Detecting Surrounding Users by Reverberation Analysis with a Smart Speaker and Microphone Array

Naoki Yoneoka*, Yutaka Arakawa*, Keiichi Yasumoto*
* *Nara Institute of Science and Technology*, Nara, Japan
Email: {yoneoka.naoki.yk9, ara, yasumoto}@is.naist.jp

***Abstract*—Recently, smart speakers like Amazon Echo and Google Home have been spread widely. Those devices support users' life through voice interface by receiving voice commands to operate appliances and order goods to online shops. Meanwhile, it is reported that smart speakers are vulnerable to some malicious attacks which steal personal information and/or order unnecessary goods by uttering voice from a device nearby the speaker, abusing the fact that the smart speakers cannot distinguish human voice from machine voice. A new type of attack called DolphinAttack which utters ultrasonic voice inaudible to human is also reported. Therefore, a method to identify which of human or machine is sending voice commands to a smart speaker is desired. In this paper, to prevent such machine-voice based attacks to a smart speaker in absence of residents, we propose a system consisting of a speaker and a microphone array to detect the existence of a human nearby, supposing it can be incorporated in a smart speaker in the future. In our proposed system, the speaker emits sonar sound generated based on Orthogonal Frequency Division Multiplexing (OFDM) in all directions, the microphone array with 8 channels attached on top of the speaker receives the reflected sound, and the human existence is judged by comparing the reflected sound with that measured in the same environment without human. Through experiments with a prototype system, we confirmed that our proposed system can detect the human existence by measuring the reflected signal of 0.5 second.**

***Index Terms*—Security, Smart Speaker, Human Existence Detection**

I. Introduction

Recently, smart speakers with voice interface has been on the market by Google and Amazon and the shipment number has reached 16.8 million at the second quarter in 2018 [1]. When the market expands at the current pace, 90 million smart speakers will be spread to households by 2020 [2].

A smart speaker is a device consisting of a non-directional speaker and a multi-channel microphone array. It starts the services after receiving the triggers called Wake word (e.g., "Alexa" for Amazon Echo and "OK, Google" for Google Home) from a user. To notice the wake words, the microphone array continues receiving ambient voice as long as it is powered.

User's voice talked after the wake words is sent to the cloud through the Internet where the voice data is analyzed with a voice recognition algorithm. The smart speaker then answers through voice feedback and provides the service in response to the user's request. Smart speakers support user's daily life in a diverse way through various services including setting wake-up call of an alarm clock, appliance operations and online shopping.

Meanwhile, the fact that smart speakers are always connected to the Internet while receiving ambient voice causes a problem of inducing user's undesired operations. As an incident, a wake word uttered in a TV program caused a viewer's smart speaker to order a doll house [3]. For this problem inherent in voice interface, various possible attacks have been proposed by researchers so far. Roy et al. [4] showed that ultrasonic voice inaudible to human can control a smart speaker thanks to wide frequency range its microphone captures. Roy's method required two speakers, but Zhang et al. [5] extended it to use only one speaker, then the method has become an easily executable attack. The ultrasonic voice based attacks are done with speech synthesis, but Google and Microsoft provide algorithms to identify individual user's voice and thus can avoid accepting voice commands by synthesized voice. However, Diao et al. [6] found that voices uttered to the voice assistant function of a smartphone are stored in the smartphone and/or cloud and playing back these voices arbitrarily can cause undesired operations of the smart speaker. Therefore, to protect smart speakers from these kinds of attacks, it is required to judge if the voice is uttered by the user or not. In general, it is difficult to judge which of the user or the machine (speaker) utters voice only by sound analysis. Thereby, we employ an idea of judging existence of the human near the smart speaker. A possible way of realizing this idea is adding a camera and/or a motion sensor to a smart speaker. However, it increases cost of the smart speaker and impairs its diffusion.

In this paper, aiming to prevent machine-voice based attacks to a smart speaker in absence of residents, we propose a method to judge existence of human near a smart speaker by using only a speaker and a microphone array of the smart speaker. In the proposed method, a sonar sound is transmitted from the speaker of a smart speaker in all directions, and the existence of a human is judged by analyzing the reverberation sound. We generate the sonar sound using Orthogonal Frequency Division Multiplexing (OFDM) because the precise reception time can be calculated from the received signals in OFDM. Although the received sound at the microphone is mixed sounds consisting of the sonar sound coming directly

from the speaker and that coming after the reflections at the room wall, human and other obstacles, the direct sound can be easily removed because its delay time is constant. From the remaining reverberation sounds, we measure the average power of the sounds received by each microphone element. This average power at some microphone element should be different between the cases with and without human (caused by reflections at human), then by setting an appropriate threshold, we can determine the existence of human.

We developed a prototype of the proposed system with a smart speaker (SONY LF–S50G) and 8ch microphone array (TAMAGO–03 manufactured by System In Frontier Inc.). We conducted an experiment with the prototype placed in the center of a room, where in both cases with and without a human, 30 seconds sonar sounds were emitted. As a result, we found that there is a clear difference between the cases and the proposed method can identify the existence of a human with analysis of 0.5 sec sound signal which is much shorter than the state-of-the art method [7].

II. Related Work

RF-based user detection approaches are widely used in the context of indoor localization. Bocca et al. [8] used 30 or more sensors indoor and their communication network to detect users and estimated 90 % of users within 1 m error. Adib et al. [9] used two antenna arrays of $1\,\mathrm{m} \times 2\,\mathrm{m}$ long and estimated 99 % of users within 1 m error. RF-based approach can detect multiple users with high accuracy in indoor situation. On the other hand, these RF-based approaches require special or large antennas, thus its versatility is poor.

As another approach, Nandakumar et al. [10] proposed an acoustic approach for user detection in indoor localization. It transmits acoustic pulses based on SOund Navigation And Ranging (*i.e., SONAR*) which can be easily emitted and received by a speaker and two microphones of smart phone to detect user. By using a smartphone to emit/receive acoustic signals, this method can be widely used, but it cannot detect multiple users.

Recently, a new acoustic user detection system using SONAR has been developed. Alanwar et al. [7] designed a system to detect users assuming the use of smart speakers. They proposed the system to emulate smart speakers using omni-directional speaker and 8ch microphone array, and transmit SONAR pulse in indoor situation. The system calculates several statistical features from received signals and uses a clustering algorithm to detect users. In this approach, the system achieved estimation accuracy of 93.13 % using 16 seconds of received signal as a time window and 50 of statistical features. However, if the time window is longer or shorter than 16 seconds, the accuracy decreases, then it requires at least 16 seconds to detect users after starting the measurement for estimation.

In our study, we design and develop a new method to enable a smart speaker to detect nearby users in shorter time using SONAR.

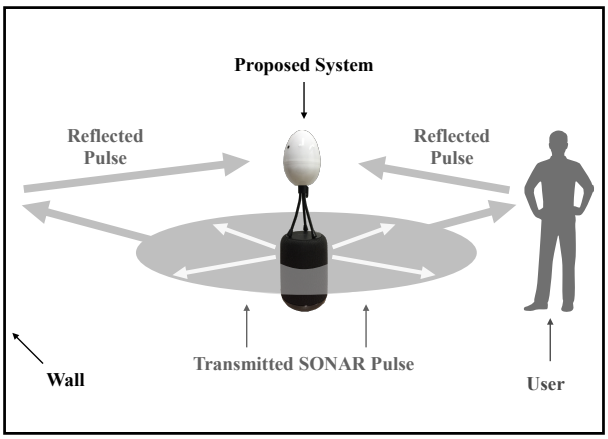

Fig. 1: Schematic diagram of the proposed system.

Fig. 2: Speaker.

Fig. 3: Mic array.

III. Human Existence Detection System

In this section, first we describe the purpose of the proposed system and then describe the detailed design of the system.

A. Purpose

As we addressed in Sect. I, there are potential threats by malicious attackers to manipulate smart speakers.

It is desirable to detect (1) existence of person in the room and (2) position of the person. With (1), the system can detect a strange situation that some voice command is sent to smart speaker despite of user's absence. With (2), attacks using inaudible voice like DolphinAttack can be detected (assuming that the positions of the person and the device emitting inaudible voice are different). In this paper, we focus on (1).

B. System Design

We propose a SONAR based person detection system. Our system consists of hardware devices and utilizes a SONAR-based pulse sound. Fig. 1 shows the schematic diagram of our system.

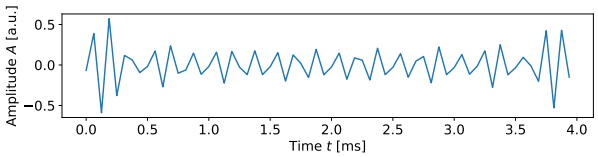

Fig. 4: SONAR pulse.

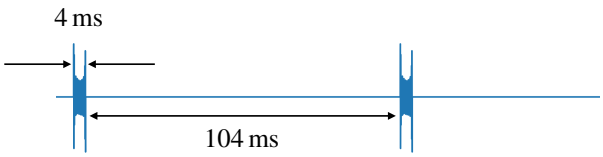

Fig. 5: The length and interval of SONAR pulse.

Fig. 6: The room used for data collection.
Proposed system is located at the center.

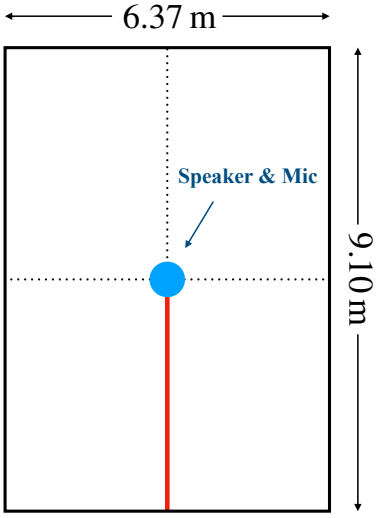

Fig. 7: The size of the room.
Blue circle is the position of the proposed system,
and red line shows the direction of the microphone's ch1.

Devices: We assemble our proposed system from a multiple channel microphone array[1] and an omni-directional speaker. This combination is similar to commercially available smart speakers such as Amazon Echo and Google Home. By emulating a smart speaker, we can deal with raw sound data captured with microphones easier than using an off-the-shelf smart speaker.

Fig. 2 and Fig. 3 show the speaker and the microphone array constituting the proposed system. For our purpose, it is desirable to produce a sound in the horizontal direction. Thus we selected Sony Corporation's LF-S50G as the omni-directional speaker which has a cylindrical shape and has speakers around it. However, LF-S50G has only two microphones (although Amazon Echo has an array of seven microphone elements). Therefore, we selected System Infrontia Corporation's TAMAGO-03 as the microphone array. TAMAGO-03 is a microphone array with eight microphone elements equally spaced in the azimuthal direction on the horizontal plane. The sampling frequency is 16 kHz, and all eight channels can be sampled synchronously.

SONAR pulse: In our system, the SONAR pulse shown in Fig. 4 is used. This SONAR pulse is generated by adding eight sine waves with the same power and phase taken at intervals of 250 Hz from 6000 Hz to 7750 Hz. Since this configuration is based on the Orthogonal Frequency Division Multiplexing (OFDM) method and it is known that the correlation between the original and the received signals is strong when using the SONAR pulse generated by this method. Thus, it is easy to estimate the signal reception time [10]. The sampling frequency of the SONAR pulse is set to 16 kHz in accordance with the sampling frequency of TAMAGO-03. Also, the time window per single SONAR pulse is 4 ms. As shown in Fig. 5, during operation of the proposed system, SONAR pulses are continuously transmitted at fixed time intervals of 108 ms (4 ms for emitting sonar pulse and 104 ms silence interval) as shown in Fig. 5 and based on [10].

[1]Since we focus only on person's existence detection in this paper, using a single non-directional microphone would be enough. However, taking into account user's position estimation to be done in the future, we used a microphone array.

IV. Data Collection Environment and Method

In this section, we describe in detail the environment and the method to collect data for evaluating user detection performance with the proposed system. The goal of the experiment is to know if the person existence detection is possible or not for various conditions in terms of distance from person to the speaker and the number of persons in the room.

A. Environment

Assuming that our system is used in indoor environments, we arranged our experiment in an indoor room which has similar size to the actual home where obstacles such as furniture exists and the air conditioner is operating.

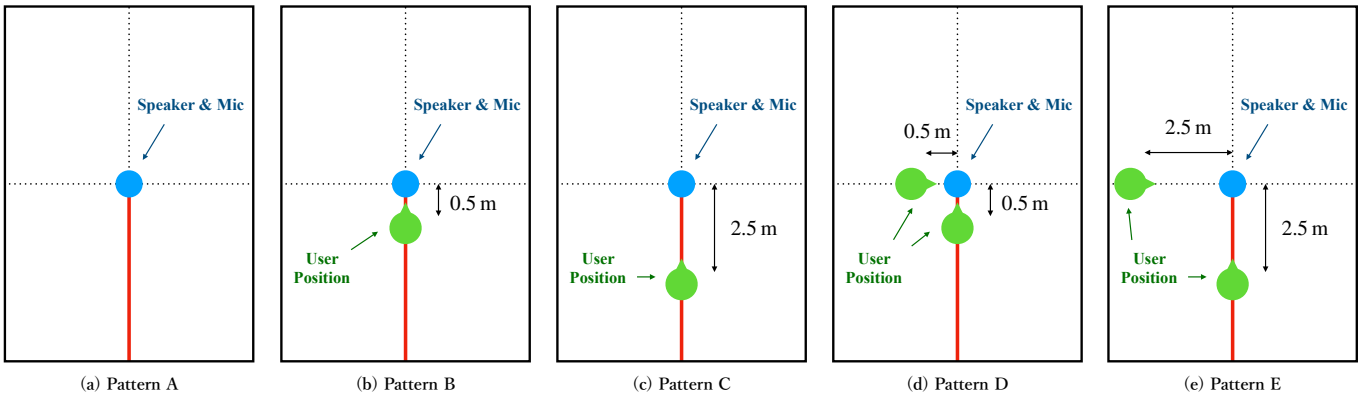

Fig. 8: Surrounding person placement patterns used for data collection

Fig. 9: Photo of Pattern B

Fig. 10: Photo of Pattern C

Fig. 6 and Fig. 7 show the photo of the room and its size, respectively, where we placed the proposed system at the center of the room.

B. Patterns of surrounding person's location

The experiment was carried out in three patterns of surrounding persons (no person, one person, two persons). Also, we evaluate the effect of distance to the surrounding persons. These patterns are named Patterns A, B, C, D and E, respectively, and these configurations are shown in Fig. 8. The photos of Pattern B and Pattern C are shown in Fig. 9 and Fig. 10, respectively. The red line in Fig. 8 is the direction in front of the microphone element ch1 painted black in Fig. 3.

In all patterns, the experiment was conducted for 30 seconds, and the SONAR pulse continued to be transmitted at regular intervals. Each pattern is described below.

Pattern A - No persons: This is the experimental situation where no users exist in the room. As shown in Fig. 8 (a), only our proposed system is put at the center of the room, and emitted the SONAR pulse for 30 seconds.

Pattern B - 1 person stands $0.5\,\mathrm{m}$ *away from the system:* We asked a participant to stand in the front direction of the microphone element ch1 $0.5\,\mathrm{m}$ away from the system, as shown in Fig. 8 (b). As shown in Fig. 9, this distance is the closest distance to the system in domestic environments. Since the analysis method is carried out assuming that the room environment is steady for 30 seconds, we asked the user to keep standing for 30 seconds while the SONAR pulse is being emitted.

Pattern C - 1 person stands $2.5\,\mathrm{m}$ *away from the system:* We asked a participant to stand in the front direction of the microphone element ch1 $2.5\,\mathrm{m}$ away from the system, as shown in Fig. 8 (c). As shown in Fig. 10, this distance is the farthest distance from the system in domestic environments. We asked the user to stand the upright immovable state for 30 seconds while emitting the SONAR pulse.

Pattern D - 2 persons stand $0.5\,\mathrm{m}$ *away from the system:* We asked two participants to stand in the front direction of the microphone element ch1 $0.5\,\mathrm{m}$ away and in the front direction of the microphone element ch7 $0.5\,\mathrm{m}$ away from the system, respectively, as shown in Fig. 8 (d). We asked the participants to keep standing for 30 seconds while the SONAR pulse is

being emitted.

Pattern E - 2 persons stand $2.5\,\mathrm{m}$ *away from the system:* We asked two participants to stand in the front directions of the two microphone elements ch1 and ch7 $2.5\,\mathrm{m}$ away from the system, respectively, as shown in Fig. 8(e). We asked the participants to keep standing for 30 seconds while the SONAR pulse is being emitted.

V. Analysis Method and Results

In this section, we describe the analysis method applied to the data obtained in Sect. IV. We first explain details of data processing and analysis, and then show analysis results with discussion.

A. Analysis Method

The analysis consists of 4 steps listed below.

1) Correlation calculation of received sound and pulse
2) Correlation filtering and separation into processing units
3) Baseline acoustic characteristic calculation and its subtraction
4) Mean power calculation and comparison

Calculation was carried out independently for each pattern from Step 1 to 4, and at the end of Step 4, the average power of cross-correlation obtained in each pattern is compared with other patterns. If the average power of patterns B–E is greatly different from that of pattern A (no person), we regard that the person existence detection is possible. Each step is described in detail below.

Step 1 Correlation calculation of received sound and pulse: The sound received by the microphone array includes the SONAR pulse transmitted from the omni-directional speaker, but in a noisy environment with many obstacles like home, it is difficult to calculate the reception time of the SONAR pulse from the received sound signal. Therefore, by calculating the cross correlation between the received sound and the SONAR pulse, we make it easy to calculate the reception time of the SONAR pulse. Cross correlation $\phi(t)$ at t-th sample in the received sound signal is given by

$$\phi(t) = \frac{1}{N_s} \sum_{i=0}^{N_s-1} x(t+i)s(i),$$

where $x(t)$ is the sound value of t-th sample in the received sound signal, and $s(i)$ is the sound value of i-th sample in SONAR pulse, and N_s is the number of samples (i.e., 16KHz×4ms=64) in SONAR pulse. When there is a waveform with high correlation with SONAR pulse at t-th sample in the received sound, $\phi(t)$ shows a large value, then we can calculate the reception time of SONAR pulse from the maximal value of $\phi(t)$.

Step 2 Correlation filtering and separation into processing units: Among the eight kinds of frequency bands of the SONAR pulse (see Sect. III-B) used in the experiment, the frequency band which can be received with the largest sound volume among them is regarded as the frequency band with the least loss and used for analysis. Therefore, cross correlation is applied to the band pass filter to remove noises other than the frequency band of interest. Furthermore, this cross correlation, time series data for about 30 seconds are separated into SONAR pulse receptions called *cross correlation processing units* or just *units* with length of 108ms. Thus, cross correlation is calculated for each unit in the received sound signal at each microphone channel.

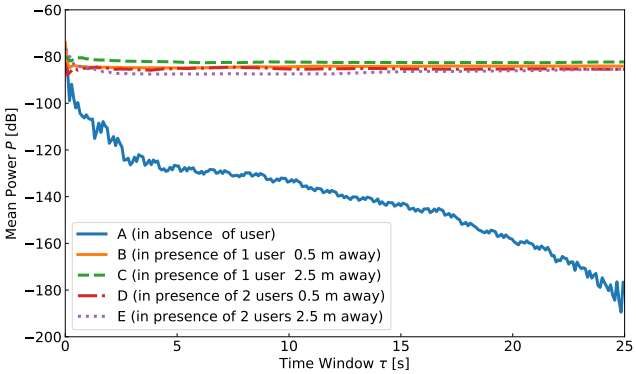

Fig. 11: Mean Power for Different Patterns and Different Time Window Widths

Step 3 Baseline acoustic characteristic calculation and its subtraction: The received sound includes both direct sound coming from the speaker and indirect sound coming after reverberated in the environment (e.g., reflection at obstacles), and it is considered that direct sound is not related to persons' existence. Therefore, we average the cross correlation values calculated for all units in the received signal (30 seconds) in pattern A (case of no person) for each of the eight channels. The derived vector (8 channels) of values called *cross correlation vector* is used as the baseline for the case of no person (pattern A). By subtracting this baseline vector from the cross correlation vector calculated from the received signals in other pattern, we can cancel the direct sound effect and the acoustic characteristics of the environment.

Step 4 Mean power calculation and comparison: First, the difference between the correlation value and the baseline value is calculated for each unit in the received signals at each channel. To eliminate fluctuation due to phase difference, etc., we average the calculated differences over all units in the signal received at each channel. Then, we obtain carrier power by calculating averaged square of the averaged differences over channels. Since the value of carrier power varies depending on the number of units, this value is calculated as a function of the number of units (i.e., time window of the received sound used for calculation).

Finally, person existence is judged by comparing the average power obtained from the received sound to that of pattern A.

B. Result

Fig. 11 shows the mean power P derived for different patterns (A–E) and different time window (108ms to 25s by 108ms step, multiples of 108ms). The figure shows that

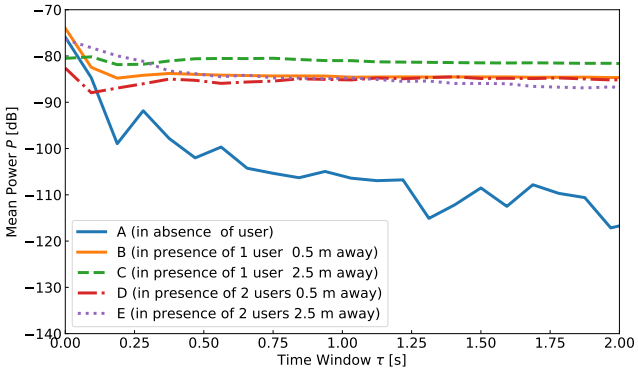

Fig. 12: Magnified graph in Fig. 11 for $0 \leq \tau \leq 2$.

the average power of pattern A is clearly smaller than the other patterns (B–E) regardless of the time window width. Moreover, we see that the power decreases as the time window width increases. This is because we calculated mean square of differences between cross correlation value of each unit in the received sound signal and that of the baseline (averaged over all units in 30 seconds) and the environmental noise varies from time to time. This result suggests that the larger time window is more robust against versatile environmental noise.

Fig. 12 shows the figure enlarging the region of τ between 0 and 2 sec in Fig. 11. P of pattern A can be clearly separated from P in other patterns by setting τ to over 0.5 sec and the threshold to around -95dB.

C. Discussion

Since power P with and without person(s) is clearly separable from each other even when τ is less than 1 s, by appropriately setting the average power threshold (e.g., -95dB). That means with our method existence of persons can be judged within 1 sec since the SONAR pulse is emitted. Since the state-of-the-art method [7] takes 16 s to detect the human since the SONAR pulse emission started, our proposed method detects persons much more quickly.

Meanwhile, Figs. 11 and 12 also show that it is difficult to distinguish between patterns B–E. That means, with the current approach, it is difficult to detect the position of the user. Our current approach just calculates the mean power by averaging the values obtained from all microphone channels, but we need to analyze difference between those channels to detect the user position (direction and distance to the user).

VI. Conclusion

In this paper, to protect a smart speaker from attacks by machine voice, we proposed a system to detect persons around the smart speaker. In the proposed system, first the SONAR pulse emitted from the speaker is received by a multi-channel microphone array, second the influence of the direct sound coming from the speaker is eliminated from the received signal, third the average power is calculated for the residual reverberation sound, then it is used as a reference in detecting existence of persons. The average power is calculated sequentially from the start of SONAR pulse emission, and by judging whether the average power falls below the predetermined threshold or not, our system determines existence of person(s) in a reasonably short time. In our experiment, the time of person detection by the proposed system was 0.5 sec, much quicker than the state-of-the-art method, and we found that our method based on the average power is effective for quick person existence detection.

Acknowledgements

This study was partly supported by Grant-in-Aid for Scientific Research 17KT0080 and 15KK0011 (Fund for the Promotion of Joint International Research (Fostering Joint International Research)).

References

[1] Canalys, "Global smart speaker shipments grew 187% year on year in q2 2018, with china the fastest-growing market," https://www.canalys.com/static/press_release/2018/Press-release-160818-global-smart-speaker-shipments-grew.pdf, accessed: 11 Jan. 2019.

[2] Ministry of Internal Affairs and Communications, Japan, "2018 WHITE PAPER Information and Communications in Japan," http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/pdf/n1100000.pdf, accessed: 11 Jan. 2019 (in Japanese).

[3] A. Liptak, "Amazon's alexa started ordering people dollhouses after hearing its name on tv," https://www.theverge.com/2017/1/7/14200210/amazon-alexa-tech-news-anchor-order-dollhouse, The Verge, accessed: 11 Jan. 2019.

[4] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 2–14.

[5] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 103–117.

[6] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*, 2014, pp. 63–74.

[7] A. Alanwar, B. Balaji, Y. Tian, S. Yang, and M. Srivastava, "Echosafe: Sonar-based verifiable interaction with intelligent digital agents," in *Proceedings of the 1st ACM Workshop on the Internet of Safe Things*, 2017, pp. 38–43.

[8] M. Bocca, O. Kaltiokallio, N. Patwari, and S. Venkatasubramanian, "Multiple target tracking with rf sensor networks," *IEEE Transactions on Mobile Computing*, vol. 13, no. 8, pp. 1787–1800, 2014.

[9] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person localization via rf body reflections." in *NSDI*, 2015, pp. 279–292.

[10] R. Nandakumar, A. Takakuwa, T. Kohno, and S. Gollakota, "Covertband: Activity information leakage using music," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 87, 2017.