Hiding an Ear in Plain Sight: On the Practicality and Implications of Acoustic Eavesdropping with Telecom Fiber Optic Cables

Youqian Zhang*§, Zheng Fang*§, Huan Wu*‡✉, Sze Yiu Chau†, Chao Lu*, and Xiapu Luo*✉
*The Hong Kong Polytechnic University †The Chinese University of Hong Kong
‡ Technological and Higher Education Institute of Hong Kong

***Abstract*—Optical fibers are widely regarded as reliable communication channels due to their resistance to external interference and low signal loss. This paper demonstrates a critical side channel within telecommunication optical fiber that allows for acoustic eavesdropping. By exploiting the sensitivity of optical fibers to acoustic vibrations, attackers can remotely monitor sound-induced deformations in the fiber structure and further recover information from the original sound waves.**

This issue becomes particularly concerning with the proliferation of Fiber-to-the-Home (FTTH) installations in modern buildings. Attackers with access to one end of an optical fiber can use commercially available Distributed Acoustic Sensing (DAS) systems to tap into the private environment surrounding the other end. However, because the optical fiber alone is not sensitive enough to airborne sound, we introduce a "Sensory Receptor" that improves acoustic capture. Our results demonstrate the ability to recover critical information, such as human activities, indoor localization, and conversation contents, raising important privacy concerns for fiber-optic communication networks.

I. Introduction

An optical fiber, a flexible, transparent medium made from glass or plastic, is widely known for its ability to transmit light across long distances with minimal loss. It has revolutionized modern communications, enabling rapid data transmission over extended ranges, and now forming the backbone of high-speed internet, connecting regions, and continents across long distances [1]. Unlike electrical cables, which can emit radio-frequency (RF) signals that might be intercepted (e.g., TEMPEST attacks [2], [3] and crosstalks [4], [5]), optical fibers do not produce any RF emissions, thus making people believe that optical fibers are inherently more reliable transmission medium that poses fewer side-channel risks than their electrical counterparts [6], [7].

This study will challenge the assumption by showing a critical privacy problem within optical fibers that can be exploited to eavesdrop on personal information, including human activities and private conversations. The inherent sensitivity of optical fibers to external vibrations [8] provides a potential attack surface: Sound waves could cause tiny deformations in the optical fiber's structure; these deformations further result in slight phase shifts in the laser signals transmitted back and forth through the optical fiber; as a result, it is possible to recover acoustic information from these phase changes.

Indeed, the widespread adoption of Fiber-to-the-Home (FTTH) technology [9] in modern buildings across many countries/places[1] could intensify this concern. FTTH installations wire optical fibers directly into residential and commercial spaces to provide high-speed internet access. While one end of a fiber resides within the user's room, the other end is situated remotely at an optical distribution point [11], [12], [13]. By connecting the other end to a commercially off-the-shelf Distributed Acoustic Sensing (DAS) system (see detailed explanation in Section II-B), an attacker could remotely capture acoustic information from the victims' premises. It is essential to mention that, in many cases, multiple optical fibers are installed, each belonging to different internet service providers (ISPs). Usually, only one fiber is in active use, while the others remain unused (which are also known as "dark fibers" [14], [15]), running along walls, ceilings, and other interior structures. These fibers could potentially serve as unintended channels for eavesdropping.

Yet, implementing such an optical-fiber-based eavesdropping attack in practical scenarios is far from straightforward as described above. While the concept of acoustic information leakage through optical fibers has been qualitatively discussed since 2012 by Grishachev [16], [17] and others [18], these studies largely remain theoretical. More recently, in 2022, Hao et al. [19] demonstrated such an attack where both the optical fiber and the sound source were placed in close proximity on the same stainless steel experimental plate, and even in such an idealized setting, recovering meaningful acoustic information proved highly challenging. Additionally, this setting does not reflect practical conditions, where sound propagates through air or standard building materials rather than being directly coupled to the optical fiber. The attenuation in the propagation may make the attack more difficult. To date, the question

§ These authors contributed equally to this work.
✉ Corresponding Authors: Xiapu Luo (csxluo@comp.polyu.edu.hk), Huan Wu (huan.wu@vtc.edu.hk).

Network and Distributed System Security (NDSS) Symposium 2026
23-27 February 2026, San Diego, CA, USA
ISBN 979-8-9919276-8-0
https://dx.doi.org/10.14722/ndss.2026.240546
www.ndss-symposium.org

[1] In 2024, the penetration rate, defined as the proportion of households that have actively subscribed to and are using FTTH services, varied: United Arab Emirates (99.5%), South Korea (96.6%), China (93.6%), Hong Kong (89.9%), Singapore (87.5%), United States (28.7%), United Kingdom (26.3%) [10].

of whether such attacks can be successfully realized in real-world scenarios remains unanswered, leaving a critical gap in understanding its practical feasibility.

In this work, we will fill the gap by demonstrating the attacks under more practical scenarios. To better understand the potential threat, we present a threat model that is abstracted from realistic optical fiber network scenarios, detailing system functionality as well as the capabilities and limitations of an attacker (Section III). We then ask a yet unanswered question: "Can linearly laid fibers hear well enough?" To explore this, we conduct preliminary experiments to demonstrate both the capabilities and limitations of the optical fibers in an indoor context, and we find that a linearly laid optical fiber alone can hardly capture fine-grained acoustic information such as human speech (Section IV). Based on our observations from the preliminary experiments, to achieve effective acoustic monitoring, we identify the following four challenges (denoted as **C1–C4**) that need to be addressed.

C1: An Effective Structure to Capture Sound: As sound waves propagate through the air to reach the optical fiber, they attenuate rapidly, making it difficult for the sound waves to cause any sufficient deformation in the optical fiber. A primary challenge lies in developing a physical structure (which we call a "Sensory Receptor") that can amplify subtle pressure fluctuations, thereby enhancing the fiber's sensitivity to acoustic vibrations.

C2: Sound Recovery from Fiber Deformations: Even with a sensory receptor that enhances sensitivity, recovering sound waves from the resulting structural deformations in the fiber presents its own difficulties. The challenge lies in understanding the limits of this approach, including identifying the maximum range and volume of sound that can be captured with sufficient clarity.

C3: Evaluating Adequacy for Sound Recovery: A key technical question is how to assess the performance of the amplification structure itself. For successful eavesdropping, the sensory receptor must be sufficient to capture signals of interest, such as speech or specific sound patterns. It is crucial to use appropriate metrics and methods for evaluating whether the structure consistently captures usable audio signals.

C4: Practical Performance and Usable Information: Testing the practical performance of this eavesdropping approach in realistic settings is crucial. This challenge entails determining the specific types of information that can be consistently and reliably extracted. Understanding the limits of data fidelity, such as clarity of speech or detail of sound sources, helps to determine the overall effectiveness and the privacy implications.

Further, we provide detailed solutions, namely, **S1–S4**, to tackle corresponding challenges, guiding through our approach from principles to experimental validation.

S1 and S2: We introduce an effective sensory receptor to tackle the challenges of capturing and recovering acoustic signals. We quantitatively model and parameterize the process of eavesdropping through optical fibers, laying the groundwork for further research into the risks, as well as potential mitigation strategies. (Section V)

S3: We characterize a practical implementation of the proposed sensory receptor and demonstrate the fidelity of the recovered acoustic signals by comparing them with reference signals across different cases. (Section VI)

S4: By employing our proposed sensory receptor, or a combination of them, we can effectively recover multiple types of information. Additionally, integrating state-of-the-art deep learning algorithms allows us to push the limits of this attack further, uncovering detailed relationships between the amount of recoverable information, sound source volume levels, and distance. (Section VII and Section VIII).

Note that the goal of our work is to turn the interesting physical phenomena (i.e., optical fibers as sensors capturing vibrations) into a practical, end-to-end privacy attack, and demonstrate for the first time the success as well as limitations of such attacks through thorough and realistic experiments. We demonstrate that it is possible to infer human activities with performance exceeding that of random guessing, localize sound sources with an average error on the order of tens of centimeters, and capture spoken conversations by retaining over 80% of the information within $2\,\mathrm{m}$. These findings illustrate the fine-grained level of information that can potentially be recovered through such an optical-fiber-based method. Some reconstructed audio samples can be found at https://osf.io/wna5d/overview?view_only=c4203a45b5ae4238904d0627ebe8a561

II. Background

This section provides background on optical fiber sensing (OFS), and a type of OFS known as Distributed Acoustic Sensing (DAS).

A. Optical Fiber Sensing

When an optical fiber is subjected to external interference, the light wave transmitted in it will be modulated by external fields so that its characteristic parameters, such as intensity, phase, and polarization state, change accordingly. As a result, there is an opportunity to detect the changes in these characteristic parameters and further restore the external variations to achieve the sensing function. Optical fiber sensing can be classified into two categories: one is based on specialty optical fibers, and the other is based on standard optical fibers.

Specialty optical fibers are those carefully engineered to enhance sensing sensitivity or enable new sensing parameters and applications [20], [21]. They are not utilized for data transmission in telecommunication networks due to high loss or incompatibility with standard transmission equipment. In this study, since we are considering the scenario of telecommunications, we do not use specialty optical fibers for the purpose of eavesdropping. On the other hand, utilizing standard telecommunication optical fibers as sensing media is attractive because they have been extensively laid both underground and under the sea, connecting buildings and spanning continents. As of 2025, there are nearly 1.4 million kilometers of submarine

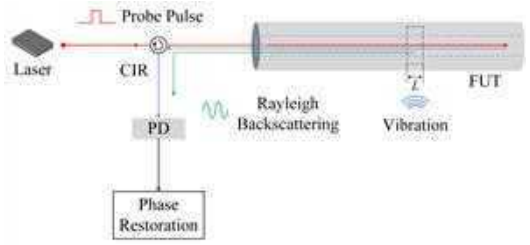


Fig. 1: Diagram of the DAS system. CIR: circulator, PD: photo-detector, FUT: fiber under test.

cables in service [22]. These telecommunication cables have already demonstrated capabilities for seismic detection [23], wildlife monitoring [24], traffic flow estimation [25], etc.

B. Basics on Distributed Acoustic Sensing

The most widely used sensing technology based on standard optical fibers is called Distributed Optical Fiber Sensing (DOFS). In DOFS, laser lights propagate through an optical fiber, and because of inherent scattering phenomena [8], such as Rayleigh scattering, the lights scatter everywhere along the optical fiber and reflect back to the transmitter. This unique scattering property allows for "distributed" sensing signals to be collected along the entire length of the optical fiber.

DAS is a typical example of DOFS. The structure and principles of the DAS system are briefly illustrated in Figure 1. Probe pulses from a laser are pumped through a circulator (CIR) into the fiber under test (FUT). When external vibrations induce stress on the fiber, changes in the phase of Rayleigh backscattering occur, in response to strain variations. This phase shift pattern is captured by a photo-detector (PD), enabling the system to retrieve acoustic wave parameters, such as frequency and amplitude, through phase restoration. Since DAS can detect real-time strain changes by demodulating the phase change of Rayleigh scattering, potentially allowing for the detection of sound waves occurring in the vicinity of the optical fiber, effectively turning the optical fiber into a covert listening device. In our study, we employ a commercial DAS system. Our work is the first to demonstrate the potential of using DAS in conjunction with telecommunication optical fibers to extract fine-grained information, such as human conversations, beyond the coarse, large-scale vibrations targeted in previous applications (e.g., [23], [24], [25]).

III. Threat Model

In this section, we introduce a system model that illustrates the eavesdropping scenario, and an attacker model.

A. System Model

A common Fiber-to-the-Home (FTTH) network is established using a point-to-multipoint infrastructure, which is also known as a passive optical network (PON), as depicted in Figure 2. This type of network originates from the Optical Line Terminal (OLT), managed by the ISPs. From the OLT, a fiber optic cable extends to a splitter, which distributes optical signals to various customers, and which is known as

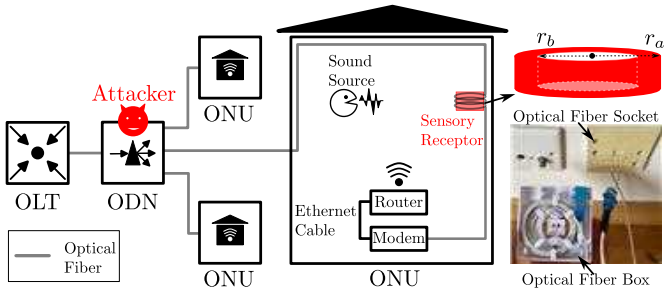


Fig. 2: The optical fiber network starts from OLT, extending to a splitter (ODN), and further connecting to an optical modem at the users' home (ONU).

the Optical Distribution Network (ODN). At the customer end, the Optical Networking Unit (ONU) interfaces with the ODN through optical fibers, receiving and processing signals to provide services to individual customers. Within the ONU, the fiber connection terminates at an optical modem, where optical signals are converted into Ethernet signals. For example, a router will disseminate data across local networks.

The wiring of optical fiber in a room can vary depending on the layout and design of the space. In modern buildings, it is common for the optical fiber to be channeled within the walls or run overhead along the ceilings, offering a concealed route. Alternatively, the fiber can be routed along the baseboards, providing an unobtrusive pathway. The wiring typically ends at an optical fiber outlet, as shown in Figure 2. Additionally, any excess length of optical fiber outside the outlet is usually gathered into an optical fiber box, as shown in Figure 2.

B. Attacker Model

In our model, we assume that an attacker has access to both the victim's premises (i.e., ONU) and the ODN. Such access can realistically be achieved, as FTTH deployments often involve physical access during installation, upgrades, or troubleshooting [26]. For example, this access may be obtained by an insider within the ISP, such as a technician or subcontractor, or alternatively, by attackers impersonating these roles, or through compromised third-party service providers, which are approaches that have been observed and documented in prior incidents [27].

To achieve effective eavesdropping, the attacker must overcome a key limitation: standard optical fibers may not be sensitive enough to air-borne sounds like human speech (see details in Section IV). To address this, the attacker can construct a sensory receptor, onto which the optical fiber is wound. This structure can enhance the fiber's ability to capture sound vibrations. Details of the sensory receptor's design and functionality will be discussed in subsequent sections. Note that the attacker can disguise the sensory receptor as the ordinary optical fiber box, as shown in Figure 2. This subtle camouflage allows the sensory receptor to blend in with other networking equipment for home/business, reducing the risk of raising suspicion. An example of the camouflage is demonstrated in our case study later (i.e., Figure 12 in Section VIII).

At ODN, the attacker identifies the specific fiber connected to the victim's room and links it to their own equipment, i.e., a Distributed Acoustic Sensing (DAS) device, which as mentioned before is capable of measuring the phase shift of light traveling through the fiber. With the fiber connection established, the attacker can use the available signal processing techniques for the phase-shift data, reconstructing the captured sound waves. By applying deep learning models, the attacker might even recognize speech content and other information.

Optical Fiber versus Other Sensors: Given a threat model where the attacker has physical access to the victim's premises, indeed, the attacker can perform other privacy attacks, such as wiretapping potentially sensitive network traffic. Directly listening to voice-based conversations through optical fibers is a new possibility enabled by our proposed attack; however, we acknowledge that it is not the only nor the most powerful one (see Section X for a discussion of other side-channel eavesdropping methods). Further, one might question why an attacker would not use conventional sensors such as microphones or cameras. Unlike microphones, which require electricity and may emit detectable radio-frequency (RF) signals (e.g., during analog-to-digital conversions [28], whether wired or wireless), optical fibers operate without electricity and do not emit RF signatures, making them invisible to standard RF scanners and electromagnetic detection tools [28]. Moreover, while hidden microphones and cameras have become common focal points in privacy audits and surveillance countermeasures [29], such as Technical Surveillance Countermeasures (TSCM) sweeps or bug sweeps, the acoustic sensing capability of optical fiber is relatively obscure to the public, or even professionals, and this obscurity increases the stealth of such attacks. In addition, defenders can deploy ultrasonic jammers to disrupt the microphones, while the performance of the optical-fiber-based method is not significantly affected (see more details in Section VIII-B3). Although this optical-fiber-based method may appear niche, it has value in high-stakes settings, such as corporate boardrooms, government and diplomatic facilities, where the use of conventional surveillance devices is heavily scrutinized and tightly controlled. In such contexts, the undetectable and unconventional nature of optical-fiber-based eavesdropping makes it a strategically potent tool for adversaries seeking to extract sensitive information without raising alarms.

IV. Can Linearly Laid Fibers Hear Well Enough?

What remains untested yet is whether these standard optical fibers, which are used for telecommunications in indoor environments, can capture detailed sound information, such as identifying the nature of the sounds, or any sensitive/critical information they may carry. To explore this, we conducted preliminary experiments.

A. Preliminary Experimental Setup

Our experiments were in a room with a wood floor. The total length of the standard telecommunication optical fiber is more than $5\,\mathrm{km}$, coiled on a big optical fiber spool designed for collecting kilometer-length optical fibers. *Note that the total length of the optical fiber does not directly correspond to the distance between the attacker and the victim. To avoid confusion, the maximum distance we will evaluate in this work is around* $50\,\mathrm{m}$*, as demonstrated by the case study in Section VIII.* As shown in Figure 3, we arranged the last 4 meters of the fiber along the baseboard of the room in an L-shaped configuration. The fiber was securely fixed to the baseboard using adhesive tape.

The other end of the fiber was connected to a DAS. The data gathered by the DAS system was then processed through a computer to reconstruct the audio signal (the specific method for signal reconstruction is discussed in Section V). We selected three equidistant points on the fiber within the room for signal analysis, located at 5014m, 5016m, and 5018m from the DAS system. It is worth noting that the optical fiber functions as thousands of independent sensing points, each capturing only local vibrations that deform a specific segment of the optical fiber. Our sensing points are within the $4\,\mathrm{m}$ tail of the optical fiber. Vibrations elsewhere along the fiber (including those on the spool) do not affect or interfere with the deformations measured at the tail.

B. Capturing Sound from Loudspeaker

A loudspeaker, on an acoustic foam that prevents sound propagation through the ground, is placed around $1\,\mathrm{m}$ away from the optical fiber, and plays a sound at $80\,\mathrm{dB}$[2] (approximately the volume of normal human loud speech [30], [31]) within the frequency range of $100\,\mathrm{Hz}$ to $1000\,\mathrm{Hz}$ (within the range of frequency of human speech [31]). However, no discernible audio signal could be recovered from the data collected by the DAS system. This failure is attributed to the fact that sound, as a pressure wave propagating through air, is attenuated quickly. Also, the thickness of the optical fiber is at a micrometer scale, and the sound wave induces insufficient deformation in the optical fiber.

C. Capturing Sound of Walking

We marked 12 points along the fiber, as shown in Figure 3, and a leather shoe hit these points 10 times, involving a "heel-to-toe" pattern. The sound levels of the footsteps, measured with a decibel meter, are 76 dB on average. We collected signals at the three selected points along the fiber and from each point, we successfully reconstructed the footstep vibrations.

The reconstructed time-domain and frequency-domain signals are shown in Figure 3. We observed 12 steps, and the frequency-domain analysis revealed that the vibrations primarily occurred in the $10\,\mathrm{Hz}$ to $100\,\mathrm{Hz}$ range. This experiment demonstrates that walking-induced vibrations can be captured well. This is because the vibration from footsteps

[2]The sound pressure level in dB is defined and explained in Section V-A. We did experiments and found that the sound level of human speech is $83.8\,\mathrm{dB}$ on average (ranging from $51.1\,\mathrm{dB}$ to $97.2\,\mathrm{dB}$), and more details are presented in Appendix A. In addition, most previous studies, as discussed in Section X, used sound levels higher than $80\,\mathrm{dB}$. While internet searches may suggest that normal speech is between 60 and $70\,\mathrm{dB}$, these measurements are typically taken from a distance and do not reflect the actual sound level at the source.

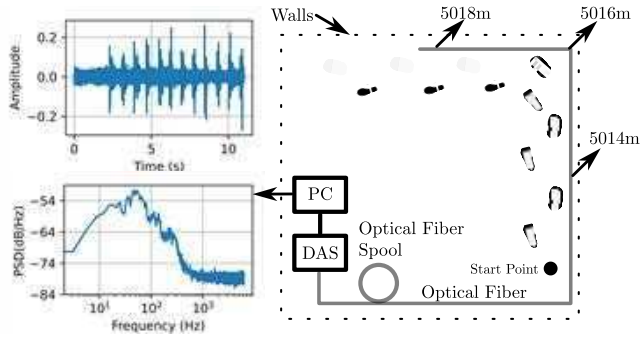


Fig. 3: Preliminary experimental setup. An optical fiber is deployed along the baseboard of a room. A DAS is used to collect vibration caused by sound sources, and a computer (PC) is used to recover the acoustic sound, as shown on the left.

transmits through the floor to the baseboard, causing sufficient deformations for the optical fiber to capture.

D. Observations

As shown above, the linearly laid optical fibers are effective at capturing structure-borne mechanical vibrations (e.g., walking), but inadequate for capturing air-borne acoustic signals like human speech. This limitation suggests that, to achieve finer sensitivity to sound waves, especially for applications including human speech, it is necessary to modify the system in a way that enhances its responsiveness to air-borne acoustic waves. Hence, we design the sensory receptor as follows.

V. Design of Sensory Receptor

To address **C1** and **C2**, we begin by finding a structure of the sensory receptor to capture the sound and recover it.

Although the sound pressure on a small segment of optical fiber in the perpendicular direction is minimal, if it can be converted into a longitudinal strain along the fiber and accumulated, it could result in a more noticeable deformation. Inspired by previous designs of fiber-optic microphones [32] and accelerometers [33], which used specialty optical fibers (see details in Section II-A), a similar effect can be achieved by winding the telecommunication optical fiber around a cylindrical hollow structure, as illustrated in Figure 2. This approach achieves both a directional transformation and an accumulative effect: changes in the cylinder's diameter (caused by sound waves) translate into stretching and contracting forces along the fiber's length, while the coiling allows a longer fiber segment to be subjected to the strain. In the following subsections, we will model and explain how acoustic information is recoverd.

A. Sound Propagation

For a point source with sound pressure of p, its sound pressure level (SPL) P in dB can be represented by $P = 20\log\frac{p}{p_0}$, where $p_0 = 20\times10^{-6}\text{Pa}$ is the reference sound pressure in air. Let d represent the distance between the sound source and the receiver, i.e., sensory receptor. The attenuation in sound pressure level, ΔP, due to spreading over a spherical surface [34] is given by: $\Delta P = 10\log(\frac{1}{4\pi d^2})$. This relationship shows that if the distance d is doubled, the sound pressure level decreases by approximately 6 dB. The sound pressure level at the receiver is then $P_r = P+\Delta P$. Note that the pressure p_r on the receiver can be expressed as: $p_r = p_0 10^{\frac{P_r}{20}}$. Substituting the previous equations into the equation of p_r, we obtain

$$p_r = \frac{p}{d}\cdot\frac{1}{2\sqrt{\pi}} \tag{1}$$

Equation 1 indicates that the sound pressure p_r at the sensory receptor is directly proportional to the initial sound pressure p at the source, and conversely, inversely proportional to the distance d.

B. Deformation and Phase Change

Consider a hollow cylinder with an external radius r_a and an internal radius r_b. When a change in sound pressure, Δp_r, is applied, it causes a small expansion in the hollow cylinder's outer radius. The change in r_a is given by [35]:

$$\Delta r_a = \frac{\Delta p_r\cdot r_a}{E}(\frac{r_a^2+r_b^2}{r_a^2-r_b^2}-v) \tag{2}$$

where E is Young's modulus of the material, describing its elasticity; v is the Poisson ratio, which measures the tendency of the material to expand in directions perpendicular to the applied force.

Now, let's consider a fiber of length L wound tightly around this cylinder. The pressure-induced expansion of its outer radius Δr_a results in a proportional change in the length of the fiber, denoted ΔL. This relationship is expressed as $\frac{\Delta r_a}{r_a} = \frac{\Delta L}{L}$. Note that the phase of the light propagating the fiber with refractive index n is $\phi = \frac{2\pi}{\lambda}nL$, where λ is the optical wavelength in vacuum [36]. A change in the fiber's length ΔL will cause a corresponding change in the phase of light, $\Delta\phi$, given by $\Delta\phi = \frac{2\pi}{\lambda}n\Delta L$. Substituting $\Delta L = L\frac{\Delta r_a}{r_a}$ into the equation of $\Delta\phi$, we get $\Delta\phi = \frac{2\pi}{\lambda}n'L\frac{\Delta r_a}{r_a}$, where $n' = 1-\frac{n^2}{2}(p_{12}-v_f(p_{11}+p_{12}))$. Note that p_{12} and p_{11} are the strain-optic coefficients, which describe how the refractive index n changes with strain, and v_f is the Poisson ratio of the fiber [36]. Substituting the expression for Δr_a from Equation 2, we obtain $\Delta\phi = \frac{2\pi}{\lambda}n'L\frac{\Delta p_r}{E}(\frac{r_a^2+r_b^2}{r_a^2-r_b^2}-v)$. Finally, according to Equation 1, we can express $\Delta p_r = \frac{\Delta p}{d}\cdot\frac{1}{2\sqrt{\pi}}$, assuming d is not changed, and substitute it into the equation of $\Delta\phi$ above and get:

$$\Delta\phi = \frac{\sqrt{\pi}}{\lambda}n'L\frac{\Delta p}{d\cdot E}(\frac{r_a^2+r_b^2}{r_a^2-r_b^2}-v) \tag{3}$$

C. Sound Information Recovery

As indicated by Equation 3, each phase change corresponds directly to a change in the sound pressure at that moment, so tracking these phase changes over time essentially captures the oscillating pattern of the original sound wave.

The phase change sequence $[\Delta\phi_0, \Delta\phi_1, ..., \Delta\phi_i]$ represents the sound-induced variations in the fiber length caused by the

original sound signal, each corresponding to a specific point in time. Note that these values are sampled at a consistent interval. To avoid aliasing error, according to the Nyquist-Shannon sampling theorem [37], we assume that the sampling rate is greater than or equal to twice the highest frequency component in a sound wave to be measured. To recover the sound signal, we can calculate the cumulative sum of these incremental phase changes over time. Let us denote the recovered signal as $s = [s_0, s_1, ..., s_i]$, where

$$s_i = \Sigma_{k=0}^{i} \Delta\phi_k.$$

Note that to mitigate the impact of external vibrations introduced during the return path of the reflected light to the DAS, we can select a reference point on the optical fiber that is physically close to the sensing point but located outside the sensory receptor. Supported by our prior experiments, vibrations at such a reference point cause minimal mechanical deformation, resulting in an extremely small phase shift. As a result, the phase difference measured between the sensing point and the reference primarily captures the signal of interest. More importantly, when the light scattered from both points travels back to the DAS through the same optical fiber, since the speed of light is orders of magnitude faster than any mechanical vibration, any external vibration affecting the return path tends to be common-mode. Thus, by measuring the differential phase between two adjacent points, such common-mode noise naturally cancels out in the subtraction, making the return-path impacts negligible in the sound information recovery. In Section VIII-B1, we experimentally demonstrate and analyze that the effects of external interference on the optical fiber as conduit are minimal.

D. Limitations Due to Noise and Saturation

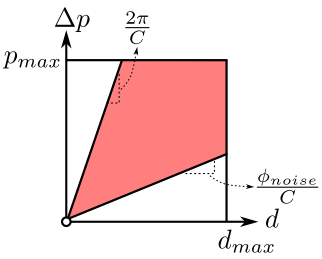
To make the attack practical, we need to carefully consider the effect of noise, which inevitably affects the phase change measurements used to recover sound signals. Various factors, such as environmental and system noise, introduce noise into the phase changes, which we denote as $\phi_{noise} \geq 0$. The maximum phase change measurable by the system is limited to 2π. The maximum sound pressure is denoted as p_{max}, and the maximum distance between the receptor and the furthest sound source allowed in the space is denoted as d_{max}.

For the proposed method to work reliably, the phase changes must lie within the range $\phi_{noise} < \Delta\phi \leq 2\pi$. By substituting Equation 3 into the inequality, and introducing a constant $C = \frac{\sqrt{\pi}}{\lambda} n' L \frac{1}{E} (\frac{r_a^2 + r_b^2}{r_a^2 - r_b^2} - v)$, where $C > 0$, we can rewrite the inequality as:

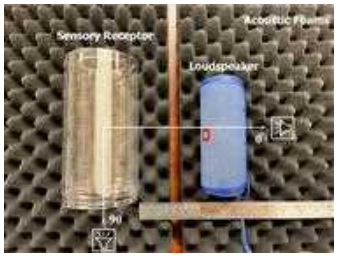
$$\frac{\phi_{noise}}{C} < \frac{\Delta p}{d} \leq \frac{2\pi}{C}$$

where $0 < \Delta p \leq p_{max}$ and $0 < d \leq d_{max}$.

This relationship, illustrated in Figure 4a, defines a bounded region (the shaded area) within which the attack is feasible. Below the curve $\Delta p = \frac{\phi_{noise}}{C} d$, the sound pressure is too weak for the system to distinguish it from noise, making sound recovery unreliable. Above the curve $\Delta p = \frac{2\pi}{C} d$, the phase changes exceed 2π, leading to saturation, and preventing sound recovery. This analysis reveals how the attack is limited by the laws of physics. A resourceful attacker can attempt to adjust C by tuning system parameters such as the fiber length L or the material properties to affect n'. Nevertheless, even with these limitations, a meaningful privacy attack can be achieved with off-the-shelf commodity optical fibers.



(a)



(b)

Fig. 4: (a) Within the shaded region (red), an attacker can achieve effective sound capture and recovery. (b) A testbed for the characterization.

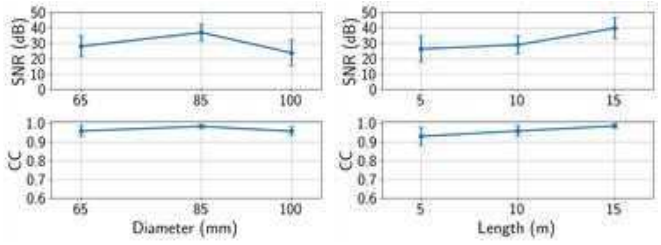
VI. Characterizing Sensory Receptor

To address **C3**, we assess the performance of the sensory receptor across various cases, demonstrating its consistent capability to recover high-quality signals. We start by selecting an appropriate material for the sensory receptor. We considered different types of materials, including polyethylene terephthalate (PET), resin, polyamide (PA), etc. By experiments, PET (as shown in Figure 4b) was ultimately chosen for two reasons: first, its transparency, which helps concealment, and second, its ability to capture high-quality sound signals, as further discussed in this section.

A. Sensory Receptor Performance

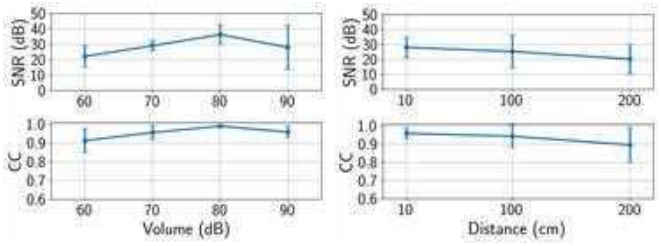
To demonstrate the performance and discuss the impact of associated parameters, we evaluated the quality of the restored signals by comparing them with sinusoidal signals at different frequencies.

1) Testbed: A testbed is shown in Figure 4b. Both the loudspeaker and the sensory receptor are placed on acoustic foams to ensure that the sound waves travel through the air. The ambient noise level in the room is between $50\,\mathrm{dB}$ and $60\,\mathrm{dB}$. The loudspeaker volume ranges from $60\,\mathrm{dB}$ to $90\,\mathrm{dB}$, and the frequency of the sound varies from $100\,\mathrm{Hz}$ to $1000\,\mathrm{Hz}$, which include the human speech volume range and frequency band as discussed in Section IV-B. The distance between the loudspeaker and the sensory receptor varies from $10\,\mathrm{cm}$ to $200\,\mathrm{cm}$, and the angle between them spans from $0°$ to $90°$. Note that all these parameters will be varied so as to study their impacts. The optical fiber is wrapped on the sensory receptor manually, and one end is connected to a DAS. Note that the laser's wavelength used in the experiments is $1550\,\mathrm{nm}$, per the norms for telecommunications. The impacts of optical wavelength on performance will not be further discussed in this paper because the optical wavelength and optical network structure used in the communication system are well-matched and cannot be changed arbitrarily.



(a) (b)

Fig. 5: SNR and CC versus (a) outer diameter; (b) wrapping fiber length.



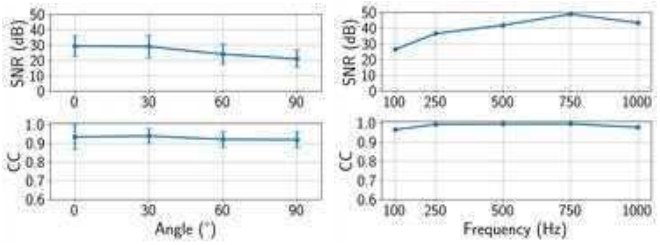
(a) (b)

Fig. 6: SNR and CC versus (a) volume; (b) distance.

2) Metrics for Evaluation: Two commonly used metrics were selected to quantitatively assess signal quality: the signal-to-noise ratio (SNR) and the correlation coefficient (CC). SNR (in dB) compares the level of a desired signal to the level of background noise [38]. The higher the SNR value, the better the signal quality, meaning that there is more useful information (signals) than unwanted data (noise). CC is a numerical measure of linear correlation, representing the similarity between two signals [39]. The value of CC ranges between -1 and 1, with values closer to 1 representing a higher similarity.

3) Intrinsic Parameters: The analysis focused on two parameters related to the fiber-wrapped cylinder: the outer diameter of the cylinder and the length of the wrapping fiber.

Outer Diameter: Three sizes were chosen, which are $65\,\mathrm{mm}$, $85\,\mathrm{mm}$, and $100\,\mathrm{mm}$. Note that all cylinders have the same thickness, i.e., $r_a - r_b = 0.2\,\mathrm{mm}$. So this is equivalent to studying the impacts of $(\frac{r_a^2+r_b^2}{r_a^2-r_b^2} - v)$. In Figure 5a, averaged results of various (other) parameters are presented, and they indicated that both SNR and CC levels were highest at an outer diameter of $85\,\mathrm{mm}$. However, according to Equation 3, the introduced phase variation should be more significant under a bigger outer diameter. To explain this result, we analyzed the noise level in the three groups of data. It was found that the root mean square (RMS) rises from 0.12 to 3.96 as the outer diameter increases from $65\,\mathrm{mm}$ to $100\,\mathrm{mm}$. The result could be explained by the fact that the larger the outer diameter, the greater the air mobility inside the hollow cylinder, which also introduces a greater noise level. In addition, satisfactory performance was also observed with an SNR of $18\,\mathrm{dB}$ and a CC of 0.92 at a diameter of $65\,\mathrm{mm}$. Considering the stealthiness of the sensory receptor, the $65\,\mathrm{mm}$ cylinder was selected for subsequent experiments.



(a) (b)

Fig. 7: SNR and CC versus (a) angle; (b) frequency.

Length of Wrapping Optical Fiber: $5\,\mathrm{m}$, $10\,\mathrm{m}$, and $15\,\mathrm{m}$ of fibers were wrapped for comparison; the results are shown in Figure 5b. It can be seen that the longer the fiber, the better the quality of signal restoration. When L is $15\,\mathrm{m}$, CC can be as high as 1 across all frequencies. This is because more points on the fiber experience the pressure of sound waves, matching our modeling in Equation 3.

4) External Parameters: Regarding the sound source, we picked volume, distance, angle, and frequency for evaluation.

Volume: It can be noticed in Figure 6a that the SNR and CC improve as the volume increases. Beyond $80\,\mathrm{dB}$, the mean of SNR and CC deteriorate slightly; however, they are still high enough to indicate a good enough sound recovery quality.

Distance: In Figure 6b, it can be seen that as the distance increases, the SNR and CC decrease as well. Even at $2\,\mathrm{m}$ away, SNR and CC maintain around $20\,\mathrm{dB}$ and 0.9, respectively.

Angle: We placed the loudspeaker $1\,\mathrm{m}$ away from the sensory receptor and moved it around the sensory receptor to vary the angle at which the sound waves reach it. It can be noticed from Figure 7a that the SNR is better when the angle is smaller. It can be construed that the smaller the angle, the larger the area of contact between the sound wave and the cylinder, and the more intense the pressure introduced. There is no obvious trend with respect to CC. Overall, the results indicate that signal capture and restoration perform well across different angles.

Frequency: The frequency response is shown in Figure 7b, and the SNR can be maintained above $25\,\mathrm{dB}$ in the tested frequency range while the CC is consistently above 0.95.

The results above prove that the crafted optical fiber sensory receptor matches our modeling, and it can capture and restore the signal well in various cases. We decided to wrap $15\,\mathrm{m}$ of optical fiber around a $65\,\mathrm{mm}$ outer diameter PET hollow cylinder as a discrete sensory receptor for further experiments.

VII. Experiments of Eavesdropping

In this section, we aim to address **C4**, assessing the range of information our approach can capture in practical settings. We begin with *sound event detection*, which involves analyzing collected sounds to identify domestic activities or events occurring in the environment. This step provides a broad overview of the sounds present without focusing on precise details. Building on this, we proceed to *indoor localization*, where we determine the spatial origin of detected sounds. By identifying the approximate location of sound sources,

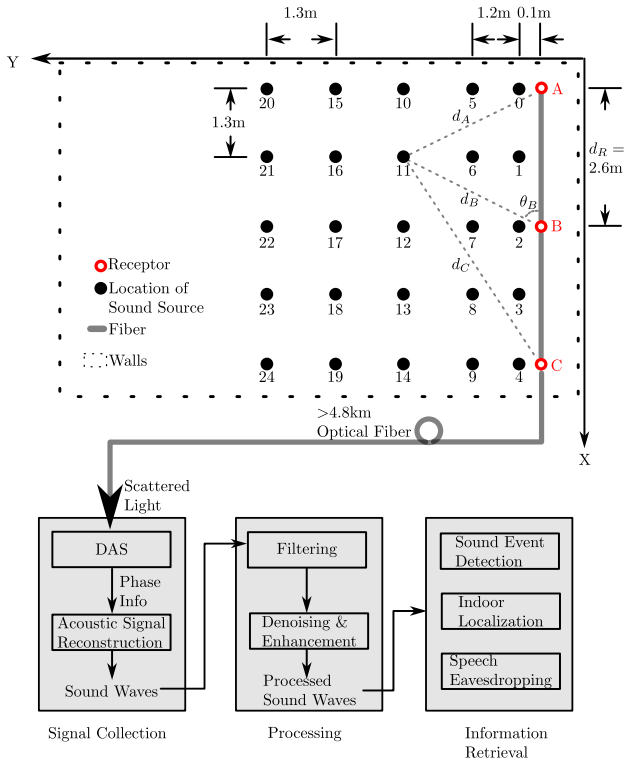


Fig. 8: In our experimental setup, A, B, and C are receptors. There are multiple points where a loudspeaker is placed to play sound. The scattered lights are then collected by a DAS and reconstructed into sound waves and further processed, and finally, information is retrieved.

it allows an adversary to map the layout of activity within the monitored space. Finally, we focus on the most privacy-critical level of information extraction: *speech eavesdropping*. Here, we aim to identify specific speech content, allowing the recovery of actual spoken words or phrases.

A. Setup

We first present the layout of our experimental environment, and then the workflow of data collection.

1) Layout: The experiment was conducted in a room measuring approximately $8\,\mathrm{m}$ in length and $6\,\mathrm{m}$ in width. As shown in Figure 8, three receptors were positioned along one of the shorter walls, equally spaced at intervals of $2.6\,\mathrm{m}$. With three receptors, it is possible to triangulate the position based on the time or phase differences in the signals detected at each receptor, providing a spatial estimate. We utilized a single optical fiber over $4.8\,\mathrm{km}$ in length. On each receptor, $15\,\mathrm{m}$ of fiber was tightly wound to form the sensing structure, while an additional $5\,\mathrm{m}$ segment of fiber remained uncoiled between adjacent receptors to maintain separation. The distances from the measurement points on the receptors to the DAS system, are $4840\,\mathrm{m}$, $4820\,\mathrm{m}$, and $4800\,\mathrm{m}$, respectively, labeled as points A, B, and C.

We marked 25 positions at intervals of 1.2–$1.3\,\mathrm{m}$ to place the sound sources (i.e., loudspeakers), as shown in Figure 8, numbered 0 to 24. We defined point A as the origin, with

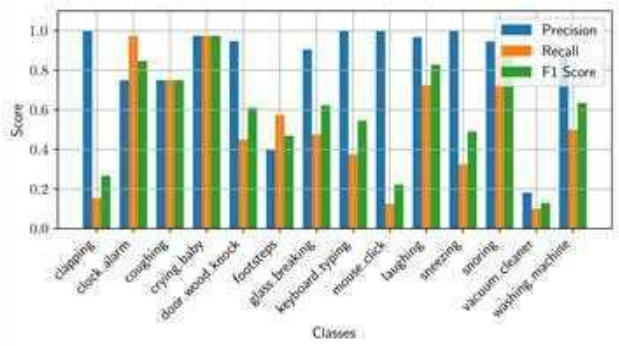


Fig. 9: Precision, recall, and F-1 score of detection results of different sound events.

the x-axis running along the width of the room and the y-axis along its length. Using this coordinate system, we could express the positions of the measurement points and sound sources in the room. For instance, the coordinates of points A, B, and C are (0, 0), (2.6, 0), and (5.2, 0), respectively. Similarly, the coordinates of the 25 sound source locations could be defined. Note that the loudspeaker was maintained at a constant volume of approximately $80\,\mathrm{dB}$ hereafter, unless stated otherwise.

2) Workflow: After the scattered light is captured by the DAS, the system follows a structured workflow to extract information from the collected signals. As shown in Figure 8, the process begins with "Signal Collection". In this stage, the DAS detects phase variations, which are induced by sound waves interacting with the optical fiber as discussed previously. A computer then processes these phase variations to reconstruct the corresponding sound waves.

Following the signal collection stage, the reconstructed sound waves proceed to the "Processing" stage, where they first go through a "Filtering" step to remove unwanted high-frequency and low-frequency noise components. A Butterworth filter is used to keep the frequency band between $50\,\mathrm{Hz}$ and $3000\,\mathrm{Hz}$. After filtering, the signals may undergo "Denoising & Enhancement", where techniques are applied to further reduce noise that overlaps with the frequency band of the sound source, so as to further enhance sound quality.

The final stage is "Information Retrieval", where the processed sound waves are analyzed to extract specific types of information for sound event detection, indoor localization, and speech eavesdropping.

B. Sound Event Detection

We begin by introducing the sound source used in our experiments, followed by a brief introduction to the detection models employed for this task. Next, we describe the metrics used for evaluation, and finally, we present a detailed analysis of the results.

1) Sound Sources of Domestic Activities: The loudspeaker was used to play 14 selected sound clips from the ESC-50 [40] dataset, which is commonly used for sound event classification, focusing on sounds associated with domestic activities such as clock alarms, coughing, keyboard typing, washing

TABLE I: Sound Event Detection Accuracy of Different Models at Different Distances

Models	Accuracy			
	Ref	0.1 m	1 m	2 m
BEATs [42]	0.96	0.53	0.11	0.06
HTS-AT [43]	0.97	0.39	0.06	0.05
Efficient-AT [45], [46]	0.97	0.44	0.11	0.07
Our Fine-tuned	0.97	**0.83**	**0.50**	**0.43**

machines, etc. Each sound category included 40 individual sound clips. We positioned the loudspeaker at distances of $0.1\,\mathrm{m}$, $1\,\mathrm{m}$, and $2\,\mathrm{m}$ from point A.

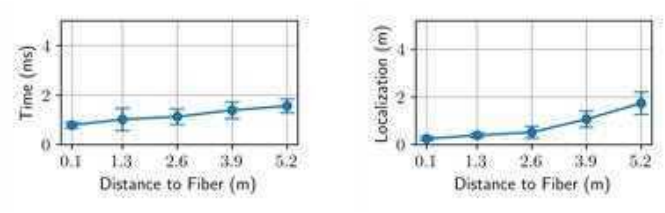
2) Detection Models: Sound event classification begins with transforming audio into mel spectrograms, creating a time-frequency representation that captures the audio's key characteristics, and further being processed by deep learning models. The most advanced methods for analyzing these spectrograms fall into two categories. The first category uses Transformer-based architectures [41], e.g., the state-of-the-art models BEATs [42] and HTS-AT [43]. The second category, uses Convolutional Neural Network (CNN) [44], specifically a model called Efficient-AT [45], [46]. We selected the three aforementioned state-of-the-art sound event classification models: BEATs, HTS-AT, and Efficient-AT. Importantly, we used these models in their original form, without any fine-tuning on the dataset ESC-50 [40].

3) Detection Evaluation Metrics: We use three common metrics to evaluate detection results. Precision measures the proportion of correctly predicted positives out of all predicted positives: $\frac{\text{True Positives}}{\text{True Positives+False Positives}}$. Recall measures the proportion of actual positives correctly identified: $\frac{\text{True Positives}}{\text{True Positives+False Negatives}}$. F1-Score is the harmonic mean of precision and recall, balancing the above two metrics: $2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$. In addition, to compare the performance of different models on our data, and the benchmark, we also include the accuracy, which represents the proportion of correctly classified samples (both positive and negative) out of the total samples: $\frac{\text{True Positives+True Negatives}}{\text{Total Number of Samples}}$.

4) Results and Analysis: We began by using the BEATs model for sound event classification on the reconstructed sounds from fiber vibration.

Performance of BEATs: Figure 9 presents precision, recall, and F1 scores for each class. The "clock_alarm", "crying_baby", and "snoring" classes achieve high scores across all three metrics, indicating high accuracy and consistency in classification. On the other hand, classes like "keyboard_typing" and "mouse_click" exhibit low recall. This discrepancy is likely due to these sounds' power being approximately 10 dB weaker than others, resulting in weaker vibrations captured by the fiber. This lower signal-to-noise ratio (SNR) affects the recall for these classes. A more detailed confusion matrix on the classification results is presented in Figure 16 in Appendix B.

Performance across Different Models: Next, we extended the experiment to include the other two models and tested data collected at various speaker-to-fiber distances. Additionally, we used the original dataset as a reference and employed



(a) (b)

Fig. 10: With the distance between the sound source and the fiber increases, (a) the estimation error of time difference increases, and (b) the estimation error of localization increases.

accuracy as the evaluation metric. The results are shown in Table 1. It can be observed that these models perform well on the original audio data; however, classification accuracy halves when applied to sound reconstructed from fiber vibrations, especially when the event happens at distances greater than 1 m, where accuracy falls to 0.05. This decline can be attributed to two main factors. First, these models were not fine-tuned on the dataset, so they lack familiarity with the characteristics of sounds collected and reconstructed through fiber optics. Second, as the distance increases, the signal-to-noise ratio (SNR) decreases, leading to poor recognition of certain features in the spectrogram.

Finetuning to Improve Accuracy: To address the issues above, we fine-tuned the Efficient-AT model on our dataset, using 80% of the data for fine-tuning and the remaining 20% for testing; we repeat this process with five-fold cross-validation and report the averaged results in Table 1 under the row labeled "Our Fine-tuned". After the fine-tuning (200 epochs), the accuracy improved to 0.83. Although this accuracy is still lower than that achieved with the original audio data, it is sufficient for an attacker to infer information about the activities occurring within a room. At farther distances, the accuracy drops to 0.43, but this still surpasses the random-guessing probability of 0.07 for 14 classes, indicating adequate performance.

C. Indoor Localization

With the reconstructed sound from our sensory receptor, placing three receptors can be used to locate the sound's position in a room using the time difference of arrival. The detailed description of the localization method is presented in Appendix C.

1) Localization Error Metric: We use Euclidean distance between the estimated position $(\overline{x}_S, \overline{y}_S)$ and the ground truth (x_S, y_S) to measure the performance of the localization method, and the metric is calculated as: $\Delta d = \sqrt{(x_S - \overline{x}_S)^2 + (y_S - \overline{y}_S)^2}$. A smaller Euclidean distance indicates a more accurate estimate.

2) Results and Analysis: We began with the time difference estimation and then the location estimation.

Time Difference Estimation: Apart from the sound source at location 1, where the time estimation is far beyond a normal value, the averaged absolute estimation error of time difference for all other locations is $1.19\,\mathrm{ms}$, with a standard

error of $0.14\,\mathrm{ms}$. With the sound source leaving the fiber, the estimation of time becomes less and less accurate, as shown in Figure 10a. This is due to lower sound pressure on the receptor and reduced SNR at greater distances, making it harder to detect the sound's onset. The details on the estimation of the time difference are shown in Table III in Appendix C.

Location Estimation: Based on the estimation of the time difference, the position of the sound source can be further solved as explained in Appendix C, and the estimation of the position of the sound source is presented in Table IV. The averaged estimation error of localization is $0.77\,\mathrm{m}$, with a standard error of $0.17\,\mathrm{m}$. From Figure 10b, it can be found that the estimation error becomes larger while the sound source is moving away from the fiber. This is attributed to the increase in estimation error in time difference. However, in an area of $27.04\,\mathrm{m}^2 (= 5.2\,\mathrm{m} \times 5.2\,\mathrm{m})$, the error of localization is below $1\,\mathrm{m}$, which still presents a significant privacy risk in an indoor setting.

D. Speech Eavesdropping

This section further demonstrates the feasibility and the limitations of eavesdropping on human speech.

1) Sound Sources of Human Speech: For the eavesdropping on human speech, the loudspeaker played audio samples from the Librispeech [47] dataset, which is widely used in automatic speech recognition research. From the training subset of this dataset, we randomly selected 15 male and 15 female speech clips, and a matching set of clips was also selected from the testing subset. Thus, in total 60 clips were chosen. In addition, we extended the loudspeaker placement from $2\,\mathrm{m}$ to distances of $3\,\mathrm{m}$ and $4\,\mathrm{m}$ (for more discussion on increased distance and obstacles, please see Appendix E).

2) Speech Recognition Principles and Models: Note that the goal of the attacker is to learn the speech contents of the victim. Automatic speech recognition (ASR) using transformer models has become a dominant approach in recent years due to its ability to transcribe spoken language into text. Transformer-based ASR leverages the self-attention mechanism, which allows the model to focus on relevant parts of an audio sequence, regardless of distance in time. State-of-the-art ASR models, such as whisper-large-v3 [48], canary-1b [49], and parakeet-tdt-1.1b [50], employ transformers that are trained on massive amounts of unlabeled audio data, learning robust audio representations before fine-tuning on labeled datasets for transcription. Note that these models have already learned features from the speech clips in the training dataset of Librispeech. However, they have not been exposed to, nor have they learned features from, the speech clips in the testing dataset.

3) Speech Recognition Evaluation Metrics: Word Error Rate (WER) is one of the most common metrics for evaluating the accuracy of automatic speech recognition systems [51]. It is calculated as the sum of substitutions (S), deletions (D), and insertions (I) divided by the total number of words in the reference transcript (N): $\mathrm{WER} = \frac{S+D+I}{N}$. A lower WER indicates better accuracy.

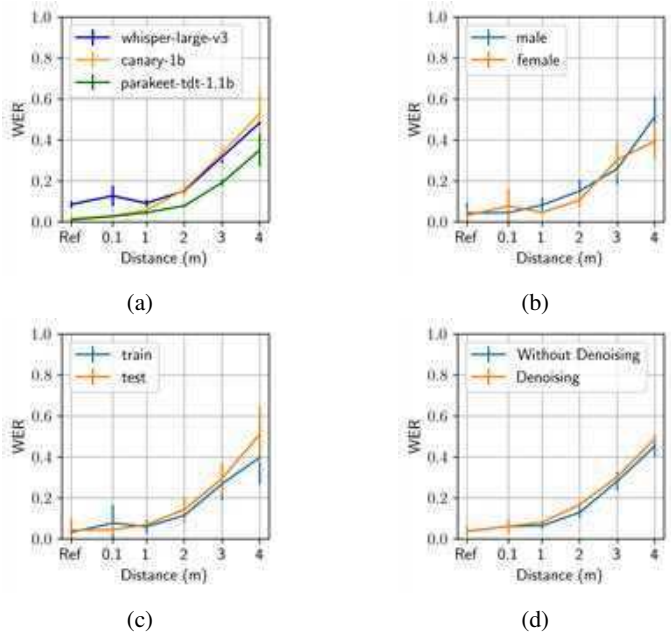


Fig. 11: Combined analysis of Word Error Rate (WER): (a) WER vs. distance, (b) WER by gender, (c) WER by training vs. testing, and (d) WER with and without denoising.

4) Results and Analysis: We first applied the ASR models on the selected clips (clean audios) from Librispeech, and the WER corresponds to "Ref" in the x-axis tick in Figure 11. The averaged WER value of these clips is around 0.07.

Effectiveness of Different Models: From the experimental results, as shown in Figure 11a, we observe that up to 1 meter, these models can achieve a WER below 0.1, meaning only 10 incorrect words per 100 words. At 3 meters, the WER remains at approximately 0.3, while at 4 meters it increases to around 0.5. These results indicate that our system can achieve relatively reliable eavesdropping within a 3-meter range. Among the models tested, parakeet-tdt-1.1b consistently achieves the lowest WER, making it a prime choice for attackers looking to maximize transcription accuracy.

Impacts of Voice Pitch: We further analyze the impact of voice pitch, broken down by speaker gender, by averaging results across all models, and the results are presented in Figure 11b. In our dataset, male voices are between $100\,\mathrm{Hz}$ to $150\,\mathrm{Hz}$, while female voices are in the $150\,\mathrm{Hz}$ to $300\,\mathrm{Hz}$ range. Across most distances, including $1\,\mathrm{m}$, $2\,\mathrm{m}$, and $4\,\mathrm{m}$, female voices tend to have lower WER, suggesting more accurate recognition. This may be due to the lower frequencies of male voices, which are more likely to overlap with low-frequency system noise, thereby reducing recognition accuracy.

Impacts of Familiarity with Speech: Next, we consider the models' performance based on familiarity with the data source, distinguishing between training and testing data results. Since the training data has been previously learned by the model, it tends to achieve lower WER with this data compared to the testing data, particularly as the distance increases, as shown in Figure 11c. This result implies that if the attacker can obtain

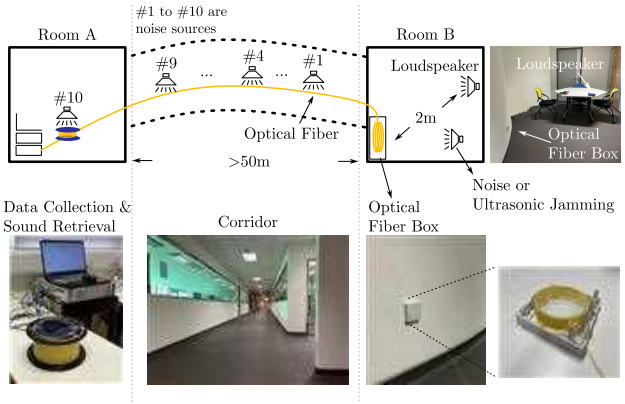


Fig. 12: The optical fiber with yellow coating is wound in an optical fiber box in a meeting room (Room B), and in the other room (Room A), the sound signal is reconstructed.

TABLE II: WER of Speech Recognition in Office Scenario.

WER↓	Min	Max	Mean	Std Dev
Wall	0.13	0.31	0.22	0.07
Ground	0.11	0.29	0.20	0.08
Desk	0.04	0.15	0.09	0.05

some voice recording from the victim so as to finetune their models, it is more effective to recognize the contents at a distance.

Impacts of Denoising and Speech Enhancement

We applied a state-of-the-art AI noise reduction tool, ensemble-enhance [52], to perform denoising and speech enhancement on our audio data. The spectrograms of before and after noise removal are shown in Figure 17 in Appendix D. Speech recognition results, as shown in Figure 11d, indicate that denoising has minimal impact on performance. This is likely because these models are trained to be inherently robust to noise, effectively integrating denoising and enhancement within their architecture. For an attacker, this is advantageous, as it eliminates the need for additional signal processing to achieve effective information extraction using the latest speech recognition models in our proposed system.

VIII. Case Study: Evaluation in Office Scenario

To further evaluate the effectiveness of our optical fiber-based eavesdropping approach in real-world scenarios, we deployed standard telecommunication optical fiber across two office rooms separated by more than $50\,\mathrm{m}$ as shown in Figure 12.

In Room B, around $3\,\mathrm{m}$ of optical fiber is coiled around a sensory receptor (diameter: $65\,\mathrm{mm}$, height: $25\,\mathrm{mm}$, made of PET material) and housed within a typical fiber optic box, commonly used in FTTH installations for excess fiber storage, as mentioned in Section III-A. Note that the optical fiber wound on the sensory receptor is much shorter (than $15\,\mathrm{m}$ that is used for performance measurement) because the receptor itself is smaller, allowing it to be concealed within the optical fiber box. As shown in Section VI-A3, a shorter fiber degrades the quality of the recovered sound, and it is therefore expected and reasonable that the recovered sound quality here may not be as good as in the earlier measurements. This optical fiber box is affixed to the base of the wall. A loudspeaker is placed on a table near the center of Room B, approximately $2\,\mathrm{m}$ away from the box. The sound volume is set at $80\,\mathrm{dB}$, playing the human speech as that used in Section VII-D.

The two rooms are connected via a corridor, through which the optical fiber is routed. In Room A, at the other end of the optical fiber, we conduct the data collection and speech retrieval.

A. Impacts of Optical Fiber Box Placements

We consider three different placements. One is that there is no direct contact between the loudspeaker and the fiber box to avoid direct mechanical coupling. We refer to this arrangement as the "Wall" configuration. For comparison, we also place both the loudspeaker and the optical fiber box directly on the floor, maintaining the same 2-meter separation. This is referred to as the "Ground" configuration. Additionally, we test a "Desk" configuration, where the loudspeaker remains on the table while the optical fiber box is attached to the underside of the table. During the experiments, the corridor experiences regular foot traffic and is adjacent to an active construction site, introducing considerable ambient noise and contributing to the non-trivial acoustic environment.

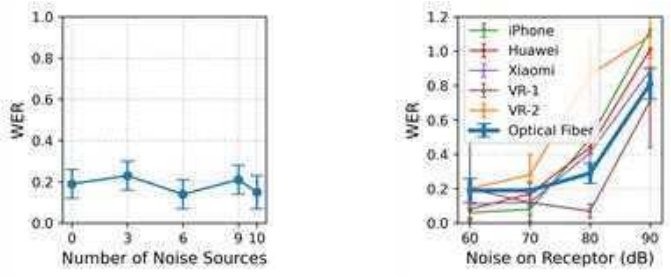
The speech recognition results are shown in Table II. Among them, the "Desk" configuration achieved the best performance, with a mean WER of 0.09, and the lowest minimum and maximum WERs (0.04 and 0.15, respectively). In contrast, the "Wall" and "Ground" setups exhibited higher WERs. The "Desk" setup also had the lowest standard deviation (0.05), indicating more consistent recognition results compared to the other two configurations. The key reason for the Desk configuration's superior performance lies in the way acoustic signals are transmitted. When the optical fiber box is attached under the desk, vibrations from the loudspeaker are efficiently transmitted through the desk surface, which can conduct vibrations more directly and with less attenuation than air or other structures, resulting in a stronger and clearer acoustic signal being coupled to the fiber. This efficient transmission leads to a higher signal-to-noise ratio and, consequently, better speech recognition accuracy.

These results indicate that, in the office scenario, the averaged WER is approximately 0.17, suggesting that around 80% of the original speech information can be successfully preserved and recognized.

B. Impacts of Noise and Ultrasound Jamming

Keeping the "Wall" configuration, we conducted additional experiments using controlled acoustic noise sources, including loudspeakers and a commercial off-the-shelf ultrasonic jammer. We examined the following three experimental conditions.

1) Noise along the Optical Fiber as Conduit: Nine noise sources (generated by loudspeakers) were placed at roughly equal intervals along the optical fiber routed through the



(a) (b)

Fig. 13: (a) WER at different numbers of noise sources along the optical fiber as the conduit. (b) WER at different noise levels for the optical-fiber-based method and the microphones.

Fig. 14: WERs of sound captured by microphones and the optical fiber under ultrasonic jamming.

corridor, as illustrated in Figure 12. They are labeled from #1 to #9 from Room B to Room A. The optical fiber was taped to the loudspeaker diaphragms. A tenth loudspeaker (#10) was placed on the blue optical fiber spool (in Room A), holding the excess optical fiber. All the loudspeakers play white noise, and their volumes are set at their maximum ($> 90\,\mathrm{dB}$).

The experiments were conducted when no pedestrians were present. We activated the noise sources in groups: first (#1, #4, #7), then (#2, #5, #8), and finally (#3, #6, #9), increasing the number of active sources from 3 to 9 in steps of 3. To ensure that the noise sources indeed affect the optical fiber conduit, we used DAS to verify that they increased the noise power by approximately $33\,\mathrm{dB}$ at each point, relative to the condition without any noise sources activated. As shown in Figure 13a, under the impacts of controlled noise sources, the average WER remained approximately 0.19. Adding the tenth source (#10) on the spool yielded a WER of 0.15 ± 0.08. We tested statistical significance with the Mann–Whitney U test (a t-test is not chosen because normality may not hold), using a 99% confidence level. The null hypothesis was that "the WER under n controlled noise sources ($n = 3, 6, 9, 10$) on the conduit is the same as the WER with no controlled noise ($n = 0$)". The p-values were 0.46, 0.28, 0.57, and 0.46, all above 0.01. Thus, we do not have sufficient evidence to reject the null hypothesis. This indicates that external noise along the optical fiber conduit has only a small impact on the WER, thereby supporting our theoretical explanation in Section V-C.

2) Noise around the Sensory Receptor: A loudspeaker was placed near the sensory receptor to generate white noise at 70, 80, and $90\,\mathrm{dB}$, with the sound pressure level measured at the sensory receptor instead of the loudspeaker. Note that the ambient noise is around $60\,\mathrm{dB}$ as mentioned previously. For comparison, we also evaluated five microphone-equipped devices: an iPhone 13, a Huawei Mate 30, a Xiaomi Note 9 Pro, and two mini voice recorders (denoted as VR-1 and VR-2). The microphones are placed where the sensory receptor is, and tested one by one.

As shown in Figure 13b, increasing environmental noise near the sensory receptor increases WER. At $70\,\mathrm{dB}$, WER remains at 0.19. At $90\,\mathrm{dB}$, WER rises up to 0.93, indicating severe information loss. Microphones exhibit a similar trend, except for the VR-1, which outperforms the optical-fiber-based method under noisy conditions. When the noise level is $70\,\mathrm{dB}$, the smartphones' microphones achieve an averaged WER below 0.1, which is about half of the optical-fiber-based method. However, when the noise is increased to $80\,\mathrm{dB}$, the optical-fiber-based method exhibits a WER of 0.3, while most of the microphones rise above 0.4. When the noise level reaches $90\,\mathrm{dB}$, WER values for microphones greater than 1 arise because the ASR system inserts words that are not present in the original speech, whereas the WER of the optical-fiber-based method remains below 1. This suggests that, in general, microphones perform better in low-noise environments, but the optical fiber approach has a slight advantage in high-noise conditions.

3) Ultrasonic Jamming: In addition to TSCM sweeps, defenders can deploy ultrasonic jammers to disrupt microphones. The jammers are effective because most commercial microphones and their front-end amplifiers exhibit nonlinearity. Strong ultrasonic inputs can intermodulate into the audible band via these nonlinearities, producing in-band interference that obscures speech [53], as shown in Figure 18 of Appendix F. We are interested in whether such jamming affects our optical-fiber-based method. To evaluate this, we used a commercially available ultrasonic jammer and placed it $1\,\mathrm{m}$ from both the sensory receptor and the microphones. Figure 14 shows that microphones are highly sensitive to ultrasonic jamming: their WERs reach or exceed 1 under jamming. In contrast, the optical-fiber-based system shows no significant change in WER, even when the ultrasonic jammer is placed as close as $10\,\mathrm{cm}$ away from the sensory receptor in our experiments. This implies that our approach of eavesdropping can effectively evade active ultrasonic jamming.

Under jamming, a determined microphone-based attacker might attempt to remove the jamming-induced noise by any available means. Whether the noise can be effectively removed to recover intelligible audio remains an open question. Some studies report no improvement in WER after applying noise-reduction techniques such as deep neural networks or Wiener filtering [54], [55], whereas more recent work shows that noise removal still leaves average WER above 0.29 [53] or 0.50 [56]. On the other hand, our optical-fiber-based method, without electrical stages at the sensory receptor, is inherently much less susceptible to ultrasonic jamming. The results here suggest that the optical-fiber-based method remains viable even when the

potential victims of eavesdropping are cautious and equipped with commercial off-the-shelf jammers.

IX. Potential Mitigation Methods

Mitigation strategies can focus on controlling both laser light reflections and cable installations to reduce the risks of eavesdropping through optical fibers. One approach is to minimize the detectability of Rayleigh backscattering, which the eavesdropping relies on, by increasing Fresnel reflections that can saturate the photodetector (see Figure 1) and create a "dead zone" where DAS cannot detect anything [57], [58]. Using polished connectors [59] can be an effective means to introduce significant Fresnel reflections. Additionally, for systems using separate fibers for transmitting and receiving, users can install optical isolators [1] on each channel, allowing light to travel in only one direction, preventing scattered light from returning to potential attackers. In terms of weakening the sensitivity of optical fiber to vibrations, several installation practices are recommended. Users should ensure that fiber cables are installed in a way that avoids excess length within rooms or keeps them from looping around or touching objects, which can unintentionally amplify vibrations. Adding soundproofing materials to walls and ceilings, especially in areas where fiber optic cables run, can help block external sounds from reaching the cables.

X. Related Work

We review state-of-the-art acoustic eavesdropping via optical, motion, and radio-frequency (RF) side channels, with discussions on practical constraints such as distance, line of sight, and recognition quality. Because prior work reports the recognition quality using different metrics, it is difficult to make direct comparisons; we will present comparisons in terms of intelligibility [60] (ranging between 0 and 1; the higher, the better), accuracy, and WER.

Optical. Early work used high-speed cameras to capture minute object vibrations induced by sound and reconstruct audio from the video signal, up to about $2\,\mathrm{m}$ away and, for short utterances, near-perfect transcription (WER = 0) [61]. Subsequent methods improved efficiency by up to around 100 times, with a trade-off in the recovery quality [62], [63]. More work further showed that commodity cameras can recover intelligible speech (intelligibility > 0.8) [64], [65], [66], [67]. Electro-optical sensing extends range to roughly $35\,\mathrm{m}$ but with reduced intelligibility (< 0.5) [68], [69], [70]. Another approach illuminates a target (or a nearby proxy) using laser beams and demodulates the reflected beam to audio [71], [72], [73], [74]. Some lab studies reported WER = 0 at $10\,\mathrm{m}$ [73], and commercial devices even claim up to $500\,\mathrm{m}$ [75]. A common limitation across optical methods is the need for a line of sight between the attacker and the target surface, which is difficult to achieve under our threat model.

Motion. Micro-Electro-Mechanical Systems (MEMS) sensors (e.g., accelerometers and gyroscopes) can leak speech through vibration coupling, allowing eavesdropping with accuracy often below 0.8 in early work [76], [77], later improved to accuracy > 0.8 [78], [79] or WER < 0.1 [80] with stronger coupling assumptions and advanced models. However, Anand et al., [81] pointed out that unless a strong loudspeaker shares the same surface with the sensors (creating a strong mechanical path), inference is not practical in typical real-world conversation scenarios. Additionally, actuators themselves can act as unintended sensors: vibration motors [82], read/write heads of hard disks [83], and camera stabilizers [84] have been exploited, reaching speech-recognition accuracy up to 0.88 [82]. These attacks typically require close proximity between the sound source and sensor, which is suitable if the attacker can run an application on the victim device.

RF. Wi-Fi signals, for example, have been utilized to profile mouth movements [85] and detect loudspeakers' vibrations [86], where the distance between the sound source and the signal transmitter is around $2\,\mathrm{m}$ and the accuracy is above 0.8. More recent work has focused on millimeter-wave (mmWave) radar [87], [88], [89], [90], [91], [92], [93], [94], Radio Frequency Identification (RFID) [95], or by collecting RF emanations from a microphone [96]. Across these studies, the average source-to-attacker distance is about $3.6\,\mathrm{m}$, with some systems reaching up to $8\,\mathrm{m}$ [88], [92]. The accuracy of the speech recognition can be as high as 0.94 at $1\,\mathrm{m}$ [95] or WER = 0.06 at $2\,\mathrm{m}$ [96]. Because RF can penetrate common building materials, many setups operate through walls; however, they still require a direct propagation path, which is often achieved by steering a directional antenna or beam toward the source, even through a wall.

Comparison and Limitations Our method can recover speech at a source-to-receptor distance of $2\,\mathrm{m}$ with WER around 0.2. This performance is not as good as the best prior results of the state-of-the-art side-channel methods, but it can be regarded as a trade-off of attack distance, line-of-sight requirements, and stealthiness. Our attack leverages telecom optical fibers as a passive, low-profile conduit: the attacker can be $50\,\mathrm{m}$ away, requires no line of sight to the victim environment, and resists commercial ultrasonic jammers, as demonstrated in our case study.

Regarding the limitations of our method, the acoustic sensitivity of optical fibers depends on the design of sensory receptors, while other sensors, like microphones, are commercially available and readily capable of capturing airborne sound with high fidelity. Another limitation is the reliance on expensive and specialized equipment, i.e., DAS systems (ranging from several thousand to tens of thousands USD), creating a higher barrier to entry compared to cheaper microphones and other sensors. Moreover, fiber-optic cables are fragile and prone to accidental damage. If a cable is broken, the eavesdropping will immediately cease to operate, and restoring the attack would require physical re-installation.

XI. Conclusion

The work presents a study on how standard telecommunication optical fibers can be exploited for acoustic eavesdropping. We not only discuss the theoretical basis for detecting sound-induced deformations in optical fibers, but also propose a phys-

ical "sensory receptor" structure to amplify airborne sounds for more effective eavesdropping. Our experimental results show that one can recover detailed information, from human activities and indoor localization to recognizable speech. Our work also emphasizes the need for continuous security assessments of widely deployed technologies like FTTH and motivates the development of stronger protections to counteract emerging side-channel threats.

XII. Ethical Consideration

In this section, we present a discussion of the ethical considerations for this work, structured around the four key principles outlined in the Menlo Report [97].

Respect for Persons: To measure the sound intensity of human speech, we recruited 11 volunteers who orally consented to the use of their sound intensity data for research purposes. Only the intensity of their everyday speaking voices is measured by a sound level meter, and no sensitive information is recorded. Additionally, the sound sources, including human activities and speech, are derived from publicly available datasets that have been carefully curated and reviewed to ensure compliance with ethical standards and privacy considerations. All experiments are conducted within a controlled laboratory environment, ensuring that no individuals are directly impacted by the research. To safeguard the well-being of our researchers during the experiments, strict safety protocols are followed. Essential protective equipment, including earmuffs with a noise reduction function, is provided to mitigate potential hazards.

Beneficence: Regarding the *benefits*, this work provides valuable insights and potential mitigation methods that can lead to improved privacy measures and better protection against eavesdropping attacks through optical fibers. By quantifying range, WER, and environmental constraints, we reduce uncertainty, promote defensive investments, and help prevent misuse. Revealing the feasibility and limits of optical-fiber acoustic eavesdropping enables stakeholders (e.g., ISPs, enterprises, standards bodies) to assess risk and adopt mitigations (e.g., installation practices, optical isolation, managed reflections, policy, and audit controls), which can in turn improve the trust in optical communication infrastructures and thus benefit society as a whole.

Regarding the *risks* or adverse effects of this work, the technique could be misused by personnel with physical access to fiber endpoints (e.g., rogue insiders) to covertly extract speech or activities. We therefore omit implementation details that would materially lower the barrier to weaponization (e.g., exact DAS devices). Although telecom operator practices are outside users' control, we also provide home-applicable mitigations, such as minimizing in-room fiber slack and preventing contact with resonant structures, which users can readily adopt and effectively mitigate the risk.

Justice: The research is conducted in an equitable and unbiased manner, including the selection of the volunteers, and focuses solely on the technical properties of optical fiber systems. It does not target or involve any specific groups, organizations, or individuals. By employing a controlled experimental setup, the study ensures that no one is unfairly exposed to risks or burdens resulting from the research. The design of the project prioritizes fairness and seeks to distribute the benefits of the findings broadly, without favoring or disadvantaging any particular group.

Respect for Law and Public Interest: This study is carried out in full compliance with all relevant laws and regulations governing scientific research, data collection, and cybersecurity. No unauthorized access to communication networks or private data is involved at any stage of the research. No unethical/illegal voice recorders were purchased. The use of a controlled experimental environment ensures that the study does not infringe upon privacy laws, wiretapping regulations, or any other legal protections related to communication systems.

Acknowledgment

This work was supported in part by a grant from HKPolyU (No. 1-ZVG0), as well as grants from the CUHK IE department (project code: GRF/23/SYC and GRF/24/SYC).

References

[1] J. M. Senior and M. Y. Jamro, *Optical Fiber Communications: Principles and Practice*. Pearson Education, 2009.

[2] W. Van Eck, "Electromagnetic Radiation from Video Display Units: An Eavesdropping Risk?" *Computers & Security*, vol. 4, no. 4, pp. 269–286, 1985.

[3] Y. Long, Q. Jiang, C. Yan, T. Alam, X. Ji, W. Xu, and K. Fu, "EM Eye: Characterizing Electromagnetic Side-channel Eavesdropping on Embedded Cameras," *Proceedings of ACM NDSS*, 2024.

[4] K.-H. Gonschorek and R. Vick, *Electromagnetic Compatibility for Device Design and System Integration*. Springer Science & Business Media, 2009.

[5] I. Giechaskiel, K. B. Rasmussen, and K. Eguro, "Leaky Wires: Information Leakage and Covert Communication between FPGA Long Wires," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 15–27.

[6] M. P. Fok, Z. Wang, Y. Deng, and P. R. Prucnal, "Optical Layer Security in Fiber-optic Networks," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 725–736, 2011.

[7] B. Wu, B. J. Shastri, and P. R. Prucnal, "Secure Communication in Fiber-optic Networks," in *Emerging trends in ICT security*. Elsevier, 2014, pp. 173–183.

[8] Z. Fang, K. Chin, R. Qu, and H. Cai, *Fundamentals of Optical Fiber Sensors*. John Wiley & Sons, 2012.

[9] T. Koonen, "Fiber to the Home/Fiber to the Premises: What, Where, and When?" *Proceedings of the IEEE*, vol. 94, no. 5, pp. 911–934, 2006.

[10] R. Montagne and D. Dichiarante, "FTTH/B Global Ranking," FTTH Council Europe, 2024.

[11] ITU Telecommunication Standardization Sector, "G. 983.1– "Broadband Optical Access Systems Based on Passive Optical Networks (PON)"," https://www.itu.int/rec/T-REC-G.983.1/en, 2005, accessed: 2025-07-11.

[12] ——, "ITU-T Recommendation G.984: Gigabit-capable Passive Optical Networks (GPON): General characteristics," https://www.itu.int/rec/T-REC-G.984.1, 2003, Accessed: 2025-07-11.

[13] 802.3 WG - Ethernet Working Group, "IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications—Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks," https://standards.ieee.org/ieee/802.3ah/3179/, IEEE, 2004, accessed: 2025-07-11.

[14] N. J. Lindsey, S. Yuan, A. Lellouch, L. Gualtieri, T. Lecocq, and B. Biondi, "City-scale Dark Fiber DAS Measurements of Infrastructure Use during the COVID-19 Pandemic," *Geophysical Research Letters*, vol. 47, no. 16, p. e2020GL089931, 2020.

[15] E. F. Williams, M. R. Fernández-Ruiz, R. Magalhaes, R. Vanthillo, Z. Zhan, M. González-Herráez, and H. F. Martins, "Distributed sensing of microseisms and teleseisms with submarine dark fibers," *Nature Communications*, vol. 10, no. 1, p. 5778, 2019.

[16] V. Grishachev, "Detecting Threats of Acoustic Information Leakage through Fiber Optic Communications," *Journal of Information Security*, vol. 3, no. 2, pp. 149–155, 2012.

[17] ——, "Threat Model of the Speech Information Confidentiality in Modern Office Based on Convergence of Functions of Optical Networks," *Photonics Russia*, pp. 90–103, 2017.

[18] V. Grishachev, Y. Kalinina, and O. Kazarin, "Fiber-Optic Channel of Voice Information Leakage," in *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2019, pp. 1512–1514.

[19] H. Hao, Z. Pang, G. Wang, and B. Wang, "Indoor Optical Fiber Eavesdropping Approach and Its Avoidance," *Optics Express*, vol. 30, no. 20, pp. 36 774–36 782, 2022.

[20] K. O. Hill and G. Meltz, "Fiber Bragg Grating Technology Fundamentals and Overview," *Journal of lightwave technology*, vol. 15, no. 8, pp. 1263–1276, 1997.

[21] H. Liu, D. J. J. Hu, Q. Sun, L. Wei, K. Li, C. Liao, B. Li, C. Zhao, X. Dong, Y. Tang *et al.*, "Specialty Optical Fibers for Advanced Sensing Applications," *Opto-Electronic Science*, vol. 2, no. 2, pp. 220 025–1, 2023.

[22] TeleGeograghy, "Submarine Cable Map," https://www.submarinecablemap.com/, accessed: 2025-01-22.

[23] F. Walter, D. Gräff, F. Lindner, P. Paitz, M. Köpfli, M. Chmiel, and A. Fichtner, "Distributed Acoustic Sensing of Microseismic Sources and Wave Propagation in Glaciated Terrain," *Nature communications*, vol. 11, no. 1, p. 2436, 2020.

[24] L. Bouffaut, K. Taweesintananon, H. J. Kriesell, R. A. Rørstadbotnen, J. R. Potter, M. Landrø, S. E. Johansen, J. K. Brenne, A. Haukanes, O. Schjelderup, and F. Storvik, "Eavesdropping at the Speed of Light: Distributed Acoustic Sensing of Baleen Whales in the Arctic," *Frontiers in Marine Science*, vol. 9, 2022.

[25] C. Wiesmeyr, C. Coronel, M. Litzenberger, H. J. Döller, H.-B. Schweiger, and G. Calbris, "Distributed Acoustic Sensing for Vehicle Speed and Traffic Flow Estimation," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2596–2601.

[26] J. Hayes, "FTTH - Customer Premises Installation," https://www.thefoa.org/tech/ref/appln/FTTH-prem.html, The Fiber Optic Association, Inc., 2021, accessed: 2025-07-11.

[27] Deloitte, "Tapping of Fibre Networks," https://zybersafe.com/wordpress/wp-content/uploads/2017/04/Deloitte_Fiber_tapping_Q1_2017_English.pdf, Deloitte Touche Tohmatsu Limited, 2017, accessed: 2025-07-11.

[28] R. Zhou, X. Ji, C. Yan, Y.-C. Chen, W. Xu, and C. Li, "DeHiREC: Detecting Hidden Voice Recorders via ADC Electromagnetic Radiation," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 3113–3128.

[29] U.S. Senate Committee on Commerce, Sicence, & Transportation, "Smart Devices, Appliances with Hidden Microphones, Cameras Must Be Disclosed to Consumers," https://www.commerce.senate.gov/2023/7/smart-devices-appliances-with-hidden-microphones-cameras-must-be-disclosed-to-consumers, accessed: 2025-01-22.

[30] M. Florentine, *Loudness*. Springer, 2010, p. 215.

[31] P. Gramming, "Vocal Loudness and Frequency Capabilities of the Voice," *Journal of Voice*, vol. 5, no. 2, pp. 144–157, 1991.

[32] M. J. Murray, A. Davis, and B. Redding, "Fiber-wrapped Mandrel Microphone for Low-noise Acoustic Measurements," *Journal of Lightwave Technology*, vol. 36, no. 16, pp. 3205–3210, 2018.

[33] H. Zheng, H. Wu, C. Y. Leong, Y. Wang, X. Shen, Z. Fang, X. Cheng, J. Cui, D. Ma, Y. Miao, L. Zhou, M. Yan, J. Sun, H.-Y. Tam, X. Ding, and C. Lu, "Enhanced Quasi-Distributed Accelerometer Array Based on ϕ-OTDR and Ultraweak Fiber Bragg Grating," *IEEE Sensors Journal*, vol. 23, no. 16, pp. 18 176–18 182, 2023.

[34] C. Brown and T. Riede, *Comparative Bioacoustics: An Overview*. Bentham Science Publishers, 2017, pp. 62–115.

[35] W. C. Young, R. G. Budynas, A. M. Sadegh *et al.*, *Roark's Formulas for Stress and Strain*. McGraw-hill New York, 2002, vol. 7.

[36] C. D. Butter and G. Hocker, "Fiber Optics Strain Gauge," *Applied Optics*, vol. 17, no. 18, pp. 2867–2869, 1978.

[37] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[38] H. Kwakernaak, R. Sivan, and R. C. Strijbos, *Modern Signals and Systems*. Prentice-Hall, Inc., 1991.

[39] J. R. Taylor, "An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements," 1997.

[40] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[41] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Neural Information Processing Systems*, 2017.

[42] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio Pre-Training with Acoustic Tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[43] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[45] F. Schmid, K. Koutini, and G. Widmer, "Efficient Large-scale Audio Tagging via Transformer-to-CNN Knowledge Distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[46] ——, "Dynamic Convolutional Neural Networks as Efficient Pre-trained Audio Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[47] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[48] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-scale Weak Supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[49] Nvidia, "canary-1b," https://huggingface.co/nvidia/canary-1b, accessed: 2024-11-04.

[50] ——, "parakeet-tdt-1.1b," https://huggingface.co/nvidia/parakeet-tdt-1.1b, accessed: 2024-11-04.

[51] A. C. Morris, V. Maier, and P. D. Green, "From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition." in *Interspeech*, 2004, pp. 2765–2768.

[52] Resemble.AI, "resemble-enhance," https://github.com/resemble-ai/resemble-enhance, accessed: 2024-11-04.

[53] M. Gao, Y. Chen, Y. Li, L. Zhang, J. Liu, L. Lu, F. Lin, J. Han, and K. Ren, "A Resilience Evaluation Framework on Ultrasonic Microphone Jammers," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1914–1929, 2023.

[54] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, "Wearable Microphone Jamming," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[55] Y. Chen, M. Gao, Y. Liu, J. Liu, X. Xu, L. Cheng, and J. Han, "Implement of a Secure Selective Ultrasonic Microphone Jammer," *CCF Transactions on Pervasive Computing and Interaction*, vol. 3, no. 4, pp. 367–377, 2021.

[56] Z. Yu, L. Tang, K. Wang, X. Tang, and H. Ge, "Dynamic Ultrasonic Jamming via Time–Frequency Mosaic for Anti-Eavesdropping Systems," *Electronics*, vol. 14, no. 15, p. 2960, 2025.

[57] F. Liu, T. Gu, and Z. Huang, "Dead Zone Fault Detection Optimization Method for Few-Mode Fiber Links Based on Unexcited Coupled Higher-Order Modes," in *Photonics*, vol. 11, no. 5. MDPI, 2024, p. 433.

[58] X. Bao and Y. Wang, "Recent Advancements in Rayleigh Scattering-Based Distributed Fiber Sensors," *Advanced Devices & Instrumentation*, vol. 2021, 2021.

[59] R. Biswas, "All Fiber Optic Sensor with Reference to Different Reflectors," *Results in Optics*, vol. 11, p. 100392, 2023.

[60] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[61] 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.*, 2014.

[62] D. Zhang, J. Guo, X. Lei, and C. Zhu, "Note: Sound Recovery from Video using SVD-based Information Extraction," *Review of Scientific Instruments*, vol. 87, no. 8, 2016.

[63] D. Zhang, J. Guo, Y. Jin, and C. Zhu, "Efficient Subtle Motion Detection from High-speed Video for Sound Recovery and Vibration Analysis using Singular Value Decomposition-based Approach," *Optical Engineering*, vol. 56, no. 9, pp. 094 105–094 105, 2017.

[64] G. Zhu, X.-R. Yao, Z.-B. Sun, P. Qiu, C. Wang, G.-J. Zhai, and Q. Zhao, "A High-speed Imaging Method Based on Compressive Sensing for Sound Extraction using A Low-speed Camera," *Sensors*, vol. 18, no. 5, p. 1524, 2018.

[65] L. Franklin and D. Huber, "Exploiting Camera Rolling Shutter to Detect High Frequency Signals," in *Applications of Digital Image Processing XLII*, vol. 11137. SPIE, 2019, pp. 85–94.

[66] H. Shindo, K. Terano, K. Iwai, T. Fukumori, and T. Nishiura, "Noise-reducing Sound Capture Based on Exposure-time of Still Camera," in *Proceedings of the 23rd International Congress on Acoustics*, 2019, pp. 2893–2899.

[67] A. Yoshida, H. Shindo, K. Terano, T. Fukumori, and T. Nishiura, "Interpolation of Acoustic Signals in Sound Capture with Rolling-shuttered Visual Camera," in *Forum Acusticum*, 2020, pp. 39–45.

[68] B. Nassi, R. Swissa, J. Shams, B. Zadov, and Y. Elovici, "The Little Seal Bug: Optical Sound Recovery from Lightweight Reflective Objects," in *2023 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2023, pp. 298–310.

[69] 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.

[70] B. Nassi, Y. Pirutin, R. Swisa, A. Shamir, Y. Elovici, and B. Zadov, "Lamphone: Passive Sound Recovery from a Desk Lamp's Light Bulb Vibrations," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4401–4417.

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

[72] 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.

[73] K. Doi, K. Ooi, and T. Sugawara, "Acoustic Eavesdropping Attack Using Self-Mixing Laser Interferometer," in *Proceedings of the 2024 Workshop on Attacks and Solutions in Hardware Security*, 2024, pp. 23–35.

[74] P. Walker and N. Saxena, "Laser Meager Listener: A Scientific Exploration of Laser-based Speech Eavesdropping in Commercial User Space," in *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022, pp. 537–554.

[75] Argo-A Security LLC. (2025) Long-range laser listening device. Accessed: 19 Nov 2025. [Online]. Available: https://argoasecurity.com/index.php?route=product/product&path=8&product_id=263

[76] 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.

[77] J. Han, A. J. Chung, and P. Tague, "PitchIn: Eavesdropping via Intelligible Speech Reconstruction using Non-Acoustic Sensor Fusion," in *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2017, pp. 181–192.

[78] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, "Spearphone: A Lightweight Speech Privacy Exploit via Accelerometer-Sensed Reverberations from Smartphone Loudspeakers," in *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, pp. 288–299.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] Y. Long, P. Naghavi, B. Kojusner, K. Butler, S. Rampazzi, and K. Fu, "Side Eye: Characterizing the Limits of POV Acoustic Eavesdropping from Smartphone Cameras with Rolling Shutters and Movable Lenses," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1857–1874.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] P. Hu, W. Li, R. Spolaor, and X. Cheng, "mmEcho: AmmWave-based Acoustic Eavesdropping Method," in *Proceedings of the ACM Turing Award Celebration Conference-China 2023*, 2023, pp. 138–140.

[91] 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.

[92] 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 mmWaveDevice," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 338–351.

[93] C. Wang, F. Lin, H. Yan, T. Wu, W. Xu, and K. Ren, "VibSpeech: Exploring Practical Wideband Eavesdropping via Bandlimited Signal of Vibration-based Side Channel," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 3997–4014.

[94] R. Zhao, L. J.-T. Yu, T. Li, Z. Jiang, C. Zhang, C. Wu, H. Zhao, and E. C. Ngai, "SPACE: Speaker Adaptation for Acoustic Eavesdropping using mmWave Radio Signals," *IEEE Transactions on Mobile Computing*, 2025.

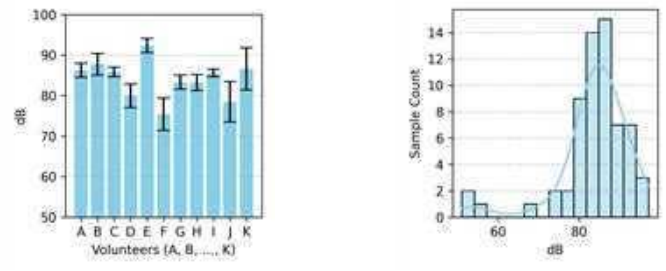
[95] C. Wang, L. Xie, Y. Lin, W. Wang, Y. Chen, Y. Bu, K. Zhang, and S. Lu, "Thru-the-wall Eavesdropping on Loudspeakers via RFID by Capturing Sub-mm Level Vibration," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–25, 2021.

[96] A. Onishi, S. H. Bhupathiraju, R. Bhatt, S. Rampazzi, and T. Sugawara, "Sound of Interference: Electromagnetic Eavesdropping Attack on Digital Microphones Using Pulse Density Modulation," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025.

[97] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The Menlo Report," *IEEE Security & Privacy*, vol. 10, no. 2, pp. 71–75, 2012.

[98] B. D. Steinberg, "Principles of Aperture and Array System Design: Including Random and Adaptive Arrays," *New York*, 1976.

[99] L. J. Ziomek, "Three Necessary Conditions for the Validity of the Fresnel Phase Approximation for the Near-field Beam Pattern of An



(a) (b)

Fig. 15: Sound intensity of human speech: (a)mean and standard deviation of the sound level of different volunteers, (b) a distribution of all recorded samples.

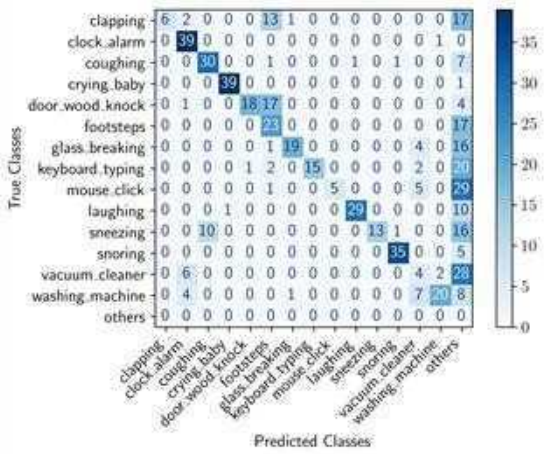


Fig. 16: Confusion matrix of the sound event detection results.

Aperture," *IEEE Journal of Oceanic Engineering*, vol. 18, no. 1, pp. 73–75, 1993.

[100] J. G. Ryan and R. A. Goubran, "Array Optimization Applied in the Near Field of A Microphone Array," *IEEE Transactions on speech and Audio Processing*, vol. 8, no. 2, pp. 173–176, 2000.

[101] Y. R. Zheng, R. A. Goubran, M. El-Tanany, and H. Shi, "A Microphone Array System for Multimedia Applications with Near-field Signal Targets," *IEEE Sensors Journal*, vol. 5, no. 6, pp. 1395–1406, 2005.

[102] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

APPENDIX

A. Sound Level of Human Speech

To measure the sound intensity of human speech, we recruited 11 volunteers who orally consented to the use of their sound intensity data for research purposes. We then recorded the intensity of their everyday speaking voices. A sound level meter was used for these measurements, with a sampling frequency of 1Hz. Each volunteer was asked to count from 1 to 10 at their normal speaking volume. During the recording, the microphone of the sound level meter was positioned 1–2 cm from the volunteer's mouth to ensure accurate measurement of the speech sound pressure level.

For each volunteer, we recorded the sound intensity over a 5-second interval. The statistical analysis of the collected data shows that the average intensity was $83.8\,\mathrm{dB}$. The maximum recorded value was $97.2\,\mathrm{dB}$, and the minimum value was $51.1\,\mathrm{dB}$. We further visualize the sound intensity of all volunteers in Figure 15a, and a distribution of all sound intensity samples in Figure 15b. It can be observed that the most sound intensities fall within the range above $80\,\mathrm{dB}$. These results provide a representative range and average for normal human speech intensity.

B. Confusion Matrix of Sound Event Detection

The confusion matrix shown in Figure 16 provides a detailed view of the classification results. The x-axis represents the predicted classes, while the y-axis represents the true classes. Each class contains 40 samples, so the total count in each row sums up to 40. Notably, some output labels do not fit into the selected 14 categories, leading to the inclusion of an "others" category to capture instances that were not classified into any of the predefined classes. The model performs well in certain categories, achieving high counts along the diagonal, which indicates correct predictions. However, there are several off-diagonal values that reflect misclassifications. For example, the "keyboard_typing", "mouse_click", and "vacuum_cleaner" classes show some level of confusion with other classes, suggesting that the model might have difficulty distinguishing between similar sound features in these categories.

C. Localization Methods

When sound signals are located within the radial distance of $\frac{2\times d^2}{\lambda}$, where λ is the wavelength of the sound signal, and d is the distance between sound source and the receptor, this is a near-field scenario, and a spherical sound wave needs to be considered [98], [99], [100], [101]. Our setup is a near-field scenario.

The distance between the receptors is denoted as d_R, and the distances from the sound source S, at (x_S, y_S) to points A, B, and C are d_A, d_B, and d_C, respectively. Let u represent the speed of sound in air. The time difference for the sound to travel between point A to point B, and between point C and B can be expressed as:

$$\tau_{AB} = \frac{d_A - d_B}{u},\ \tau_{CB} = \frac{d_C - d_B}{u}.$$

Let the angle between the line connecting the point of the sound source to B and the x-axis be denoted as θ_B. Using the law of cosines, we can derive the following equations:

$$d_A^2 = d_R^2 + d_B^2 - 2 \times d_R \times d_B \times \cos(\theta_B)$$
$$d_C^2 = d_R^2 + d_B^2 - 2 \times d_R \times d_B \times \cos(\pi - \theta_B).$$

Given τ_{AB}, τ_{CB}, d, and u, solving these four equations above allows us to determine d_A, d_B, d_C, and θ_B. It's important to note that we only retain the solution where θ_B is less than π. Knowing the coordinates of point B as (x_B, y_B), the estimated position of the sound source can then be determined as:

$$(\overline{x}_S, \overline{y}_S) = (x_B - d_B \times \cos(\theta_B), d_B \times \sin(\theta_B))\,.$$

Thus, accurately estimating τ_{AB} and τ_{CB} from the signals measured at points A, B, and C is sufficient for determining

TABLE III: Estimation of Time Difference of Arrivals

Sound Loc.	Ground Truth τ_{AB} (ms)	Ground Truth τ_{CB} (ms)	Estimation $\overline{\tau}_{AB}$ (ms)	Estimation $\overline{\tau}_{CB}$ (ms)
0	-7.36	7.64	-7.9	7.7
1	0.00	7.64	-1.1	0.3
2	7.36	7.36	6.5	7.2
3	7.64	0.00	7.3	0.2
4	7.64	-7.36	7.5	-6.5
5	-4.73	7.22	-5.5	5.3
6	0.00	6.68	-0.4	7
7	4.73	4.73	4.3	4.8
8	6.68	0.00	5.9	0.3
9	7.22	-4.73	7.3	-4.7
10	-3.17	6.28	-2.4	7.8
11	0.00	5.24	-0.4	5.5
12	3.17	3.17	2.7	3.2
13	5.24	0.00	4	0.1
14	6.28	-3.17	5.9	-2.7
15	-2.32	5.33	-2.3	7.6
16	0.00	4.13	0.6	2.7
17	2.32	2.32	1.7	2.9
18	4.13	0.00	3.1	0
19	5.33	-2.32	5.2	-2.6
20	-1.81	4.53	-3	5.1
21	0.00	3.35	1.7	3.6
22	1.81	1.81	2.1	3.4
23	3.35	0.00	3.1	-0.2
24	4.53	-1.81	3.7	-2.7

TABLE IV: Estimation of Sound Sources' Locations

Sound Loc.	Ground Truth (m) x_S	Ground Truth (m) y_S	Estimation (m, except θ_B in rads) $\overline{x}_S$	$\overline{y}_S$	d_A	d_B	d_C	θ_B	Δd (m)
0	0	0.1	-5.28	0.00	5.14	7.88	10.54	0.26j	NA
1	1.3	0.1	4.86	-24.00	-24.49	-24.11	-24.00	1.48	NA
2	2.6	0.1	2.48	0.20	2.49	0.23	2.73	1.02	0.16
3	3.9	0.1	3.86	0.53	3.90	1.37	1.44	2.74	0.43
4	5.2	0.1	5.04	0.11	5.04	2.44	0.19	3.10	0.16
5	0	1.3	36.46	-32.61	-48.92	-47.01	-45.17	0.77	NA
6	1.3	1.3	1.21	0.92	1.53	1.67	4.09	0.59	0.38
7	2.6	1.3	2.50	1.35	2.84	1.35	3.02	1.50	0.11
8	3.9	1.3	3.81	1.80	4.22	2.17	2.28	2.16	0.51
9	5.2	1.3	4.94	0.81	5.01	2.48	0.85	2.81	0.55
10	0	2.6	1.12	0.00	0.64	1.48	4.18	0.47j	NA
11	1.3	2.6	1.15	2.39	2.65	2.79	4.70	1.03	0.26
12	2.6	2.6	2.47	2.79	3.73	2.79	3.90	1.53	0.23
13	3.9	2.6	3.85	3.89	5.47	4.08	4.12	1.88	1.29
14	5.2	2.6	5.11	2.88	5.86	3.82	2.88	2.29	0.30
15	0	3.9	0.98	0.00	0.82	1.62	4.25	0.26j	NA
16	1.3	3.9	1.75	5.44	5.72	5.51	6.45	1.42	1.61
17	2.6	3.9	2.23	3.80	4.40	3.82	4.82	1.47	0.38
18	3.9	3.9	3.90	5.61	6.83	5.76	5.76	1.80	1.71
19	5.2	3.9	5.56	4.34	7.05	5.25	4.35	2.17	0.57
20	0	5.2	-1.05	5.26	5.36	6.40	8.17	0.96	NA
21	1.3	5.2	2.08	3.12	3.75	3.16	4.41	1.41	2.22
22	2.6	5.2	2.25	3.02	3.77	3.04	4.22	1.46	2.20
23	3.9	5.2	4.06	5.97	7.23	6.15	6.08	1.81	0.79
24	5.2	5.2	9.44	14.32	17.16	15.87	14.94	2.02	NA

the location of the sound source. We used the Generalized Cross-Correlation (GCC) method [102], together with a careful manual check, to obtain τ_{AB} and τ_{CB}, and further, estimate the ground truth on the sound source's position.

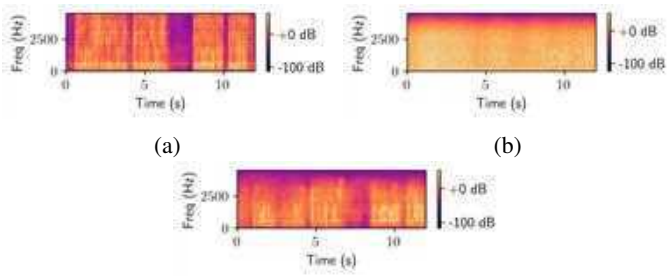
As shown in Table III, it presents the ground truth values of the time difference of arrivals, i.e., τ_{AB} and τ_{CB}, and their estimations $\overline{\tau}_{AB}$ and $\overline{\tau}_{CB}$. In Table IV, it presents the ground truth values of the coordinates of sound sources (x_S and x_S), and the estimation ($\overline{x}_S$, $\overline{y}_S$, d_A, d_B, d_C, and θ_B). The estimation error is presented in the column labeled Δd (m), where NA indicates the estimation is not a valid solution of the method.

D. Impacts of Denoising and Speech Enhancement

Figure 17 shows (a) the spectrogram of the original audio, (b) the spectrogram of the audio recovered from optical fiber vibration, and (c) the spectrogram after denoising and speech enhancement applied to (b). The processed signal appears more prominent in the spectrogram.

E. Increased Distance and Obstacles

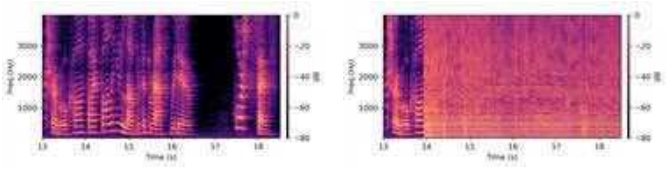
We investigated the maximum effective eavesdropping range, finding that at $8\,\mathrm{m}$, WER approaches 1, indicating



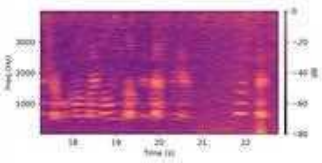
(a) (b)

(c)

Fig. 17: The spectrogram of an (a) original audio, and a recovered audio (b) without denoising, and (c) with denoising.



(a) (b)



(c)

Fig. 18: (a) Spectrogram of a female voice snippet recorded on an iPhone 13 with no jamming (b) Spectrogram of the same snippet recorded on the iPhone 13, and ultrasonic jamming began around 14.0s. (c) Spectrogram of the same snippet recorded by the optical fiber while the ultrasonic jamming is always on.

nearly complete transcription failure. This is primarily due to an increased insertion count I, resulting in high word misrecognition. We also experimented with placing receptors within a ceiling 3 meters above the ground or in partitioned areas. The ceiling or partition material was approximately 3-$5\,\mathrm{cm}$ thick and made of wood. In this setup, the WER for all models exceeded 0.9, indicating that it is very challenging to extract meaningful information from the recovered audio signal. This is due to the walls or ceiling absorbing most of the energy, causing severe attenuation of the sound signal, which leaves the recovered audio dominated by noise. It also shows that trivially hiding behind ceilings/walls is unlikely to work effectively, and how to hide the sensory receptor remains an open problem.

F. Impacts of Ultrasonic Jamming

We visualized the effect of ultrasonic jamming on the microphone signal, as shown in Figure 18. When the jamming begins at approximately $14.0\,\mathrm{s}$ (Figure 18b), the original sound becomes masked by the jamming. The optical fiber is also affected (Figure 18c), but its low-frequency components are preserved, remaining sufficiently informative for speech recognition.