A Fly on the Wall - Exploiting Acoustic Side-Channels in Differential Pressure Sensors

Yonatan Gizachew Achamyeleh[1], Mohamad Habib Fakih[1], Gabriel Garcia[1], Anomadarshi Barua[2], Mohammad Abdullah Al Faruque[1]
[1]*Dept. of Electrical Engineering and Computer Science, University of California, Irvine, CA, USA*
{*yachamye, mhfakih, gegarci1, alfaruqu*}*@uci.edu*
[2]*Dept. of Cyber Security Engineering, George Mason University, VA, USA. abarua8@gmu.edu*

***Abstract*—Differential Pressure Sensors are widely deployed to monitor critical environments. However, our research unveils a previously overlooked vulnerability: their high sensitivity to pressure variations makes them susceptible to acoustic side-channel attacks. We demonstrate that the pressure-sensing diaphragms in DPS can inadvertently capture subtle air vibrations caused by speech, which propagate through the sensor's components and affect the pressure readings. Exploiting this discovery, we introduce BaroVox, a novel attack that reconstructs speech from DPS readings, effectively turning DPS into "a fly on the wall." We model the effect of sound on DPS, exploring the limits and challenges of acoustic leakage. To overcome these challenges, we propose two solutions: a signal-processing approach using a unique spectral subtraction method and a deep learning-based approach for keyword classification. Evaluations under various conditions demonstrate BaroVox's effectiveness, achieving a word error rate of 0.29 for manual recognition and 90.51% accuracy for automatic recognition. Our findings highlight the significant privacy implications of this vulnerability. We also discuss potential defense strategies to mitigate the risks posed by BaroVox.**

***Index Terms*—Differential pressure sensors; Side-channel attacks; Privacy**

1. Introduction

Differential Pressure Sensors (DPS) have become ubiquitous in various environments, ranging from industrial facilities and cleanrooms to residential buildings, offices, hotels, and hospitals [1]–[3]. These sensors are designed to measure minute pressure differences between two points, enabling precise control and monitoring of critical systems such as HVAC, airflow management, and room pressure regulation [2], [4]. While DPS find applications in various sectors, their role is particularly critical in the semiconductor industry, where they are essential for maintaining cleanroom integrity. However, the widespread deployment of DPS has inadvertently introduced a hidden vulnerability that can be exploited for eavesdropping.

In many real-world applications, audio systems, including speakers and intercoms, are often installed in close proximity to DPS. This practice is driven by various factors, such as the need for effective communication, audio notifications, or entertainment purposes. For instance, intercoms are used in industrial cleanrooms to facilitate coordination among workers without compromising the controlled environment [2].

While the co-location of audio systems and DPS serves practical purposes, it unintentionally creates an acoustic side channel that attackers can exploit. DPS's high sensitivity to pressure variations makes them susceptible to unintended acoustic coupling. When sound waves from nearby speakers or intercoms impinge upon the DPS, they induce minute vibrations on the sensor's diaphragm, causing measurable changes in its output. This unintended interaction effectively transforms the DPS into a makeshift microphone, allowing potential attackers to eavesdrop on confidential conversations or recover sensitive audio information.

In this paper, we introduce BaroVox, a novel side-channel attack that exploits DPS's acoustic vulnerability to recover speech from their output signals. BaroVox leverages audio systems' close proximity to DPS, which is a common deployment scenario in various real-world settings. By carefully analyzing the pressure variations captured by the DPS, BaroVox enables the reconstruction of speech signals, effectively turning these ubiquitous sensors into unintended listening devices.

We argue that if an attacker manages to get the pressure reading of this DPS, they can process the pressure data and partially reconstruct speech to still confidential information. The main challenge in realizing BaroVox lies in extracting intelligible audio from the low-bandwidth, noisy signals captured by DPS, which are primarily designed for measuring pressure differences rather than recording sound. The sensor's non-linear frequency response adds another layer of complexity. BaroVox offers a unique **Pressure-Acoustic Transformation (PAT)** to mitigate these challenges, presenting two approaches for eavesdropping conversations.

The first design solution employs PAT for speech reconstruction and then focuses on enhancing the reconstructed speech's signal-to-noise ratio (SNR). This enhancement is achieved by integrating multiple digital signal processing techniques, comprising a novel spectral subtraction method, normalization, high-pass filtering, and equalization. In the spectral subtraction phase, first, our design divides the recon-

structed speech into percussive and harmonic segments using median filtering [5], [6]. Then, after evaluating the statistical property of noise in the target environment, we apply tailored spectral noise removal to both components, adjusting the degree of removal for each before their reintegration. This results in an improved SNR of the speech while keeping the integrity of other speech elements.

The second design solution leverages deep learning techniques to extract pertinent audio features, classifying words from a specific vocabulary dataset. Our Automatic Speech Recognition (ASR) model introduces **denoising autoencoder** and **equalization layers** on top of ResNet to address the challenges posed by low SNR and non-linear frequency response of the sensor. This solution can be preferred by attackers when more precision is desired, and focus is required on specific critical keywords.

We extensively evaluate the effectiveness of BaroVox through real-world experiments, considering various factors that influence the success of the attack. We evaluate the first solution's capability in Manual Speech Recognition (MSR) by engaging 18 volunteers to transcribe 20 reconstructed speeches from pressure data. The performance metrics are Mean Opinion Score (MOS) [7] and Word Error Rate (WER) [8]. Our results demonstrate BaroVox's alarming efficacy. The proposed signal processing pipeline achieves a WER of 0.29, comprehending over 70% of the discourse. The second solution's efficacy was gauged on its word classification prowess within a restricted vocabulary dataset. Our ASR model achieved an impressive 90.51% accuracy for a 35-keyword, speaker-independent classification.

These findings highlight the serious privacy implications of BaroVox. Attackers can exploit this vulnerability to eavesdrop on sensitive conversations, compromise confidential information, or invade privacy without physical access to the targeted premises. The implications are particularly severe in critical environments such as industrial facilities, corporate offices, and healthcare institutions, where the loss of sensitive data can have far-reaching consequences. Our main **contributions** are summarized as follows:

(1) We discover a previously overlooked acoustic side-channel vulnerability in deploying DPS in proximity to sound sources.

(2) We characterize the limits and challenges of speech recovery from pressure reading. We propose **BaroVox** to overcome these challenges. To the best of our knowledge, *BaroVox is the first side-channel attack on pressure sensors*.

(3) We evaluate BaroVox on MSR and ASR tasks and show an attacker can partially reconstruct a speech with 0.29 WER using MSR and 90.51% accuracy using ASR. We present sample reconstructed audio for demonstration here: BaroVox.

(4) We propose practical countermeasures and defense strategies to mitigate the risks associated with BaroVox.

2. Background

In this section, we discuss DPSs and their applications across various domains, as well as the privacy implications of deploying DPS in close proximity to audio systems.

2.1. Physics of Differential Pressure Sensors

The structure of a DPS is illustrated in Fig. 1. A DPS consists of three fundamental components: 1) pressure ports, 2) a pressure-sensing element, and 3) a transducer. Pressure ports are openings in the sensor where pressure gets applied and are connected directly to the pressure-sensing element.

The pressure-sensing element is a force collector constructed of a flexible diaphragm such as a thin semiconductor material film, a silicon membrane, or another material that responds to the measured pressure. *In our attack, we show that the thin and flexible nature of the diaphragm enables the sensing element to respond to tiny vibrations created by sound waves. We then utilize these vibrations collected by the sensing element to reconstruct speech.*

Various physical mechanisms, such as thermal mass-flow, capacitive sensing, or piezoresistive sensing, could form the basis of the transducer element. Of all DPSs, thermal mass-flow and piezoresistive sensing-based DPSs are widely used in various applications. However, thermal mass-flow-based DPSs are preferable due to their high measuring accuracy even on long connecting hoses [9]. For this reason, we used a thermal mass-flow DPS to show the feasibility of our attack. Specifically, we utilized the SDP800 [3] DPS from Sensirion, and we will provide a detailed explanation of its structure below. *Importantly, the vulnerability we have discovered is inherent to the fundamental design of DPS, specifically the flexible diaphragm used as a force collector. This diaphragm-based sensing mechanism is common across various DPS types. This makes all DPSs vulnerable to our attack as they can pick up tiny vibrations.*

Thermal mass-flow DPS: Fig. 1 depicts the structure of a thermal mass-flow DPS. This DPS uses a pressure-sensing element which consists of a thin semiconductor diaphragm (Silicon Nitrate film), two temperature sensors, and a heating element. As air moves across this diaphragm, it prompts a temperature differential between the temperature sensors. This temperature shift, which correlates with the mass flow rate of the air, is then adeptly translated by the transducer into an electrical signal, reflecting the differential pressure.

Electronics inside of DPSs: DPS has other electronic circuits to condition and relay signals. These circuits process the signal derived from the sensing element, ultimately generating an output indicative of the measured differential pressure. The circuitry amplifies the output signal, including any pressure differential captured by the pressure-sensing element due to sound vibration. As illustrated in Fig.1, this processing sequence involves components like an amplifier, a Digital Signal Processor (DSP), and an Analog-to-Digital Converter (ADC). Once digitized, the signal is directed to a microcontroller via the Communication Interface (CI). The ADC's sampling rate might constrain data transfer to the controller, influencing the sampling rate of the pressure reading — an aspect detailed in Sec. 5.3.

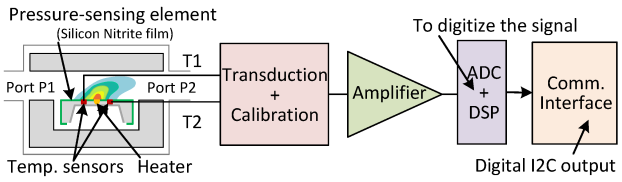

Figure 1. Components of a DPS.

2.2. DPS's Critical Role in Controlled Environments

DPS are widely used in applications that require monitoring the pressure difference between two points in a system. These sensors regulate airflow and maintain specific pressure levels within controlled environments. DPS offers high sensitivity, accuracy, and reliability.

While DPS find applications in various sectors, their role is particularly critical in the semiconductor industry, where they are essential for maintaining cleanroom integrity. Cleanrooms in semiconductor manufacturing serve as highly controlled environments designed to minimize airborne contaminants [10], [11]. Recent contamination incidents at major companies like Samsung and TSMC, resulting in losses exceeding $1 billion [12], [13], underscore the critical nature of maintaining cleanroom integrity.

Besides cleanrooms, DPSs are commonly used in HVAC systems for buildings, offices, and hotels. Monitoring and controlling airflow enable efficient temperature regulation, air quality management, and energy optimization. DPSs are also employed in smart home systems, where they contribute to creating a comfortable and energy-efficient living environment. DPSs are also used in healthcare settings, commonly in Negative Pressure Rooms (NPRs), to measure the negative pressure in the facility [14].

To contextualize our BaroVox attack, we examine the deployment of DPSs and sound systems in semiconductor cleanrooms as an example of a secure pressure-regulated facilities. The following sections dissect the cleanroom components pertinent to this attack, focusing on the integration of pressure sensors and audio systems.

2.3. Components of a real-world cleanroom

Cleanrooms may vary in design and specifications across different organizations. However, the primary objective remains consistent: to prevent contamination by maintaining a sterile environment. Fig. 2 illustrates a standard cleanroom design. The process begins with intake vents that channel outdoor air into the HVAC system. This system filters and conditions air by adjusting temperature and humidity, which are crucial for semiconductor production. Air pumps and compressors then push the air through HEPA filters, which remove contaminants, ensuring a clean environment. The RPM system is instrumental in constantly overseeing the pressure differences across cleanroom areas. The Room Pressure Monitoring (RPM) feeds data from the DPS into the HVAC, directing adjustments in airflow. The ensuing subsections spotlight the deployment of DPSs and sound systems within cleanrooms, given their significance to BaroVox.

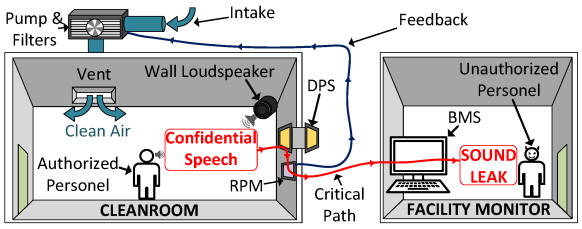

Figure 2. Components of a semiconductor cleanroom.

2.3.1. Sound Systems in Cleanrooms and Their Deployment. Cleanroom operations *demand* robust communication for coordination and safety adherence. Consequently, these environments often feature comprehensive sound systems [2]. Intercoms serve a dual purpose of ensuring clear communication while reducing contamination risk. Additionally, cleanroom personnel often use protective clothing with integrated microphones, enabling efficient communication without removing protective gear, thus preserving the controlled environment. Wall-mounted speakers are strategically placed to broadcast announcements and facilitate communications.

These audio systems, crucial for operational efficiency, are frequently installed in close proximity to DPS due to space constraints and the need for integrated environmental control and communication. Moreover, the audio systems are channels for disseminating ***proprietary and confidential data*** — encompassing ***process parameters, recipes, design information, and quality control protocols***.

2.3.2. DPS Deployment in Cleanrooms. DPSs are typically mounted in the cleanroom walls or close to the Building Management System (BMS). Each DPS employs a dual-port system connected to sampling tubes, strategically positioned to monitor pressure differentials between the cleanroom and adjacent spaces. One sensor port often connects to the cleanroom interior, while the other links to a reference point, usually via sampling tubes, enabling accurate pressure differential measurements. The RPM system utilizes these DPS to continuously monitor cleanroom pressure, transmitting readings to the BMS.

2.4. Privacy implications of DPS deployment

In this research, we show that the deployment of DPS in close proximity to audio systems, such as intercoms and speakers, raises significant privacy concerns. While this co-location is often necessary for facilitating communication [2], [15], [16], it inadvertently creates an acoustic side-channel that can be exploited by attackers.

The high sensitivity of DPS to pressure variations makes them vulnerable to unintentional acoustic coupling. Sound waves from nearby audio systems can induce minute vibrations on the sensor's diaphragm, causing measurable changes in the DPS output. Consequently, the DPS effectively

functions as a makeshift microphone, allowing attackers to analyze the output variations and partially reconstruct the original audio signal. This enables eavesdropping on confidential conversations.

The privacy implications of this vulnerability are particularly concerning in environments where confidential discussions occur, such as cleanrooms, corporate meetings, healthcare consultations, or personal conversations in residential settings. Attackers may leverage this to gather valuable intelligence, compromise trade secrets, or invade personal privacy, all without the need for physical intrusion. Amid the rising IP war, this can be costly with IP and national security at stake. Privacy implications of BaroVox in cleanrooms are discussed in detail in Appx. A.

3. Related work

In this section, we comparatively discuss BaroVox with state-of-the-art works in the following two categories.

Side-channel speech eavesdropping: Side-channel eavesdropping has been considerably studied in academic literature [17]–[26]. Several works, such as [17]–[21], [23], [27]–[30], have investigated the vulnerabilities of motion sensor eavesdropping, most notably in mobile devices. Sami et al. [31] devised a novel acoustic side-channel threat using the lidar sensors found in consumer-grade robot vacuums. Roy et al. [32] looked into the practicality of using a mobile device's vibration motor as a sound sensor. Nassi et al. [33] exploit the vibrations of a hanging bulb inside a room to retrieve sound from desk lamp light bulbs through an optical side-channel attack. [34]–[40] are other works that focus on acoustic signal eavesdropping. These studies raised public awareness of the feasibility of recovering sound by analyzing non-acoustic data. Our work takes a unique approach to side-channel attacks by demonstrating the ***first-ever use of this technique on pressure sensors***, adding a new dimension to this field of study.

Attacks on pressure sensors: Tu et al. [41] reveal threats of Electromagnetic interference spoofing attacks on tire pressure sensors to over/under-inflate car tire. Barua et al. designed malicious music to create resonance in pressure sensors to fool the pressure sensor used in the RPM systems of a negative pressure room (NPR) to turn NPR's negative pressure into positive one [42]. Our research distinguishes itself by highlighting the sensitivity of pressure sensors across a range of frequencies beyond the resonant frequency. Moreover, while both papers engage DPS sensors, they diverge significantly in focus and methodology. Barua et al. exploit resonant frequencies to manipulate pressure readings and create hazardous conditions — an integrity attack. In contrast, our work reveals the potential of DPS as a covert channel for leaking acoustic information — a confidentiality attack. Additionally, we demonstrate the potential to extract sound signals from minor fluctuations in pressure readings, a new avenue for exploiting pressure sensors.

4. Threat Model

Fig. 3 depicts an overview of our attack model, outlining the attacker's target system, goals, capabilities, and assumptions.

Attacker's Target: We consider a scenario in which the attacker targets a sensitive conversation or meeting taking place in an environment equipped with DPSs, such as a corporate boardroom, a secure research facility, a cleanroom in a semiconductor manufacturing plant, or any pressure-regulated room. The participants in the conversation are unaware of the potential for eavesdropping through the DPSs and assume that their discussion is confidential.

Attacker's Goal: The attacker aims to reconstruct speech stealthily or recover sensitive information from the environment by exploiting the acoustic side-channel vulnerability in DPSs. The attacker could use this data for malicious activities, including espionage, blackmail, and theft.

Attacker's Capabilities: The attacker is assumed to know the type and characteristics of the DPS deployed in the target environment. This information can be obtained through technical specifications or by studying similar systems as discussed in Sec. 5. The attacker also possesses the necessary technical skills and resources to process and analyze the captured pressure data using signal processing and machine learning techniques.

Attack Scenarios: BaroVox can be used for either *targeted eavesdropping*, where the attacker targets specific individuals, conversations, or events, or *broad-spectrum eavesdropping*, where the attacker indiscriminately monitors the environment to recover any valuable or sensitive audio content.

These attacks can be executed across various settings. In industrial and corporate environments, attackers can intercept confidential discussions about trade secrets or strategic plans. The vulnerability extends to private spaces, where DPSs in smart home or hotel HVAC systems could capture personal conversations. In healthcare facilities, government buildings, or military installations, BaroVox could be used to eavesdrop on confidential patient-doctor conversations or gather classified intelligence. The attack's effectiveness may vary based on factors like DPS model and environmental conditions, but the range of potential scenarios underscores the critical need to address this acoustic side-channel vulnerability.

Attacker's Access Level: We assume that the attacker does not have physical access to the targeted environment or the ability to tamper with the DPS hardware. Instead, the attacker relies on remote access to the pressure sensor readings, either through a compromised system, a malicious insider, or by intercepting the sensor data transmitted over a network.

Attackers may obtain pressure reading data by exploiting personnel with clearance to access pressure sensor logs but not necessarily the targeted environment itself, such as a rogue employee, visitor, or maintenance worker. Our attack model focuses on situations where security measures protecting pressure log data are less stringent than those within the targeted environment.

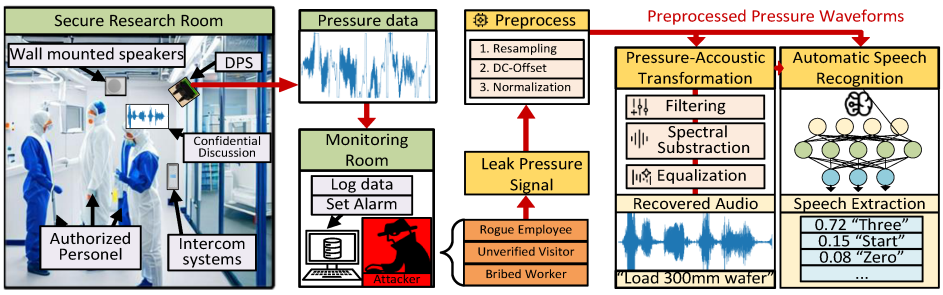

Figure 3. Overview of the attack model.

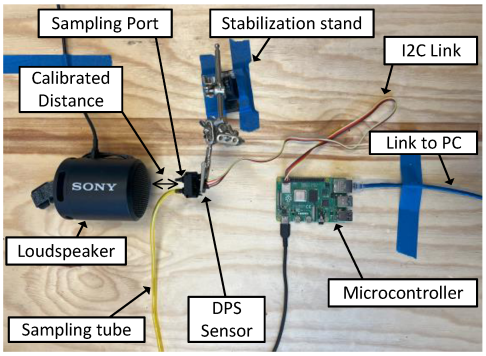

Figure 4. Our experimental setup for a feasibility study.

Attackers may also tamper with the pressure monitoring system during delivery or installation, enabling it to transmit data via the Interne. Modern RPMs often have internet modules that allow remote connections, which can be exploited by the attacker to intercept data without being physically present [43]. Attackers may also access archived pressure sensor logs, which are often overlooked but preserved for diagnostics or maintenance. Studies have shown that attackers exploit vulnerabilities in log security, using techniques like mapping sensor locations through security alerts or bypassing log protections [44], [45].

Assumptions: We assume that the DPSs in the target environment are deployed in proximity to audio sources, such as speakers, intercoms, or areas where conversations occur. This assumption is based on the common practice of integrating audio systems with DPSs for various purposes.

5. Feasibility Study

Exploiting DPSs' acoustic side-channel vulnerability to recover speech signals requires a thorough understanding of the sensor's behavior, sampling rate, and frequency response. Sensor datasheets often conceal these details, which presents challenges for attackers attempting to reconstruct audio from pressure readings. Our analysis aims to highlight these challenges and demonstrate the attack's feasibility.

5.1. Experimental setup

The experimental setup is depicted in Fig. 4. To showcase our attack's feasibility, we use the SDP800 DPS [3] (see 2.1), securely mounted to avoid movement-related noise. A Sony XB13 [46] speaker is placed 5 cm directly underneath one of the pressure ports to play the audio clips. To prevent sound signals from influencing the pressure in the other port, we connect a sampling tube to it, isolated by positioning its opening a meter away, shielded by a hard surface. Signals from the sensor are read using a Raspberry Pi 3 Model B [47]. We employ a custom Python script to convert the pressure readings into a format an attacker can process.

5.2. Primary observation

Fig. 5 compares the Short-Time Fourier Transform (STFT) spectrum of the word "one" captured by an iPhone 13 microphone (right) and the SDP800 DPS (left). The spectrum demonstrates the variations in power across different audio frequency components over time. Notably, the DPS's response is discernibly feebler and less extensive than the microphone's, especially in frequencies surpassing 0.4 kHz. Distortions are observable even in zones of ostensibly strong signals (0.4 - 0.85 kHz). Furthermore, the DPS fails to register signals exceeding 0.9 kHz. A comparative analysis of the Fast Fourier Transform (FFT) plots in Fig. 6 echoes these observations, with the DPS showing subpar responsiveness to signals above 0.4 kHz. This indicates a notable loss in acoustic fidelity when reconstructing speech from the DPS data (also see Fig. 7).

5.3. Challenges

A successful attack requires recovery of information lost from pressure readings, at least partially. Here, we examine the challenges attackers may face during this process.

5.3.1. Low sampling rate. The sensor's low sampling rate is a significant challenge in executing the BaroVox attack.

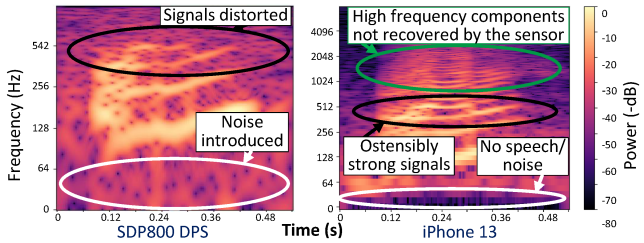

Figure 5. STFT plot of a spoken word "one".

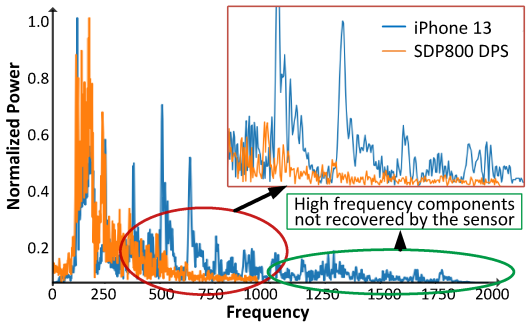

Figure 6. FFT plot of a spoken word "one".

For example, the SDP800 sensor used in our experiment has a theoretical maximum sampling rate of 2.2 kHz [3]. However, as demonstrated in Fig. 6, it only captures frequencies up to approximately 0.9 kHz, with notably weak signals within the 0.6 - 0.9 kHz range. Consequently, given Nyquist's theorem [48], the sensor's effective sampling rate is approximated at 1.8 kHz, a shortfall of 0.4 kHz from its potential. This limitation restricts speech recovery within the sub-0.9 kHz range, significantly below the 4 kHz threshold necessary for intelligible speech [49], [50], causing aliasing and complicating the recovery stage.

5.3.2. Non-linear frequency response. DPS can be depicted as a second-order dynamic system since it employs an elastic diaphragm for pressure force collection [51]. This leads to a non-linear frequency response to sound waves of varying frequencies. To investigate this, we construct an audio file with a sine sweep wave ranging from 1 Hz to 2 kHz. We play the audio file through a speaker using the setup specified in Sec. 5.1. We then reconstruct audio from pressure reading and analyze the data using FFT plots. FFT plots of the reconstructed audio (Fig. 7) show reduced signal strength across frequencies compared to the original. The original audio has consistent power up to 2 kHz, whereas the reconstructed signal loses power as frequency increases, dropping sharply after 0.4 kHz. At frequencies above 0.65 kHz, its power nears mere noise levels. These insights can be leveraged to equalize the signal by boosting the strength of the affected frequencies (see Sec.6.3.3).

5.3.3. Low Signal-to-Noise Ratio (SNR). DPSs yield non-zero readings even in the absence of sound due to the presence of ambient noise and pressure fluctuations. Acoustic disturbances like machinery operation, air conditioning noise, and door activities further affect the readings. These ambient noises introduce unwanted interference to the sensor's data, lowering the SNR and degrading the quality of reconstructed speech. To mitigate this issue, we applied filters and spectral subtraction techniques to improve the audio recovery quality (see Sec. 6.3).

6. Attack Design and Implementation

This section outlines the attack system design. We begin with a mathematical model to demonstrate sound waves' impact on DPSs and introduce two BaroVox design solutions.

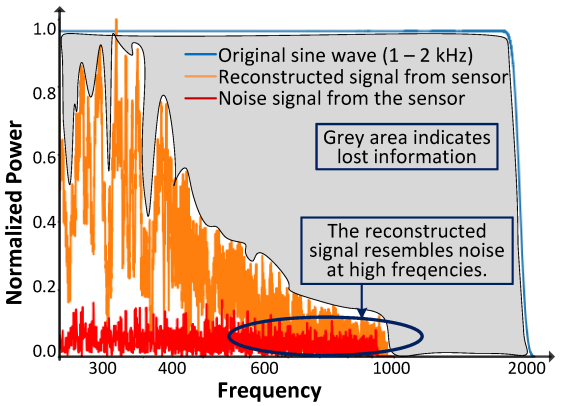

Figure 7. Non-linear frequency response of the SDP800 DPS for a sine sweep wave with a frequency range of 1 to 2 kHz.

6.1. Modeling effects of sound on DPSs

Sound waves, as disturbances propagating through matter, create varying local pressures by compressing and expanding air. Thus, sound can be modeled as a pressure wave. If a sound is played at a frequency f with an initial phase ϕ and a wavelength of λ, then the change in pressure due to sound ($\Delta P_s(t)$) at a distance x from the sound source can be represented:

$$\Delta P_s(t,x) = A(x,f,\mu) \cdot P_{smax} \cdot sin(kx \pm ft + \phi) \quad (1)$$

where P_{smax} is the maximum pressure change due to sound and $A(x,f)$ represents the attenuation of a sound wave, which depends on distance x and frequency f of the audio source and noise μ.

If there is a DPS at a distance $x = x_0$ from a sound source, we model the perturbation in the reading of the DPS as a result of a sound wave as follows:

$$P(t) = P_o(t) + \Delta P_s(t, x_0) \quad (2)$$

where $P(t)$ is the total measured pressure by the DPS, and $P_o(t)$ is the original pressure and noise read by the DPS without sound.

6.2. Design solutions: overview

BaroVox offers two design approaches to address challenges discussed in Sec. 5.3. Both designs pivot on a Pressure-Acoustic Transformation (PAT) outlined below. Fig. 3 summarizes how these designs integrate into the attack model.

Pressure-Acoustic Transformation (PAT): PAT converts DPS readings into wave files. PAT reduces the influence of the normal differential pressure of the target environment using Eqn. 3.

$$S(t) = P(t) - \bar{P}(t) \quad (3)$$

where $S(t)$ is the amplitude of the sound wave at time t, $P(t)$ is the instantaneous pressure reading from the DPS and $\bar{P}(t)$ is the mean value of $P(t)$. This levels the DC offset due to the normal differential pressure of the room by subtracting

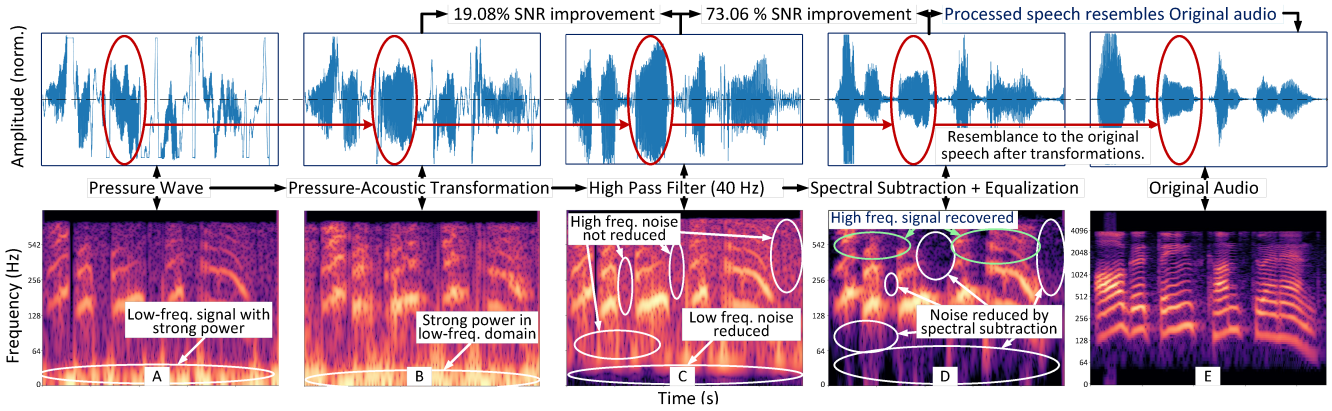

Figure 8. System Design I: Design overview and its effect on a speech recorded by a pressure sensor.

the mean of the pressure signal ($\bar{P}(t)$) from $P(t)$. $S(t)$ is further refined using either of the design solutions discussed below.

Design Solution I (DS-I): DS-I aims to enhance the SNR of reconstructed speech ($S(t)$) using a series of cascaded digital signal processing techniques for improved audio quality. It employs a smart spectral subtraction method in conjunction with standard normalization, high-pass filtering, and DC-offset reduction techniques. This approach proves valuable for attackers when the retrieved speech is substantially distorted by noise, efficiently mitigating noise and other audio signal distortions. Furthermore, DS-I is not bound by vocabulary size, providing flexibility in recognizing a wide range of words and phrases. Sec. 6.3 discusses **DS-I** in detail.

Design Solution II (DS-II): DS-II develops an Automated Speech Recognition system employing Deep Learning. It utilizes SpeechCommands datasets and utilizes a classification model. Attackers may prefer DS-II when more precision is desired and focus is required on specific critical words, numbers, or instructions in the target environment setting.

6.3. Design solution I (DS-I)

Fig. 8 provides a comprehensive overview of the techniques used in DS-I. Below, we delve into a deeper analysis of each technique.

6.3.1. High-pass filter. A high-pass filter refines the recovered audio ($S(t)$) by enhancing its SNR. We initially acquire samples from the DPS without sound to observe acoustical disturbances induced by various environmental conditions in the target environment. The FFT of this noise data, shown in Fig. 7 (red line), displays dispersed noise energy across frequency components. However, the speech recorded by an iPhone 13 in Fig. 5 (right) does not contain speech data in frequencies less than 40 Hz. In contrast, the audio from the SDP800 DPS in Fig. 5 (left) is filled with noise below 40 Hz, which is undesirable. Therefore, we apply a 3rd-order Butterworth high-pass filter with a cutoff frequency of 40 Hz to remove the low-frequency component of the noise. If not removed fully during PAT (see Eqn. 3), the DC offset generated by the ambient pressure in the room would likewise be nullified using the filter.

To further study the filter's influence, we record a short three-second speech and reconstructed the audio from the sensor data using Eqn. 3. The speech reads: "*All good things come to an end.*" The resulting SNR value of the reconstructed speech is 7.464 dB. With the high pass filter, SNR rises to 8.888 dB – 19.08% improvement. This effect is evident in Fig. 8 (C)'s spectrum and waveform plot. Post PAT transformation from pressure wave (A) to speech data (B), prominent power in the low-frequency domain emerges. However, after the high-pass filter application, this noise diminishes (C), making the waveform more congruent with the original audio.

Given the noise in other frequency ranges overlaps with speech data (see Fig. 5 (right)), a more sophisticated method to eliminate this is discussed in the upcoming subsection.

6.3.2. Spectral subtraction. Spectral subtraction subtracts an estimate of the noise spectrum from the speech spectrum to get a denoised spectrum. Direct subtraction of noise from speech is theoretically ideal but can distort elements of the speech signal with noise-like characteristics. To address this, we differentiate between two sound types within the speech signal: percussive and harmonic components, and their susceptibility to noise.

Percussive components: Characterized by their brief, non-steady nature, percussive sounds lack a clear pitch or tonal quality [52]. Examples include consonants like plosives and non-pulmonic sounds produced by rapid, irregular vibrations within the vocal tract [53]. With sharp attack and decay times, these components exhibit noise-like spectral characteristics.

Harmonic components: In contrast, harmonic components possess a clear, identifiable pitch or tone, produced by steady-state vibrations of the vocal cords [52]. These components encompass vowels and voiced consonants like *'l'*, *'n'*, and *'r'*, exhibiting periodic characteristics [54].

Given the noise-like attributes of percussive sounds, they are inherently vulnerable to distortion during spectral subtraction, impacting the speech signal's integrity [55], [56]. Consequently, while employing spectral subtraction,

emphasis should be on mitigating noise within harmonic components. Subtraction should be *cautiously applied to percussive elements to preserve crucial signal information*. Fig. 11 in Appx B.2 illustrates the spectral subtraction process. Hereafter, we comprehensively analyze each part of the process separately.

Algorithm 1: PHS using median filtering.

Input: s: Input audio, W: Window size, H: Hop size, n_iter: Number of iterations
Output: P: Complex spectrogram of percussive component, H: Complex spectrogram of the harmonic component

1 $S[n\_iter, K] \leftarrow$ STFT of s with window size W and hop size H, where N is the time frame index, and K is the frequency index
2 $S\_mag \leftarrow$ magnitude spectrogram of S
3 $S\_phase \leftarrow$ phase spectrogram of S
4 **for** $n = 1$ *to* n_iter **do**
5 Compute the harmonic component energy envelope: $E_h[n,k] \leftarrow \sum_{j=1}^{\infty} S\_mag[n, j \cdot k]$
6 Compute the percussive component energy envelope: $E_p[n,k] \leftarrow \max_{j \neq 0, j \cdot k \leq K} S\_mag[n, j \cdot k]$;
7 Compute the percussive mask: $M_p[n,k] \leftarrow \frac{E_p[n,k]}{E_h[n,k]+E_p[n,k]}$
8 Compute the harmonic mask: $M_h[n,k] \leftarrow \frac{E_h[n,k]}{E_h[n,k]+E_p[n,k]}$
9 Compute the complex spectrogram of the percussive component: $P[n,k] \leftarrow M_p[n,k] \cdot S[n,k] \cdot \ exp(i \cdot S\_phase[n,k])$
10 Compute the complex spectrogram of the harmonic component: $H[n,k] \leftarrow M_h[n,k] \cdot S[n,k] \cdot \ exp(i \cdot S\_phase[n,k])$
11 **return** P, H

Percussive-Harmonic Separation (PHS): PHS is the initial step in spectral subtraction (see Fig. 11 (B)). We use a technique based on median filtering to separate the speech signal into harmonic and percussive components [5], [6]. **Median filtering** is a signal processing technique that aims to remove noise from a signal by replacing each sample with the median value of a group of neighboring samples. Algorithm 1 shows the technique we used for PHS. First, the STFT of the signal is calculated *(line 1)*, followed by computing the magnitude and phase of the spectra *(lines 2-3)*. Median filtering is then applied to the magnitude spectra across sequential frames, enhancing percussive components and suppressing harmonic ones *(line 5)*. Subsequently, median filtering across frequency bins is performed on magnitude spectra to bolster harmonic components while suppressing percussive events *(line 6)*. We use the two resulting median-filtered spectrograms to generate masks *(lines 7-8)*. These masks are applied to the original spectrogram to separate the harmonic and percussive parts of the signal *(lines 9-10)*. The algorithm yields two signals: one containing only percussive components and one containing only harmonic components *(line 11)* (also see Fig. 11 (B)).

Characterizing noise: Before implementing noise reduction, it is vital to model the noise properties [57], [58]. To characterize the statistical features of noise, we capture a segment of noise in the target environment and produce its power spectrum.

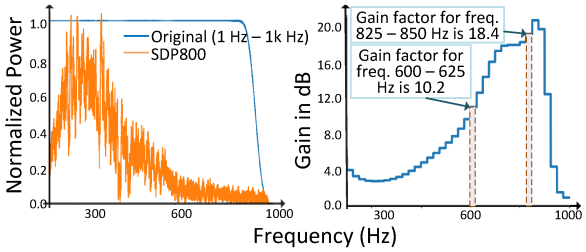

Figure 9. (Left) Frequency response of the SDP800 DPS for a sine sweep wave (1 Hz - 2 kHz). (Right) The gain factor for each frequency band.

Spectral subtraction: We then apply spectral subtraction only to the harmonic signal using an estimate of the residual noise spectrum obtained above (see Fig. 11 (C)). This step results in an enhanced harmonic signal with minimized noise artifacts. Prior to applying spectral subtraction to the percussive component, the residual noise spectrum is downscaled to prevent compromising the speech signal quality. The optimal downscale factor is determined through subjective listening tests. Subsequently, a denoised speech signal is reconstructed by combining the adjusted harmonic and percussive signals. Fig. 8 (D) demonstrates the amalgamated effect of spectral subtraction and equalization (see Sec. 6.3.3) on the spectrum and waveform plot, highlighting the noise reduction across various frequency ranges. Notably, applying spectral subtraction alone leads to an 11.6% increase in the SNR of the reconstructed speech, improving it from 8.888 to 9.921 dB. The ensuing section will address the final step of DS-I: equalization.

6.3.3. Equalization. As noted in Sec. 5.3.2, DPS exhibits a non-linear frequency response. To rectify this, we implement equalization, aiming to establish a desired tonal balance and sound quality. This process involves adjusting the gain of various frequency components of the signal to offset any deviations from a flat frequency response. We use **frequency-domain equalization** technique to modify the sensor's frequency response. This technique efficiently amplifies or diminishes specific frequency ranges without influencing other portions of the signal [59].

Frequency-Domain Equalizer (FDE): We design our FDE using a bank of bandpass filters. The design process involves choosing the right number of filters, setting their center frequencies and bandwidths, and defining the necessary gain or attenuation for each frequency [60]. An experiment was conducted to define these parameters: a sine wave with frequencies ranging from 1 to 1 kHz was played and recorded using the SDP800 DPS. Normalized FFT plots of both the original and the DPS-reconstructed signals are presented in Fig. 9 (left). It is clear that the sensor's non-linear response affected the reconstructed signal.

Filter specifications: The number of bandpass filters required is determined by various parameters, including the signal's frequency range (1 kHz in our case), desired frequency resolution, and the filter bank's complexity. A balance between filter bank complexity and required frequency

resolution is needed for optimal computational efficiency and equalization accuracy. Our approach uses 40 bandpass filters, each with 25 Hz equal-width bands. Subjective tests indicated that increasing the filter count did not significantly improve sound quality, guiding our decision for 40 filters.

To calculate the gain at each frequency band, we used the reconstructed signal and the original sine wave. We put them through Fourier transformation to get their frequency spectrums. The gain for each band is computed as the ratio of the original sine wave's amplitude to the amplitude of the reconstructed signal at the band's center frequency. Fig. 9 depicts the gain factor for each frequency band. These gain factors were then applied to the output signal for each frequency band to produce an equalized signal that balanced the non-linear response of the DPS. Fig. 8 depicts the result of the equalization process, in which the higher frequencies of the input signal are amplified after passing through the equalizer, increasing the SNR value by 55.04% from 9.921 dB to 15.382 dB.

6.4. Design solution II (DS-II)

DS-I offers enhanced audio with a lower SNR, enabling attackers to perform partial manual speech recognition. However, there are times when the attacker wants to get a more precise and accurate recognition of words spoken inside the target environment. For this purpose, we develop an Automated Speech Recognition (ASR) DL model. The ASR solution classifies the pressure signals into their textual representations.

6.4.1. Task and dataset selection. With DS-II in operation, the attacker aims to **automatically extract and categorize** spoken words from the captured pressure signals. Given communication inside sensitive environments is often restricted for clarity and brevity, we want our model to learn to classify pressure wave signals into a limited set of **keywords**. For this, we focus on types of speech containing critical information; specifically, we target spoken digits and commands. For example, the sentence "Load 3.1-millimeter wafers" presents insight into both the manufacturing process and product specification within a semiconductor manufacturing cleanroom through the command "load" and the number "3.1," respectively. We focus on digit and command classification using the **SpeechCommands** [61] dataset. SpeechCommands contains common speech commands from multiple speakers in various environments. An overview of the datasets is available in Appx. B.3.

Considerations on vocabulary limitations: While it is evident that real-world applications would present a more diverse vocabulary, the current task is focused on unveiling the latent risks tied to this unexplored side-channel attack. Employing a constrained vocabulary for this proof-of-concept might not entirely reflect the breadth of real-world scenarios, yet it aptly illustrates the potential of the proposed approach. Generating a comprehensive dataset is undoubtedly resource-intensive and costly. However, it is crucial to emphasize that the primary aim here is to spotlight potential security vulnerabilities. In a scenario where confidential information is the target, it is reasonable to assume that attackers, understanding the value of the information at stake, would not balk at investing in a robust dataset to realize their illicit objectives.

6.4.2. Dataset transformation. As the dataset involves perfect *microphone* recordings of speech utterances, we use an **acoustic-pressure transformation strategy** to transform them into *pressure-wave* datasets. Given sound clips and keyword labels, denoted as $(x_i(t), y_i)$, we play the speech through a speaker in the proximity of a pressure sensor. The data is collected in a research facility, with varying distances and orientations between the speaker and the pressure sensor, as discussed in Sec. 7. This creates a new dataset that pairs each sound signal $x_i(t)$ with the recorded pressure $P_i(t)$. We then process $P_i(t)$ by using **PAT** (see Sec. 6.2) yielding the signal $S_i(t)$. We create a final dataset by pairing the reconstructed speech signal $S_i(t)$ with their original keyword labels, forming pairs $(S_i(t), y_i)$. An attacker can use this new dataset to train models.

6.4.3. ASR model architecture. While numerous time series classification models are available in the literature [62]–[64], not all of them are well-suited for our specific application due to the challenges posed by the low SNR and the low sampling rate of the sensor. We build upon *ResNet* [65], [66], known for extracting information from sparse signals [67]. Our contribution centers around our ASR model, designed to address the challenges posed by low SNR and non-linear frequency response of the DPS. This includes *learnable* ***denoising autoencoder*** and ***equalization layers***, providing a robust solution to these specific issues. It also employs a spectrogram representation defined by FFT bins, window length, and hop size parameters. We sweep these parameters to find the optimal receptive field for our transformed dataset. The full explanation of each spectrogram parameter and the model architecture is included in Appx. C.

7. Evaluation

This section evaluates BaroVox's design solutions across various metrics and scenarios. The experiments are conducted in a seminar room of an anonymous research lab, simulating real-world conditions while ensuring a controlled environment for data collection.

7.1. Methodology and metrics

7.1.1. Manual Speech Recognition (MSR). We evaluate DS-I using MSR. For this task, we recruited 18 volunteers from our institution. The survey includes people with diverse linguistic backgrounds, with only eight considering English as their primary language. Others speak languages from Africa, South Asia, and East Asia. We use the setup discussed in Sec. 5.1 to prepare the evaluation dataset. We record 20 sentences focused on general conversations, scientific theories, and semiconductor fabrications using an iPhone 13

microphone and DPS. We use DS-1 to perform pressure-acoustic transformation, remove background noise, and improve the overall quality of the speech. We then use the **word error rate (WER)** and the **mean opinion score (MOS)** metrics for evaluation. For both metrics, each volunteer is kept in a quiet room for listening.

WER: WER is a standard measure for speech recognition tasks [8]. To calculate WER, we play the reconstructed audio data from the DPS separately for each volunteer and request them to transcribe it based on their comprehension. WER is computed by comparing the transcription to the actual spoken words and counting errors using the formula: $WER = (S + D + I)/N$. S, D, and I denote the number of incorrectly transcribed, missing, and incorrectly added words by the volunteer, while N represents the total number of words in the actual spoken recording. A lower WER indicates greater intelligibility of the reconstructed audio.

MOS: MOS serves as a subjective measure of the perceived speech *quality* [7]. MOS assesses how well volunteers can comprehend the content of the reconstructed speech. Participants are instructed to listen to the reconstructed speech first and then the original audio. Subsequently, they rate the content-wise similarity between the two on a scale of 1 to 5. For instance, if volunteers perceive that the reconstructed audio is understandable and resembles the original audio content, they assign a score of 5. Conversely, if they believe that the reconstructed speech differs significantly from the original speech, they assign a score of 1. This approach allows us to quantify how effectively our system reconstructs speech that is understandable and akin to the original content.

7.1.2. Automatic Speech Recognition (ASR). We evaluate DS-II on automated classification tasks through the transformed SpeechCommands [61] datasets. We mainly use **accuracy** as the evaluation metric to measure our model's ability to accurately and truthfully recognize the given classes.

7.2. Results for MSR

7.2.1. Performance on general speech. Table 5 in Appx. D shows the average **WER** and **MOS** of each ground truth sentence over the responses of the 18 volunteers. On average, participants had a WER of **0.35**, and a MOS of **4.09/5**. A WER of 0.35 means volunteers can reconstruct more than 60% of the speech effectively, a significant achievement from the attacker's perspective. A MOS of 4.09/5 also shows volunteers perceived a notable content resemblance between the reconstructed speech and the original audio. These results show that humans can identify and reconstruct everyday speech partially from pressure-wave signals.

7.2.2. Performance on sensitive information. To assess the impact of targeted eavesdropping on sensitive information, we focus on sentences containing confidential data related to semiconductor manufacturing processes. A detailed discussion of the key components of a cleanroom, including the deployment of DPSs and sound systems, is provided in Appx. A to contextualize the BaroVox attack in such environments. Table 1 shows the aggregated responses of the 10 volunteers on the chosen sentences and the type of confidentiality breach each sentence target. The results are an average **WER** and **MOS** of **0.45** and **3.36/5**, respectively. These results confirm that humans can partially understand speech related to semiconductor fabrication from pressure-wave signals even without extensive field knowledge. The results are comparatively lower than the performance on general speech, and this is due to the participants' limited familiarity with semiconductor cleanroom contexts. To validate that, we conduct the survey on the remaining 8 volunteers, but at this time, we provide them with a short text about the semiconductor manufacturing process. The content of the text is available in Appx. E.1. The survey results in an average $\mathcal{WER}$ score of **0.29** and $\mathcal{MOS}$ score of **3.71**. The score for each individual sentence is provided in Tab. 1. Given an attacker's presumed familiarity with the content in Appx. E.1, we infer that employing the BaroVox attack could enable them to accurately reconstruct over 70% of cleanroom speech. In the semiconductor industry, even fragments of information about processes or specifications can be valuable to competitors, making this level of reconstruction particularly concerning.

We demonstrate the effect of DS-I on some of the sentences that we used in the survey in the following link: BaroVox.

Ground Truth Sentence	**WER** $\mathcal{WER}$[1]	**MOS** $\mathcal{MOS}$[1]	**Breach Type**[2]
Run the diffusion process at three hundred kelvin for nine minutes	0.38/0.20	3.45/3.50	TC
Use process B for the trench isolation	0.62/0.32	3.00/3.88	TC
Etch the pattern using SF6 plasma	0.40/0.25	3.55/3.25	TC
Sputter a one hundred and fifty nano meter layer of Aluminum copper alloy for contacts	0.48/0.45	3.00/3.38	TC
The defect rate for lot number seven is two percent	0.30/0.15	3.27/3.88	TC&QC
Overheating issues impacted four percent of the latest batch	0.28/0.23	3.64/4.00	QC
Decrease metal layer flowrate	0.67/0.65	3.00/2.75	TC
Use different etching gas for oxide layer	0.40/0.29	3.36/3.50	TC
Perform a defect analysis on the wafer batch	0.39/0.29	3.55/3.50	TC&QC
Reduce the concentration of the cleaning solution.	0.53/0.29	3.82/4.50	TC&QC
Average	0.45/0.29	3.36/3.71	

1: $\mathcal{WER}$ & $\mathcal{MOS}$: Volunteers' scores after receiving context.
2: TC - Trade secret, QC - Quality control
NB: IRB exemption approval obtained to conduct the survey.
TABLE 1. PERFORMANCE ON SEMICONDUCTOR-FOCUSED SENTENCES.

Models	**FFT size**	**Our ASR Model**	**ResNet**
SpeechCommand64	64	80.37%	68.16
SpeechCommand128	128	86.82%	70.35
SpeechCommand256	**256**	**90.51%**	80.14

TABLE 2. PERFORMANCE METRICS OF ASR.

7.3. Results for ASR

7.3.1. Performance on the testing dataset. Table 2 demonstrates the performance of the model when trained across various FFT bin sizes, specifically 256, 128, and 64 bins. A clear correlation between the number of bins and performance is observed, with larger bins yielding improved results. In particular, models using 256 FFT bins (**SpeechCommand256**) outperform the others, achieving accuracy of 90.51%. This is not surprising since larger bins imply a more receptive field, which provides better frequency resolution. We compare our model's performance to others who conducted word classification on the original SpeechCommand dataset. We find that the current state-of-the-art ML technique achieves 98.32% accuracy [68]. This difference is deemed acceptable, given the speech signals used in the original model have a higher sampling rate (16 kHz). To demonstrate the impact of our contribution — the addition of *denoising autoencoder* and *equalization layers* — we trained and tested the unmodified ResNet model, and as shown in Table 2, the performance drops by 10.37% percent. Using a 2 m sampling tube reduces our model's performance to 72.8% (see Appx. F). Subsequently, we evaluate BaroVox in different scenarios that influence the recovered speech quality.

7.3.2. Performance analysis under various scenarios. **Evaluation setup**. We investigate BaroVox's responses to speaker volume, distance, and orientation variations. Our experiments cover distances from 5cm to 2m and sound levels from 65dB to 90dB, reflecting common deployments in cleanrooms and other DPS environments. This range of configurations allows us to evaluate the attack's viability across realistic scenarios. Initially, the model's generalized parameters led to suboptimal performance due to their lack of specificity to the nuanced conditions of each scenario. Without fine-tuning, the model struggles with reduced SNRs during lower volume levels or increased distances, impairing its classification accuracy. Different orientations present additional challenges by introducing phase variations and amplitude alterations, which the unadjusted model could not effectively handle. To mitigate these performance issues, fine-tuning is deemed essential. The process involves retraining our ASR model using scenario-specific datasets, each reflecting the unique acoustical and signal characteristics associated with particular volume, distance, and orientation variations. Through this focused retraining, the model's parameters are recalibrated, allowing it to more accurately navigate and adapt to the challenges within each scenario, leading to improved performance.

Varying the volume of the audio source. To investigate the speaker volume's effect (measured in dB) on audio recovery, we record pressure readings at 90 dB, 80 dB, and 65 dB of volume. Table 3 reveals a significant correlation between volume and the classification model's efficacy. As the volume decreases from 90 dB to 65 dB, there's a notable decline in accuracy from 90.51% to 81.38%. Furthermore, Table 3 shows fine-tuning the model for this specific scenario boosted ASR model performance by a huge margin.

Factors		**Performance** (in %) **Fine-tunned / Unmodified** **Speaker Orientation**		
Volume	**Distance**	0°	90°	180°
90dB	5cm	90.51 / 90.51	82.16 / 13.87	82.19 / 10.15
	50cm	78.75 / 7.24	55.00 / 4.27	69.81 / 7.24
	1m	78.27 / 2.48	23.27 / 2.57	30.76 / 3.41
80dB	5cm	86.65 / 12.91	86.47 / 13.61	57.45 / 9.47
	50cm	46.21 / 6.45	43.50 / 3.82	31.13 / 6.62
	1m	68.59 / 2.42	19.94 / 1.97	49.25 / 2.58
65dB	5cm	81.38 / 17.13	75.44 / 12.39	80.83 / 8.98
	50cm	65.68 / 6.19	77.36 / 3.46	55.34 / 6.03
	1m	56.30 / 2.34	25.80 / 2.43	33.71 / 3.00

TABLE 3. PERFORMANCE OF BAROVOX IN DIFFERENT SCENARIOS.

Varying the distance of the audio source. We explore the effect of increasing speaker-DPS distance on attack accuracy, testing at distances of 5cm, 50cm, and 1m. The results are depicted in Table 3. Unsurprisingly, the model's performance varies inversely to the distance, given that sound amplitudes drop off with respect to the square of the distance. While the attack's effectiveness decreases with distance, this is a fundamental limitation shared by all sensor-based side-channel attacks that rely on sound waves. However, our attack model demonstrates robust performance up to 2m, which is comparable to or exceeds the effective range of many other sensor-based side-channel attacks. This range is sufficient for numerous real-world scenarios, particularly in cleanrooms and healthcare settings where DPS and sound sources are often in close proximity.

To further evaluate the BaroVox's robustness at a larger distance, we experimented by putting the speaker at a distance of 2m from the DPS. The classifiers' accuracy was 36.09%, much higher than random-guess accuracy. Nevertheless, the attacker must employ improved ASR models to overcome the distance limitation. We defer this task to future research, but the motivation is explained in Appx. F.

Varying the orientation of the sound source from the DSP. Given that human speech and speakers produce directed sound waves, we probe how a source's angle to the sensor influences ASR performance. We mount the pressure sensor at 0°, 90°, and 180° (when the speaker and DPS face similar direction) from the speaker and we report the result in Table 3. The results indicate that there exists a considerable effect on accuracy. The accuracy is higher at 0° because the sound wave's energy will be focused in the direction of the DPS, potentially creating a strong vibration. The performance reduced to 82.19% accuracy at 180°.

8. Discussion

8.1. Potential Outcomes and their Implications

Our study reveals a critical security vulnerability in DPSs, demonstrating the feasibility of extracting speech from pressure data.

8.1.1. Information leakage across industries. Our evaluation demonstrates significant potential for information leakage. The Mean Opinion Score (MOS) of 3.36/5 for semiconductor-focused sentences indicates substantial semantic content in the reconstructed speech. The semiconductor industry exemplifies the potential impact of BaroVox. Given past IP theft incidents and the presence of sensitive information in cleanrooms (Appendices A.2 and E), even fragments of reconstructed information could be highly valuable to competitors. In healthcare settings, similar reconstruction rates could lead to breaches of patient confidentiality, potentially violating regulations and trust.

8.1.2. Automated threat scaling. Our ASR model's 90.51% accuracy on the SpeechCommand256 dataset represents a significant threat escalation. This high accuracy enables potential automated, large-scale eavesdropping in real-world scenarios. It could facilitate continuous monitoring, data mining of partially reconstructed speech, and contextual attacks based on identified key terms.

The combination of accurate speech reconstruction, high-performance ASR, and increasing connectivity of DPS in smart building systems creates a scenario where automated eavesdropping becomes a tangible threat.

8.1.3. Attack effectiveness in various conditions. Our comprehensive evaluation of BaroVox reveals its effectiveness across various real-world conditions. The attack demonstrates high accuracy at close range (90.51% at 5cm) and maintains significant effectiveness up to 1m (78.27% after fine-tuning), with detectable speech components persisting at 2m. This performance curve aligns with many real-world DPS deployments. Notably, BaroVox adapts well to different volume levels, maintaining 81.38% accuracy at conversational volume (65dB) at close range. The attack also shows remarkable resilience to source orientation, with accuracy remaining above 82% even at perpendicular and opposite angles to the sound source.

These results indicate BaroVox's adaptability to diverse environments, from industrial settings to quieter spaces like offices or healthcare facilities. When compared to other sensor-based side-channel attacks, BaroVox demonstrates comparable or superior performance, suggesting that DPS are as vulnerable to acoustic side-channel attacks as sensors more commonly associated with such vulnerabilities. The significant improvements achieved through fine-tuning (e.g., from 2.48% to 78.27% at 1m) indicate potential for further enhancements, implying that the attack's effectiveness could increase with more sophisticated processing techniques. In conclusion, BaroVox presents a practical and adaptable attack vector, with its performance characteristics closely aligning with real-world DPS deployment scenarios, underscoring the urgent need for comprehensive countermeasures.

8.1.4. Technological trends and future enhancements. The increasing integration of DPS into IoT and smart building systems amplifies BaroVox's potential impact. The trend towards networked, often under-secured sensors significantly expands the attack surface. Future enhancements could further increase the threat level, including advanced signal processing techniques like adaptive noise cancellation, machine learning improvements such as attention mechanisms or transformers, and multi-sensor fusion in environments with multiple DPS. These improvements could extend the attack's effective range beyond 2m and enhance its performance in challenging acoustic environments.

8.2. Limitation

While our research demonstrates the significant potential of BaroVox, it's important to consider the context and limitations of our findings. The attack's effectiveness depends on the attacker's ability to access pressure sensor readings, which varies across different deployment scenarios. This aspect highlights the importance of secure data handling practices in DPS-equipped environments.

The performance of BaroVox is influenced by the specific characteristics of the DPS employed and environmental factors such as the proximity of sound sources. Our experiments with the ASR model in DS-II showed that fine-tuning might be necessary to adapt to diverse scenarios, indicating the attack's adaptability but also the need for scenario-specific optimizations.

Our controlled experiments, while providing crucial insights, represent a first step in understanding this vulnerability. Real-world environments may present additional complexities not fully captured in our current study. Nonetheless, our work serves as a foundation for understanding the risks associated with the acoustic side-channel vulnerability in DPS and highlights the need for further research and development of countermeasures.

8.3. Countermeasures

Potential defenses against BaroVox attacks include:

Sound dampening. Dampening the sound wave by putting a sound-dampening material around the pressure ports is an inexpensive countermeasure. We conduct sound-dampening experiments using 3 materials: acrylic sheet, foam, and paper box. Using these materials, the performance of BaroVox decreased to 4.13%, 3.95%, and 4.04%, respectively.

Filtering. Incorporating a low-pass filter into the electronic components of the DPS can help mitigate the attack. The low-pass filter smooths the pressure readings and removes higher-frequency data representing speech, making it more difficult for an attacker to extract sensitive information. In our experiments, a third-order Butterworth low-pass filter with a cutoff frequency of 40 Hz successfully prevented the attack.

Increasing audio source distance. The proximity of the audio source to the DPS impacts the attack's efficacy. Our research indicates that placing the DPS at a distance exceeding 3.5 m from sound sources effectively thwarts the attack.

9. Conclusion

We introduce BaroVox, a novel side-channel attack that exploits the acoustic vulnerabilities of DPS to reconstruct speech from pressure readings. Our two design solutions, focusing on signal processing and deep learning, demonstrate the effectiveness of BaroVox in recovering sensitive information. The implications of this attack extend beyond information leakage, potentially impacting finances, competitiveness, and security. Our work highlights the need for increased awareness and development of countermeasures to mitigate the risks posed by BaroVox.

Acknowledgment

The authors would like to thank our shepherd and the anonymous reviewers for their valuable comments, which greatly improved this paper. We also express our gratitude to Harsh Thomare for his contributions during the early stages of this research. This work was partially supported by the National Science Foundation (NSF) under award ECCS-2028269. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] A. Abacus, "Pressure sensors: The design engineer's guide," 2017, retrieved from Avnet website. [Online]. Available: https://my.avnet.com/abacus/solutions/technologies/sensors/pressure-sensors/

[2] C. Standard, British and B. ISO, "Cleanrooms and associated controlled environments—," 2004.

[3] T. S. Company, "Datasheet sdp8xx-digital differential pressure sensor," 2019, . (Accessed: 05-21-2024). [Online]. Available: https://sensirion.com/media/documents/90500156/6167E43B/Sensirion_Differential_Pressure_Datasheet_SDP8xx_Digital.pdf

[4] C. Bradford, "Understanding your hvac system: Building pressure monitoring and control," Buildings IOT website, 2024, accessed: 05-21-2024. [Online]. Available: https://www.buildingsiot.com/blog/understanding-your-hvac-system-building-pressure-monitoring-and-control-bd

[5] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, vol. 13, 2010, pp. 1–4.

[6] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals." in *ISMIR*, 2014, pp. 611–616.

[7] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[8] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic speech recognition errors detection and correction: A review," *Procedia Computer Science*, vol. 128, pp. 32–37, 2018.

[9] F. S. AG, "How to decide on piezoresistive or thermal measuring principle," *AZoSensors*, October 13 2020. [Online]. Available: https://www.azosensors.com/article.aspx?ArticleID=1723

[10] D. W. Cooper, "Particulate contamination and microelectronics manufacturing: an introduction," *Aerosol Science and Technology*, vol. 5, no. 3, pp. 287–299, 1986.

[11] H. Kitajima and Y. Shiramizu, "Requirements for contamination control in the gigabit era," *IEEE transactions on semiconductor manufacturing*, vol. 10, no. 2, pp. 267–272, 1997.

[12] A. Shilov, "TSMC's Fab 14B Photoresist Material Incident: $550 Million in Lost Revenue," *AnandTech*, 2019. [Online]. Available: https://www.anandtech.com/show/13975/tsmcs-fab-14b-photoresist-material-incident-550-million-in-lost-revenue

[13] L. Yap, "Samsung's Production Plant Contaminated Resulting in $560 Million Loss," *Tech Critter*, 2018. [Online]. Available: https://www.tech-critter.com/samsung-manufacturing-plant-contamination/

[14] S. L. Miller, N. Clements, S. A. Elliott, S. S. Subhash, A. Eagan, and L. J. Radonovich, "Implementing a negative-pressure isolation ward for a surge in airborne infectious patients," *American journal of infection control*, vol. 45, no. 6, pp. 652–659, 2017.

[15] L. Audio, "Healthcare operating theatre audio," 2023, accessed: 05-21-2024. [Online]. Available: https://www.litheaudio.com/healthcare-operating-theatre-audio.html

[16] Zenitel, "Cleanroom intercom station ip-cror datasheet," 2023, accessed: 05-21-2024. [Online]. Available: https://www.zenitel.com/print/pdf/node/4584

[17] 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.

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

[19] 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.

[20] 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.

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

[22] S. Rokka Chhetri and M. A. Al Faruque, "Side channels of cyber-physical systems: Case study in additive manufacturing," *IEEE Design & Test*, vol. 34, no. 4, pp. 18–25, 2017.

[23] P. Marquardt, A. Verma, H. Carter, and P. Traynor, "(sp) iphone: Decoding vibrations from nearby keyboards using mobile phone accelerometers," in *Proceedings of the 18th ACM conference on Computer and communications security*, 2011, pp. 551–562.

[24] Y. Zhang, R. Yasaei, H. Chen, Z. Li, and M. A. Al Faruque, "Stealing neural network structure through remote fpga side-channel analysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4377–4388, 2021.

[25] M. A. Al Faruque, S. R. Chhetri, A. Canedo, and J. Wan, "Acoustic side-channel attacks on additive manufacturing systems," in *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*, 2016, pp. 1–10.

[26] S. R. Chhetri, A. Canedo, and M. A. A. Faruque, "Confidentiality breach through acoustic side-channel in cyber-physical additive manufacturing systems," *ACM Trans. Cyber-Phys. Syst.*, vol. 2, no. 1, Dec. 2017. [Online]. Available: https://doi.org/10.1145/3078622

[27] 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*, 2020, pp. 23–26.

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

[29] M. Gao, Y. Liu, Y. Chen, Y. Li, Z. Ba, X. Xu, J. Han, and K. Ren, "Device-independent smartphone eavesdropping jointly using accelerometer and gyroscope," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[30] M. Gao, L. Zhang, L. Shen, X. Zou, J. Han, F. Lin, and K. Ren, "Kite: exploring the practical threat from acoustic transduction attacks on inertial sensors," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 696–709.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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," *arXiv preprint arXiv:2301.10056*, 2023.

[36] 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.

[37] 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.

[38] B. Nassi, R. Swissa, Y. Elovici, and B. Zadov, "The little seal bug: Optical sound recovery from lightweight reflective objects." *IACR Cryptol. ePrint Arch.*, vol. 2022, p. 227, 2022.

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

[40] 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.

[41] Y. Tu, V. S. Tida, Z. Pan, and X. Hei, "Transduction shield: A low-complexity method to detect and correct the effects of emi injection attacks on sensors," in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 901–915.

[42] A. Barua, Y. G. Achamyeleh, and M. A. Al Faruque, "A wolf in sheep's clothing: Spreading deadly pathogens under the disguise of popular music," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 277–291.

[43] Primex, "Environmental monitoring solutions: Room pressure monitoring," https://www.primexinc.com/en/solutions/environmental-monitoring/room-pressure-monitoring, 2021, accessed: 05-21-2024.

[44] S. Lee, W. Choi, H. J. Jo, and D. H. Lee, "How to securely record logs based on arm trustzone," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, pp. 664–666.

[45] R. Paccagnella, K. Liao, D. Tian, and A. Bates, "Logging to the danger zone: Race condition attacks and defenses on system audit frameworks," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1551–1574.

[46] Sony, "Xb13 extra bass™ portable wireless speaker," 2021. [Online]. Available: https://www.sony.com/electronics/support/res/manuals/5025/f505fde429679d4719abd78c1a231ac4/50254675M.pdf

[47] R. Group, "Datasheet: Raspberry Pi 3 Model B," https://us.rs-online.com/m/d/4252b1ecd92888dbb9d8a39b536e7bf2.pdf, 2018, accessed: 05-21-2024.

[48] P. Vaidyanathan, "Generalizations of the sampling theorem: Seven decades after nyquist," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 9, pp. 1094–1109, 2001.

[49] D. A. Heide and G. S. Kang, "Speech enhancement for bandlimited speech," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 1. IEEE, 1998, pp. 393–396.

[50] E. Villchur, "Signal processing to improve speech intelligibility in perceptive deafness," *The Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1646–1657, 1973.

[51] S. A. Whitmore and B. Fox, "Improved accuracy, second-order response model for pressure sensing systems," *Journal of aircraft*, vol. 46, no. 2, pp. 491–500, 2009.

[52] P. Ladefoged and K. Johnson, *A course in phonetics*. Cengage learning, 2014.

[53] M. Gilman, "The science of voice and the body," *The Oxford handbook of music and the body*, pp. 62–78, 2019.

[54] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech communication*, vol. 16, no. 2, pp. 175–205, 1995.

[55] E. Cano, M. Plumbley, and C. Dittmar, "Phase-based harmonic/percussive separation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[56] H. Tachibana, N. Ono, H. Kameoka, and S. Sagayama, "Harmonic/percussive sound separation based on anisotropic smoothness of spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2059–2073, 2014.

[57] L. Zhang, F. Schlaghecken, J. Harte, and K. L. Roberts, "The influence of the type of background noise on perceptual learning of speech in noise," *Frontiers in Neuroscience*, vol. 15, p. 646137, 2021.

[58] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2480–2480, 2004.

[59] S. Li, W. Yuan, J. Yuan, B. Bai, D. W. K. Ng, and L. Hanzo, "Time-domain vs. frequency-domain equalization for ftn signaling," *IEEE transactions on vehicular technology*, vol. 69, no. 8, pp. 9174–9179, 2020.

[60] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Communications Magazine*, vol. 40, no. 4, pp. 58–66, 2002.

[61] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *ArXiv e-prints*, Apr. 2018. [Online]. Available: https://arxiv.org/abs/1804.03209

[62] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[63] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[64] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," 2016.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[66] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[67] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1987–2000, 2021.

[68] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[69] F. Cobo, D. Grela, and A. n. Conchal, "Airborne particle monitoring in clean room environments for stem cell cultures," *Biotechnology Journal: Healthcare Nutrition Technology*, vol. 3, no. 1, pp. 43–52, 2008.

[70] D. Holbrook, "Controlling contamination: the origins of clean room technology," *History and Technology*, vol. 25, no. 3, pp. 173–191, 2009.

[71] T. Times, "TSMC engineer charged with stealing trade secrets," *Taipei Times*. [Online]. Available: https://www.taipeitimes.com/News/biz/archives/2017/05/03/2003669834

[72] U.S. Attorney's Office Northern District of California, "Chinese citizen sentenced for economic espionage, theft of trade secrets, and conspiracy," *Department of Justice*. [Online]. Available: https://www.justice.gov/usao-ndca/pr/chinese-citizen-sentenced-economic-espionage-theft-trade-secrets-and-conspiracy

[73] U.S. Attorney's Office District of Connecticut, "Three Chinese nationals arrested in scheme to steal and illegally export military-grade carbon fiber from the U.S." *Department of Justice*. [Online]. Available: https://www.justice.gov/usao-ct/pr/three-chinese-nationals-arrested-scheme-steal-and-illegally-export-military-grade

[74] U.S. Department of Justice, "Lexington man and semiconductor company indicted for theft of trade secrets," https://shorturl.at/imqyG, 2018, accessed: 05-21-2024.

[75] A. Instruments, "Alpha 161 Low-Cost Differential Pressure Transducer datasheet," 2021, accessed: 05-21-2024. [Online]. Available: https://www.alphainstruments.com/product/model-161-differential-pressure-transmitter/

[76] T. S. . Co. (2017) Testo 6383 data sheet. Accessed: 05-21-2024. [Online]. Available: https://static-int.testo.com/media/97/f5/9593ea116ffe/testo-6383-EN.pdf

[77] I. Siemens Industry, "Room pressure monitor, technical specification sheet," 2020, accessed: 05-21-2024. [Online]. Available: https://sid.siemens.com/v/u/A6V10322677

[78] N. AG, "PASCAL-ST/ZB Accurate & long-term stable measurement," 2016, accessed: 05-21-2024. [Online]. Available: https://cesstech.com/wp-content/uploads/2023/07/Product-flyer_PascalST_ZB_EN_005391_00-1.pdf

[79] I. Dwyer Instruments, "Room status monitor," 2021, accessed: 05-21-2024. [Online]. Available: https://www.dwyer-inst.com/PDF_files/RSME.pdf

[80] SensoScientific, Inc., "Differential pressure sensor data sheet," https://www.laboratory-equipment.com/media/asset-library/d/i/differential-pressure-sensor-sensoscientific-data-sheet.pdf, 2017, accessed: 05-21-2024.

[81] E. C. Knight, S. P. Hernandez, E. M. Bayne, V. Bulitko, and B. V. Tucker, "Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks," *Bioacoustics*, vol. 29, no. 3, pp. 337–355, 2020. [Online]. Available: https://doi.org/10.1080/09524622.2019.1606734

[82] S. Chang, H. Park, J. Cho, H. Park, S. Yun, and K. Hwang, "Subspectral normalization for neural audio data processing," 2021.

[83] K. Y. Kamal, "The silicon age: Trends in semiconductor devices industry," *Journal of Engineering Science and Technology Review*, 2022, accessed: 05-21-2024. [Online]. Available: https://www.researchgate.net/publication/360851950_The_Silicon_Age_Trends_in_Semiconductor_Devices_Industry

[84] C. Nader, W. Van Moer, K. Barbe, N. Bjorsell, and P. Handel, "Harmonic sampling and reconstruction of wideband undersampled waveforms: Breaking the code," *IEEE transactions on microwave theory and techniques*, vol. 59, no. 11, pp. 2961–2969, 2011.

[85] C. Nader, N. Björsell, and P. Händel, "Unfolding the frequency spectrum for undersampled wideband data," *Signal Processing*, vol. 91, no. 5, pp. 1347–1350, 2011.

[86] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," 2021.

[87] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2022.

[88] R. Hasani, M. Lechner, T.-H. Wang, M. Chahine, A. Amini, and D. Rus, "Liquid structural state-space models," *arXiv preprint arXiv:2209.12951*, 2022.

Appendix A. Basics of Cleanrooms

To assess the impact of targeted eavesdropping on sensitive information, this section explains cleanroom and IP in the semiconductor manufacturing industry, including the implication of side-channel eavesdropping in these types of secure environments.

A.1. Cleanroom in semiconductor industry

Cleanrooms are controlled environments designed to filter out pollutants, ensuring that airborne contaminants remain at acceptable low-level concentrations [2], [69]. While employed across various sectors [70], cleanrooms are paramount in semiconductor manufacturing where even a single speck of dust can drastically compromise chip quality [10], [11]. Consequently, cleanrooms are integral to assuring semiconductor product integrity. Emphasizing their importance, contamination mishaps at giants like Samsung and TSMC have previously led to staggering combined losses of more than $1 billion [12], [13]

A.2. Cleanroom and Intellectual Property (IP)

Cleanrooms not only protect the semiconductor industry from contaminants but also play a vital role by protecting proprietary products from unauthorized access and tampering. With ***IP and national security*** at stake, a lapse in cleanroom security can be costly, especially amid the rising IP war. Numerous theft incidents underscore this; for instance, TSMC grappled with a major trade secret theft attempt [71], and arrests have been made regarding IP theft from US semiconductor firms [72]–[74]. To counter these threats, rigorous security protocols are enforced.

A.3. Pressure sensors used in cleanrooms

DPSs are at the heart of maintaining a cleanroom's integrity for chip manufacturing. These sensors are crucial in regulating the airflow and maintaining a specific pressure level within the cleanroom. Specifically, the cleanroom must be maintained at a higher static pressure than adjacent spaces to prevent contaminants from entering. To achieve this, a Room Pressure Monitoring (RPM) system with an integrated pressure sensor offers real-time differential pressure tracking between distinct points. Table 4 shows popular manufacturers' RPM systems used in semiconductor cleanrooms, all predominantly incorporating DPSs, providing evidence of DPSs' industry prevalence.

S	RPM	Manufacture	Type
1	Alpha 161 [75]	Alpha Inst.	Differential
2	Testo 6383 [76]	Testo	Differential
3	Siemens 547-203 [77]	Siemens	Differential
4	PASCAL-ST/ZB [78]	Novasina	Differential
5	RSME-B-003 [79]	Dwyer	Differential
6	B20-200-OTA [80]	Sensoscientific	Differential

TABLE 4. DIFFERENTIAL PRESSURE SENSORS USED IN CLEANROOMS.

Appendix B. Miscellaneous

B.1. Modeling effects of sound on DPS

Fig. 10 shows the model of sound signal a s a pressure wave and its effect on a pressure reading.

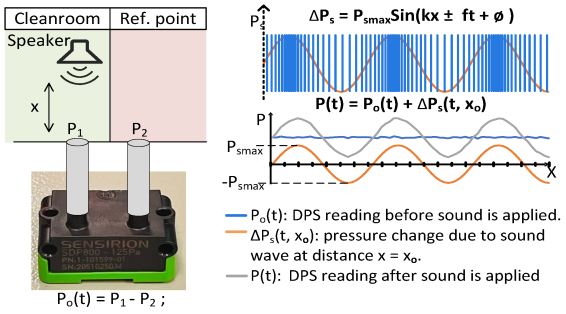

Figure 10. Modeling sound signal as a pressure wave.

B.2. Percussive-Harmoinic Separation using Median Filtering

Fig 11 visually illustrates the spectral subtraction process.

B.3. SpeechCommands dataset composition

The SpeechCommands v2 dataset [61] is a collection of spoken commands in English, consisting of $105,829$ utterances across 35 different words and phrases, such as "yes", "no", "stop", "go", "bed", "bird", "tree", and "wow". The dataset includes recordings from $2,618$ different speakers, spanning a wide range of ages and genders. The dataset was recorded in various acoustic conditions, including different background noise and reverberation levels, and contains both clean and noisy recordings. A complete list of the number of utterances per word can be found in [61].

Appendix C. ASR Model Architecture Summary

C.1. Spectrogram parameters

This section provides an overview of the primary parameters of the $Spectrogram$ transform, which plays a vital role in time series signal processing. Understanding these parameters, including $NFFT$, window length, and window hop, is essential for effectively analyzing and interpreting time series data. For a more in-depth analysis of the effect of these parameters, we refer our readers to [81].

(1) **Number of FFT bins** ($NFFT$) is a parameter in the $Spectrogram$ process determining the frequency resolution of the spectrogram. A higher $NFFT$ value offers better frequency resolution but increased computational complexity. Alternatively, if the NFFT size is increased, the time resolution of the spectrogram will decrease. This occurs because the window used for each time segment becomes larger. As a result, rapid changes in the signal may be overlooked or smoothed out, leading to a loss of temporal information. It is important to tune the $NFFT$ value to fit the specific purpose of the application. Typically, $NFFT$ is set to a power of 2 for optimization. In our experiment $NFFT$ size of 256 results in better accuracy.

(2) **Window length** refers to the length of the window function applied to audio segments in the $Spectrogram$ process. It affects both time and frequency resolutions, with longer windows providing better frequency resolution at the expense of time resolution. Typically, the window length is equal to or larger than $NFFT$.

(3) **Window hop** defines the distance between adjacent windows in the $Spectrogram$ process, influencing time resolution and computational complexity. A smaller window hop increases overlap and time resolution but also raises computational complexity.

C.2. Architecture description

Our ASR model is a 2D convolutional neural network designed for spectrogram-based audio processing. It is inspired by the ResNet [65], [66] architecture and consists of a series of normal and transition blocks. Our contribution centers around the challenges posed by low SNR and non-linear frequency response of the DPS. Our ASR model introduces *learnable* ***denoising autoencoder*** and ***equalization layers*** to the ResNet architecture to provide a robust solution to the specific issues posed by the sensor. The normal

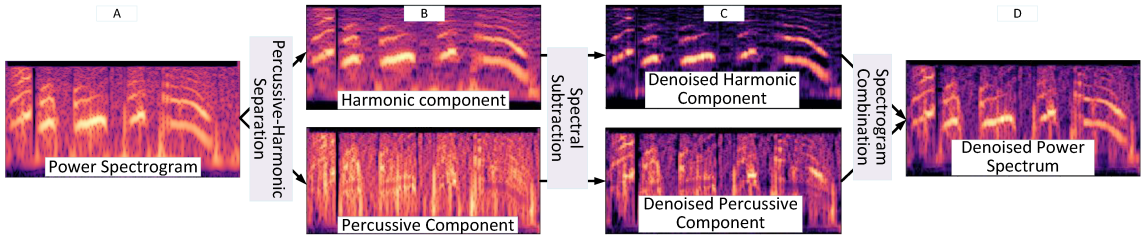

Figure 11. Spectral subtraction by Percussive-Harmonic Separation using median filtering technique.

blocks contain a residual connection, while the transition blocks are responsible for reducing the spatial dimensions and increasing the number of channels. The model employs SubSpectralNorm [82], a normalization technique that applies batch normalization across sub-bands in the frequency domain, improving the model's performance on spectrogram-based tasks. ResNet also utilizes depth-wise separable convolutions for increased efficiency and reduced computational complexity. The final layers include a depth-wise convolution, a 1x1 convolution, and a head convolution for classification. The model is suitable for various audio tasks, such as speech recognition or sound event detection, where the input is a time-frequency representation of the audio signal.

Appendix D. Performance on General Speech-Focused Sentences

Table 5 shows the average WER and MOS of each ground truth sentence over the responses of the 18 volunteers.

Ground Truth Sentence	WER	MOS
An object will remain at rest.	0.54	3.89
For every action in nature there is an equal and opposite reaction.	0.49	3.89
Here is my password 7 5 6 2 3.	0.29	4.25
He said the weather will be cold today.	0.43	4.06
The white egg is bigger than the green one.	0.21	4.36
A picture is worth a thousand words.	0.20	4.22
A journey of a thousand miles begins with a single step.	0.31	4.11
A bird in the hand is worth two in the bush.	0.30	3.75
Actions speak louder than words.	0.27	4.50
Never put off until tomorrow what you can do today.	0.43	3.86
Average	0.35	4.09

NB: IRB exemption approval obtained to conduct the survey.

TABLE 5. PERFORMANCE ON GENERAL SPEECH-FOCUSED SENTENCES.

Appendix E. Semiconductor Manufacturing Process

E.1. Process description

As part of our survey, we provided volunteers with the following information to help them understand the manufacturing process. Semiconductor manufacturing is the process of creating electronic components from semiconductor materials, such as silicon. These components are used in a wide range of electronic devices, from smartphones to computers to cars. The key stages of this process are explained as follows [83]:

- Wafer fabrication: In this step, silicon wafers are created by growing a single crystal of silicon and slicing it into thin, circular wafers.
- Photolitography: A beam of UV light of specific wavelength is passed through a template mask onto a layer of photoresistive material on the waffer to carve a pattern onto the material.
- Etching: Chemicals are used to remove material from the wafer, leaving only the desired pattern behind.
- Deposition: A layer of material is deposited onto the wafer, either by chemical vapor deposition or physical vapor deposition, to create specific features.
- Packaging: The individual chips are cut from the wafer and packaged into the final product.

Appendix F. Future Work

Effect of Distance on ASR Model Performance: Based on the findings in Table 3, we have discovered that the distance between the DPS and the sound source affects the accuracy of our ASR model. With increasing distance, the sensor registers a weaker signal strength. Although this presents a limitation to the current effectiveness of the BaroVox approach, further study is warranted. As the STFT plots in Fig. 12 suggest, even with reduced signal strength due to increased distance (from 0.5m to 2m), the pressure sensor continues to detect speech signals. There's potential for refining the model by accounting for these distance variations.

Vocabulary Limitations of the Current Model: The current model is bound by vocabulary constraints, limiting its speech recognition capability. A strong attacker can enhance the capabilities by creating a complete speech-to-text translation system. Addressing challenges such as frequency spectrum unfolding for under-sampled pressure sensor data could provide a pathway to navigate the sampling rate constraints of the DPS [84], [85].

Exploration of Advanced ASR Models: We plan to build upon other advanced speech recognition models [86]–[88]. Leveraging these models alongside digital signal pro-

cessing techniques may offer avenues to mitigate existing limitations.

Impact of Sampling Rate: In this study, the exploration of sampling rate's influence on model performance was limited due to space constraints. An in-depth investigation into this aspect remains a subject for future research.

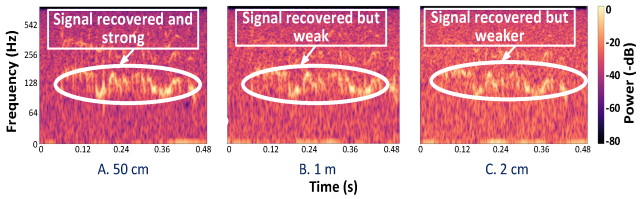

Figure 12. STFT plot of speech signals recovered from pressure readings when the sound source is put at various distances from the sensor.