Watchdog: Detecting Ultrasonic-Based Inaudible Voice Attacks to Smart Home Systems

Jian Mao, *Member, IEEE*, Shishi Zhu, Xuan Dai, Qixiao Lin, and Jianwei Liu

***Abstract*—Internet of Things is a critical infrastructure component as well as an enabling technology to support the fast-developing cross-region, cross-application, and diversified collaborative smart city services that require systematic cooperation among multiple smart city systems. Speech recognition-based voice controllable systems become one of the most popular interfaces in smart devices. However, it has been proved that attackers can hide their voice commands via modulating them on ultrasonic carriers and carry out inaudible voice attacks to manipulate voice controllable devices (e.g., mobile phone) unnoticeably. Although there are defense suggestions to enhance the hardware or add new modules of microphones, it is impractical to change the hardware design of all voice-controllable devices developed by different manufactures. In this article, we validate the effectiveness of ultrasonic-based inaudible voice attacks to voice-controllable smart home devices and propose a signal-processing-based hidden voice attack detection approach. Our approach uses an independent device that deploys a two-step lightweight detecting algorithm to identify the attack signals. We simulate our algorithm and make a prototype implementation of the proposed approach. The simulation results illustrate the correctness of the detection algorithm and the experiments show that our approach can detect the ultrasonic-based inaudible voice attack effectively.**

***Index Terms*—Inaudible audio attack, Internet of Things (IoT), signal-processing-based detection, smart home, smart speakers.**

I. Introduction

INTERNET of Things (IoT) has a wide range of applications due to its distributed feature [1]–[3]. A large number of sensors and actuators at a hardware layer in an IoT application can gather numerous data or respond to decision instructions [4], [5]. For example, voice controllable techniques are widely adopted in smart home devices. Speech recognition-based voice controllable systems become one of the most popular user interfaces, which greatly facilitate the interactions between users and smart devices [6]. Meanwhile, these voice-controlled interfaces also introduce vulnerabilities that may cause serious security problems [7]–[10]. Attackers tend to skillfully design attack signals to fool devices, but not to alert users. Besides, as there exist recognition gaps between humans and devices [11], Carlini *et al.* [12] extracted features of voice commands needed for the speech recognition algorithm, added some noise to make people confused and attacked voice controllable systems successfully. Also, previous work (e.g., Dolphin attack [13]) already proved that attackers may hide their voice commands by modulating them on ultrasonic carriers and carry out inaudible voice attacks to manipulate mobile devices unnoticeably. Such ultrasonic modulating-based inaudible voice attacks are different from the prior work on obfuscated voice attacks, which exploit the nonlinear property of microphones.

Since the Dolphin attack is based on microphones' hardware defect, it does not require any changes or contact to the target, nor is it limited to a specific device. As long as the microphone used by a voice controllable device is nonlinear, such an attack will be achieved on this device. Driven by the similarity between the microphones of smart home devices and those discussed in the Dolphin attack, we consider the possibility of implementing such an attack on smart home systems by analyzing the architecture of built-in microphones. Based on our experiment, such an inaudible attack is also effective in smart home systems, which is equipped with voice controllable devices. Moreover, in a smart system, smart devices may be controlled by other nodes according to the rules assigned by the permission system. Once a voice controllable smart device is compromised, the effect will be amplified due to the chain reaction.

There are some proposed defense solutions against the inaudible voice attack. First, it is easy to think of using speaker verification to reject the attacker's voice commands. But this kind of defense can be bypassed. The speaker recognition algorithms are not robust enough and conducting brute force attacks to a trained device is not difficult [13]. Also, as long as the voice segment of the victim is obtained, simple speech concatenative synthesis can splice the attack command of the victim's voice. On the other hand, different from personal electronics, there are always several family members involved in a smart home system. In such a scenario, training the voice assistant repeatedly will be very inconvenient for them.

Second, it is obvious that recovered voice commands from the attack and normal signal will possess different features, so researchers attempt to train classifiers to distinguish

Manuscript received February 29, 2020; revised April 26, 2020; accepted May 19, 2020. Date of publication May 26, 2020; date of current version September 15, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0802400, in part by the Beijing Natural Science Foundation under Grant 4202036, in part by the National Natural Science Foundation of China under Grant U11733115 and Grant 61871023, and in part by the Opening Project of Shanghai Key Laboratory of Integrated Administration Technologies for Information Security under Grant AGK2019001. *(Corresponding author: Jian Mao.)*

Jian Mao, Xuan Dai, Qixiao Lin, and Jianwei Liu are with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China (e-mail: maojian@buaa.edu.cn; daixuan@buaa.edu.cn; linqx529@buaa.edu.cn; liujianwei@buaa.edu.cn).

Shishi Zhu is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: zss@buaa.edu.cn).

Digital Object Identifier 10.1109/JIOT.2020.2997779

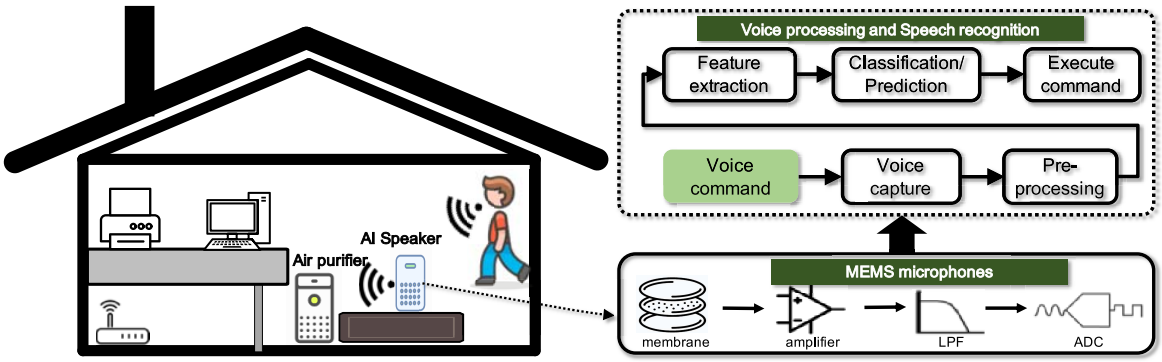

Fig. 1. Voice control system in smart home scenario.

them [12], [13]. However, the characteristics of the normal signal are similar to the attack one, which will raise false alarms. For example, in [13], they leveraged the difference between the recovered (demodulated) attack signal and the normal one in high frequency, ranging from 500 to 1000 Hz, to train a classifier. However, when a normal voice command is spoken by the user whose spectrum is similar to the recovered signal attack in 500–1000 Hz, it will cause false alarms. Furthermore, it requires internal changes to the device's speech recognition system and leads to delay since every command will take effect after prediction. Although there are defense suggestions in [13] to enhance the hardware or add new modules of microphones, it is impractical to change the hardware design of all voice controllable devices developed by different manufactures. Therefore, how to detect these inaudible voice attacks is still a challenge.

In this article, we validate the effectiveness of ultrasonic-based inaudible voice attacks on voice controllable smart home devices and propose a signal-processing-based approach of inaudible voice attack detection. Instead of modifying the hardware or software design of microphones proposed by prior work, our approach uses an independent device and deploys a two-step lightweight detecting algorithm to identify the attack signals. We simulate our algorithm and make a prototype implementation of the proposed approach. The simulation results illustrate the correctness of the detection algorithm and the real-world experiments show that our approach can detect the ultrasonic-based inaudible voice attack effectively.

In summary, we make the following contributions in this article.

1) We analyze the hardware structure of voice controllable smart home devices. Taking the microphone circuit of Mi AI speaker as an example, we find the similar nonlinearity exploited in the Dolphin attack. We validate the inaudible voice attack on smart home devices and analyze the attack.
2) We propose a lightweight detection scheme that uses signal processing to analyze the ambient ultrasound at home and detect the inaudible attack signal. Our defense solution needs no change in the internal structure, and increase no delay of the victim device.
3) We evaluate the feasibility of the detection method by simulation and hardware implementation. The evaluation result shows that after the detection process, the waveform of the attack ultrasound signal is different from the waveform of the normal one, proving the effectiveness of our detection approach.

Article Organization: The remainder of this article is organized as follows. Section II introduces the inaudible attack to smart home systems. Section III presents our signal-processing-based defense solution. Section IV presents the system simulation and result analysis. We implement and evaluate our approach in Section V. The closely related work and limitations are discussed in Sections VI and VII. We conclude this article in Section VIII.

II. Inaudible Attack to Smart Home Systems

In this section, we will analyze the nonlinearity of MEMS microphones and model the inaudible voice attack, including the generation of attack ultrasonic signals and the recovering phase of the ultrasonic signal by a nonlinear microphone.

A. Background

1) Analysis of Voice Controllable Systems: The process of the voice signal in a voice controllable system (VCS) is shown in Fig. 1. The speech recognition includes five procedures.

1) *Voice Capture:* The voice signal is first collected and converted from mechanical vibration to the analog electrical signal.
2) *Preprocessing:* The analog electrical signal is amplified, filtered, and converted to a digital signal for further processing.
3) *Feature Extraction:* The acoustic features necessary for speech recognition algorithms are extracted, such as mel-frequency cepstral coefficients (MFCCs), one of the most widely used acoustic features.
4) *Classification/Prediction:* The acoustic features are inputted to the speech classifier and predicted as the corresponding text command.
5) Finally, the command instructs the device to complete operations.

Nowadays, microelectrical–mechanical system (MEMS) [14], [15] microphones occupy almost all mainstream IoT devices. Almost all smart speakers use MEMS microphone array as audio input [16], including Amazon

Echo [17], [18], Google Home [19], Mi AI speaker [20], DingDong smart home speaker, and so on. Therefore, we principally discuss MEMS microphones fabricated in smart home devices here.

2) Nonlinearity of Microphones: As shown in Fig. 1, a typical digital MEMS microphone consists of a membrane, an amplifier, a low-pass filter (LPF), and an ADC. So the audio signal is processed by the following steps in a microphone. First, the audio signal goes through the diaphragm and is converted into the electrical signal. Then, the weak electrical signal cross is amplified by an audio amplifier. Next, the high-frequency components (e.g., ultrasound) of the signal is filtered out by a low-pass filter. Finally, the analog signal is converted into a digital signal through an ADC for the sake of subsequent processing.

Generally speaking, the sampling rate of ADC is from 44.1 to 48 kHz, and the cut-off frequency of the LPF is typically 20 kHz, which is enough to identify any frequency in the human audible range from 20 Hz to 20 kHz. Theoretically, it should eliminate signals greater than 20 kHz due to the presence of the LPF. But in practice, the working mechanism of the microphone makes it possible for the device to "hear" ultrasound. Specifically, the diaphragm and the amplifier of microphones have inherent nonlinearity, which will cause the spectrum of the input signal to move, so that the device can analyze the ultrasound [21]. The nonlinear model can be assumed as follows:

$$s_{\text{out}}(t) = As_{\text{in}}(t) + Bs_{\text{in}}^2(t). \tag{1}$$

Here, we ignore the higher order terms whose influence is too small compared to the second term.

B. Model of Inaudible Voice Attack

1) Attack Signal Generation: First, attackers should prepare voice commands (e.g., "open the door") as baseband signals. Since in the smart home scenario, the application of speaker recognition is not extensive and effective so that trained systems can be brute forced, so attackers can use the existing text-to-speech (TTS) system (e.g., Baidu TTS) to generate voice commands. Based on this voice command, attackers carefully design the attack signal. A valid attack signal must have two characteristics. On the one hand, it should be inaudible. On the other hand, nonlinear microphones can recover voice commands from the attack signal. So attackers have to utilize amplitude modulation to modulate the voice commands to the ultrasonic section.

2) Attack Via Nonlinearity: Due to the nonlinearity of the microphone, it can recover voice commands from the attack signal. We model the recovering phase as follows.

First, set the attack voice signal as $m(t)$, whose frequency ranges from 20 Hz to 20 kHz. Then, $m(t)$ is modulated by AM whose carrier frequency is fc. After that, the spectrum of the modulated signal is moved to the ultrasound segment, resulting in

$$s_{\text{mod}}(t) = m(t)\cos(2\pi f_c t) + \cos(2\pi f_c t). \tag{2}$$

Theoretically, the spectral components that appear after the voice passes through the nonlinear system should be as follows:

$$\begin{aligned} s_{\text{out}} &= s_{\text{mod}} + s_{\text{mod}}^2 \\ &= (1 + m(t))\cos\omega_c t + \cos^2\omega_c t + 2m(t)\cos^2\omega_c t \\ &\quad + m^2(t)\cos^2\omega_c t \\ &= (1 + m(t))\cos\omega_c t + \frac{1}{2}(\cos 2\omega_c t + 1) \\ &\quad + \frac{1}{2}m^2(t)(\cos 2\omega_c t + 1) + m(t)(\cos 2\omega_c t + 1). \end{aligned} \tag{3}$$

Then, s_{out} will pass through the low-pass filter, whose cut-off frequency is 20 kHz. After the LPF, the spectrum will remain the original audio signal $m(t)$ and spectral components corresponding to $m(t) * m(t)$. Despite the interference, the attack can still be achieved due to the fault tolerance of the speech recognition algorithm. We conduct simulations and demonstrate this in Sections IV and V.

C. Existing Inaudible Voice Attack

Smart home systems use voice control mechanisms to enable automation operations. However, such voice controllable devices also bring huge security and privacy risks. Recent attacks proposed by Zhang *et al.* [13] and Song [22] both exploited the nonlinearity of microphones to attack smartphone and popular speech recognition systems. This kind of attack successfully hides the malicious voice commands into the ultrasonic portion. Although inaudible to people, the attack ultrasound can create "shadows" in the speech frequency range due to the inherent nonlinearity of the microphones assembled on all voice controllable devices. Such inaudible attacks utilize hardware defects, so it is difficult to detect and defend.

In the smart home scenario, the attack signal of the ultrasonic segment can be generated and transmitted by ultrasonic washing machines, ultrasonic sensors, and so on. Similarly, the smart home devices equipped with the MEMS microphone are under the threat of the inaudible attack. We implement this attack and conduct attacks to the Mi AI speaker, as described in Section V. Based on the analysis of the attack signal, we propose our defense solution against the inaudible attack.

III. Signal-Processing-Based Defense Solution

A. Our Observation and Approach Overview

By observing and analyzing the spectrum of the attack ultrasound, we find a distinct characteristic that the ultrasound has the high center frequency, which is not common in smart home scenarios, except for a small number of electrical appliances, such as ultrasonic washing machines. Based on the observation, this kind of high central frequency can be used as a critical feature for detecting the ultrasonic-based inaudible voice attacks. Because normal ultrasound transmitted by some household applications (e.g., ultrasonic washing machine) may be captured, the high center frequency cannot act as the determinant but a suspicious feature for attack detection. Despite that, we can issue a low-level warning to notify users that the devices may under an inaudible attack. Further processes should be implemented to received ultrasound for attack detection. Comparing Fig. 2 with Fig. 3, there are differences in the

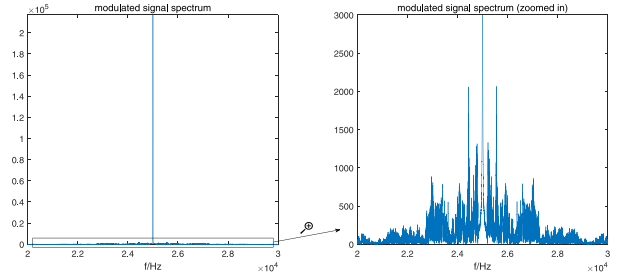

Fig. 2. Spectrum of the attack ultrasound whose center frequency falls at 25 k.

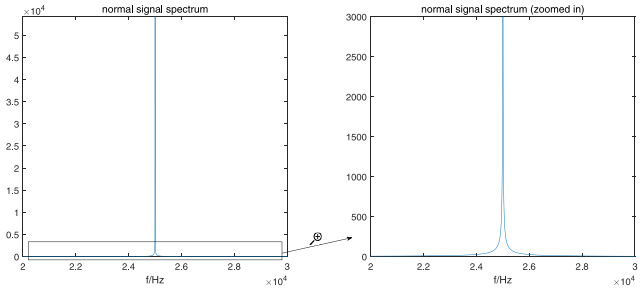

Fig. 3. Spectrum of the normal ultrasound whose center frequency falls at 25 k.

spectrum between two sides of the central frequency, which come from the voice command modulated in the ultrasound. Therefore, we can distinguish the attack ultrasound form the normal one based on this. We demodulate it with the central frequency and eventually obtain the baseband signal modulated in it. It is evident that the baseband signal will be a voice command when the received ultrasound is an attack. When we ascertain that a voice command exists, an advanced alarm (e.g., sending a message) should be issued to inform users of the ongoing inaudible attack.

B. Ultrasonic-Based Inaudible Voice Attack Detection

The defense solution consists of four phases.

1) *Phase 1 (Ultrasound Capture):* In this phase, we receive all environmental ultrasound, which may have modulated harmful voice commands.
2) *Phase 2 (Suspicious Frequency Detection):* The second phase is to determine whether the center frequency f_c of the received ultrasonic signal falls within the suspicious attack frequency range. If yes, our approach issues a low-level alert to notify the owner that devices may be under attack, and continues with the next phase; if not, our approach returns to Phase 1.
3) *Phase 3 (Demodulation and Capturing Baseband Signal):* In this phase, we demodulate the ultrasonic signal with f_c, and filters out the components that out of the speech frequency range to obtain the baseband signal.
4) *Phase 4 (Malicious Voice Detection):* In the last phase, our approach uses voice activity detection (VAD) [23], [24], which is one of the most commonly used voice detection methods. If the baseband signal is detected as a voice signal, the received ultrasonic signal will be considered as an attack signal and an advanced alert should be issued. It is also worth noting that VAD should be based on current ambient noise, so it is also necessary to collect ambient noise and calculate the noise threshold in real time.

The defense scheme we come up here aims to independently perform detection against inaudible attack without changing the internal structure or increasing the algorithm delay of the victim devices. The whole defense system illustrated in Fig. 4 consists of five modules: 1) an ultrasonic receiver; 2) a preprocessing module; 3) a noise threshold extraction module; 4) a detection module; and 5) an alarm module. We describe each individual module as follows.

1) Ultrasonic Receiver: The ultrasonic receiver, as the "ear" of the defense device, incessantly monitors the environmental ultrasound. When an ultrasonic wave is detected, the receiver begins to receive it until the end. Then subsequent processes will be performed. However, there are some limitations. If attackers issue an overlong attack ultrasonic signal, such as a 1-min attack which only contains a 4-s malicious voice command. The detection process will be blocked at the receiving part, causing an unbearable time delay. Hence, we determine a maximum time limit τs to force the process to advance. Specifically speaking, if the duration of the ultrasonic signal exceeds τs, the data of the first τs will be intercepted for subsequent processing first. Then the following data will be temporarily stored in a queue until the end of the ultrasonic signal or reaching the τs time limit again, which causes another interception and temporary storage. Another question is that if the attack ultrasound signals are short but dense ultrasound signals, whether it will cause the subsequent processing unable to keep pace with receiving. This situation may make the cache queue longer and longer until the memory is out. In practice, we find that the time taken for processing is proportional to the length of the received signal, meaning that the corresponding processing time for short data is also short. In Section V, it is also proved that the processing is always faster than receiving.

2) Preprocessing Module: The preprocessing module contains an anti-aliasing LPF, an amplifier, and an ADC. The relationship between the cut-off frequency f_d of the LPF and the sampling rate f_s of the ADC is

$$f_s \geq 2f_d. \tag{4}$$

After this module, the raw ultrasound is amplified and converted to digital signal $c(t)$ for the convenience of further process.

3) Noise Threshold Extraction Module: This module is the auxiliary of the detection module because the environmental noise plays an important role in a speaker's speech recognition process. We employ a microphone to collect ambient noise for calculating and updating the noise threshold T in real time. The calculation of T is divided into two parts: 1) initialization threshold T_0 and 2) real-time update of T.

1) *Initialization Threshold T_0:* For the convenience of processing, we divide the noise into frames, one frame duration is 20 ms [24]. We collect k frame noise and calculate each energy E_j separately

$$E_j = \frac{1}{s}\sum_{i=1}^{s} x_j^2(i) \tag{5}$$

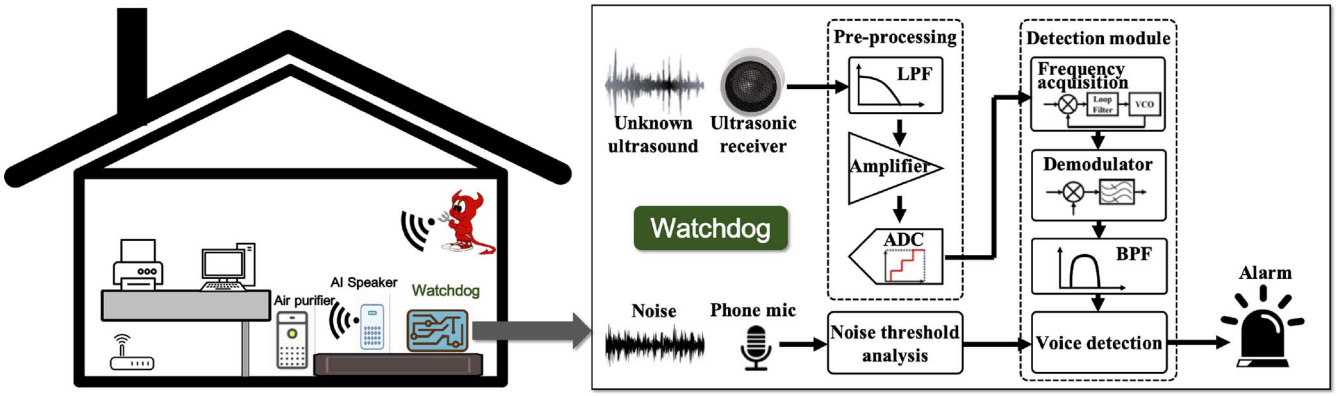

Fig. 4. Overall architecture of our approach.

Algorithm 1 Process of Defense Solution

Input: $c(t)$ - a preprocessed ultrasonic signal
Output: *flag* - the alarm flag

```
/* initialize */
flag ← False
/* get the center frequency of c(t) */
f_c ← FFT{c(t)}
/* if f_c is within the frequency range of an attack */
if f_c ∈ F_Arange then
    /* demodulate the c(t) by f_c */
    c'(t) = DEM_{f_c}{c(t)}
    issue a low-level warning
    /* pass a LPF */
    m(t) ← c'(t) ⊗ h(t),
        where |H(f)| = { 1, 300 < f < 3400Hz ; 0, else }
    /* detect m(t) is a voice command or not by VAD */
    flag_isVoice ← VAD{m(t)}
    if flag_isVoice == True then
        flag ← True
return flag
```

where $x_j(i)$ is the ith sample of the jth frame and s is the number of samples per frame. Then store the energy of k frames in a queue. T_0 is determined by average energy $\overline{E_0}$ of the k frames

$$T_0 = f\left(\overline{E_o}\right) \tag{6}$$

where

$$\overline{E_o} = \frac{1}{k}\sum_{j=1}^{k} E_j. \tag{7}$$

2) *Real-Time Update Strategy of T:* When a new frame is received, its energy is calculated and queued, meanwhile, discard the oldest frame. T is updated by the average energy of the k frame in the current queue every time the new frame is received.

4) Detection Module: The detection module contains four submodules: 1) a frequency analysis module; 2) a demodulator; 3) a band-pass filter (BLF); and 4) a voice detection module. We summarize the process in Algorithm 1.

First, we perform spectral analysis on the preprocessed signal $c(t)$ to obtain the central frequency f_c. Only when f_c is within the F_{Arange}, the frequency range of an attack signal, the following step will be implemented. We issue a low-level warning and demodulate $c(t)$ by f_c and issuing a low-level warning, that is

$$c'(t) = \text{DEM}\{c(t)\} = c(t) \cdot \cos(2\pi f_c t). \tag{8}$$

Then, we remove the frequency component out of normal human voice range (e.g., 300–3400 Hz) from $c'(t)$ by the BLF to get $m(t)$. If $c(t)$ is an attack signal, $m(t)$ will be a voice signal whose shape is similar to the original malicious voice command. On the contrary, $m(t)$ will be random noise if $c(t)$ is not an attack signal.

Here, we utilize short-time energy (STE)-based VAD to detect $m(t)$ as a voice command or not. The VAD workflow is as follows.

1) Downsample the frequency of $m(t)$ to 8 kHz, to satisfy Nyquist's theory.
2) Take 20 ms as one frame duration, so the number of samples per frame is 160.
3) Since the filters are not ideal filters, the mean of $m(t)$ will not strictly equal to 0, as shown in Figs. 9 and 10. When calculating the STE of each frame, we need to subtract each sample value from the mean of each frame
$$q_j'(i) = q_j(i) - \overline{q_j} \tag{9}$$
where $q_j(i)$ is the ith sample of the jth frame. $\overline{q_j}$ is the mean of the jth frame. $q_j'(i)$ is the result of normalized $q_j(i)$.
4) Calculating the energy Q_j of each frame
$$Q_j = \frac{1}{160}\sum_{i=1}^{160} q_j'^2(i). \tag{10}$$
5) Compare Q_j to T, the noise threshold of the noise extraction threshold module. Only when the baseband signal is detected as a voice signal under current environmental noise, the voice command is capable of activating or controlling the device. So, if $Q_j > T$, we can think of the jth frame as a speech.
6) Determine whether $m(t)$ contains an attack command based on p, the number of detected speech frames. When

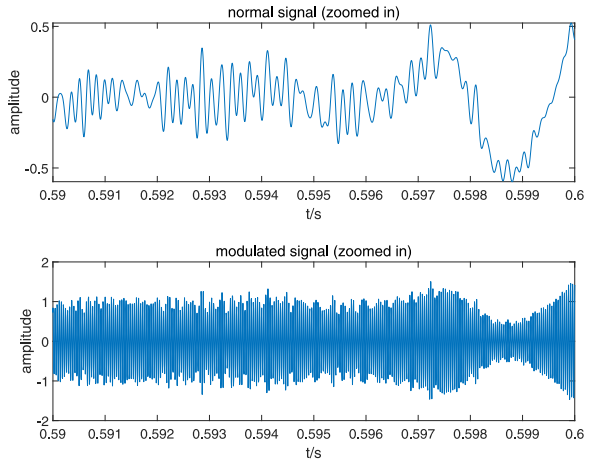

Fig. 5. Attack signal generation. From top to bottom are the zoomed-in normal voice signal and the zoomed-in attack ultrasonic signal.

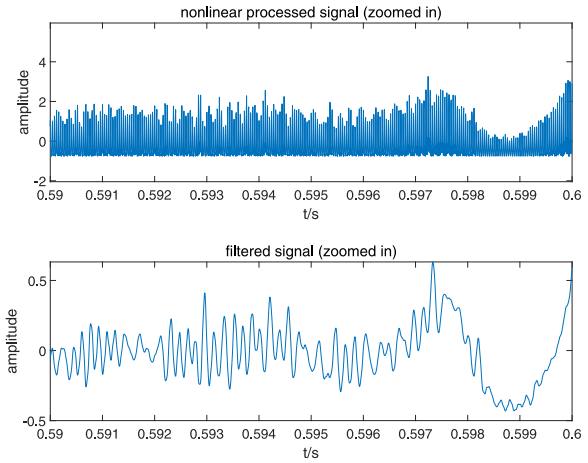

Fig. 6. Process of attack signal in a nonlinear microphone. From top to bottom are the zoomed-in signal after the modulated signal passes through the nonlinear module and the zoomed-in filtered signal of the above signal.

p is greater than a given threshold T_p, $m(t)$ can be considered as containing an attack command, which should trigger an advanced alert.

5) Alarm Module: This module is a feedback to users. If the flag is true, it can use beeping to warn the users at home, and furthermore send messages to users who are not at home.

Compared with the previous defense suggestion [13], our approach does not need to change the internal structure and algorithm of the victim device, and will not cause an extra delay of the speech recognition process. Also, it will not mistakenly detect normal voice commands as attack signals, because the ultrasonic receiver only receives ultrasonic signals. At the same time, the benign ultrasound transmitted by some household ultrasonic facilities will not cause false alarms because the baseband signal recovered from the suspicious ultrasound cannot be detected as a voice signal.

IV. System Simulation and Analysis

A. Simulation Analysis of Ultrasonic-Based Voice Attacks

As most smart home audio devices rarely have speaker recognition, such as the Mi AI speaker, anyone can control them only if the decibels are high enough. We use the audio commands generated by Baidu TTS as the normal signal (e.g., "Xiaoai Tongxue" and "open the light"). Then, the normal signal is AM-modulated to 25 kHz in MATLAB as the attack signal since the best attack frequency range is 20–50 kHz according to prior work [13], [22]. The zoomed-in time-domain waveform of them is shown in Fig. 5.

From the analysis in Section II, there is harmonic interference after an attack signal pass through the nonlinear microphone. To identify the effectiveness of the attack, we simulate the nonlinearity of the microphone circuit in MATLAB. The nonlinear model is formulated in (1). The modulated signal first passes through the nonlinear module, then the signal segments beyond the audible range (20 Hz–20 kHz) are filtered out. The zoomed-in time-domain waveform of them is shown in Fig. 6. We can find that the recovered signal is very similar to the original audio commands. When we play the recovered audio, we can clearly recognize the original command. This demonstrates that the inaudible attack can be achieved in theory.

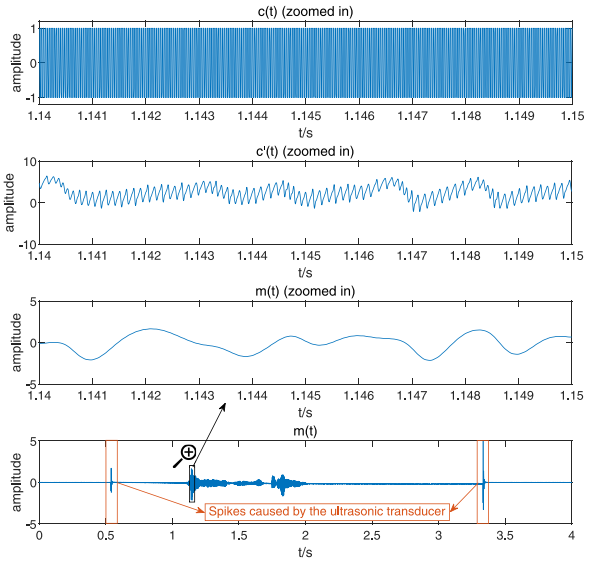

Fig. 7. Waveform of the attack signal during the process. From top to bottom are the zoomed-in received ultrasound $c(t)$, the zoomed-in demodulated signal $c'(t)$, and the zoomed-in and the whole filtered signal $m(t)$.

B. Simulation Analysis of Our Approach

We simulate the detection process in the laptop which is attached to an ultrasonic receiver [25]. The attack signal is AM modulated to 25 kHz from the normal signal "Xiaoai Tongxue," as shown in Fig. 5. We set the recording sampling rate of the laptop to 192 kHz and begin to receive ambient ultrasound. When a suspicious ultrasonic signal is received, it will be preprocessed to digital signal $c(t)$. We perform the fast Fourier transform (FFT) on $c(t)$ to get the center frequency f_c. We found that f_c is equal to 24 999 Hz which falls within the best attack frequency range, so we continue the following steps. $c(t)$ is demodulated by f_c to obtain $c'(t)$. Then we filter out the components beyond the speech range (typically 300–3400 Hz) to obtain baseband signal $m(t)$. The zoomed-in waveform of the above signal is presented in Fig. 7. Also,

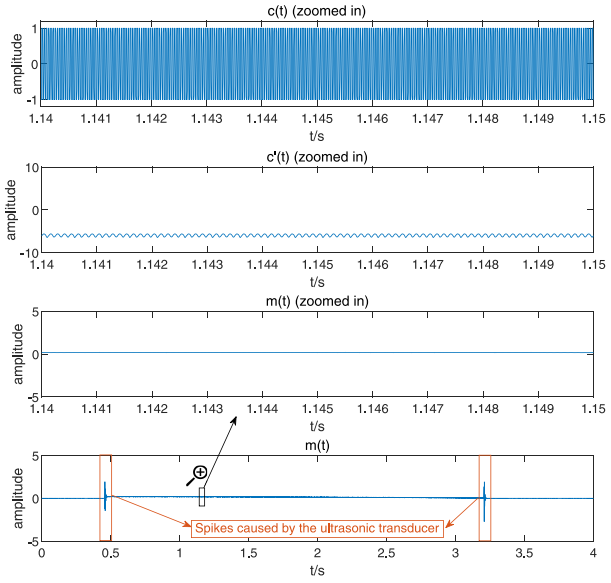

Fig. 8. Waveform of the single-frequency signal during the process. From top to bottom are the zoomed-in received ultrasound $c(t)$, the zoomed-in demodulated signal $c'(t)$, and the zoomed-in and the whole filtered signal $m(t)$.

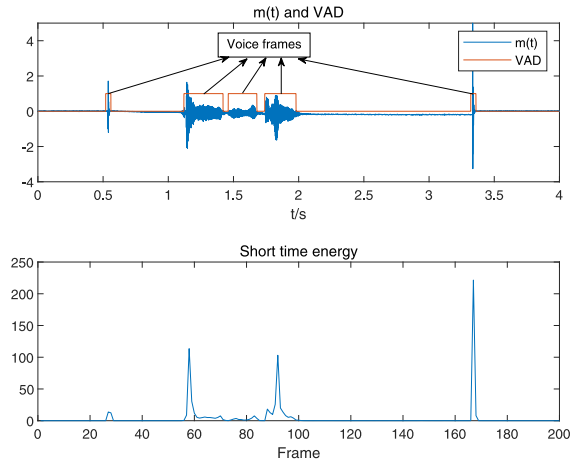

Fig. 9. STE of attack signal and frames detected as voice ones.

we generate the single-frequency signals at 25 kHz as a contrast. Its zoomed-in waveform is presented in Fig. 8. From both of them, we noted that there are two spikes in the front and back of the recovered baseband signal $m(t)$, as framed in Fig. 7, which cannot be filtered out by filters. Based on the mechanism of ultrasonic transducers [26], it is due to the unsteadiness of the ultrasonic transducer we used when it begins or stops to vibrate. The spikes will influence the determination of T_p because they will be detected as voice frames under the mechanism of STE-based VAD.

The last step to identify the attack signal is to determine whether $m(t)$ is a voice signal or not. We implement STE-based VAD here. One frame duration is 20 ms. With the current environmental noise of the laboratory, we determine that $T = 1$. For instance, we amplify $m(t)$ ten times and show the STE of $m(t)$ (recovered signals of both attack signal and single-frequency signal) and highlight the frames detected as voice frame in Figs. 9 and 10. The spikes we describe

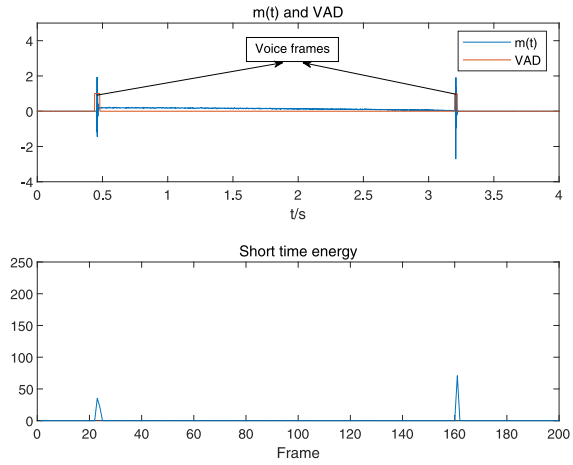

Fig. 10. STE of single-frequency signal and frames detected as voice ones.

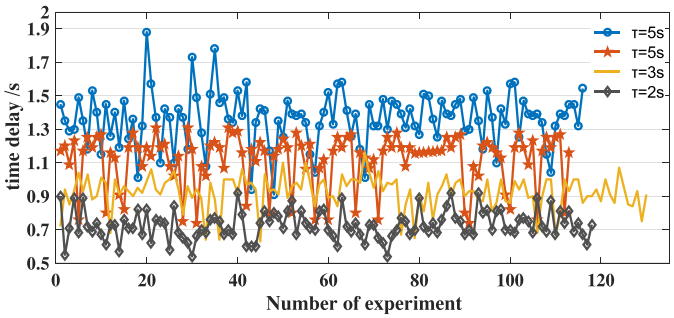

Fig. 11. Time delay under different time limits after hundreds experiments.

TABLE I
AVERAGE TIME DELAY UNDER DIFFERENT TIME LIMITS

τ/s	5	4	3	2
Average time delay $(\overline{t_d})$/s	1.362	1.117	0.924	0.728

above cannot be filtered out and their amplitude is always big enough to be detected as speech. In Fig. 9, we observe that the STE of voice components and spikes are greater than T. In Fig. 10, there are no voice components except for the spikes. The threshold T_p is based on multiple experiments on $m(t)$ of both attack signals and inoffensive signals. Specifically speaking, we consider ultrasound as an attack signal if more than ten frames are detected as voice.

One of the performance metrics that we considered is the delay of detection. We define the delay t_d as the time interval from the end of the reception to the completion of attack detection. Since in theory, the length of the processing is proportional to the length of the received signal, the delay will be proportional to the length of the received signal, too. To better understand the delay, we launch the attack signals repeatedly and set a maximum time limit τ to force the process to advance. The processing time delay obtained at this time is the time delay t_d required to process τ seconds of data. The result is illustrated in Fig. 11 and the average time delay under different time limits are listed in Table I.

From Fig. 11 and Table I, we can conclude the following.

1) The time delay due to the processing is proportional to the length of received signals and the time costs of the

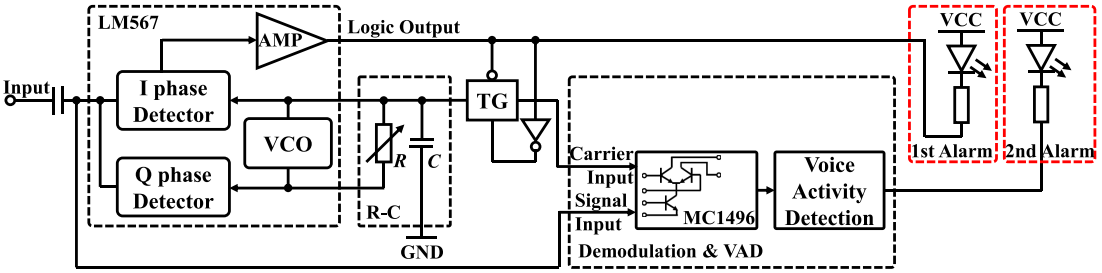

Fig. 12. Implementation of our defense solution: watchdog.

process are always shorter than that of receiving. So, if an attacker continuously transmits short but dense ultrasound signals, the subsequent processing can keep pace with the receiving. Also, the detection device is robust under an overlong attack ultrasound launched intentionally by an attacker due to the maximum time limit τ we introduce to force the process to advance. Since the process is faster than receiving, there is only one cache waiting to be processed, which will not cause a block of detection.

2) Once we determine the maximum time limit of τ, we can analyze the maximum detection time interval t_m from the smart home devices being attacked to the detection device completing the attack detection. Such a worst situation will happen when the attacker modulates the malicious command in the first t_0s of the attack signal and makes the attack signal overlong to reach the maximum time limit τ. So, the attack is realized first but the detection is finished $(\tau - t_0)$ s later (here, we do not consider the response time of the smart speaker). For instance, if $\tau = 5$ s, based on result listed in Table I, the maximum detection time interval t_m is

$$t_m = \tau - t_0 + \overline{t_d} = 5 \text{ s} - t_0 + 1.362 \text{ s} < 6.362 \text{ s} \quad (11)$$

where $\overline{t_d}$ is the average time delay under τ. Such time delay is tolerable. It is enough to aware users that smart home devices are under attack and make them take measures in time.

V. Implementation and Evaluation

A. Implementation of Ultrasonic-Based Attack to Smart Devices

We perform real-world experiments to evaluate the generated inaudible voice commands. In our experiment, the Mi AI speaker acts as the main victim device, and a laptop attached to a narrow-band ultrasonic transducer [25] serves as the attack device. In general, laptops can only transmit and receive signals within the audible sound range. Since the sampling rate of their sound cards is often 44.1 or 48 kHz, the prefilter will filter out the frequencies beyond the audible sound range. To transmit ultrasound, we choose the Lenovo IdeaPad S300 whose sound card's sampling rate can reach 192 kHz after setting. Here, we AM-modulate the normal signal to 25 kHz. All of the implements are performed in our laboratory with the average background noise about 57 dB.

First, we choose the activation command "Xiaoai Tongxue" as attack audio and validate the effectiveness of the attack. The Mi AI speaker is activated at a max distance of about 105 cm. Attackers can achieve more distant attacks with a higher power of attack signals or use more professional attack devices.

Then, we choose the recognition command "Sing a song" as attack audio and validate the effectiveness of the attack. The Mi AI speaker is activated at a max distance of about 126 cm.

Finally, we verify the control of household appliances. A light and an air purifier are bound to the Mi AI speaker. When the attack command is "Turn on the light/air purifier," the light and the air purifier are successfully turned on. It represents that applications connected to the smart speaker can be easily controlled by the attacker, so the security of our smart home is under threat.

Practice has confirmed that although the recovered signal has interference, it can still be accepted by the device to cause waking up and further operation.

B. Implementation of Our Approach: Watchdog

To better protect the security of smart home systems, we design an independent device to practice our defense solution. we use the LM567 tone decoder and the MC1496 multiplier to verify the proposed signal-processing-based defense solution. The simplified diagram of hardware implementation is as shown in Fig. 12.

In the framework, different dotted boxes represent different modules, including a phase-locked loop designed by the LM567, a resistor–capacitance circuit (the RC component), a coherent demodulation module with the VAD function, and two alarm modules. In particular, received signals and an external frequency are input into the LM567. The received signals may be the unknown ultrasound and the external frequency $\tilde{f_c}$ is generated by the RC component, i.e., $\tilde{f_c} = (1.1/RC)$. Note that the external frequency can be adjusted between 0.1 Hz and 500 kHz, and the detection bandwidth is set within $\pm 14\%$ of the external frequency.

In the LM567, a voltage-controlled oscillator (VCO) drives an in-phase (I phase) and quadrature-phase (Q phase) detector to determine the central frequency of the input signal f_c. It is worth mentioning that we use a phase-locked loop on the independent detection device, rather than the spectral analysis in simulation, to obtain the carrier frequency. If the input signal frequency matches the external frequency within the detection bandwidth, the LM567 provides a transistor

TABLE II
ATTACKS BASED ON AUDIO AND DEFENSE SOLUTION

Scheme	Attack surface					Defense solution					Black or white box		Audible	Attack target
	A1	A2	A3	A4	A5	S1	S2	S3	S4	S5	Black	White		
Did you hear that? [33]	×	√	√	×	×	×	×	×	×	×	√	×	√	Speech Commands classification Model implemented in the TensorFlow software framework
Hidden voice commands [12]	×	√	√	×	×	×	×	√	√	×	√	√	√	Google Nows speech recognition system
Stealing voices [34]	×	×	×	×	√	√	×	×	×	×	×	√	√	Speaker verification algorithms and human verification
Backdoor [21]	√	×	×	×	×	×	√	√	√	×	×	√	×	Devices equipped with nonlinear microphones
Dolphinattack [13] [22]	√	×	×	×	×	×	√	√	√	×	×	√	×	7 popular speech recognition systems across 16 common voice controllable system platforms
Walnut [35]	√	×	×	×	×	×	×	×	×	×	×	√	√	MEMS accelerometers and systems that employ on these sensors
Speake(a)r [36]	√	×	×	√	×	×	×	×	×	×	×	√	×	Common codecs in PCs (e.g., Realtek codec chips)
Note$_1$: A1-Hardware defects; A2-Algorithm defects; A3-Acoustic feature; A4-OS levels defects; A5-Imitating Note$_1$: S1-VSButton [37]; S2-Hardware improvement [13]; S3-Machine learning [13] [12]; S4-Alert [12]; S5-Speaker verification [12]														

switch to ground output. Otherwise, the LM567 outputs a high impedance (High-Z) state. The logical expression of frequency detection using the LM567 is as follows:

$$\text{output} = \begin{cases} 0, & \text{if } \frac{\left|f_c - \tilde{f}_c\right|}{f_c} \leq 14\% \\ Z, & \text{otherwise.} \end{cases} \tag{12}$$

When the LM567 provides a ground output, we can use an LED to early warn users. The illuminated LED indicates there may be an inaudible attack due to the presence of an ultrasonic frequency. This is called as the low-level alarm. However, it may cause false positives due to other innocuous interference ultrasound existing in the ambient environment. As a result, we have to demodulate the received signals and detect whether there exists a voice signal. Here, we exploit the logic output of the LM567 to control a transmission gate (TG). The output of the TG is the local oscillator frequency generated by the RC component when the LM567 provides a ground output (i.e., the TG is open). Otherwise, the output of the TG outputs is a High-Z. Here, the local oscillator frequency can be considered as the local carrier. The local carrier and the received signals input into the multiplier (MC1496) simultaneously. Then, we can use the MC1496 to obtain demodulated signals based on the coherent demodulation principle.

Finally, we exploit the VAD method to detect voice signals in the demodulated signals. The VAD module can output a logical variable to indicate whether there exists a voice signal. The ground output of the VAD module represents the presence of a voice signal. We use another LED to further warn users. This is called an advanced alarm. The lighting LED means that the existing ultrasonic signal does carry a voice signal. The voice control system may suffer an inaudible attack.

VI. RELATED WORK

In recent years, attacks to smart home devices are on the rise [27], [28]. Apthorpe *et al.* [29] demonstrated that an ISP or other network observer can infer sensitive in-home activities by analyzing Internet traffic from smart homes containing commercially available IoT devices even when the devices use encryption. Sivaraman *et al.* [30] showed how an attacker can infiltrate the home network via a doctored smartphone app. Coppolino *et al.* [31] showed how a famous MANET attack, the sinkhole attack, can be adapted to ZigBee networks and perform a cyber–physical attack to the ZigBee network. Abrishamchi *et al.* [32] provided an overview on side-channel attacks with emphasis on vulnerabilities in the smart home since built-in sensors, including cameras, microphones, motion detectors, and activity loggers, increase privacy concerns due to data leakage.

Voice has become one of the most important ways of interacting in smart homes, so the attacks against speech recognition systems are also endless. Alzantot *et al.* [33] designed attacks against speech classification model [38] that can fool the speech recognition system to produce incorrect results. Carlini *et al.* [12] extracted the features of audio commands needed for the speech recognition algorithm and added some cluttered noise to make people confused while enough for speech classifier to predict the malicious commands. It utilized the gap between human and machine [11]. Trippel *et al.* [35] investigated the damage of the digital integrity of the capacitive MEMS accelerometer with intentional acoustic interference injection. Guri *et al.* [36] introduced a new type of espionage malware, "SPEAKE(a)R," which can covertly turn the headphones, earphones, or simple earbuds connected to a PC into microphones when a standard microphone is not present, muted, taped, or turned off. Mukhopadhyay *et al.* [34] used voice morphing techniques to transform voice with any arbitrary message into a victim's voice so as to control the victim's device without a hitch. Roy *et al.* [21] utilized the nonlinearity of microphones in smart devices to play harmful commands at a frequency that is inaudible to humans but created a shadow in the audible range of the microphone. Zhang *et al.* [13] and Song [22] both exploited the same mechanism to attack our smartphone and popular speech recognition systems. This article considers the possibility of the attack to the smart home system by inaudible commands. To summarize, as enumerated in Table II, the attacks based on audio mainly consist of five

types: 1) attack due to the hardware defects; 2) attacks to speech classification model; 3) attacks by leveraging acoustic feature; 4) attack due to OS levels defects; and 5) attack by imitating victim's voice.

Researchers also proposed a series of defense solutions for attacks on smart home devices and speech recognition systems. Lei *et al.* [37] disclosed three security vulnerabilities, which root in the insecure access control of home digital voice assistants, and devise a virtual security button (VSButton), which leverages the WiFi technology to detect indoor human motions because attacks are mainly launched while victims are not at home. But obviously, users do not always move before speaking so that the VSButton will not accept normal commands when users are motionless. Also, even when users at home, they cannot notice the inaudible attack. Zhang *et al.* [13] not only proposed the Dolphin attack but also provided hardware and software defense solutions. But the hardware defense solution should change the internal structure of microphones, which is costly. Also, the software defense solution leverages the spectral difference between recovered signal and normal voice to train a classifier. But such a solution also needs to change the internal structure of devices and will cause extra delay of speech recognition process since every voice commands should be tested before going into effect. Carlini *et al.* [12] explored three defense approaches for hidden voice commands.

1) *Alert:* Notify the user every time when a voice command is accepted. But the problem is that people may feel annoyed or even overlook this prompt because they will get used to it.
2) *Speaker Verification:* But the speaker verification algorithm is not perfect at present, so that can be violently cracked. Also, as long as the voice segment of the attacker is obtained, a simple speech synthesis can synthesize the attack commands. Finally, in the smart home application scenario, different from personal electronics, there are always multiple family members, so that training the voice assistant becomes a burden to users.
3) *Classifier:* Train a classifier to distinguish human-generated voice from machine-generated voice. However, the flaw of training such a classifier is talked about above.

To sum up, the audio attacks and defense solutions are listed in Table II. Due to the limitation of existing defense solutions, in this article, we come up with a signal-processing-based detection and warning strategy to ensure that voice controllable equipment is protected from an inaudible attack.

VII. Discussion

Detection Rather Than Prevention: We put forward a signal-processing-based detection solution to detect the inaudible attack and alert users. From the defense design presented in Section III, our approach aims to detect attacks simultaneous with the attack to smart home devices, instead of completing detection and prevention before the attack. Preventing audio attacks must modify to all voice controllable devices. Whether it is a hardware or software change, the cost is too high and can cause extra delays for the normal process. So, we design an independent device here to preform detection wherever users need. In order to improve the validity and timeliness of detection, we can design a low-level alarm to alert the user that the device may be under attack when suspicious high-frequency signals are detected. At this point, the attack has not realized and the user has foreseen the possibility of the attack. Also, improve the performance of the detection device can reduce the delay of detection t_d.

Jamming Attack to Detection Device: Attacks may intentionally design ultrasound modulated noise or say meaningless voice clips rather than voice commands and transmit to smart home scene [39]. It will not cause an attack on the smart home devices but will cause the detection device to keep on alerting. Such bug emerges because the detection device cannot differentiate noise and malicious voice commands modulated in attack signals if only relies on the STE-based VAD method. Once the energy of noise exceeds the threshold T, it will be detected as voice. So, further work can apply a more efficient VAD method, such as hidden Markov model-based VAD and zero-crossing rate-based VAD. Fortunately, the continuously alerting also indicate that something harmful has happened in your home.

VIII. Conclusion

In this article, we validate the effectiveness of ultrasonic-based inaudible voice attacks to voice controllable smart home devices and propose a signal-processing-based detecting approach. Our approach uses an independent device and deploys a two-step lightweight detecting algorithm to identify the attack signals without changing the hardware or software design of microphones. The simulation results illustrate the correctness of the detection algorithm and the real-world experiments show that our approach can detect the ultrasonic-based inaudible voice attack effectively.

References

[1] D. Yu *et al.*, "Implementing abstract MAC layer in dynamic networks," *IEEE Trans. Mobile Comput.*, early access, Feb. 4, 2020, doi: 10.1109/TMC.2020.2971599.
[2] Y. Huo, C. Hu, X. Qi, and T. Jing, "LoDPD: A location difference-based proximity detection protocol for fog computing," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1117–1124, Oct. 2017.
[3] D. Yu *et al.*, "Stable local broadcast in multihop wireless networks under SINR," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1278–1291, Jun. 2018.
[4] D. Yu, L. Ning, Y. Zou, J. Yu, X. Cheng, and F. C. M. Lau, "Distributed spanner construction with physical interference: Constant stretch and linear sparseness," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2138–2151, Aug. 2017.
[5] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, and W. Lv, "Edge computing security: State of the art and challenges," *Proc. IEEE*, vol. 107, no. 8, pp. 1608–1631, Apr. 2019.
[6] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Trans. Depend. Secure Comput.*, vol. 15, no. 4, pp. 577–590, Jul./Aug. 2018.
[7] G. Yuan and C. Poellabauer, "An overview of vulnerabilities of voice controlled systems," in *Proc. 1st Int. Workshop Security Privacy Internet Things (IoTSec)*, 2018, pp. 1–4.
[8] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Trans. Netw. Sci. Eng.*, early access, Apr. 24, 2018, doi: 10.1109/TNSE.2018.2830307.

[9] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial IoTs," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 968–979, May 2020.
[10] H. Sheng *et al.*, "Mining hard samples globally and efficiently for person re-identification," *IEEE Internet Things J.*, early access, Mar. 13, 2020, doi: 10.1109/JIOT.2020.2980549.
[11] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: Exploiting the gap between human and machine speech recognition," in *Proc. USENIX Conf. Offensive Technol.*, 2015, p. 16.
[12] N. Carlini *et al.*, "Hidden voice commands," in *Proc. 25th USENIX Security Symp. (USENIX Security)*, Austin, TX, USA, 2016, pp. 513–530.
[13] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible voice commands," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2017, pp. 103–117.
[14] J. Tiete, F. Dominguez, B. da Silva, A. Touhafi, and K. Steenhaut, "MEMs microphones for wireless applications," in *Proc. Wireless MEMS Netw. Appl.*, 2016, pp. 177–195.
[15] J. Weigold, T. Brosnihan, J. Bergeron, and X. Zhang, "A MEMs condenser microphone for consumer applications," in *Proc. IEEE Int. Conf. Micro Electro Mech. Syst. (MEMS)*, Istanbul, Turkey, 2006, pp. 86–89.
[16] Z. I. Skordilis, A. Tsiami, P. Maragos, G. Potamianos, L. Spelgatti, and R. Sannino, "Multichannel speech enhancement using MEMs microphones," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 2729–2733.
[17] Amazon. (Jul. 2018). *Amazon Alexa*. [Online]. Available: https://developer.amazon.com/zh/alexa
[18] P. Dempsey, "The teardown Amazon echo digital personal assistant [teardown consumer electronics]," *Eng. Technol.*, vol. 10, no. 2, pp. 88–89, 2015.
[19] A. Nijholt, "Google home: Experience, support and re-experience of social home activities," *Inf. Sci.*, vol. 178, no. 3, pp. 612–630, 2008.
[20] Mi. (Jul. 2018). *The Open Platform of Xiaomi MI AI*. [Online]. Available: https://xiaoai.mi.com/
[21] N. Roy, H. Hassanieh, and R. R. Choudhury, "BackDoor: Making microphones hear inaudible sounds," in *Proc. 15th Annu. Int. Conf. Mobile Syst. Appl. Services (MobiSys)*, 2017, pp. 2–14.
[22] P. M. L. Song, "Inaudible voice commands," 2017. [Online]. Available: arXiv:1708.07238.
[23] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.
[24] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the Internet," in *Proc. 5th IEEE Int. Conf. High Speed Netw. Multimedia Commun.*, 2002, pp. 46–50.
[25] J. Technology. (Jul. 2018). *Ultrasonic Transducer*. [Online]. Available: http://www.jinci.cn/en/goods/112.html
[26] F. Akasheh, T. Myers, J. D. Fraser, S. Bose, and A. Bandyopadhyay, "Development of piezoelectric micromachined ultrasonic transducers," *Sensors Actuators A Phys.*, vol. 111, no. 2, pp. 275–287, 2004.
[27] X. Zheng, Z. Cai, and Y. Li, "Data linkage in smart Internet of Things systems: A consideration from a privacy perspective," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 55–61, Sep. 2018.
[28] X. Zheng, Z. Cai, J. Yu, C. Wang, and Y. Li, "Follow but no track: Privacy preserved profile publishing in cyber-physical social systems," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1868–1878, Dec. 2017.
[29] N. Apthorpe, D. Reisman, S. Sundaresan, A. Narayanan, and N. Feamster, "Spying on the smart home: Privacy attacks and defenses on encrypted IoT traffic," 2017. [Online]. Available: arXiv:1708.05044.
[30] V. Sivaraman, D. Chan, D. Earl, and R. Boreli, "Smart-phones attacking smart-homes," in *Proc. ACM Conf. Security Privacy Wireless Mobile Netw.*, 2016, pp. 195–200.
[31] L. Coppolino, V. Dalessandro, S. Dantonio, L. Levy, and L. Romano, "My smart home is under attack," in *Proc. IEEE 18th Int. Conf. Comput. Sci. Eng.*, Oct. 2015, pp. 145–151.
[32] M. A. N. Abrishamchi, A. H. Abdullah, A. D. Cheok, and K. S. Bielawski, "Side channel attacks on smart home systems: A short overview," in *Proc. 43rd Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2017, pp. 8144–8149.
[33] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," in *Proc. NIPS Mach. Deception Workshop*, 2018, pp. 1–6.
[34] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *Proc. Eur. Symp. Res. Comput. Security*, 2015, pp. 599–621.
[35] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "WALNUT: Waging doubt on the integrity of MEMs accelerometers with acoustic injection attacks," in *Proc. IEEE Eur. Symp. Security Privacy*, 2017, pp. 3–18.
[36] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, "Speake(a)r: Turn speakers to microphones for fun and profit," 2016. [Online]. Available: arXiv:1611.07350.
[37] X. Lei, G. H. Tu, A. X. Liu, C. Y. Li, and T. Xie, "The insecurity of home digital voice assistants—Amazon Alexa as a case study," 2018. [Online]. Available: arXiv:1712.03327.
[38] L. Deng *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 8604–8608.
[39] Y. Huo, Y. Tian, L. Ma, X. Cheng, and T. Jing, "Jamming strategies for physical layer security," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 148–153, Feb. 2018.