Reconstructing Intelligible Speech from the Pressure Sensor Data in HVACs

1st Tarikul Islam Tamiti
Cyber-Security Engineering
George Mason University
USA
ttamiti@gmu.edu

2nd Biraj Joshi
Cyber-Security Engineering
George Mason University
USA
bjoshi2@gmu.edu

2nd Rida Hasan
Cyber-Security Engineering
George Mason University
USA
rhasan6@gmu.edu

1st Anomadarshi Barua
Cyber-Security Engineering
George Mason University
USA
abarua8@gmu.edu

***Abstract*—Pressure sensors are an integrated component of modern Heating, Ventilation, and Air Conditioning (HVAC) systems. As these pressure sensors operate within the 0–10 Pa range, support high sampling frequencies of 0.5-2 kHz, and are often placed close to human proximity, they can be used to eavesdrop on confidential conversation, since human speech has a similar audible range of 0-10 Pa and a bandwidth of 4 kHz for intelligible quality. This paper presents WaLi, which reconstructs intelligible speech from the low-resolution and noisy pressure sensor data by providing the following technical contributions: (i) WaLi reconstructs intelligible speech from a minimum of 0.5 kHz sampling frequency of pressure sensors, whereas previous work can only detect hot words/phrases. WaLi uses complex-valued conformer and Complex Global Attention Block (CGAB) to capture inter-phoneme and intra-phoneme dependencies that exist in the low-resolution pressure sensor data. (ii) WaLi handles the transient noise injected from HVAC fans and duct vibrations, by reconstructing both the clean magnitude and phase of the missing frequencies of the low-frequency aliased components. Extensive measurement studies on real-world pressure sensors show an LSD of 1.24 and NISQA-MOS of 1.78 for 0.5 kHz to 8 kHz upsampling. We believe that such levels of accuracy pose a significant threat when viewed from a privacy perspective that has not been addressed before for pressure sensors.**

***Index Terms*—HVAC, pressure sensor, eavesdropping, complex-valued network, magnitude and phase reconstruction.**

1. Introduction

The integration of smart sensors into Heating, Ventilation, and Air Conditioning (HVAC) systems has revolutionized building automation and energy management. Today's HVAC systems utilize a network of smart sensors to monitor environmental and operational parameters in real-time. However, such a rich ecosystem of smart sensors is often considered a "double-edged sword" since they can be used to leak private information. To understand what level of leakage is appropriate, the key question boils down to: *What types of sensors can leak private information from today's HVAC systems and how much information can be inferred from a given sensor data in HVAC systems?*

To answer the above question, we need to point out that pressure sensors are an integrated component of HVAC systems. Pressure sensors often operate within the pressure range of 0-10 Pa and have a high sampling frequency of 0.5-2 kHz [1]–[3], which is essential for the dynamic control of fans, dampers, and air handling units for real-time monitoring and fast response in today's HVAC systems [4]–[6]. These pressure sensors are often placed close to humans and installed in room walls, near diffusers, or within ventilation grilles where human occupancy is high to control indoor air quality (IAQ), to monitor thermal comfort, or room-level pressure regulation in spaces like offices, hospitals, or cleanrooms. As these pressure sensors operate within the 0–10 Pa range, support high sampling frequencies of 0.5-2 kHz, and are often placed close to human proximity, they can be used to eavesdrop on confidential conversation, since human speech has a similar audible range of 0-10 Pa [7] and a bandwidth of 4 kHz for intelligible quality.

To eavesdrop on a confidential conversation with intelligible quality, the attacker must reconstruct speech from pressure sensor data with unrestricted vocabulary by answering the following two questions: ***(1)*** *Will it be possible to capture intelligible speech of 4 kHz bandwidth with 0.5-2 kHz sampling frequencies using pressure sensors? And* ***(2)*** *How do we handle transient noise injected from HVAC fans, duct vibrations, physical shocks, or turbulent airflow while reconstructing intelligible speech from pressure sensor data located in HVAC systems?*

In this paper, we answer the above two questions by proposing WaLi (**Wa**ll can **Li**sten), which can recover intelligible speech by reconstructing bandwidth up to 4 kHz from a lowest sampling frequency of 500 Hz from pressure sensor in transient noisy conditions. WaLi employs the following two technical strategies to answer the above two questions:

Strategy 1. With a sampling rate of 0.5-2 kHz, high-frequency speech components are severely aliased and truncated in pressure sensors, while a few lower pitch frequencies are preserved. Unfortunately, a few low pitches are not sufficient to provide perfect intelligibility [8] as most of the formants at high frequencies will be missing from the bandwidth. To reconstruct intelligible speech from the severely aliased spectrogram of pressure sensors, the missing high frequencies should be reconstructed from the

low-frequency aliased components. To make this happen, WaLi employs Conformers [9], combining the strengths of Convolutional Neural Networks (CNNs) and Transformers to capture local and global dependencies between low-frequency pitches and missing high frequencies. In addition to conformers, WaLi employs Complex Global Attention Block (CGAB) to capture the long-range inter-phoneme correlations that exist among pitches and harmonics along both the time and the frequency axes in a spectrogram. Prior work [10] published in USENIX only captures correlations along the time axis.

Strategy 2. To prevent the transient noise of HVAC from impacting speech reconstruction, WaLi reconstructs both the clean magnitude and phase of the missing frequencies from the low-frequency aliased components. To jointly reconstruct clean magnitude and phase, WaLi is designed as a complex-valued network, which can enhance both the amplitude and phase of pressure sensor data using complex-valued time-frequency (T-F) spectrogram [11]. As complex-valued T-F spectrograms have both magnitude and phase information, WaLi takes the complex-valued T-F spectrograms as input and removes noisy phases from them, motivated by the fact that phase plays a crucial role in speech enhancement [12]. Prior works [10], [13], [14] only use real-valued T-F spectrograms. Therefore, they cannot reconstruct clean phase under noisy conditions. WaLi employs complex multi-resolution STFT loss to reconstruct clean magnitude and phase from the noisy pressure sensor data.

We extensively evaluate the effectiveness of WaLi through real-world experiments, using five evaluation metrics - LSD, NISQA-MOS, PESQ, STOI, and SI-SDR. The full form of these terms is given in Section 7.2. Our findings highlight the serious privacy implications of pressure sensors in HVAC systems. The implications could be particularly severe in critical environments, such as industrial facilities, corporate offices, and healthcare institutions, where eavesdropping can expose sensitive information and compromise privacy. Our main contributions are summarized as follows:

1. We propose WaLi, an acoustic eavesdropping system that uses pressure sensor data to reconstruct intelligible user speech. To the best of our knowledge, WaLi is the first method that recovers intelligible speech with an unconstrained vocabulary rather than recognizing individual hot words/phrases from the pressure sensor data.

2. We use complex-valued architecture to jointly reconstruct both the clean magnitude and phase from the noisy pressure sensor data and extensively evaluate the effectiveness of WaLi using five evaluation metrics - LSD, NISQA-MOS, PESQ, STOI, and SI-SDR. A demonstration of the reconstructed audio is provided in the following link: WaLi.

2. Related Work

Acoustic eavesdropping: Acoustic eavesdropping using different sensors in different modalities is extensively explored in the literature. For example, lasers [15], [16], inertial measurement units (IMU) [13], [17]–[20], wireless signals [21]–[26], optical sensors [27], [28], vibration motors [29], and hard drives [30] are explored to reveal great threats to speech privacy. mmEcho [21], mmEve [31], and mmSpy [32] use mmWave sensors to capture vibrations and reconstruct speech from vibration data. However, these works cannot reconstruct speech with full intelligibility from the narrowband vibration data. IMU sensors [18], [19] and gyroscopes [20] are used to eavesdrop using the vibration induced by sound in IMUs and gyroscopes. They can do digit inference, gender recognition, and limited hot-word reconstruction. However, these works still suffer from the narrowband condition of vibration-based side channels and can only recover band-limited sound with damaged speech intelligibility. GlowWorm [33] and Lamphone [28] employ optical sensors to record vibrations and are used to recover speech with limited intelligibility. However, they do not recover clean phases, and, as a result, they will not work in transient noise conditions.

Recently, AccEar [13] and VibSpeech [10] revealed the possibility of using generative adversarial networks to reconstruct intelligible speech with unrestricted vocabulary by learning the mapping between low-frequency pitches and missing high frequencies. Our proposed WaLi solves the following two problems of AccEar [13] and VibSpeech [10]:

1. AccEar [13] assumes that abundant ground-truth speech from the target victim is available for training, which is not realistic. Compared to AccEar [13], our WaLi does not require any {audio, pressure sensor data} pair from a specific target victim for training to recover intelligible audio. This makes our attack model realistic, as it is not practical to have access to the victim's speech for training.

2. Compared to VibSpeech [10], our WaLi can reconstruct intelligible audio in the presence of transient noise injected from HVAC fans, duct vibrations, physical shocks, or turbulent airflow. VibSpeech [10] has not been tested in the presence of a such transient noise. Moreover, our WaLi does not require SpkEnc-type encoders and extra vocoders like VibSpeech, as our WaLi use complex-valued network to extract features from magnitude and phase of the speech.

Pressure sensor-based eavesdropping: Recently, BaroVox [34] proposed to use pressure sensors to eavesdrop on speech in the cleanroom. The main differences between WaLi and BaroVox are: (1) BaroVox only recognizes hot words or phrases, whereas WaLi can reconstruct intelligible speech with unrestricted vocabulary. (2) BaroVox has not been tested for the transient noise condition. Therefore, it is not realistic for the real-world scenario in HVAC systems.

3. Background

3.1. Human Voice and Its Pressure Range

The human voice generates sound by producing acoustic waves that propagate through the air. These waves create fluctuations in sound pressure, typically measured in Pascals (Pa). A comparative study of the voice pressure range is given in Table 1 for different human voice conditions.

Table 1 indicates that the pressure of speech signals is typically restricted between 0 to 10 Pa for a wide range of intelligible speech. Therefore, to successfully eavesdrop

Condition	in dB	in Pa
Quiet whisper	40 dB	0.02 Pa
Conversation	60 dB	0.2 Pa
Loud speech	70 dB	0.6 Pa
Shouting	100-115 dB	6-10 Pa

Table 1. PRESSURE RANGES FOR DIFFERENT VOICE CONDITIONS.

on speech signals, pressure sensors should also need to be sensitive in the same range of 0 to 10 Pa.

3.2. Pitch, Phonemes, and Intelligible Bandwidth

The human vocal folds vibrate at different frequencies. The strongest and slowest vibration is the pitch, and the faster vibrations that occur simultaneously are called harmonics. The pitch and harmonics generated by the vocal folds are selectively converted into phonemes [8].

Phonemes are the basic sound units of speech [8]. The frequencies in the phonemes are distributed particularly in terms of formants, pitches, and harmonics. For example, English has 44 phonemes, and the vowel phonemes of English have a frequency range of $\sim$4000 Hz, where the first few formants carry a significant amount of energy. The stops phonemes have a broadband energy often above 1000 Hz with an aperiodic burst. The fricative phonemes have energy concentrated in 4-8 kHz. Therefore, to capture intelligible speech with all phonemes, the sampling frequency should be at least 16 kHz (twice the maximum frequency content, 8 kHz). However, vowel phonemes and critical formants fall within the 300-4000 Hz range. Therefore, for speech intelligibility, only 8 kHz sampling is sufficient, as up to 4000 Hz bandwidth is enough for intelligible speech reconstruction.

3.3. Pressure Sensor in HVACs

Differential pressure sensors (DPSs) are the state-of-the-art (SOTA) sensors for HVAC systems due to their better control, accurate measurement, and reliable operations [35]. Hence, in the rest of the paper, we use the DPS and the pressure sensor terms interchangeably. DPSs measure the difference between any two pressure levels.

DPSs have an elastic diaphragm, which collects the input pressure from the environment. The elastic diaphragm is placed between two pressure input ports P_1 and P_2 (see Fig. 1(Left)). The diaphragm senses the differential pressure P_1 - P_2 applied to the pressure input ports by changing its shape. The change in the shape of the diaphragm is converted to a proportional output voltage by using a transducer and a Wheatstone bridge. *The diaphragm is sensitive to acoustic pressure and can pick up sound pressure when someone speaks.* In real-world HVAC systems, sampling tubes are connected with either one or both input ports of DPSs. The other end of the sampling tube is connected to pressure pickup devices (see Fig. 1 (Right)), which senses pressure in the environment.

3.4. Pressure Range and Sampling Frequencies of Differential Pressure Sensors (DPSs) in HVACs

Sections 3.1 and 3.2 point out that for eavesdropping using DPSs in HVACs, the pressure sensors should be sensitive

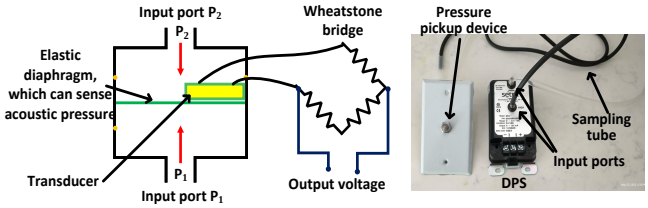

Figure 1. (Left) Internals of a DPS. (Right) The accessories connected with DPSs in a real-world HVAC system.

within 0-10 Pa and have a sampling frequency close to 8 kHz. In HVAC systems, DPSs operating within the 0–10 Pa range and supporting high sampling frequencies within 0.5-2 kHz are typically used to achieve precise environmental control and energy efficiency. These sensors are used in a variety of applications where small pressure changes must be detected accurately and quickly. Common use cases include air filter monitoring, cleanroom pressure regulation, static duct pressure control, and variable air volume (VAV) system management. High-frequency sampling enables real-time responsiveness to transient pressure fluctuations, which is essential for dynamic control of fans, dampers, and air handling units. By providing fast and accurate data, these sensors support improved indoor air quality, occupant comfort, and operational efficiency. A summary of the pressure sensor range and sampling frequencies in HVACs is given in Table 2. Table 2 shows that pressure sensors in HVACs are sensitive to the audible pressure range of 0-10 Pa.

Application	Pressure Range	Sampling Frequency	Function
Air Filter Monitor [36]	0–150 Pa	$\sim$0.7 kHz	Detect pressure drop to signal filter clogging
Duct Static Pressure [37]	0–200 Pa	$\sim$1 kHz	Ensure optimal airflow and energy use
VAV Control [37]	0–200 Pa	$\sim$2 kHz	Adjust airflow with occupancy or thermal need
Pressure Balancing [38]	0–50 Pa	$\sim$0.5 kHz	Stabilize pressure in connected indoor spaces

Table 2. HUMAN AUDIBLE PRESSURE RANGE AND SAMPLING FREQUENCIES OF PRESSURE SENSORS IN HVACS.

3.5. Impact of Speech on Pressure Sensors

Table 2 indicates that DPSs have a sampling frequency of 0.5-2 kHz. However, at least 8 kHz sampling is required to recover intelligible speech. Therefore, a natural question is: *Will it be possible to capture intelligible speech with at least 0.5 kHz sampling frequency using pressure sensors?*. The pitch frequencies for a male speaker vary between 85-180 Hz, whereas for a female speaker, they vary between 165-255 Hz [39]. Although the 0.5 kHz sampling rate can capture pitches and a few harmonics, it is not sufficient to provide perfect intelligibility [8], as most harmonics at high frequencies will be missing from the bandwidth.

Fig. 2 shows the importance of various frequencies in the intelligibility of speech signals. Evidently, the higher frequency components that involve the use of fricatives and other consonants are critical for higher intelligibility (see Fig. 2 (Left)). Unfortunately, with a sampling rate of 0.5 kHz, high-frequency speech components are severely aliased

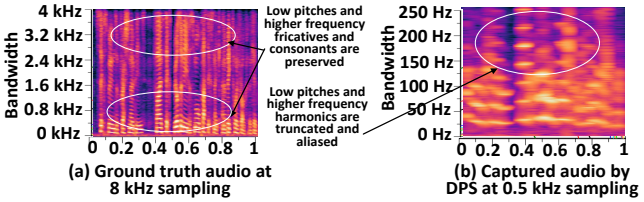

Figure 2. (Left) Low pitch and high frequency consonants are preserved. (Right) High-frequency fricatives get severely aliased.

in pressure sensors (see Fig. 2 (Right)). WaLi provides a means to eavesdrop speech in full intelligibility by enhancing the severely aliased signals from pressure sensors in HVACs. This will seriously hamper the confidentiality of safety-critical systems, as nowadays most critical infrastructures have some form of HVAC in their design.

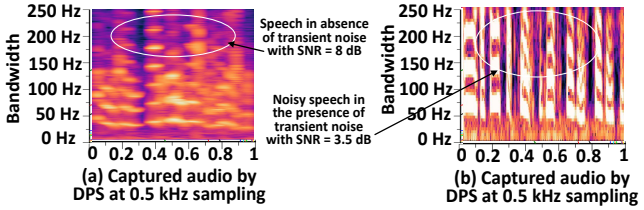

Figure 3. (Left) Clean speech in the absence of noise. (Right) Noisy speech in the presence of transient noise.

3.6. Transient Noise and Phase Reconstruction

DPSs operating within a low-pressure range of 0–200 Pa and at high sampling frequencies of 0.5-2 kHz, the impact of noise on DPSs becomes particularly significant. The presence of transient noise sources, such as HVAC fans, duct vibrations, physical shocks, or turbulent airflow, injects substantial noise into DPSs in the form of pressure spikes and fluctuating readings. The noisy readings from DPSs impact the intelligible speech reconstruction for successful eavesdropping. Although earlier speech reconstruction methods [10], [13], [34] focus on estimating the magnitude spectrum of missing higher frequencies, the phase spectrum is equally important [12] to produce natural and intelligible audio, particularly under transient noise conditions in real-world HVAC systems. Fig. 3 shows the impact of noise on pressure sensor data at 0.5 kHz sampling frequency. Transient noise corrupts the phase and magnitude of the speech signal in HVAC systems and change the Signal-to-Noise Ratio (SNR) from 8 dB to 3.5 dB.

Application Area	Sensor Location	Purpose
Differential control [40]	Wall and ceiling near doorways	Maintain pressurization in isolation rooms in hospitals
Controlled ventilation [41]	Inside duct or vent near occupants	Adjust airflow based on occupancy level
Lab pressure monitor [42]	Inside room or partition walls	Protect from contamination by maintaining air barriers
Residential IAQ [43]	Return air ducts in living areas	Balance pressure and airflow for occupant comfort

Table 3. DPSs PLACED CLOSE TO HUMANS CAN DO EAVESDROPPING.

4. Pilot Study

Here, we discuss the feasibility of eavesdropping on intelligible speech using DPSs in HVAC systems.

Proximity to sound sources and humans: DPSs are an integrated component of today's HVAC systems. However, to eavesdrop using DPSs, the DPSs should be close to humans or a sound source; otherwise, the feasibility of the acoustic eavesdropping using DPSs will be reduced. We identify a few locations in HVACs, summarized in Table 3, where DPSs are placed close to humans.

Table 3 indicates that DPSs placed close to humans are typically used for indoor air quality (IAQ) control, thermal comfort monitoring, or room-level pressure regulation in spaces like offices, hospitals, or cleanrooms. These sensors operate in low-pressure ranges (e.g., 0–250 Pa) and are often installed in room walls, near diffusers, or within ventilation grilles where human occupancy is high. To support this claim, we survey an anonymous cleanroom in our educational premises and point out the location of DPSs near the human occupancy that is shown in Fig. 4.

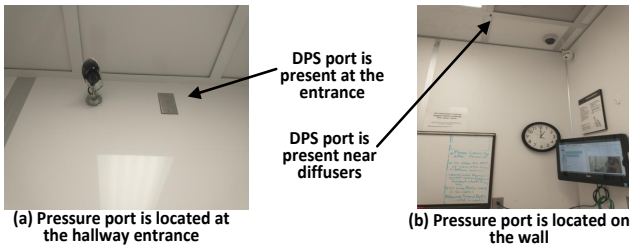

Figure 4. Pressure ports are located at the hallway entrance and inside rooms of a lab facility in human proximity.

Fig. 4 supports the fact that DPSs can be found in room entrances and inside rooms in practical HVAC systems in human proximity. Therefore, they can be a source of eavesdropping. The main challenges are that DPSs are noisy and use a sampling frequency of 0.5-2 kHz (refer to Table 2). Therefore, we propose WaLi to recover full intelligible speech from the aliased pressure sensor data by reconstructing both amplitude and phase of the speech signal.

A pilot experiment: We prepare a testbed as a pilot study to support our idea of eavesdropping using pressure sensors with an unrestricted vocabulary. We use an industry-used DPS from Setra with part# 26410R1WD11T1E [44]. It has two input ports. We connect two vinyl sampling tubes with inner diameters of 3/16" and 5/16" [45], a pressure pickup device with part# A-417A [46] with the two input ports (see Fig. 1). A volunteer speaks from 0.5 m distance from the pressure pickup device. We record the output data from the DPS sensor using an oscilloscope with a sampling frequency of 1 kHz. The recorded data from the DPS are shown in Fig. 5 with its matching ground-truth audio. The ground truth audio is sampled at 8 kHz.

In Fig. 5, the time domain waveform indicates a strong correlation between the ground truth audio and the pressure sensor data. We measure the intelligibility of the pressure sensor data by taking the ground truth as a reference using a metric named Perceptual Evaluation of Speech Quality

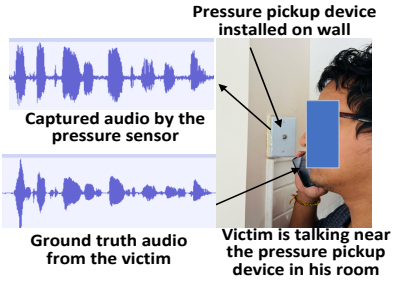

Figure 5. Victim is talking close to the pressure pickup device in his room.

(PESQ). PESQ is widely used to evaluate how a degraded speech signal compares to a reference (ground truth) signal, closely modeling human perception. The PESQ for the pressure data is 0.98. PESQ can have values between -0.5 to 4.5, where -0.5 means low perceptual quality and 4.5 means perceptually close to the ground truth. The 0.98 of PESQ for the pressure data indicates that the DPS captures low-grade perceptual audio, which is not intelligible (to be intelligible PESQ value of more than 1.4 is required). Therefore, it indicates that DPS can be used to eavesdrop by an attacker if the attacker can reconstruct intelligible audio from the severely aliased and low-grade perceptual data of DPSs.

5. Threat Model

We discuss the components of the threat model below.

a) Attacker's goal: The attacker can eavesdrop on the conversation with unrestricted vocabulary using HVAC systems where DPSs are located close to human occupancy. In addition to eavesdropping a natural conversation, the attacker can also understand how many people are present inside of a room/facility, how many of them are males and females, what types of sound sources are present, and from the sound sources, the attacker can also identify what types of activities are going on inside the room/facility. The attacker collects pressure data from DPSs and reconstructs intelligible audio from the noisy and aliased pressure data using our WaLi. WaLi uses complex-valued models to reconstruct both magnitude and phase to generate intelligible audio for at least 0.5 kHz sampling frequency.

b) Attacker's access level: Prior works with the capability to spy on natural conversations use malicious apps to access inertial sensors [10], [13] to collect data for eavesdropping. These attacks require the installation of such apps to initiate the eavesdropping. In contrast, our attack model exploits the normal behavior of HVACs without installing additional software on HVACs. Access to collect pressure data from DPSs is possible in the following scenarios.

First, an attacker disguised as a malicious employee or maintenance person can access pressure data from the Building Management System (BMS) software dashboard, as in modern buildings, pressure sensors are integrated into the BMS using standard protocols, such as Modbus TCP, KNX, and LonWorks.

Second, in many cases, the BMS is handled by a third-party vendor, HVAC contractors, or system integrator, especially in commercial buildings, hospitals, labs, and large campuses. These third-party vendors manage complex integration, security and control, compliance, and 24/7 monitoring. In many cases, authorities often outsource teams to provide continuous support and alert handling. An attacker disguised as one of these third-party vendors can access sensitive pressure data via a web-based interface, historical logs, or an Open Platform Communication (OPC) server.

Third, some commercial HVAC units (e.g., rooftop units, air handling units, variable air volume units) have onboard controllers and internal sensors. Service technicians often access these units for diagnostics in case of failure and troubleshooting. An attacker disguised as one of the technicians can access the pressure data through the control panel and diagnostic port via network protocols.

Fourth is a potential supply chain attack, which has been rumored to occur in the past [47]–[50] and was recently demonstrated to be feasible [51], involving tampering with HVAC systems during delivery or installation. In such an attack, a competitor could intercept the equipment and modify DPSs by embedding a data collection device with a radio transmitter inside, and then allow the delivery or installation to proceed as planned to the commercial facility.

c) Non-invasive attack: Our eavesdropping attack using the HVAC system is non-invasive as it is performed without making direct physical contact with the target DPS. The attacker doesn't need to directly open or physically touch DPSs. However, we expect that attackers can examine the behavior of a similar sensor subjected to acoustic impacts before initiating an actual attack.

d) Attacker's resources and cost: We assume that the attacker knows how the HVAC system works and how the data can be collected from pressure sensors.

e) Assumption: Considering the practicality of the attack, we assume that the attacker has sufficient pressure sensor data collected from the victim and does not have matching ground truth audio from the victim. WaLi can reconstruct the audio even if the speaker is different. This makes our attack model flexible compared to prior works [13], [34], where prior works assume that clean ground-truth speech is available via a microphone or a phishing call.

6. WaLi ARCHITECTURE DESIGN

WaLi is designed to achieve the following objectives:

a) WaLi processes the input pressure sensor data in the complex-valued T-F spectrum and enhances both the amplitude and phase of the pressure sensor data to reconstruct intelligible speech from the aliased pressure sensor data.

b) WaLi handles transient interference in noisy conditions of the pressure sensor data using joint magnitude and phase reconstruction using a complex-valued network.

c) WaLi reconstructs speech from the victim's pressure sensor data without using the ground truth audio of the victim.

WaLi adopts a complex-valued U-Net model and processes the incoming low-resolution and noisy pressure data using the complex-valued T-F spectrogram. We will first

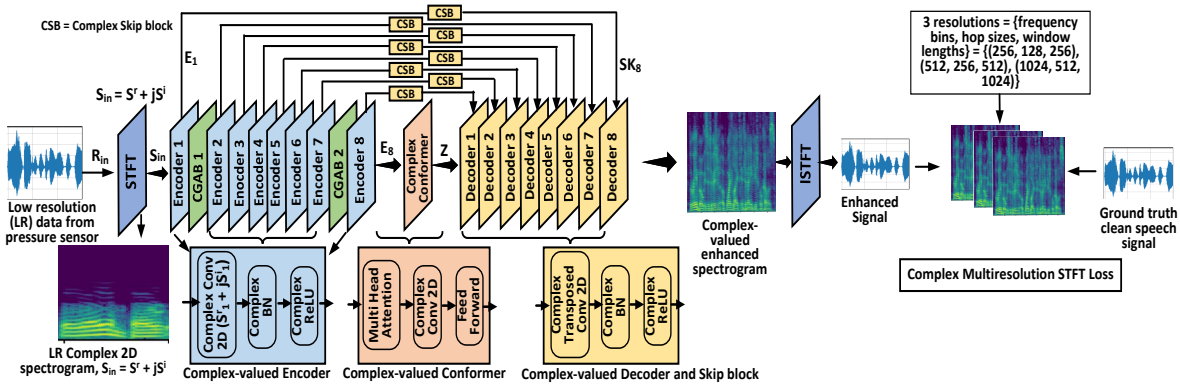

Figure 6. WaLi has complex-valued encoders, decoders, complex-valued skip blocks, CGAB, and complex multireoslution STFT loss.

explain the reasons behind the use of the complex-valued T-F spectrogram and network.

Why the complex-valued spectrogram and network: A complex-valued spectrogram is a T-F representation of a signal that contains both magnitude and phase information, not just the energy or amplitude as in a standard real-valued spectrogram. A complex-valued T-F spectrogram is typically generated by the Short-Time Fourier Transform (STFT) and then outputs a complex value $S(t, f)$ at each time-frequency bin, following Eq. 1.

$$S(t, f) = A(t, f) * e^{j\phi(t,f)} \tag{1}$$

where $A(t, f)$ is the magnitude and $\phi(t, f)$ is the phase of the spectrum. Prior works [10], [13] only use real-valued T-F spectrograms. Therefore, they cannot predict phase information and hence, need a vocoder to generate audio from real-valued spectrograms. Moreover, real-valued T-F spectrogram-based methods typically use Mean Square Error (MSE) loss for magnitudes and cannot use phase-reconstruction loss functions to improve the phase quality while reconstructing speech from the low-resolution pressure sensor data. Therefore, motivated by the fact that phase plays a crucial role in speech enhancement [12], WaLi adopts a joint reconstruction of magnitudes and phases. As complex-valued spectrograms have both magnitude and phase information (refer to Eqn. 1), WaLi takes complex-valued T-F spectrograms as input and generates complex-valued T-F spectrograms at its output. To process the complex-valued T-F spectrograms, we design WaLi as a complex-valued network having U-Net as its backbone.

WaLi architecture: The detailed architecture of the proposed WaLi is shown in Fig. 6. The network consists of four main components: **(i)** a total of 16 full complex-valued encoder-decoder blocks (i.e., 8 encoders and 8 decoders), **(ii)** complex-valued skip blocks, **(iii)** complex-valued conformer in the bottleneck layer, and (iv) complex-valued global attention blocks - CGAB. The complex domain processing by our proposed WaLi has the potential to recover clean phases from the noisy pressure sensor and to reconstruct magnitudes from the aliased spectrum (i.e., pressure sensor is sampled at a minimum of 0.5 kHz).

6.1. Complex-Valued Encoders

The complex-valued spectrogram of the pressure sensor data is transformed to real and imaginary components, and then the real and imaginary components are combined into a complex-valued tensor, which serves as input to the first complex encoder block (E1). Formally, the input low-resolution pressure data R_{in} is first transformed into STFT spectrogram, denoted by S_{in} in Fig. 6. Here, $S_{in} (= S^r + jS^i) \in \mathbb{C}^{F \times T}$ is a complex-valued spectrogram, where F denotes the number of frequency bins and T denotes the number of time frames.

Typically, each encoder in a U-Net extracts hierarchical features from the input or previous layers by downsampling the features using a convolution operation. WaLi adopts complex-valued encoders, which is built upon complex-valued convolution to ensure successive extraction of both magnitude and phase from the complex-valued T-F spectrogram. Complex convolution is the key difference between a complex-valued encoder and a real-valued encoder. Formally, the complex-valued T-F spectrogram S_{in} is fed into 2D complex convolution layers of the first encoder to produce feature $S_1 \in \mathbb{C}^{F \times T \times C}$, where C is the number of channels. If a complex kernel is denoted by $W = W_r + jW_i$, the complex convolution is defined by Eqn. 2.

$$\begin{aligned} S_1^r &= W_r * S_{in}^r - W_i * S_{in}^i + b_r, \\ S_1^i &= W_r * S_{in}^i - W_i * S_{in}^r + b_i, \end{aligned} \tag{2}$$

where $*$ denotes the convolution, S_1^r & S_1^i are real and imaginary parts of S_1, and b_r & b_i are bias terms.

The convolution output is then normalized using Complex Batch Normalization (CBN) for stable training. CBN is an extension of traditional batch normalization to complex-valued neural networks, where inputs, weights, or features are complex numbers (i.e., having both real and imaginary parts). Instead of treating real and imaginary parts independently, CBN treats them as a 2D vector and normalizes using a 2×2 covariance matrix using Eqn. 3.

$$\Sigma_{S_1} = \begin{bmatrix} \text{Var}(S_1^r) & \text{Cov}(S_1^r, S_1^i) \\ \text{Cov}(S_1^i, S_1^r) & \text{Var}(S_1^i) \end{bmatrix} \tag{3}$$

where Σ_{S_1} is the covariance matrix of S_1. Then the CBN is calculated by Eqn. 4.

$$\text{CBN} = \Sigma_{S_1}^{-\frac{1}{2}}(S_1 - \mu) \tag{4}$$

where $\Sigma_{S_1}^{-\frac{1}{2}}$ is the inverse square root matrix of Σ_{S_1} used to decorrelate and normalize the input S_1, and μ is the complex mean vector of S_1.

Next, the output from CBN passes through a complex ReLU activation for adding nonlinearity. Formally, first encoder output, denoted by E_1 can be expressed by Eqn. 5.

$$E_1 = \text{Complex ReLU}(\text{CBN}(S_1^r + jS_1^i)) \tag{5}$$

The output of the first encoder E_1 is given as input to the 2nd encoder, and 2nd encoder's output to the third encoder, and so on. Every complex encoder has a similar complex 2D convolution (Eqn. 2), CBN (Eqn. 4), and complex ReLU activation (Eqn. 5).

6.2. Complex-Valued Conformer in Bottlenecks

We use complex-valued conformers in the bottleneck layer of our WaLi to capture local and global dependencies *among consecutive spectrograms*. A conformer [9] combines the strengths of CNNs and Transformers. Traditional Transformers are great at capturing long-term dependencies, but struggle with local patterns that exist in each phoneme. Convolutional layers, on the other hand, excel at local context but lack long-range modeling. The conformer blends both by stacking self-attention and convolutional modules in each layer, resulting in a highly expressive and efficient model for sequential input. Our complex-valued conformer optimally balances global context with fine-grained local information in phase and amplitude domains for the successful reconstruction of intelligible audio from pressure data.

The output of the last (eighth) encoder block E_8 is expressed as $E_8 \in \mathbb{R}^{B\times C\times F\times T}$ and fed as input into the conformer block. Here, B denotes the batch size, C the number of channels, F the spectral dimension, and T the temporal dimension. Next, we rearrange E_8 using a permutation and reshaping so that the conformer processes each channel independently. For a given channel, the complex-valued conformer comprises complex multi-head self-attention (MHSA), complex feed-forward (FF), and complex convolutional modules, inspired by [9]. The complex MHSA uses complex-valued queries, keys, and values over the real and imaginary parts. The complex FF module uses complex linear layers, complex ReLU activation, and CBN.

6.3. Complex-Valued Decoder

The output from the bottleneck layer, denoted by $\mathbf{Z}$ (i.e., latent space), is given as input to the first decoder. The decoder reconstructs the T-F representation from the latent space by employing a complex transposed convolution by upsampling by a factor of two at each decoder block. For a latent complex tensor $\mathbf{Z} = \mathbf{Z}_r + j\,\mathbf{Z}_i$, the transposed convolution is formulated as:

$$\begin{aligned}\tilde{\mathbf{Y}}_1^r &= \mathbf{W}_r^T * \mathbf{Z}_r - \mathbf{W}_i^T * \mathbf{Z}_i + \mathbf{b}_r,\\ \tilde{\mathbf{Y}}_1^i &= \mathbf{W}_r^T * \mathbf{Z}_i + \mathbf{W}_i^T * \mathbf{Z}_r + \mathbf{b}_i,\end{aligned} \tag{6}$$

where $\tilde{\mathbf{Y}}_1^r$ & $\tilde{\mathbf{Y}}_1^i$ are real and imaginary parts of $\tilde{\mathbf{Y}}_1$ (i.e., output after the transposed convolution in the first decoder), and b_r & b_i are bias terms. $\tilde{\mathbf{Y}}_1$ is then normalized by CBN and activated by a complex ReLU activation as Eqn. 7.

$$D_1 = \text{Complex ReLU}\Big(\text{CBN}\big(\tilde{\mathbf{Y}}_1^r + j\,\tilde{\mathbf{Y}}_1^i\big)\Big) \tag{7}$$

where D_1 is the output of the first decoder, which is given as input to the 2nd encoder, and so on. Every complex decoder has a similar complex 2D convolution (Eqn. 6), CBN (Eqn. 4), and complex ReLU activation (Eqn. 7).

6.4. Complex Skip Block

A skip connection in our proposed WaLi passes high-dimensional features from complex-valued encoders to the appropriate decoders. This enables the model to preserve the spatial features, which may lost during the down-sampling operation, and guides the network to propagate from encoders to decoders. WaLi implements skip blocks in complex domains, inspired by [52], to enable the proper flow of complex features from the encoder's output to decoders. Each complex skip block applies a complex convolution on the encoder output, followed by a CBN and a complex ReLU activation. Formally, the complex skip block's output, denoted by SK_n is (i.e, n is 1 to 8):

$$SK_n = \text{Complex ReLU}(\text{CBN }(\text{Complex Conv}(E_n))) \tag{8}$$

where Complex Conv is implemented following Eqn. 2, and E_n is the output from the nth encoder (i.e, n is 1 to 8).

6.5. Complex Global Attention Block (CGAB)

Long-range correlations exist along both the time and the frequency axes in a complex-valued T-F spectrogram. As the audio signal embedded in the pressure sensor data is a time series signal, inter-phoneme correlations exist along the time axis. Moreover, harmonic correlations also exist among pitch and formants along the frequency axis. As a convolution kernel is limited by its receptive field, standard convolutions cannot capture global correlations on the time and frequency axes in a complex-valued T-F spectrogram. Therefore, we propose the Complex Global Attention Block (CGAB) to capture long-range correlation from the T-F axes. Please note that Frequency Transformation Blocks (FTBs) used in [10] published in USENIX, 2024, don't work along both the T-F axes.

The detailed implementation of our proposed CGAB is shown in Fig. 7. CGAB provides attention to the time and frequency axes of a complex-valued T-F spectrogram by following two steps:

Step 1 - Reshaping along the T-F axes: The output E_n from the encoder is decomposed in 2 steps by CGAB into two tensors: one along the time axis and another along the frequency axis. Formally, E_n, which has a feature dimension of $C \times F \times T$, is given at the input of CGAB. At the first stage of reshaping, E_n parallelly reshaped into $C.T$

vectors with dimension $C \cdot T \times F$ and into $C.F$ vectors with dimension $C \cdot F \times T$. This reshaping is done using 2D complex convolution, CBN, and complex ReLU activation, followed by vector reshaping. In the second stage of reshaping, $C \cdot T \times F$ is reshaped into $1 \times T \times F$ and $C \cdot F \times T$ is reshaped into $1 \times F \times T$ using 1D complex convolution, CBN, complex ReLU activation followed by vector reshaping. The tensors with dimension $1 \times F \times T$ capture the global harmonic correlation along the frequency axis and $1 \times T \times F$ capture the global inter-phoneme correlation along the time axis. The captured features along the T-F axes and the original features from E_n are point-wise multiplied together to generate a combined feature map with a dimension of $C \times T \times F$ and $C \times F \times T$ along T and F axes, respectively. This point-wise multiplication captures the inter-channel relationship between the encoder's output and complex time and frequency axes.

Step 2 - Global attention along the T-F axes: It is possible to treat the spectrogram as a 2D image and learn the correlations between every two pixels in the 2D image. However, this is computationally too costly and is not realistic. On the other hand, ideally, we can use self-attention [53] to learn the attention map from two consecutive complex T-F spectrograms. But this might not be necessary. Because, on the time axis in each T-F spectrogram, when calculating SNR, the same set of parameters in the recursive relation are used, which suggests that temporal correlation is time-invariant among consecutive spectrograms. Moreover, harmonic correlations are independent in the consecutive spectrograms [54].

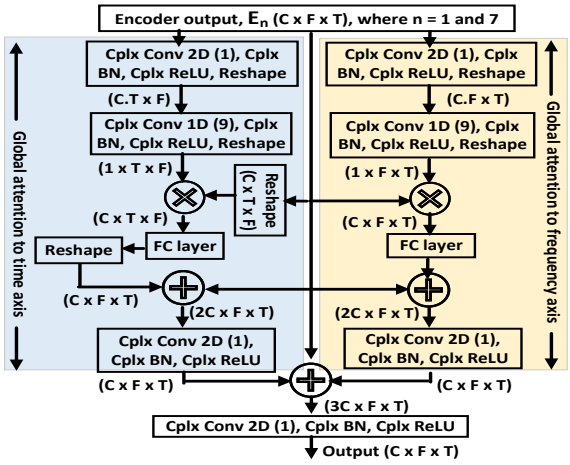

Figure 7. CGAB captures global time and frequency correlations. Here, Cplx = complex, and Conv = convolution.

Based on this understanding, we propose a self-attention technique along the T-F axes within each spectrogram, without considering correlations among consecutive spectrograms. Specifically, attention on frequency and time axes are implemented by two separate fully connected (FC) layers. Along the time path, the input and output dimensions of FC layers are $C \times T \times F$. Along the frequency path, the input and output dimensions of FC layers are $C \times F \times T$. FC layer learns weights from complex T-F spectrograms and technically is different from the self-attention [53] operation. To capture interchannel relationships among the input E_n and output of FC layers, concatenation happens, followed by 2D complex convolutions, CBN, and complex ReLU activation. Finally, the learned weights from the T-F axes are concatenated together to form a unified tensor, which holds joint information on the T-F global correlations from each spectrogram.

We use only two CGABs - one after the 1st encoder, and another after the 7th encoder.

6.6. Complex Multiresolution STFT Loss

In this work, we use *Complex Multi-Resolution STFT Loss* that measures the differences between clean and enhanced signals in the complex T-F domain (i.e., amplitude and phase). Unlike standard magnitude-only loss [55], we propose a complex multiresolution STFT loss that separately evaluates the loss on real and imaginary parts of STFT across multiple resolutions, thereby capturing fine and coarse spectral details in complex domains. At first, spectral convergence loss L_{SC} [56] and log STFT magnitude loss L_{mag} [56] are calculated on both real and imaginary parts of the enhanced signal and ground truth speech data. Let us define the L_{SC} and L_{mag} calculated on real and imaginary STFT data as $\{L^r_{SC}, L^i_{SC}\}$ and $\{L^r_{mag}, L^i_{mag}\}$, respectively. Assuming we have S different STFT resolutions, we aggregate the losses by averaging over the resolutions for both real and imaginary parts following Eqn. 9.

$$
\begin{aligned}
L_{\mathrm{r}} &= \frac{1}{S}\sum_{s=1}^{S}\left(L^r_{\mathrm{SC}} + L^r_{\mathrm{mag}}\right) \\
L_{\mathrm{i}} &= \frac{1}{S}\sum_{s=1}^{S}\left(L^i_{\mathrm{SC}} + L^i_{\mathrm{mag}}\right)
\end{aligned}
\tag{9}
$$

We use three resolutions, such as frequency bins = [256, 512, 1024], hop sizes = [128, 256, 512], and window lengths = [256, 512, 1024] to calculate L_r and L_i. The overall complex multiresolution STFT loss, $L^{complex}_{STFT}$ is the sum of L_r and L_i. The joint optimization in complex T-F domain on magnitue and phase improves the *quality and intelligibility* of the reconstructed speech from pressure sensor data.

7. Evaluation

7.1. Pressure Data and Audio Corpus

We use VCTK (version 0.92) [57], a multi-speaker English corpus for evaluation. The dataset contains ∼43600 audio traces sampled at 48 kHz from 110 individuals (i.e., balanced for males and females). We downsample the dataset to 8 kHz for evaluation. Each audio clip has a duration ranging from 2s to 7s. We standardize all audio clips to 4s by either zero-padding or silence trimming. Following [58], only the mic1 data are used for the experiments, and p280 and p315 are omitted due to technical issues. We play the audio at 60 dB through a loudspeaker and collect the corresponding pressure data via a DPS sensor (SDP810-125PA). The speaker is placed at a 5 cm distance from one of the pressure ports to play the audio clips. To prevent

	500 Hz to 8 kHz					1 kHz to 8 kHz					2 kHz to 8 kHz				
	L↓	N↑	S↑	P↑	ST↑	L↓	N↑	S↑	P↑	ST↑	L↓	N↑	S↑	P↑	ST↑
Raw pressure data	3.45	0.84	6.24	0.87	0.71	3.15	0.95	8.54	0.98	0.75	2.95	1.27	9.87	1.14	0.79
Reconstructed	1.24	1.78	8.78	1.51	0.75	1.16	1.95	10.04	1.61	0.79	1.01	2.17	11.38	1.95	0.82

Table 4. EVALUATING WALI FOR RECONSTRUCTING INTELLIGIBLE AUDIO FROM PRESSURE SENSOR DATA FOR 500 HZ, 1 KHZ, AND 2 KHZ SAMPLING FREQUENCIES TO 8 KHZ UPSAMPLING FOR 70 DB AUDIO. HERE, L = LSD, N = NISQA-MOS, S = SI-SDR, P = PESQ, AND ST = STOI.

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. Signals from the sensor are collected as .edf and next converted to .wav format using a custom Python script. The sampling frequency of the pressure sensor is varied in between 500 Hz and 2 kHz. We use sinc interpolation to upsample before the audio reconstruction to ensure that the system input and output have the same shape. Note that in a real case, the speech contents may be different from the spoken ones during the attack phase. Thus, for testing purposes, we use 11 different speakers, who are not present in the training. The models are trained offline with an NVIDIA 4090 GPU and deployed on a desktop for speech recovery.

7.2. Evaluation Metrics

We comprehensively evaluate WaLi using the following five different metrics.

a) Log-Spectral Distance (LSD): LSD measures the difference between the log-magnitude spectra of a ground truth signal and a reconstructed one from the pressure data.

b) Non-Intrusive Speech Quality Assessment - Mean Opinion Score (NISQA-MOS): NISQA-MOS is used to estimate the perceived quality of reconstructed audio without needing a reference signal. It predicts a MOS similar to human ratings with values between 0 to 5.

c) Scale-Invariant Signal-to-Distortion Ratio (SI-SDR): SI-SDR measures how close a reconstructed signal is to a clean target signal, ignoring any difference in gain.

d) Perceptual Evaluation of Speech Quality (PESQ): PESQ is widely used to evaluate how a reconstructed signal compares to a reference (clean) signal, closely modeling human perception. PESQ has values between -0.5 to 4.5

e) Short-Time Objective Intelligibility (STOI): STOI is specifically designed to predict how understandable speech is to human listeners, especially in the presence of noise. STOI has values between 0 to 1.

7.3. Overall Performance

To intuitively observe the performance of WaLi, we compared among ground truth speech, raw data from the pressure sensor, and reconstructed speech by WaLi, shown in Fig. 8. We can see that the frequency components above 250 Hz are absent in the pressure sensor data (see Fig. 8(b)) as we use a sampling frequency of 0.5 kHz in DPS at this time. The ground truth audio is sampled at 8 kHz (4 kHz bandwidth), shown in Fig. 8(a). From Fig. 8(c), we can see that the high-frequency components are reconstructed up to 4 kHz by WaLi. The reconstructed spectrogram shows a high similarity to the ground truth one.

Moreover, to quantitatively measure WaLi's performance, we vary the sampling frequency of the pressure sensor on three scales, such as 500 Hz, 1 kHz, and 2 kHz. We reconstruct the audio to 8 kHz upsampling, to evaluate WaLi for high upsampling ratios [55]. The detailed comparison is shown in Table 4, which shows the average value of the metrics over all the test speakers. The reconstructed audio by WaLi achieves a 2.78x improvement in LSD (i.e., 3.45 vs 1.24), 2.11x increase in NISQA-MOS (i.e., 0.84 vs 1.78), 1.4x increase in SI-SDR (i.e., 6.24 vs 8.78), 1.73x increase in PESQ (i.e., 0.87 vs 1.51), and 1.05x increase in STOI (i.e., 0.71 vs 0.75) between the raw pressure data and the reconstructed audio for 500 Hz sampling frequency.

Please also note that the metrics improve for 2-8 kHz compared to 1-8 kHz and 0.5-8 kHz because for 2-8 kHz the upsampling ratio is around 4, for 1-8 kHz the upsampling ratio is 8, and for 0.5-8 kHz the upsampling ratio is 16. The lower the upsampling ratio, the greater the performance gain for the reconstruction model.

Individual test speaker: In addition to average values, Fig. 9 (Left) shows the individual values of each test speaker for LSD and NISQA-MOS. The individual values indicate that WaLi performs consistently for all unseen test speakers. This indicates that WaLi can effectively reconstruct the high-frequency components and suppress artifacts, and improve the speech intelligibility significantly.

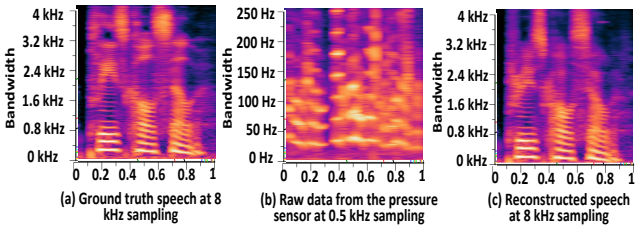

Figure 8. The reconstructed spectrogram from the pressure sensor data in (c) shows a high similarity to the ground truth one in (a).

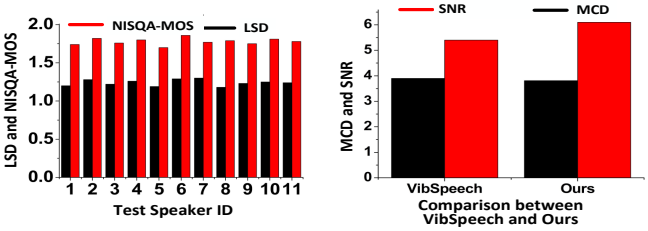

Figure 9. (Left) LSD and NISQA-MOS for individual test speakers. (Right) Comparison between WaLi and VibSpeech [10] in terms of MCD and SNR.

Comparison with VibSpeech [10]: We also compare between WaLi and VibSpeech [10] in terms of SNR and Mel-Ceptral Distortion (MCD) [10] in Fig. 9 (Right), which indicates that WaLi performs better (MCD = 3.81 and SNR = 6.1) than VibSpeech (MCD = 3.9 and SNR = 5.4).

Please note that WaLi is tested with different speakers, which are not seen by the model during the training time. Moreover, the VCTK dataset consists of everyday-use audio clips, which are not limited to a certain specialized vocabulary such as [34]. Therefore, WaLi can recover intelligible audio with unrestricted vocabulary irrespective of specific applications.

7.4. Phase Reconstruction in Non-Noisy Condition

As WaLi is a complex-valued model, it can directly reconstruct phase information from the low-resolution pressure sensor data. Previous works used the Griffin-Lim algorithm [13] as a vocoder to reconstruct the phase. To compare the performance of our phase reconstruction in the non-noisy condition, we reconstruct audio from only the magnitude of our model with the Griffin-Lim algorithm, and then compare it with the reconstructed audio from our model. A comparison summary is shown in Table 5 for a 500 Hz to 8 kHz reconstruction.

Phase	L↓	N↑	S↑	P↑	ST↑
Griffin-Lim	1.26	1.71	7.65	1.47	0.74
WaLi	1.24	1.78	8.78	1.51	0.75

Table 5. EVALUATING WALI FOR RECONSTRUCTING PHASE AT NON-NOISY CONDITION. HERE, L = LSD, N = NISQA-MOS, S = SI-SDR, P = PESQ, AND ST = STOI.

Table 5 indicates that our proposed complex-valued WaLi is similar/slightly better in phase reconstruction compared to the Griffin-Lim algorithm in the non-noisy condition. However, the performance of WaLi is much better under noisy conditions that is discussed in the next section.

7.5. Phase Reconstruction in Noisy Conditions

DPSs, operating within a low-pressure range of 0-200 Pa and at high sampling frequencies of 0.5-2 kHz, are sensitive to transient noise sources (see Section 3.6). Therefore, clean phase reconstruction becomes particularly significant in noisy conditions. To test the performance of phase reconstruction under noisy conditions, we use the DCASE challenge dataset [59] to add transient mechanical noise, such as fans, valves, pump noise, air leaks, tool drops, etc., to the clean VCTK dataset. Noise is randomly added within a range of -7 to 40 dB. Then, we retrain WaLi with a {noisy, clean} pair dataset. A comparison summary is shown in Table 6 for a 500 Hz to 8 kHz reconstruction in noisy conditions. Table 6 indicates that our proposed complex-valued WaLi is much better in phase reconstruction compared to the Griffin-Lim algorithm in transient noisy condition. The reason behind this is that WaLi handles both magnitude and phase jointly and recovers both clean magnitude and phase under noisy conditions from the clean reference signal using a complex multi-resolution loss function. On the other hand, Griffin-Lim-based models [13] do not learn clean phase reconstruction as these models are magnitude-only models and cannot handle phase jointly.

7.6. Impact of Sound Volume

We use a loudspeaker to simulate a speaker and vary the volume of the loudspeaker from 60 dB to 85 dB with a 5 dB increment. We inject the audio into the pressure sensor ports through the 1m long sampling tube from a 5 cm distance. The result is shown in Fig. 10 (Left) for LSD and NISQA-MOS for 500 Hz to 8 kHz upsampling. We can see that the LSD and NISQA-MOS improve rapidly between 60 dB and 70 dB. However, after 70 dB, LSD and NISQA-MOS improve steadily rather than a sharp improvement. The reason behind this is that the pressure sensor's sensitivity increases with the increase of volume up to 70 dB. After 70 dB, the increase in volume does not drastically change the low-resolution frequency components. Therefore, the improvement is not sharp after 70 dB; rather, a steady ascent occurs. The LSD is close to 1, and NISQA-MOS is close to 2 for a wide volume range from normal conversation (i.e., 60 dB) to loud speech. It indicates that WaLi is capable to reconstruct intelligible audio from normal conversation (i.e., 60 dB) to loud speech in real-world scenario.

Phase	L↓	N↑	S↑	P↑	ST↑
Griffin-Lim	1.38	1.43	6.17	1.21	0.70
WaLi	1.26	1.71	8.14	1.47	0.74

Table 6. EVALUATING WALI FOR RECONSTRUCTING PHASE AT TRANSIENT NOISY CONDITION. HERE, L = LSD, N = NISQA-MOS, S = SI-SDR, P = PESQ, AND ST = STOI.

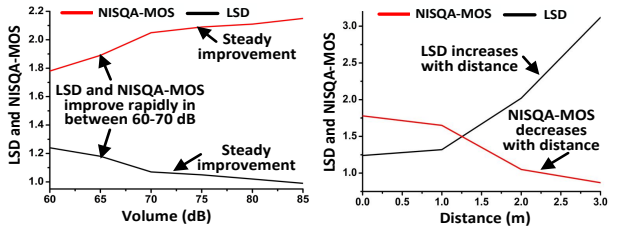

Figure 10. (Left) Relationship with speaker volume and (Right) speaker distance with the reconstruction performance.

7.7. Impact of Speaker Distance

We vary the distance of a speaker from 0 m (touching the pressure sensor's input port) to 3 m from the target pressure sensor. The result is shown in Fig. 10 (Right) for LSD and NISQA-MOS for 500 Hz to 8 kHz upsampling for 60 dB audio. In acoustics, the energy of a sound wave radiating from a point source decreases as the distance increases following the inverse-proportional law [60]. Therefore, the improvement in LSD and NISQA-MOS decreases with increasing distance from the audio source to the pressure sensor. It is also clear from Fig. 10 (Right) that WaLi performs well up to 1 m distance. After 1 m, the reconstructed audio has severely degraded intelligibility.

7.8. Impact of Speaker Orientation

We use a loudspeaker to simulate a speaker and place the loudspeaker at 0°(in front of the loudspeaker), 90°(right of the loudspeaker), and 180°(left of the loudspeaker) of the pressure sensor. The loudspeaker is 5 cm away from the pressure sensor and generates a sound of 60 dB. The results are shown in Fig. 11 for LSD and NISQA-MOS for 500 Hz to 8 kHz upsampling. The LSD and NISQA-MOS are higher at 0° orientation, as at 0°, the audio is directed

to the pressure sensor ports with minimum loss, potentially creating a strong vibration. Performance is the same at both 90°and 180°orientations, but lower compared to 0°.

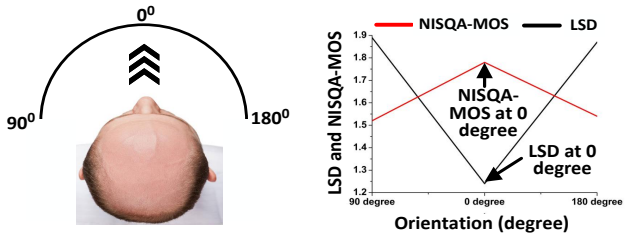

Figure 11. Impact of orientation of the pressure sensor on the reconstruction quality of audio from the pressure sensor data.

7.9. Fine-Tuning on Victim's Ground Truth Audio

Please note that WaLi can reconstruct the audio even if the speaker is different as WaLi is tested with different speakers, which are not seen by the model during the training time. This makes our attack model more flexible compared to prior works [10], [13], where prior works assume that clean ground-truth speech of the victim is available via a microphone or a phishing call. However, if the attacker could manage the victim's {ground truth, pressure sensor data} pair, the accuracy of the models would be greatly improved by WaLi. To evaluate this, we incorporate a small amount of {ground truth, pressure sensor data} pair from the victim to fine-tune the pre-trained network. As shown in Table 7, the performance gradually improves if the amount of data from the victim increases. For example, with just one minute of data, we observe significant performance gains, and after five minutes of additional data, the performance gain is minimal. This proves that WaLi can be fine-tuned with the victim's audio, if possible, in a stronger assumption of the attack model. Table 7 is for 60 dB audio for 500 Hz to 8 kHz upsampling.

Fine tuning data size	L↓	N↑	S↑	P↑	ST↑
0 min (before fine-tuning)	1.24	1.78	8.78	1.51	0.75
1 min	1.03	1.89	10.23	1.72	0.77
5 min	0.98	2.04	11.45	1.91	0.79
7 min	0.97	2.07	11.53	1.92	0.79

Table 7. EVALUATING WALI AFTER FINE-TUNING ON VICTIM SPEAKER'S GROUND TRUTH AUDIO.

7.10. Testing With Less Than 500 Hz Sampling

We test WaLi while reconstructing speech from 250 Hz sampling frequency of the pressure sensor data, shown in Fig. 12. As only a few pitch frequencies are present within 125 Hz bandwidth of the 250-Hz-sampled data, the reconstructed speech in Fig. 12(c) is not intelligible with a NISQA-MOS value of 0.48. Therefore, the attacker should have access to DPSs that have at least 500 Hz sampling for proper intelligibility. However, the 250-Hz-sampled data may be used for hot word detection (i.e., our future work).

8. Limitation and Discussion

Feasibility of the Attack: Sections 7.6, 7.7, and 7.8 indicate that the distance, orientation, and volume of the target

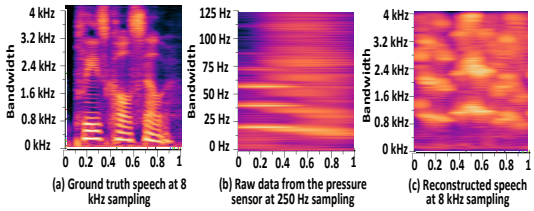

Figure 12. Reconstructed speech at (c) from 250 Hz sampling frequency at (b) has severely degraded intelligibility.

speaker can restrict the effectiveness of the eavesdropping attack using DPSs. For a successful attack, the target victim should be at least 1 m close to the pressure sensor, have a pressure level of 60 dB or more, and should be closer to 0° orientation. This situation exists in numerous scenarios (see Section 4), where DPSs are often installed in room walls, near diffusers, or within ventilation grilles, where human occupancy is high. If a target victim speaker unknowingly comes close to the pressure ports and continues a private conversation, the attacker can easily eavesdrop on the victim's conversation using our WaLi.

Sampling frequency limit: WaLi achieves intelligible speech recovery with unrestricted vocabulary when the sampling frequency of the pressure sensor is 500 Hz or higher. If the sampling frequency of the DPSs is less than 500 Hz, the bandwidth of the speech is lower than 250 Hz, which is extremely narrow for an intelligible speech reconstruction. Therefore, the reconstruction quality degrades significantly at a sampling frequency lower than 500 Hz. A possible direction is to find the inner relationship between the pitch and the low-order harmonics to make WaLi work for a sampling frequency lower than 500 Hz.

9. Countermeasures

Audio damping: The simplest method of preventing eavesdropping using DPSs is to damp the audio using a physical damping device at the input of the pressure ports. The damping device should be such that it does not affect the normal pressure measurement of the pressure sensors. However, it should dampen the audio significantly. One solution is to use a pressure sampling tube longer than 1 m to dampen the audio significantly without impacting the normal pressure measurement. Another solution is to enclose the pressure pickup device using a box-like enclosure filled with sound-damping foam to dampen the audio. Both of these countermeasures are cheap and easy to adopt in a critical infrastructure where confidentiality might be an issue.

Reducing sampling frequency of DPSs: Another intuitive solution is to reduce the sampling frequency of pressure sensors in HVAC systems. However, we need to keep in mind that often reducing sampling frequency is not possible in some specific applications, where small pressure changes must be detected accurately and quickly for a stable control process. Therefore, for these closely controlled applications, an audio damping device should be used instead of reducing the sampling frequency of pressure sensors.

10. Conclusion

In this paper, we expose a new speech threat that adversaries can recover intelligible audio up to 8 kHz from severely aliased pressure sensor data, having a sampling frequency greater than 500 Hz. If a target victim speaker unknowingly comes close to the pressure ports and continues a private conversation, the attacker can eavesdrop on the victim's conversation using our WaLi. WaLi can reconstruct audio even if the speaker is different and are not seen by the model during the training time. This makes our attack model more flexible compared to previous work. Using our WaLi, an attacker can secretly listen natural conversation behind the wall which is least unexpected. Moreover, we comprehensively evaluate WaLi using five metrics that have not been done before for speech eavesdropping tasks.

References

[1] S. AG, "Qbm2030-5 differential pressure sensor," https://hit.sbt.siemens.com/RWD/app.aspx?action=ShowProduct&key=S55720-S245&module=Catalog, accessed: May 23, 2025.

[2] H. I. Inc., "Abp2 series board mount pressure sensors," https://prod-edam.honeywell.com/content/dam/honeywell-edam/sps/siot/en-us/products/sensors/pressure-sensors/board-mount-pressure-sensors/basic-abp2-series/documents/sps-siot-abp2-series-datasheet-32350268-en.pdf, accessed: May 23, 2025.

[3] CFSensor, "Xgzp6887d pressure sensor," https://www.endrich.com/Datenbl%C3%A4tter/Sensoren/XGZP6887D-Pressure-Sensor-V2.5.pdf, accessed: May 23, 2025.

[4] R. Ghosh, S. Baldi, I. Michailidis, and R. Babuska, "Multi-rate control for energy-efficient operation of hvac systems," *Journal of Building Performance Simulation*, vol. 9, no. 3, pp. 251–263, 2016.

[5] A. Kelman, Y. Ma, and F. Borrelli, "Model predictive control of thermal energy storage in building cooling systems," *2011 IEEE Conference on Decision and Control and European Control Conference*, p. 392–397, 2011.

[6] V. L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. E. Cerpa, M. D. Sohn, and S. Narayanan, "Energy efficient building environment control strategies using real-time occupancy measurements," *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys)*, p. 19–24, 2009.

[7] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. John Wiley & Sons, 2000.

[8] G. Fant, *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[10] C. Wang, F. Lin, H. Yan, T. Wu, W. Xu, and K. Ren, "{VibSpeech}: Exploring practical wideband eavesdropping via bandlimited signal of vibration-based side channel," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 3997–4014.

[11] D. P. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3. IEEE, 2016, pp. 483–492.

[12] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9458–9465.

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

[14] S. Zhang, Y. Liu, and M. Gowda, "I spy you: Eavesdropping continuous speech on smartphones via motion sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–31, 2023.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] P. Hu, W. Li, R. Spolaor, and X. Cheng, "mmecho: A mmwave-based acoustic eavesdropping method," in *Proceedings of the ACM Turing Award Celebration Conference-China 2023*, 2023, pp. 138–140.

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

[23] 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.

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

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

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] Y. G. Achamyeleh, M. H. Fakih, G. Garcia, A. Barua, and M. A. Faruque, "A fly on the wall–exploiting acoustic side-channels in differential pressure sensors," *arXiv preprint arXiv:2409.18213*, 2024.

[35] A. Gourab and D. R. Parhi, "Iot-based smart hvac system: Control and monitoring," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6484–6492, 2021.

[36] S. S. Technology, "Hv series differential pressure sensors," https://superiorsensors.com/applications/hvac-af/, 2024, accessed: 2025-05-31.

[37] S. AG, "Sdp1108-w7 differential pressure sensor," https://sensirion.com/resource/datasheet/sdp1108-w7, 2012, accessed: 2025-05-31.

[38] "Guardian space pressure monitor," 2022, https://paragoncontrols.com/wp-content/uploads/2021/07/SPM-1000-IOM.pdf.

[39] R. D. Kent and C. Read, *The Speech Sciences*. San Diego, CA: Singular Publishing Group, 1997.

[40] M. Herrick, "Planning and maintaining hospital air isolation rooms," *Health Facilities Management Magazine*, 2017. [Online]. Available: https://www.hfmmagazine.com/articles/2671-planning-and-maintaining-hospital-air-isolation-rooms

[41] D. Dougan, "Guide for placing probes to measure airflow and psychrometric data in ducts and plenums," https://knowledgelibrary.ifma.org/guide-for-placing-probes-to-measure-airflow-and-psychrometric%-data-in-ducts-and-plenums/, 2024.

[42] D. Instruments, "Importance of sensor stability in healthcare isolation rooms," https://blog.dwyer-inst.com/2021/03/10/importance-of-sensor-stability-in-healthcare-isolation-rooms/, 2021.

[43] Senseware, "Air duct iaq system monitors air moving through ducts," *Buildings*, 2021. [Online]. Available: https://www.buildings.com/smart-buildings/article/33018367/air-duct-iaq-system-monitors-air-moving-through-ducts

[44] I. Setra Systems, "Model 264 differential pressure transducer, part no. 26410r1wd11t1e," https://www.setra.com/products/pressure/model-264-low-differential-pressure-transducer, 2025, accessed: 2025-05-14.

[45] "Clear vinyl tubing," 2022, https://www.homedepot.com/p/UDP-3-16-in-I-D-x-5-16-in-O-D-x-20-ft-Clear-Vinyl-Tubing-%T10007004/304185167.

[46] "Static pressure pickup," 2022, https://www.dwyer-inst.com/Product/Pressure/RoomStatusMonitors/SeriesRSME#accessories.

[47] J. Appelbaum, L. Poitras, M. Rosenbach, C. Stöcker, J. Schindler, and H. Stark, "Inside TAO : documents reveal top NSA hacking unit," *Der Spiegel*, 12 2013.

[48] L. Van der Velden, "Leaky apps and data shots: Technologies of leakage and insertion in NSA-surveillance," *Surveillance & Society*, vol. 13, no. 2, pp. 182–196, 2015.

[49] B. Snyder, "Snowden: The NSA planted backdoors in cisco products," *InfoWorld*, vol. 15, 2014.

[50] S. R. Chhetri *et al.*, "Tool of spies: Leaking your ip by altering the 3d printer compiler," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[51] P. Swierczynski, M. Fyrbiak, P. Koppe, A. Moradi, and C. Paar, "Interdiction in practice—Hardware Trojan against a high-security USB flash drive," *Journal of Cryptographic Engineering*, vol. 7, no. 3, pp. 199–211, 2017.

[52] V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang, "Skipconvnet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping," in *Interspeech 2020*, 2020, pp. 3935–3939.

[53] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. I, 2017.

[54] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[55] H. Liu, W. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech super-resolution," in *Interspeech*, 2022.

[56] Q. Tian, Y. Chen, Z. Zhang, H. Lu, L. Chen, L. Xie, and S. Liu, "Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis," *arXiv preprint arXiv:2011.12206*, 2020.

[57] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[58] R. Kumar, K. Kumar, V. Anand, Y. Bengio, and A. Courville, "Nu-gan: High resolution neural upsampling with gan," *arXiv preprint arXiv:2010.11362*, 2020.

[59] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[60] "Sound waves — university physics volume 1," 2016, https://courses.lumenlearning.com/suny-osuniversityphysics/chapter/17-1-sound-waves/.