NUSGuard: Smart Device Anti-Eavesdropping Protection Based on Near-Ultrasonic Interference

Xiaoxiao Qiao, Man Zhou, *Member, IEEE*, Hongwei Li, Xiaojing Zhu, Zhihao Yao, Houzhen Wang, and Xiaojing Ma

***Abstract*—Voice assistants (VAs) have become ubiquitous in smart devices, and are highly valued for their ability to perform a variety of tasks through voice interaction, offering users hands-free convenience. However, the always-on microphones of VAs have raised significant privacy concerns in recent years. In this paper, we propose and implement NUSGuard, a novel and practical anti-eavesdropping system. To our knowledge, it is the first system to utilize the built-in speakers of commercial off-the-shelf (COTS) devices for anti-eavesdropping, thereby eliminating the need for dedicated ultrasonic transmitters. Specifically, it exploits human ears' insensitivity to near-ultrasonic signals and the inherent non-linearity of mic to inject jamming noises into the microphones of unauthorized smart devices. Furthermore, we propose a robust mixed-noise scheme and a lexical-level automatic jammer control strategy, effectively disrupting unauthorized recordings while maintaining seamless voice interaction with authorized VA devices. Extensive digital and real-world experiments have demonstrated NUSGuard's superior performance in terms of jamming effectiveness and security.**

***Index Terms*—Security and privacy protection, anti-eavesdropping, nonlinearity, near-ultrasound.**

I. Introduction

VOICE assistants (VAs) like Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana have become ubiquitous on smart VA devices, including smartphones and home systems. These assistants enable users to control home appliances via voice commands, providing a hands-free experience. This convenience has contributed to VAs becoming an essential component of many users' daily lives. To ensure real-time responsiveness, VA devices often capture or transmit all incoming audio signals for processing [1]. Once captured, voice data becomes difficult to track and control, raising significant privacy concerns [2]. This has led to the question of whether always-online smart devices are turning into constant recorders of our private lives.

Manufacturers claim that VAs only record and send data to the cloud when activated by specific trigger words like "Hey, Siri". However, numerous reports have highlighted the risks of VAs being flawed for eavesdropping. In 2018, an Amazon Echo user found that Alexa recorded a private conversation and sent it to his colleague [3]. A report [4] revealed that Amazon staff regularly listen to Alexa recordings for improvements, posing significant privacy risks. Likewise, Iqbal et al. [5] concluded that Amazon was collecting and sharing voice interactions from Echo speakers with third parties to deduce user preferences, further exacerbating privacy concerns. Moreover, smart speakers like Echo enable remote calls between user-owned devices without recipients explicitly accepting, leading to potential eavesdropping by anyone with access to the user's account [6]. Furthermore, a documented hardware defect opened a security vulnerability in the VA, enabling eavesdropping through unintended recordings [7].

While anti-eavesdropping technologies for VAs are continually evolving, they still face constraints. A simple mitigation involves using a physical mute button to control the VA's activation, but this undermines the benefit of hands-free convenience. Some manufacturers have developed sound masking systems (*e.g.*, VoiceArrest [8]) that inject audible white noise with specially designed frequencies and volumes to mask speech for voice privacy. However, it requires users to tolerate high-energy, disruptive ambient noise. Moreover, recent studies [9], [10] have demonstrated that an attacker can retrieve the original speech from white noise-masked recordings using speech enhancement techniques. Electromagnetic interference (EMI) presents a potential anti-eavesdropping method, but it faces several limitations, including high deployment costs, limited concealability, and a dependence on prior knowledge of the target device [11], [12]. Ultrasonic jamming [13], [14], [15], [16], [17], [18] uses carefully designed ultrasound as carriers to modulate low-frequency audible noise into the ultrasonic frequency range before emission. Due to the intrinsic nonlinearity of commercial microphones, noise in the ultrasonic spectrum can be demodulated into the audible frequency range by

Received 12 January 2025; revised 10 June 2025 and 17 July 2025; accepted 17 July 2025. Date of publication 24 July 2025; date of current version 31 July 2025. The work of Man Zhou was supported by NSFC under Grant 62202180. The work of Xiaojing Ma was supported in part by the Major Research Plan of Hubei Province under Grant 2023BAA027 and in part by the Key Research and Development Plan of Hubei Province under Grant 2024BAB049. The associate editor coordinating the review of this article and approving it for publication was Dr. Yansong Gao. *(Xiaoxiao Qiao and Man Zhou contributed equally to this work.) (Corresponding author: Man Zhou.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Natural Sciences Ethics Committee of Wuhan University under Application No. WHU-NS-IRB2024003.

Xiaoxiao Qiao, Xiaojing Zhu, and Houzhen Wang are with the School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: qiaoxx@whu.edu.cn; hsiaoching_chu@whu.edu.cn; whz@whu.edu.cn).

Man Zhou, Hongwei Li, and Xiaojing Ma are with Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: zhouman@hust.edu.cn; hongweili@hust.edu.cn; lindahust@hust.edu.cn).

Zhihao Yao is with the Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: zhihao.yao@njit.edu).

Digital Object Identifier 10.1109/TIFS.2025.3592558

microphones (*i.e.*, generating the "projected" signals), thereby obfuscating unauthorized recordings. Although effective and user-friendly, ultrasonic methods have limited practicality due to requiring dedicated ultrasonic transmitters or arrays.

In this work, we are the first to explore the use of near-ultrasonic signals emitted by COTS devices for anti-eavesdropping protection, which eliminates the need for dedicated ultrasonic transmitters and exhibits greater practicality and generalizability in real-world deployments. To achieve effective and robust jamming, three key challenges must be addressed: (1) Most commercial devices incorporate various noise suppression mechanisms, making it essential to design a robust and reliable jamming scheme. While energy masking schemes [13], [15] provide strong jamming, they are susceptible to modern denoising algorithms. Phoneme-level information masking [17] is more resilient to denoising but suffers from constrained corpus and poor cross-linguistic generalizability. (2) Unlike traditional ultrasonic noise injection, near-ultrasonic noise injection relies on the limited inaudible bandwidth of commercial devices. Moreover, due to manufacturing constraints, the speakers and microphones of commercial devices typically exhibit different and non-flat frequency responses within the limited bandwidth. Consequently, it is challenging to ensure the inaudibility of the interference while simultaneously maximizing the intensity of the "projected" noise across a variety of devices. (3) To maintain seamless user–device interaction, NUSGuard must dynamically pause jamming during user activity and resume it after the interaction ends. Moreover, implementing such real-time control on commercial devices introduces multiple system-level challenges, such as constrained power availability, diverse VA operational modes, Bluetooth communication latency, and security risks from failed wake-ups or false detections.

To address these challenges, we have introduced the following innovative solutions: (1) We propose a novel mixed-noise scheme inspired by multi-speaker conversations, offering strong robustness, jamming effectiveness, and cross-linguistic generalizability. Specifically, it uses multi-level speech sequences to disrupt speech within the same frequency band, leveraging the natural bandwidth and variability of human speech. Furthermore, by analyzing human speech patterns and automatic speech recognition (ASR) systems, we design noise enhancement techniques to further enhance the continuity and complexity of the jamming noise. (2) We study the frequency response curves of multiple commercial devices, in conjunction with the energy-dense frequency band of human speech (300 $\sim$ 3400 Hz) [19]. It identifies an optimal near-ultrasonic carrier window that maximizes the "projected" noise across various devices while maintaining perceptual transparency. Additionally, we employ the upper sideband amplitude modulation (USB-AM) and carefully optimize the modulation index to balance jamming performance and hardware limitations. (3) We present a new lexical-level jammer control strategy including adaptive jamming activation, pausing, and resumption. Our pausing mechanism combines keyword spotting with feedback detection to achieve high accuracy and low false alarms, while capturing responses from the target device to mitigate safety risks in real-world deployments. The activation and resumption strategies account for jammers' power constraints and VA devices' operational modes, ensuring robust performance across deployment scenarios. Our main contributions in this work are summarized as follows.

- We propose NUSGuard, a practical and reliable anti-eavesdropping system using near-ultrasonic signals. To the best of our knowledge, this is the first work to eliminate the demand for specialized ultrasonic transmitters, allowing any COTS speaker device with wireless communication capabilities to function as a jammer.
- We design an effective mixed noise scheme and a novel lexical-level automatic jammer control strategy to effectively disrupt unauthorized recordings, while flexibly preserving normal voice interactions between users and smart VA devices.
- We implement a system prototype using contemporary commercial mobile devices and demonstrate through comprehensive experiments that NUSGuard achieves excellent performance in both jamming effectiveness and usability.

II. Related Work

A. Anti-Eavesdropping Protection

To protect voice privacy, researchers have developed methods to prevent unauthorized eavesdropping. Microphone jammers can be categorized into three primary types: electromagnetic interference (EMI) [20], audible noise-based jamming [8], and ultrasonic jamming [13], [14], [15], [16], [17], [18]. Recent studies [11], [12] have demonstrated that EMI can be exploited to inject audio signals by leveraging the nonlinearity of microphones. However, it typically requires high-gain directional antennas and carefully tuned coupling paths, limiting its practicality in real-world use. While effective, the audible noise-based jamming method [8] faces a significant challenge in determining an appropriate volume for audible noise. Furthermore, studies [9], [10] have demonstrated that the white noise commonly used in audible noise-based jamming is not robust against modern speech enhancement techniques. More recently, ultrasonic noise injection has emerged as a prominent research area. BackDoor [13] exploited the intrinsic nonlinearity of microphones to inject additional low-frequency aliasing signals, thereby degrading the quality of recordings. MicShield [15] protected speech privacy without handicapping the VAs based on white noise. Patronus [16] introduced the selective jamming mechanism, which allows authorized devices to recover the original speech from the jammed recordings. A phoneme-based noise masking method [17] shows stronger robustness against speech enhancement than typical high-energy noise techniques. However, its limited confusability and poor cross-linguistic generalizability, reduce its suitability for multilingual use or time-constrained registration scenarios. MicFrozen [18] canceled speech signals and introduced noises to disrupt the speech eavesdropped by the adversary. These solutions necessitate specialized ultrasonic jammers. In contrast, our

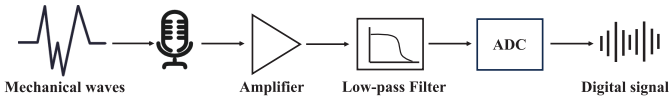


Fig. 1. The workflow of sound capture by the microphone and the nonlinear effect in the microphone.

system employs COTS speakers to achieve effective and robust interference against unauthorized recordings.

B. Near-Ultrasonic Applications

In previous studies, near-ultrasonic signals have been widely adopted in ranging and location [21], [22], [23], two-factor authentication [24], [25], [26], [27], unobtrusive communication [28], [29], [30], acoustic tracking [31], [32], [33], and inaudible attacks [34], [35], [36], [37]. In particular, acoustic sensing typically involves emitting near-ultrasonic signals from a speaker and analyzing the reflected echoes from sensed objects (*e.g.*, driver breath monitoring [38] and facial expression recognition [39]). In the field of inaudible attacks, NUIT [36] is the first to utilize near-ultrasonic signals to wage attacks against voice control systems. Notably, NUSGuard is the first study to apply near-ultrasonic signals for anti-eavesdropping protection.

III. Preliminaries

In this section, we describe the principles and simulation techniques used for the nonlinear effects in microphones.

A. Nonlinearity in Microphone

The process of sound capture by a mic is illustrated in Figure 1. Initially, the mic converts mechanical waves generated by air vibrations into electrical signals. These signals are then processed through an amplifier, a low-pass filter (LPF), and an analog-to-digital converter (ADC). Although amplifiers are designed for linear operation, their performance exhibits nonlinear characteristics when operating outside the audible frequency range. Previous studies [13], [40] have demonstrated that carefully designed ultrasound can generate "projected" signals within the audible frequency range by exploiting the nonlinear flaws in commercial mic amplifiers. This phenomenon enables the injection of inaudible sound into microphone's recordings. Assuming the input signal to the microphone is $S_{in}(t)$, the output can be modeled as:

$$S_{\text{out}}(t) = D_1 S_{\text{in}}(t) + D_2 S_{\text{in}}^2(t) + D_3 S_{\text{in}}^3(t) + \cdots, \quad (1)$$

where D_i represents the gain coefficient of the i-th order term. Typically, the gain coefficients of higher-order terms are sufficiently weak and can be disregarded. To inject inaudible signals into the mic, we use amplitude modulation to modulate a random audible signal $s(t)$ (300 $\sim$ 3400 Hz) onto a high-frequency carrier $\cos(2\pi f_c t)$, where $f_c \geq 17$ kHz. Since the gain of third-order and higher-order terms is negligible, the output signal of the amplifier can be represented as:

$$S_{\text{out}}(t) = D_1(1 + s(t))\cos(2\pi \cdot f_c t) + \frac{D_2}{2}(s^2(t) + 2s(t)) + \frac{D_2}{2}(1 + s^2(t) + 2s(t))\cos(2\pi \cdot 2f_c t) + \frac{D_2}{2}. \quad (2)$$

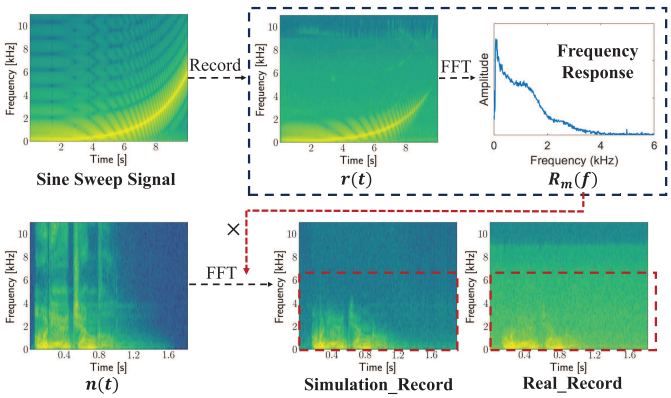


Fig. 2. Simulation of the nonlinear effects in a system using JBL speaker as the transmitter and the iphone 14 Plus as the receiver.

Note that S_{out} consists of four components: the first-order component, the audible component, the high-frequency component at $2f_c$, and the Direct Current (DC) component $D_2/2$. The linear term lies in the near-ultrasonic band (above 17 kHz) and is typically filtered by low-sampling-rate VA devices. The high-frequency and DC components are removed by the microphone's low-pass filter. The audible component, especially the $2s(t)$ term, passes through the filter and introduces the noise signal into the recordings.

B. Simulation of Nonlinear Effect

Due to manufacturing limitations, microphones often exhibit a non-flat frequency response, leading to some distortion in the demodulated audio signals. To accurately characterize a microphone's conversion of acoustic energy at different frequencies, we employ a nonlinear effect modeling technique [41]. This technique captures the microphone's frequency response to simulate its reception of sound signals across various frequencies in physical scenarios. Specifically, a sine sweep signal is used as the base signal. After carrier modulation and transmission, the microphone receives this signal. The received signal $r(t)$ undergoes a Fourier transform to obtain the frequency response $R_m(f)$. For any arbitrary signal $n(t)$, we can use this frequency response to simulate nonlinear distortion by $N(f) \cdot R_m(f)$, where $N(f)$ is the Fourier transform of $n(t)$, as illustrated in Figure 2.

IV. Threat Model

We assume that users operate the VA device in environments where private conversations may occur. The adversary's objective is to intercept and record these private conversations. We further assume that the adversary has access to recordings from smart home VA devices and is aware of the presence of the jammer. We assume that a capable adversary, upon obtaining the disrupted recordings, will employ denoising techniques to recover the original speech. We consider the following four typical denoising strategies:

Mix Source Separation (MSS). The adversary leverages the statistical independence or non-Gaussianity of speakers' voices (referred to as sources) to separate one or more sources from multiple simultaneously recorded mixed audio signals.

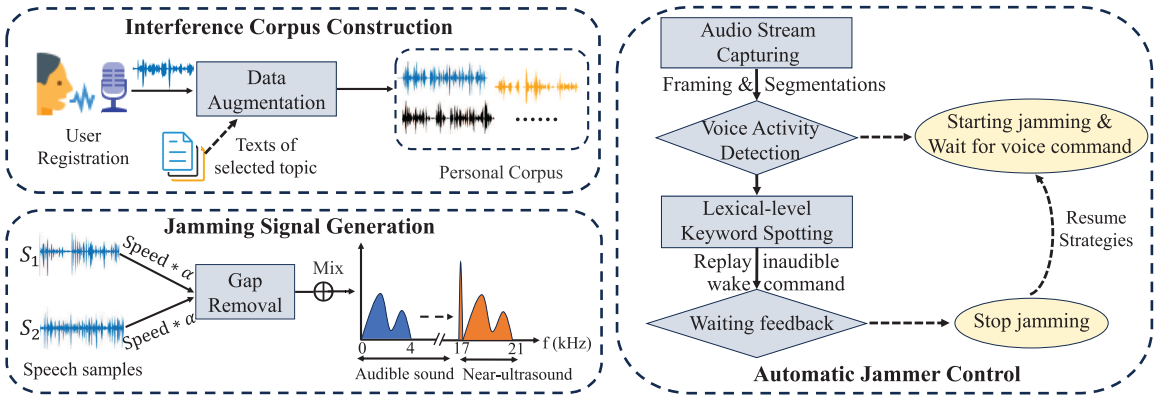


Fig. 3. Overview of NUSGuard.

This denoising method does not require prior knowledge of either the source signals or the mixing process.

Frequency Analysis. By applying spectrum analysis techniques (*e.g.*, the Short-Time Fourier Transform (STFT)), adversaries can precisely locate the specific frequency bands where speech components predominantly reside. Subsequently, they can use filters to effectively eliminate noise within those specific frequency bands.

Speech Separation. Due to the distortion caused by nonlinear effects in real-world scenarios, even noises generated from the target user's corpus may still be vulnerable to speech separation. This technique typically relies on deep learning models trained on large datasets of mixed and clean speech, enabling them to effectively learn and identify distinct patterns associated with different speakers. An adversary can exploit these models to isolate the target speaker's voice from other concurrent sounds in the recording, thereby extracting the original speech from the jammed recording.

Adaptive Noise Filtering. Adaptive filter denoising dynamically adjusts its parameters to accommodate changes in input signal characteristics. By inputting the jammed recording and reference noise, the adaptive filter continuously updates its weights to minimize the error signal, approximating the clean speech and allowing adversaries to remove noise from the disrupted recording.

V. System Design

A. Overview

Figure 3 illustrates three phases of NUSGuard: Interference Corpus Construction, Jamming Signal Generation, and Automatic Jammer Control. In our design, the user's mobile device, which generates the jamming noise, is designated as the "controller", and the Bluetooth speaker serves as the "jammer".

1) Interference Corpus Construction: To prevent an adversary from separating speech from noise by exploiting their distinguishing features, NUSGuard generates a jamming signal derived from the user's own speech corpus. During this phase, the user performs a one-time registration on the controller by reading predefined sentences. We further employ data augmentation to expand the user's corpus, thereby avoiding repetitive noise samples.

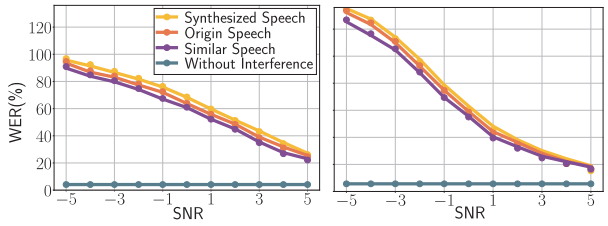
2) Jamming Signal Generation: After constructing the user's speech corpus, our system randomly selects multiple speech sequences from the corpus and applies noise enhancement strategies to generate mixed noises. The noises are then modulated into the near-ultrasonic frequency range to enable inaudible noise injection via commercial speakers.

3) Automatic Jammer Control: To effectively disrupt eavesdropping while preserving normal interactions with VA devices, we introduce an automatic jammer control strategy that incorporates adaptive jamming activation, pausing, and resumption. To minimize energy consumption, the controller activates jamming only when speech is detected. Subsequently, it pauses the jamming signal and activates the target VA upon detecting the user's wake command using our lexical-level jamming pausing strategy. Once the interaction concludes, jamming is resumed based on the resumption strategy.

B. Interference Corpus Construction

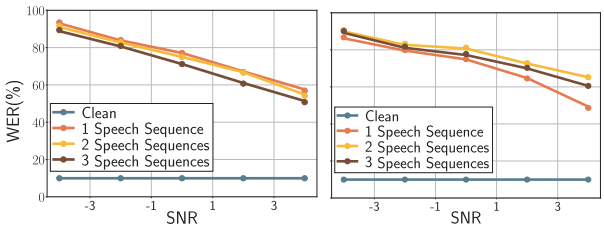
1) User Registration: To mitigate the risk of adversaries leveraging distinct voiceprint features between speech and noise for separation, NUSGuard generates noises using each user's personal corpus. During the registration process, users are prompted to record several speech samples ($n \geq 3$) with the Harvard Sentences dataset [42]. To enrich their corpus, users have the option to provide additional recordings at their convenience.

2) Data Augmentation: While NUSGuard generates jamming noises based on the user's corpus, the diversity of these noises is limited by the available volume of user speech data. A smaller corpus increases the risk of repetitive noise, potentially allowing adversaries to extract the original speech from distorted recordings. To mitigate this risk, we employ OpenVoice [43], an advanced speech synthesis technology, to expand the personal corpus. Users can select the theme of the synthetic text based on their specific scenarios, such as economics, culture, daily life, *etc*. Figure 4 compares the jamming effectiveness in three cases: interference noise generated from synthesized user speech data, interference noise based on the target user's original corpus, and interference noise derived from speech data with the highest similarity to the user's voice. It can be observed that synthesized speech-based noise and original corpus-based noise outperform those derived from the



(a) Wenet (b) Whisper

Fig. 4. Comparison of the jamming effectiveness between speech synthesis and the other two interference strategies.



(a) Before speech separation (b) After speech separation

Fig. 5. The impact of the number of mixed speeches on jamming effectiveness. We employed the SepFormer model, and trained three distinct models to target noise containing 1, 2, and 3 overlapping speech sources, respectively.

speech data with the highest similarity to the user's voice. This indicates that applying synthesized speech to expand the user's speech data does not diminish the interference effectiveness.

C. Jamming Signal Generation

After constructing the user's speech corpus, NUSGuard generates and modulates the jamming signal based on this personalized corpus.

1) Interference Noise Design: Recent studies have shown that modern ASR systems are typically trained using additive noise augmentation to improve robustness [44]. The training and evaluation inputs to such systems in the frequency domain can be modeled as:

$$|Y_{\text{tr}}| = |S_{\text{tr}} + N_{\text{tr}}|, \quad |\hat{S}_{\text{eval}}| = |S_{\text{eval}} + D_{\text{eval}}|. \tag{3}$$

Here, $|\cdot|$ denotes the signal magnitude. The subscripts "tr" and "eval" denote training and evaluation respectively; S_{tr} is the clean speech, N_{tr} is the additive noise used for training, and D_{eval} denotes the distortion in enhanced speech. When $D_{\text{eval}} \not\sim N_{\text{tr}}$, *i.e.*, the distortion differs from the training noise distribution, the input distribution between training and testing becomes mismatched, often leading to a significant degradation in recognition accuracy.

Building on this insight, we propose a set of noise enhancement strategies to enhance the jamming effectiveness. Specifically, based on the selected theme, we randomly sample speech segments from the user's corpus. We then apply a speed adjustment factor α to accelerate these speeches, making them sound more disruptive. Moreover, we remove gaps from the accelerated speeches to prevent prolonged silence periods that could allow eavesdroppers to capture speech content. The final jamming noise is generated by mixing these processed speeches. The techniques involved are as follows:

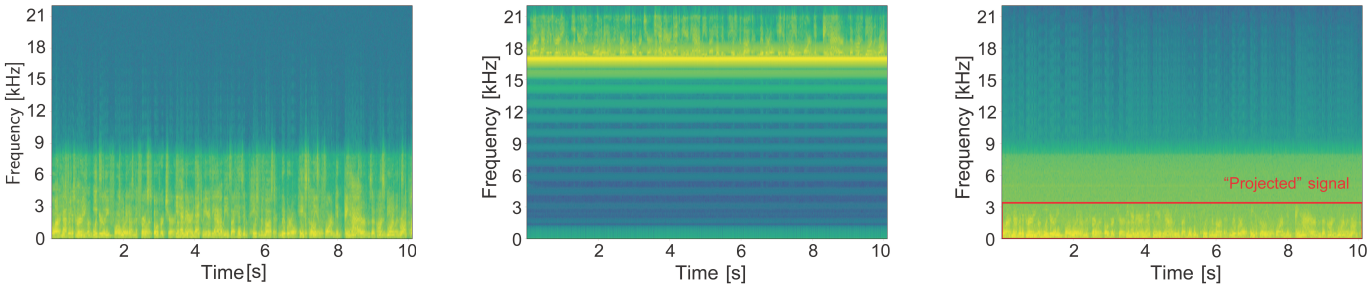
Noise Acceleration. Accelerating the interference noise packs more semantic information into a given timeframe, thereby enhancing jamming effectiveness. To maintain a balance between effectiveness and naturalness, we experimentally determine the optimal range [1.4, 1.8] for the random speed adjustment factor α and apply it to each speech component of the noise.

Gap Removal. Natural speech often contains frequent pauses, which could compromise jamming effectiveness by allowing clear speech to be captured. To address this, we eliminate segments where the signal energy falls below a predefined threshold, thereby ensuring that the generated noises are continuous and effective.

Speeches Mixing. As depicted in Figure 5, increasing the number of mixed speech can generate more complex noise, but can also dilute the energy allocated to each component, thus decreasing the overall disruption to speech recognition systems (SRS). In contrast, using too few mixed speeches may achieve stronger interference effectiveness, but renders the noise easier to comprehend and more vulnerable to speech separation models, potentially exposing the original speech. To achieve a balance between interference effectiveness and resilience against speech separation models, we set the number of mixed speech sources to two. Due to the broad frequency bandwidth and variability of human speech, our scheme effectively disrupts other co-occurring speech signals within the same frequency band. This increases the interpretation difficulty for SRS and enhances robustness against denoising methods.

2) Noise Modulation: We tested the frequency responses of two types of speakers (JBL Go3, and JBL Clip4) and three microphones (Apple, Samsung, and Xiaomi). Specifically, we played a single-frequency signal with progressively increasing frequency and analyzed the speaker's performance and the frequency distribution in the microphone recordings to determine the optimal frequency range. The results show that frequencies above 17 kHz become increasingly inaudible from the speakers. The tested microphones exhibited better frequency responses below 21 kHz. Consequently, we selected the optimal frequency range of 17 ~ 21 kHz.

Through amplitude modulation, we render the mixed noise inaudible to humans while ensuring it can still be effectively captured by adversarial microphones. As previously mentioned, the available inaudible bandwidth of commercial speakers ranges from 17 kHz to 21 kHz, amounting to just 4 kHz. This bandwidth is too narrow for the double-sideband amplitude modulation (DSB-AM) scheme. Additionally, due to the non-flat frequency response of microphones, the "projected" noise in recordings using lower-sideband amplitude modulation (LSB-AM) is too weak, leading to poor interference effectiveness. To address these issues, we therefore employ the upper-sideband amplitude modulation (USB-AM) method to modulate noise into the near-ultrasonic frequency range. Figure 6 shows the modulated noise and the corresponding recording using the USB-AM method. As observed, the energy of the recorded jamming noise within the energy-concentrated frequency band (300 ~ 3400 Hz) is sufficiently strong to effectively interfere with unauthorized recordings.



(a) Origin jamming noise (b) USB-AM modulated jamming noise (c) Recorded jamming noise

Fig. 6. The spectrogram of the original noise, the USB-AM modulated noise, and the recorded noise by a quad-microphone array.

Assuming our jamming noise is $n(t)$ and the carrier signal is $\cos(2\pi f_c t)$, where f_c = 17 kHz. Then the modulated noise S_{in} can be expressed as follows:

$$S_{\text{in}} = (1 + n(t))\cos(2\pi f_c t) - \hat{n}(t)\sin(2\pi f_c t). \tag{4}$$

Applying the nonlinear model of the microphone amplifier, and after processing through the LPF and removing the DC component, the output of the microphone is obtained:

$$\begin{aligned} S_{\text{out}} &= D_1 S_{\text{in}}(t) + D_2 S_{\text{in}}^2(t) \\ &\overset{\text{LPF}}{=} D_1(1 + n(t))\cos(2\pi f_c t) - D_1\hat{n}(t)\sin(2\pi f_c t) \\ &\quad + \frac{D_2}{2}(2n(t) + n^2(t) + \hat{n}^2(t)). \end{aligned} \tag{5}$$

From the above derivation, it is clear that our mixed noise $n(t)$ will remain in the unauthorized recordings of spy microphones, degrading the recording quality and effectively jamming attempts to capture clean speech.

D. Automatic Jammer Control

To maintain the hands-free convenience provided by VA devices, while also accounting for practical constraints such as energy consumption and diverse VA operating modes, we introduce a comprehensive automatic jammer control strategy. It involves three key components: Jamming Activation, Jamming Pausing, and Jamming Resumption.

1) Jamming Activation: Operating the jammer continuously places a substantial power burden on both the controller and the jammer. Additionally, injecting jamming signals when the user is not speaking could inadvertently trigger VA responses, degrading the user experience. To address these challenges, NUSGuard incorporates voice activity detection (VAD) method [45]. When detecting that the user is speaking, the controller (*e.g.*, smartphone) sends modulated near-ultrasonic noises to the jammer via Bluetooth. The jammer then injects noises into nearby spy microphones.

2) Jamming Pausing: While an always-on jammer can effectively prevent unauthorized eavesdropping when the user is speaking, it also hinders user interactions with smart VA devices. To mitigate this issue, we design the pausing strategy that incorporates lexical-level keyword spotting and a feedback-waiting mechanism. Notably, the attenuation of near-ultrasound in air limits the interference coverage. In our scenario, the controller (*e.g.*, smartphone) is typically close to the user and often lies at the edge of the jammer's effective interference radius or even outside it. Therefore, we use the controller to monitor for user's wake commands and subsequently control the jammer's playback.

Upon detecting a wake command, the controller instructs the jammer to pause interference and play a pre-modulated inaudible wake command to activate the target VA device. Through experimentation, we found that detecting and playing longer wake commands (*e.g.*, "Hey Siri") takes approximately 0.75 seconds, which does not significantly degrade the user experience. Following the playback of the inaudible wake command, the controller uses an MFCC-CQCC-GMM model to detect a device feedback signal within a predefined time window (approximately 5 seconds) to determine whether the target VA device has been activated. If feedback is detected, the jammer remains paused until the interaction ends; otherwise, interference playback resumes.

This feedback-waiting mechanism mitigates two critical edge cases that could otherwise compromise user privacy. First, when the user's wake command is correctly detected but the inaudible wake-up fails, the jammer would mistakenly pause, leaving the user's speech unprotected. Notably, even audible wake-up commands in everyday use can occasionally fail. Second, in cases involving false wake-command detection followed by a failed inaudible wake-up attempt, spontaneous user speech could be misinterpreted as VA interaction, causing a prolonged pause of interference and consequently exposing private content. The proposed mechanism effectively mitigates these risks by verifying actual device feedback before pausing interference.

3) Jamming Resumption: Once the VA returns to an inactive state, NUSGuard continues to monitor for voice activity to decide whether to resume interference. However, accurately determining when the VA switches to an inactive state is challenging due to its various operational modes. In this section, we consider the follow-up mode and the non-follow-up mode.

Many modern VA devices implement follow-up mode [46], where the VA remains active after an initial interaction, awaiting further commands without requiring the wake command. The VA will automatically transition to an inactive state if no further interaction occurs within a predefined period. For this mode, we simulate the VA's workflow to identify when it becomes inactive. Specifically, using VAD [45] technology and a preset silence period, we consider the VA inactive if no voice activity in this silence period. In non-follow-up mode, the VA responds to a single command and then immediately

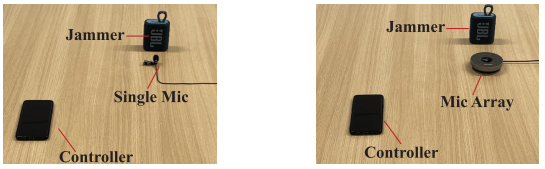


Fig. 7. The prototype of NUSGuard.

returns to the inactive state once the command execution is complete. Typically, the VA uses VAD and semantic content recognition to determine when a user command ends and begins responding. However, deploying a semantic content recognition model on the controller is time-consuming and computationally expensive, with the VA's response time being unpredictable. Hence, we resume monitoring voice activity once the VA begins to respond and resume jamming when voice activity is detected again. Specifically, we distinguish between the user and VA by comparing the Mel-frequency cepstral coefficients (MFCC) features of the captured signal frames to those of the target user's speech. This enables us to pinpoint the exact moment the VA begins to respond.

VI. Evaluation

A. Experimental Setup

1) Implementation: We implement a prototype of NUSGuard to evaluate its performance using existing hardware, as illustrated in Figure 7. NUSGuard mainly consists of two components: 1) a smartphone acting as the controller, and 2) a Bluetooth speaker serving as the jammer. Our experiments have tested three different Bluetooth speakers: JBL Go3, JBL Clip4, and Xiaomi Speaker. The evaluated eavesdroppers include a single-microphone recorder, a HIKVISION 4-microphone array [47], smartphones (iPhone 14 Plus and Xiaomi 12 Pro), and smart speakers (Google Home Mini and Amazon Echo Dot). All experiments were conducted in a public conference room. By default, we use the JBL Clip4 as the jammer. The near-ultrasound carrier frequency is set to 17 kHz.

Our implementation primarily focuses on the controller, which is responsible for constructing the interference corpus, generating and modulating the jamming noises, and performing automatic jammer control. As depicted in the pipeline in Figure 3, the controller generates the interference corpus and noises based on user selection. Then it manages the playback and pausing of the Bluetooth jammer through our jammer control strategy.

2) Dataset: The Harvard Sentences dataset [42] is used for user registration, while the LibriSpeech dataset [48] is utilized to create the test dataset for the jamming experiments. The Alexa dataset [49] is used to test the user's wake-up success rate. For the test dataset, we curate both clean speech and mixed noise segments for each user. Specifically, we prepared 60 clean speech samples, including 20 daily conversational samples, 20 sensitive-content samples, and 20 passphrase-related samples to ensure dataset diversity. For all these samples, we generated corresponding mixed noise samples based on our noise scheme. In subsequent experiments, we use this test dataset by default.

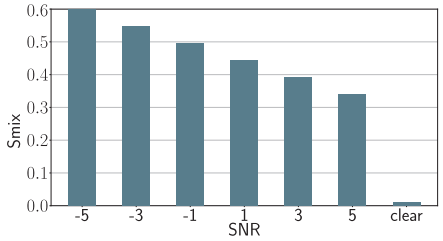


Fig. 8. Changes in intelligibility after jamming.

3) Evaluation Metrics: Our system measures the jamming effectiveness across three main aspects:

Signal-to-Noise Ratio. Signal-to-noise ratio (SNR) is the most commonly used metric to measure noise intensity, expressed as the ratio of signal power to noise power. We use it to evaluate the intensity of the jamming noise.

Speech Intelligibility. We assess speech intelligibility using a composite score derived from six objective metrics:

$$S_{\text{mix}} = 1 - \boldsymbol{\omega} \cdot [\text{PESQ}, \text{LLR}, \text{WSS}, \text{NCM}, \text{CSII}, \text{STOI}]^T, \quad (6)$$

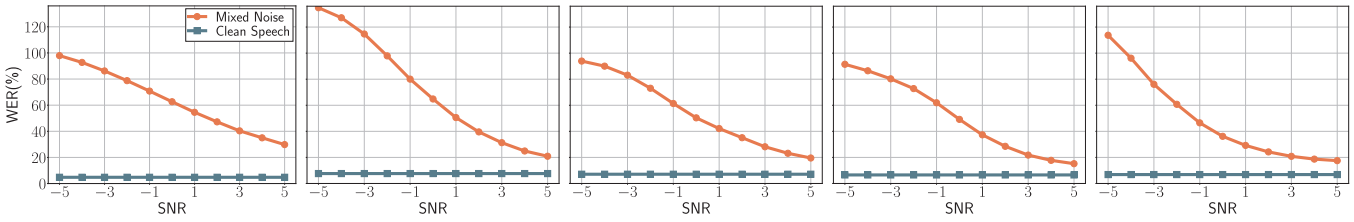
where $\omega = [0.18, 0.13, 0.06, 0.2, 0.21, 0.22]$ [10]. These metrics collectively characterize various aspects of perceived speech quality and intelligibility. Each metric is min-max normalized to the range [0, 1] before aggregation. For negatively correlated metrics (*e.g.*, LLR and WSS), a transformation of 1 – value is applied prior to weighting. Consequently, the final intelligibility score S_{mix} ranges from 0 to 1, where higher values correspond to lower intelligibility.

Word Error Rate. We evaluate the impact of our noise on the recognition performance of automatic speech recognition (ASR) systems using Word Error Rate (WER), which is calculated by comparing the reference transcription with the hypothesis transcription. We test the jamming effect by using two open-source ASR systems: Wenet [50] and Whisper [51], as well as three commercial ASR systems: Tencent ASR [52], Xunfei ASR [53] and Google STT [54].

B. Jamming Effectiveness

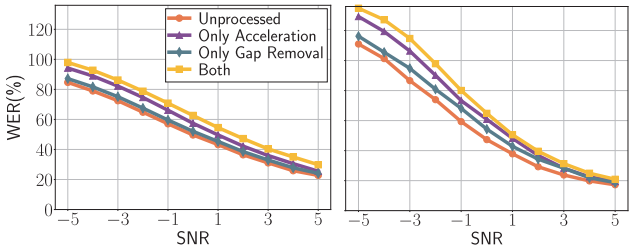
Modern smart devices are typically equipped with noise suppression technologies, such as those designed for white noise [17]. To fairly compare the jamming performance of different noise schemes, we apply nonlinear effect modeling (as discussed in Section III-B) to the modulated near-ultrasonic noises. The simulated noise closely resembles what a microphone would capture in real conditions. We then mix the simulated noise with clean speech in the digital domain to assess jamming effectiveness.

1) Overall Performance: We test the jamming performance of our noise across different SNRs and use the corresponding clean audio for comparison. Figure 8 shows the intelligibility results after interference. As the SNR decreases, the S_{mix} of the jammed recordings increases accordingly, indicating a corresponding degradation in speech intelligibility. Even at an SNR of 5, the S_{mix} reaches 0.34, a value substantially higher than that of clean speech. Given that eavesdroppers often use automatic speech recognition to extract users' private information, the accuracy of speech recognition becomes crucial. Therefore, we further test the average WER of different ASR



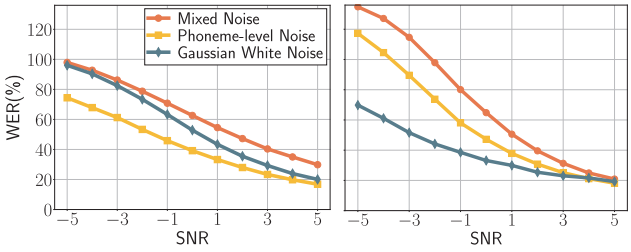
(a) Wenet (b) Whisper (c) Tencent ASR (d) Google STT (e) Xunfei ASR

Fig. 9. Recognition results of different ASR systems.



(a) Wenet (b) Whisper

Fig. 10. The impact of noise enhancement.



(a) Wenet (b) Whisper

Fig. 11. Results of different ASR systems under different noise schemes.

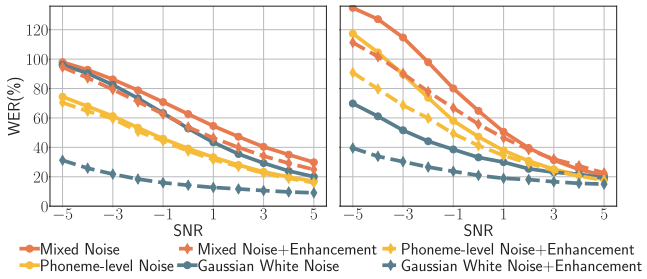


Fig. 12. Comparison of different noise schemes against speech enhancement under Wenet (left) and Whisper (right) ASR systems.

systems under jamming conditions. As shown in Figure 9, when the SNR $\leq$ -5, the WERs of different ASR systems are generally close to 100%. Even excessive insertion and substitution errors in Whisper and Xunfei ASR systems result in the WER exceeding 100%. As the power of the original speech gradually increases, the WER decreases. When the SNR $\geq$ 6, it becomes increasingly difficult to interfere with the spy microphones successfully.

2) Impact of Interference Noise Design: The noise design of NUSGuard consists of three modules: noise acceleration, gap removal, and speech mixing. We first investigate the impact of noise acceleration and gap removal on jamming performance. Figure 10 presents the recognition error rates of different ASR systems under four conditions: unprocessed, only noise acceleration, only gap removal, and both processed. At the higher SNR levels (SNR $\geq$ 4), the average WERs show minimal differences due to the relatively low noise energy. As the noise energy increases, the method combines noise acceleration and gap removal, demonstrates superior jamming capabilities. In the Whisper ASR system, the average WER for the combined method is 134.85%, compared to 111.01% for the unprocessed scheme. Similarly, within the Wenet system, the combined method results in a 13.25% higher error rate compared to the unprocessed scheme.

We then compare the jamming effectiveness of mixed noise with other noise schemes (*i.e.*, phoneme-level noise and Gaussian white noise) at the same power level. As shown in Figure 11, at an SNR of -5, the WER for mixed noise under the two ASR systems is 97.93% and 134.85%, respectively, phoneme-level noise achieves WERs of 74.41% and 117.33%, and Gaussian white noise results in WERs of 96.04% and 69.83%. Even as the SNR increases, the mixed noise still maintains the highest WER.

3) Robustness Against Speech Enhancement: Simplistic jamming noise is often susceptible to removal by speech enhancement algorithms. To evaluate NUSGuard's robustness against the speech enhancement method, we apply the FRCRN [55] to enhance jammed audio and then feed the enhanced signals into ASR systems. The WER results before and after speech enhancement for three schemes are shown in Figure 12. It can be observed that the effectiveness of all three noise schemes diminishes after speech enhancement, yet our noise scheme remains more effective than the others. In the Wenet ASR system, the WER of mixed noise and phoneme-level noise only decreased slightly. At an SNR of 5, the mixed noise still resulted in a WER of 24.92%, while the phoneme-level noise achieved a WER of 15.95%. In comparison, the white noise scheme exhibited the weakest robustness against speech enhancement. At SNR = -5, the WER of white noise dropped from 96.04% to 31.19%, and at SNR = 5, it dropped to only 9.06%, approaching the performance of clean audio.

C. Performance in Real-World

1) Performance of Different Jammers: In this experiment, we compare the impact of different jammers on the performance of NUSGuard in a real-world scenario. Specifically, we test the jamming effectiveness of three different smart speakers (JBL Clip4, JBL Go3, Xiaomi) at an SNR of 0. The results are presented in Figure 13. We find that the power difference of the jammers slightly affects the jamming performance of

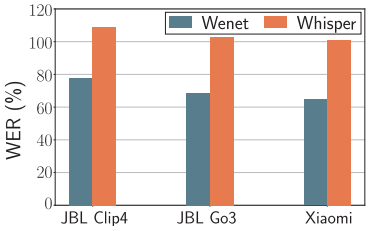


Fig. 13. Recognition results of different ASR systems.

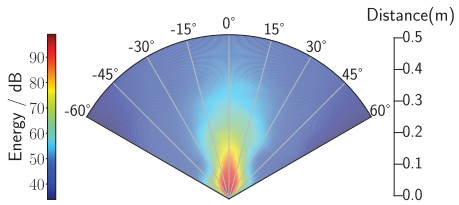


Fig. 14. Energy distribution of our noise using the JBL Clip4 as the jammer.

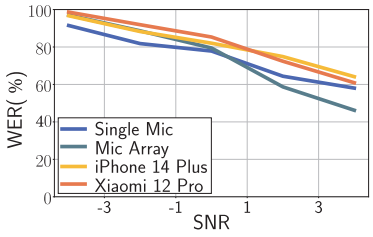


Fig. 15. Performance against different eavesdroppers.

TABLE I

THE AVERAGE WERS (%) / S_{MIX} BEFORE AND AFTER SPEECH ENHANCEMENT IN REAL-WORLD SCENARIOS

	SNR				
	[-5,-3]	[-3,-1]	[-1,1]	[1,3]	[3,5]
W/o enhancement	97.02 / 0.90	88.64 / 0.88	79.45 / 0.86	58.70 / 0.81	46.21 / 0.80
W/ enhancement	97.61 / 0.91	89.04 / 0.87	82.58 / 0.88	64.67 / 0.83	53.35 / 0.82

NUSGuard. Overall, however, all three jammers demonstrated satisfactory performance.

To further assess the interference coverage of NUSGuard in practical application, we measure the energy distribution generated by NUSGuard at the maximum volume. Specifically, we test the energy of the demodulated signal at 5 cm intervals across nine different angles, ranging from -60° to 60° in 15° increments, and smooth the collected data using interpolation. The energy distribution using the JBL Clip4 as the jammer is shown in Figure 14. The propagation characteristics of near-ultrasonic waves and the speaker arrangement of the jammer significantly impact the system's interference performance at different angles. At a distance of 30 cm directly in front of the jammer, the recorded noise energy is approximately 60 dB. Given that the volume of a normal conversation in a quiet environment is typically 60-70 dB, we recommend placing the jammer within 20 cm of a potential eavesdropping device to ensure effective interference.

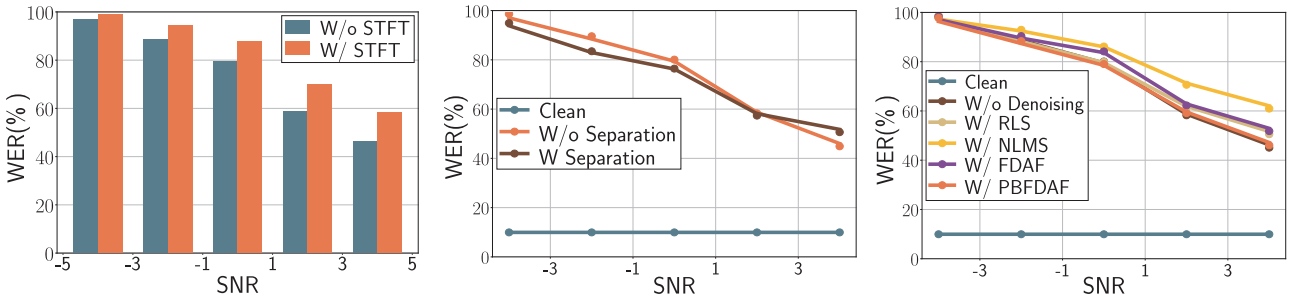
2) Performance Against Different Eavesdroppers: We evaluate the jamming performance of NUSGuard against various types of eavesdroppers, including devices with recording access (*e.g.*, a single-mic device, a 4-mic array, and smartphones) and those without recording access (*e.g.*, Google Home Mini and Amazon Echo Dot).

For devices with recording access (mic recorders and smartphones), we quantify jamming effectiveness by analyzing the average WERs from the recordings at different SNRs. Specifically, we randomly select 200 clean speech samples from our test dataset and play them through a speaker. The eavesdroppers were placed next to the jammer (JBL Clip 4), and we adjusted the eavesdropper position to obtain jammed recordings at different SNRs. Given the challenges of achieving a stable and accurate SNR in a real-world scenario, we calculate the average WER of Wenet within fixed SNR intervals to evaluate the actual jamming performance, as shown in Figure 15. At low SNRs ([-5,-3]), all devices exhibit high average WERs (above 89%). As the SNR increases, WERs gradually decrease across all devices. In real-world deployments, the jammer is typically placed closer to the eavesdroppers than the user, resulting in relatively low SNRs. The results also support this scenario. At SNRs within the range of [-1, 1], all devices achieve average WERs above 70%, with average WERs of 77.84% for the single-mic recorder, 79.45% for the mic array, 82.03% for the iPhone 14 Plus, and 85.32% for the Xiaomi 12 Pro. These results demonstrate that our jamming strategy can effectively preserve users' speech privacy under realistic deployment conditions.

For devices without recording access (Google Home Mini and Amazon Echo Dot), we first activated the device and then played the jamming noise while issuing follow-up commands (*e.g.*, "What's the weather today?" or "Tell me a joke."). By observing whether the commands were executed, we validated the jamming performance. We observed that during jamming, these devices consistently failed to respond. These results collectively demonstrate the robustness and generalizability of NUSGuard's jamming performance across a diverse set of eavesdroppers.

3) Robustness Against Speech Enhancement: Furthermore, we also evaluate the robustness of NUSGuard against speech enhancement in a real-world scenario. The experimental results are presented in Table I. In contrast to the results observed in the digital domain, the average WERs increase across all SNR ranges in real-world scenarios, and there is no significant improvement in speech intelligibility. We speculate this is due to the influence of variable factors such as environmental noise in the physical domain, which complicates the relationship between noise and clean speech. This complexity makes it more difficult for speech enhancement models to distinguish and filter out interference signals effectively. When attempting to improve signal quality, the speech enhancement model may inadvertently amplify or alter the interference, leading to an increase in recognition error rates.

4) Human Intelligibility of Eavesdropped Recordings: After obtaining IRB approval, we randomly recruited 15 volunteers aged between 22 and 32, with high English proficiency (IELTS score above 7.5), to assess the impact of NUSGuard on the human intelligibility of eavesdropped recordings. The test focused on two primary aspects: speech recognition correctness and speech quality rating. At each SNR level, we



(a) Frequency analysis (b) Speech separation (c) Adaptive filter denoising

Fig. 16. The average WERs of Wenet ASR system after applying different denoising methods.

TABLE II
THE AVERAGE WER (%) AND PERCEPTUAL RATINGS FROM VOLUNTEERS

	SNR					
	[-5,-3]	[-3,-1]	[-1,1]	[1,3]	[3,5]	clear
Average WER(%)	92.28	85.76	74.43	67.31	60.49	18.92
Score	1.51	2.06	2.55	2.72	3.08	4.66

TABLE III
THE AVERAGE WERS (%) BEFORE/AFTER MIX SOURCE SEPARATION

	SNR				
	[-5,-3]	[-3,-1]	[-1,1]	[1,3]	[3,5]
Without BSS	97.02	88.64	79.45	58.70	46.21
With BSS	97.52	88.88	79.77	58.87	46.23

randomly selected approximately a one-minute audio sample for the volunteers to recognize and rate, with the option to replay and pause. During the experiment, volunteers listened to the audio in a controlled quiet environment. For the recognition task, they transcribed the audio they heard, and we calculated the WERs by comparing their transcriptions to the original texts. Additionally, the volunteers rated the speech quality on a scale from 1 (poor) to 5 (excellent). The combined feedback from the volunteers, including both recognition results and quality ratings, is presented in Table II. The results indicate that it is also highly challenging to reconstruct speech content from jammed recordings via human auditory perception.

D. Security Analysis

In this section, we evaluate the security of NUSGuard against various denoising methods in real-world scenarios. We consider four denoising methods: MSS, frequency analysis, speech separation, and adaptive filter denoising. Given that modern smart VA devices are equipped with the microphone array, we use recordings captured by the microphone array in real-world scenarios as the test dataset.

1) MSS: Adversaries capture multi-channel recordings using multiple microphones and apply the fast Independent Component Analysis (ICA) algorithm to separate speech signals and noises in the time domain. As depicted in Table III, denoising using MSS does not result in significant improvement in recognition accuracy. Conversely, the average WERs increase across all SNR ranges. We attribute this is due to the wide bandwidth and rich variability of our mixed noise, which effectively disrupts other speech signals within the same frequency band. Moreover, the complex nonlinear effect in real-world scenarios likely disrupts the statistical independence of the sound sources, resulting in the poor performance of MSS.

2) Frequency Analysis: Adversaries may employ Short-Time Fourier Transform (STFT) to obtain the time-frequency spectrum of recordings and attempt to remove high-power noise. We experiment with various power thresholds to achieve optimal denoising. We find that only low power thresholds (0.05) can achieve relatively better denoising results. This is because our noises do not rely on the masking effect of high-power signals to suppress speech. The energy distribution of our noise closely matches that of the original speech, making it difficult to effectively distinguish between noise and speech using energy thresholds alone. The changes in recognition error rates further support this observation, as shown in Figure 16a. It is evident that after applying the frequency analysis, the average WERs increase by approximately 10% across all SNR intervals. At an SNR of -5, the WER even approached 100%. Thus, it is challenging for adversaries to remove the interference signal without discarding a substantial portion of the actual speech signal.

3) Speech Separation: To evaluate NUSGuard's resistance against speech separation in real-world scenarios, we employ the SOTA speech separation model [56], fine-tuned on the Librispeech dataset targeting noise containing two mixing speeches. The denoising results are shown in Figure 16b. At SNR < 2, the average WERs across various SNR levels slightly decrease after denoising, however, they still remain significantly higher than the error rate of clean speech. At higher SNRs, the average WERs increase after denoising. Results indicate that NUSGuard exhibits strong resistance to speech separation in practical applications.

4) Adaptive Filters Denoising: Adversaries exploit several common adaptive filters for denoising, including the normalized least mean square (NLMS) [57], recursive least square (RLS), frequency-domain adaptive filter (FDAF) [58], and partitioned block FDAF (PBFDAF) [59]. We use the pre-modulation noise as the reference signal, imposing high demands on the attacker's capabilities. For each interfered recording in the test dataset, we select the optimal parameter for each adaptive filter by parameter search to achieve the best possible denoising effect. Figure 16c compares the speech

TABLE IV
EVALUATION OF TWO JAMMER CONTROL STRATEGIES

	Accuracy (%)	Recall Rate (%)	FPR (%)
Lexical-level (Ours)	98.39	97.00	0.22
Phoneme-level	75.59	66.00	14.83

TABLE V
WAKE-UP SUCCESS RATES (%) ACROSS DIFFERENT DEVICES

	Xiaomi 12 Pro	iPhone 14 Plus	Google Home Mini	Amazon Echo Dot
Wake-up success rate	99.20	98.80	98.80	98.20

recognition results before and after applying these adaptive filters. Clearly, with optimal parameters, the average WERs of jammed recordings increase across all SNR ranges for the three filters, except for PBFDAF. And there is little improvement in the recognition results for the jammed recordings even using the best-performing PBFDAF. At SNR $\leq$ 0, applying PBFDAF slightly reduces the WER by about 1%. At SNR > 0, the average WERs after denoising are higher than before. We believe that this is due to the complexity of real-world acoustic environments and variable factors such as recording distance, which prevent simple models, such as adaptive filters, from effectively capturing the complex nonlinear characteristics of actual environments, resulting in poor denoising performance.

E. Usability Analysis

1) Evaluation of Jammer Control Strategy: We first evaluated the detection performance of our lexical-level jamming pausing strategy across multiple wake commands and compared it with the phoneme-level strategy. We constructed a test set for each wake command consisting of 200 target wake commands and 200 non-target utterances, including other wake commands and daily conversational speeches. During the experiment, the system was configured to detect a specific wake command (*e.g.*, "Hey Siri"). We then sequentially played the utterances from the test set and recorded the system's detection results for analysis. Notably, due to the lack of publicly available phoneme datasets beyond "Alexa", the phoneme-level experiments were conducted primarily using the Alexa wake command.

We assessed the performance of each strategy using three metrics: pause accuracy, recall rate, and false positive rate. The results, presented in Table IV, show that our lexical-level jamming pausing strategy achieved an accuracy of 98.39%, a recall rate of 97.00%, and a false positive rate of 0.22%, indicating highly reliable detection. We observed that in some test samples, the target wake commands themselves occasionally failed to activate actual VA devices, which slightly affected overall accuracy. In contrast, the phoneme-level strategy exhibited significantly lower performance, with an accuracy of 75.59%, a recall rate of 66.00%, and a considerably higher false positive rate of 14.83%. This degradation is primarily due to the lower accuracy of phoneme recognition and the over-fitting tendency introduced by fine-tuning on phonemes from the target wake command. These factors collectively lead to a higher false positive rate, which not only reduces the overall jamming effectiveness but also shortens the jammer's lifespan. Overall, the lexical-level strategy provides more robust protection for user voice privacy and causes minimal disruption to the user experience in real-world deployments compared to the phoneme-level approach.

Additionally, to assess the practical effectiveness of the proposed feedback-waiting mechanism, we conducted real-world experiments involving 500 human speech samples and 500 machine-generated utterances collected from various VA devices. The results show that the mechanism achieves an accuracy of 93.8%, a recall rate of 95.6%, and a precision rate of 92.28%, demonstrating its practical utility and reliability.

2) End-to-End Evaluation on Commercial VA Devices: To evaluate the practical usability of NUSGuard for devices equipped with voiceprint authentication, we tested whether normal user-device interactions could be maintained under our jamming control strategy (*i.e.*, user wake-up success rate). The evaluation involved various commercial VA devices, including smartphones (iPhone 14 Plus and Xiaomi 12 Pro) and smart speakers (Google Home Mini and Amazon Echo Dot). To evaluate the wake-up success rate of our strategy, we recorded 100 wake commands for each target device and confirmed that these commands could successfully activate the device without interference. We then modulated these wake commands onto the near-ultrasonic carrier and played them. Each experiment was repeated five times.

As shown in Table V, the average wake-up success rates for Xiaomi 12 Pro and iPhone 14 Plus were 99.2% and 98.8%, respectively. The Google Home Mini and Amazon Echo Dot achieved success rates of 98.8% and 98.2%. Based on the recordings from devices with recoding access, the demodulated inaudible wake commands exhibited only minimal distortion. This distortion has limited impact on wake-up success rates and did not hinder the overall interaction between users and their devices.

VII. DISCUSSION

A. Potential Risk of Noise Exposure

According to Equation 2, for microphones with a sampling rate greater than 42 kHz, near-ultrasonic noise (the linear term in the equation) will be retained in the recordings after passing through the microphones' filtering system, as shown in Figure 17. In our targeted threat model, the adversary is assumed to only have access to cloud-stored recordings from the VA device. Based on our investigation [60], [61], most commercial voice assistants employ a 16 kHz sampling rate to balance recognition accuracy and reduce data transmission and computational costs. Therefore, in our considered scenario, the potential risk of noise exposure due to retained near-ultrasonic linear components does not exist. Nonetheless, to assess the robustness of our jamming design against a more capable adversary with physical access, we further explored the impact of this potential risk.

Specifically, assuming the adversary is aware of this potential risk, they could deploy a spy microphone with a sampling

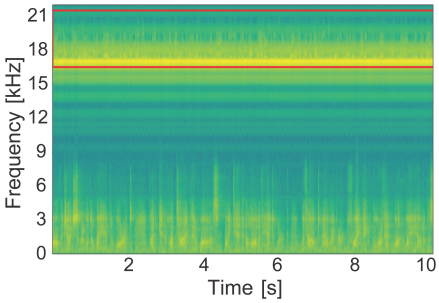


Fig. 17. A jammed recording received by microphone with a sampling rate of 48 kHz. The SNR of this recording is approximately 0. Notably, the frequency range of 17 kHz to 21 kHz contains the retained linear components.

TABLE VI
DENOISING RESULTS USING THE EXTRACTED LINEAR COMPONENT

	SNR				
	[-5,-3]	[-3,-1]	[-1,1]	[1,3]	[3,5]
Without Denoising	96.85	87.65	78.22	64.30	52.51
With Denoising	92.04	72.56	64.36	50.07	40.46

rate greater than 42 kHz in the target user's environment to capture noise information. To explore the impact of this potential risk on our system's interference performance, we attempt to extract the linear component from the recordings and use it as the reference signal, inputting it into the PBFDAF filter along with interfered recordings at various SNRs. The filter parameters are set to N = 256, M = 96, and mu = 0.01. The denoising results using the extracted linear component are shown in Table VI. Due to the correlation between the retained linear component in the jammed recordings and the "projected" noise in the audible frequency range, this denoising method proves more effective across all SNR ranges compared to the previous four denoising methods. Additionally, as the SNR increases, the reduction in the average WER after separation becomes more pronounced. For example, in the SNR range of -5 to -3, the average WER decreases by about 4%, while for SNR $\geq$ 0, the average WERs drop by nearly 10%. Nonetheless, the overall error rates remain high.

Furthermore, the adversary might employ deep learning techniques, inputting the reference noise and jammed recording into a model to denoise. However, the complexities of real-world acoustic environments and the requirement for extensive, fine-grained training data present significant challenges to an attacker.

B. Dilemma Between Jamming Effectiveness and Inaudibility

Ideally, due to the non-flat frequency response of commercial microphones, the lower the carrier frequency relatively, the stronger the noise components in the jammed recordings. A lower carrier frequency increases the likelihood that the modulated noise will occupy a frequency range to which the microphone is more responsive, resulting in a stronger demodulated noise component in the recordings. However, practical tests indicate that when the carrier frequency is set to 16 kHz and the near-ultrasonic noise is played at a volume level of 60, some volunteers could still perceive the presence of the near-ultrasonic signal and believe that it could affect their daily activities. When the carrier frequency is set to 17 kHz, we continuously adjust the jammer's volume and ask volunteers standing 0.5 meters away to provide feedback on the signal's audibility. At a maximum volume, no volunteers reported hearing an uncomfortable sound. Additionally, in the evaluation section, we have demonstrated the jamming effectiveness of this carrier frequency.

C. Integrated Jammer-Controller Design

A promising direction for simplifying the deployment of NUSGuard involves integrating the controller and jammer into a single unit, which would reduce system complexity and enhance portability. One feasible design is to place the controller on the back side of the speaker unit, outside the main jamming lobe, so that it remains less affected by the interference while still able to detect user wake commands.

To verify the practicality of this configuration, we conducted a preliminary experiment by physically attaching the controller behind the jammer. User speech and wake commands were played to test whether the system could operate as intended. The results show that the controller can still accurately detect wake commands and reliably control the pausing and resumption of the noise. Even when the user was positioned in front of the jammer and issued wake commands from a moderate distance, the system remained functional as long as the user moderately raised their voice. These findings suggest that an integrated jammer-controller design is technically feasible and merits consideration for future hardware implementations, particularly if commercial manufacturers provide support.

D. Future Work

Although NUSGuard achieves effective anti-eavesdropping protection using commercial off-the-shelf devices, there are still several aspects that deserve further exploration. First, the non-flat frequency response of commercial devices can cause noise distortion during injection, which reduces the masking effect of the noise on user speech. In future, we will further consider noise distortion caused by microphone nonlinearity when constructing the system model, and design corresponding compensation methods to mitigate its impact. In addition, NUSGuard injects jamming noise into all microphones within the effective interference range to prevent potential eavesdropping. However, this mechanism may also affect authorized devices. In the future, we plan to introduce a selective jamming mechanism. For example, authorized devices could leverage prior knowledge of the noise and active noise cancellation techniques to eliminate interference in real time. In future work, we will continue to improve our approach by advancing system modeling and enhancing its applicability in practical scenarios.

VIII. CONCLUSION

In this paper, we propose NUSGuard, a novel, practical, and reliable anti-eavesdropping protection system. To the best of our knowledge, NUSGuard is the first to leverage the built-in speakers of commercial devices to effectively

disrupt VAs' unauthorized recordings without specialized ultrasonic transmitters. We design a robust mixed noise scheme and lexical-level automatic jammer control strategy, ensuring strong interference while preserving normal voice interactions between users and smart VA devices. Extensive experiments demonstrate NUSGuard's superior performance in terms of jamming effectiveness, security, and usability, applicable to actual VA devices. Furthermore, NUSGuard demonstrates strong resistance to various denoising techniques, including MSS, frequency analysis, speech separation, and adaptive filtering, making it a cost-effective and secure solution suitable for integration into existing commercial smart devices.

REFERENCES

[1] T. Moynihan. (2016). *Alexa and Google Home Record What You Say. But What Happens to That Data*. [Online]. Available: https://www.wired.com/2016/12/alexa-and-google-record-your-voice

[2] Z. Wang, K. Liu, J. Hu, J. Ren, H. Guo, and W. Yuan, "AttrLeaks on the edge: Exploiting information leakage from privacy-preserving co-inference," *Chin. J. Electron.*, vol. 32, no. 1, pp. 1–12, Jan. 2023.

[3] D. Lee. (2018). *Amazon Alexa Heard and Sent Private Chat*. [Online]. Available: https://www.bbc.com/news/technology-44248122

[4] A. Hern. (2019). *Amazon Staff Listen to Customers' Alexa Recordings, Report Says*. [Online]. Available: https://www.theguardian.com/technology/2019/apr/11/amazon-staff-listen-to-customers-alexa-recordings-report-says

[5] U. Iqbal et al., "Tracking, profiling, and ad targeting in the Alexa echo smart speaker ecosystem," in *Proc. ACM IMC*, 2023, pp. 569–583.

[6] J. Moore. (2023). *Alexa, Who Else is Listening*. [Online]. Available: https://www.welivesecurity.com/2023/02/09/alexa-who-else-is-listening/

[7] M. Kunze. (2022). *Turning Google Smart Speakers Into Wiretaps for $100k*. [Online]. Available: https://downrightnifty.me/blog/2022/12/26/hacking-google-home.html

[8] MPS. (2024). *Voicearrestsound Masking System*. [Online]. Available: https://mpsacoustics.com/voicearrest-sound-masking-system

[9] P. Walker and N. Saxena, "Evaluating the effectiveness of protection jamming devices in mitigating smart speaker eavesdropping attacks using Gaussian white noise," in *Proc. Annu. Comput. Secur. Appl. Conf.*, Dec. 2021, pp. 414–424.

[10] Y. Chen et al., "Big brother is listening: An evaluation framework on ultrasonic microphone jammers," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2022, pp. 1119–1128.

[11] Y. Zhang and K. Rasmussen, "Detection of electromagnetic interference attacks on sensor systems," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 203–216.

[12] Z. Xu, R. Hua, J. Juang, S. Xia, J. Fan, and C. Hwang, "Inaudible attack on smart speakers with intentional electromagnetic interference," *IEEE Trans. Microw. Theory Techn.*, vol. 69, no. 5, pp. 2642–2650, May 2021.

[13] N. Roy, H. Hassanieh, and R. Roy Choudhury, "BackDoor: Making microphones hear inaudible sounds," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2017, pp. 2–14.

[14] Y. Chen et al., "Wearable microphone jamming," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–12.

[15] K. Sun, C. Chen, and X. Zhang, "'Alexa, stop spying on me!' speech privacy protection against voice assistants," in *Proc. ACM SenSys*, 2020, pp. 298–311.

[16] L. Li, M. Liu, Y. Yao, F. Dang, Z. Cao, and Y. Liu, "Patronus: Preventing unauthorized speech recordings with support for selective unscrambling," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, Nov. 2020, pp. 245–257.

[17] P. Huang et al., "InfoMasker: Preventing eavesdropping using phoneme-based noise," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2023.

[18] M. Gao, Y. Chen, Y. Liu, J. Xiong, J. Han, and K. Ren, "Cancelling speech signals for speech privacy protection against microphone eavesdropping," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2023, pp. 1–16.

[19] Wikipedia. (2024). *Voice Frequency*. [Online]. Available: https://en.wikipedia.org/wiki/Voice_frequency

[20] D. F. Kune et al., "Ghost talk: Mitigating EMI signal injection attacks against analog sensors," in *Proc. IEEE Symp. Secur. Privacy*, May 2013, pp. 145–159.

[21] C. Peng, G. Shen, and Y. Zhang, "BeepBeep: A high accuracy acoustic ranging system using COTS mobile devices," in *Proc. 5th ACM Conf. Embedded Netw. Sensor Syst.*, 2007, pp. 1–14.

[22] W. Mao, W. Sun, and M. Wang, "DeepRange: Acoustic ranging via deep learning," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–23, 2020.

[23] Y. Wang et al., "FaceOri: Tracking head position and orientation using ultrasonic ranging on earphones," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–12.

[24] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "EchoPrint: Two-factor authentication using acoustics and vision on smartphones," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 321–336.

[25] M. Zhou, Q. Wang, Q. Li, W. Zhou, J. Yang, and C. Shen, "Securing face liveness detection on mobile devices using unforgeable lip motion patterns," *IEEE Trans. Mobile Comput.*, vol. 23, no. 10, pp. 9772–9788, Oct. 2024.

[26] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, "EchoHand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 2931–2945.

[27] M. Zhou et al., "PressPIN: Enabling secure PIN authentication on mobile devices via structure-borne sounds," *IEEE Trans. Depend. Secure Comput.*, vol. 20, no. 2, pp. 1228–1242, Mar. 2023.

[28] S. Ka et al., "Near-ultrasound communication for TV's 2nd screen services," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2016, pp. 42–54.

[29] M. Zhou, Q. Wang, K. Ren, D. Koutsonikolas, L. Su, and Y. Chen, "Dolphin: Real-time hidden acoustic signal capture with smartphones," *IEEE Trans. Mobile Comput.*, vol. 18, no. 3, pp. 560–573, Mar. 2019.

[30] K. Qian, Y. Lu, and Z. Yang, "{*AIRCODE*}: Hidden {*screen-camera*} communication on an invisible and inaudible dual channel," in *Proc. USENIX NSDI.*, 2021, pp. 457–470.

[31] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 15–28.

[32] A. Wang and S. Gollakota, "MilliSonic: Pushing the limits of acoustic motion tracking," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–11.

[33] D. Li, J. Liu, S. I. Lee, and J. Xiong, "FM-track: Pushing the limits of contactless multi-target tracking using acoustic signals," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, Nov. 2020, pp. 150–163.

[34] 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 *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2015, pp. 142–154.

[35] M. Zhou et al., "Stealing your Android patterns via acoustic signals," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1656–1671, Apr. 2021.

[36] Q. Xia, Q. Chen, and S. Xu, "Near-ultrasound inaudible trojan (Nuit): Exploiting your speaker to attack your microphone," in *Proc. USENIX Secur.*, 2023, pp. 4589–4606.

[37] M. Zhou et al., "PrintListener: Uncovering the vulnerability of fingerprint authentication via the finger friction sound," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2024.

[38] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "BreathListener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals," in *Proc. 17th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2019, pp. 54–66.

[39] G. Wang, Q. Yan, S. Patrarungrong, J. Wang, and H. Zeng, "FacER: Contrastive attention based expression recognition via smartphone earpiece speaker," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2023, pp. 1–10.

[40] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible voice commands," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 103–117.

[41] X. Li, C. Yan, X. Lu, Z. Zeng, X. Ji, and W. Xu, "Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2024.

[42] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Jan. 1969.

[43] Z. Qin, W. Zhao, X. Yu, and X. Sun, "OpenVoice: Versatile instant voice cloning," 2023, *arXiv:2312.01479*.

[44] P. Wang, K. Tan, and D. L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 39–48, 2020.

[45] Wiseman. (2021). *Py-Webrtcvad*. [Online]. Available: https://github.com/wiseman/py-webrtcvad
[46] Amazon. (2024). *What is Follow Up Mode*. [Online]. Available: https://www.amazon.com/gp/help/customer/display.html?nodeId=GX7EJ9WHEPYBV94J
[47] (2016). *Hikvision Omnidirectional Conference Microphone VM1*. [Online]. Available: https://www.openchinacart.com/es/marketplace/1365313524505
[48] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[49] A. ANHARI. (2018). *Alexa Dataset*. [Online]. Available: https://www.kaggle.com/datasets/aanhari/alexa-dataset
[50] Z. Yao et al., "WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*, Aug. 2021, pp. 4054–4058.
[51] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.
[52] Tencent. (2024). *Tencent ASR*. [Online]. Available: https://cloud.tencent.com/product/asr
[53] IFLYTEK. (2024). *Xunfei ASR*. [Online]. Available: https://www.xfyun.cn/services/lfasr
[54] Google. (2024). *Google Speech-to-Text*. [Online]. Available: https://cloud.google.com/speech-to-text
[55] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9281–9285.
[56] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.
[57] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 8, no. 5, pp. 508–518, 2002.
[58] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
[59] J. Páez Borrallo and M. Garcia Otero, "On the implementation of a partitioned block frequency domain adaptive filter (PBFDAF) for long acoustic echo cancellation," *Signal Process.*, vol. 27, no. 3, pp. 301–315, Jun. 1992.
[60] G. A. SDK. (2025). *Google Assistant*. [Online]. Available: https://developers.google.com/assistant/sdk/reference/rpc/google.assistant.embedded.v1alpha1
[61] A. Interface. (2025). *Audioinput Interface*. [Online]. Available: https://alexa.github.io/alexa-auto-sdk/docs/explore/features/core/AudioInput/