: The spectrogram generated by the generator contains rich amplitude information for high-frequency components but lacks any phase-related information, making it challenging to directly reconstruct playable audio. To address this, we employ the classic phase reconstruction algorithm called Griffin-Lim (GL) to recover the magnitude spectrum back to audio with reasonable phase information. The GL algorithm initiates with a random phase spectrum and utilizes the phase spectrum along with the input magnitude spectrum to synthesize speech using iSTFT. Subsequently, STFT is performed on the synthesized speech to obtain a new magnitude spectrum and phase spectrum. The new magnitude spectrum is then discarded, and the original magnitude spectrum is combined with the new phase spectrum to synthesize speech again using iSTFT. Through successive iterations, the estimated phase spectrum converges to values close to the real phase spectrum, resulting in an audio reconstruction with minimal distortion.

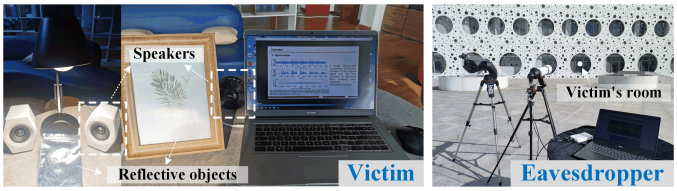
VI. Evaluation

In this section, we first introduce the experimental setup, datasets, and evaluation metrics. Next, we conduct an overall performance evaluation of `EchoLight`. Finally, we provide a comprehensive analysis of the impact of different influencing factors on the attack.

A. Implementation and Experiment Setup

Fig. 6(a) presents our experimental setup. We placed different reflective objects on the victim's desk, and the attacker directed the telescope towards these reflective objects in the room. The photodetector captured the reflected ambient light through the telescope. During this process, when a loudspeaker played a sound or a live human spoke, the attacker could detect the vibrations caused by the sound and subsequently reconstruct the sound signal present in the room.

The reflective materials employed in the experiment consisted of mirror, photo frame, packing bag, file folder, and gift box, as shown in Fig. 6(b). The sampling rate of DAQ is set to 16 kHz. In accordance with the Nyquist sampling theorem, audio signals with a maximum frequency of 8 kHz could be reconstructed. For attacks beyond 15 meters, we used



(a) Attack scenario

(b) Reflective object

Fig. 6. Experimental setup of for EchoLight eavesdropping attack.

the CELESTRON NexStar 127 SLT telescope; for those within 15 meters, the CELESTRON Libra 805 telescope was used.

B. Data Collection

We evaluate the performance of the EchoLight on the VCTK [42] dataset, which contains speech data from 110 English speakers with different accents. Each speaker reads approximately 400 sentences. For our evaluation, we select audio recordings from 6 of these speakers (3 males and 3 females) as our training and testing data. These 6 speakers are labelled User1 to User6, and the details of the dataset are given in Tab. I. Note that there is no overlap between the training and test data for each user. Specifically, the audio from User1 to User4 is used for training and evaluation, while the audio from User5 to User6 is used for cross-user evaluation. In addition, we use a piece of music called "Let It Be" to evaluate EchoLight's performance in reconstructing music.

TABLE I
THE DATASET USED FOR EVALUATING ECHOLIGHT

	Sex	Duration(s)	Training length(s)	Testing words
User1	female	1283	1155	329
User2	female	1182	1064	317
User3	male	1626	1463	342
User4	male	1182	1064	327
User5	male	118	0	304
User6	female	124	0	272
Music	male	243	218	-

C. Evaluation Metrics

We utilize Mel Cepstral Distortion (MCD), Log Spectral Distance (LSD), and Mean Opinion Score (MOS) as evaluation metrics to assess the performance of EchoLight.

MCD and LSD are used to measure the degree of difference between two speech signals. MCD quantifies the distortion between two sets of MFCCs, while LSD calculates the distance between their logarithmic spectrograms. The combined use of MCD and LSD provides a more comprehensive assessment of the quality of synthesized speech. Lower values of MCD and LSD indicate that the synthesized speech is closer to the original speech. Speech recognition systems generally accept two sounds if their MCD values are below 8.

TABLE II
MOS SCORES AND RATINGS

Score	Rating
5	Excellent: Recovered all of the original speech
4	Good: Recovered most of the original speech
3	Fair: Recovered half of the original speech
2	Poor: Recovered little of the original speech
1	Bad: Recovered none of the original speech

MOS is obtained through human subjective evaluation. As shown in Tab. II, scores range from 1 to 5, where 1 indicates very poor reconstruction quality and 5 indicates very good reconstruction quality. We recruited 16 volunteers (8 males and 8 females) to rate our reconstructed audio. These volunteers participated in our experiments without any compensation.

D. Overall Performance

We evaluate the effectiveness of the EchoLight in reconstructing three categories of sound using two different ambient light sources. The three categories of sound include speech played through loudspeakers, music played through loudspeakers and live human voices. For the ambient light sources, we use both natural sunlight and an indoor desk lamp. For these experiments, we choose a packing bag as a reflective object and position the telescope 15 meters away from it. The loudspeakers are placed 5 centimeters from the reflective object and the sound level is set to 90 dB.

1) Recovering speech: We played an 8 s audio from User1 through the loudspeaker in the victim's room, with the audio content being "Please call Stella. Ask her to bring these things with her from the store. Six spoons of fresh snow peas." After denoising the optical signal, we utilized the bandwidth extension model to reconstruct the high-frequency components. The model was pre-trained using User1's audio, and there was no overlap between the training and testing audio (except for common words such as "a," "the," "of," etc.).

Fig. 7 (left) presents the spectrograms of the original audio, the optical signal, and the reconstructed audio using the bandwidth extension model. We observed that the reconstructed audio exhibited a high similarity to the original audio compared to the acquired optical signal. This indicates the effectiveness of our bandwidth extension model in learning low-frequency features to recover high-frequency components. Additionally, during the outdoor experiment, the wind speed was categorized as level two, with wind speeds ranging from approximately 1.6 to 3.3 m/s. Fortunately, these moderate wind conditions did not significantly impact the performance of EchoLight.

2) Recovering music: We used the loudspeakers to play the first 8s segment of the music "Let It Be", while the remaining portion is used for model training. The reconstruction result is shown in Fig. 7 (right). It can be seen that EchoLight has excellent reconstruction performance even for music.

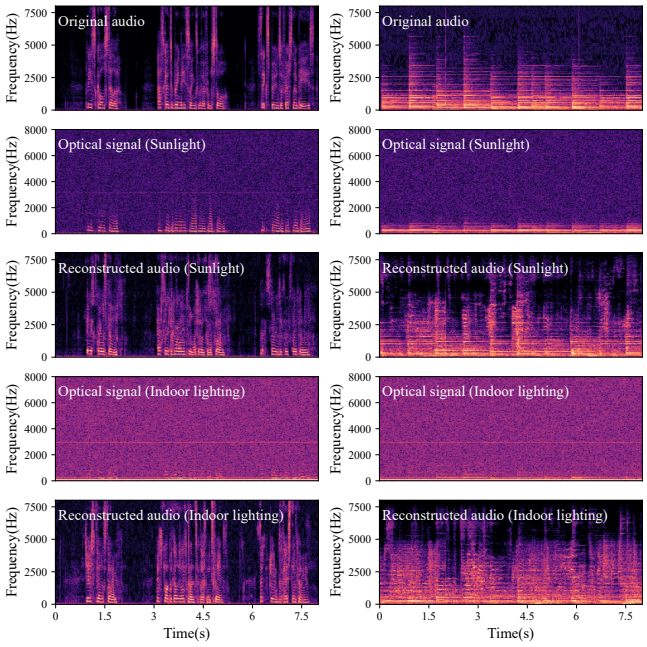


Fig. 7. Recovering speech (left) and music (right) using sunlight and indoor lighting as environment light sources.

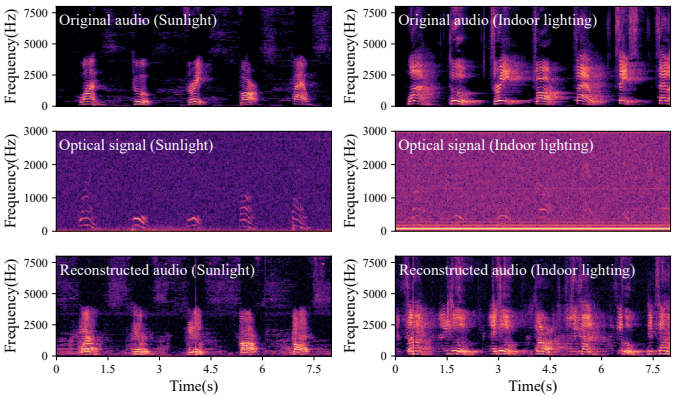
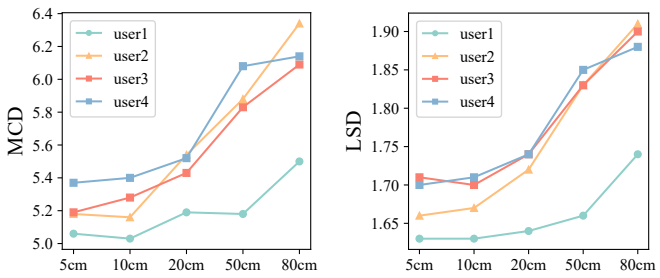


Fig. 8. Recovering live human voices using sunlight (left) and indoor lighting (right) as environment light source.



(a) MCD (b) LSD

Fig. 9. The impact of different sound source distances.

Furthermore, using sunlight as an ambient light source for audio reconstruction yielded better results than using indoor lights as an ambient light source. This improvement can be attributed to sunlight's much higher light intensity compared to indoor light sources. Consequently, eavesdropping attacks using light sources with higher light intensity are capable of recovering higher-quality audio.

3) Recovering live human voices: Given that the victim might not only use the loudspeaker to play the sound of an online meeting but also engage in conversation with other people, we conducted an additional test. The victim was asked to speak while standing at a distance of 5 centimeters from the reflective object. The speech content involved counting from "one". Simultaneously, the voice was recorded using the mobile phone's microphone to obtain the original audio. Finally, the first 8 s segment of the audio is used for testing, while the remaining portion is used for model training. Fig. 8 presents the test results using sunlight and indoor desk lamp light as ambient light sources. The experimental results demonstrate that `EchoLight` exhibits commendable reconstruction capabilities for live human speech.

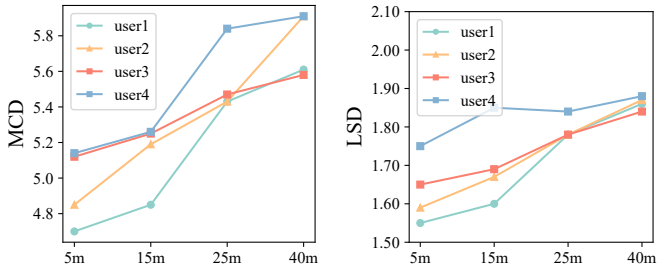
In summary, `EchoLight` demonstrates a remarkable ability to effectively reconstruct sound close to the victim in the room, whether it is speech or music played by loudspeakers or live human voices.

E. Experimental Analysis

Next, we will investigate the impact of different settings on `EchoLight`. For this exploration, we select the indoor desk lamp as the light source and place the loudspeaker at a distance of 5 centimeters from the reflective object, with the sound level set to 80 dB. The telescope is placed 15 meters away from the reflective object. Throughout our experimental analysis, when a particular setting is not specified, we use this basic setting as our default configuration.

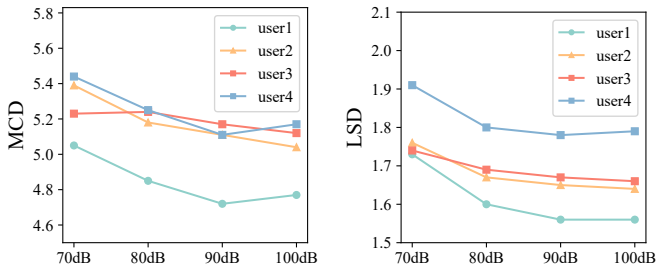
1) Impact of sound source distances: In a real attack scenario, the distance between the loudspeakers and reflective objects may vary. To investigate the impact of sound source distances, we conducted experiments using a packing bag as the reflective object. The loudspeaker was placed at different distances from the reflective object while playing the test audio from User1 to User4. The loudspeaker's sound level (80 dB) and the distance between the telescope and the reflective object (15 meters) were set as the default configuration. The results in Fig. 9 show that `EchoLight` can produce intelligible audio even with the source distance set to 80 centimeters (with an average MCD and LSD of 6.02 and 1.86).

2) Impact of attack distances: A longer attack distance offers potential advantages to an attacker in performing the `EchoLight` eavesdropping attack. To evaluate the impact of various attack distances, we conducted systematic experiments, placing the telescope at different distances from the reflective object. In these experiments, a mirror served as the reflective object. The results depicted in Fig. 10 demonstrate that `EchoLight` can intelligibly reconstruct sound at an attack distance of 40 meters, with average MCD and LSD values measuring 5.75 and 1.86, respectively.



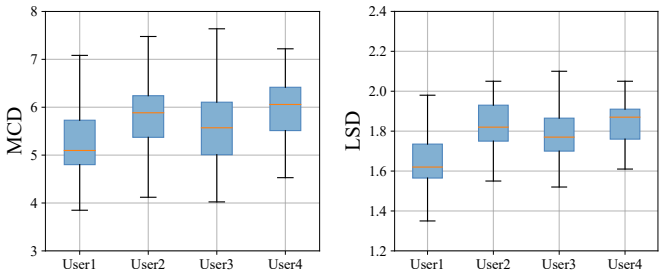
(a) MCD (b) LSD

Fig. 10. The impact of different attack distances.



(a) MCD (b) LSD

Fig. 11. The impact of different sound levels.



(a) MCD (b) LSD

Fig. 12. The impact of moving audio sources.

3) Impact of sound levels: The Speech Intelligibility Index (SII) standard provides guidelines for different levels of vocal effort. As mentioned in [43], the levels of normal, raised, loud speech and shouting are about 62.3 dB SPL (59.2dBA), 68.4 dB SPL (66.4 dBA), 74.8 dB SPL (73.9 dBA) and 82.3 dB SPL (82.2 dBA) respectively. It's important to recognise these levels as they represent real-life scenarios where people might raise their voices in a noisy environment. In order to investigate the impact of sound levels on `EchoLight`, we conducted experiments using a mirror as the reflective object and set the loudspeaker to different sound levels.

The evaluation results are depicted in Fig. 11. Even at a sound level of 70 dB, where the response frequency of the optical signal is only about 600 Hz, the audio reconstructed by bandwidth extension still exhibits a certain degree of intelligibility, demonstrating that `EchoLight` has good performance in covert eavesdropping scenarios.

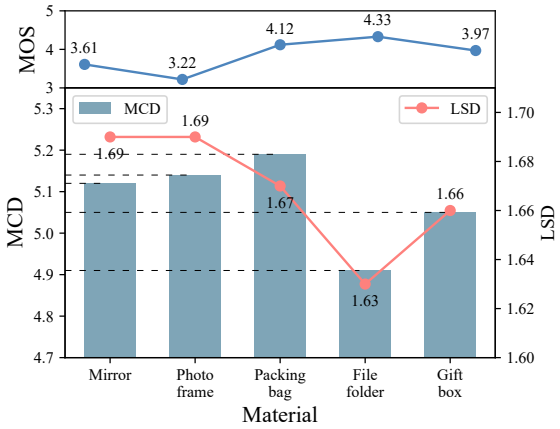


Fig. 13. The performance of audio reconstruction on different reflective materials.

4) Impact of moving audio sources: In a real-world scenario, the victim may move around the room. In order to evaluate the effect of moving audio sources on `EchoLight`, we simulated this by moving the loudspeaker back and forth at a speed of 20 cm/s while maintaining a distance of 10 centimeters to 50 centimeters from a reflective object. For this experiment, we used a packing bag as the reflective object, and the evaluation results are shown in Fig. 12.

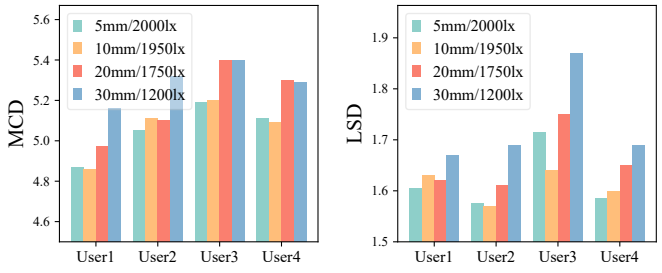
The medians of the MCD and LSD values are within 6 and 2 respectively, although the range of variation is relatively large. This variance can be attributed to the performance of the loudspeaker, which improves as it gets closer to the target and deteriorates as it gets further away. Nevertheless, `EchoLight` demonstrates the ability to reconstruct intelligible audio from a moving audio source within a range of 80 centimeters.

5) Impact of materials: Due to the unique properties of different materials, various reflective objects respond to sound waves to varying degrees. Consequently, we conducted a comprehensive evaluation using different materials. The averaged results for User1-User4 are presented in Fig. 13.

The experimental results indicate that the file folder exhibits the best performance among the tested materials. This superiority can be attributed to its strong reflective effect and its ability to vibrate effectively under the influence of sound waves, resulting in lower MCD and LSD values and higher subjective evaluation scores. In contrast, the photo frame exhibits the highest LSD value and the lowest MOS value. This is primarily due to the weak elastic deformation ability and reflective effect of the photo frame's surface. As a result, the vibration generated by the photo frame is very feeble, and only a small fraction of the light is reflected back. Additionally, the mirror also exhibits a high LSD value, possibly due to its strong reflectivity but limited vibration capability.

In real attack scenarios, selecting an object with both a good reflective effect and vibration ability as the attack medium can significantly enhance the attack's effectiveness.

6) Impact of light intensity: In real attack scenarios, it is common for light to pass through glass media, and the inten-



(a) MCD (b) LSD

Fig. 14. The impact of light intensity.

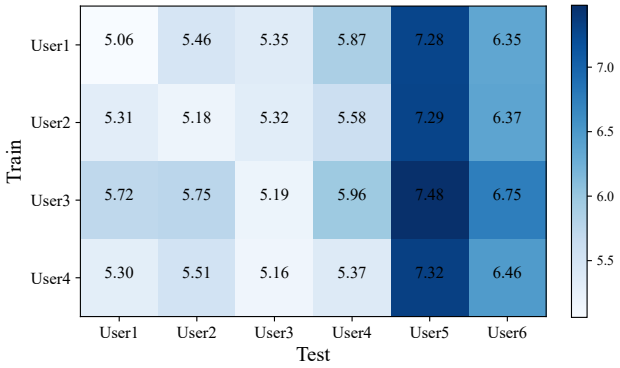


Fig. 15. Cross-user MCD evaluation.

sity of light decreases as the thickness of the glass increases. To investigate the impact of light intensity on EchoLight, we used a mirror as the reflective object and placed different thicknesses of glass between the light source and the mirror. At the same time, the light flux at the reflected object was measured separately. The results in Fig. 14 demonstrate a gradual decline in EchoLight's performance as the light flux decreases. However, even under the condition of a glass thickness of 30 mm (i.e. a decrease in light flux to 1200 lx), EchoLight still performs well.

7) Impact of cross-user training: Since different users may have varying pronunciation styles, and attackers might not have access to the victim's voice samples beforehand, it is essential to perform cross-user training and testing to assess the generalization capability of EchoLight. We performed separate training on User1 to User4 and evaluated the trained model on each user from User1 to User6. The experiment was conducted using the packing bag as the reflective object.

The evaluation results are shown in Fig. 15. It can be seen that optimal audio reconstruction performance is achieved by testing with User1 and training the model exclusively on User1's data. Similar trends can be seen for the other users. Nevertheless, even in cross-user training and testing scenarios, the MCD values remain consistently below 8. This demonstrates that EchoLight is capable of effectively executing attacks across different users. In addition, we observed that the performance with respect to User5 is slightly below average, a phenomenon that may be due to pronounced differences in speaking speed or accent between User5 and Users 1 to 4.

VII. Discussion

EchoLight is able to perform eavesdropping attacks in a variety of lighting conditions, including both sunlight and indoor lighting scenarios. The high intensity of sunlight often produces better results than indoor desk lamp light, with effectiveness depending on the specific light intensity. While the sunlight EchoLight attack can provide high-quality audio reconstruction, it presents certain challenges: (1) The movement of the sun, which is beyond human control, results in constant changes in illumination and reflection angles. As a result, attackers may need to dynamically adjust attack angles during an eavesdropping operation, especially if the time of eavesdropping is relatively long. (2) EchoLight's performance can be affected by changes in weather conditions, with adverse weather or nighttime conditions potentially resulting in a failed eavesdropping attempt. Despite these limitations, EchoLight retains its effectiveness when using indoor desk lamp lights or similar ambient light sources, ensuring the recovery of intelligible audio. On clear and sunny days, attackers can still choose sunlight to initiate an EchoLight eavesdropping attack, achieving the highest quality audio reconstruction.

In addition, the characteristics of reflective objects, such as their ability to vibrate and their reflective properties, significantly affect the effectiveness of EchoLight attacks. In fact, the passive vibration capacity of everyday objects, such as packing bag, and file folder, is generally adequate. Another limitation is the minimal sound pressure hitting on these reflective objects.

VIII. Conclusion

In this work, we have introduced EchoLight, an innovative method for ambient light eavesdropping that is capable of effectively recovering sounds occurring in the vicinity of a target. EchoLight leverage ambient light, which is either reflected or diffused by a range of reflective objects commonly found in everyday life, to recover sound. To enhance the audio quality and reconstruct the missing high-frequency components present in the optical signal, we implemented a bandwidth extension model based on the cGAN architecture. Our evaluation results demonstrate the effectiveness of EchoLight in reconstructing intelligible sounds under various settings, such as different sound source distances, different reflective materials, and moving audio sources. Additionally, EchoLight has the ability to launch attacks up to 40 meters away from the victim.

IX. Acknowledgement

This work is supported by the National Key Research and Development Program of China (No. 2021YFB3100400), National Natural Science Foundation of China (Grant No. 62202276, 62232010, 62202274, 62302274) and Department of Science and Technology of Shandong Province (No. 2022HWYQ-038, 2023TSGC0105).

REFERENCES

[1] P. Hu, W. Li, R. Spolaor, and X. Cheng, "mmecho: A mmwave-based acoustic eavesdropping method," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2022, pp. 836–852.

[2] P. Hu, W. Li, Y. Ma, P. S. Santhalingam, P. Pathak, H. Li, H. Zhang, G. Zhang, X. Cheng, and P. Mohapatra, "Towards unconstrained vocabulary eavesdropping with mmwave radar using gan," *IEEE Transactions on Mobile Computing*, 2022.

[3] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 11–20.

[4] C. Wang, F. Lin, T. Liu, K. Zheng, Z. Wang, Z. Li, M.-C. Huang, W. Xu, and K. Ren, "mmeve: eavesdropping on smartphone's earpiece via cots mmwave device," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 338–351.

[5] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu, and K. Ren, "mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 820–829.

[6] S. Basak and M. Gowda, "mmspy: Spying phone calls using mmwave radars," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1211–1228.

[7] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 14–26.

[8] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 593–604.

[9] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 130–141.

[10] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer." in *NDSS*, vol. 2020, 2020, pp. 1–18.

[11] P. Hu, H. Zhuang, P. S. Santhalingam, R. Spolaor, P. Pathak, G. Zhang, and X. Cheng, "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1757–1773.

[12] 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)*. IEEE, 2018, pp. 1000–1017.

[13] J. Han, A. J. Chung, and P. Tague, "Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2017, pp. 181–192.

[14] J. Choi, H.-Y. Yang, and D.-H. Cho, "Tempest comeback: A realistic audio eavesdropping threat on mixed-signal socs," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1085–1101.

[15] B. Nassi, Y. Pirutin, T. Galor, Y. Elovici, and B. Zadov, "Glowworm attack: Optical tempest sound recovery via a device's power indicator led," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 1900–1914.

[16] B. Nassi, Y. Pirutin, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Real-time passive sound recovery from light bulb vibrations," *Cryptology ePrint Archive*, 2020.

[17] M. G. Kuhn, "Optical time-domain eavesdropping risks of crt displays," in *Proceedings 2002 IEEE Symposium on Security and Privacy*. IEEE, 2002, pp. 3–18.

[18] D. Balzarotti, M. Cova, and G. Vigna, "Clearshot: Eavesdropping on keyboard input from video," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 170–183.

[19] Y. Xu, J. Heinly, A. M. White, F. Monrose, and J.-M. Frahm, "Seeing double: Reconstructing obscured typed input from repeated compromising reflections," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013, pp. 1063–1074.

[20] R. P. Muscatell, "Laser microphone," *The Journal of the Acoustical Society of America*, vol. 76, no. 4, pp. 1284–1284, 1984.

[21] M. Chounlakone and J. Alverio, "The laser microphone," 2002.

[22] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: eavesdropping via lidar sensors," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 354–367.

[23] S. Sami, S. R. X. Tan, Y. Dai, N. Roy, and J. Han, "Lidarphone: acoustic eavesdropping using a lidar sensor," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 701–702.

[24] 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.

[25] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 301–315.

[26] H. A. C. Maruri, P. Lopez-Meyer, J. Huang, W. M. Beltman, L. Nachman, and H. Lu, "V-speech: noise-robust speech capturing glasses using vibration sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–23, 2018.

[27] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, 2016, pp. 57–69.

[28] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, "{SPEAKE (a) R}: Turn speakers to microphones for fun and profit," in *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, 2017.

[29] A. Kwong, W. Xu, and K. Fu, "Hard drive of hearing: Disks that eavesdrop with a synthesized microphone," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 905–919.

[30] C. Wang, L. Xie, Y. Lin, W. Wang, Y. Chen, Y. Bu, K. Zhang, and S. Lu, "Thru-the-wall eavesdropping on loudspeakers via rfid by capturing sub-mm level vibration," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–25, 2021.

[31] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "Uwhear: through-wall extraction and separation of audio vibrations using wireless signals," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 1–14.

[32] C. Wang, F. Lin, Z. Ba, F. Zhang, W. Xu, and K. Ren, "Wavesdropper: Through-wall word detection of human speech via commercial mmwave devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–26, 2022.

[33] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," 2014.

[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)*. IEEE, 2018, pp. 1–5.

[35] M. Backes, T. Chen, M. Dürmuth, H. P. Lensch, and M. Welk, "Tempest in a teapot: Compromising reflections revisited," in *2009 30th IEEE Symposium on Security and Privacy*. IEEE, 2009, pp. 315–327.

[36] M. Backes, M. Dürmuth, and D. Unruh, "Compromising reflections-or-how to read lcd monitors around the corner," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 158–169.

[37] S. Chakraborty, W. Ouyang, and M. Srivastava, "Lightspy: Optical eavesdropping on displays using light sensors on mobile devices," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 2980–2989.

[38] K. Mowery, S. Meiklejohn, and S. Savage, "Heat of the moment: Characterizing the efficacy of thermal {Camera-Based} attacks," in *5th USENIX Workshop on Offensive Technologies (WOOT 11)*, 2011.

[39] D. Shukla, R. Kumar, A. Serwadda, and V. V. Phoha, "Beware, your hands reveal your secrets!" in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 904–917.

[40] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[41] B. Barras, "Sox: Sound exchange," *Flash informatique*, no. 9, pp. 3–6, 2012.

[42] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[43] A. S. 5, "Methods for calculation of the speech intelligibility index," *American National Standard*, 1997.