VibSpeech: Exploring Practical Wideband Eavesdropping via Bandlimited Signal of Vibration-based Side Channel

Chao Wang[1,2], Feng Lin[1,2*], Hao Yan[1,2], Tong Wu[1,2], Wenyao Xu[3], and Kui Ren[1,2]

[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]School of Cyber Science and Technology, Zhejiang University
[3]University at Buffalo, the State University of New York
[*]Corresponding author

Abstract

Vibration-based side channel is an ever-present threat to speech privacy. However, due to the target's frequency response with a rapid decay or limited sampling rate of malicious sensors, the acquired vibration signals are often distorted and narrowband, which fails an intelligible speech recovery. This paper tries to answer that when the side-channel data has only a very limited bandwidth (<500Hz), is it feasible to achieve a wideband eavesdropping based on a practical assumption? Our answer is YES based on the assumption that a short utterance (2s-4s) of the victim is exposed to the attacker. What is most surprising is that the attack can recover speech with a bandwidth of up to 8kHz. This covers almost all phonemes (voiced and unvoiced) in human speech and causes practical threat. The core idea of the attack is using vocal-tract features extracted from the victim's utterance to compensate for the side-channel data. To demonstrate the threat, we proposed a vocal-guided attack scheme called VibSpeech and built a prototype based on a mmWave sensor to penetrate soundproof walls for vibration sensing. We solved challenges of vibration artifact suppression and a generalized scheme free of any target's training data. We evaluated VibSpeech with extensive experiments and validated it on the IMU-based method. The results indicated that VibSpeech can recover intelligible speech with an average MCD/SNR of 3.9/5.4dB.

1 Introduction

Sound-induced vibration is a common phenomenon and prevalent when a sound source (e.g., a loudspeaker) produces acoustic waves. Researchers have revealed great threats posed by these sound-related vibrations via different sensing methods, e.g., optical sensors [14, 34, 38–40, 45], motion sensors [5, 8, 18, 21, 37], wireless devices [9, 19, 20, 36, 50–52], and vibration-sensitive components [28, 44]. For a protected zone like a soundproof room, millimeter-wave (mmWave) sensors can be exploited for through-wall eavesdropping due to the advantages of penetrability and high precision.

However, current mmWave-based works [19, 56] either require an ideal reflector with a wideband frequency response or require abundant training data from the target to achieve satisfactory performance. These methods suffer narrowband frequency responses and the strong assumption of targets' training data. With a laser vibrometer (Figure 1), we found that narrowband frequency response is often the case due to the forced damped vibration [15] induced by excitation audio.

In this paper, we reveal a new speech threat that adversaries can recover wideband (up to 8kHz) intelligible speech from acquired narrowband (<500Hz) vibration signals on loudspeakers. An attacker can use a mmWave sensor to capture the sound-related vibrations in a soundproof zone and leverage the proposed wideband-speech-recovery scheme called *VibSpeech* to recover intelligible speech. The core idea of our work is using vocal-tract features extracted from a pre-obtained utterance of the victim to compensate for acquired narrowband signals during the attack and recover wideband speech. Compared with prior works, our proposed attack is

(1) *Free of a tough requirement of wideband frequency response of the target.* Frequency response plays an important role in vibration-based speech recovery. However, an ideal target with wideband frequency response is not always available. Our study (Section 5) indicates that due to the characteristic of the forced vibration, the vibration amplitude on loudspeakers can suffer over 20dB degradation when the excitation frequency is above 500Hz. A consequence is that high-frequency vibration is overwhelmed by noise and practically unavailable. VibSpeech focuses on the prevailing narrowband status quo while holding the goal of wideband speech recovery.

(2) *Free of training data from the target.* Although using a neural network to learn the mapping between narrowband vibration signals and wideband audio may merge the gap between the two signals, such a method requires a large training dataset of the target to achieve satisfactory performance, which is impractical to conduct in real cases. One of our goals is to free the narrowband-wideband speech transition from the abundant training data of the target. Only with a short utterance (2s-4s) of the victim, our work can recover wideband

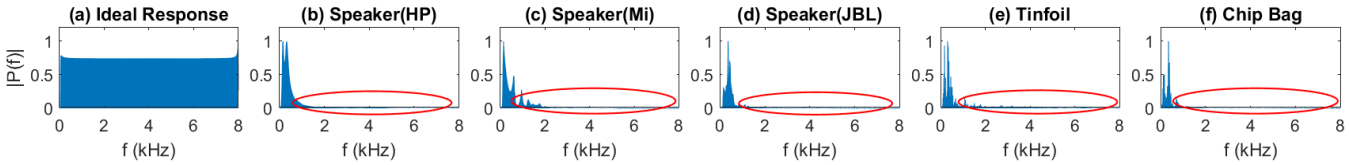

Figure 1: Frequency response of sound-forced vibrating objects measured by a laser vibrometer (Figure 4(a)). Compared with (a) the ideal frequency response, pervasive uneven response on objects (b)-(f) challenges a wideband and intelligible speech recovery from the distorted and narrowband vibration signals. Please refer to Section 5 for a detailed investigation.

intelligible speech during the attack.

(3) *A through-wall and non-invasive attack.* Although existing works revealed that the vibration signals can be acquired using lasers, cameras, optical sensors, and motion sensors, these methods either cannot penetrate opaque obstacles or work in an invasive manner of deriving data from the target via malware. Our mmWave-based attack can break through soundproof protections in a non-invasive manner, and achieve a better resolution than the ones based on other radio-frequencies techniques such as WiFi and ultra-wide-band radio.

(4) *An intelligible speech recovery with a bandwidth of up to 8kHz.* Although many works revealed that attackers can use narrowband vibration signals to infer speech contents (i.e., word recognition), such a method is restricted to specific vocabulary and cannot recover arbitrary speech that is not contained in the vocabulary. By contrast, VibSpeech can recover hearable speech with high intelligibility and quality.

To achieve the attack, we aim to solve the following challenges. **First**, how to eliminate the distortion and artifacts in acquired narrowband vibration signals? In a real case, the collected vibration signal often suffers distortions and artifacts due to the target's uneven frequency response and noise (Section 5). To solve this problem, we proposed adaptive multi-bandpass filtering to suppress distortion and inter-harmonic noise in the narrowband signal (Section 6.2). **Second**, how to merge the bandwidth gap between the vibration signal and intelligible speech? Formants of consonants can be up to 4k-8kHz (Figure 2). Narrowband audio with incomplete formants can be blurred and unintelligible. To solve this challenge, we designed a vocal-guided scheme by fusing narrowband speech and the target's vocal-tract features. The vocal-tract features are extracted via a *SpkEnc* network that only requires a short utterance (2s-4s) from the victim (Section 6.1). Further, the wideband speech is reconstructed via a bandwidth-extension network (Section 6.3) and a generalized vocoder (Section 6.4). **Third**, how to free the narrowband-wideband speech transition from the labor-intensive and impractical assumption of training data collected from the target? To achieve this goal, we proposed to manipulate the bandwidth of public audio datasets to generate narrowband audio as training data. This training strategy makes the attack free of the requirement of abundant data from the target to achieve satisfactory performance and becomes a general scheme that can also applied to other vibration side channels, e.g., IMUs (Section 8).

Our contributions are summarized as follows:

- We revealed a new speech threat that when the vibration side-channel data has a limited bandwidth as low as 500Hz, it is feasible to recover wideband (8kHz) intelligible speech. The attack is based on the assumption that a short (2s-4s) utterance of the victim is exposed to the attacker.
- We first investigated the narrowband characteristic of the vibration-based side channel. We proposed a vocal-guided attack scheme (VibSpeech) and used a mmWave sensor to penetrate soundproof protections to demonstrate the threat.
- We performed extensive experiments to evaluate the proposed attack. The results indicate that VibSpeech can recover wideband and intelligible speech with an overall MCD/SNR score of 3.9/5.4dB. The evaluation under different conditions shows the robustness of the attack. We also validated the generalization of the attack on the IMU-based method. Audio samples are available at https://demo-online.github.io/VibSpeech/.

2 Related Work

Sound-induced vibrations on objects can be measured by sensors or vibration-sensitive devices for speech recovery, e.g., lasers [38, 45], inertial measurement units (IMU) [5, 8, 18, 21, 37], wireless signals [19, 20, 36, 49, 51–53], optical sensors [14,34,40], vibration motors [44], and hard drives [28]. These works revealed great threats to speech privacy. Table 1 shows the comparison with typical and state-of-the-art work. Different features are involved to discuss their pros and cons, such as the dependency on the target's prior data, invasiveness, through-wall capability, sensing distance, and the bandwidth of recovered speech. Considering that metrics used in these works are different, common metrics (e.g., SNR) are reported for quantitative comparison. Furthermore, we conducted an experiment under the same experimental setting (Section 7.13) for a fair comparison with related methods [20, 56].

Optical-based: Laser microphone has been studied since the 1980s [38]. Adversaries can use laser transceivers to measure sound-induced vibrations on objects to recover sound. Sami *et al.* [45] revealed that the Lidar sensor on a robot

Table 1: Comparison with prior work on vibration-based side channel.

Sensor	Work	Target	Target's prior data	Non-invasive	Through-wall (opaque)	Sensing Distance	Speech Bandwidth	Performance
Optical	[14, 38, 40]	Loudspeaker	-	✔	✘	Far	Narrow	SNR:7.4dB [40]
	[34, 45]	Loudspeaker	Moderate	✘	✘	Moderate	Narrow	Acc:81% [34]
IMU	[37]	Loudspeaker	Heavy	✘	✘	Close	Narrow	Acc:65% [37]
	[8, 21]	Smartphone	Moderate	✘	✘	Close	Narrow	MCD:4.8 [21]
mmWave	[9, 50]	Smartphone	Light	✔	✘	Moderate	Narrow	PSNR: 20dB [50]
	[20, 56]	Loudspeaker	Heavy	✔	✔	Moderate	Narrow	SNR:0.2dB/MCD:10.3 [1]
	VibSpeech	Loudspeaker	Light	✔	✔	Moderate	Wide	SNR:5.4dB/MCD:3.9

[1] The result was acquired under the same experimental setting with VibSpeech. More details are introduced in Section 7.13.

cleaner can sense the speech-related vibration on objects for eavesdropping if the attacker hacks the robot cleaner and accesses the sensor data. Besides the laser, visible light can also be used for vibration measurement. The Visual Microphone [14] decoded audio from videos via a high-speech camera. Lamphone [40] and Glowworm [39] used electro-optical sensors and telescopes to capture the light bulb's vibration and the light changes of power indicators to recover speech of around 7.4dB SNR. With a professional lens, the sensing distance can be enlarged over 10m but can only work without opaque walls. Recently, Long *et al.* [34] revealed that pixel changes in smartphone images can also be used for speech decoding. It achieves 81% accuracy for digit recognition. Optical-based methods can sense delicate vibrations and often have a longer sensing distance than other methods with professional lenses for enhancement. But this kind of method cannot penetrate opaque blockage for sensing. Although VibSpeech has a lower SNR of 5.4dB than the state-of-the-art (7.4dB) due to the penetrating attenuation, it still indicates satisfactory performance for speech recovery.

IMU-based: Motion-sensor-based attacks have been widely studied these years [5, 8, 18, 21, 37]. Gyrophone [37] exploited smartphone gyroscopes to capture the external loudspeaker's vibration for digit inference. AccelEve [8] achieved word reconstruction and gender inference via the built-in accelerometer to collect the vibration of the smartphone's loudspeaker. Recently AccelEar [21] exploited generative adversarial networks to recover audible speech from accelerometer data. For the IMU-based attacks, an intelligible speech recovery is challenging due to the low sampling rate of IMUs. Although recent work [21] revealed the possibility of using a more powerful network to learn the mapping between the distorted vibration signal and the corresponding utterance, they still required abundant vibration data and ground-truth audio from the specific target to achieve satisfactory performance. Compared with current IMU-based work, we do not rely on the strong assumption of abundant training data from the specific target while holding the attack goal of recovering intelligible speech. For the motion-sensor-based attacks, another drawback is their intrusive manner, i.e., requiring malicious applications on the target to collect vibration data. IMU-based methods can cause threats to smartphones besides external loudspeakers. This kind of attack can be achieved with low-cost motion sensors but often requires physical contact with sound sources for vibration measurement. Compared with non-invasive methods like optical-based and mmWave-based ones, the IMU-based methods also require to hack the target for data collection. VibSpeech works in a non-invasive manner and achieves a lower MCD of 3.9 (better performance) than the state-of-the-art (4.8) of IMU-based methods [21].

Wireless-based: Wireless signals have been exploited by researchers for speech inference [9, 19, 20, 30–33, 48–50, 52, 53]. An advantage of wireless signals is that it can penetrate soundproof obstructions. Benefiting from a shorter wavelength, mmWave sensors can acquire more dedicated vibrations for speech recovery. Work [19] proposed a phase-calibration method to recover vibration signals. However, an ideal object is not always available and current work still cannot essentially solve the narrowband dilemma that is pervasive in vibration-based side channels. Our work does not rely on the ideal reverberator but leverages the prevalently-narrowband vibrating signals. mmSpy [9] and mmEve [50] are good works to reveal the vibration-based side channel on smartphones that can cause threats to speech played by the earpieces. They found that the built-in earpieces can induce vibration on the smartphone backshell, which can be used to infer words or recover narrowband sound. However, these works still suffer the narrowband condition of vibration-based side channels and can only recover bandlimited sound with damaged speech intelligibility. For comparison, VibSpeech seeks to solve the pervasively narrowband feature of vibration-based side channels and can recover 8kHz-wideband sound that shows practical threats. Besides, VibSpeech is a generalized framework for different vibration-sensing methodologies, e.g.,mmWave sensors, laser vibrometers, and IMUs. Compared with laser-based and IMU-based methods, the mmWave-based methods often have a moderate sensing distance of several meters(<10m) but can penetrate opaque walls for non-contact vibration sensing. It is also a non-invasive manner without hacking the target. However, current works still suffer the narrowband limitation of vibration-based side channels and rely on heavy training data from the target to achieve

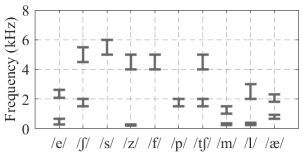

Figure 2: Frequency range of phoneme formants (F1, F2).

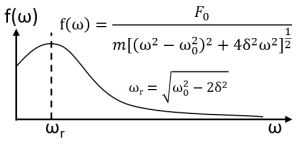

Figure 3: Freq. response of forced damped vibration [4].

satisfactory performance. Recent work [20, 56] used neural networks to learn the mapping between the acquired vibration signals and corresponding audio. However, they required both abundant vibration data and ground-truth audio collected from the victim to achieve satisfactory performance. This constrains prior work to a target-dependent attack with the strong assumption of requiring a large training set from the victim. Without sufficient training data from the target, the performance can get worse with SNR/MCD scores of 0.2dB/10.7. By contrast, VibSpeech uses vocal-tract features to compensate for narrowband audio instead of learning the mapping. It achieves better performance (5.4dB/3.9) with less prior data (2-4s audio) from the target. This frees the attack from a tough requirement of training data from the victim.

3 Background

Speech Production. Human speech is produced by the cooperation among vocal organs (e.g., lungs, vocal cords, and articulators). The lungs pump the air through the vocal cords to cause an acoustic resonance and produce low-frequency components (e.g., pitch and its low-order harmonics) of the speech. The resonating air further passes through the vocal tract composed of the articulators (e.g., the tongue, mouth, and lips), and is fine-tuned to generate higher frequency components (e.g., formants) to make a hearable speech. The **phoneme** is the basic unit of pronunciation. In the case of English, there are seven categories of phonemes, i.e., vowels (e.g., "/e/" in bet), fricatives (e.g., "/s/" in this), stops (e.g., "/p/" in cap), affricates (e.g., "/tʃ/" in change), nasals (e.g., "/m/" in make), glides (e.g., "/l/" in lip), and diphthongs (e.g., "/æ/" in at). Each phoneme is produced by different configurations of human vocal organs and has different formant frequency bands [10, 35] as shown in Figure 2 (F1 and F2 mean the first and second formants). The combination of phonemes with transitions (e.g., bigrams) helps people distinguish the words. *A damaged frequency band of phonemes (e.g., losing high frequencies) will make the speech blurred and unintelligible for human hearing.*

mmWave-based Vibration Sensing. Millimeter-wave sensors operating in frequency-modulated continuous-wave (FMCW) mode are widely used in automatic driving and industrial monitoring. Benefiting from the millimeter-level wavelength, the mmWave sensor can measure delicate vibrations from the phase of demodulated intermediate-frequency (IF) signals. The reflected signal can be taken as a delayed replica of the transmitted signal. Given a transmitted signal with a frequency slope of S, the IF signal demodulated by the sensor can be represented by $A\sin(2\pi f_0 t + \phi_0)$. The frequency f_0 of the IF signal indicates the distance d of the target from the sensor $d = \frac{f_0 c}{2S}$ where c is the speed of light in a vacuum. To acquire the target's vibration signals, we can apply fast Fourier transform on the IF signal (called *range-FFT*) and acquire the rang-bin (i.e., f_0) corresponding to the target. Then the vibration signal Δd can be derived from the phase change $\Delta\phi$ of the range-bin $\Delta d = \frac{\lambda\Delta\phi}{4\pi}$ where λ is the wavelength of the transmitted signal. To acquire the range-bin corresponding to the target, we can apply a high-pass filter (f_c=80Hz) on all derived phase sequences and calculate the power density. Then we choose the one with the highest score as the derived vibration signal. The sampling rate f_s of the vibration signal is determined by the chirp rate f_{chirp}, i.e., $f_s = f_{chirp}$.

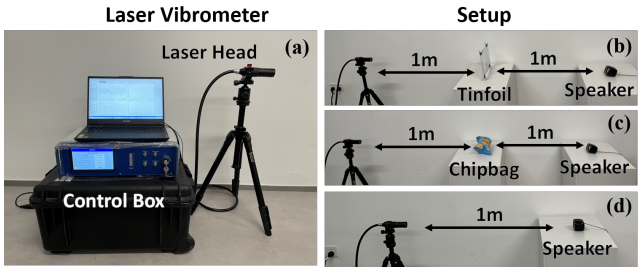

Figure 4: The laser vibrometer and experimental setup to measure the frequency response of vibrating objects.

4 Threat Model

Attack Scenario and Goal. We consider an attack scenario where an individual is having an online conversation in a soundproof room. The individual uses a loudspeaker to play the speech. The speech may contain private information and secrets. An attacker who is interested in the speech launches eavesdropping from the outside of the isolated zone. The attacker can use a portable mmWave sensor to penetrate the sound insulator and capture the sound source vibration for speech recovery. To cause a practical threat, the attacker aims to overcome the limitations of the distorted narrowband vibration signals, and recover wideband intelligible speech.

Assumption. Considering the practicality of the attack, we do not assume the attacker has plenty of vibration data and audio data collected from the victim to train a target-dependent model. By contrast, we assume that the attacker can acquire a short (2s-4s) utterance of the victim before launching the attack. The utterance can be collected by a hidden microphone or acquired in cyberspace, e.g., by making phishing calls to record seconds of audio of the victim.

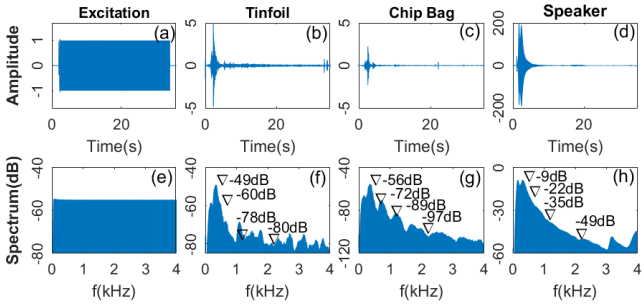

Figure 5: Frequency response measured by the laser vibrometer. (a)(e) show the waveform and spectrum of played chirp audio. (b)(c)(d) show measured vibration on objects. (f)(g)(h) show corresponding spectrums. The marked points in (f)(g)(h) indicate the maximum magnitude, and the magnitudes when the excitation frequency equals 500Hz, 1kHz, and 2kHz.

5 Characteristic of Vibration-based Side Channel: the Narrowband Nature

In this section, we first introduce the theoretical model of sound-related vibration and reveal the significant nature of the vibration-based side channel, i.e., *the sound-related vibration can attenuate with the increasing frequency of audio signals, resulting in uneven frequency response and ubiquitously-narrowband vibration signals*. Then we introduce our investigation and observations about the narrowband phenomenon via a professional laser vibrometer and a mmWave sensor.

Frequency Response of Vibration-based Side Channel. Frequency response characterizes the relationship between the object's vibration and the excitation signal in the frequency domain. Sound-induced vibration (e.g., on passive diaphragms) or coil-driven vibration (e.g., on a loudspeaker) can be taken as forced damped vibration [15]. Given an excitation signal with acoustic-wave function $P(t) = P_0\cos(\omega t+\phi)$, the radiation force on the object can be represented by $F = F_0\cos(\omega t+\phi)$ where $F_0 = AP_0$ and A is the object's equivalent surface area. Then the forced vibration on the object m can be characterized by the typical spring-oscillator model [4] and the forced displacement is formulated by

$$x(t) = \frac{F_0}{m[(\omega^2-\omega_0^2)^2+4\delta^2\omega^2]^{\frac{1}{2}}}\cos(\omega t+\phi), \qquad (1)$$

where δ and ω_0 are constants and known as the damping coefficient and the natural angular frequency of the object, respectively. Eq.1 indicates that when $\omega > [\omega_0^2-2\delta^2]^{\frac{1}{2}}$, the displacement amplitude $\frac{F_0}{m[(\omega^2-\omega_0^2)^2+4\delta^2\omega^2]^{\frac{1}{2}}}$ of the vibrating object degrades greatly with the increasing frequency $f=\frac{\omega}{2\pi}$ of the excitation signal, as shown in Figure 3. When the vibrating frequency increases to the point where the induced displacement is overwhelmed by noise, the captured vibration loses high-band components and causes a narrowband result.

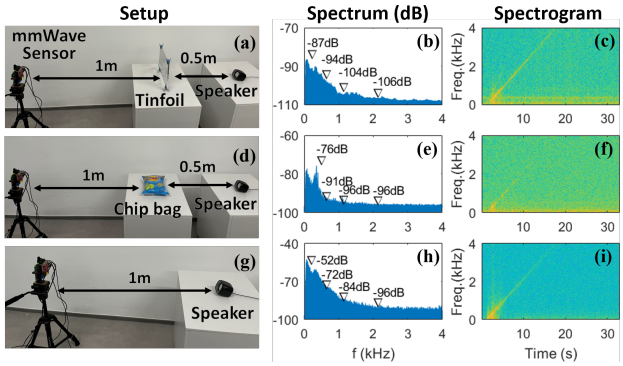

Figure 6: Frequency response measured by a mmWave sensor.

Laser Vibrometry. To quantitatively investigate the characteristic of the vibration-based side channel, we used a laser vibrometer (LV-FS01 [3]) with a displacement resolution smaller than 1nm, to measure sound-related vibration as shown in Figure 4. Tested objects included tinfoil, a chip bag, and an HP loudspeaker. The object vibrated due to the excitation of chirp audio (80-8kHz, 68dB) played by the loudspeaker. The frequency response of tested objects is shown in Figure 5. Compared with the flat spectrum of the excitation signal shown in Figure 5(e), we can observe that the spectrums of vibrating objects have an attenuating magnitude with increasing frequency as shown in Figure 5(f)(g)(h). Compared with the strongest response at the frequency band below 1kHz, the attenuation of the frequency response achieves 11dB-16dB when f=500Hz, 26dB-33dB when f=1kHz, and 31dB-41dB when f=2kHz. *The result validated that, with the increasing frequency of excitation audio, the amplitude of induced vibration on objects will decrease. On one hand, the uneven frequency response causes distortions in captured signals (i.e., distorted speech) compared with the ground-truth audio. On the other hand, the attenuated high-frequency components are prone to be overwhelmed by background noise resulting in narrowband signals. The narrowband dilemma of the vibration-based side channel causes loss of abundant speech information in the high-frequency band, and challenges intelligible speech recovery.*

mmWave Sensing. As shown in Figure 6, we used a mmWave sensor (AWR1843Boost) to capture the vibration of the same objects mentioned above. The same chirp audio was played as the excitation signal to acquire the frequency response. The vibration signals are derived from the mmWave signals as introduced in Section 3. The spectrogram and spectrum of acquired vibration signals are shown in Figure 6. From Figure 6(c)(f)(i), we can observe that the power density of the vibration signal decreases with increasing frequency and is overwhelmed by background noise when the frequency reaches a specific high-frequency band. Compared with the strongest response shown in Figure 6(h), we can find that the frequency response of the loudspeaker suffers attenuation of

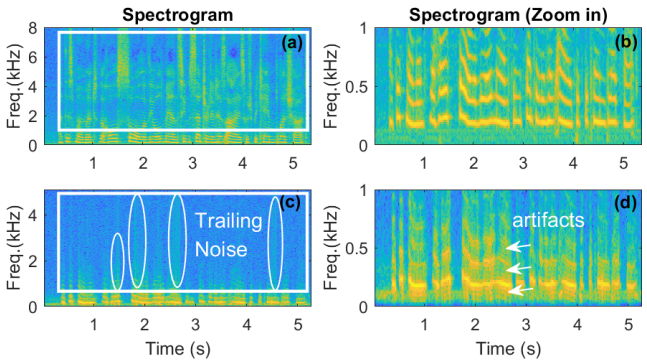

Figure 7: Through-wall results. (a) and (b) show the spectrograms of played audio. (b) shows the 0-1kHz band of (a). (c)(d) show spectrograms of mmWave-recovered audio. (d) shows the 0-1kHz band of (c).

20dB/32dB/44dB at the frequency of 500Hz/1kHz/2kHz.

mmWave Sensing (Thru-wall). To further investigate the impact of wall blockage on captured vibration signals, we set the mmWave sensor outside a room to penetrate the soundproof glasses and sense the loudspeaker's vibration from a distance of 3m (Figure 10(a)). The loudspeaker played speech audio that contained frequency components of up to 8kHz, i.e., "at this moment, the whole soul of the old man seemed centred in his eyes which became bloodshot". Spectrograms of the excitation audio and raw mmWave-recovered audio (i.e., derived vibration signals) are shown in Figure 7. Compared with the original audio as shown in Figure 7(a), we can find that the recovered audio shown in Figure 7(c) suffers a significant loss of high-frequency response, especially the frequencies above 500Hz. The result is consistent with our observation and analysis mentioned above, i.e., the vibration-derived audio has a limited bandwidth due to the uneven frequency response with rapid decay of the vibration-based side channel. The high-band speech components are prone to be buried in the noise resulting in unintelligible speech. We can also find that there is trailing noise in the mmWave-recovered audio as shown in Figure 7(c). Besides, for the 0-1kHz band of the mmWave-recovered audio as shown in Figure 7(d), there are artifacts between the pitch and harmonics due to the sensor noise. These artifacts degrade the signal-to-noise ratio of the speech and make speech blurred, which should be suppressed.

Short Summary. Based on the above investigation, we can find that *(1) the amplitude of sound-related vibration actually does not follow linearity with the excitation audio, but degrades with the increasing frequency of the excitation audio. This also explains the narrowband results in prior work on vibration-based side channels [14, 20, 40, 56]. Such characteristic of the vibration-based side channel causes distortion in the vibration-derived audio and restricts the recovered speech to the dilemma of a narrow band. (2) Although mmWave sensing is feasible for through-wall vibration sensing, the mmWave-recovered audio still suffers the loss of high-band response (especially for f >500Hz) and artifacts in the available low-frequency band (f <500Hz), which challenge intelligible and wideband speech recovery.*

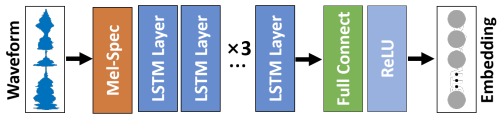

Figure 8: The structure of *SpkEnc*.

6 VibSpeech: A Wideband Speech Recovery Scheme for Vibration-based Side Channel

System Overview. In this section, we introduce how to overcome the aforementioned challenges. We propose to use the vocal-tract features of the speaker to compensate for the compromised and narrowband speech. We first introduce *SpkEnc* to extract speaker embeddings that characterize the speaker's vocal-tract features (Section 6.1). The feature extraction only requires a short utterance of around 2s-4s from the victim. Before feeding the narrowband vibration data for wideband speech recovery, we preprocess the vibration signal to suppress the distortion and artifacts in the low-frequency band (Section 6.2). Then a vocal-guided bandwidth extension network is proposed to compensate for the narrowband signal using extracted vocal-tract features (Section 6.3), and produce an enhanced spectrogram. Finally, a vocoder transforms the enhanced spectrogram into audible waveforms (Section 6.4).

6.1 Vocal-tract Feature Extraction

Deep-learning-based vocal-tract feature extraction has shown a powerful ability to characterize the features of human vocal-tracts [7, 11, 23, 47]. An ideal way of vocal-tract feature extraction is to collect all the phonemes/bigrams from the speaker for the feature extraction. However, in a real case, speech contents of acquired utterances from the victim can be arbitrary and cannot be promised on expected phonemes. Thus, a key challenge of the vocal-tract feature extraction is how to make the extraction independent of the pronounced phonemes (i.e., utterance-independent features) of the speaker and achieve an accurate extraction based on a short utterance with acceptable duration. To achieve this goal, we designed a speaker-encoder network *SpkEnc*. We trained *SpkEnc* with the GE2E loss [47] which computes the loss on each embedding and aims to push the embedding result to the centroid of the same speaker on the feature space and away from ones of other speakers. For a satisfactory generalization performance, we randomly chose fragments from utterances of different individuals. The training process randomly takes utterances from different speakers without the requirement of utterance alignment to achieve an

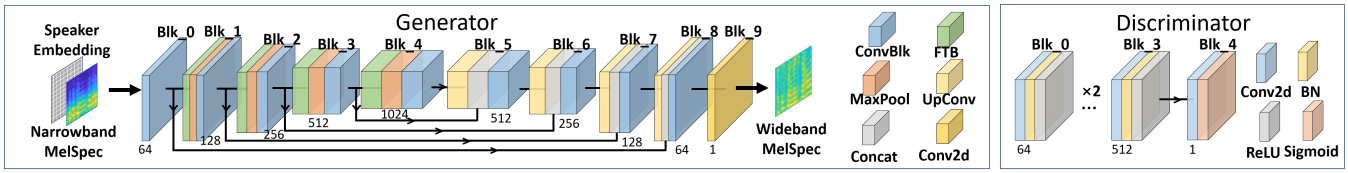

Figure 9: Vocal-guided bandwidth extension. The network parameters are detailed in Appendix C.

utterance-independent embedding extraction. The GE2E loss is detailed in Appendix A. The network consists of six layers of Long-Short-Term-Memory (LSTM) blocks followed by a full-connected layer as the projection layer. The feature size of each LSTM layer and the projection layer is 256 which is consistent with the embedding size. A Rectified Linear Unit (ReLU) outputs the final embedding from projected features.

Training: We used a public dataset (Voxceleb2 [13]) to train *SpkEnc* with 5.9 million steps for convergence. During training, we applied voice-activity-detection (VAD) [25] to remove internal silence parts and segmented the utterance into fragments of 1.6 seconds with an overlap of 0.8 seconds. The 40-band mel spectrogram of each fragment is fed into *SpkEnc* to generate the embedding of each fragment. The final speaker embedding is derived by calculating the L2-normalization of these embedding vectors. Note that the network for vocal-tract feature extraction is trained offline in advance and free of target's training data. Based on the trained *SpkEnc*, the attacker can collect a short utterance from the victim and feed the utterance into *SpkEnc* to derive the victim's speaker embeddings. During the attack phase, the attacker uses the extracted embeddings for wideband speech reconstruction(Section 6.3).

6.2 Preprocess

As investigated in Section 5, besides the limited bandwidth, the vibration signal also suffers distortion and artifacts due to the uneven frequency response of objects and the sensor noise in the raw vibration signal. Thus, before feeding the narrowband signals for wideband speech reconstruction, we need to correct the distortion and remove the artifacts first. The derivation of vibration signals from mmWave data is introduced in Section 3. From Figure 7(d), we can find that the artifacts exist between the pitch and its harmonics. For such inter-band noise, an intuitive solution is to apply a finite impulse response (FIR) filter with multiple passbands to remove the noise. However, finding proper filter parameters (e.g., center frequency) is hard and impractical to achieve considering that the frequency of pitch and harmonics can change according to different pronunciations. To address this problem, we propose an adaptive bandpass-filtering algorithm based on the feature of human speech. The core idea is based on the fact that for a short fragment, the pitch often shows a high power density in the low-frequency band (80-255Hz) while the harmonics are multiples of the pitch. Thus, we can first acquire fragments of the speech (*OverlapSegment*) and estimate the fundamental frequency f_0 for each fragment s_n (*PitchEstimation*). Then we apply multiple bandpass-filtering $BPF(s_n, f_c)$ on the harmonics by setting the center frequency f_c of the filter into multiples of f_0, i.e., $f_c = M \cdot f_0$. To suppress the low-band distortion, a correction coefficient is applied on m-th harmonic based on its power density $|S_m|$. The final result is acquired by adding the segments with overlap (*OverlapAdd*). The algorithm is shown in Algorithm 1 (Appendix B). After the preprocessing, the artifacts and distortion in the raw vibration signals are suppressed. The result is fed into the vocal-guided bandwidth extension to derive a wideband spectrogram.

6.3 Vocal-guided Bandwidth Extension

The core idea of VibSpeech is to use the extracted vocal-tract features of the victim to compensate for acquired narrowband vibration signals. The key is how to fuse the narrowband spectrogram and extracted vocal-tract features (i.e., speaker embeddings mentioned in Section 6.1) to acquire a wideband spectrogram. To achieve this goal, we designed an attention-based neural network as shown in Figure 9 to fuse the speaker embedding and narrowband signal. Specifically, we used adversarial training to improve the generality of the model. The generator takes an encoder-decoder structure that shows superior performance in pixel-to-pixel transformation, as the backbone. The shortcut connection between every two blocks helps to maintain details in the upper level [43]. To capture non-local correlations among harmonics in the time-frequency (T-F) domain, we adopted a frequency-transformation-block (FTB) for each block. The core of FTB is to learn a transformation matrix applied on the frequency axis and has a full-frequency receptive field [55]. Its inner convolution layers predict an attention map in the T-F domain and help to capture the essential correlation among speech harmonics using the channels with speaker embeddings. The discriminator aims to justify the consistency of two inputs and output the predicted label, i.e., real or fake. It consists of four convolution blocks to extract pixel-wise features of inputs, and a Sigmoid layer to predict the result. The network parameters are introduced in Appendix C.

Training: During the training of the generator, we adopt the mel loss and adversarial loss to improve the distortion of reconstructed spectrograms. We use the BCE loss to update the discriminator. Details of the loss functions are in-

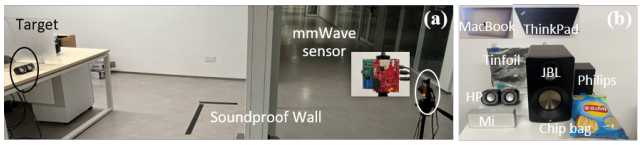

Figure 10: (a) A room with the soundproof wall (two layers of 1cm-thick glass). (b)Tested vibrating objects in the evaluation.

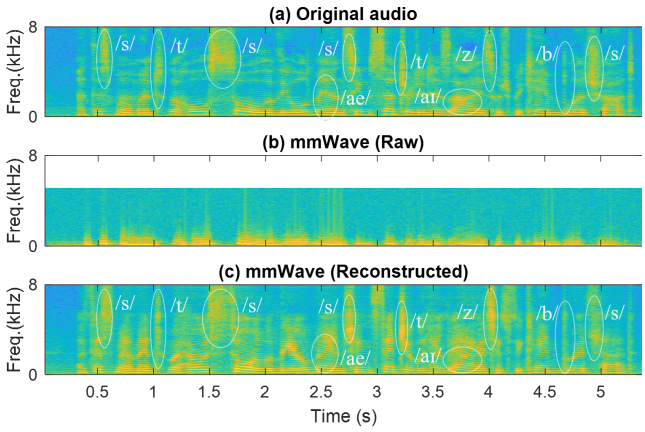

Figure 11: The spectrograms of (a) original audio, (b) raw-recovered audio via mmWave, and (c) reconstructed audio by VibSpeech. (Speech: *at this moment the whole soul of the old man seemed centred in his eyes which became bloodshot*)

troduced in Appendix C. The generator takes narrowband spectrogram and speaker embedding as inputs to reconstruct lost formants in the high-frequency band. The narrowband spectrograms used for training are derived from public audio datasets (train-clean-100/train-clean-360/train-other-500 of LibriSpeech [41]) by applying low-pass filtering (with a cut-off frequency f_c of 500Hz) to each audio trace. The generator and the discriminator are updated alternatively. The network was trained with 61k steps for convergence. During the attack, the attacker feeds the preprocessed narrowband signal (Section 6.2) and pre-extracted speaker embedding (Section 6.1) into the trained generator to acquire a wideband spectrogram.

6.4 Vocoder

To recover hearable voice, we need to transform the enhanced spectrogram (Section 6.3) into waveforms in the time domain. The transformation can be achieved by a vocoder. In recent years, generated adversarial network (GAN)-based vocoders such as WaveRNN [24], MelGAN [27], HiFiGAN [26], and BigVGAN [29], have shown great improvement in the naturalness and fidelity of reconstructed speech compared with traditional methods (e.g., Griffin-Lim vocoder [16]). Thus, we adopt the state-of-the-art structure, i.e., BigVGAN [29], which achieves superior performance and generality for audio synthesis. The vocoder is composed of six blocks of transposed 1-D convolution followed by an anti-aliased multi-periodicity composition (AMP) module. To model complex waveforms, the AMP module contains multiple signal components with learnable periodicities and adopts a low-pass filter to reduce the high-frequency artifacts. **Training:** The vocoder takes the mel spectrogram as the input and produces hearable audio waveforms. We used public datasets (train-clean-100/train-clean-360/train-other-500 of LibriSpeech [41]) to train the vocoder. The mel-spectrograms are produced by applying a 128-mel-transform on the result of a short-time Fourier transform. The model was trained with 110k steps for convergence.

7 Evaluation

7.1 Setup and Metrics

We used samples from a widely-used speech corpus LibriSpeech (dev-clean) [41] for evaluation. The dataset contains 2,073 audio traces sampled at 16kHz from 40 individuals (i.e., 20 males and 20 females). We played the audio (68dB) via an HP loudspeaker and collected corresponding vibration data via a mmWave sensor (AWR1843Boost) as shown in Figure 10(a). The derived vibration signal is at a sampling rate of 10.2kHz and resampled at 16kHz for further processing. Note that in a real case, the speech contents for vocal-tract feature extraction may be different from the spoken ones during the attack phase. Thus, for each tested individual, we randomly chose one trace for the vocal-tract feature extraction, and used the left samples to evaluate the performance of the speech reconstruction. The models were trained offline with four NVIDIA RTX A6000 GPUs and deployed on a laptop (ThinkPad) for speech recovery. For robustness experiments (Section 7.4-7.9), the used corpus includes 663 samples of the first five males and five females in the 40-individual dataset.

Mel-Ceptral Distortion (MCD): The MCD is a widely used metric to quantify the speech distortion between the recovered audio and the original audio. Considering the target waveform x_{targ} and the reference waveform x_{ref}, the MCD can be calculated by

$$MCD = \frac{10}{T \ln 10} \sum_{t=0}^{T-1} \sqrt{2 \sum_{d=1}^{D} [c_d^{targ}(t) - c_d^{ref}(t)]^2}, \quad (2)$$

where $c_d^{targ}(t)$ and $c_d^{ref}(t)$ denote the MFCC [17] of x_{targ} and x_{ref} respectively. T and D mean the total frames and MFCC dimensions. Typically, the recovered audio with an MCD value smaller than 8 indicates satisfactory performance [54]. A lower MCD score indicates better speech intelligibility.

Frequency-weighted Signal-to-Noise Ratio (SNR): To quantify the gain and quality of the reconstructed speech, we used the frequency-weighted SNR [22] calculated by the average gain across multiple frequency bands of the signal:

$$SNR = \frac{10}{N} \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} W(k,n) \log_{10} \frac{|S(k,n)|^2}{(|S(k,n)|-|\hat{S}(k,n)|)^2}}{\sum_{k=1}^{K} W(k,n)}, \quad (3)$$

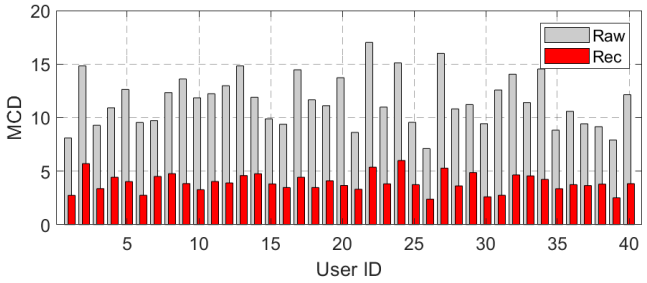

Figure 12: Raw: the score of raw-recovered audio. Rec: the score of reconstructed audio by VibSpeech.

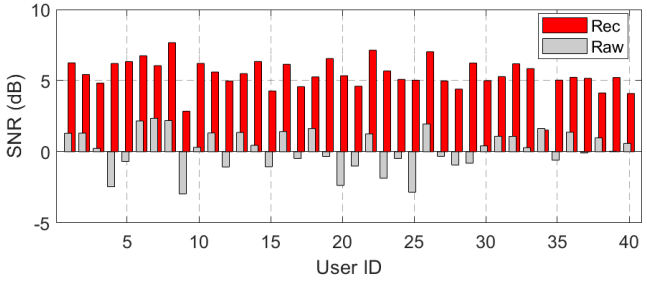

Figure 13: Raw: the score of raw-recovered audio. Rec: the score of reconstructed audio by VibSpeech.

where $S(k,n)$ and $\hat{S}(k,n)$ are weighted spectrums in the k-th frequency band at the n-th frame of original and reconstructed speech. $W(k,n)$ is the weight placed on the k-th frequency band where $W(k,n) = |S(k,n)|^{0.2}$. N is the total number of frames. A higher SNR indicates a better speech quality.

7.2 Overall Performance

To intuitively observe the performance of the wideband speech recovery, we compared the spectrograms of original audio, raw-recovered audio, and reconstructed audio by VibSpeech, as shown in Figure 11. Compared Figure 11(a) with Figure 11(b), we can observe that the frequency components above 500Hz in the raw-recovered audio are absent due to the narrowband frequency response of the vibration-based side channel. From Figure 11(c), we can observe that the high-frequency components are recovered with a bandwidth of up to 8kHz (e.g., consonants /s/ and /z/) after the reconstruction. The reconstructed spectrogram shows a high similarity with the original one in (a). The overall MCD and SNR scores are shown in Figure 12 and Figure 13, respectively. Overall, VibSpeech-reconstructed audio achieves an average MCD/SNR score of 3.9/5.4dB while the raw-recovered audio only achieves an average MCD/SNR score of 11.5/0.2dB. This indicates that VibSpeech can effectively correct the distortion and suppress artifacts in raw-derived vibration signals, and improve the speech intelligibility and quality significantly.

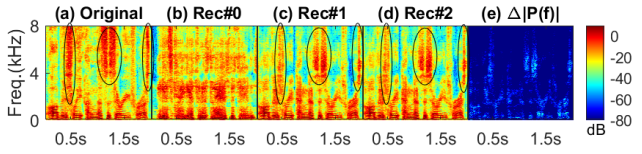

Figure 14: (a) Original audio of speaker A. (b) Reconstructed audio without speaker embedding (speaker A). (c)(d) show results with speaker embeddings extracted from two different utterances of speaker A. (e) Difference between (c) and (d).

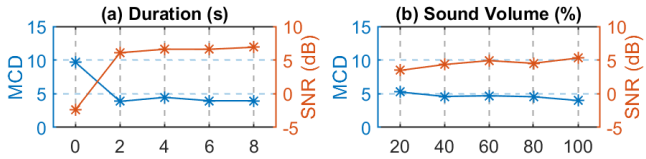

Figure 15: Impact of (a)speaker embedding and (b)volume.

7.3 Impact of Speaker Embedding

To investigate the impact of the speaker embedding, we used different durations of the audio trace for speaker embedding extraction, including 0s (random noise), 2s, 4s, 6s, and 8s. The tested dataset is the one we constructed in Section 7.1. For each tested individual, we randomly chose a trace and selected one of the fragments for the embedding extraction. The collected vibration data corresponding to the left samples are used for the speech reconstruction. The calculated MCD and SNR scores are shown in Figure 15(a). (1) Overall, the MCD score decreases and the SNR score increases with the increasing duration of the embedding audio. (2) We can observe that without the speaker embedding from the victim (i.e., 0s), the recovered audio achieves a limited performance with an MCD score of 9.7 and SNR of -2.4dB. (3) With a short utterance ($\geqslant$2s) for speaker embedding extraction, the reconstructed audio achieves an MCD score lower than 4.5 and an SNR higher than 6.1dB, which indicates satisfactory speech intelligibility and quality. The results indicate that VibSpeech can effectively extract speaker embeddings from the short utterance of the victim to recover wideband speech.

To intuitively observe the impact of the speaker embedding, we used random noise and two different utterances from the same speaker for speaker embedding extraction and compared the spectrograms of reconstructed audio. The results are shown in Figure 14. (1) Compared the original spectrogram (Figure 14(a)) with the reconstructed spectrogram without speaker embedding (Figure 14(b)), we can find although the latter maintains low-band speech components, the reconstructed high frequencies are random-like instead of consistent with formants as shown in the original spectrogram. (2) Figure 14(c) and Figure 14(d) show the reconstructed spectrograms with speaker embeddings extracted from two different utterances of the same speaker. Compared with the original

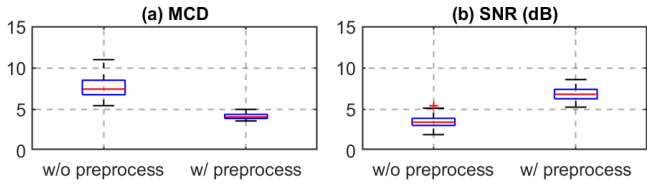

Figure 16: Impact of preprocessing.

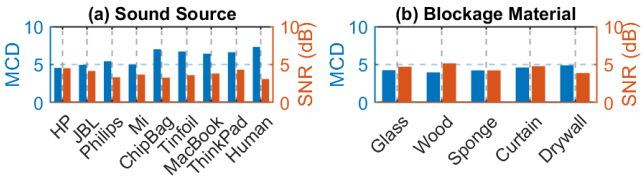

Figure 17: Impact of (a) oscillators and (b) blockages.

audio shown in Figure 14(a), we can observe that the two reconstructed audio in Figure 14(c)(d) contain most of the high-band components of up to 8kHz and the two spectrograms both show high similarity with the original spectrogram. This indicates that VibSpeech can accurately reconstruct lost formants in high band and recover wideband speech. Figure 14(e) shows the difference between the spectrograms shown in Figure 14(c)(d). We can observe that the difference is subtle, which indicates robust speaker-embedding extraction.

7.4 Impact of Sound Volume

We played audio by setting the volume into different values including 20% (63dB), 40% (68dB), 60% (72dB), 80% (78dB), and 100% (84dB). We conducted the experiment in the same scenario as shown in Figure 10(a) with a sensing distance/angle of 3m/0°. The calculated MCD and SNR scores are shown in Figure 15(b). We can observe that the performance is satisfactory in a normal sound volume (e.g., 20%-40%). We also noticed that the performance does not show great promotion but a steady improvement as the sound volume increases. The possible reason is that even though the sound volume increases, this does not change the narrowband frequency response of the vibrating target. Thus, the improvement is not as significant as we expected but shows a smooth ascent. Overall, the reconstructed audio by VibSpeech has an average MCD score of below 5.3 and an average SNR score of above 3.5dB, which indicates satisfactory performance with intelligible speech recovery.

7.5 Impact of Preprocessing

A key training strategy for vocal-guided wideband extension (Section 6.3) is generating narrowband speech from public audio datasets by applying low-pass filtering. However, as investigated in Section 5, the difference between the narrowband audio and narrowband vibration signal is that the raw vibration signals suffer distortions and artifacts in the low band. Thus, we proposed *Preprocess* in Section 6.2 to merge the gap between normal narrowband audio and distorted vibration signal. To study the impact of *Preprocess*, we compared the results of speech reconstruction with and without *Preprocess* and kept other settings the same. The tested dataset is the one we constructed in Section 7.1. As shown in Figure 16(a), we can find that the MCD score degrades from 7.7 to 4.1 when the speech is reconstructed with *Preprocess*. As shown in Figure 16(b), the SNR increases from 3.4dB to 6.7dB with over 3.3dB improvement. This indicates the *Preprocess* in Section 6.2 can effectively suppress the distortions and artifacts in the low band of vibration signals and thus contributes to speech recovery with better intelligibility and quality.

7.6 Different Oscillators

We chose different targets for speech recovery, including loudspeakers (HP, JBL, Philips, and Mi), passive films (chip bag and tinfoil, 40cm from the HP loudspeaker), built-in loudspeakers of laptops (MacBook and ThinkPad), and human speaker (male, 40cm away from the tinfoil). We placed the mmWave sensor outside the soundproof room (Figure 10(a)) with a sensing distance/angle of 3m/0° and captured the vibration of the in-room target. The loudspeakers and laptops played audio with an SPL of 68dB. The human speaker read the same corpus with around 72dB SPL and induced vibrations on the tinfoil. A microphone recorded the audio at the same time. From Figure 17(a), we can observe that the MCD ranges from 3.5 to 7.3 and the SNR score ranges from 3.1dB to 4.5dB. The results on loudspeakers are better than the ones on passive films and human speaker (MCD/SNR: 7.3/3.1dB). The reason is that the vibration amplitude on loudspeakers is larger than the sound-induced vibration amplitude on passive films by loudspeakers or a human speaker, which results in a higher SNR on the reflected mmWave signals. This also reveals that compared with passive films, loudspeakers can be more prone to be a compromised target. We can also find that the results on external loudspeakers are better than the ones on built-in loudspeakers of laptops (6.4/3.8dB, 6.6/4.3dB). The reason is that the solid-propagated vibration on the laptop shell can suffer more attenuation compared with the former. Overall, VibSpeech can both cause threats to passive films and loudspeakers. When passive films are not available in the attack scenario, VibSpeech can still cause threats to the commonly used loudspeakers and recover intelligible speech.

7.7 Blockage Materials

We performed experiments by penetrating different materials for vibration sensing. Tested materials include soundproof glass, wood (0.5cm-thick), sponge (1.5cm-thick), curtain, and drywall (2cm-thick). We placed the mmWave sensor 3m away

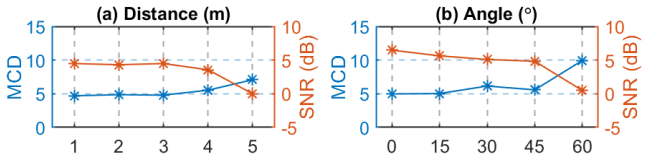

Figure 18: Impact of sensing (a) distance and (b) angle.

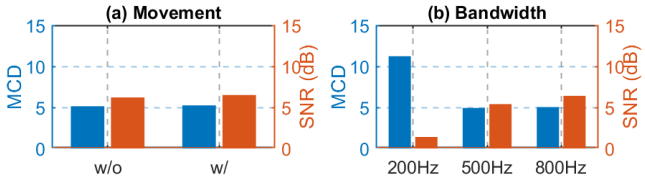

Figure 19: Impact of (a) movement and (b) raw bandwidth.

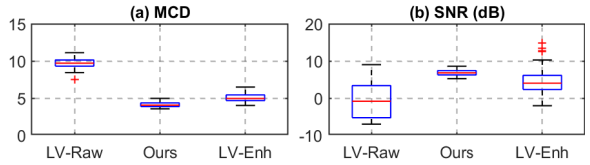

Figure 20: Comparison with laser vibrometry.

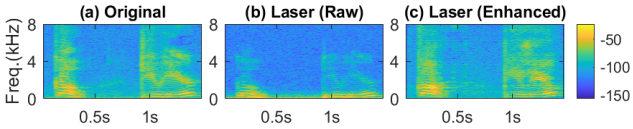

Figure 21: Spectrograms of (a) original, (b) laser-recovered, and (c) VibSpeech-enhanced audio (*"go, do you hear"*).

from the HP loudspeaker with an angle of 0°, and penetrated these materials for speech recovery. The loudspeaker played audio with an SPL of 68dB. The result is shown in Figure 17(b). We can observe that although the performance varies across different materials, the overall MCD and SNR scores are below 5.0 and above 3.9dB, respectively. This indicates that VibSpeech is resistant to common-used soundproof materials and can recover intelligible speech.

7.8 Sensing Distance and Angle

The SNR of reflected mmWave signals can change with the sensing distance and angle, and may affect the performance of speech recovery. Thus, we performed a quantitative experiment penetrating the soundproof wall (Figure 10(a)) to investigate the impact of sensing distance and angle. The distance refers to the distance between the mmWave sensor and the target (i.e., an HP loudspeaker). The angle refers to the angle between the orientation of the loudspeaker and the transmitted-beam direction of the sensor. The loudspeaker played audio with an SPL of 68dB. The results are shown in Figure 18(a)(b). We can find that when the distance increases up to 5m and the angle reaches 60°, the performance shows a noticeable degradation. The reason is that the attenuating reflection due to the increasing distance and angle causes a lower SNR on the received signals and the delicate vibrations are prone to be overwhelmed by the noise floor. But overall, with a sensing distance of around 4m and an angle within 45°, the calculated MCD and SNR scores are lower than 6.2 and above 4.8dB, which indicates a satisfactory performance.

7.9 Impact of Movement

VibSpeech extracts the range-bin of the target after the range-FFT and removes irrelevant range-bins to eliminate interferences, e.g., background movements. Considering that there can be moving subjects around the target during the attack, we conducted a comparison experiment to investigate the impact of surrounding movements. We asked a volunteer to sit on a chair and type randomly at a distance of 1m from the targeted loudspeaker (HP) in the room (Figure 10(a)). Another volunteer is asked to pace back and forth at a distance of 1.5-2.5m from the target. The loudspeaker played audio with an SPL of 68dB. The result is shown in Figure 19(a). We can find that when there are movements around the target (i.e., w/), the MCD score achieves 5.2 which is close to the score of 5.1 when without (i.e., w/o) the movements. The calculated SNR scores under the two conditions are also very close (i.e. 6.2dB for w/o and 6.5dB for w/). The result indicates that VibSpeech is resistant to surrounding movements in the attack scenario.

7.10 Bandwidth of Raw Narrowband Signal

To investigate the impact of the bandwidth of raw-acquired vibration signals, we performed low-pass filtering with different cut-off frequencies f_c on the raw vibration data to acquire narrowband signals with different bandwidths. We used the constructed dataset in Section 7.1. For each experiment, we retrained the bandwidth-extension model (Section 6.3) by setting f_c of the low-pass filter to 200Hz/500Hz/800Hz while keeping other processing (e.g., SpkEnc, Preprocess, and the vocoder) in Section 6 unchanged. We calculated scores of reconstructed audio under different f_c. As shown in Figure 19(b), we can observe that the performance is improved when f_c increases from 200Hz to 800Hz. Because a higher bandwidth of raw-vibration signal contains more speech components and contributes more to the bandwidth extension.

We also notice that the performance is limited when the raw vibration signal has a bandwidth of 200Hz with MCD>11.2 and SNR<1.4dB. The possible reason is that the limited bandwidth compromised the pitch considering the pitch ranges from 80Hz to 255Hz. Besides, we can notice that there is a significant improvement when the bandwidth of vibration signals increases from 200Hz to 500Hz. This indicates that a

Table 2: Performance on the Common Voice corpus.

Performance	Overall	A-C	B-C	C-C
MCD	3.7	9.1	10.2	3.3
SNR	6.1dB	0.1dB	0.2dB	5.9dB

wider bandwidth of the narrowband vibration signal can bring great improvement in the speech recovery and the performance is satisfactory when only with a bandwidth of 500Hz. To investigate the root cause of this result, we used a motion sensor to measure a male speaker's throat vibration when he spoke. We found that frequency components of human vocal-cords vibration can be around 500-600Hz rather than just the pitch (as shown in Figure 26 in Appendix E). Note that these basic frequencies induced by human vocal cords pass through the vocal tract to produce intelligible speech. Thus, when the narrowband (<200Hz) signal loses these frequency components, the produced high-frequency components after the vocal tract can be compromised. This is possibly another reason for the limited performance of VibSpeech when the raw signal has a bandwidth below 200Hz. Note that such narrow bandwidth (<200Hz) is rare for common sound sources as we have investigated in Section 5 and Section 7.6. Overall, VibSpeech can recover intelligible speech even though the raw vibration signal has a narrow bandwidth as low as 500Hz.

7.11 Different Vocal Tracts

To further investigate the performance of VibSpeech across different vocal tracts, we tested VibSpeech on another public dataset (i.e., Common Voice [6]). There were 400 utterances from twenty individuals (10 males and 10 females) involved. We made sure that none of the samples was used to train the model of VibSpeech. We used the trained model in Section 7.1 and kept the setup the same for speech recovery. The calculated MCD/SNR scores of recovered speech are shown in Table 2. We can find that the overall performance is satisfactory with average MCD/SNR scores of 3.7/6.1dB. The result validates the generalization ability of VibSpeech across different vocal tracts. To better understand the impact of vocal-tract features (i.e., speaker embeddings), we chose samples of three individuals and used the embedding from individual A (male), individual B (female), and individual C (male) to compensate for the narrowband speech of individual C. The results of reconstructed speech with embeddings of A, B, and C are denoted as A-C, B-C, and C-C, respectively. We can find that VibSpeech performs well with embedding C (3.3/5.9dB). But the performance degrades when with embeddings from another person (e.g., A or B). Considering that VibSpeech used target's vocal-tract features (i.e., speaker embedding) to compensate for the narrowband speech, the degradation possibly results from the differences in vocal tracts between C and A/B. The result indicates that currently vocal-tract features of the target are still required to achieve satisfactory performance.

7.12 End-to-End Attack

We performed an end-to-end attack to demonstrate the attack process of VibSpeech by penetrating the soundproof wall shown in Figure 10(a) with a sensing distance/angle of 3m/0°. We asked a volunteer to sit in the room and play the speech audio (68dB SPL) of five males and five females chosen from the 40 individuals via an HP loudspeaker. For each individual, we randomly selected one audio trace to extract his/her speaker embedding and used the remaining samples for testing. The used models for speech recovery are the ones pre-trained in Section 6. During the attack, we used the mmWave sensor to transmit and receive mmWave signals from the room outside, and derived captured vibrations of the loudspeaker on a laptop. (1) The derived vibration signals were first fed into the *Preprocess* (Section 6.2) to suppress the low-band distortion and artifacts. (2) Then the preprocessed signal and corresponding speaker embedding were fed into the vocal-guided bandwidth extension module (Section 6.3) to recover the high-frequency band. (3) Finally, the enhanced spectrogram was fed into the vocoder (Section 6.4) to generate an audible waveform.

To fully understand the performance of the end-to-end attack, we compared the performance of VibSpeech with the laser vibrometry. For the laser vibrometry, we performed the experiment in a line-of-sight condition (1m) without blockage as the same configuration shown in Figure 4(d) and used the same speech corpus as the one for VibSpeech. We calculated the scores of raw laser-recovered audio (*LV-Raw*), VibSpeech-recovered audio (*Ours*), and the laser-recovered audio but enhanced by our method (*LV-Enh*, i.e., we precessed the audio of *LV-Raw* with VibSpeech as mentioned above to acquire the audio of *LV-Enh*). The results are shown in Figure 20.

We can find that the average MCD/SNR of the laser vibrometry (*LV-Raw*) achieves 9.7/ -0.1dB which indicates the worst performance. By contrast, VibSpeech (*Ours*) performed better with an average MCD/SNR score of 4.1/6.8dB. The reason is that although the laser vibrometry has a higher resolution than the mmWave sensor for vibration sensing, the recovered audio is still distorted and has large attenuation in the high-frequency band due to the uneven frequency response of the vibration-based side channel (Section 5). The attenuated high-frequency components can cause mel distortions in the laser-recovered audio and result in poor performance. Enhanced by our proposed scheme, the performance of the laser vibrometry (*LV-Enh*) is improved with an average MCD/SNR score of 4.3/5.6dB. Figure 21 shows the result of *LV-Raw* and *LV-Enh*. Compared with the original audio shown in Figure 21(a), we can find that high-frequency components (>500Hz) of the raw laser-recovered audio attenuate significantly as shown in Figure 21(b) while the enhanced audio shown in Figure 21(c) has a wider bandwidth which contributes to higher speech intelligibility. This result also validates that VibSpeech can be applied to laser vibrometry for speech enhancement.

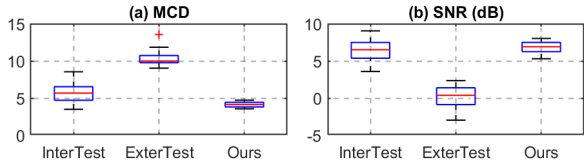

Figure 22: Compared with target-dependent speech recovery.

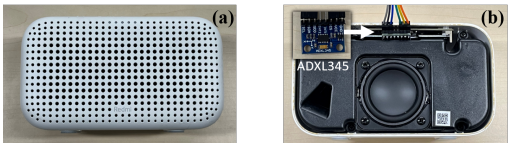

Figure 23: (a) We played audio via a smart speaker (Redmi). (b) The accelerometer captured the vibration of the speaker.

7.13 Comparison with Target-dependent Speech Recovery

To achieve intelligible speech recovery, an intuitive method is to learn the mapping between narrowband vibration signals and wideband audio of the victim using an end-to-end neural network [20, 56]. In such a case, the attacker is required to collect both vibration (narrowband) data and ground-truth (wideband) audio from the target at the same time to train the model, i.e., a target-dependent attack. Besides, such trained models represent the mapping between limited combinations of phonemes in the training set. But for the same combination of phonemes, the interval and duration of each phoneme can be different for two individuals. Next, we compared our method quantitatively with the one mentioned above.

Quantitative Experiment. For a fair quantitative comparison with related methods [20, 56], we used the same dataset constructed under the same experimental setting in Section 7.1. For the end-to-end learning, we used the widely-used Unet structure [20, 56] to learn the mapping between the spectrograms of vibration signals and corresponding ground-truth audio, and used the vocoder in Section 6.4 to generate audio waveforms. We separated the 40-individual dataset into three parts, 80% data from the first 20 individuals as the training set, the remaining 20% data from the first 20 individuals as a testing set (*InterTest*), and the data of the remaining 20 individuals as another testing set (*ExterTest*). For VibSpeech, we used the models in Section 6 and used the same dataset as *ExterTest* to show the performance on untrained targets.

Result Analysis. To compare the performance of above methods, we calculated the MCD and SNR scores of recovered audio and the results are shown in Figure 22. For the end-to-end learning, we can observe that when the testing data is from individuals included in the training set (*InerTest*), the performance is satisfactory with an MCD/SNR of 5.7/6.5dB. However, for unseen targets that are not included in the training set (*ExterTest*), the performance is poor with the MCD score increasing to over 10.3 and the SNR degrading to below 0.2dB. The result illustrates the limitation of the end-to-end learning on the cross-target attack. For unseen targets, our method outperforms the former with an MCD score lower than 4.7 and an SNR higher than 5.3dB, which indicates an intelligible and high-quality speech recovery. The results indicate that our method has a better generality for unseen targets which is more practical in a real attack case.

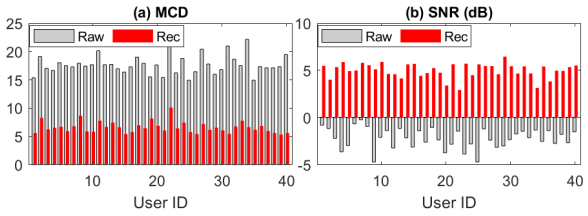

Figure 24: Overall performance of the IMU-based method.

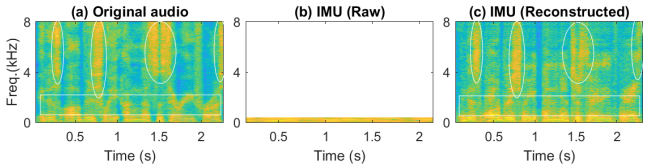

Figure 25: (a)Original, (b)IMU-recovered, and (c)VibSpeech-enhanced audio (*"it is obviously unnecessary for us"*).

8 Extension to IMU-measured Vibration

In this section, we applied the proposed scheme (Section 6) to IMU-measured vibration signals for speech recovery. Modern commercial acoustic devices [1] are equipped with accelerometers to enable tap gesture controls to pause/resume music or detect position changes of the device. Adversaries e.g., malicious service suppliers or man-in-the-middle attackers, may use the acquired vibration data from the motion sensor for eavesdropping. Due to the inaccessibility of the motion-sensor data on commercial loudspeakers, we attached an accelerometer (ADXL345) to a smart speaker's motherboard as shown in Figure 23 to reveal the threat.

End-to-End Wideband Speech Recovery. We played audio of the same corpus in Section 7.1 via the smart speaker (68dB SPL) and acquired corresponding accelerometer data over the SPI interface of a Raspberry Pi 4B at a sampling rate of 1kHz. For each individual, we randomly chose a sample and used 3s of audio for his/her speaker embedding extraction (Section 6.1). The remaining samples of the individual are used for testing. The detailed process for each audio reconstruction is as follows. (1) We upsampled the vibration signal to 16kHz and preprocessed (Section 6.2) the data to remove artifacts and correct low-band distortions. (2) We performed the vocal-guided bandwidth extension (Section 6.3) with the

preprocessed vibration data and the extracted speaker embedding, and acquired the enhanced spectrogram. (3) We fed the enhanced spectrogram into the vocoder (Section 6.4) and acquired the reconstructed audio. Note that all the models are the ones introduced in Section 6 without re-training.

Result Analysis. For comparison, we calculated scores of recovered audio from raw vibration data (*Raw*) and reconstructed audio by VibSpeech (*Rec*). The results are shown in Figure 24. From Figure 24(a), we can observe that the audio recovered from the raw IMU-captured vibration data suffers significant distortions with an average MCD score of 17.7 and SNR of -2.2dB. The reason is that the sensor can only recover frequency components lower than 500Hz due to the limited sampling rate of 1kHz according to the Nyquist theorem [42]. The absence of higher frequency components causes severe distortions and a poor SNR of the recovered audio. However, after the enhancement by VibSpeech, the IMU-recovered audio has a wider bandwidth of up to 8kHz, and thus achieves a better MCD score of 6.5 and SNR of 5.1dB. Figure 25(a)(b)(c) show the spectrograms of original audio (ground-truth), raw-IMU-captured vibration signal, and VibSpeech-reconstructed audio based on the raw IMU data, respectively. Comparing (b) with (a) in Figure 25, we can find that the raw vibration signal captured by the IMU loses high-frequency components above 500Hz. After reconstruction (Figure 25(c)), the lost speech formants are recovered which contributes to better speech intelligibility and quality.

9 Limitation and Discussion

Vibration Signal with a Narrower Bandwidth. VibSpeech achieves wideband speech recovery with up-to-8kHz bandwidth for narrowband vibration signals below 500Hz. However, we find that the performance is not satisfactory when the raw narrowband signal has an extremely limited bandwidth under 200Hz. As analyzed in Section 7.10, the vocal-cord-induced components (denoted as *source components*) can be around 500Hz, which pass through the human vocal tract to generate higher frequencies and finally produce intelligible speech. Thus, a narrowband signal with incomplete source components may cause a limited performance on wideband speech recovery by VibSpeech. The frequency aliasing should also be considered when with a low sampling rate. To make VibSpeech work under a narrower bandwidth (e.g., smartphone IMUs), a potential improvement is to investigate the inner relationship between the pitch and low-order harmonics.

A General Model for Wideband Speech Recovery. Recovering wideband audio from band-limited signals has always been a challenging problem, especially for the vibration-based side channels. Besides addressing the uneven frequency response, another key challenge lies in how to reconstruct the high-band components accurately without information loss or bias. From an attacker's point of view, an ideal solution is to design a general model that can recover high-band speech components of an arbitrary victim without any prior knowledge about the victim. However, there is a fact that the high-band components are determined mostly by human vocal-tract features (i.e., the classic source-filter model [46]) that are diverse among people. This introduces a contradiction between the goal of a **general** model and the **uniqueness** of human pronunciation. VibSpeech reveals the threat that vocal-tract features can be used to extend the bandwidth of narrowband audio once the attacker acquires a short (2s-4s) utterance from the victim. This reminds the public that people should be aware of their casual speech leakage to avoid exposure of their vocal-tract features, e.g., be careful with unknown calls or posting voice samples on the Internet.

10 Countermeasures

RF-based methods. To defend against the malicious mmWave sensor, (1) an intuitive method is to deploy electromagnetic shielding to block the transmitted mmWave signals, such as a Faraday cage. However, it is costly to deploy the shielding materials for a conference room with hundreds of square meters. By contrast, (2) passive smart surfaces [12] can be a potentially cost-effective solution to defend against malicious mmWave signals. The user can deploy the tag near the loudspeaker to manipulate reflected mmWave signals and disturb the acquired vibration in the reflected signal. To actively defend against the attack, (3) the user can use a wireless jammer to interfere with the malicious receiver. However, such defense requires parameters of the malicious device [50], such as operating frequency band and chirp duration.

Acoustic-based methods. The mmWave-based attack in this paper works in a scenario where the user plays audio via a loudspeaker. To avoid speech leakage via the outside vibration-based side channel, the user can choose a headset for confidential communication. However, note that VibSpeech not only works in the mmWave sensing methodology but also works for other vibration-based side channels, such as IMU-based speech recovery. Nowadays, IMUs are widely used in headsets and earphones [2] for touch control and position estimation, which may still be able to acquire the inner vibrations and become potential attack surfaces.

11 Conclusion

In this paper, we revealed a new speech threat that adversaries can recover wideband (up to 8kHz) intelligible speech from narrowband (<500Hz) signals of vibration-based side channels, when a short utterance (2s-4s) of the victim is exposed to the attacker. We proposed a vocal-guided general scheme (VibSpeech) and a mmWave-based prototype to demonstrate the threat. We evaluated VibSpeech with extensive experiments and validated its generality on the IMU-based method.

Acknowledgments

The authors would like to thank our Shepherd and all the anonymous reviewers for their insightful comments. This paper is partially supported by the National Key R&D Program of China (2020AAA0107700) and National Natural Science Foundation of China (62032021 and 62372406).

References

[1] Amazon launches new echo dot with motion and temperature sensors. https://www.thehindu.com/sci-tech/technology/amazon-launches-new-echo-dot-with-motion-and-temperature-sensors/article66571261.ece, 2023. [Online; accessed on October 3, 2023].

[2] Sony WH-1000XM4. https://helpguide.sony.net/mdr/wh1000xm4/v1/en/contents/TP0002928786.html, 2023. [Online; accessed on October 3, 2023].

[3] Laser vibrometer LV-FS01. https://sunnyinnova.com/products/13523.html, 2024. [Online; accessed on February 15, 2024].

[4] JD Achenbach, J Bjarnason, and T Igusa. Effect of a vibrating substructure on acoustic radiation from a cylindrical shell. 1992.

[5] S Abhishek Anand and Nitesh Saxena. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 1000–1017. IEEE, 2018.

[6] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[7] Sasan Asadiabadi and Engin Erzin. A deep learning approach for data driven vocal tract area function estimation. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 167–173. IEEE, 2018.

[8] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *NDSS*, volume 2020, pages 1–18, 2020.

[9] Suryoday Basak and Mahanth Gowda. mmspy: Spying phone calls using mmwave radars. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1211–1228. IEEE, 2022.

[10] Jacob Benesty, M Mohan Sondhi, Yiteng Huang, et al. *Springer handbook of speech processing*, volume 1. Springer, 2008.

[11] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who are you (i really wanna know)? detecting audio deepfakes through vocal tract reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2691–2708, 2022.

[12] Xingyu Chen, Zhengxiong Li, Baicheng Chen, Yi Zhu, Chris Xiaoxuan Lu, Zhengyu Peng, Feng Lin, Wenyao Xu, Kui Ren, and Chunming Qiao. Metawave: Attacking mmwave sensing with meta-material-enhanced tags. In *The 30th Network and Distributed System Security (NDSS) Symposium*, volume 2023, 2023.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[14] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J Mysore, Fredo Durand, and William T Freeman. The visual microphone: Passive recovery of sound from video. 2014.

[15] Peter A Dourmashkin. *Classical mechanics*. John Wiley & Sons, Incorporated, 2013.

[16] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[17] Shikha Gupta, Jafreezal Jaafar, WF Wan Ahmad, and Arpit Bansal. Feature extraction using mfcc. *Signal & Image Processing: An International Journal*, 4(4):101–108, 2013.

[18] Jun Han, Albert Jin Chung, and Patrick 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*, pages 181–192, 2017.

[19] Pengfei Hu, Wenhao Li, Riccardo Spolaor, and Xiuzhen Cheng. mmEcho: A mmwave-based acoustic eavesdropping method. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 836–852. IEEE Computer Society, 2022.

[20] Pengfei Hu, Yifan Ma, Panneer Selvam Santhalingam, Parth H Pathak, and Xiuzhen Cheng. Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 11–20. IEEE, 2022.

[21] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth Pathak, Guoming Zhang, and Xiuzhen Cheng. Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1757–1773. IEEE, 2022.

[22] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.

[23] Corentin Jemine. Real-time-voice-cloning. *University of Liége, Liége, Belgium*, page 3, 2019.

[24] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[25] Jong Hwan Ko, Josh Fromm, Matthai Philipose, Ivan Tashev, and Shuayb Zarar. Limiting numerical precision of neural networks to achieve real-time voice activity detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2236–2240. IEEE, 2018.

[26] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[27] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[28] Andrew Kwong, Wenyuan Xu, and Kevin Fu. Hard drive of hearing: Disks that eavesdrop with a synthesized microphone. In *2019 IEEE symposium on security and privacy (SP)*, pages 905–919. IEEE, 2019.

[29] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

[30] Qianru Liao, Yongzhi Huang, Yandao Huang, Yuheng Zhong, Huitong Jin, and Kaishun Wu. MagEar: eavesdropping via audio recovery using magnetic side channel. In *MobiSys*, pages 371–383, 2022.

[31] Feng Lin, Chao Wang, Tiantian Liu, Ziwei Liu, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. High-quality speech recovery through soundproof protections via mmwave sensing. *IEEE Transactions on Dependable and Secure Computing*, 2023.

[32] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 97–110, 2021.

[33] Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenyao Xu, Ming-Chun Huang, and Kui Ren. WavoID: Robust and secure multimodal user identification via mmwave-voice mechanism. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, 2023.

[34] Yan Long, Pirouz Naghavi, Blas Kojusner, Kevin Butler, Sara Rampazzi, and Kevin Fu. Side Eye: Characterizing the limits of pov acoustic eavesdropping from smartphone cameras with rolling shutters and movable lenses. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023.

[35] Antonia Maxon and Diane Brackett. The hearing-impaired child: Infancy through high school years. 1992.

[36] William McGrath. Technique and device for through-the-wall audio surveillance, October 6 2005. US Patent App. 11/095,122.

[37] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. Gyrophone: Recognizing speech from gyroscope signals. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 1053–1067, 2014.

[38] Ralph P Muscatell. Laser microphone. *The Journal of the Acoustical Society of America*, 76(4):1284–1284, 1984.

[39] Ben Nassi, Yaron Pirutin, Tomer Galor, Yuval Elovici, and Boris 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*, pages 1900–1914, 2021.

[40] Ben Nassi, Yaron Pirutin, Raz Swisa, Adi Shamir, Yuval Elovici, and Boris Zadov. Lamphone: Passive sound recovery from a desk lamp's light bulb vibrations. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4401–4417, 2022.

[41] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[42] Emiel Por, Maaike van Kooten, and Vanja Sarkovic. Nyquist–shannon sampling theorem. *Leiden University*, 1(1), 2019.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[44] Nirupam Roy and Romit Roy Choudhury. Listening through a vibration motor. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 57–69, 2016.

[45] Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. Spying with your robot vacuum cleaner: eavesdropping via lidar sensors. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 354–367, 2020.

[46] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[47] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[48] Chao Wang, Feng Lin, Zhongjie Ba, Fan Zhang, Wenyao Xu, and Kui Ren. Wavesdropper: Through-wall word detection of human speech via commercial mmwave devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–26, 2022.

[49] Chao Wang, Feng Lin, Tiantian Liu, Ziwei Liu, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. mmPhone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 820–829. IEEE, 2022.

[50] Chao Wang, Feng Lin, Tiantian Liu, Kaidi Zheng, Zhibo Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Kui Ren. mmEve: eavesdropping on smartphone's earpiece via cots mmwave device. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 338–351, 2022.

[51] Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani 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*, pages 1–14, 2020.

[52] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 130–141, 2015.

[53] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao 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*, pages 14–26, 2019.

[54] Chen Yan, Guoming Zhang, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. The feasibility of injecting inaudible voice commands to voice assistants. *IEEE Transactions on Dependable and Secure Computing*, 18(3):1108–1124, 2019.

[55] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9458–9465, 2020.

[56] Feng Yiwen, Zhang Kai, Wang Chuyu, Xie Lei, Ning Jingyi, and Chen Shijia. mmEavesdropper: Signal augmentation-based directional eavesdropping with mmwave radar. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.

A Loss Function of SpkEnc in Section 6.1

The Generalized End-to-End (GE2E) loss [47] characterizes the similarity of embeddings from different speakers by building a similarity matrix. Considering the embeddings $e_{ij}(1 \leqslant i \leqslant N, 1 \leqslant j \leqslant M)$ of M utterances from N speakers, the similarity matrix $S_{ij,k}$ is calculated from a two-by-two comparison of all embeddings e_{ij} against the embedding centroid $c_k(1 \leqslant k \leqslant N)$ for every speaker:

$$S_{ji,k} = \begin{cases} \omega \cdot \cos(e_{ji}, c_j^{(-i)}) + b & \text{if } k = j; \\ \omega \cdot \cos(e_{ji}, c_k) + b & otherwise, \end{cases} \tag{4}$$

where $c_k = \frac{1}{M}\sum_{m=1}^{M} e_{km}$, $c_j^{(-i)} = \frac{1}{M}\sum_{m=1}^{m \neq i} e_{jm}$. ω and b are learnable parameters of the network. Then the loss on each

embedding e_{ji} and GE2E loss L_G can be calculated by

$$L(e_{ji}) = S_{ji,j} - \log \sum_{k=1}^{N} \exp(S_{ji,k}), L_G = \sum_{j,i} L(e_{ji}). \quad (5)$$

B The Algorithm in Section 6.2

Algorithm 1: Artifacts&Distortion Suppression

Input: s: raw vibration signal
Output: $\hat{s}$: processed signal

$N_{seg} = 512, N_{overlap} = 256, bw = 20, f_c = 500$;
$s_0, s_1, ..., s_N = OverlapSegment(s, N_{seg}, N_{overlap})$;
for $n = 1, ..., N$ **do**
    $f_0 = PitchEstimation(s_n)$;
    $\hat{s_n} = \sum_{m=1}^{M} \frac{|S_1|}{|S_m|} BPF(s_n, m \cdot f_0, bw)$;
$\hat{s} = OverlapAdd(\hat{s_1}, \hat{s_2}, ..., \hat{s_N}, N_{seg}, N_{overlap})$;
$\hat{s} = LPF(\hat{s}, f_c)$;
return $\hat{s}$

Table 3: Parameters of the generator.

Block	Parameter
Blk_0	ConvBlk(2,64)
Blk_1	FTB(64,128), MaxPool, ConvBlk(64,128)
Blk_2	FTB(128,64), MaxPool, ConvBlk(128,256)
Blk_3	FTB(256,32), MaxPool, ConvBlk(256,512)
Blk_4	FTB(512,16), MaxPool, ConvBlk(512,1024)
Blk_5	UpConv(1024,512), Concat, ConvBlk(1024,512)
Blk_6	UpConv(512,256), Concat, ConvBlk(512,256)
Blk_7	UpConv(256,128), Concat, ConvBlk(256,128)
Blk_8	UpConv(128,64), Concat, ConvBlk(128,64)
Blk_9	Conv2d(64,1)

Table 4: Parameters of the discriminator.

Block	Parameter
Blk_0	Conv2d(2,64), BatchNorm(64), ReLU
Blk_1	Conv2d(64,128), BatchNorm(128), ReLU
Blk_2	Conv2d(128,256), BatchNorm(256), ReLU
Blk_3	Conv2d(256,512), BatchNorm(512), ReLU
Blk_4	Conv2d(512,1), Sigmoid

C Details about the Model in Section 6.3

Discriminator Loss: Considering a predicted label x and a true label y, the Binary Cross Entropy (BCE) loss

$$L_{BCE} = -(y \log(x) + (1-y) \log(1-x)) \quad (6)$$

We denote the narrowband mel-spectrogram, reconstructed mel-spectrogram, and the ground-truth mel-spectrogram as m_{nb}, m_{re} and m_{gt}, respectively. The loss of the discriminator

$$L_D = L_{BCE}(D(m_{re}, m_{nb}), 0) + L_{BCE}(D(m_{gt}, m_{nb}), 1), \quad (7)$$

where $D(\cdot)$ represents the output of the discriminator.
Generator Loss: The loss of the generator L_G consists of two parts, i.e., the adversarial loss L_{adv} and the mel loss L_{mel}:

$$L_{adv} = L_{BCE}(D(m_{re}, m_{nb}), 1), \quad (8)$$

$$L_{mel} = L_1(m_{re}, m_{gt}), L_G = L_{adv} + 0.5 \cdot L_{mel}. \quad (9)$$

Table 3 and Table 4 show the model parameters.

D Details about the Used Speech Corpus

The dataset used to train models in Section 6.3 and Section 6.4 includes *train-clean-100*, *train-clean-360*, and *train-other-500* from LibriSpeech [41]. The test set for evaluation (Section 7) includes *dev-clean* subset from LibriSpeech. Note that there is no overlap among these datasets to ensure the experimental result is target-independent.

Table 5: Training sets in Section 6.3 and Section 6.4.

dataset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Table 6: Test set in the evaluation (Section 7).

dataset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40

E Frequency of Human Throat Vibration

We attached an accelerometer to a male speaker's throat area to capture the vibration signal when the speaker said /m/.

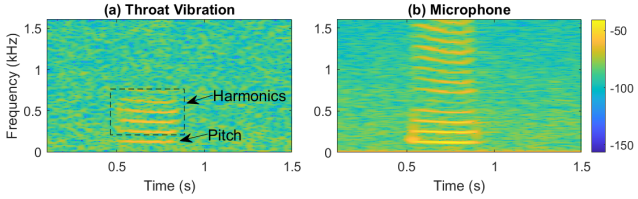

Figure 26: Spectrograms of (a) throat vibration and (b) microphone audio when a male speaker pronounced /m/.