Cancelling Speech Signals for Speech Privacy Protection against Microphone Eavesdropping

Ming Gao[1,2,†], Yike Chen[1,2,†], Yajie Liu[1,2], Jie Xiong[3], Jinsong Han[1,2,✉], Kui Ren[1,4]
[1]Zhejiang University, China
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, China
[3]University of Massachusetts Amherst, USA
[4]Zhejiang Provincial Key Laboratory of Blockchain and Cyberspace Governance, China
{gaomingppm,chenyike,yajie,hanjinsong,kuiren}@zju.edu.cn,jxiong@cs.umass.edu

ABSTRACT

Ultrasonic microphone jammers protect speech privacy from being eavesdropped by leveraging microphones' non-linearity. However, existing jammers merely introduce independent noises and are vulnerable to capable adversaries who adopt advanced denoising techniques. We propose a novel jammer, namely MicFrozen. It reduces the signal-to-noise ratio (SNR) at the adversary's microphone from two perspectives, i.e., cancelling speech signals and adding noises that are difficult to be removed. It effectively cancels out the protected speech signals at the adversary without compromising the delivery of the signal to the targeted individual. MicFrozen further adds coherent noises that are coupled with the speech signals to resist removal by the adversary. Extensive evaluations show that MicFrozen can cause a low SNR (-13.6 dB) at the adversary and up to 96.9% of speech signals are unrecognized at the adversary even if state-of-the-art denoising techniques are adopted by the adversary. Comprehensive experiments demonstrate the effectiveness of MicFrozen confronted by capable adversaries.

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**.

KEYWORDS

Privacy protection, Anti-eavesdropping

† Ming Gao and Yike Chen contribute equally in this paper.
✉Jinsong Han is the corresponding author.

ACM MobiCom '23, October 2–6, 2023, Madrid, Spain

ACM ISBN 978-1-4503-9990-6/23/10…$15.00
https://doi.org/10.1145/3570361.3592502

ACM Reference Format:
Ming Gao, Yike Chen, Yajie Liu, Jie Xiong, Jinsong Han, Kui Ren. 2023. Cancelling Speech Signals for Speech Privacy Protection against Microphone Eavesdropping. In *The 29th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '23), October 2–6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3570361.3592502

1 INTRODUCTION

Eavesdropping via microphones poses a serious threat to privacy. Spy microphones can be used for eavesdropping [72], forging voiceprint [85], and inferring physical keystrokes [62]. Ubiquitous devices such as smartphones and voice assistants further exacerbate this threat [3, 71, 77]. Ultrasonic microphone jammers (UMJs) [15, 48, 63, 69] are proposed as a user-friendly solution to prevent microphone-based eavesdropping. UMJs modulate jamming noises on ultrasounds that are inaudible to humans. The noises are then injected into spy microphones, leveraging microphones' nonlinear characteristic [1, 24] to combat eavesdropping. In this case, adversaries cannot easily obtain private information from the polluted sound recordings.

Unfortunately, the capability of adversaries is underestimated and the state-of-the-art (SOTA) UMJs are still vulnerable to adversarial countermeasures. Existing methods introduce diverse noises to combat eavesdropping. However, these noises are independent of speech signals. Although these noises mask private speech signals, the original speech signals still exist in the recordings without being distorted. Capable adversaries can leverage such independence [18, 61] to remove the jamming noises. Advanced noise removal techniques such as blind source separation (BSS) can support private information recovery from jammed recordings. Therefore, adversaries are capable of bypassing the UMJs and still extracting private information from recordings. For instance, without adversarial countermeasures, Sun *et al.* [69] achieved a low speech recognition rate of no more than 6% in the jammed recordings. However, after countermeasures were applied, the jamming performance degrades significantly with 75% of words recognized [69]. This is a general

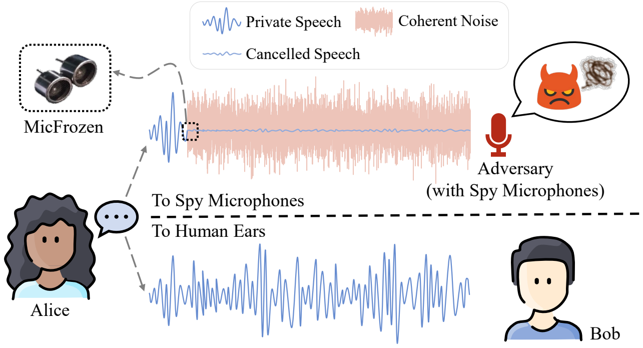


Figure 1: MicFrozen leverages ultrasound to jam spy microphones without affecting human hearing. It couples complex coherent noises and cancels speech signals actively for reducing SNR in illegal recordings.

problem associated with the SOTA UMJs [12], which greatly limits UMJs' jamming capability in practice.

Instead of merely increasing the intensity of the independent noises, we propose to improve UMJs' design from two aspects to reduce the signal-to-noise ratio (SNR) of private speech signals at the adversary. The first aspect is to weaken the intensity of speech signals at the adversary through signal cancellation. The second aspect is to further cover speech residues with dedicated noises which are difficult to be removed. The basic idea is illustrated in Fig. 1. Specifically, inspired by active noise cancellation (ANC) techniques [10, 11], we design an anti-speech signal to weaken speech signals recorded at spy microphones. It has the same frequency and amplitude but opposite phase as the speech signal. We measure the inverse channel from the sound source to a known microphone, based on which the inverse channel to the spy microphones is coarsely estimated. The obtained inverse channel is used to generate the anti-speech signal carried by ultrasound and injected into the spy microphone. On the other hand, to defend against adversarial countermeasures (e.g., removing the added noise), we replace the independent noises used in SOTA UMJs with carefully designed coherent noises that are difficult to be removed.

To realize such a UMJ, we need to overcome two core design challenges. *i) Inverse channel estimation:* Traditional ANC techniques [10, 11] can only cancel sound in a target area. It requires feedback microphones placed near the target area to estimate the inverse channel for effective cancellation. However, in our task, the spy microphones' locations are unknown. *ii) Noise generation:* Adversaries would take measures to separate and remove the jamming noises in time, frequency, or spatial domains. Designing noises which can resist various denoising techniques is non-trivial.

To address the above challenges, we present MicFrozen, a novel UMJ design to prevent potential eavesdropping behaviors. Different from traditional methods which place the reference microphone near the target area, we show that by placing the reference microphone close to the sound source, we can cancel out the speech signal at the adversary by a significant amount (e.g., 80%) without a need of knowing the spy microphone's location. On the other hand, MicFrozen generates coherent noises that are coupled with speech signals to improve the resilience against adversarial countermeasures such as signal separation and removal. The dependence of the coherent noises with respect to the protected speech signals degrades the performance of denoising techniques significantly [18, 61]. Note that as the proposed system does not just rely on signal cancellation but also adds noises to prevent eavesdropping, even if the signal is not completely cancelled out, the end performance is not greatly affected.

We prototype MicFrozen using commercial off-the-shelf (COTS) hardware. We conduct comprehensive experiments to evaluate the performance and resilience of the system. MicFrozen yields a low SNR of -13.6 dB in the illegal recordings at the spy microphone, leading to less than 4% of words being correctly recognized by automatic speech recognition (ASR) algorithms and humans. It still maintains an SNR of -14.3 dB and prevents over 86.9% of speech signals from being recognized against various denoising techniques. The demo of MicFrozen is presented in [76].

We summarize the contributions of MicFrozen as follows:

- We propose a novel UMJ, MicFrozen, to protect speech privacy. It enhances protection effectiveness by cancelling speech signals and adding coherent noises simultaneously.
- We probe the acoustic propagation with a mathematical model. It supports the inverse-channel estimation without the need of knowing the spy microphone's location. MicFrozen accordingly generates anti-speech signals in a real-time and adaptive manner for cancelling private speech signals.
- We prototype MicFrozen using low-cost COTS components with extensive evaluations on commercial microphones. We validate its effectiveness and resilience against capable adversaries, achieving an adversarial speech recognition rate below 4%.

2 ACOUSTIC NON-LINEARITY

Various microphone jammers, based on electromagnetic interference (EMI) [46], audible sound [52, 58], or ultrasound [14, 15, 48, 63, 66, 69], are proposed to obfuscate audio signals recorded at spy microphones to protect private information. Among them, the EMI-based jammers require prior knowledge about the target spy devices (e.g., the frequencies of the EMI signals [46]), while the audible ones interfere with the user's normal conversation. In comparison, UMJs leverage the ubiquitous non-linear property of microphones and work on the inaudible bands. Non-linearity occurs when the input sound $x(t)$ occupies a high frequency over 25 kHz. We have the output recording $y(t) = A_1x(t) + A_2x^2(t) + A_3x^3(t) + \cdots$, where A_i is the gain. Here, terms with A_i ($i \geq 3$) can be ignored due to the low power.

Given an input $x(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t)$, ($f_1, f_2 >$ 20 kHz), microphones would output the sum of six tones of frequencies f_1, f_2 (i.e., derived from the term $A_1x(t)$), $2f_1$, $2f_2$, $|f_1 - f_2|$, and $f_1 + f_2$ (i.e., derived from the term $A_2x^2(t)$). Among these components, the term $|f_1 - f_2|$ survives after the low-pass filter processing in microphones. As a result, two ultrasonic signals produce an audible sound in microphones.

By leveraging this acoustic non-linearity, UMJs can inject audible noises into the spy microphone. SOTA UMJs transmit two signals together, i.e. the modulated noise $n(t)\cos(2\pi f_c t)$ and the carrier $\cos(2\pi f_c t)$, where $n(t)$ is the low-frequency noise and f_c is the carrier frequency. The spy microphone receives the added noise as

$$N(t) = A_2[n(t) + 0.5n^2(t)], \tag{1}$$

in which the high-power noise component makes the spy microphone 'deaf' [65].

Various UMJs are proposed leveraging the acoustic non-linearity to inject noises into spy microphones. Unfortunately, they all adopt independent noises, i.e., noises whose frequencies vary following specific patterns [15, 48] or Gaussian noises [17, 28, 63, 69, 74, 75]. They are vulnerable against advanced denoising techniques.

3 THREAT ANALYSIS

We assume that an adversary aims at eavesdropping on victims' speech. The adversary can place microphones near the victims. Furthermore, a capable adversary could be aware of the existence of UMJs and try to avoid being interfered [69].

In practice, the adversary will not give up eavesdropping when capturing audios full of noise. Instead, the adversary would employ denoising techniques for obtaining clean speeches. Generally, the adversary could separate jamming noises from the speech signals in three different domains, i.e., time domain, frequency domain, and spatial domain.

Time domain. The time-domain characteristics of noises allow temporal denoising [83]. Relying on multiple microphones, blind speech separation (BSS) [51], a representative signal separation method, is widely utilized for separating independent signals.

Frequency domain. The adversary could observe the frequency characteristics of jamming noise using spectrum analysis techniques (e.g., Fast Fourier Transform (FFT)). The adversary could employ high-pass, low-pass, or band-pass filters to remove noise in specific frequency bands.

Spatial domain. The spatial difference between signal/noise sources enables the adversary to remove jamming noises. Beamforming is a representative spatial domain method widely used to enhance signal strength coming from specific directions. Multiple microphones are used to distinguish the arrival angle of signals and recover targeted audio signals [69].

Sniffer-assisted techniques. A recent work [32] presents a new method to capture the detailed features of jamming noises modulated on ultrasound. Accordingly, the adversary could utilize denoising methods such as adaptive noise filtering to separate and remove noises.

In practice, users have no knowledge of what kind of denoising countermeasures the adversary may use. Therefore, a powerful UMJ needs to protect private speech signals against diverse denoising methods.

4 SYSTEM OVERVIEW

We briefly introduce the mechanism of UMJs and our two key design modules, i.e., the speech signal cancellation module and the coherent noise coupling module.

We denote the speech signal to be protected as $s(t)$, and the illegal recording acquired by an adversary as $r(t)$. Without jammers, the adversary would receive $r(t) = s(t) + n_0(t)$, where $n_0(t)$ is the inherent channel noise. SOTA UMJs introduce additional noise $N(t)$ in the form of $\frac{1}{2}n^2(t) + n(t)$ as illustrated in Eq. 1 and the signal received at the adversary can be represented as

$$r(t) = s(t) + N(t) + n_0(t). \tag{2}$$

Thus, the SNR in recordings drops from $\frac{|s(t)|}{|n_0(t)|}$ to $\frac{|s(t)|}{|N(t)+n_0(t)|}$. SOTA UMJs adopt high-power ultrasound to carry various jamming noises $n(t)$. They use Gaussian noise [17, 28, 63, 69, 74, 75] and noises whose frequencies vary following specific patterns such as hopping [15] and sweeping [48]. However, these adopted noises are independent of speech signals that still exist in illegal recordings without being distorted. Hence, SOTA UMJs are ineffective against denoising techniques adopted by more sophisticated adversaries [12, 69].

Instead, we protect speech privacy from two perspectives, i.e., reducing the intensity of speech signals received at the spy microphone and introducing more effective noises. We estimate the inverse channel to generate anti-speech signals $-\hat{s}(t)$ in real time to cancel speech signals, namely *Speech Signal Cancellation* module, in Sec. 5. We redesign jamming noises. The coherent noises $N(s(t))$ are coupled with speech signals and are difficult to be removed by denoising techniques, namely *Coherent Noise Coupling* module in Sec. 6. Thus, Eq. 2 can be rewritten as

$$r(t) = s(t) - \hat{s}(t) + N(s(t)) + n_0(t). \tag{3}$$

Therefore, the SNR in illegal recordings sharply decreases to

$$SNR = \frac{|s(t) - \hat{s}(t)|}{|N(s(t)) + n_0(t)|}. \tag{4}$$

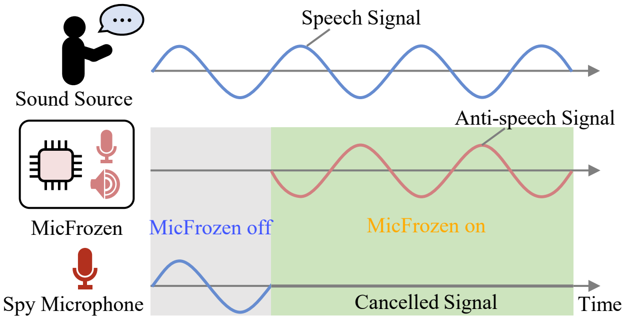


Figure 2: MicFrozen generates anti-speech signals to cancel the private speech signals and thus the spy microphone would capture no information.

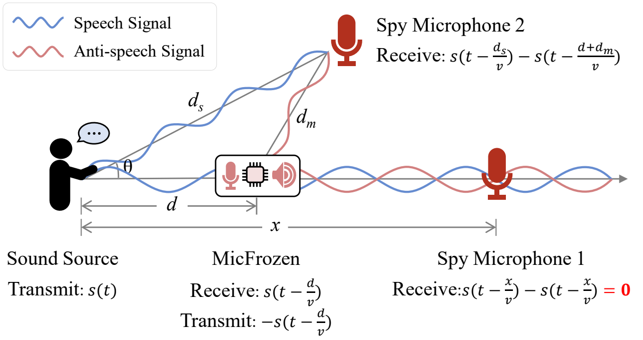

Speech Signal
Anti-speech Signal
Spy Microphone 2
Receive: $s(t-\frac{d_s}{v})-s(t-\frac{d+d_m}{v})$
d_s
d_m
θ
d
x
Sound Source
Transmit: $s(t)$
MicFrozen
Receive: $s(t-\frac{d}{v})$
Transmit: $-s(t-\frac{d}{v})$
Spy Microphone 1
Receive:$s(t-\frac{x}{v})-s(t-\frac{x}{v})=0$

Figure 3: Sound propagation in one-dimension (Spy Microphone 1) and multi-dimension (Spy Microphone 2).

With joint efforts from the two aspects, MicFrozen achieves a more resilient jamming performance. Moreover, we also consider practical issues with solutions proposed in Sec. 7.

5 SPEECH SIGNAL CANCELLATION

We cancel speech signals to reduce the leakage of speech at spy microphones. The basic idea is illustrated in Fig. 2. We propose a mathematical model to support global inverse-channel estimation. We then realize real-time anti-speech signal generation for speech signal cancellation.

5.1 Model of Inverse Channel

We model the inverse channel for signal cancellation. The model enables global inverse-channel estimation and supports timely anti-speech signal generation.

Traditional ANC techniques [10, 11] can cancel noises in a small area (usually within 30 cm^2) by placing a feedback microphone near the target area. However, in our scenario, the spy microphone's location is unknown and traditional methods do not work well. To jam a larger area, we propose to place the reference microphone near the sound source. Through theoretical modeling and experiments, we show that this deployment strategy can effectively jam a larger area without a need to know the spy microphone's location. We first consider a one-dimensional sound propagation model and then extend this model to multiple dimensions.

As shown in Fig. 3, MicFrozen is placed at a distance of d from the sound source and the spy microphones are placed at a distance of x ($x > d$) away. The speech signal arriving at MicFrozen can be represented as $s(t - d/v)$, where v is the speed of sound propagation in the air. To cancel $s(t)$, MicFrozen plays the anti-speech signal as follows,

$$-\hat{s}(t) = -s(t - \frac{d}{v}). \tag{5}$$

5.1.1 One-dimensional Cancellation. In the 1-D scene, Spy Microphone 1, the sound source, and MicFrozen are located on the same line. We denote $\alpha_L(x)$ as the multiplicative attenuation function of audible sound. In this scene, the path length difference between source-spy and MicFrozen-spy is a constant. The signal received at MicFrozen resulted from this constant path difference can be denoted as $s(t - d/v)$. If now MicFrozen transmits an opposite signal $-s(t - d/v)$, the received signal at Spy Microphone 1 becomes

$$r(t, x) = \alpha_L(x)[s(t - \frac{x}{v}) - s(t - \frac{d}{v} - \frac{x-d}{v})] = 0. \tag{6}$$

In this scene, the speech signal is completely cancelled out.

5.1.2 Multi-dimensional Cancellation. In reality, the spy microphone does not necessarily have to be located on the same line as the sound source and MicFrozen. We extend the 1-D scene to the two-dimensional (2-D) scene, where the sound source, MicFrozen, and Spy Microphone 2 are on a 2-D plane.

As shown in Fig. 3, the speech signal from the source and its opposite signal from MicFrozen propagate along different paths to Spy Microphone 2. We denote the distance from the spy microphone to the sound source as d_s and that to MicFrozen as d_m. The received signal at Spy Microphone 2 is

$$r(t, x) = \alpha_L(x)[s(t - \frac{d_s}{v}) - s(t - \frac{d_m + d}{v})]. \tag{7}$$

For this received signal, the phase difference between the speech signal and the anti-speech signal is not 180° anymore. The extra phase difference $\Delta\phi$ caused by the path length difference can be represented as

$$\Delta\phi = \frac{2\pi f_s}{v}(\sqrt{d_s^2 + d^2 - 2 \cdot d_s \cdot d\cos\theta} + d - d_s), \tag{8}$$

where f_s is the frequency of the speech signal. For human voice with a frequency of 200 Hz, $\Delta\phi$ is below 10° when θ is within the range of −30° to 30°. Such a small phase difference has little effect on signal cancellation. Even if the spy microphone is not on the same line, provided that θ is not big, the cancellation scheme leveraging the inverse channel measured by a reference microphone is still effective. In this case, we do not need to know the spy microphone's location.

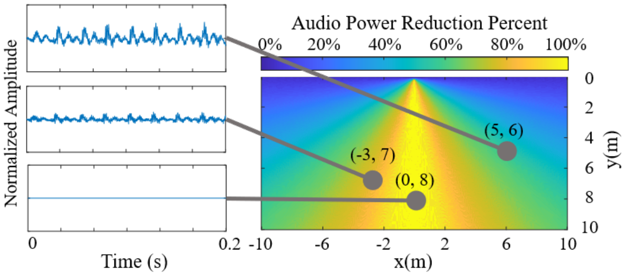


Figure 4: Simulation effect of anti-speech signals for 2-D signal cancellation. MicFrozen locates at (0, 0.2)m.

Based on the two equations, a better cancellation performance can be achieved with a smaller d. Thus, we suggest placing MicFrozen near the sound source for a more effective signal cancellation at the spy microphone. By also taking the latency requirement (Sec. 5.3) into consideration, we place MicFrozen 20 cm away from the sound source.

We further illustrate the performance of speech signal cancellation in 2D using MATLAB simulation [13]. We place a sound source at (0, 0) in a 20 $m \times 10$ m free space. The sound source plays speech signals from the audio database AudioMNIST [9] and Librispeech [59]. Another omnidirectional sound source emits inverse signals at (0, 0.2 m). To simplify the issue, we ignore the additional distance and direction attenuation during propagation. Heatmap in Fig. 4 shows the cancellation effect of anti-speech signals, where the yellow color indicates a better cancellation and the blue color indicates a worse one. 100% speech signal cancellation occurs along the line parallel to the y-axis from the first sound source to the second sound source emitting inverse signals, which matches the theoretical analysis in Sec. 5.1.1. When the spy microphone deviates from this line (i.e., $\theta \geq 0$), the cancelling performance gradually decreases. Nevertheless, it still supports signal cancellation with a power reduction over 80% within an angle range of 55° ($-27.5°$ to $+27.5°$). When the transducer is located at (0, 0.4 m), the coverage angle range is just slightly decreased to 45°. This simulation result shows that MicFrozen is able to cancel the speech signal in a relatively large area. Note that the proposed system does not only rely on this speech signal cancellation module but further adds coherent noise to the residual signals.

5.2 Anti-speech Signal Modulation

We modulate the anti-speech signals on ultrasound for practical implementation. However, during the demodulation process, distortion can occur and harmonics are induced. We therefore modulate the ultrasound a second time to make sure that the speech signal $-\hat{s}(t)$ carried could be demodulated without inducing distortion or harmonics. If we directly replace $n(t)$ with $-\hat{s}(t)$ in Eq. 1, the additional term $\hat{s}(t)^2$ remains after cancellation between $s(t)$ and $-\hat{s}(t)$. This $\hat{s}(t)^2$ term also contains information of $s(t)$ and a capable adversary could recover speech signals from it. To address this issue, we reshape the signal as follows,

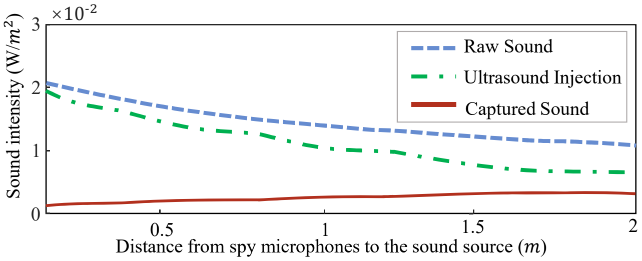


Figure 5: Using the anti-speech signal, we can significantly reduce the intensity of sound received by spy microphones even if it cannot be cancelled completely.

$$\begin{aligned} n(t) &= -\hat{s}(t) - 0.5\hat{s}^2(t), \\ N(t) &= -\hat{s}(t) + 0.5\hat{s}^3(t) + 0.125\hat{s}^4(t), \end{aligned} \tag{9}$$

in which the power of the high-order component $0.5\hat{s}^3(t) + 0.125\hat{s}^4(t)$ is too low to be recognized by human ears or voice assistants. Hence, these high-order terms can be ignored and injected signal is just $-\hat{s}(t)$.

We further consider the difference between audible sound and ultrasound attenuation. In real-world scenes, ultrasound attenuates faster than audible sound. We denote the overall attenuation of high-frequency ultrasound as $\alpha_H(x)$, which consists of the attenuation caused by two ultrasounds and non-linearity coefficient A_2. Then we rewrite Eq. 6 as

$$\begin{aligned} r(t,x) &= \alpha_L(x) \cdot s(t - x/v) - \alpha_H(x) \cdot s(t - d/v - (x-d)/v) \\ &= [\alpha_L(x) - \alpha_H(x)] \cdot s(t - x/v). \end{aligned} \tag{10}$$

We implement a pilot experiment to validate the presented model. A speaker serves as the sound source and emits audio signals whose intensity is approximately the average loudness of human voice (0.02 W/m^2 [63]). Another speaker is placed close-by and emits anti-speech signals. A spy microphone is placed at the sound source's transmission direction. We vary its distance to the source from 10 cm to 200 cm at a step size of 5 cm. As shown in Fig. 5, with the aid of the anti-speech signals, the signal intensity at the spy microphone is significantly reduced. More evaluations on practical 2D setup are presented in Sec. 8.3.

5.3 Real-time Signal Generation

Real-time signal generation is critical for cancellation. If signal processing takes too much time and anti-speech signals do not arrive in time, speech signals cannot be cancelled out.

We design a real-time anti-speech signal generation method, as shown in Fig. 6. It consists of four modules, i.e., a reference microphone M_r, a processor, an anti-speech speaker, and a

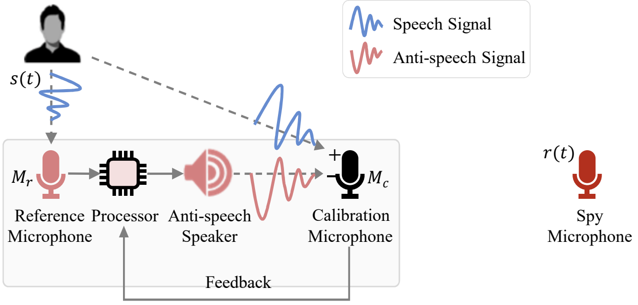


Figure 6: Architecture of the real-time anti-speech signal generation.

calibration microphone M_c. We denote the air channel from the source to M_r and M_c as h_{sr} and h_{sc}, respectively. The reference microphone is placed near the sound source and the signal received at M_r is $h_{sr}*s(t)$ where $*$ denotes the convolution operation. Then the processor generates the anti-speech signal using an adaptive filter h_{rc}, i.e., $-\hat{s}(t) = h_{rc}*(h_{sr}*s(t))$. The anti-speech signal will be modulated to the ultrasonic band following Eq. 9 and emitted out. It will be demodulated due to acoustic non-linearity at M_c. Note that in a short term, M_c is not required but the phase errors can accumulate over time. M_c then feeds back the error

$$e(t) = h_{ac} * (h_{rc} * (h_{sr} * s(t))) + h_{sc} * s(t), \tag{11}$$

where h_{ac} is the overall channel from the anti-speech speaker to M_c, including the air channel and non-linearity coefficient. To achieve effective cancellation, $e(t)$ should be close to 0. We design h_{rc} as follows,

$$h_{rc} = h_{ac}^{-1} * h_{sc} * h_{sr}^{-1}. \tag{12}$$

To estimate these time-variant channels, traditional ANC method leverages normalized least mean square (NLMS) algorithm [31, 34, 79, 82]. It keeps adjusting the estimated channel parameters following a stochastic gradient descent iteration strategy to reduce the value of $e(t)$.

To compensate for the latency, we utilize the lookahead ANC (LANC) method [67]. It separates the anti-speech speaker and M_r by a few inches and connects them using a wire or wireless link. Such LANC method works because electromagnetic signal travels much faster (3×10^8 m/s) than sound (340 m/s) in the air. Combining the NLMS algorithm, the adjustment of h_{rc} follows the steepest gradient descent on the square of $f(t)$ ($h_{rc}^{new} = h_{rc}^{old} - \frac{\mu}{2}\frac{\partial f^2(t)}{\partial h_{rc}}$), where the step size μ controls the speed of gradient descent. In algorithm implementation, we randomly initialize a finite vector with M taps as h_{rc}, and then iterate each filter tap $h_{rc}(k)$ as

$$h_{rc}^{new}(k) = h_{rc}^{old}(k) + \mu(\delta + ||x(k)||^2)^{-1} x(k) * f(t-k). \tag{13}$$

By enlarging the source-MicFrozen distance, we obtain a better ANF to generate anti-speech signals. Note that there is a trade-off between the signal cancelling performance and system latency requirement. A smaller source-MicFrozen distance brings more accurate cancellation but also poses a higher bar to meet the latency requirement. Based on our experience, a distance of 20 cm maintains a good balance. Accurate cancellation can be achieved while the time latency is still large enough for our system to complete all the processing. We evaluate the system latency in Sec. 8.4 to demonstrate that MicFrozen can meet the latency requirement.

In practice, although we do not have the knowledge of the channel with respect to the spy microphone, it is not a big issue because our system is composed of two modules, i.e., the speech signal cancellation module and the coherent noise coupling module. Even if speech signal cancellation is not perfect, it does not affect the overall system performance that much as powerful noises will be added.

6 COHERENT NOISE COUPLING

We develop coherent noises to replace independent noises adopted in SOTA UMJs, which enable MicFrozen to be more resilient against adversarial denoising techniques. We adopt the non-linear mixture to generate coherent noises and then redesign the noises to resist removal techniques.

6.1 Non-linear Mixture

We generate coherent noises by adopting the non-linear mixture method [99]. Its basic idea is to use a differentiable bijective mapping that fuses speech signals with random noises [36, 44, 70]. Because the generated coherent noises share similar features as the speech signals, it is very difficult to remove them in the time domain [21, 70].

To be specific, we leverage a common non-linear mixture function [99] to generate coherent noises as follows,

$$n_{mix}(t) = \boldsymbol{A}_{1\times 2}(t) \cdot sgm[s(t),\ n_M(t)]^{\mathrm{T}}, \tag{14}$$

where $\boldsymbol{A}_{1\times 2}(t)$ is a nonsingular random mixing matrix, $n_M(t)$ is a random noise, and $sgm[\cdot]$ is a sigmoid activation function, where for a given vector input $\boldsymbol{x}(t)$, we have

$$sgm[\boldsymbol{x}(t)] = 1 - e^{-\boldsymbol{x}(t)}(1 + e^{-\boldsymbol{x}(t)})^{-1}. \tag{15}$$

By involving $n_{mix}(t)$, the spy microphone would receive $y(t) = n_{mix}(t) + s(t)$. For obtaining clean speeches, an adversary needs to find an inverse function $G(\cdot)$ that meets

$$G(y(t)) = [s(t),\ n_M(t)]^{\mathrm{T}}. \tag{16}$$

Note that the problem of recovering two 1D vectors (i.e., $s(t)$ and $n_T(t)$) from one (i.e., $y(t)$) is undetermined [35]. That is, there are infinite number of possible solutions for Eq. 16. Denoising techniques using features in time domain to find the optimal solution under the assumption of signal independence [51]. However, these techniques are ineffective here.

It has been proved that there are still infinite inverse functions $G(\cdot)$ with countless pairs of $[s(t),\ n_M(t)]^{\mathrm{T}}$ which can meet the requirement of statistical independence [21, 44, 70]. These methods thus often obtain local-optimal solutions. For example, Almeida [2] exploits mutual information to optimize $G(\cdot)$ but only obtains meaningless noises rather than the target signals. Although adding constraint conditions such as time correlation [98, 99], signal non-stationarity [36], the structure of mixing models [44, 70], and regularization [70] is helpful to reduce the indeterminacy, an adversary is not able to collect enough prior information and MicFrozen could further modify the random noise $n_M(t)$ in a time-varying manner. Therefore, BSS-like denoising techniques do not work well in separating coherent noises.

6.2 Inseparable Noise Design

To resist various denoising techniques that an adversary could adopt as discussed in Sec. 3, we redesign the above coherent noise to increase their correlation with raw speech signals in all domains and resist sniffer-assisted techniques.

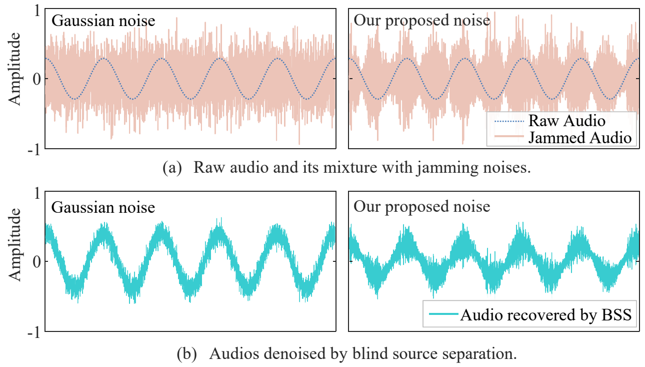
6.2.1 Time Domain Inseparability. Although the non-linear mixture generates a coherent noise, a strictly synchronous coherent noise can cause a high computational cost. Fortunately, we could maintain weak synchronization by convoluting $n_{mix}(t)$ with another random noise $n_T(t)$. The resultant noise $n_{co} = n_T(t) * n_{mix}(t)$ will couple with speech signals within a time delay of several seconds. In other words, for an arbitrarily delayed speech signal $s(t-\tau)$, there is always a corresponding delayed component in $n_{co}(t)$ to couple with. In this way, we keep the inseparability between the proposed coherent noise and speech signals in time domain.

6.2.2 Frequency Domain Inseparability. To enhance the correlation between the proposed noise and speech signal in frequency domain, we introduce a frequency-domain noise, denoted as $\mathscr{F}[n_F(t)] * \mathscr{F}[s(t)] = \mathscr{F}[n_F(t) \cdot s(t)]$, where $n_F(t)$ is a random noise and $\mathscr{F}(\cdot)$ is the Fourier transform. Such a convolution makes the noise coupled with speech signals. High/low/band-pass/stop filters can only eliminate additive noises. They would result in severe distortion of speech signals that coupled with convolutive noises. Other frequency-domain denoising methods such as independent component analysis [87] are also ineffective due to the strong dependence on convolutive components [51, 57]. The proposed noise can be expressed in time domain as $n_F(t) \cdot s(t)$, and the coherent noise in Eq. 14 becomes

$$n_{co}(t) = n_T(t) * \boldsymbol{A}_{1\times 3}(t) \cdot sgm[s(t),\ n_M(t),\ n_F(t) \cdot s(t)]^{\mathrm{T}}, \quad (17)$$

where $\boldsymbol{A}_{1\times 3}(t)$ is a nonsingular random mixing matrix.

6.2.3 Spatial Domain Inseparability. To combat space-domain separation, we suggest placing MicFrozen close to the sound source. Such arrangements ensure the overlap of sound sources and signals emitted from MicFrozen in spatial domain, which enhances the resilience against spatial domain separation.



(a) Raw audio and its mixture with jamming noises.

(b) Audios denoised by blind source separation.

Figure 7: Performance comparison of a Gaussian noise (left) and our proposed noise (right) defending against the same denoising BSS algorithm.

6.2.4 Resilience against Sniffing. A wide-band noise can effectively defend against sniffing methods. It causes the captured noise distorted [32]. The frequency-domain convolution can expand the bandwidth of $n_F(t) \cdot s(t)$ to over 12 kHz in Sec. 6.2.2. In addition, ultrasonic sniffers can capture suspicious ultrasounds. However, recalling that the ultrasonic signal is a mixture of anti-speech signal and coherent noise, adversaries cannot infer the two components directly from the captured signal, not to mention that the required wide passband ultrasonic microphone is expensive and easy to be caught due to its large size [4].

In short, we proposed coherent noises (i.e., Eq. 17) that can effectively resist adversarial removal. Note that the coherent noises do not need to meet any latency requirement.

6.3 Resilience Verification

We compare the performance of the proposed coherent noise with the typical jamming noise in SOTA UMJs [17, 28, 63, 74, 75], i.e., Gaussian noise. A tone signal with a frequency of 100 Hz acts as the raw audio to be protected. Here, we adopt the Gaussian noise with a bandwidth of 1 kHz. The two noises are both carried by ultrasound of 39 kHz. The spy microphones obtain a low SNR of -29.4 dB against the Gaussian noise and -25.0 dB against our proposed noise, as shown in Fig. 7(a). To remove noise and extract the raw audio, a BSS denoising method using four microphones is implemented. It increases the SNR of the audio jammed by Gaussian noise to 5.6 dB. Other jamming noises [15, 48, 69] achieve an even worse performance with an SNR higher than 6 dB for the recovered audio signal. As compared in Fig. 7(b), the SNR of the recovered audio signal from our proposed coherent noise still remains low, i.e., -1.1 dB. This result demonstrates the resilience of MicFrozen. More detailed evaluations of its defending performance against various advanced denoising techniques are presented in Sec. 8.2.2.

6.4 Avoid Collision between Modules

The coherent noise coupling module may collide with the speech signal cancellation module. The coherent noises are likely to be captured by the feedback microphone and block the convergence of the LANC algorithm. To avoid collisions between the two modules, we add a denoising module before the feedback. Since MicFrozen has the knowledge of the coherent noise, an ANF is used to remove the noise and recover the speech signal cancellation result. However, this solution increases the computational latency. Therefore, beamforming is used to constrain the coverage of noise and avoid covering the feedback microphone. Here, we propose a simple method, i.e., use a transducer array to realize directional noise injection. In our future study, parametric acoustic array [25] will be considered to reduce the array size.

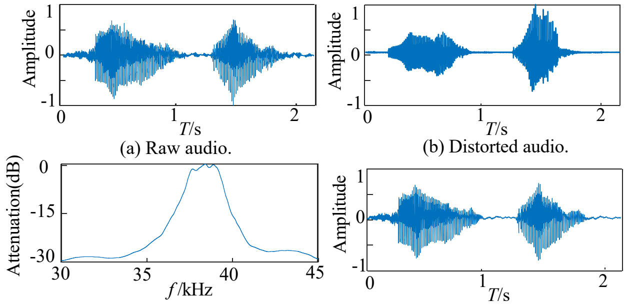
7 PRACTICAL SYSTEM DESIGN

To make our system work well in real-world settings, we further tackle several practical issues in this section.

7.1 Carrier Frequency Selection

For efficient injection of the anti-speech signal and the coherent noise into spy microphones, the ultrasonic carrier wave needs to be carefully selected. We select the most effective carrier frequency that produces injections with the highest demodulated amplitude. However, the most effective frequency differs among microphones [92]. Current commercial microphones are categorized as condenser microphones and dynamic microphones. Portable recording devices usually use condenser microphones due to their small size. Fortunately, MicFrozen does not require any knowledge of the spy microphones such as the non-linearity coefficient. It is true that an accurate measurement of the non-linearity coefficient of the spy microphone can present the best cancellation performance. However, one observation is that the non-linearity coefficients across different microphones are not that significant. Hence, we can just measure the non-linearity coefficients of several microphones and use the average value to approximate the non-linearity coefficient of the spy microphone. This simple scheme works reasonably well and based on our experiments, the achieved percentage of unrecognized words against eavesdropping decreases from 96.9% to 86.9% when we use the average value. This simple averaging ensures that MicFrozen is effective against most microphones and avoids the need of acquiring any knowledge of the spy microphone.

We test the non-linearity of 20 ECM microphones and 20 MEMS ones (produced by Panasonic, Hosiden, Harman, and Bosung), along with 20 COTS recording devices. We test two typical sampling rates, i.e., 48 kHz and 96 kHz. We choose a 39 kHz carrier to defend against spy microphones with sampling rates less than 48 kHz, and an 80 kHz carrier against those with sampling rates below 96 kHz. The power ratio of ultrasonic injections to speech signals is restricted to below -5.3 dB. Similarly, we can adopt multiple carriers with higher frequencies against eavesdropping devices with higher sampling rates (e.g., 192 kHz). In practice, MicFrozen transmits noises demodulated on multiple carriers to trigger non-linearity.



(a) Raw audio. (b) Distorted audio.

(c) Frequency response of 40kHz transducer. (d) Modified audio free from distortion.

Figure 8: Signal modification against distortion by counteracting the attenuation in different frequencies.

7.2 Carrier Phase Synchronization

The phase difference between ultrasonic carriers would distort the injections and undermine the performance of speech signal cancellation.

We reconsider the input of microphone $x(t)$ as follows,

$$x(t) = \cos(2\pi f_c t + \varphi) + n(t)\cos(2\pi f_c t), \tag{18}$$

where φ is the phase difference between two carriers. Then the received signal $r(t)$ becomes

$$r(t) = \frac{1}{2}n^2(t) + \cos(\varphi)n(t). \tag{19}$$

What is worse is that the frequency offset of different signal sources makes φ vary over time. Thus the NLMS algorithm cannot be applied for distortion compensation. To address the impact of φ , we adopt a phase detector [41] to eliminate the phase difference to synchronize carrier phases.

7.3 Signal Modification

To ensure the cancellation performance, we should guarantee the anti-speech signals are not distorted during the process. This requires ultrasonic transducers to have a wide bandwidth, i.e., larger than 4 kHz. However, most COTS transducers have a 3-kHz flat working bandwidth [42]. This causes signal distortion, i.e., the skew of signals in sideband. We present one example audio and the distorted version after demodulation in Fig. 8(a) and Fig. 8(b).

To handle this problem, we measure the precise frequency response of the used transducer, as shown in Fig. 8(c). Then we apply an inverse filter to counteract the frequency attenuation of the transducer. Fig. 8(d) demonstrates that our method can significantly mitigate the effect of distortion. The power of the demodulated signal decreases due to filtering and an amplifier can be used to make up for the loss.

8 EVALUATION

We implement MicFrozen with COTS devices and evaluate its performance in defending against practical eavesdropping methods. All experiments follow the IRB protocol approved.

8.1 Experiment Setup

Hardware. In MicFrozen, we design an ultrasound speaker. It consists of a transducer array and a power amplifier. The designed speaker is able to emit ultrasounds whose frequencies range from 35 to 85 kHz at an unweighted sound pressure level of 110 dB measured 20 cm away. It transmits the anti-speech signal, the coherent noise, and ultrasonic carriers simultaneously. An NI USB-4431 signal processor is adopted for signal input/output. A laptop executes the codes written in Labview for signal generation (both the anti-speech signal and the coherent noise) and analysis. The ultrasonic carriers are created using a signal generator (SIGLENT SDG1020). Two ADMP401 microphone modules act as the reference and calibration microphones.

Sound Source. A loud speaker (EDIFIER M230 Portable Speaker) serves as the source. The speech signals are derived from two open-source audio datasets, i.e., AudioMNIST [9] and Librispeech [59]. The sound pressure of speech signals is set at 65 dB, which is the intensity level of people's normal conversation.

Spy Microphones. We use 20 ECM microphone modules, 20 MEMS ones (produced by Panasonic, Hosiden, Harman, and Bosung), and 7 COTS recording devices as the spy microphones. In the experiment, we test each of them and present the average results. We also employ professional recording devices and directional microphones on the adversary side to evaluate the system performance.

Evaluation Metrics. We evaluate the effectiveness of jammers from two aspects, i.e., accuracy and difficulty of speech recognition. Specifically, we apply Cooperative Word Error Rate (CWER) for accuracy measurement and Signal-to-noise Ratio (SNR) for difficulty measurement. CWER is the percentage of words that are missed or incorrectly recognized. We use three ASRs (Google STT [29], CMU Sphinx [39], and iFLYTEK [40]) and recruit five volunteers for recognition. SNR reflects the speech reduction and noise jamming performance, which is commonly used in signal quality evaluation. A higher CWER or a lower SNR indicates a better jamming performance. The metrics with a subscript 'raw' (i.e., CWER_{Raw} and SNR_{Raw}) mean that the measurement is based on raw recordings, while CWER and SNR mean that the recordings are denoised with the following methods.

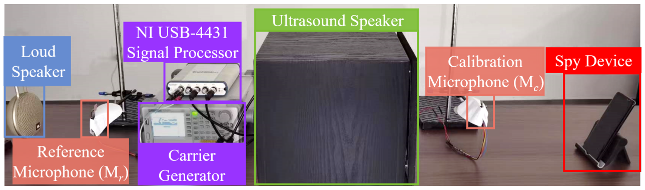

Loud Speaker
NI USB-4431 Signal Processor
Ultrasound Speaker
Calibration Microphone (M_c)
Spy Device
Reference Microphone (M_r)
Carrier Generator

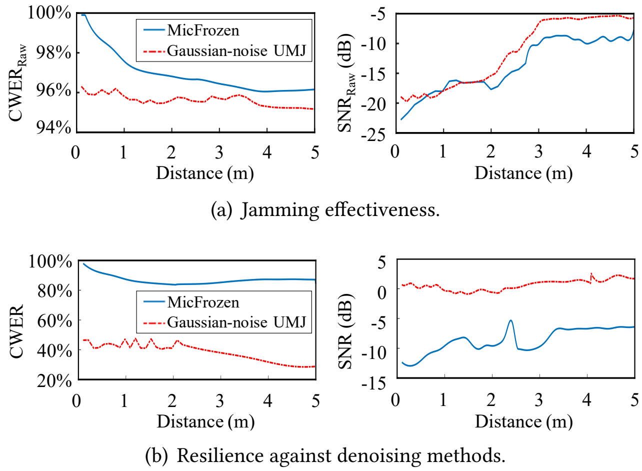
Figure 9: The prototype of MicFrozen.

Denoising Methods. In the experiment, we introduce an adversary with the ability to use four representative denoising methods as presented in Sec. 3. 1) BSS. The adversary captures multi-channel audios with four microphones and uses the fast independent component analysis (ICA) [33] algorithm to separate speech signals and noises in time domain. 2) Frequency filter. The adversary acquires the time-frequency spectrum of a recording by applying the short time-frequency transform (STFT) operation and removes high-power noises. 3) Beamforming. The adversary utilizes the time difference of arrival (TDOA) to separate the sound coming different directions using a four-microphone array and a delay-and-sum algorithm. 4) An adaptive noise filter (ANF) assisted by sniffers. The adversary uses assistant tools (i.e., sniffers) to acquire jamming noises by transmitting ultrasounds [32] and remove them with an ANF based on the NLMS algorithm [31].

8.2 Overall Performance

We first evaluate the effectiveness of MicFrozen in the line-alignment scenario. We place the sound source, the reference microphone M_r, the ultrasound speaker, and the calibration microphone M_c on the same line, as shown in Fig. 9. The distance between the sound source and M_r is 10 *cm*. The distance between M_r and the ultrasonic speaker is 20 *cm*, and that between the ultrasonic speaker and M_c is 10 *cm*.

Baseline. We implement three other UMJs for comparison. They adopt a Gaussian noise with a bandwidth of 4 kHz, a single-frequency noise sweeping from 0 to 4 kHz, and a single-frequency noise hopping within 4 kHz as their jamming noises. For the sake of fairness, we set the sum power of their transmitted noises to be the same. Though these UMJs achieve similar performances, i.e., a CWER_{Raw} above 90%, the CWER of these UMJs drops to 55.7%, 31.2%, and 41.9% respectively when denoising methods are applied. Note that the CWER of the Gaussian noise varies with its bandwidth and reaches a peak of 55.7% when the bandwidth is 4 kHz, which is the best among these UMJs. Considering the fact that Gaussian noise is the most popular mechanism used in commercial UMJs [17, 28, 74, 75], we select the UMJ



(a) Jamming effectiveness.

(b) Resilience against denoising methods.

Figure 10: Jamming effectiveness and resilience of MicFrozen in comparison with a Gaussian-noise UMJ.

using a 4-kHz-bandwidth Gaussian noise as the baseline.

8.2.1 Jamming Effectiveness. We first evaluate the jamming effectiveness of MicFrozen when the spy microphone is located at different distances from the sound source. As shown in Fig. 10(a), MicFrozen outperforms the Gaussian noise UMJ in both SNR_{Raw} and CWER_{Raw} at almost all distances. Although the performance of MicFrozen drops faster in terms of CWER_{Raw} because it suffers more from the attenuation of ultrasound, it provides a more effective defense against eavesdroppers within a distance of 5 *m*. Such a long protecting distance supports practical anti-eavesdropping.

8.2.2 Resilience against Denoising Methods. From the attacker's perspective, we apply denoising methods on all audios collected from the spy microphones and choose those with the most effective denoising effect to evaluate their SNR and CWER. Fig. 10(b) reports the CWER and SNR of protected speech signals after denoising. The signals protected by MicFrozen achieve an average CWER of 86.9% and maintains an SNR lower than -8.6 dB against denoising. As a comparison, Gaussian noise UMJ can only protect 47.7% words at most and the speech SNR is over 0 dB after denoising. The result shows the resilience of MicFrozen against denoising methods.

8.3 Performance on 2D coverage

Under the same setup as the line scenario, we extend our evaluation to a 2D area. The area is a sector with a radius of 5 *m* and an angle of ±60°. Here, the speaker of MicFrozen is located at the center. The CWER of MicFrozen in this area is plotted in Fig. 11. Within the tested sector area, MicFrozen always maintains a high CWER over 72.3%. In particular, we find that MicFrozen can jam spy microphones placed 3.5 *m* away within ±60° with a CWER over 85%. Moreover, at a distance of 5 *m*, MicFrozen still retains a high CWER (≥ 90%) in the sector with a ±30° angle. These results demonstrate that MicFrozen can leverage the trade-off between distance and coverage angle to balance its performance. In addition, the area can be further extended by deploying multiple MicFrozen systems. With a sector coverage of 120°, three MicFrozen systems can achieve a 360° protection.

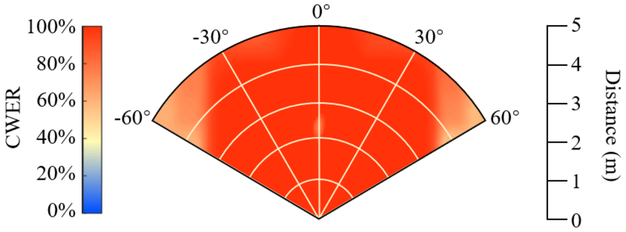


Figure 11: 2D Performance against denoising methods.

8.4 Latency

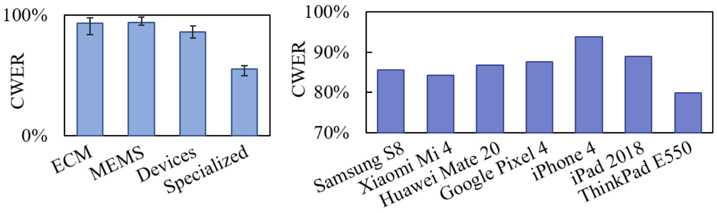
We use the timer of Labview to measure the computational latency of signal processing. We keep testing MicFrozen in jamming different speech signals for 2 hours. The average latency is 0.50 *ms* with a standard deviation of 0.093 *ms*. With this latency, the smallest distance between the sound source and MicFrozen is 13.6 *cm*, which is practical in real-world implementation. We also test the latency of an Arduino (UNO R3) that is expected to be adopted as the processor in our future wearable implementation. The latency is 0.59 *ms* with a standard deviation of 0.097 *ms* which can guarantee real-time protection.

8.5 Impact of Sound Source

We now evaluate the impact of sound source orientation and the number of sound sources.

8.5.1 Orientation. We vary the angle between the direction of the sound source and that of MicFrozen from 0° to 90° at a step size of 5°. The results show that an orientation difference within 30° does not affect the jamming performance of MicFrozen. When the angle increases to 55°, MicFrozen can still protect over 80% words at a distance of 4 *m*. Even under a larger angle (≥ 55°), spy microphones within 3 *m* are effectively jammed. In short, MicFrozen is robust against orientation change of sound sources.

8.5.2 Directional Sound Source. Sometimes users may adopt directional speakers as sound sources. We employ a Honeywell TKU-P30A directional speaker as the sound source. Its transmission angle is about ±63° and the maximum transmission distance is 7 *m*. In such an area, the CWER of MicFrozen is still over 71.9%. We also observe that within the angle of ±31°, the jamming effectiveness is not affected.



(a) Microphone categories. (b) Recording devices in our everyday life.

Figure 12: Performance against different microphones.

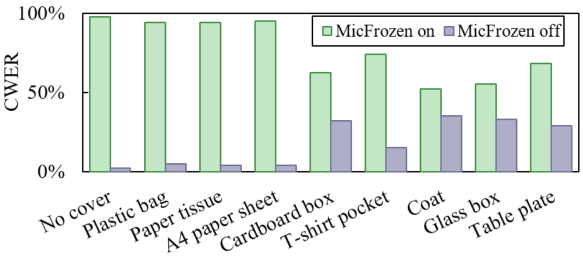


Figure 13: Performance under physical covers.

8.5.3 Multiple Sound Sources. We recruit five volunteers to sit around a 1.4 *m* × 1.4 *m* table and talk freely for 10 minutes. Then they change seats randomly for three times. MicFrozen is placed 20 *cm* away from one edge of the table. We measure the CWER and find that within a sector with a ±32° opening angle and a 3 *m* radius, the CWER is no less than 80%. We can arrange multiple MicFrozen systems for a larger coverage based on practical requirements. A more complex scenario is when the spy microphone is placed among the users. We plan to design a wearable version of MicFrozen in our future work so the reference microphone M_r could be pined to the collar of the user and the MicFrozen devices would move with the user's mouth. The wearable version of MicFrozen system can thus be carried by each user to protect speech privacy in such complex scenarios.

8.6 Impact of Spy Microphones

In this section, we evaluate the effect of microphone diversity on system performance.

8.6.1 Ordinary Devices. We first test several mobile devices that can be used for eavesdropping. We place those spy microphones 0.5 *m* away from the sound source. As shown in Fig. 12(a), the average CWER is 93.0% against ECM microphone modules, 93.7% against MEMS ones, and 85.8% against recording devices used in our everyday life. We further detail the defending effectiveness against these recording devices, including smartphones (Samsung S8, Xiaomi Mi 4, Huawei Mate 20, Google Pixel 4, and iPhone 4), an iPad 2018, and a laptop (ThinkPad E550). Fig. 12(b) shows that the achieved CWER is always over 75% in the presence of MicFrozen.

8.6.2 Specialized Microphones. We choose three professional recording devices (HIKVISION Bluetooth, BY-BM3051s, and GMTD GM-A905) as spy microphones. When the adversary employs these specialized devices, MicFrozen achieves a CWER of 58.5%, 57.3%, and 49.9% respectively after adversarial denoising. As a comparison, a UMJ using Gaussian noises performs much worse, with a CWER below 13%.

We also involve directional microphones for eavesdropping. We test three directional microphones, i.e., BY-BM3051s in directional mode, RODE VideoMic Go, and SONY ECM-GZ1M. For each microphone, we vary the distance and angle between it and the sound source. The results show that MicFrozen achieves a CWER ≥ 43.1% within a 4 *m* radius and ±22° angle range. If the eavesdropper is closer to our system (i.e., 1 *m*), the CWER increases to over 55.6% while the angle range expends to ±40°.

In short, MicFrozen is able to resist diverse spy microphones in various scenes. Although there is a performance degradation against professional spy microphones, MicFrozen is still much more powerful than SOTA UMJs.

8.6.3 Physical Cover. Spy microphones may be hidden inside physical covers. The poor penetration ability of ultrasound through those covers degrades the performance of UMJs including MicFrozen. We select eight commonly seen covers made of different materials for evaluation. The thin covers include a plastic bag (0.2 *mm* thickness), a paper tissue, an A4 paper sheet, and a cardboard box (3 *mm* thickness). The thick ones include a T-shirt, a coat, a glass box (1 *cm* thickness), and a wooden table plate (3 *cm* thickness). The spy microphones are placed 0.5 *m* away from MicFrozen. As shown in Fig. 13, thin materials have little impact on MicFrozen with CWERs all above 98%. On the other hand, the thick covers decrease the CWER to about 40%. To mitigate the impact of physical covers, besides removing them, we can increase the power of ultrasounds.

8.7 Impact of Environment

We consider two practical cases, where the users are speaking outdoors, and eavesdropped under non-line-of-sight (NLoS) conditions.

8.7.1 Outdoor Environment. We conduct an outdoor experiment to demonstrate the robustness of MicFrozen against environment diversity. The experiment is conducted outside our department building during the period of 12:00 am and 2:00 am. Midnight period is chosen to minimize the effect of ambient acoustic noise which is measured to be around 43 dB. The setup is the same as that described in Section 8.2. MicFrozen achieves a high CWER_{Raw} of 98.0% against spy microphones at 1 *m* away, 97.9% at 3 *m* away, and 87.7% at

5 *m* away. Even various denoising methods are applied, the achieved CWER only slightly decreases to 96.4%, 86.8%, and 81.0% at 1, 3, and 5 meters away respectively. MicFrozen can jam spy microphones placed 3 *m* away within ±60° with CWERs over 84.0%. The results demonstrate that MicFrozen can effectively protect speech privacy in outdoor environments.

8.7.2 NLoS Scenario. We evaluate the performance of the proposed system against spy microphones behind walls, i.e., under NLoS conditions. We consider walls made of three commonly seen materials, i.e., wood, (hollow) glass, and concrete. All the three walls have a same thickness of 4 *cm* (the hollow glass [43] consists of two 1 *cm* glass sheets and a 2 *cm* air gap). We put sound source and MicFrozen at one side of the wall, and the spy microphone at the other side. The distance from the MicFrozen system to the wall is 0.2 *m*, and that from the spy microphone to the wall is 0.1 *m*. Due to the wall attenuation, even when MicFrzoen is turned off, the achieved CWER at spy microphones is pretty high, i.e., 32.2%, 78.9%, and 48.3% for walls made of wood, glass, and concrete respectively. As a comparison, the CWER further increases to 61.3%, 92.5%, and 51.0% respectively after MicFrozen is turned on. Based on the above results, we can see that the protection effect of MicFrozen is degraded in NLoS conditions, which remains a limitation we plan to address in our future work.

8.8 Impact of Movement

We consider three types of simple movements: 1) only the sound source moves, 2) only MicFrozen moves, and 3) only the spy microphone moves. The distance between MicFrozen and the sound sources is kept as 30 *cm*, and the spy microphones is placed 3 *m* away. For the first type of movement, we ask one volunteer to act as the sound source and walk in two different patterns, i.e., randomly walking and following a pre-defined trajectory (regularly). For the second type of movement, we ask one volunteer to stand still, acting as the sound source, and another volunteer to hold the MicFrozen, moving regularly or randomly. For the third type of movement, we ask one volunteer to stand still, acting as the sound source, and another volunteer to hold the spy microphone, moving regularly or randomly. The influence of MicFrozen's movements is most significant, with CWERs always larger than 81.8% no matter which motion pattern is used. In comparison, the CWER achieved when sound source, MicFrozen and spy microphone are all stationary is 92.1%. We note that the movement of spy microphone has little influence, with an average CWER of 90.5% for regular walking and 89.2% for random walking. The movement of sound source causes the CWER to drop to 83.0% for regular walking and 84.3% for random walking respectively. These results demonstrate the robustness of MicFrozen against movements.

8.9 Power Consumption

We use a Digital Power Meter [22] to measure the power consumption of MicFrozen. Overall, the average consumption in 2 hours is 0.81 W. Its three major components, i.e., the microphone, ultrasound speaker, and processor, consume 0.0129 W, 0.5982 W, and 0.1933 W power respectively. In particular, we also consider two extreme cases. Such low power consumption facilitates MicFrozen to achieve a long battery life in real-world deployment.

8.10 Comparison with SOTA UMJs

We have demonstrated the superiority of our proposed noise over others in Sec. 8.2. Here, we further compare MicFrozen with SOTA UMJs [15, 48, 63]. When denoising methods are not applied, all UMJs are effective, i.e., achieving CWERs above 90%. However, once denoising method is applied, the performance degrades significantly. In particular, for UMJs enhanced with Gaussian noise [63], sweeping-frequency noise [48], and hopping-frequency noise [15], the CWER drops to 51.1%, 7.1%, and 41.2% respectively. Compared with the SOTA UMJs, MicFrozen's CWERs are always over 85% even when different types of denoising methods are applied.

9 DISCUSSION

Coverage and Ultrasound Power. Ultrasound attenuates faster than audible sound. For the same transmission power, the transmission range of ultrasounds is smaller than audible sounds. Increasing the transmission power can extend the jamming range and mitigate this issue. However, we need to make sure that the transmission power is lower than the maximum power allowed by regulations. The sound pressure limit suggested by the International Non-Ionizing Radiation Committee is 110 dB SPL [19]. In our experiment, we limit the sound pressure level of the transmitted ultrasound to be lower than 95 dB half meter from MicFrozen to ensure human safety. Such transmission power is high enough to jam spy microphones within a range of several meters.

Multiple Spy Microphones. Adversaries can use multi-microphone denoising methods for advanced attacks. Generally, multi-microphone denoising methods can be grouped into two categories: beamforming and BSS [26], which have been discussed in Sec. 3. We evaluated the performance MicFrozen against two representative methods, i.e., TDOA-based beamforming and ICA-based BSS. MicFrozen is able to defend against these two methods, with a CWER over 79%. Although there are some advanced BSS schemes (e.g., empirical mode decomposition (EMD) [38, 55] and ensemble-EMD (EEMD) based BSS [84]), they mainly focus on special cases or improving the performance in terms of computational overhead, scalability, and latency, instead of that of denoising. For example, EMD BSS is utilized to solve the issue that the number of observers is less than that of sources, and EEMD BSS addresses the defeats of EMD BSS in mode mixing [84]. Our experiment shows that even if advanced methods are applied, the achieved CWER is only slightly decreased (i.e., less than 10%).

Ultrasonic Spy Microphones. Sniffing with ultrasonic spy microphones to obtain the noise is a potential countermeasure against UMJs. To effectively capture the ultrasonic signal, the adversary needs to know the frequency band of the ultrasonic signal. In practice, the adversary has no knowledge of which frequency band to sniff. We can thus adopt the frequency hopping scheme to further reduce the possibility of being sniffed. We can also utilize a cryptographic pseudo-random number generator [23] to choose a random frequency to modulate our signals. Even in the worst case in which the adversary obtains the ultrasonic signals, it is a mixture of anti-speech signals and coherent noise. The adversary cannot easily extract the speech signal since the non-linear mixture design makes coherent noise inseparable from speeches [70], as discussed in Sec. 6.2. Moreover, the hardware cost of realizing such a wideband sniffing attack is high [5, 6]. Adopting a time-changing design of coherent noise can effectively defend against the adversary attack. In MicFrozen, we can vary the key parameters in the noise generation function (i.e., Eq. 17) in Sec. 6.2.1. We can use a random number generator [16] or random permutation algorithms [20] to generate the parameters. Even if the adversary has full knowledge of the noise generation method, it is still difficult to obtain all the parameters to infer the noise added.

Mobility Attacks. The spy microphone can move to different locations having different angles with respect to the MicFrozen system. The amount of cancellation is different as the cancellation signal varies. It is true that the amount of jamming due to cancellation signal varies when the spy microphone moves. However, the effect of jamming noise is not affected much. Therefore, the overall performance of the proposed system is not affected much by the mobility attack. According to our experiment, MicFrozen still achieves a high CWER over 88% on average against such mobility attacks.

10 RELATED WORK

Eavesdropping and Defense. To defend against eavesdropping via microphones, users can adopt hardware devices to generate jamming noise. Based on noise generated, the methods can be broadly grouped as EMI-based [46], audible sound-based [52, 58] and ultrasound-based [15, 48, 63, 66, 69]. Adversaries also conduct microphone-free eavesdropping through fine-grained vibration sensing to recover sound/speech content. Various sensors [7, 27, 30, 54], millimeter wave signals [86], ultra-wideband signals [81] and even hard disk drivers [47] have been successfully utilized to eavesdrop through vibration sensing.

Applications of Acoustic Non-linearity. Microphones exhibit the square-law non-linearity [1, 24]. Based on this property, malicious inaudible commands are injected to voice assistants [64, 68, 88, 92] and several countermeasures are proposed [32, 64, 90]. This property is also utilized to defend against eavesdropping [14, 15, 63, 66, 69], localization [49], device fingerprint [97], and communication [63, 91].

Applications based on Acoustic Signals. Abundant acoustic (both audible and ultrasonic) signals can also be exploited for authentication [37, 96], localization [60, 93], distance measurement [73, 94], device tracking [8, 95], face recognition [45, 53], and behavior sensing [50, 56, 78, 80, 89].

Ultrasonic Microphone Jammers. Basically, UMJs exploit the acoustic non-linearity of microphones to protect speech privacy. Backdoor [63] is the first work of UMJs. After comparing several types of noise, Backdoor selects an 8-kHz-bandwidth Gaussian noise as the jamming noise. Chen *et al.* [15] enable a wearable UMJ with multiple ultrasonic speakers for a wider jamming angular range. MicShield [69] is a jamming system which could protect voice assistants being used to eavesdrop on users' privacy. Patronus [48] supports selective jamming. It allows authorized devices to record but prohibits the other (illegal) microphones by employing narrow-band chirp noise. Nevertheless, all the prior UMJs jam microphones merely by adding random noise (Gaussian noise, chirp noise, or hopping noise). They have been proven to be vulnerable to denoising methods [12]. In comparison, we utilize speech signal cancellation and improve the inseparability of added noise. Therefore, MicFrozen realizes an effective and resilient defense against eavesdropping threats from sophisticated adversaries.

11 CONCLUSION

We propose MicFrozen, a novel UMJ against eavesdropping. It produces anti-speech signals to cancel private speech signals and generates coherent noises to further cover the cancellation residue. MicFrozen is capable of estimating the inverse-channel in a real-time manner for speech cancellation without a need to know the spy microphone's location. MicFrozen outperforms SOTA UMJs in protecting speech privacy against illegal recording and speech recovery.

ACKNOWLEDGES

This paper is partially supported by the National Key R&D Program of China (2021QY0703), National Natural Science Foundation of China under grant U21A20462 and 62032021, Research Institute of Cyberspace Governance in Zhejiang University, and Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005).

REFERENCES

[1] Muhammad Abuelma'atti. Analysis of the effect of radio frequency interference on the DC performance of bipolar operational amplifiers. *IEEE Transactions on Electromagnetic Compatibility*, 45:453–458, 2003.

[2] Luis Almeida. Linear and nonlinear ICA based on mutual information. In *IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium*, 2000.

[3] Russakovskii Artem. Google is permanently nerfing all home minis because mine spied on everything i said 24/7. https://www.androidpolice.com/2017/10/10/google-nerfing-home-minis-mine-spied-everything-said-247/#1, 2021.

[4] Avisoft Bioacoustics. Condenser Ultrasound Microphone. http://www.avisoft.com/ultrasound-microphones/cm16-cmpa/, 2021.

[5] Avisoft Bioacoustics. Condenser ultrasound microphone. https://www.avisoft.com/ultrasound-microphones/cm24-cmpa/, 2022.

[6] Avisoft Bioacoustics. UltraSoundGate 116Hme. https://www.avisoft.com/ultrasoundgate/116hme/#71165, 2022.

[7] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kaili Ren. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *Network and Distributed System Security Symposium*, 2020.

[8] Yang Bai, Nakul Garg, and Nirupam Roy. SPiDR: Ultra-low-power acoustic spatial sensing for micro-robot navigation. In *International Conference on Mobile Systems, Applications and Services*, 2022.

[9] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *ArXiv preprint:1807.03418*, 2018.

[10] Sreeram Chakravarthy and Sen Kuo. Application of active noise control for reducing snore. In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006.

[11] Chengyuan Chang, Shingtai Pan, and Kuochun Liao. Active noise control and its application to snore noise cancellation. *Asian Journal of Control*, 15(6):1648–1654, 2013.

[12] Yike Chen, Ming Gao, Yimin Li, Lingfeng Zhang, Li Lu, Feng Lin, Jinsong Han, and Kui Ren. Big brother is listening: An evaluation framework on ultrasonic microphone jammers. In *IEEE International Conference on Computer Communications*, 2022.

[13] Yuxin Chen. Wearable microphone jamming. https://github.com/y-x-c/wearable-microphone-jamming/, 2020.

[14] Yuxin Chen, Huiying Li, Steven Nagels, Zhijing Li, Pedro Lopes, Ben Y. Zhao, and Haitao Zheng. Understanding the effectiveness of ultrasonic microphone jammer. *CoRR*, abs/1904.08490, 2019.

[15] Yuxin Chen, Huiying Li, Shan-Yuan Teng, Steven Nagels, Zhijing Li, Pedro Lopes, Ben Y. Zhao, and Haitao Zheng. Wearable microphone jamming. In *International Conference on Human Factors in Computing Systems*, 2020.

[16] J-L Danger, Sylvain Guilley, and Philippe Hoogvorst. High speed true random number generator based on open loop structures in fpgas. *Microelectronics journal*, 40(11):1650–1656, 2009.

[17] Detective Store. Speech jammer tower-a for blocking professional microphones and counter-surveillance. https://www.detective-store.com/speech-jammer-tower-a-for-blocking-proffessional-microphones-counter-surveillance-1516.html, 2021.

[18] Shumin Dong, Bo Zhao, Ying Wang, and Tong Zou. A method of blind separation for coherent source based on single vector sensor. In *International Conference on Computer, Information and Telecommunication Systems*, 2017.

[19] Francis A Duck. Medical and non-medical protection standards for ultrasound and infrasound. *Progress in biophysics and molecular biology*, 93(1-3):176–191, 2007.

[20] Richard Durstenfeld. Algorithm 235: random permutation. *Communications of the ACM*, 7(7):420, 1964.

[21] Jan Eriksson and Visa Koivunen. Blind identifiability of class of nonlinear instantaneous ICA models. In *European Signal Processing Conference*, 2002.

[22] Everfine. Pf9800 digital power meter. http://www.everfine.net/en/productsinfo.php?cid=65&id=320, 2022.

[23] Aurélien Francillon and Claude Castelluccia. Tinyrng: A cryptographic random number generator for wireless sensors network nodes. In *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks and Workshops*, 2007.

[24] Javier Gago, Josep Balcells, David GonzÁlez, Manuel Lamich, Juan Mon, and Alfonso Santolaria. EMI susceptibility model of signal conditioning circuits based on operational amplifiers. *IEEE Transactions on Electromagnetic Compatibility*, 49(4):849–859, 2007.

[25] Woonseng Gan, Jun Yang, and Tomoo Kamakura. A review of parametric acoustic array in air. *Applied Acoustics*, 73(12):1211–1219, 2012.

[26] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich Golan, and Alexey Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017.

[27] Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. Deaf-aid: mobile IoT communication exploiting stealthy speaker-to-gyroscope channel. In *International Conference on Mobile Computing and Networking*, 2020.

[28] Global TSCM Group, Inc. Anti-recoding system with battery pack. https://www.globaltscmgroup-usa.com/, 2021.

[29] Google Cloud. Speech-to-text: Automatic speech recognition. https://cloud.google.com/speech-to-text, 2021.

[30] Jun Han, Albert Jin Chung, and Patrick Tague. Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2017.

[31] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley press, 1996.

[32] Yitao He, Junyu Bian, Xinyu Tong, Zihui Qian, Wei Zhu, Xiaohua Tian, and Xinbing Wang. Canceling inaudible voice commands against voice control systems. In *International Conference on Mobile Computing and Networking*, 2019.

[33] Jeanny Herault and Christian Jutten. Space or time adaptive signal processing by neural models. In *AIP Netural Networks for Computing*, 1987.

[34] Yoichi Hinamoto and Hideaki Sakai. Analysis of the filtered-X LMS algorithm and a related new algorithm for active control of multitonal noise. *IEEE Transaction Speech Audio Processing*, 14(1):123–130, 2006.

[35] Morris W. Hirsch and Stephen Smale. Differential equations, dynamical systems, and linear algebra. *Academic Press*, 1974.

[36] Shahram Hosseini and Yannick Deville. Blind separation of parametric nonlinear mixtures of possibly autocorrelated and non-stationary sources. *IEEE Transactions on Signal Processing*, 62(24):6521–6533, 2014.

[37] Long Huang and Chen Wang. Notification privacy protection via unobtrusive gripping hand verification using media sounds. In *International Conference on Mobile Computing and Networking*, 2021.

[38] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London*, 454(1971):903–995, 1998.

[39] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alexander I. Rudnicky. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *IEEE International Conference on Acoustics Speech and Signal*

Processing Proceedings, 2006.

[40] iFLYTEK Co., Ltd. iFLYTEK open platform-an artificial intelligence platform focusing on intelligent speech interaction wihch provides solutions for global developers. https://global.xfyun.cn/, 2021.

[41] Chanyoung Jeong, Dongho Choi, and Changsik Yoo. A fast automatic frequency calibration (AFC) scheme for phase-locked loop (PLL) frequency synthesizer. In *IEEE Radio Frequency Integrated Circuits Symposium*, 2009.

[42] Jinci Technologies. Product review. http://www.jinci.cn/en/goods/112.html, 2021.

[43] JingGlass Inc. Hollow glass. http://www.jingglass.com/news/Hollow-glass.html, 2017.

[44] Christian Jutten, Massoud Babaie-Zadeh, and Shahram Hosseini. Three easy ways for separating nonlinear mixtures? *Signal Processing*, 84(2):217–229, 2004.

[45] Kaustubh Kalgaonkar and Bhiksha Raj. Recognizing talking faces from acoustic Doppler reflections. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2008.

[46] Denis Kune, John Backes, Shane Clark, Daniel Kramer, Matthew Reynolds, Kevin Fu, Yongdae Kim, and Wenyuan Xu. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *IEEE Symposium on Security and Privacy*, 2013.

[47] Andrew Kwong, Wenyuan Xu, and Kevin Fu. Hard drive of hearing: Disks that eavesdrop with a synthesized microphone. In *IEEE Symposium on Security and Privacy*, 2019.

[48] Lingkun Li, Manni Liu, Yuguang Yao, Fan Dang, Zhichao Cao, and Yunhao Liu. Patronus: Preventing unauthorized speech recordings with support for selective unscrambling. In *International Conference on Embedded Networked Sensor Systems*, 2020.

[49] Qiongzheng Lin, Zhenlin An, and Lei Yang. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *International Conference on Mobile Computing and Networking*, 2019.

[50] Jian Liu, Yan Wang, Gorkem Kar, Yingying Chen, Jie Yang, and Marco Gruteser. Snooping keystrokes with mm-level audio ranging on a single phone. In *International Conference on Mobile Computing and Networking*, 2015.

[51] Shoji Makino, Te-Won Lee, Shoji Makino, and Hiroshi Sawada. *Blind speech separation.* Dordrecht: Springer Netherlands, 2007.

[52] Masking Privacy Simple. VoiceArrest Sound Masking Features. https://mpsacoustics.com/sound-masking/, 2021.

[53] Phillip McKerrow and Kok Kai Yoong. Classifying still faces with ultrasonic sensing. *Robotics and Autonomous Systems*, 55(9):702–710, 2007.

[54] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. Gyrophone: Recognizing speech from gyroscope signals. In *USENIX Security Symposium*, 2014.

[55] Bogdan Mijovic, Maarten De Vos, Ivan Gligorijevic, Joachim Taelman, and Sabine Van Huffel. Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis. *IEEE transactions on biomedical engineering*, 57(9):2188–2196, 2010.

[56] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. FingerIO: Using active sonar for fine-grained finger tracking. In *CHI Conference on Human Factors in Computing Systems*, 2016.

[57] Francesco Nesta, Piergiorgio Svaizer, and Maurizio Omologo. Convolutive bss of short mixtures by ICA recursively regularized across frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):624–639, 2011.

[58] Oeler Industries, Inc. Sound Masking Systems. https://www.oeler.com/sound-masking-systems/, 2020.

[59] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[60] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. BeepBeep: A high accuracy acoustic ranging system using COTS mobile devices. In *International Conference on Embedded Networked Sensor Systems*, 2007.

[61] Yanmin Qian, Chao Weng, Xuankai Chang, Shuai Wang, and Dong Yu. Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*, 19(1):40–63, 2018.

[62] Soundarya Ramesh, Rui Xiao, Anindya Maiti, Jong Taek Lee, Harini Ramprasad, Ananda Kumar, Murtuza Jadliwala, and Jun Han. Acoustics to the rescue: Physical key inference attack revisited. In *USENIX Security Symposium*, 2021.

[63] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *International Conference on Mobile Systems, Applications, and Services*, 2017.

[64] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *USENIX Symposium on Networked Systems Design and Implementation*, 2018.

[65] Manfred R. Schroeder, Bishnu S. Atal, and Joseph L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66:1647–1652, 1979.

[66] Hao Shen, Weiming Zhang, Han Fang, Zehua Ma, and Nenghai Yu. Jamsys: Coverage optimization of a microphone jamming system based on ultrasounds. *IEEE Access*, 7:67483–67496, 2019.

[67] Sheng Shen, Nirupam Roy, Junfeng Guan, Haitham Hassanieh, and Romit Roy Choudhury. MUTE: Bringing IoT to noise cancellation. In *Conference of the ACM Special Interest Group on Data Communication*, 2018.

[68] Liwei Song and Prateek Mittal. Poster: Inaudible voice commands. In *ACM conference on computer and communications security*, 2017.

[69] Ke Sun, Chen Chen, and Xinyu Zhang. "Alexa, stop spying on me!": Speech privacy protection against voice assistants. In *International Conference on Embedded Networked Sensor Systems*, 2020.

[70] Anisse Taleb. A generic framework for blind source separation in structured nonlinear models. *IEEE Transactions on Signal Processing*, 50(8):1819–1830, 2002.

[71] The Guardian. Apple apologises for allowing workers to listen to siri recordings. https://www.theguardian.com/technology/2019/aug/29/apple-apologises-listen-siri-recordings, 2019.

[72] The Guardian. Ukraine prime minister offers resignation after leaked recording. https://www.theguardian.com/world/2020/jan/17/ukraine-prime-minister-oleksiy-goncharuk-offers-resignation-after-leaked-recording, 2020.

[73] Yu-Chih Tung and Kang G. Shin. EchoTag: Accurate infrastructure-free indoor location tagging with smartphones. In *International Conference on Mobile Computing and Networking*, 2015.

[74] U-spy Store. Portable ultrasonic microphone defeater. https://uspystore.com/portable-ultrasonic-microphone-defeater/, 2021.

[75] U-spy Store. Silent ultrasonic microphone defeater. https://uspystore.com/silent-ultrasonic-microphone-defeater/, 2021.

[76] Unname. Demos of the prototype of micfrozen on anti-eavesdropping. https://youtu.be/8g2UykkPZ-M, 2022.

[77] VRT NWS. Google employees are eavesdropping, even in your living room. https://www.vrt.be/vrtnws/en/2019/07/10/google-employees-are-eavesdropping-even-in-flemish-living-rooms/, 2019.

[78] Junjue Wang, Kaichen Zhao, Xinyu Zhang, and Chunyi Peng. Ubiquitous keyboard for small mobile devices: Harnessing multipath fading for fine-grained keystroke localization. In *International Conference on*

Mobile Systems, Applications, and Services, 2014.

[79] Kuo Wang and Wei Ren. Convergence analysis of the multi-variable filtered-X LMS algorithm with application to active noise control. *IEEE Transaction Signal Processing*, 47(4):1166–1169, 1999.

[80] Xuyu Wang, Runze Huang, and Shiwen Mao. Sonarbeat: Sonar phase for breathing beat monitoring with smartphones. In *International Conference on Computer Communication and Networks*, 2017.

[81] 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 *ACM Conference on Embedded Networked Sensor Systems*, 2020.

[82] Stefan Werner, Marcello Campos, and Paulo Diniz. Partial-update NLMS algorithms with data-selective updating. *IEEE Transaction on Signal Processing*, 52(4):938–949, 2004.

[83] Wikipedia. Finite impulse response. https://en.wikipedia.org/w/index.php?title=Finite_impulse_response&oldid=1044909269, 2021.

[84] Zhaohua Wu and Norden E Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in adaptive data analysis*, 1(1):1–41, 2009.

[85] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66:130–153, 2015.

[86] 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 *International Conference on Mobile Systems, Applications, and Services*, 2019.

[87] Makoto Yamada, Gordon Wichern, Kazunobu Kondo, Masashi Sugiyama, and Hiroshi Sawada. Noise adaptive optimization of matrix initialization for frequency-domain independent component analysis. *Digital Signal Processing*, 23(1):1–8, 2013.

[88] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. SurfingAttack: Interactive hidden attack on voice assistants using ultrasonic guided waves. In *Network and Distributed System Security Symposium*, 2020.

[89] Sangki Yun, Yi-Chao Chen, and Lili Qiu. Turning a mobile device into a mouse in the air. In *International Conference on Mobile Systems, Applications, and Services*, 2015.

[90] Guoming Zhang, Xiaoyu Ji, Xinfeng Li, Gang Qu, and Wenyuan Xu. Eararray: Defending against DolphinAttack via acoustic attenuation. In *Annual Network and Distributed System Security Symposium*, 2021.

[91] Guoming Zhang, Xiaoyu Ji, Xinyan Zhou, Dong-lian Qi, and Wenyuan Xu. UltraComm: High-speed and inaudible acoustic communication. In *Quality, Reliability, Security and Robustness in Heterogeneous Systems*, 2019.

[92] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *ACM Conference on Computer and Communications Security*, 2017.

[93] Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. SwordFight: Enabling a new class of phone-to-phone action games on commodity phones. In *International Conference on Mobile Systems, Applications, and Services*, 2012.

[94] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *International Conference on Mobile Systems, Applications, and Services*, 2017.

[95] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. BatTracker: High precision infrastructure-free mobile device tracking in indoor environments. In *ACM Conference on Embedded Network Sensor Systems*, 2017.

[96] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. Echoprint: Two-factor authentication using acoustics and vision on smartphones. In *International Conference on Mobile Computing and Networking*, 2018.

[97] Xinyan Zhou, Xiaoyu Ji, Chen Yan, Jiangyi Deng, and Wenyuan Xu. Nauth: Secure face-to-face device authentication via nonlinearity. In *IEEE Conference on Computer Communications*, 2019.

[98] Andreas Ziehe, Motoaki Kawanabe, Stefan Harmeling, and Klaus-Robert Müller. Separation of post-nonlinear mixtures using ace and temporal decorrelation. In *International Workshop on Independent Component Analysis and Blind Signal Separation*, 2001.

[99] Andreas Ziehe, Motoaki Kawanabe, Stefan Harmeling, and Klaus-Robert Müller. Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation. *The Journal of Machine Learning Research*, 4:1319–1338, 2003.