Big Brother is Listening: An Evaluation Framework on Ultrasonic Microphone Jammers

Yike Chen†, Ming Gao†, Yimin Li, Lingfeng Zhang, Li Lu, Feng Lin, Jinsong Han*, Kui Ren
Zhejiang University, China
{chenyike, gaomingppm, 3160102338, li.lu, flin, hanjinsong, kuiren}@zju.edu.cn, ninalym13@gmail.com

***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 in 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.**

I. Introduction

Eavesdropping or recording via microphones has always been a serious privacy threat. Nowadays, ubiquitous smart devices, such as smartphones and voice assistants (VA), are reported to eavesdrop on private speeches and pass recordings along to third-party [1]–[3], which exacerbates this threat.

To combat microphone-enabled eavesdropping, researchers proposed the ultrasonic microphone jammers (UMJs) [4]–[9]. Compared with the conventional electromagnetic and audible jammers [10], 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 [11], [12]. With this property, ultrasounds that are imperceptible to human ears over the air, would leak energy into the audible spectrum when arriving microphones [13]. This noise migrates 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 [4]–[7].

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, a primer design of MicShield, a representative UMJ, was reported to be vulnerable to beamforming-based eavesdropping attack (with 75.0% jammed fragments recovered and recognized by adversaries) [6]. Other noise elimination methods are also effective in jamming reduction in practice according to our experiments in Sec. V-C. 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 [1]. The adversary infer 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 observation in mind, we explore and summarize realistic threats from sophisticated adversaries into three perspectives, including *ambient information*, *semantic comprehension*, and *collaborative recognition*. First, even if an 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 that 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) [14], CMU Sphinx [15], and iFLYTEK [16], acute to distinct words. Although their recognition results are variant due to different intrinsic models and algorithms [17], a smart adversary can integrate their results [17] to recover more information, even

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

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 also refine the eavesdropping model by complementing the adversary with a practical denoising ability. Correspondingly, we propose a comprehensive evaluation framework that leverages 12 metrics to cover the above attacking surfaces. As for *ambient information*, we employ three *intensity* metrics [18], [19] to cover both the verbal and non-verbal factors. As for *semantic comprehension*, we adopt six *intelligibility* metrics [20]–[25] 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 quantify the defensive effectiveness of four representative UMJs [4]–[7]. 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 a comparative analysis and present several constructive suggestions for future UMJ designs.

Our contributions are summarized as follows:

- We propose a comprehensive framework for evaluating UMJs' defence effectiveness. Involving 12 metrics, it covers potential threats as much as possible in real eavesdropping attacks, enabling a thorough evaluation on UMJs.
- 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 vulnerabilities of existing UMJs against sophisticated adversaries in real-world scenarios.
- We outline the key factors influencing UMJs' performance and summarize constructive suggestions for the improvement and future design of UMJs based on our comparative analysis. We also release our source code [26] to facilitate the anti-eavesdropping research.

II. Background

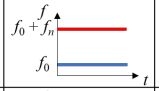
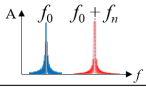
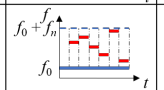
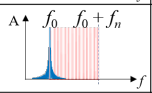
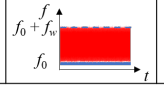
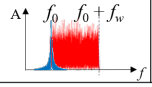
A. Principle of UMJs

An UMJ leverages microphone non-linearity [7] for jamming, where the inner amplifier exhibits square-law nonlinear characteristics [11], [13], especially when the input frequency is above 25 kHz. Hence, some high-frequency signals are demodulated to low-frequency signals intentionally. Concretely, suppose an audio input $x(t)$, and 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 [27].

TABLE I
THE COMMON CATEGORIES OF JAMMING SIGNALS

Category	Frequency-time ($f-t$)	Amplitude-frequency (A – f)	Representatives
Tone Signal(s)			
Dynamic Single Frequency Noise (DSFN)			Wearable Jammer[4] Patronus[5]
White Gaussian Noise (WGN)			MicShield[6] Backdoor[7]

An 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 $others$ represents 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 the microphone, making the private speech unrecognizable by virtue of the masking properties [28]. With the aforementioned properties, the UMJs induce the modulated ultrasonic signals to inject the arranged noise into a spy microphone.

B. Classification of Jamming Signals

The essential differences among UMJs [4]–[7] lay 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. I.

Tone signal. The previous example, $m(t) = cos[2\pi(f_c + f_b)t]$, is a typical tone signal. Furthermore, an 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 hop frequency signal [4], [5] is a representative. It scrambles discretely and randomly at a predetermined interval as following,

$$m(t) = cos[2\pi(f_0 + a[\lfloor \frac{t}{p} \rfloor])t], \tag{5}$$

where f_0 is the sum of the carrier frequency f_c and a 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.

White Gaussian noise (WGN). Some approaches [6], [7] recommend WGN for jamming, whose energy is distributed over a broad ultrasonic spectrum. We have

$$m(t) = gwn(f_0, f_0 + f_w), \tag{6}$$

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

We categorize the representative UMJs [4]–[7] according to the above standards. The systems we evaluated in this paper are all based on these representative prototypes.

III. 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.

A. Threat Model

We follow the well-known STRIDE threat model [29] to refine the threat model. Here, the information disclosure model [29] 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 a VA. He can deploy spy devices at suitable places for articulate and complete recordings. Even if illegal recordings are awash with jamming noise, he would recover and extract the private information by the means including but not limited to those detailed in Sec. III-B. Furthermore, he 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 the private information from jammed recordings. He would result in three-fold threats (See Sec. III-C) using various methods (See Sec. III-B).

B. Noise Elimination Methods

Different from the assumption that an adversary would give up once jammed [4], [7], we point out that he would endeavor for noise elimination and information extraction. We consider two noise elimination algorithms, i.e., blind source separation (BSS) [30] and filtering within the adversary's capability.

1) BSS: BSS 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 demands 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.

2) Filtering: Filtering requires basic knowledge of noise characteristics, such as frequency distribution. Unfortunately, the adversary could analyze and 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. Furthermore, we exploit the coupling of ultrasound for a real-time noise distributions analysis. Therefore, the adversary can conduct an *adaptive notch filter (ANF)* for noise elimination. Specifically, an ultrasonic tone at the frequency f_1 can couple with jamming signals and introduce a low-frequency component at the frequency $|f_1 - f_c|$. This component can imply the frequency distribution of jamming noise [31]. With such a reference, the adversary could use ANF for anti-jamming.

In practice, adversaries may utilize multiple techniques concurrently. Besides the above methods, sophisticated noise reduction methods including traditional statistic-based [32] and advanced learning-based ones [33] are also popular. Unfortunately, results in Sec. V-C demonstrate that existing UMJs [4]–[7] are extremely vulnerable to the above-mentioned methods, let alone other sophisticated techniques.

C. Three-perspective Architecture on Eavesdropping Threats

We organize eavesdropping threats 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), the adversary can surmise that the victim is in an office if the spy microphone captures noise emitted from printers. He could further draw the victim's daily routine.

The adversary can semantically comprehend the meaning of speech from partly-unrecognizable conversations. This is because adversaries 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!'. 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 extensively studied in previous works [4]–[7], the collaboration between multiple ASRs and humans on recognition is overlooked. Utilizing different intrinsic models, current ASRs [14]–[16] are acute to distinct words and generate different recognition results [17]. Moreover, 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. III-B to exacerbate eavesdropping threats. On account of such a three-perspective architecture,

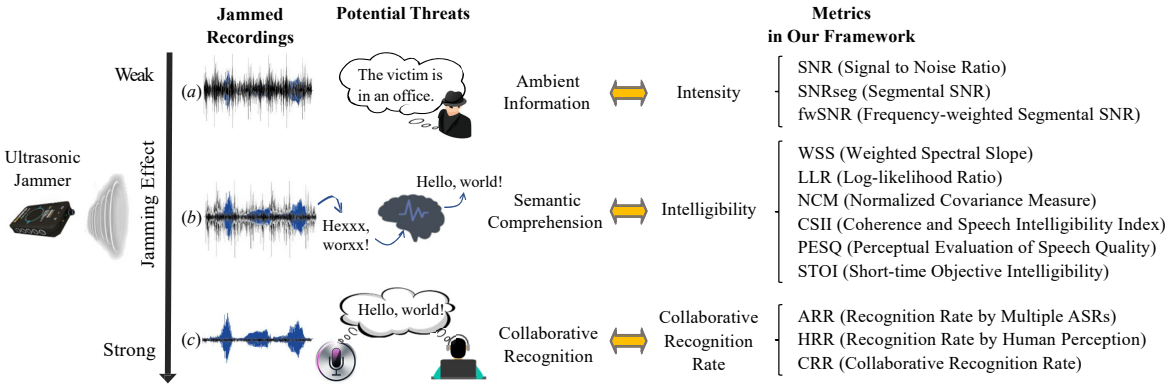


Fig. 1. Three-perspective architecture completely containing potential eavesdropping threats, consisting of ambient information, semantic comprehension, and collaborative recognition. Correspondingly, we propose an evaluation framework on UMJs' resilience, comprised of 12 metrics from these perspectives.

a comprehensive framework should concern all the above perspectives to evaluate UMJs.

IV. 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 rate.

A. Intensity

Previous evaluations based on speech recognition serve no purpose in inferring the environment, leaving a flaw of privacy leakage. Hence, a thorough assessment of UMJs ought to consider both the verbal and non-verbal factors. We leverage the *intensity* metrics [18], [19] to describe these factors. We adopt the Signal to Noise Ratio (SNR) and its derivatives to assess UMJs against ambient information leakage.

SNR is a widely used metric 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. Segmental SNR (**SNRseg**) [18] and frequency-weighted segmental SNR (**fwSNR**) [19] jointly embody the jamming intensity in short segments. In our experiment, all sounds except jamming noise are treated as signal so as to include background sounds.

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

$$S_{SNR} = \boldsymbol{w_{SNR}} \cdot [SNR\ SNRseg\ fwSNR]^T, \quad (7)$$

where w_{SNR} is a 3-dimensional weight vector. The three intensity metrics are normalized by the theoretical or experimental maximum and minimum values respectively following a linear conversion function. This matrix is set as [0.34 0.22 0.44] based on their correlations with subjective ratings of signal quality [34]. 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.

B. 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 [20]–[25]. It can be described by measuring audio distortions [25], [34]. 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 [35]. We utilize the following metrics to cover all distortions and evaluate the intelligibility of speeches comprehensively.

The propagation distortion component could be depicted by the spectral envelope difference. We utilize two basic measurements i.e. weighted spectral slope (**WSS**) [20] and log-likelihood ratio (**LLR**) [21] for description. WSS is the weighted difference between the spectral slopes in each frequency band and LLR is the likelihood ratio of features in 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**) [22] and coherence and speech intelligibility index (**CSII**) [23]. NCM is the weighted sum of normalized covariance signals in each frequency band. CSII is calculated by using coherence as features in the frequency domain. The algorithmic artifacts component is quantitatively described by perceptual evaluation of speech quality (**PESQ**) [24] and Short-time objective intelligibility (**STOI**) [25]. PESQ estimates the overall loudness difference between the noise-free and processed signals, which assesses the speech distortion introduced by artifact algorithms [36]. STOI is designed for measuring speeches processed by noise reduction and speech separation algorithms.

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

$$S_{In} = \boldsymbol{w_{In}} \cdot [WSS\ LLR\ NCM\ CSII\ PESQ\ STOI]^T, \quad (8)$$

where $\boldsymbol{w_{In}}$ is a 6-dimensional weight vector. In general, $\boldsymbol{w_{In}}$ is set as [0.06 0.13 0.20 0.21 0.18 0.22] based on their

correlations with subjective ratings of speech intelligibility [25], [34]. 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.

C. Collaborative Recognition Rate

Existing literatures [4]–[7] take the speech recognition by ASRs (**ARR**) or by human perception (**HRR**) into consideration. 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 the adversary can recognize by jointly using multiple ASRs or human perception. CRR is twofold, subsuming ARR and HRR as follows,

$$\begin{aligned} ARR &= \frac{card(R_{ASR})}{card(S)} = \frac{card(\bigcup_{i=1}^{n} R_{ASR_i})}{card(S)}, \\ HRR &= \frac{card(R_H)}{card(S)} = \frac{card(\bigcup_{j=1}^{m} R_{H_j})}{card(S)}, \\ CRR &= \frac{card(R)}{card(S)} = \frac{R_{ASR} \cup R_H}{card(S)}, \end{aligned} \tag{9}$$

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. We design a recognition score S_{CRR} as follows,

$$S_{CRR} = 1 - CRR. \tag{10}$$

It is positively correlated to the resilience of UMJs against such a man-machine collaborative attack. S_{CRR} can display the security of UMJs against eavesdroppers in speech recognition.

D. Resilience Evaluation for UMJs

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

$$S_{total} = \boldsymbol{W} \cdot [S_{SNR}\ S_{In}\ S_{CRR}]^T, \tag{11}$$

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 utilize principal component analysis (PCA) [37] to determine the weight coefficients $\boldsymbol{W}$ based on the preference of customers. We design a questionnaire to randomly collect the concern of potential customers using Likert-type scale [38], where they score the threat level in each layer by judging several descriptions. In 851 issued questionnaires, 732 participants are willing to use an UMJ and provide their preference on the defence effect toward potential adversaries. 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.

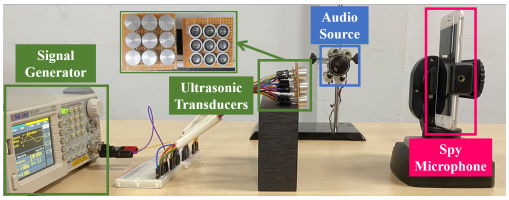


Fig. 2. Experimental setup.

V. Evaluation

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

A. 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 an UMJ: nine sets of NU40A14TR-1 [39] play the jamming signal $m(t)$ while nine sets of NU40C16TR-1 [39] 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 [40].

Recording Devices. We use three kinds of recording devices as spy microphones, i.e., a Samsung Galaxy S8, a Pixel 4, an iPhone 6s, an iPad Pro 2020, and two Thinkpad x201 laptops. The average score of an UMJ on these six devices is regarded as its final score.

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 as 65dB-SPL (dB of sound pressure level), which is the common average loudness of social conversations [7]. The power of jamming signals is set as 115dB-SPL. Because of the intrinsic hardware error, the measured values are 65.1dB-SPL (raw audio), 114.5dB-SPL (DSFN jamming signals [4], [5]), and 105.1dB-SPL (WGN signals [6], [7]). Correspondingly, SNRs retain about -49.4dB [4], [5] or -40dB [6], [7].

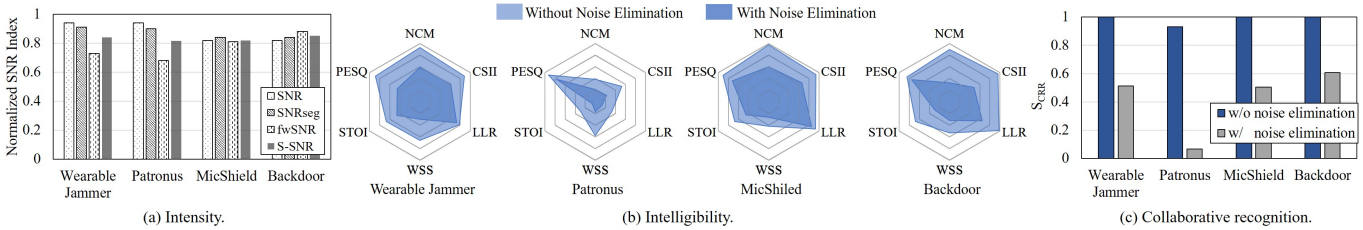
Test Audios. We play 11000 items of audio segments in total, derived from AudioMNIST [41] and Librispeech [42], including common words [7] (accounting for 94%), letters (4%), and digits (2%). The audios are randomly allocated among volunteers. In the experiment, volunteers are allowed to replay the audio until they are able to recognize or give up.

We set an adversary's practical capability as following.

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 [14], CMU Sphinx [15], and iFLYTEK [16]. The speech recognition ratios of these ASRs are claimed to exceed 80% on the raw audios [14]–[16]. (2) Humans: We randomly recruit 20 volunteers aged between 18 and 45 without any knowledge of specific selecting strategies

TABLE II
FOUR 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 [4]	[0,1] kHz DSFN hopping per 0.45 ms	0.88	0.82	0.84	1.00	0.60	0.54	0.72	0.51
Patronus [5]	[85,255] Hz DSFN hopping per 0.2 s	0.75	0.82	0.54	0.93	0.24	0.16	0.45	0.07
MicShiled [6]	4 kHz bandwidth WGN	0.89	0.82	0.85	1.00	0.60	0.56	0.73	0.51
Backdoor [7]	8 kHz bandwidth WGN	0.89	0.85	0.85	1.00	0.58	0.46	0.67	0.61



(a) Intensity. (b) Intelligibility. (c) Collaborative recognition.

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

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).

Noise Elimination. To judge the security of these representative UMJs against realistic adverse approaches, we implement four noise elimination methods as mentioned in Sec. III-B. We utilize (a) BSS with the fast independent cost analysis [43], (b) a STFT-based NF, (c) a WBSF of 2 kHz cut-off frequency, and (d) an ANF with the normalized least mean square (NLMS) algorithm [44].

B. Representative UMJs

We replicate four representative UMJs [4]–[7] under the aforementioned conditions (See Tab. II). Wearable Jammer [4] and Patronus [5] utilize the DSFN signals for jamming, while MicShield [6] 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 chirps to smooth the frequency hopping. It produces a noise that changes frequency among [85,255] Hz for every 0.2 s. MicShield employs WGN with a 4 kHz bandwidth modulated on a 40 kHz carrier. Backdoor [7] tried four kinds of jamming signals and claimed WGN is most useful. Hence, we set WGN of 8 kHz bandwidth as its jamming signals. These four representative UMJs own their unique designs in hardware platform arrangement and energy supply in previous works [4]–[9]. However, we set these parameters to the same for a fair evaluation. The influences of these designs are discussed in Sec. VII.

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 the others nearly reach 1. They seem adequate for effective acoustic privacy protection.

C. Performance and Vulnerability

Based on the proposed framework, we reexamine UMJs' defensive effectiveness against the eavesdropping threats. Fig. 3 presents the scores of these UMJs without noise elimination and their average scores suffering from the denoising means.

1) Overall Performance: Tab. II illustrates the overall performance of these four representative UMJs. The higher score in this table implies better resilience against eavesdropping threats. Results without noise elimination demonstrate that MicShield owns the best defensive effectiveness with S_{total} of 0.89, followed by Backdoor, Wearable Jammer, and Patronus in turn. However, these UMJs are vulnerable against the realistic adversary. From the comparison in Tab. II, their performances decrease by 41.8% on average after the noise elimination. MicShield, the most secure one, maintains a low score of only 0.6, which is still unavailing in face of the sophistication. Patronus scores merely 0.24. It proves barely resilient against arbitrary technique of adversarial noise elimination. Following parts will detail their vulnerability from each perspectives.

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 0.82. We observe that the contribution of each metric is unique. Wearable Jammer and Patronus gain the high SNR and SNRseg. Their bandwidths do not exceed 2 kHz. By contrast, 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. V-D2.

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

3) Intelligibility: We present the performance of these UMJs on each intelligibility metric, with radar charts in Fig. 3(b). Vividly, the smallest area indicates the worst performance of Patronus with the intelligibility score of 0.54, while others score around 0.85. Although the S_{CRR} of Patronus is just 0.07 less than others in Fig. 3(c), the difference between their S_{In}s is high up to 0.31. A slight drop from the perspective of recognition will seriously enervate the defensive effectiveness against the threat of semantic comprehension. This has strengthened our arguments to consider intelligibility metrics.

After noise elimination, S_{In}s present a significant erosion, with a decrease of over 0.24. Their rank of S_{In}s is MicShield, Wearable Jammer, Backdoor, and Patronus. Though the former three UMJs earn the seemingly similar assessment in terms of recognition, they receive diverse scores here. This demonstrates that intelligibility does not depend absolutely on the speech recognition rate. Human comprehension is significant or even dominant at this state.

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 20% and all HRRs are above 35%. Backdoor 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. Fig. 4(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 statistically significant among humans. The cumulative average HRR curve as the incremental volunteers is plotted in Fig. 4(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 an UMJ from the perspective of man-machine collaborative recognition.

Accordingly, we provide experienced manufacturers and average customers with different requirements. The experienced 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 an 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 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.

D. Impact of Signal Categories and Parameters

We analyze the impact of each parameter. It explains the striking different performances among existing UMJs with different signal categories and parameters. We conclude the key factors and provide a reference for the further UMJ design.

We introduce the fuzzy entropy (FsEn) [45], a widely used measurement of the disorder degree. We use it to compare the complexity of jamming signals from a statistical point of view. After necessary preprocessing [46], we have

$$FsEn(m,r) = ln\Phi^m(r) - ln\Phi^{m+1}(r), \tag{12}$$

where Φ is the mean of the degree matrix of membership on top of the elements in the principal diagonal, m is the window length, and r is a parameter, generally set as the quotient of m divided by the standard deviation of the data [46].

1) The Category of Jamming Signals : We design three prototypes of UMJs utilizing different jamming signal. Apart from hop frequency signals and WGN signals in Sec II-B, we introduce the sweep frequency signals, another representative of DFSN signals. It repeats a linear continuous chirp regularly. 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). As illustrated in Fig. 6(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.

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. 6(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

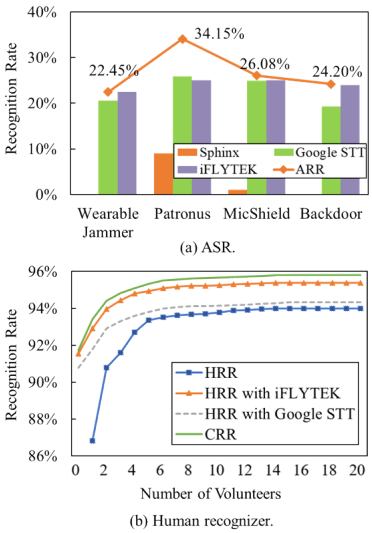


Fig. 4. Impact of recognizers.

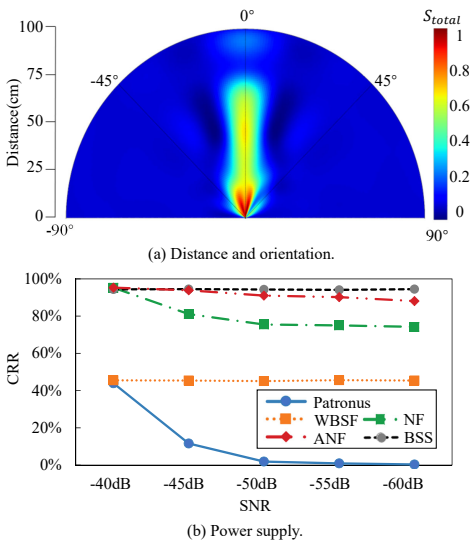


Fig. 5. Impact of Implementation.

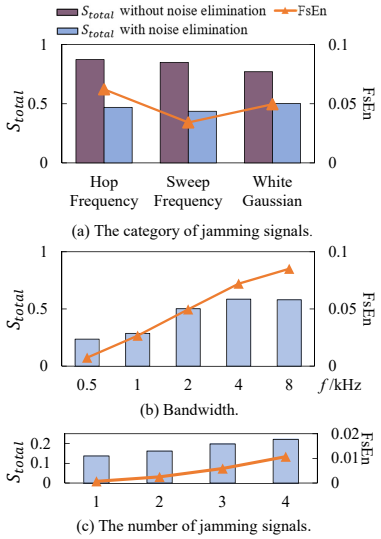


Fig. 6. Impact of jamming parameters.

disperses the noise energy, possibly leading to poor performance. In addition, the wide bandwidth demands transducers of an excellent frequency response. Otherwise, there is an unexpected energy loss. Such hardware limitations may be the chief cause of the drop at 8 kHz in Fig. 6(b) as well as the vulnerability of Backdoor.

3) The 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. Fig. 6(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.

In brief, a high FsEn brings about better defensive effectiveness, accordant with the trend in Fig. 6(b)(c). FsEn can serve as the baseline for the selection of parameters. Note that it cannot be used to compare the UMJs with different signal categories. Besides, it is unwise to increase complexity blindly considering the performance degradation at 8 kHz in Fig. 6(b).

E. Impact of Implementation

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

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 to the optimization of the layout. We rotate the spy microphone around the fixed UMJs at different distances in simulation. The average S_{total} among UMJs using representative signal categories [4]–[7] is distributed, as illustrated in Fig. 5(a). These UMJs just work at a narrow angle within 75cm. They get high marks on the UMJs' central axis but show a sharp decrease downstream. An 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 [4], but it is still unable to cope with noise elimination methods from adversaries.

2) Power supply: Power seals the upper limit of a UMJ's performance. The higher power will directly increase the jamming intensity. Taking Patronus as an example, its SNR is set to change from -40dB to -60dB at the interval of -5dB. Its performance on CRR without and with each noise elimination technique is shown in Fig. 5(b). It seems more effective along with power supply, but appears observably vulnerable to adversarial noise eliminations. Specifically, CRRs tend to significantly increase against BSS, NF, and ANF with the peak of 94.4%, even SNR is set to -60dB. The UMJ's resilience improves slightly as SNR decreases, but with an actual decline of CRR by just 0.5 percent against BSS. By contrast, CRR against WBSF holds the line at a low level in spite of the increasing noise power (a decreasing SNR). Briefly, the increase of power is positive to the resilience of an UMJ, but the improvement is extremely limited.

3) Carrier frequencies: Different kinds of spy microphones vary in non-linearity frequency response. They show diverse performances jammed by noise of different frequency. Therefore, the selection of carrier frequencies might result in jamming performance changes. We reduplicate Wearable Jammer [4] with carrier 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 the adjustment of carrier frequencies affect UMJs' effectiveness but benefits barely the resilience again adversarial noise elimination methods.

4) Unique platforms of existing UMJs: Previous works adopted several strategies about actual hardware/system design for performance improvements. Wearable Jammer [4] uses a 3D circular array design to increase spatial coverage. Patronus [5] includes a reflection layer to increase coverage.

MicShield [6] 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. V-E2 stress that the SNR decrease barely benefits the resilience of UMJs. We duplicate these original proposals and repeat experiments in Sec. V-C. They remain vulnerable to adversaries with an average S_{total} of about 0.45. In addition, MicShield [6] requires the frequent listening of users' speeches. The risk of privacy leakage is just transferred from voice assistants to their system.

VI. 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 jamming noise' complexity as elaborated in Sec. V-D3 and expands the jamming coverage. Particularly, it requires more spy microphones and higher cost to benefit UMJs' resilience against BSS.

Appropriate Bandwidth. The superior performance of 4-kHz-bandwidth jamming signals in Fig. 6(b) demonstrates the importance of the bandwidth. An appropriate bandwidth implies high efficiency in privacy protection. Moreover, UMJs can dynamically strategize about energy allocation and bandwidth according to the distribution of protected speeches. 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. 5(b). Jamming noises are removed by an adversary without difficulty. Conversely, coherent noises 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.

VII. Discussion

Microphones difference. Recent advances have proved that UMJs transmitting ultrasounds with frequencies between 25kHz and 50kHz can jam off-the-shelf recording devices [4]–[7], [47]. We have discussed the effect of carrier frequencies in Sec. V-E3. 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 60kHz, can touch off their non-linearity. Thus, UMJs should leverage multiple 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. V-E1), the jamming coverage can be promoted utilizing a higher power supply and multiple transmitters. The specific platforms as well promote it. However, as analyzed in Sec. V-E, 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 [10], [48]. Fortunately, our framework can be adapted for the evaluation on these jammers utilizing noise to pollute eavesdropped recordings, 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.

VIII. Related Work

Microphone Jammers. Typically, there are three categories of jammers to combat this covert microphone-based eavesdropping: the electromagnetic, audible, and UMJ. Electromagnetic jammers [10] require prior knowledge about the target devices. Noisy signals from audible jammers [48] can be heard by users. As a comparison, UMJs overcome the above shortcomings and are promising in anti-recording [4]–[9].

Acoustic Non-linearity. A microphone exhibits square-law nonlinear characteristics [11], [13]. *DolphinAttack* [27] initially accomplishes inaudible command injection on VAs. Effective attacks are proposed further to expand the coverage [49], [50]. He *et al.* [31] present an active inaudible-voice-command cancellation as a defence. In contrast, *Backdoor* [7] leverages this property for anti-eavesdropping. Researches on its prospect are conducted, such as wearable implementation [4], jamming coverage [8], [9], and selective jamming [5], [6].

IX. Conclusion

We design a comprehensive evaluation framework toward 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 vulnerabilities. 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. We have provided public access to our code [26].

Acknowledgment

This paper is partially supported by National Key R&D Program of China (No. 2021QY0703), National Natural Science Foundation of China (No. U21A20462, 61872285, 62032021, 61772236, 62172359, 61972348 and 62102354), Fundamental Research Funds for the Central Universities (No. 2021FZZX001-27), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (No. 2018R01005), Zhejiang Key R&D Plan (No. 2019C03133), Research Institute of Cyberspace Governance in Zhejiang University, Ant Group Funding (No. Z51202000234), and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.

[17] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (microsoft api, google api and cmu sphinx)," *Int. J. Eng. Res. Appl*, vol. 7, no. 03, pp. 20–24, 2017.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] EchoZju, "Big-brother," https://zenodo.org/badge/latestdoi/304488420, 2022.

[27] 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.

[28] 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.

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

[30] S. Makino, T.-W. Lee, S. S. Makino, and H. Sawada, *Blind speech separation*, 1st ed. Dordrecht: Springer Netherlands, 2007.

[31] 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.

[32] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[33] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[34] 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.

[35] 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.

[36] 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.

[37] K. P. F.R.S., "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.

[38] S. G. M. and A. R. A. Jr., "Analyzing and interpreting data from likert-type scales," *Journal of graduate medical education*, p. 541, 2013.

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

[40] SIGLENT Technologies, "Sdg1000 series function/arbitrary waveform generators," https://www.siglenteu.com/waveform-generators/, 2021.

[41] 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.

[42] 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.

[43] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural models," in *AIP Netural Networks for Computing*, 1987.

[44] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[45] 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.

[46] 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.

[47] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "Eararray: Defending against dolphinattack via acoustic attenuation," in *Annual Network and Distributed System Security Symposium*, 2021.

[48] Oeler Industries, "Sound masking device," https://www.oeler.com/sound-masking-systems/, 2021.

[49] 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.

[50] L. Song and P. Mittal, "Poster: Inaudible voice commands," in *ACM conference on computer and communications security*, 2017.