A Resilience Evaluation Framework on Ultrasonic Microphone Jammers

Ming Gao, Yike Chen, Yimin Li, Lingfeng Zhang, Jianwei Liu, Li Lu, *Member, IEEE,* Feng Lin, *Senior Member, IEEE,* Jinsong Han, *Senior Member, IEEE,* and Kui Ren, *Fellow, IEEE*

Abstract—Covert eavesdropping via microphones has always been a major threat to user privacy. Benefiting from the acoustic non-linearity property, the ultrasonic microphone jammer (UMJ) is effective in resisting this long-standing attack. However, prior UMJ researches underestimate adversary's attacking capability in reality and miss critical metrics for a thorough evaluation. The strong assumptions of adversary unable to retrieve information under low word recognition rate, and adversary's weak denoising abilities in the threat model make these works overlook the vulnerability of existing UMJs. As a result, their UMJs' resilience is overestimated. In this paper, we refine the adversary model and completely investigate potential eavesdropping threats. Correspondingly, we define a total of 12 metrics that are necessary for evaluating UMJs' resilience. Using these metrics, we propose a comprehensive framework to quantify UMJs' practical resilience. It fully covers three perspectives that prior works ignored to some degree, i.e., ambient information, semantic comprehension, and collaborative recognition. Guided by this framework, we can thoroughly and quantitatively evaluate the resilience of existing UMJs towards eavesdroppers. Our extensive assessment results reveal that most existing UMJs are vulnerable to sophisticated adverse approaches. We further outline the key factors influencing jammers' performance and present constructive suggestions for UMJs' future designs.

Index Terms—Microphone jamming, Anti-eavesdropping, Privacy protection

1 Introduction

Eavesdropping or recording via microphones has always been a serious privacy threat. Nowadays, ubiquitous smart devices, such as smartphones and voice assistants (VAs), are reported to eavesdrop on private speeches and pass recordings along to third-party [2]–[4], which exacerbates this threat.

To combat microphone-enabled eavesdropping, researchers proposed the ultrasonic microphone jammers (UMJs) [5]–[11]. Compared with the conventional electromagnetic and audible jammers [12], UMJs are promising in anti-eavesdropping without prior knowledge about the target devices nor audible disturbances by utilizing the inherent non-linearity of amplifiers inside a microphone [13], [14]. With this property, ultrasounds that are imperceptible to human ears over the air, would leak energy into the audible spectrum when arriving at microphones [15]. This noise migrating from ultrasonic bands would drown out the human voice in spy microphones' recordings. Recent advances in this field have enabled more practical designs on the UMJ over off-the-shelf devices [5]–[9].

However, existing UMJs underestimate the adversary's capability of retrieving meaningful information from noise-ruined sounds, and hence their jamming effect has been overestimated. For example, MicShield, a representative UMJ, was reported to be vulnerable to a beamforming-based eavesdropping attack (with 75.0% jammed fragments recovered and recognized by adversaries) [7]. Other noise elimination methods are also effective in jamming reduction in practice according to our experiments in Sec. 5.3. Hence, a more practical adversarial model should take the noise elimination as a fundamental attacking ability.

Besides the above overlooked capabilities, in the security assumption of prior works, the considered eavesdropping surface is too narrow in existing literatures. The adversary was assumed to recognize individual words from the recording by either human's perception or automatic speech recognizers (ASRs). Therefore, they usually leverage a defective evaluation method that mainly focuses on a single metric, i.e., the word recognition rate. Once the word recognition rate cannot exceed a pre-defined threshold, the UMJ is regarded as secure. However, non-verbal sounds and unrecognizable words also leak privacy. For instance, such an attack could position a victim's house via an unrecognizable eavesdropped audio [2]. The adversary infers the victim's region by the victim's accent and deduces that a fuzzy word 'Strxxt Sevxx, Waxsmxxsxxx' ('x' represents an unrecognizable syllable) was 'Street Seven, Waasmunster', which revealed the victim's detailed location.

With the above observations in mind, we explore and summarize realistic threats from sophisticated adversaries into three perspectives, including *ambient information*, *seman-*

- *Ming Gao, Yike Chen, Yimin Li, Lingfeng Zhang,Jianwei Liu, Li Lu, Feng Lin, and Jinsong Han (corresponding author) are with Zhejiang University, China, and with ZJU-Hangzhou Global Scientific and Technological Innovation Center, China. Email: gaomingppm@zju.edu.cn, chenyike@zju.edu.cn, ninalym13@gmail.com, lingfengzhang@zju.edu.cn, jianweiliu@zju.edu.cn, li.lu@zju.edu.cn, flin@zju.edu.cn, hanjinsong@zju.edu.cn.*
- *Kui Ren is with Zhejiang University, China, with ZJU-Hangzhou Global Scientific and Technological Innovation Center, China, with Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province, China, and with Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, China. Email: kuiren@zju.edu.cn.*
- *Ming Gao and Yike Chen contribute equally in this paper.*
- *A shorter conference version of this work were presented at the IEEE INFOCOM in 2022, "Big Brother is Listening: An Evaluation Framework on Ultrasonic Microphone Jammers" [1].*

tic comprehension, and *collaborative recognition*. First, even if a UMJ is powerful to guarantee that no verbal information would be recognized, the adversary might concentrate on the non-verbal or ambient information. For example, the background sound may expose the victim's location to the adversary. Second, the adversary can semantically comprehend the meaning of speech even though some parts of the recording are unrecognizable. This is because humans can compensate for lost information in fragmented recordings by guessing or inferring, even if these speeches are of inferior quality. Lastly, the collaboration between multiple ASRs and humans on recognition is overlooked. There are many ASRs in the market, such as Google speech to text (STT) [16], CMU Sphinx [17], and iFLYTEK [18], acute to distinct words. Although their recognition results are variant due to different intrinsic models and algorithms [19], a smart adversary can integrate their results [19] to recover more information, even if individual ASRs perform ineffectively. In addition, human's perception can further promote the recognition accuracy. With such man-machine collaboration, the adversary can maximize the recognition rate of victims' private speeches.

To comprehensively evaluate UMJs' resilience to realistic adversaries, we comprehensively investigate the threats from the above three perspectives. We refine the eavesdropping model by complementing the adversary with a practical denoising ability. Specifically, we summarize the adversary's ability on noise elimination in the following domains: time domain, frequency domain, spatial domain, and the coupling effects mixing several domains. This definition contains all the aspects adversaries can refer to, which provides a guide for defending against adversaries. Correspondingly, we propose an evaluation framework that leverages 12 metrics to cover the above attacking surfaces. For *ambient information*, we employ three *intensity* metrics [20], [21] to cover both the verbal and non-verbal factors. For *semantic comprehension*, we adopt six *intelligibility* metrics [22]–[27] to weigh how much adversaries could understand from the jammed recordings quantitatively. As for *collaborative recognition*, we define new metrics, which reflect UMJs' defensive effectiveness against adversarial man-machine collaboration. We weigh these metrics to compare UMJs on customers' convenience. Guided by this framework, we carry out a comprehensive investigation of current representative UMJs and quantify the defensive effectiveness of five representative ones [5]–[9]. Astonishingly, our assessment reveals that existing UMJs are often defeated under realistic eavesdropping. Moreover, to trigger effective countermeasures, we determine the key impact factors based on comparative analysis and present several constructive suggestions for future UMJ designs. In particular, we propose a novel method for resisting eavesdropping by combining the idea of active noise cancellation with coherent noises for a more resistant UMJ.

Our contributions are summarized as follows:

- We propose a comprehensive framework for evaluating UMJs' defence effectiveness. Involving 12 metrics, it covers potential threats as many as possible in real eavesdropping attacks, enabling a thorough and quantifiable evaluation on UMJs. We also release our source code [28] to facilitate the anti-eavesdropping research.
- We refine the adversary model of eavesdropping attacks to appraise UMJ's resilience objectively. We perform a detailed analysis of existing UMJs. The model and analysis support quantifiable evaluation on the vulnerability of existing UMJs against sophisticated adversaries in real-world scenarios.
- We outline the key factors that influence UMJs' performance and summarize constructive suggestions for the improvement and future design of UMJs based on our comparative analysis. We propose a new defending method by combining active noise cancellation with UMJs for reducing speech information leakage into spy microphones, rather than a simply masking scheme using noises in prior UMJs.

2 Background

We briefly explain how UMJs operate with the non-linearity and classify existing UMJs on account of jamming signals.

2.1 Principle of UMJs

A UMJ leverages microphone non-linearity [8] for jamming, where the inner amplifier exhibits square-law nonlinear characteristics [13], [15], especially when the input frequency is above 25 kHz. Hence, high-frequency signals are demodulated to a low-frequency signal intentionally. Supposed an audio input $x(t)$, the output $y(t)$ becomes

$$y(t) = A_1 x(t) + A_2 x^2(t), \tag{1}$$

where A_i is the gain coefficient of the i-th harmonic component $x^i(t)$. Here, higher-order harmonics ($i \geq 3$) are ignored due to their low energy [29].

A UMJ consists of several ultrasonic transducers supplied by a signal generator. In the simplest case, it exploits a pair of tones. The input of microphone $x(t)$ is

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

where $m(t)$ is the jamming signal modulated on the ultrasonic carrier $cos(2\pi f_c t)$, and f_c is the carrier frequency, higher than 25 kHz typically. A constant-frequency signal $m(t) = cos[2\pi(f_c + f_b)t]$, for example, introduces the second harmonic after an amplifier as following,

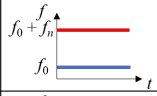
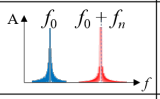
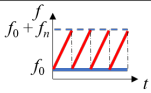
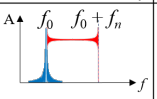
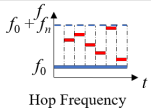
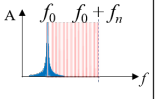
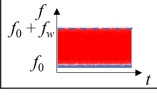
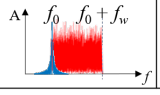
$$y(t) = A_2 + A_2 \cos(2\pi f_b t) + others, \tag{3}$$

where f_b is a bias frequency and $others$ represent high-frequency items that are removed by a low-pass filter. Afterwards, the low-frequency component $\frac{A_2}{2}\cos(2\pi f_b t)$ remains. The low-frequency residue will be recorded by microphones, making the private speech unrecognizable by virtue of the masking properties [30]. Therefore, the UMJs induce the modulated ultrasonic signals to inject the arranged noise into a spy microphone.

2.2 Classification of Jamming Signals

The essential differences among UMJs [5]–[9] lie in the categories of their jamming signal $m(t)$, which directly impact the UMJ's ability to shield voice bands. To perform an in-depth evaluation, it is necessary to classify existing UMJs according to their jamming signals, as listed in Table. 1.

TABLE 1
The common categories of jamming signals

Category	Frequency-time ($f-t$)	Amplitude-frequency (A$-f$)	Representatives
Tone Signal(s)			
Dynamic Single Frequency Noise (DSFN)	 Sweep Frequency		Patronus[6]
	 Hop Frequency		Wearable Jammer[5] Patronus[6]
White Gaussian Noise (WGN)			MicShield[7] Backdoor[8] Selective Jammer[9]

Tone signal. The previous example, $m(t) = \cos[2\pi(f_c + f_b)t]$, is a typical tone signal. Furthermore, a UMJ can employ multiple tone signals, that is,

$$m(t) = \textstyle\sum_j \cos[2\pi(f_c + f_j)t], \tag{4}$$

where f_c is the carrier frequency, j ($j \geq 1$) is the number of tone signals, and f_j is their frequency biases.

Dynamic single frequency noise (DSFN). The sweep frequency signal and hop frequency signal [5], [6] are two representatives. The sweep frequency signal repeats a linear continuous chirp regularly as follows,

$$m(t) = \cos[2\pi(f_0 + (kt \bmod m)\frac{f_n}{m-1})t], \tag{5}$$

where f_0 is the sum of the carrier frequency f_c and a bias f_b, k is the sweep rate, and f_n is the peak frequency.

The hop frequency one scrambles discretely and randomly at a preset interval as follows,

$$m(t) = \cos[2\pi(f_0 + a[\lfloor\frac{t}{p}\rfloor])t], \tag{6}$$

where f_0 is the sum of the carrier frequency f_c and a frequency bias f_b, p is the period between hopping, $a[\cdot]$ is a random sequence with a maximum value f_n, and $\lfloor\cdot\rfloor$ is the rounding down function. In practice, although the ringing effect during the frequency conversion triggers the audible sound, it is resolved by an inverse filter [8], with the jamming signals retaining their inaudibility.

White Gaussian noise (WGN). Some approaches [7]–[9] recommend WGN for jamming, whose energy is distributed over a broad ultrasonic spectrum. We have

$$m(t) = wgn(f_0, f_0 + f_w), \tag{7}$$

where $wgn(\cdot)$ is the Gaussian noise with a bandwidth f_w.

We categorize the representative UMJs [5]–[9] according to the above standards. The systems we evaluated in this paper are all based on these representative prototypes.

3 Threat Analysis

We refine the adversary model to reveal realistic threats that UMJs confront. It comprehensively analyzes the capabilities of a sophisticated but practical adversary.

3.1 Threat Model

We follow the well-known STRIDE threat model [31] to refine the threat model. Here, the information disclosure model [31] fits best and we have the following definitions:

Victims: Victims are the target devices or users to be bugged. They are protected by UMJs. Spy devices are located within the effect range of UMJs.

Adversary's capability: An adversary can plant one or more spy microphones in the vicinity of victims, or gain the microphone access of a smartphone or VA. He/she can deploy spy devices at suitable places for articulate and complete recordings. Even if illegal recordings are awash with jamming noise, he/she would recover and extract the private information by the means including but not limited to those detailed in Sec. 3.2. Furthermore, he/she may detect the existence of UMJs and perform anti-jamming treatments, such as choosing microphone deployment positions for noise elimination methods.

Winning condition: It is defined as the moment when the adversary successfully extracts private information from jammed recordings. He/she would result in three-fold threats (See Sec. 3.3) using various methods (See Sec. 3.2).

3.2 Noise Elimination Methods

Different from the assumption that an adversary would give up once jammed [5], [8], we point out that he/she would endeavor for noise elimination and information extraction. Generally, the adversary separates target audios from jamming noises leveraging the differences of features in three domains, i.e., time, frequency, and space. In addition, the out-band jamming noise would couple with ultrasound [32]. We also take denoising techniques based on such coupling effect into consideration. Therefore, we divide possible noise elimination methods into four categories as follows.

Time domain. Temporal filters exploit the distribution difference between speech signals and jamming noises. BSS [33], a representative of noise separation means, is widely utilized for separating independent source signals. It is a practical algorithm without any prior knowledge about noise. It is particularly adopted for speech separation and extraction. It profits from the mutual independence of the source signals but requires that the number of independent observers N is not less than the number of sources M. Thanks to multiple observations provided by multiple microphones, this dimension requirement is easy to fulfill.

Frequency domain. Frequency filters require basic knowledge of noise characteristics, such as frequency distribution. Unfortunately, the adversary could conclude necessary information from jammed recordings easily with the aid of advanced spectrum analysis, such as short-time Fourier transform (STFT) or discrete wavelet transformation (DWT). *Bandstop Filter (BSF)* is a preferred candidate for noise elimination. It is observed that the jamming noise has a prominent intensity. Accordingly, a *notch filter (NF)* or a *wideband bandstop filter (WBSF)* is exploited to filter out the frequency point or band with the maximum energy.

Spatial domain. The relative position between UMJs and acoustic sources to be protected would result in additional differences. Adversaries can separate signals based on features that reflect their position difference [34]. For example,

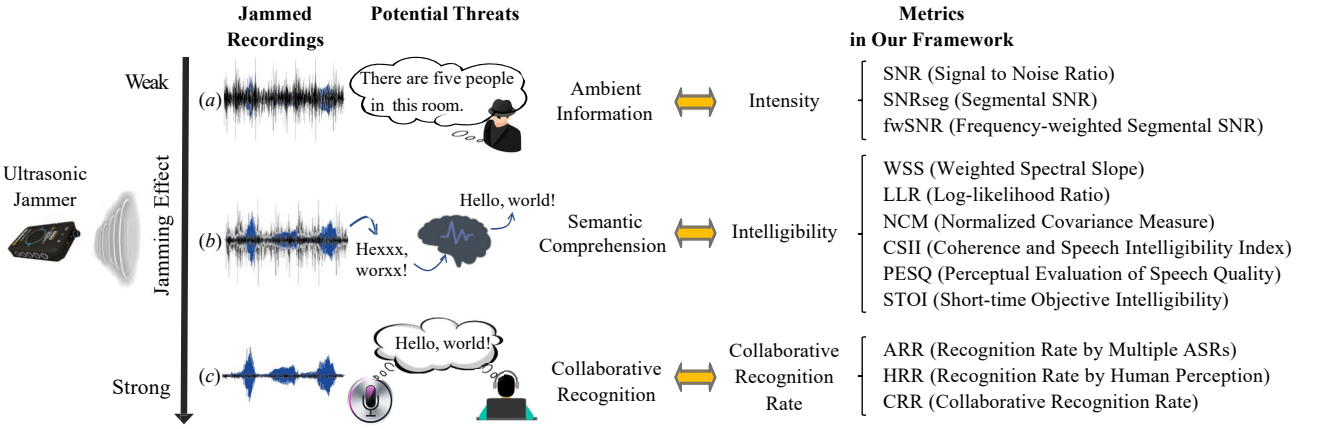


Fig. 1. Three-perspective architecture completely containing potential eavesdropping threats, consisting of ambient information, semantic comprehension, and collaborative recognition. Correspondingly, we propose an evaluation framework on UMJs' resilience, comprised of 12 metrics from three perspectives of intensity, intelligibility, and collaborative recognition.

beamforming methods using a microphone array can detect the spatial difference by distinguishing the arrival angle and thus enhance target private speeches [7].

Coupling effect. UMJs leverage ultrasonic jamming noises that would couple with external ultrasound [32]. It can detect the ultrasonic injection which is doomed to trigger the non-linearity of a microphone and energy leakage into the audio spectrum. It just needs a cheap ultrasonic transducer to broadcast the guard ultrasonic signals consisting of several tones with specific frequencies. A single guard tone with the appropriately configured frequency f_1 can introduce an extra component at the frequency $|f_1 - f_c|$, which can serve as a reference to cancel the low-frequency signals attendant upon the ultrasonic injection. In order for the automatic detection, the $|f_1 - f_c|$ component is expected to fall into the frequency band $[10\ \text{kHz}, 20\ \text{kHz}]$. Multi-tones, e.g. with the frequencies of $\{22\ \text{kHz}, 42\ \text{kHz}, 62\ \text{kHz}\}$, are utilized to capture the ultrasonic injection in a wide frequency range. According to the coupling effect, the adversary could obtain the complete distribution of jamming noises with the help of a series of external ultrasounds and thus adopt an *Adaptive Notch Filter (ANF)* for noise removal.

In practice, users have no knowledge of what kind of denoising countermeasures the adversary may use. Therefore, a secure UMJ needs to protect private speeches against diverse noise elimination methods from all the above aspects.

3.3 Three-perspective Architecture on Potential Eavesdropping Threats

We explore potential eavesdropping threats and organize them using a three-perspective architecture. It involves the perspectives of ambient information, semantic comprehension, and collaborative recognition.

Ambient information remaining in jammed recordings still exposes user privacy. In practice, a spy microphone collects speeches as well as acoustic contexts in the environment. Although the UMJ could guarantee that no verbal information would be recognized through the illegal recordings, the adversary might extract non-verbal information from a polluted recording for privacy theft. For example, in Fig. 1(a), when the adversary plants the spy microphones in the victim's bag, the adversary can surmise that the victim is in an office if the spy microphone captures noise emitted from printers. He/she could further draw the victim's daily routine. Moreover, some detailed privacy can be inferred, such as the number of people in surroundings, and the devices present in the room. Even worse, according to the leaked ambient information, the adversary could further conjecture more privacy, e.g., the identities or occupations of the victims, and the physical prints in a printer [35], [36]. Although eavesdropping on human voices and speeches is common, about which traditional UMJs are concerned, extensive research has demonstrated that the leakage of ambient information (e.g., via acoustic side channels [37]) puts users' privacy at risk [38], [39]. Therefore, a resilient UMJ should be able to resist eavesdropping on ambient information and we take ambient information into the consideration for a comprehensive evaluation.

Adversaries can semantically comprehend the meaning of speeches from partly-unrecognizable conversations. They can exploit conjectures or semantic knowledge to successfully understand some low-quality speeches. For instance, the adversary in Fig. 1(b) determines that the fuzzy fragment 'Hexxx, worxx!' (the character 'x' represents an unrecognizable syllable) is 'Hello, world!'. Moreover, the adversary might be able to infer an unrecognizable word according to contexts. Thus, unrecognizable words still risk the leakage of private information.

An adversary can pursue clear and accurate recognition of the victims' speech with the collaboration of ASRs and human labors, as shown in Fig. 1(c). Although speech recognition has been studied in previous works [5]–[9], the collaboration between multiple ASRs and humans on recognition is overlooked. Utilizing different intrinsic models, current ASRs [16]–[18] are acute to distinct words and generate different recognition results [19]. Human recognition further promotes accuracy. With a man-machine collaboration, the adversary can maximize the acquired privacy.

The adversary might utilize noise elimination techniques including the ones in Sec. 3.2 to exacerbate eavesdropping threats. On account of such a three-perspective architecture, a comprehensive framework should concern all the above perspectives to evaluate UMJs.

4 Evaluation Framework

We design a comprehensive framework to assess UMJs. As listed in Fig. 1, it embodies 12 metrics from three perspectives: intensity, intelligibility, and collaborative recognition.

4.1 Intensity

Previous evaluations based on speech recognition serve no purpose in inferring the environment, leaving a flaw of privacy leakage. Hence, a thorough assessment ought to consider both verbal and non-verbal factors. We leverage the *intensity* metrics [20], [21] to describe these factors. We adopt the Signal to Noise Ratio (SNR) and its derivatives to assess UMJs against ambient information leakage.

SNR is widely used to measure the impact of noise. Here, we employ it and its derivatives for describing the intensity of jamming noise. SNR serves as a predefined standard to guarantee impartial assessment. Moreover, we employ its two derivatives Segmental SNR (SNR_{seg}) [20] and frequency-weighted segmental SNR (fwSNR) [21] to compensate for the defect that SNR is possible to be affected by signal extremums. Image a long audio jammed by a short period of noise with high intensity, whose SNR would be low, but the audio contents can still be extracted in the period with few noise. In this case, only utilizing SNR cannot comprehensively judge the UMJ's performance from intensity. Instead, SNR_{seg} and fwSNR embody the jamming intensity in short segments. They average the signal intensity in each segmented time period or frequency band, in order to mitigate the impact of SNR extremums in a short segment. In our experiment, all sounds except jamming noise are treated as signal so as to include background sounds. These metrics can reflect the vulnerability of ultrasonic jammers when an adversary aims at both verbal audio and non-verbal information in the environment.

We merge them into one intensity score S_{SNR},

$$\text{S}_{\text{SNR}} = \boldsymbol{w}_{\text{SNR}} \cdot [\text{SNR}, \text{SNR}_{\text{seg}}, \text{fwSNR}]^T, \tag{8}$$

where w_{SNR} is a 3-dimensional weight vector. The three intensity metrics are normalized by the theoretical or experimental maximum and minimum values following a linear conversion function. This matrix is set to [0.34 0.22 0.44] based on their correlations with subjective ratings of signal quality [40]. Intensity metrics are incorporated in our framework to evaluate the UMJs' effectiveness in concealing the ambient information from eavesdropping. A higher S_{SNR} means a better jamming effect in terms of jamming intensity.

4.2 Intelligibility

Semantic comprehension should be painstakingly considered for evaluating the UMJs' defensive effectiveness. Besides recognizing individual words directly, an adversary may deduce information from jammed recordings via semantic comprehension. To quantify the information leakage, we project human comprehension into intelligibility [22]–[27]. It can be described by measuring audio distortions [27], [40]. We employ six metrics to represent three kinds of distortions. They weigh how much adversaries could understand from the jammed recordings quantitatively.

The distortion between received signals and raw speeches comprises a propagation distortion component, an additive noise component, and an algorithmic artifacts component [41]. These components are not necessarily dependent on each other [41]. Therefore, we introduce six metrics here to cover all distortions from the perspective of intelligibility, with each distortion measured by two metrics.

The propagation distortion component could be depicted by the spectral envelope difference. We utilize two basic measurements i.e. weighted spectral slope (WSS) [22] and log-likelihood ratio (LLR) [23] for description. Both of them measure the intelligibility by estimating a 'distance' from the clean speech to the degraded one, but they focus on different domains (i.e., frequency and time domains respectively). Specifically, WSS is the weighted difference between the spectral slopes in each frequency band and LLR is the likelihood ratio of features in the time domain. Since WSS and LRR are negatively correlated with intelligibility, we supersede them by their opposite here.

The additive noise component is described by normalized covariance measure (NCM) [24] and coherence and speech intelligibility index (CSII) [25]. These two metrics describe the additive noise component in terms of different statistical characteristics in the frequency domain. NCM is the weighted sum of normalized covariance signals in each frequency band. It measures the intelligibility by the covariance between the clean and degraded envelope speeches. Similarly, CSII is calculated by using coherence as features in the frequency domain. Previous works [40] show that NCM and CSII perform well in predicting speech intelligibility caused by additive noise.

The algorithmic artifacts component is quantitatively described by perceptual evaluation of speech quality (PESQ) [26] and Short-time objective intelligibility (STOI) [27]. They assess the intelligibility over a long or short period. PESQ estimates the overall loudness difference between the clean and degraded signals, which assesses the speech distortion introduced by artifact algorithms [42]. STOI is designed for measuring speeches processed by noise reduction and speech separation algorithms. It segments the clean and degraded speeches into short-time periods and calculates the correlation among the temporal envelopes of these segmented periods.

We normalize and integrate the above indexes into an intelligibility score S_{In} as follows,

$$\text{S}_{\text{In}} = \boldsymbol{w}_{\text{In}} \cdot [\text{WSS}, \text{LLR}, \text{NCM}, \text{CSII}, \text{PESQ}, \text{STOI}]^T, \tag{9}$$

where $\boldsymbol{w}_{\text{In}}$ is a 6-dimensional weight vector. In general, $\boldsymbol{w}_{\text{In}}$ is set to [0.06 0.13 0.20 0.21 0.18 0.22] based on their correlations with subjective ratings of speech intelligibility [27], [40]. We quantify the intelligibility of audio and score UMJs using these metrics. A higher S_{In} suggests that the UMJs can make illegal recordings barely comprehensible.

4.3 Collaborative Recognition Rate

Existing literature [5]–[8] consider the speech recognition by ASRs (ARR) or by human perception (HRR) separately. However, it is improper to separate human and machine. They neglect their collaboration on speech recognition.

We define the collaborative recognition rate (CRR). It reflects the real threats on illegal speech recognition. CRR

represents the rate of clear words that the adversary can recognize by jointly using multiple ASRs or human perception. CRR is twofold, subsuming ARR and HRR as follows,

$$\begin{aligned} \text{ARR} &= \frac{card(\text{R}_{\text{ASR}})}{card(\text{S})} = \frac{card(\bigcup_{i=1}^{n} \text{R}_{\text{ASR}_i})}{card(\text{S})}, \\ \text{HRR} &= \frac{card(\text{R}_{\text{H}})}{card(\text{S})} = \frac{card(\bigcup_{j=1}^{m} \text{R}_{\text{H}_j})}{card(\text{S})}, \\ \text{CRR} &= \frac{card(\text{R})}{card(\text{S})} = \frac{card(\text{R}_{\text{ASR}} \cup \text{R}_{\text{H}})}{card(\text{S})} \end{aligned} \tag{10}$$

where $card(\cdot)$ is the number of elements in a set, R_{ASR_i} is recognized by the i-th ASR, R_{H_j} is recognized by the j-th volunteer, R_{ASR} and R_{H} are words recognized by ASR and volunteers respectively, R is comprised of all words recognized, and all of them are the subsets of S, a set of the whole speeches to be identified. To keep positively correlated to the resilience of UMJs against such a man-machine collaborative attack, we design a recognition score

$$\text{S}_{\text{CRR}} = 1 - \text{CRR}. \tag{11}$$

S_{CRR} can indeed display the security of UMJs against an eavesdropper in the aspect of speech recognition.

4.4 Resilience Evaluation for UMJs

We weigh the above metrics for the convenience of unwitting consumers. They can directly compare UMJs' scores,

$$\text{S}_{\text{total}} = \boldsymbol{W} \cdot [\text{S}_{\text{SNR}},\ \text{S}_{\text{In}},\ \text{S}_{\text{CRR}}]^T, \tag{12}$$

where $\boldsymbol{W}$ is a 3-dimensional weight vector. A higher S_{total} implies better performance and robustness against eavesdropping and noise elimination methods.

We determine the weight of these metrics on the basis of the concern of customers using the Delphi method [43]. It is a popular forecasting process framework based on the results of questionnaires. Such a methodology can describe human participants' preferences and accordingly we leverage the principal component analysis (PCA) [44] to determine the weight coefficient of each metric. A questionnaire is designed to collect the preferences of potential customers using a Likert-type scale [45], where they score the threat level in each layer by judging several descriptions. The questionnaire is available in [46], which contains five questions. The first and the last question are about volunteers' overall attitudes toward privacy leakage and UMJs. The other three questions are about volunteers' concerns on three aspects, i.e., ambient information, semantic comprehension, and collaborative recognition, respectively. Each of them consists of four scenarios, and volunteers are asked to rate the threat level in each scenario, following a 7-score scale (1 represents no threat and 7 represents serious threat). It is distributed randomly to participants aged from 18 to 65. In 851 issued questionnaires, 732 participants are willing to use a UMJ and provide their preferences on the defence effect toward potential adversaries. PCA [44] is utilized to determine the weight coefficients $\boldsymbol{W}$ based on the preferences of customers. Specifically, we average the ranks of each question and get a 4×3 vector (four scenarios in three threat aspects). Then we introduce PCA to reduce its dimension to 1×3. With the survey result, we determine $\boldsymbol{W} = [0.3337\ 0.3609\ 0.3053]$ as the weights in S_{total} to evaluate the resilience of UMJs. Note that the similar weights of the three metrics (all around 0.33) indicate that users recognize all three aspects mentioned in this paper to be important. The average users are concerned approximately equally about the issues of potential private information leakage from the three perspectives. Meanwhile, in some cases, a specialized user (e.g., a UMJ designer) may pay more attention to one or two (not all) perspectives. Therefore, we provide an overall metric S_{total} to judge UMJs' performance directly for average users and recommend three metrics for a comprehensive evaluation, based on which UMJ designers could find out the defects of a UMJ and according improve it.

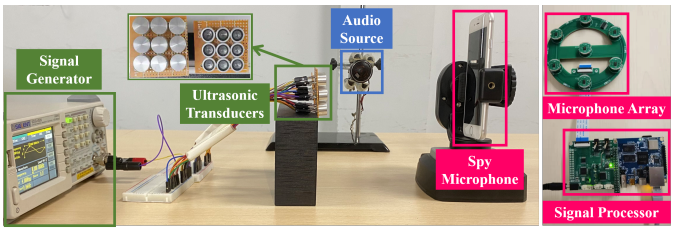


Fig. 2. Experimental setup.

5 Evaluation

Under the guidance of the above framework, we analyze the defensive effectiveness of representative UMJs thoroughly and reveal their vulnerability against an adversary.

5.1 Experiment Setup

We perform extensive experiments on existing UMJs under identical test conditions, as shown in Fig. 2.

UMJ Hardware. We use two kinds of transducers in a UMJ: nine sets of NU40A14TR-1 [47] play the jamming signal $m(t)$ while nine sets of NU40C16TR-1 [47] generate a 40 kHz ultrasonic carrier. Jamming signals depend on each UMJ. These transmitter arrays are put on a bracket and connected with a signal generator SIGLENT SDG1020 [48].

Recording Devices. We use three kinds (six models in total) of recording devices as spy microphones, i.e., a Samsung Galaxy S8, a Pixel 4, an iPhone 6s, an iPad Pro 11, and two Thinkpad x201 laptops. The average score of a UMJ on these six devices is regarded as its final score. Experimentally, the results on different tested devices are not significantly different. The standard deviations in S_{SNR}, S_{In}, S_{CRR}, and S_{total} are 0.06, 0.03, 0.04, and 0.04 respectively among six tested spy microphones. The low standard deviations indicate the reasonableness to use the average scores as the final result.

Position. A JBL 750T speaker plays the test audios and a spy microphone is 5 cm away in a quiet laboratory. The UMJ is placed next to the speaker. There is no obstacle to the line-of-sight. The spy microphone records the mixture of the raw test audios and jamming noise.

Power. The power of the raw audio is set to 65 dB-SPL (dB of sound pressure level), which is the common average loudness of social conversations [8]. The power of jamming signals is set to 115 dB-SPL. Because of the intrinsic hardware error, the measured values are 65.1dB-SPL (raw

TABLE 2
Five Representative UMJs and Their Overall Performances

Representative UMJ	Produced Noise Category	Without Noise Elimination				With Noise Elimination			
		S_{total}	S_{SNR}	S_{In}	S_{CRR}	S_{total}	S_{SNR}	S_{In}	S_{CRR}
Wearable Jammer [5]	[0,1] kHz DSFN hopping per 0.45 ms	0.86	0.84	0.76	1.00	0.41	0.41	0.49	0.32
Patronus [6]	[85,255] Hz DSFN hopping per 0.2 s	0.83	0.82	0.74	0.93	0.28	0.32	0.47	0.02
MicShiled [7]	4 kHz bandwidth WGN	0.87	0.82	0.81	1.00	0.37	0.31	0.47	0.33
Backdoor [8]	8 kHz bandwidth WGN	0.88	0.85	0.79	1.00	0.48	0.46	0.58	0.41
Selective Jammer [9]	three 4 kHz WGN covering a 12 kHz band	0.93	0.90	0.89	1.00	0.56	0.48	0.58	0.61

audio), 114.5 dB-SPL (DSFN jamming signals [5], [6]), and 105.1 dB-SPL (WGN signals [7]–[9]). Correspondingly, SNRs retain about -49.4 dB [5], [6] or -40 dB [7]–[9].

Test Audios. We play 11000 items of audio segments, derived from AudioMNIST [49] and Librispeech [50], including common words [8] (accounting for 94%), letters (4%), and digits (2%). The audios are randomly allocated among volunteers. In the experiment, volunteers are allowed to replay audios until they are able to recognize or give up.

We set an adversary's practical capability as follows.

Recognizers. We employ three ASRs and recruit 20 volunteers for the man-machine collaborative recognition on the jammed recordings. (1) ASRs: We exploit STT provided by Google STT [16], CMU Sphinx [17], and iFLYTEK [18]. The speech recognition ratios of these ASRs are claimed to exceed 80% on the raw audios [16]–[18]. (2) Humans: We randomly recruit 20 volunteers aged between 18 and 45 without any knowledge of specific selecting strategies to avoid biased impacts, whereas participants cannot bear hearing impairments and are able to recognize simple words or paragraphs. Our experiments on volunteers are validated through an institutional review (IRB) at our university.

Noise Elimination. To judge the security of these UMJs against realistic adverse approaches, we implement four noise elimination methods following the threat model in Sec. 3.2. They are representatives of denoising techniques exploiting time, frequency, spatial features, and coupling effect as follows, (a) $\mathcal{A}$-I: BSS with the fast independent cost analysis [51]. (b) $\mathcal{A}$-II: an STFT-based NF and a WBSF with 2 kHz cut-off frequency, where we choose the filter with better elimination performance as the result in each case. (c) $\mathcal{A}$-III: beamforming utilizing a circular microphone array with seven microphones (as shown in Fig. 2), in which case the distance between the tested UMJ and the speaker is set to 30 cm. We use the delay-and-sum algorithm [52] to realize beamforming. (d) $\mathcal{A}$-IV: an ANF with the normalized least mean square (NLMS) algorithm [53].

5.2 Representative UMJs

We replicate five representative UMJs [5]–[9] under the aforementioned conditions (See Tab. 2). Wearable Jammer [5] and Patronus [6] utilize the DSFN signals for jamming, while MicShield [7] exploits WGN. Specifically, Wearable Jammer issues the hop frequency signal altering randomly among [0,1] kHz for every 0.45 ms and modulated by a high-frequency carrier. Patronus uses sweeping chirp signals to smooth the frequency hopping. It produces a noise whose frequency changes among [85, 255] Hz for every 0.2 s. MicShield employs WGN with a 4 kHz bandwidth modulated on a 40 kHz carrier. Backdoor [8] tried four kinds of jamming signals and claimed WGN is the most useful. Hence, we set WGN of 8 kHz bandwidth as its jamming signals. Chen. et al. [9] proposed a selective ultrasonic microphone jammer (denoted as 'Selective Jammer' in our paper). It uses three 4 kHz WGN without overlapping to cover a 12 kHz band. These five representative UMJs own their unique designs in hardware platform arrangement and energy supply in previous works [5]–[8], [10], [11]. However, we set these parameters to the same for a fair evaluation. The influences of these designs are discussed in Sec. 7.3.

As shown in Fig. 3(c), these UMJs seem to protect the test audios from being recognized by the adversary. They maintain the low ARR and HRR, which tally with their claims about performance. Even in consideration of the man-machine collaboration, they obtain high S_{CRR}s. Patronus gets the lowest score by 0.93, while other UMJs nearly reach 1. They seem adequate for effective acoustic privacy protection.

5.3 Performance and Vulnerability

Based on the proposed framework, we reexamine UMJs' defensive effectiveness against eavesdropping. Figure 3 presents the scores of these UMJs without noise elimination and their average scores suffering from the noise elimination means that adversaries may take.

5.3.1 Overall Performance

Table 2 illustrates the overall performance of these five representative UMJs. The bigger shadow area in the charts implies better resilience against eavesdropping threats. [1] Results without noise elimination demonstrate that Selective Jammer owns the best defensive effectiveness with S_{total} of 0.93, followed by Backdoor, MicShield, Wearable Jammer, and Patronus in turn. However, these UMJs are vulnerable against the realistic adversary. From the comparison in Tab. 2, their performances decrease by 52.3% on average after the noise elimination. Selective Jammer, the most secure one, maintains a low score of only 0.56, which is still unavailing in face of the sophistication. Patronus scores merely 0.28. It proves barely resilient against arbitrary techniques of adversarial noise elimination. The following parts will detail their vulnerability from each perspective.

5.3.2 Intensity

As shown in Fig. 3(a), these UMJs share similar performance in terms of intensity without noise elimination, with S_{SNR}s of around 0.87. We observe that the contribution of each metric is unique. Wearable Jammer and Patronus achieve

1. Note that the noise elimination methods are different from those in our conference version [1], and thus, the results in Tab. 2 and Fig. 3 both differ from those in the former version.

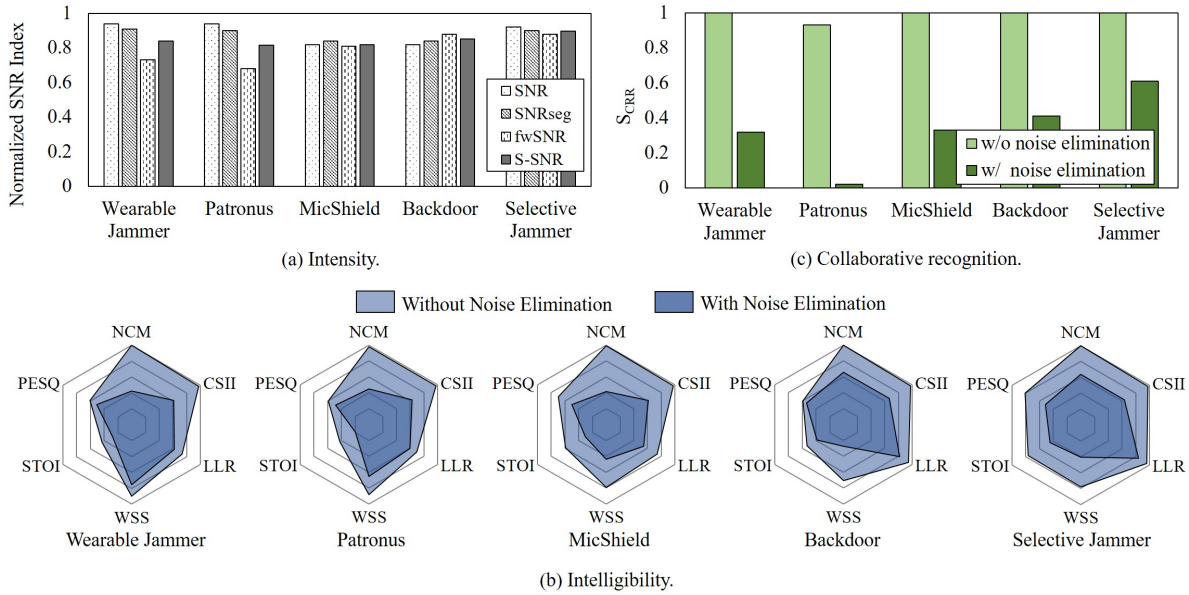


Fig. 3. Performance of five representative UMJs and their vulnerability to adversarial noise elimination.

the high SNR and SNR_{seg}. Their bandwidths do not exceed 2 kHz. MicShield and Backdoor use the broadband noise that contributes to the incline of fwSNR due to the complexity in the frequency domain. The effect of jamming signals' bandwidth is further compared in Sec. 5.5.2.

After the adversarial noise elimination, there are obvious reductions in each intensity metric. MicShield dominates among these five UMJs, possibly because of its complexity in frequency domain. Nonetheless, its SNR increases profoundly from -49.4 dB to -29.33 dB, followed by Wearable Jammer (-22.45 dB), Selective Jammer (-21.15 dB), Backdoor (-19.20 dB), and Patronus (-5.16 dB). Results indicate that most energy of jamming noise can be easily removed.

5.3.3 Intelligibility

We present the performance of these UMJs on each intelligibility metric, with radar charts in Fig. 3(b). The largest area indicates the best performance of Selective Jammer with the intelligibility score of 0.89, while others score around 0.78. The S_{In} and S_{CRR} of Patronus are both the worst among these UMJs, which indicates that the drop from intelligibility will seriously enervate the defensive effectiveness against the threat of recognition. This has strengthened our arguments to consider intelligibility metrics.

After elimination, S_{In}s present a significant erosion, with a decrease of over 0.26. The rank of S_{In} is Selective Jammer, Backdoor, Wearable Jammer, MicShield, and Patronus. Though the former four UMJs earn the seemingly similar assessment in terms of recognition, they receive diverse S_{In}s. This demonstrates that intelligibility does not depend absolutely on the speech recognition rate. Human comprehension is significant or even dominant at this state.

5.3.4 Collaborative Recognition Rate

Even in terms of speech recognition, all UMJs fail to resist adversarial noise eliminations. This means the adversary can recover most of the speech content. In Fig. 3(c), their performance decreases significantly. All ARRs outweigh 18% and all HRRs are above 20%. Selective Jammer seems to be the most resilient with its S_{CRR} just over 0.6, while Patronus performs worst again. It obscures merely 6.75% words, offering extremely less protection to users' speech privacy. In detail, its ARR, HRR, and CRR are 34.15%, 92.80%, and 93.25% respectively.

Furthermore, we analyze the influence of ASRs and volunteers on recognition. There is no doubt that an excellent ASR can increase the recognition rate in terms of ARR. Figure 5(a) illustrates that iFLYTEK is conspicuous for its efficiency, followed by Google STT, which can also offer some supplements. On the other hand, HRR depends on the ability and the number of volunteers. It seems to be uncontrolled and unpredictable. Fortunately, the difference in recognition is not obvious among humans. The cumulative average HRR curve for the incremental volunteers is plotted in Fig. 5(b). CRR reaches a high level and scarcely increases after the number of volunteers overtakes five. Nevertheless, there are still several words unrecognizable for humans but can be recognized by ASRs. This emphasizes that it is significant to take account of man-machine collaboration in speech recognition in the comprehension evaluation framework. We further ease the requirement. With the aid of iFLYTEK or Google STT, three volunteers are competent to measure the resilience of a UMJ from the perspective of man-machine collaborative recognition.

Accordingly, we provide manufacturers and average customers with different requirements. Manufacturers should pay close attention to each metric for the improvement of UMJs. Furthermore, manufacturers are obligated to assess their products based on a rich supply of experimental data and provide their S_{total}s to show the defend effectiveness quantitatively. By comparing the S_{total}s of different UMJs, the average customers can choose the appropriate UMJs. As for the average customers, we offer the low-cost measurement requirements to check the resilience of a UMJ. Experimentally, an audio consisting of at least 300 words represents adequate test audios with similar results and short test time. The average customer can access it via open source databases, or they can test on the voice captured by

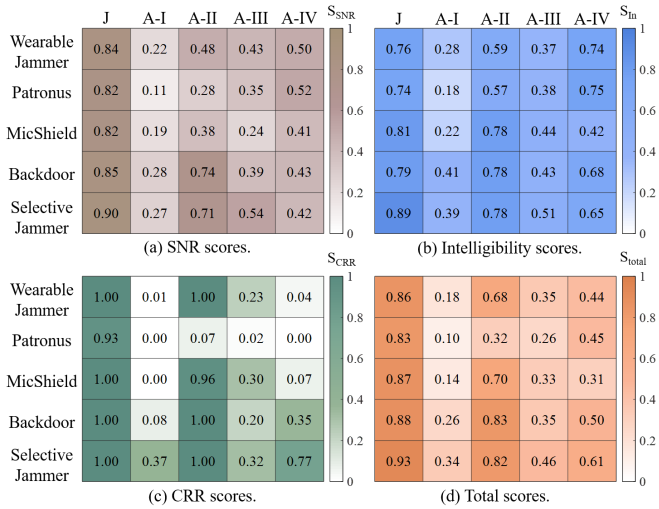


Fig. 4. Resilience of UMJs against each adversarial noise elimination.

themselves. In particular, the measurement of CRR involves several recognizers. We recommend the collaboration of one ASR and at least three human recognizers.

Briefly speaking, the existing UMJs are vulnerable against the realistic adversary. There will be a formidable task ahead of the UMJ designers. They should elaborate more ingenious jamming signals and cope with intractable adverse approaches in defence of private speeches.

5.4 Impact of Noise Elimination Methods

We perform evaluations on each adversarial noise elimination. Figure 4 illustrates the scores of representative UMJs under five conditions: jamming audios without elimination (written as J), and with four elimination methods detailed in Sec. 5.1 (i.e., $\mathcal{A}$-I, $\mathcal{A}$-II, $\mathcal{A}$-III, and $\mathcal{A}$-IV respectively).

$\mathcal{A}$-I (i.e., BSS) is the most dangerous elimination method against these UMJs, leading to an average drop of S_{total} from 0.87 to 0.20, and it is able to mitigate the impact of all representative UMJs. Though Selective Jammer utilizes multi-sources to defend against $\mathcal{A}$-I and scores highest among these UMJs, it protects only 37% of speeches. The effective noise removal of $\mathcal{A}$-I owes to the independence of noises and speeches, which is a necessary requirement for implementing BSS. The following method is $\mathcal{A}$-III (i.e., beamforming), with an average S_{total} drop of 0.52. It is also effective to all the five UMJs because the unavoidable distance difference between UMJs and the sound source makes audios separable. Compared with $\mathcal{A}$-III, $\mathcal{A}$-IV has a similar performance in reducing S_{CRR} with a drop of 0.74, while the S_{SNR} and S_{In} drop only 0.39 and 0.15 respectively. Selective Jammer has a good performance defending against $\mathcal{A}$-IV since it takes measures to avoid ultrasonic injection leaking noise information [9]. This means acquiring the information of jamming noises can help adversaries remove the noise. Among these elimination methods, $\mathcal{A}$-II performs worst. It hardly has any impacts on Wearable Jammer, MicShield, Backdoor, and Selective Jammer, while Patronus, the UMJ utilizing single tone, is vulnerable to $\mathcal{A}$-II. This means utilizing wide-band or multi-tone noises to cover a wide frequency band can defend against $\mathcal{A}$-II and other elimination methods based on frequency filter.

In general, all representative UMJs are extremely vulnerable to $\mathcal{A}$-I and $\mathcal{A}$-III. The key reason lies in the independence of existing UMJs from sound sources in terms of the time domain and location, which are ignored by prior researches on the UMJ design. Compared with the performance of Patronus against $\mathcal{A}$-II, other UMJs are more resilient, which indicates the importance of wide-band noises on UMJ's security.

5.5 Impact of Signal Categories and Parameters

To explore better jamming effectiveness, we analyze the impact of each parameter, which contributes to the strikingly different performances among UMJs. Accordingly, we conclude the key factors and provide suggestions for the future UMJ design in Sec. 7.1.

5.5.1 *Category of Jamming Signals*

We design three prototypes of UMJs utilizing different jamming signals, including sweep frequency signals, hop frequency signals, and WGN signals in Sec 2.2. We test on these three kinds of signals. The periods of the two DSFN signals are 1ms and all test jamming signals share the identical bandwidth (2 kHz).

Intuitively, a more complex jamming noise would provide more resilient defensive effectiveness. However, we cannot judge the complexity of different jamming noises with diverse parameters. To quantitatively describe the complexity of various jamming noises, we introduce the fuzzy entropy (FsEn) [54], a widely used measurement of the disorder degree. It compares the complexity of jamming signals from a statistical point of view. After necessary preprocessing [55], we have

$$FsEn(m,r) = ln\Phi^m(r) - ln\Phi^{m+1}(r), \tag{13}$$

where Φ is the mean of the degree matrix of membership on top of elements in the principal diagonal, m is the window length, and r is a parameter, generally set to the quotient of m divided by test signals' standard deviation [55].

As illustrated in Fig. 7(a), no significant correlations between FsEn and the performance of different signal categories are found. The hop frequency signal has the best performance before the noise removal, probably thanks to its randomness in the frequency domain. WGN obtains the lowest score. Energy dispersion of broadband may be to blame for this. Nevertheless, it becomes dominant after the adversarial noise elimination. Residual noise among broadband conversely improves UMJs' resilience.

5.5.2 *Bandwidth*

We further compare the influence of jamming signals' bandwidth. Taking WGN signals as example, we select five bandwidths, increasing from 500 Hz to 8 kHz with an equivalent ratio. As illustrated in Fig. 7(b), the FsEn is positively correlated with bandwidth of jamming signals. Meanwhile, the S_{CRR} increases with the bandwidth until 4 kHz, but there is a slight drop at 8 kHz. A bandwidth of 4 kHz is the best alternative for WGN signals. Jamming signals with the wider band are able to conceal more information in the audible band. However, an excessively wide-band signal disperses the noise energy, possibly leading to poor

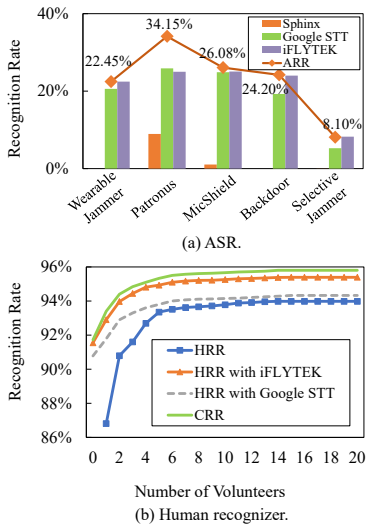


Fig. 5. Impact of recognizers.

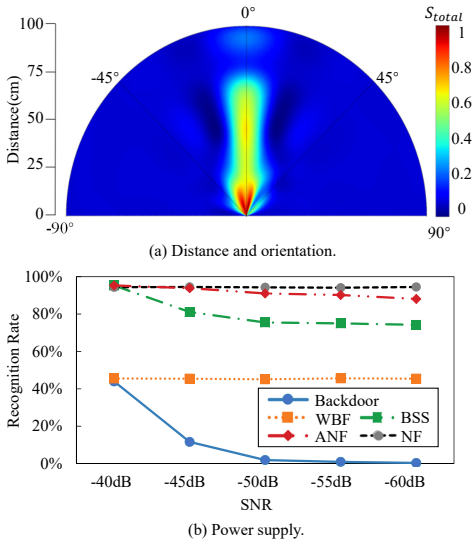


Fig. 6. Impact of Implementation.

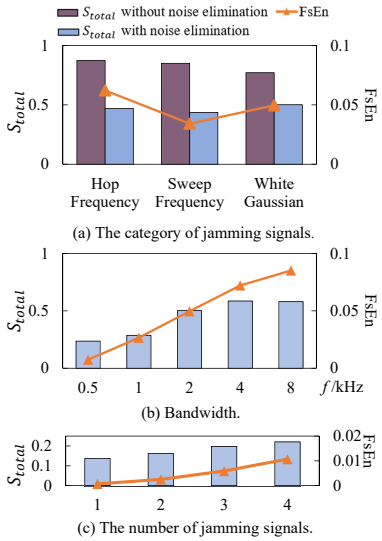


Fig. 7. Impact of jamming parameters.

performance. In addition, the wide bandwidth demands transducers of excellent frequency responses. Otherwise, there is an unexpected energy loss. Such hardware limitations may be the chief cause of the drop at 8 kHz in Fig. 7(b) as well as the vulnerability of Backdoor.

5.5.3 Number of Jamming Signals

Recalling Formula. 4, the number of jamming signals is a key parameter. We test on the tone signals with four frequencies within 2 kHz. Figure 7(c) illustrates that the FsEn multiplies with the increase of jamming signals and the multi-source UMJ behaves better. Inspired by this, an effective way for performance enhancements lies in multiple simultaneous jamming signals.

5.5.4 Human Voice as the Noise

We observe that UMJs using human voices as jamming noises are also vulnerable to BSS, with the average S_{total} dropping from 0.78 to 0.42 experimentally. The seemly complex noise, the human voice, performs unsatisfactorily. Its independence from audios to be protected is to blame for the vulnerability against BSS. Though adversaries might be confused about which one is the valuable speech, privacy is still completely leaked. Human voices are scarcely beneficial to UMJS' resilience without careful arrangement.

In brief, FsEn can serve as the baseline for the selection of parameters, where a high FsEn brings about better defensive effectiveness, accordant with the trend in Fig. 7(b) and (c). Note that it cannot be used to compare UMJs with different signal categories. Besides, it is unwise to increase complexity blindly considering the performance degradation when the bandwidth is set to 8 kHz in Fig. 7(b).

5.6 Impact of Implementation

The condition of the implementation plays a role in the defensive performance of a UMJ, including position, power supply, carrier frequency, and hardware platforms.

5.6.1 Distance and Orientation

The ultrasound is extremely sensitive to position due to its directionality. Therefore, investigating the effectiveness of UMJs under different distances and orientations is necessary for the optimization of the layout. We rotate the spy microphone around the fixed UMJs at different distances in the simulation. The average S_{total} among UMJs using representative signal categories [5]–[8] is distributed, as illustrated in Fig. 6(a). These UMJs just work at a narrow angle within 75 cm. They get high marks on the UMJs' central axis but show a sharp decrease downstream. A UMJ appears sensitive to the orientation and far from robust to the position of the spy microphones. It is an attractive option to arrange multiple UMJs around the sound source for effective and omni-directional jamming coverage [5], but it is still unable to cope with noise elimination methods.

5.6.2 Power Supply

Power seals the upper limit of a UMJ's performance. The higher power will directly increase the jamming intensity. Taking Backdoor as an example, its SNR is set to change from -40 dB to -60 dB at the interval of -5 dB. Its performance on CRR with or without each noise elimination technique is shown in Fig. 6(b). A noise along with a higher power supply seems more effective. However, in practice, it is observably vulnerable to adversarial noise eliminations as shown in Fig. 6(b). To be specific, CRRs tend to significantly decline against BSS, NF, and ANF with a peak CRR of 94.5%, even though SNR is set to -60 dB. The UMJ's resilience improves slightly as SNR increases, but with an actual decline of just 0.1 percent. By contrast, CRR against WBSF remains low in spite of the increasing power. Briefly, the increase of power is beneficial to the resilience of a UMJ, but the improvement is extremely restricted.

5.6.3 Carrier Frequencies

Different spy microphones vary in non-linearity frequency response. They show diverse performances if being jammed by the noises of different frequencies. Therefore, the selection of carrier frequencies might affect jamming effectiveness. We implement Wearable Jammer [5] with carrier

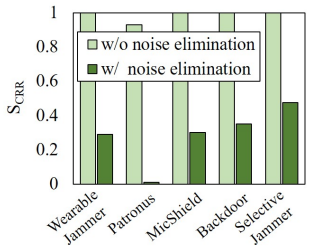


Fig. 8. Performance against digit passwords eavesdropping.

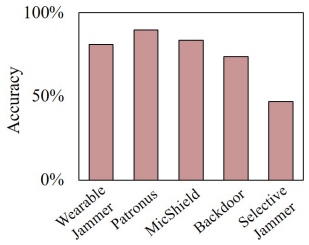


Fig. 9. Accuracy of comprehending polluted speeches.

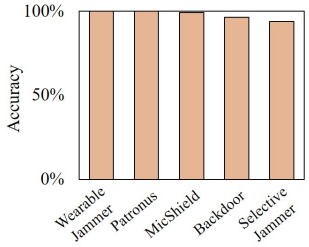


Fig. 10. Accuracy of inferring ambient information.

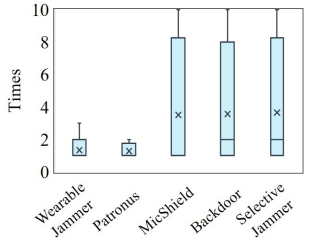


Fig. 11. Attempts needed before a successful replay attack.

frequencies ranging from 25 to 40 kHz at an interval of 1 kHz. Its average scores on multiple microphones fluctuate within (0.72, 0.8). Nevertheless, after noise elimination, its S_{total}s always drop to 0.58. Such results reflect that carrier frequencies affect UMJs' effectiveness but benefit barely the resilience against adversarial noise removal.

5.6.4 Unique Platforms of Existing UMJs

Previous works adopted several strategies about actual hardware/system design for performance improvements. Wearable Jammer [5] uses a 3D circular array design to increase spatial coverage. Patronus [6] includes a reflection layer to increase coverage. MicShield [7] employs a 2D planar circular array. In essence, these measures merely arrange the noise energy distribution with a significant SNR negative gain somewhere. However, conclusions in Sec. 5.6.2 stress that the SNR decrease barely benefits the resilience of UMJs. We duplicate these original proposals and repeat experiments in Sec. 5.3. They remain vulnerable to adversaries with an average S_{total} of about 0.45. In addition, MicShield [7] requires the frequent listening of users' speeches. The risk of privacy leakage is just transferred from voice assistants to their system.

6 Cases Studies

We carry out several cases studies to simulate some practical attacks mentioned in Sec. 3 to emphasize the vulnerability of the existing jammers and demonstrate the realistic threats from eavesdropping.

6.1 Stealing Passwords

In daily life, people may let slip their passwords, which is a threat to their account security. With the purpose of stealing passwords, the adversary would be concentrated on the relevant information, e.g., some sensitive keywords related to passwords, e.g. 'account', 'password', or a series of numbers. Once capturing those hot words, the adversary would on the alert for suspicious passwords in this context. Under that premise, the adversary's recognition rate on these certain segments will increase sharply, especially supported by multiple human recognizers, which are more sensitive to some heuristic knowledge.

We consider a common scene where passwords are composed of pure digits, which are widely used in bank accounts, electrical locks, and PINs. We select 3000 items of audio segments of spoken digits (0-9) from 60 human speakers in AudioMNIST [49] and repeat experiments for testing the performance of UMJs in protecting oral passwords under the identical conditions in Sec. 5.3. Comparing Fig. 8 with Fig. 3(c), the S_{CRR}s on words or digits seems identical among five UMJs without noise elimination. The average recognition rates of each digit are shown in Tab. 3, which shows that less than only 2% words can be recognized without noise elimination. However, after the noise elimination, these jammers are more vulnerable in the aspect of digital recognition. In accord with the assessment in Sec. 5.3.4, Patronus still performs worst, with almost all oral digits recognized, while Selective Jammer scores highest. Nevertheless, the S_{CRR} of Selective Jammer declines by 0.13 in comparison with that in Fig. 3(c). Specifically, the average S_{CRR}s in protecting each digit are: 0.40 for digit 'zero', 0.33 for 'one', 0.31 for 'two', 0.29 for 'three', 0.32 for 'four', 0.29 for 'five', 0.25 for 'six', 0.21 for 'seven', 0.22 for 'eight', and 0.26 for 'nine', respectively. A more specific description of each representative UMJ on different digits is shown in Tab. 3. Among these digits, 'seven' and 'eight' are easier to be recognized, while 'zero' is the most difficult one. Such a decline of S_{CRR}s further demonstrates the vulnerability of existing UMJs.

6.2 Comprehension on Polluted Speeches

We take the human's ability of comprehension into consideration by involving the intelligibility metrics and emphasize these metrics in some cases where an adversary can access victims' private information (including gender, birthday, occupation, interest, address, and calendars), and initiate further targeted attacks, e.g. targeted advertisement, defrauding, and impersonation. Similarly, confidential political, commercial, and military information are at risk.

Here we select 200 audios attached with two to five listening questions to simulate the real-world scenes where adversaries eavesdrop on victims' conversations. These audios are processed under identical conditions in Sec. 5. Each volunteer is asked to listen to ten audios. The relevant questions are listed in written form and volunteers need to choose the best answer from four choices, with their average accuracy shown in Fig. 9. The ranking of the five UMJs agrees with that in Fig. 3(b). This demonstrates the validity of S_{In} in our framework. Some UMJs, such as Patronus and MicShield, have close S_{In}s, but their performances in this case differ. It is because the comprehension result owes to not only intelligibility but also recognition rate.

6.3 Inferring Ambient Information

To demonstrate the necessity of involving the intensity metrics, we test existing UMJs in defending against ambient

TABLE 3
Five Representative UMJs' S_{CRR}s on Different Digits with/without Noise Elimination

Representative UMJ	Noise Elimination	S_{CRR}s on Different Digits										
		Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Overall
Wearable Jammer [5]	with	0.40	0.39	0.38	0.36	0.35	0.34	0.32	0.13	0.13	0.10	0.29
	without	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Patronus [6]	with	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	without	1.00	0.96	1.00	0.97	0.92	0.89	0.91	0.90	0.89	0.90	0.93
MicShiled [7]	with	0.46	0.43	0.35	0.33	0.36	0.33	0.16	0.12	0.12	0.33	0.30
	without	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Backdoor [8]	with	0.67	0.35	0.32	0.32	0.31	0.30	0.31	0.31	0.31	0.31	0.35
	without	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Selective Jammer [9]	with	0.42	0.46	0.45	0.44	0.53	0.46	0.47	0.48	0.53	0.52	0.48
	without	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	with	0.40	0.33	0.31	0.29	0.32	0.29	0.25	0.21	0.22	0.26	0.29
	without	1.00	0.99	1.00	0.99	0.99	0.98	0.99	0.98	0.98	0.98	0.99

information inference. The environment often contains a variety of privacy information. It may leak victims' privacy in a non-verbal way when the victims are unlocking phones [56], typing [37], opening locks [57], and watching monitors [39].

We collect two kinds of bells that are common in offices. The two bells serve as the targeted ambient private information which could imply the location of a victim. They are with different distributions in the range of [100, 5000] Hz. Each lasts 0.3 seconds and is measured as around 60 dB-SPL. We provide 200 permutations with random intervals. Volunteers are asked to point out its permutation after listening to a recovered audio from each noise elimination method. Missing and wrong selection are both recognized to the care in which a UMJ works.

As demonstrated in Fig. 10, jamming noise can cover up the ambient sound but with low S_{SNR}s. Almost all of them are detected once the noise is removed by any technique in real-world scenarios. Though the real ambient sound is likely to be various and slight, an adversary still accomplishes his/her own end by means of various advanced techniques, e.g. artificial intelligence (AI).

6.4 Replay Attacks on Voice Authentication

Replay attacks on voice-based applications using illegal recordings also threat users' privacy. Voice authentication is widely used with the fast development of voiceprint recognition. As a representative of voice-based applications, it suffers from replay attacks which play recorded passwords iteratively to spoof authentication. An adversary can take control of the victim's VA to steal money, visit a malicious website, and monitor the applications in a smartphone and smart home under the circumstances.

Volunteers are asked to register WeChat voiceprint. The whole legal process of voice authentication is recorded. If the number of attempt times is more than ten, the authentication process is failed and we record the number of times as ten. These raw recordings are capable to spoof its authentication after only one attempt. As shown in Fig. 11, the voices protected by MicShield, Backdoor and Selective Jammer fail to pass the voice authentication under some ineffective noise eliminations. However, most of the voices protected by other UMJs only need one or two attempts to pass the authentication. This means that there is a long way to go before both voiceprint and jammers are secure.

7 Discussion

We discuss the defense strategies to address the adversary in our threat model. We summarize some suggestions beneficial to the design of UMJs and propose a novel method of resisting adversarial noise removal. We also consider several practical issues on UMJs and our framework.

7.1 Suggestions on Prospective Designs

We summarize some suggestions on the subsequent design of a resilient UMJ for future reference.

Multi-source Jamming. Multiple jamming signals can be involved simultaneously with the deployment of multiple low-cost transducers. This increases the complexity of jamming noise as elaborated in Sec. 5.5.3 and expands the jamming coverage. Particularly, it requires more spy microphones and higher cost to benefit UMJs' resilience against BSS.

Overlapped Placement. Placing the UMJ close to the protected sound source can defend against malicious beamforming. According to our experiment, a MicShield with a distance of 10 cm can protect over 85% of speeches against a beamforming microphone array placed 1 m away. Another countermeasure to defend against malicious beamforming is to ensure that the noises and sounds come from the same direction. For instance, the defender can place several UMJs around the protected area, transmitting noises from every direction to the eavesdropping microphone array, which can ensure the existence of noises transmitting from the same direction with sounds wherever the source.

Appropriate Bandwidth. The superior performance of the jamming signals with a bandwidth of 4 kHz in Fig. 7(b) demonstrates the importance of bandwidth selection. An appropriate bandwidth implies high efficiency in privacy protection. Moreover, UMJs can dynamically strategize about energy allocation and bandwidth according to the

distribution of speeches to be protected. In addition, a rapid frequency alternation benefits the complexity of jamming signals statistically and UMJ's performance.

Coherent Noise. Existing UMJs employ noise that is independent of speeches. This benefits the most aggressive BBS that is shown in Fig. 6(b). Jamming noise is removed by an adversary without difficulty. Conversely, coherent noise can couple with speeches. They will be more indistinguishable from speeches, along with the promotion of resilience.

In short, the prospective UMJ design tends towards complexity, dynamism, and coherence. Particularly, we suggest a series of independent broadband noises. These noises are coupling with speeches to be protected, with the dynamic energy distribution among the appropriate bandwidth.

7.2 Active Noise Cancellation Based Proposal

Inspired by the active noise cancellation (ANC) technique used in the earphone, we propose a novel microphone-jamming method utilizing an inverse signal of speeches to reduce the recorded speeches, along with the coherent noise as mentioned above.

In our vision, the ANC-based method will exploit a signal processing module and an ultrasonic transducer array. Different from existing UMJs, we suggest the signal processing module using two microphones and a processor. One microphone is planted near the sound source to be protected and keeps capturing the private sound signal. Then the processor generates an inverse signal according to the captured signal, modulates it to the ultrasonic band, and sends it to the ultrasonic speaker for playing. The other microphone is planted in front of the ultrasonic speaker. It measures the cancellation residue and provides feedback to the processor for inverse signal generation. With such a feedforward-feedback architecture, this method will be able to generate a sophisticated inverse signal. Note that the information protected by the ANC technique is selected by the users, who can put the microphones near the sound source to be protected, for example, near a user's mouth for protecting speeches. Such a scheme can defend the sound source against eavesdropping from the perspectives of collaborative recognition and semantic comprehension, due to its effect of reducing information leakage. As for the leakage of ambient information, we suggest adopting coherent noises simultaneously to reduce SNR by both decreasing information leakage and increasing noise intensity. In comparison, prior ANC work [32] aims at canceling the ultrasonic-injected adversary voice commands to make the recording microphones unaffected. It acquires adversary voice commands in a created spectrum as the reference signal, and uses the ANC method to cancel the injected audios in the audible spectrum. Here, ANC is used for digital data processing in the users' microphones. Different from this work [32], our work aims at canceling the audio signals recorded by the spy microphones. We use the ANC method to generate an inverse analog signal, and then emit it to spy microphones for anti-eavesdropping.

We conduct a pilot study to demonstrate the feasibility of our proposal. We leverage an ultrasonic transducer array consisting of 38 NU40A16TR-1 ultrasonic transducers, half of which are used to transmit the ultrasonic carriers, and the others emit inverse signals and coherent noises. A sound vibration measurement device NI USB-4431 connected to a laptop is implemented as the processor. USB-4431 undertakes the signal input/output task and the laptop executes the code written in Labview for signal analysis and generation. Under the conditions in Sec. 5.1, our proposal achieves excellent performance. It yields a high S_{total} of over 0.8 and an S_{CRR} of 0.98 against the denoising techniques as detailed in Sec. 5.1. Such an ANC-based method outperforms all tested representative UMJs. Such a result demonstrates the effectiveness of our suggestions. We will explore the theoretical basis with an acoustic propagation model to improve our ANC-based proposal and adapt it for various real-world scenes in our future work.

7.3 Practical Considerations

Microphones difference. Recent advances have proved that UMJs transmitting ultrasounds with frequencies between 25 kHz and 50 kHz are able to jam off-the-shelf recording devices [5]–[8], [34]. We have discussed the effect of carrier frequencies in Sec. 5.6.3. However, we find that some recording devices, e.g., a Yescool A7 recorder, have no non-linearity in the band between 25 kHz and 40 kHz. Fortunately, a jamming signal with a higher frequency, i.e., over 60 kHz, can inject noises into the above devices by means of the non-linearity. Thus, as a defender, UMJs should leverage multiple frequency carriers covering a wide frequency band. This guarantees UMJs' effectiveness against spy devices with different non-linearity responses but increases the energy consumption.

Jamming coverage. Although UMJs have limited distances and narrow angle (See Sec. 5.6.1), the jamming coverage can be promoted by utilizing a higher power supply and multiple transmitters. The specific platforms in turn promote it. However, as analyzed in Sec. 5.6, such settings expand the jamming coverage but cannot benefit resilience against adversarial noise elimination methods.

Scalability. Except for UMJs, there are some other kinds of microphone jammers with unknown performances [12], [58]. Fortunately, our framework can be adapted for the evaluation on these microphone jammers utilizing noise to pollute eavesdropped recordings. It is because our framework is only correlated to the eavesdropper's ability to extract privacy from noisy recordings, regardless of how to add noise on the recordings. Hence, the framework can provide a comprehensive evaluation on existing microphone jammers and raise concern over speech privacy.

8 Related Work

Microphone Jammers. There are three types of jammers to combat covert microphone-based eavesdropping: the electromagnetic, audible, and UMJ. Electromagnetic jammers [12] require prior knowledge about the target devices. Noisy signals from audible jammers [58] can be heard by users. As a comparison, UMJs overcome the above shortcomings and are promising in anti-recording [5]–[8], [10], [11].

Acoustic Non-linearity. A microphone exhibits square-law nonlinear characteristics [13], [15]. *DolphinAttack* [29] initially accomplishes inaudible command injection on VAs.

Effective attacks are proposed further to expand the coverage [59], [60]. He *et al.* [32] present an active inaudible-voice-command cancellation as a defence. *EarArray* [34] leverages the spatial features to detect malicious injection. In contrast, *Backdoor* [8] leverages this property for anti-eavesdropping. Researches on its prospect are conducted, such as wearable implementation [5], jamming coverage [10], [11], selective jamming [6], [9], and speech protection for VAs [7]. In addition, this characteristic is leveraged for indoor localization [61], device fingerprint [62], and communication [8], [63].

9 Conclusion

We design a comprehensive evaluation framework towards the resilience of UMJs. It contains 12 metrics from perspectives of intensity, intelligibility, and collaborative recognition rate, in correspondence with the potential eavesdropping threats in real-world scenarios. Guided by the framework, we assess representative UMJs and reflect their vulnerability. We analyze the key parameters on UMJs' performances and propose suggestions for further designs. Our framework can act as a stepping-stone for thorough speech privacy protection.

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, 61872285, 62032021, 61772236, 62172359, and 61972348, Research Institute of Cyberspace Governance in Zhejiang University, Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005), Zhejiang Key R&D Plan (Grant No. 2019C03133), Ant Group Funding No.Z51202000234, and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

References

[1] Y. Chen, M. Gao, Y. Li, L. Zhang, L. Lu, F. Lin, J. Han, and K. Ren, "Big brother is listening: An evaluation framework on ultrasonic microphone jammers," in *IEEE International Conference on Computer Communications*, 2022, pp. 1119–1128.

[2] 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.

[3] S. Maheshwari, "Hey, alexa, what can you hear? and what will you do with it?" https://www.nytimes.com/2018/03/31/business/media/amazon-google-privacy-digital-assistants, 2018.

[4] R. Kiberd, "Hey siri! stop recording and sharing my private conversations," https://www.theguardian.com/commentisfree/2019/jul/30/apple-siri-voice-assistants-privacy/, 2019.

[5] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, "Wearable microphone jamming," in *International Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[6] L. Li, M. Liu, Y. Yao, F. Dang, Z. Cao, and Y. Liu, "Patronus: Preventing unauthorized speech recordings with support for selective unscrambling," in *International Conference on Embedded Networked Sensor Systems*, 2020, pp. 245–257.

[7] K. Sun, C. Chen, and X. Zhang, "'Alexa, stop spying on me!': Speech privacy protection against voice assistants," in *International Conference on Embedded Networked Sensor Systems*, 2020, pp. 298–311.

[8] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 2–14.

[9] Y. Chen, M. Gao, Y. Liu, J. Liu, X. Xu, L. Cheng, and J. Han, "Implement of a secure selective ultrasonic microphone jammer," *CCF Transactions on Pervasive Computing and Interaction*, vol. 3, no. 4, pp. 367–377, 2021.

[10] Y. Chen, H. Li, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, "Understanding the effectiveness of ultrasonic microphone jammer," *CoRR*, vol. abs/1904.08490, 2019.

[11] H. Shen, W. Zhang, H. Fang, Z. Ma, and N. Yu, "Jamsys: Coverage optimization of a microphone jamming system based on ultrasounds," *IEEE Access*, vol. 7, pp. 67 483–67 496, 2019.

[12] D. Kune, J. Backes, S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu, "Ghost talk: Mitigating emi signal injection attacks against analog sensors," in *IEEE Symposium on Security and Privacy*, 2013, pp. 145–159.

[13] M. Abuelma'atti, "Analysis of the effect of radio frequency interference on the dc performance of bipolar operational amplifiers," *IEEE Transactions on Electromagnetic Compatibility*, vol. 45, pp. 453–458, 2003.

[14] G. K. C. Chen and J. J. Whalen, "Comparative rfi performance of bipolar operational amplifiers," in *IEEE International Symposium on Electromagnetic Compatibility*, 1981, pp. 1–5.

[15] J. Gago, J. Balcells, D. GonzÁlez, M. Lamich, J. Mon, and A. Santolaria, "Emi susceptibility model of signal conditioning circuits based on operational amplifiers," *IEEE Transactions on Electromagnetic Compatibility*, vol. 49, no. 4, pp. 849–859, 2007.

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

[17] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. 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, pp. 185–188.

[18] 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.

[19] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (microsoft api, google api and cmu sphinx)," *International Journal of Engineering Research and Application*, vol. 7, no. 3, pp. 20–24, 2017.

[20] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *International Conference on Spoken Language Processing*, 1998, pp. 2819–2822.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[22] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1982, pp. 1278–1281.

[23] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[24] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[25] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The journal of the acoustical society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 749–752.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[28] EchoZju, "Evaluation Framework on UMJ," https://github.com/EchoZju/Evaluation-Framework-on-UMJ, 2022.

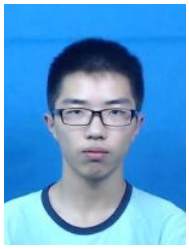
[29] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *ACM conference on computer and communications security*, 2017, pp. 103–117.

[30] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, pp. 1647–1652, 1979.

[31] L. Kohnfelder and P. Garg, "The threats to our products," *Microsoft Corporation 33*, 1999.

[32] Y. He, J. Bian, X. Tong, Z. Qian, W. Zhu, X. Tian, and X. Wang, "Canceling inaudible voice commands against voice control systems," in *International Conference on Mobile Computing and Networking*, 2019, pp. 28:1–28:15.
[33] S. Makino, T.-W. Lee, S. S. Makino, and H. Sawada, *Blind speech separation*. Dordrecht: Springer Netherlands, 2007.
[34] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "Eararray: Defending against dolphinattack via acoustic attenuation," in *Network and Distributed System Security Symposium*, 2021.
[35] M. A. Al Faruque, S. R. Chhetri, A. Canedo, and J. Wan, "Acoustic side-channel attacks on additive manufacturing systems," in *ACM/IEEE International Conference on Cyber-Physical Systems*, 2016, pp. 19:1–19:10.
[36] M. Backes, M. Dürmuth, S. Gerling, M. Pinkal, and C. Sporleder, "Acoustic side-channel attacks on printers," in *USENIX Security Symposium*, 2010, pp. 307–322.
[37] D. Asonov and R. Agrawal, "Keyboard acoustic emanations," in *IEEE Symposium on Security and Privacy*, 2004, pp. 3–11.
[38] Z. Xiao, T. Chen, Y. Liu, and Z. Li, "Mobile phones know your keystrokes through the sounds from finger's tapping on the screen," in *IEEE International Conference on Distributed Computing Systems*, 2020, pp. 965–975.
[39] D. Genkin, M. Pattani, R. Schuster, and E. Tromer, "Synesthesia: Detecting screen content via remote acoustic side channels," in *IEEE Symposium on Security and Privacy*, 2019, pp. 853–869.
[40] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.
[41] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[42] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.
[43] A. Twin, "Delphi method definition," https://www.investopedia.com/terms/d/delphi-method.asp, 2022.
[44] P. F. Karl, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.
[45] S. G. M. and A. R. A. Jr., "Analyzing and interpreting data from likert-type scales," *Journal of graduate medical education*, p. 541, 2013.
[46] EchoZju, "Big-brother," https://zenodo.org/badge/latestdoi/304488420, 2022.
[47] Jinci Technologies, "Product review," http://www.jinci.cn/en/goods/112.html, 2021.
[48] SIGLENT Technologies, "SDG1000 series waveform generators," https://www.siglenteu.com/waveform-generators/, 2021.
[49] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *arXiv preprint arXiv:1807.03418*, 2018.
[50] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.
[51] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural models," in *AIP Netural Networks for Computing*, 1987, p. 206211.
[52] S. Sur, T. Wei, and X. Zhang, "Autodirective audio capturing through a synchronized smartphone array," in *International Conference on Mobile Systems, Applications, and Services*. ACM, 2014, pp. 28–41.
[53] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.
[54] W. Chen, J. Zhuang, W. Yu, and Z. Wang, "Measuring complexity using fuzzyen, apen, and sampen," *Medical Engineering & Physics*, vol. 31, no. 1, pp. 61–68, 2009.
[55] W. Chen, Z. Wang, H. Xie, and W. Yu, "Characterization of surface emg signal based on fuzzy entropy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 2, pp. 266–272, 2007.
[56] J. Liu, Y. Wang, G. Kar, Y. Chen, J. Yang, and M. Gruteser, "Snooping keystrokes with mm-level audio ranging on a single phone," in *International Conference on Mobile Computing and Networking*, 2015, pp. 142–154.
[57] S. Ramesh, H. Ramprasad, and J. Han, "Listen to your key: Towards acoustics-based physical key inference," in *International Workshop on Mobile Computing Systems and Applications*, 2020, pp. 3–8.
[58] Oeler Industries, "Sound masking device," https://www.oeler.com/sound-masking-systems/, 2021.
[59] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *USENIX Symposium on Networked Systems Design and Implementation*, 2018, pp. 547–560.
[60] L. Song and P. Mittal, "Poster: Inaudible voice commands," in *ACM conference on computer and communications security*, 2017, pp. 2583–2585.
[61] Q. Lin, Z. An, and L. Yang, "Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices," in *International Conference on Mobile Computing and Networking*, 2019, pp. 2:1–2:16.
[62] X. Zhou, X. Ji, C. Yan, J. Deng, and W. Xu, "Nauth: Secure face-to-face device authentication via nonlinearity," in *IEEE Conference on Computer Communications*, 2019, pp. 2080–2088.
[63] G. Zhang, X. Ji, X. Zhou, D. Qi, and W. Xu, "Ultracomm: High-speed and inaudible acoustic communication," in *Quality, Reliability, Security and Robustness in Heterogeneous Systems*, 2019, pp. 184–204.

Ming Gao is a Ph.D. candidate at the school of cyber science and technology, Zhejiang University. He received the Master and Bachelor degree from Xi'an Jiaotong University. His research interests include cyber-physical security, mobile computing, and privacy protection. He is a recipient of the Best Paper Award Nomination from SenSys'21.



Yike Chen is working toward the PhD degree at the School of Cyber Science and Technology, Zhejiang University. His research interests include mobile computing and smart sensing.

Yimin Li received her M.Sc. from University College London in 2021. She was a Visiting Student with Zhejiang University from 2019 to 2020. Her research interests include wireless sensor network and privacy protection.

Lingfeng Zhang is a Master student at the School of Cyber Science and Technology, Zhejiang University, under the supervision of Prof. Jinsong Han. His research interests include cyber-physical security and smart sensing.

Jianwei Liu received the BS degree from Northwestern Polytechnical University in 2018. He received his Master degree from Xi'an Jiaotong University. He is working toward the Ph.D. degree at Zhejiang University. His research interests include RFID, mobile computing, and smart sensing. He is student member of the IEEE.

Li Lu is an Assistant Professor in the School of Cyber Science and Technology and College of Computer Science and Technology at Zhejiang University. He received his Ph.D. and B.E. degrees in Computer Science and Technology from Shanghai Jiao Tong University and Xi'an Jiaotong University, respectively. He was also a visiting research student in Wireless Information Network Laboratory (WINLAB) and Department of Electrical and Computer Engineering at Rutgers University. His research interests include IoT security, mobile sensing, ubiquitous computing and human-computer interaction. He is the recipient of ACM China SIGAPP Chapter Doctoral Dissertation Award, and First Runner-up Poster Award from ACM MobiCom 2019.

Feng Lin received the Ph.D. degree from the Department of Electrical and Computer Engineering, Tennessee Technological University, USA, in 2015. He is currently a Professor with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. He was an Assistant Professor with the University of Colorado Denver, USA, a Research Scientist with the State University of New York (SUNY) at Buffalo, USA, and an Engineer with Alcatel-Lucent (currently, Nokia). His current research interests include mobile sensing, wireless sensing, Internet of Things security, biometrics, and AI security. Dr. Lin was a recipient of the ACM SIGSAC China Rising Star Award, the Best Paper Awards from ACM MobiSys'20, IEEE Globecom'19, IEEE BHI'17, the Best Demo Award from ACM HotMobile'18, and the Best Paper Award Nomination from SenSys'21 and INFOCOM'21. He serves as an editor for IEEE Network.

Jinsong Han received his Ph.D. degree in computer science from Hong Kong University of Science and Technology in 2007. He is now a professor at the School of Cyber Science and Technology, Zhejiang University. He is a senior member of the ACM and IEEE. His research interests focus on IoT security, smart sensing, wireless and mobile computing.

Kui Ren (Fellow, IEEE and ACM) received the Ph.D. degree in electrical and computer engineering from the Worcester Polytechnic Institute. He is currently a Professor and the Associate Dean of the College of Computer Science and Technology, Zhejiang University, where he also directs the Institute of Cyber Science and Technology. Before that, he was the SUNY Empire Innovation Professor of The State University of New York at Buffalo. His H-index is 74 and his total publication citation exceeds 32 000 according to Google Scholar. His current research interests include data security, the IoT security, AI security, and privacy. He has published extensively in peer-reviewed journals and conferences and received the Test-of-Time Paper Award from IEEE INFOCOM and many Best Paper Awards from IEEE and ACM, including MobiSys 2020, Globecom 2019, ASIACCS 2018, and ICDCS 2017. He received the NSF CAREER Award in 2011, the Sigma Xi Research Excellence Award in 2012, the IEEE CISTC Technical Recognition Award in 2017, the SUNY Chancellor's Research Excellence Award in 2017, and the Guohua Distinguished Scholar Award from ZJU in 2020. He is a Clarivate Highly-Cited Researcher. He is a frequent reviewer for funding agencies internationally and serves on the editorial boards of many IEEE and ACM journals. He also serves as the Chair for SIGSAC of ACM China.