Voice Attacks to AI Voice Assistant

Seyitmammet Alchekov Saparmammedovich[1], Mohammed Abdulhakim Al-Absi[1], Yusuph J. Koni[1], and Hoon Jae Lee[2(✉)]

[1] Department of Computer Engineering, Dongseo University, 47 Jurye-ro, Sasang-gu, Busan 47011, Republic of Korea
mslchekov@gmail.com, Mohammed.a.absi@gmail.com, yusuphkoni@gmail.com

[2] Division of Information and Communication Engineering, Dongseo University, 47 Jurye-ro, Sasang-gu, Busan 47011, Republic of Korea
hjlee@dongseo.ac.kr

Abstract. Everything goes to the fact that our communication with technology will soon become almost exclusively oral. It is natural for a person to ask for something out loud and hear the answer: see how children are at ease with voice assistants. However, with new technologies - and voice control is no exception - new threats are emerging. Cybersecurity researchers tirelessly seek them out so that device manufacturers can secure their creations before potential threats turn into real attacks. However, in this paper we are going to talk about different voice attacks, which so far will hardly be able to find practical application, but the protection against which should be thought over now.

Keywords: Voice attack · AI · Voice assistant · Attack distance · Assistant · Artificial intelligence

1 Introduction

Virtual assistants can "live" on smartphones, tablets and computers (like Apple Siri) or stationary devices (like Amazon Echo and Google Home speakers). The range of their possibilities is constantly increasing: you can control music playback, find out the weather, regulate the temperature in the house, order goods in online stores…

What is the Voice Assistant?

Voice assistant is an artificial intelligence-based service that recognizes human speech and is able to perform a specific action in response to a voice command. Most often, voice assistants are used in smartphones, smart speakers, web browsers.

The functionality of voice assistants is quite diverse. What the voice assistant can do:

- conduct dialogues,
- offer quick answers to user questions,
- call a taxi,
- make calls,

M. Singh et al. (Eds.): IHCI 2020, LNCS 12615, pp. 250–261, 2021.
https://doi.org/10.1007/978-3-030-68449-5_26

- lay routes,
- place orders in the online store, etc.

Since all voice assistants have artificial intelligence, when communicating with the user, they take into account the change in his location, time of day and days of the week, search history, previous orders in the online store, etc.

- Google Now - is one of the first Google voice assistants. Works on devices with Android, iOS and Chrome browser. Likes to suggest the best routes to home, taking into account the current location of the user, offering news feeds, can analyze mail and search queries. Google Now is integrated with all Google services and some third-party applications.
- Google Assistant - is a more advanced version of the voice assistant. Can conduct dialogues and understand normal spoken language.
- Siri. Works on Apple devices only. Knows how to conduct dialogues and give recommendations, for example, where to go or which movie to watch.
- Microsoft Cortana. Available on Windows, iOS and Android. Manage reminders and calendars, track packages, set alarms, and search Bing for news, weather, and more.
- Amazon Alexa. Built into Amazon audio devices (Echo, Echo Dot, Tap) and Fire TV box. Can play music, read news, offer weather and traffic information, and voice order on Amazon.

How does it work?

The history of voice assistants begins in the late 1930s, when scientists began to attempt to recognize the voice using technology. Then two big problems got in the way of creating a quality assistant:

1. the existence of homonyms - words with the same sound, but with different meanings,
2. constant background noise from which the system must select the user's speech.

Developers are now using machine learning to solve these problems. It teaches neural networks to independently analyze the context and determine the main source of sound. However, the developers did not come to this immediately - it took at least 80 years of preparatory work.

How modern voice assistants work? Voice assistants passively read all sound signals, and for active work they need activation using a passphrase. For example, say: "Okay, Google", then you can ask your question or give a command without pauses.

At the moment of a voice request, the automatic speech recognition system (ASR system) converts the audio signal into text. This happens in four stages:

- Filtration. The system removes background noise and interference arising during recording from the audio signal.
- Digitization. Sound waves are converted into digital form that a computer can understand. The parameters of the received code also determine the quality of the recording.

- Analysis. Sections containing speech are highlighted in the signal. The system evaluates its parameters - to which part of speech the word belongs, in what form it is, how likely the connection between two words is.
- Revealing data patterns. The system includes the obtained information into a dictionary - it collects different versions of the pronunciation of the same word. To more accurately recognize new queries, assistants compare the words in them with patterns.

If, after processing the request (Fig. 1), the virtual assistant does not understand the command or cannot find the answer, he asks to rephrase the question. In some cases, additional data may be required - for example, when calling a taxi, the assistant can specify the passenger's location and destination.

Fig. 1. Processing the request

1- Device is always listening, unless muted
2- Device starts recording when the trigger word is heard, e.g. "Hey Siri or Alexa". The LED indicates recording status.
3- Recording is sent to the cloud for processing & stored
4- Response is sent back Traffic is SSL encrypted
5- Optional: Backend can send data to third-party extensions (Actions/Skills) for additional processing

2 Various Voice Attacks

- Adversarial Attacks via Psychoacoustic Hiding:

Researchers at the Ruhr University in Germany have found that voice assistants can be hacked using commands hidden in audio files that cannot be discerned to the human ear. This vulnerability is inherent in the speech recognition technology itself, which is used by artificial intelligence (Fig. 2).

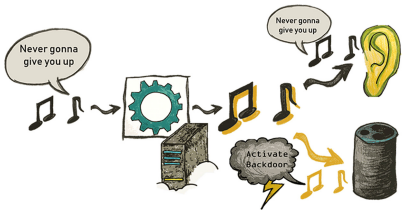

Fig. 2. various voice attacks

According to Professor Thorsten Holz, this hacking method is called "psychoacoustic hiding". With it, hackers can hide words and commands in different audio files - with music or even birdsong - that only a machine can hear. The person will hear the usual chirp, but the voice assistant will be able to distinguish something else [2].

- Laser attack.

Figure 3, researchers from the University of Michigan (USA) were able to experimentally implement a way to get remote access through a portable laser system to some iPhones, smart speakers HomePod and Amazon Echo, devices of the Google Home series using voice assistants Siri, Google Assistant and Amazon Alex. Laser beams with coded signals were sent to the place of placing the microphone in smart devices, the maximum working distance achieved in the experiment was 110 m, and the devices could thus be controlled even through a glass unit. Most speakers and gadgets took such an impact for voice commands, allowing them to perform basic actions in a smart home system that do not need additional user identification - to open a garage or a lock on a house door [3].

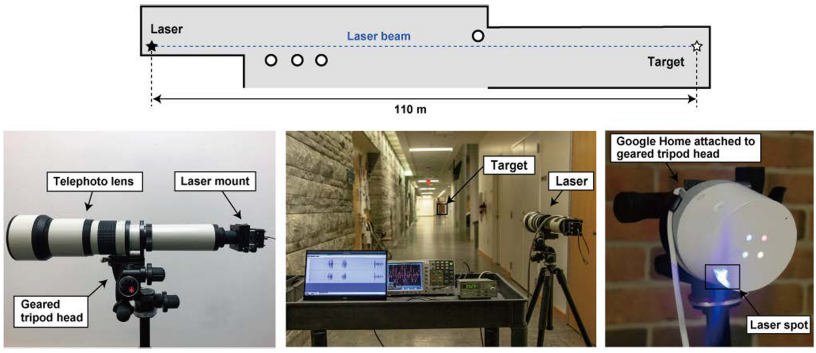

Fig. 3. Laser attack

- SurfingAttack.

Another attack technique, dubbed SurfingAttack, uses voice commands encrypted in ultrasonic waves. With the help of such waves, a potential attacker can quietly activate the voice assistant. SurfingAttack can be used to perform a limited set of actions: make calls or read text messages [4].

3 Problems

What risks do we take upon ourselves by installing a smart home system? The one that drives light bulbs and a kettle from an application on a smartphone, inside the local network and remotely. If security and lock management are tied to a smart system (as in the case of Amazon Key), then it is clear which ones. If not, then theoretically it is

possible to imagine the danger of a software failure of some coffee maker, followed by a fire, or by the influence of an intruder to gain access to your home space.

Siri itself, and other voice assistants - Cortana (Microsoft), Alexa (Amazon), Bixby (Samsung), as well as the only representative of the "masculine" Google Assistant - also do not shine with intelligence.

And this is understandable. "Smart" voice assistants are based on the architecture of neural networks and machine learning technology. It should be understood that there are about 86 billion neurons in the human brain, and in modern artificial intelligence there are only a few hundred thousand. If you count the number of neurons in the nervous system of various animals, it turns out that, as noted by the founder and head of ABBYY, David Yang, *now artificial intelligence is dumber than a bee.*

Despite this, the IQ of artificial intelligence has doubled approximately every two years.

This progression shows that sooner or later AI will still reach the human level - but this is still far from it. According to leading analyst of Mobile Research Group Eldar Murtazin, voice assistants in the next year or two will be the hallmark of premium consumer electronics, and then move to the mid-price and budget segment.

Voice assistants certainly have bright prospects. But we are talking about the future, about a new quality level, unattainable for today.

4 Voice Assistants and Security

The development of voice search and speech recognition technologies causes an ambiguous reaction from many - first of all, how safe these developments are and whether they always listen only to their owner. Voice assistants have already appeared in several notable stories:

- UC Berkeley students have found a way to launch voice assistants Siri, Alexa and Google Assistant without the owner's knowledge. To do this, it is enough to add sounds to music or video that remotely resemble human speech - the program will be enough to turn them into words and start executing the given command.
- In China, they were able to activate voice assistants using sound frequencies that the average person cannot hear.
- Burger King launched an ad with the phrase "OK Google, what is Whopper?" To which the voice assistants responded and began to read lines from Wikipedia.
- Amazon Echo ordered a dollhouse after hearing a request from a 6-year-old girl. And when this story was discussed in the news, voice assistants, taking the phrase about ordering a house as a command, began to order these houses everywhere.
- Amazon Echo owners have complained that the device starts to laugh spontaneously. It turned out that the device recognized the surrounding sounds as a command to laugh and followed it.

At the same time, Google and Amazon claim that their assistants do not turn on if they do not hear the owner's voice. Apple says Siri will never execute a command related to personal data if the iPhone or iPad is locked.

5 Comparison of Voice Attacks

In this section, we talked about different voice attacks and explained what their characteristics are, in contrast to others of the kind (Table 1).

Table 1. Comparison of voice attacks

Pros and cons	Attack name		
	1	2	3
	Psychoacoustic hiding	Laser attack	Surfing attack
Inaudible	No	Yes	Yes
Visibility	No	Yes	No
Attacking in distance and how far?	Yes	Yes	No
	Everywhere	Depends on laser power and requires line of sight	The distance should be small

If we talk about the attack "Psychoacoustic Hiding", then Hackers can play a hidden message through an app in an advertisement. Thus, they are able to make purchases on behalf of other people or steal confidential information. In the worst case, an attacker can take control of the entire smart home system, including cameras and alarms.

Attackers can use the "masking effect of sound" for their own purposes: when your brain is busy processing loud sounds of a certain frequency, then for a few milliseconds you stop perceiving quieter sounds at the same frequency. This is where the team found to hide commands to hack any speech recognition system like Kaldi, which underlies Amazon's Alexa voice assistant.

A similar principle allows you to compress MP3 files - an algorithm determines what sounds you can hear and removes anything inaudible to reduce the size of the audio file. However, hackers do not remove inaudible sounds, but replace them with the ones they need. Unlike humans, artificial intelligence like Alexa is able to hear and process every sound. He was trained so that he could understand any sound command and carry it out, whether people hear it or not.

If we talk about the disadvantages of this attack, then this is only for devices that can recognize voices. Otherwise, the message inside the audio file cannot be received by the voice assistant.

In addition, as the success of an attack improves, the original audio example affects the quality of the example. To do this, researchers [2] recommend using music or other unsuspecting sound patterns, such as the chirping of birds, that do not contain speech because speech must be dimmed, which usually results in more severe unwanted disturbances.

As for this attack, a recent discovery shows that our smart buddies can be hacked using lasers. The attack is based on the use of a photoacoustic effect, in which the absorption of changing (modulated) light by a material leads to thermal excitation of

the medium, a change in the density of the material and the appearance of sound waves perceived by the microphone membrane. By modulating the laser power and focusing the beam on the hole with the microphone, you can achieve the stimulation of sound vibrations that will be inaudible to others, but will be perceived by the microphone.

For an attack, as a rule, a simulation of the owner's voice is not required, since voice recognition is usually used at the stage of accessing the device (authentication by pronunciation "OK Google" or "Alexa", which can be recorded in advance and then used to modulate the signal during an attack). Voice characteristics can also be tampered with modern machine learning-based speech synthesis tools. To block an attack, manufacturers are encouraged to use additional user authentication channels, use data from two microphones, or install a barrier in front of the microphone that blocks the direct passage of light.

Limitations of laser attack are:

- Because of the dotted pointer, Lasers must point directly to a specific component of the microphone in order to transmit audio information.
- Attackers need a clear line of sight and a clear path for the lasers.
- Most of the light signals are visible to the naked eye and can identify intruders.
- In addition, when activated, voice control devices react loudly, which can alert nearby people of foul play.
- Accurate control of advanced lasers requires expertise and equipment. When it comes to ranged attacks, there is a high barrier to entry.

The next attack is called the SurfingAttack. How does it work? This attack allows you to remotely control the virtual assistants Google Assistant and Apple Siri using ultrasonic commands that are invisible to the human ear.

The upper bound frequency of human voices and human hearing is 20 kHz. Thus, most audio-capable devices (e.g., phones) adopt audio sampling rates lower than 44 kHz, and apply low-pass filters to eliminate signals above 20 kHz. The voice of a typical adult male has a fundamental frequency (lower) of 85 to 155 Hz, and the voice of a typical adult woman from 165 to 255 Hz.

Telephony uses a frequency band from 300 Hz to 3400 Hz to determine speech. For this reason, the frequencies between 300 and 3400 Hz are also called voice frequencies.

The person is able to hear sound vibrations in the frequency range from 16-20 Hz to 15-20 kHz.

Sound below the human hearing range is called ultrasound.

The researchers explain that, in essence, voice assistants listen to a much wider frequency than the human voice can reproduce. Because of this, they can react to ultrasonic vibrations and interpret them as voice commands. As a result, an attacker is able to discreetly interact with devices using voice assistants, intercept two-factor authentication codes, and make calls.

In SurfingAttack, the researchers showed that the ultrasonic waves can be sent not only directly, but also through materials of considerable thickness, such as a single piece of glass or even a wooden table (Fig. 4).

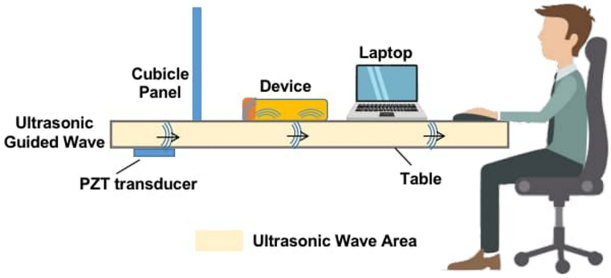

Fig. 4. SurfingAttack leverages ultrasonic guided wave in the table generated by an ultrasonic transducer concealed beneath the table [4].

Some mobile devices have customized wake-up words, SurfingAttack will not be able to activate them simultaneously, but it offers another opportunity for launching a targeted attack when the attacker learns the specific wake-up words.

Open questions of the papers (Table 2).

Table 2. Desction questions

Attack names	Suggestions from researchers
Psychoacoustic Hiding [2]	As the attack's success improves, the original audio example affects the quality of the example. To do this, the researchers recommend using music or other unsuspecting sound patterns, such as the chirping of birds, that do not contain speech, because speech must be muffled, which usually leads to more serious unwanted disturbances
Lightcommands [3]	Why the aperture vibrates from the light? At the time of writing the paper has not been studied
SurfingAttack [4]	Attacking Standing Voice Assistants. Amazon Echo and Google Home are standing voice assistants with microphones distributed across the cylinder. The current SurfingAttack cannot reach these microphones. Researchers believe this is due to the significant power loss during the power transition across the boundary of the table material and speaker material, as well as the devices' internal construction in terms of the relative position of the microphones

6 Ranged Attack

An attack using a method called "psychoacoustic concealment" [4] which can be carried out anywhere in the world using the Internet. The main goal of the attacker is to play that exactly the video or audio file in the place where the attacked Voice assistant is located. The document did not specify the exact distance, but with the attack itself, you can get access to many devices.

Lightcommands obtain control over these devices at distances up to 110 m and from two separate buildings. In addition, lightcommands have demonstrated that light can be used to control VC systems in buildings and through closed glass windows at similar distances.

In fact, the only major limitation to Light Commands attacks is the need to be within line of sight of the target device and point the light very accurately at the microphone. In the course of experiments, specialists managed to carry out Light Commands attacks from a distance of up to 110 m, including located in another building, opposite the room with the target device, or at the other end of a very long corridor.

Then aiming the laser beam at the microphone ports of the devices listed in Table 3 [3] from a distance of approximately 30 cm.

The attack distance for this SurfingAttack is very small. As shown in the Table 3 below, the MDF's attack range was significantly shorter than other table materials, which the researchers believed could also be improved by increasing signal strength. using an attack power of 1.5 W the SurfingAttack achieves a maximum attack range of 50 cm on MDF material. To improve the efficiency of SurfingAttack, an attacker can attach multiple transformers distributed across the table, which can reduce the short attack distance limit for MDF tables (Table 4).

7 Protection Against Inaudible Attacks

Manufacturers are already considering measures to protect voice-controlled devices. For example, detection of processing traces in the received signal in order to change its frequency can help from ultrasonic attacks. It would be nice to teach all smart devices to recognize the owner by voice - however, Google, which has already tested these measures in practice on its assistant, honestly warns that this protection can be bypassed using voice recording, and with proper acting skills, the timbre and manner a person's speech can be faked.

There are many solutions to protect our AI systems, such as

- Erase sensitive recordings from time to time.
- If you are not using the voice assistant, mute it.
- Turn off purchasing if not needed or set a purchase password.
- Lock the voice assistant down to your personal voice pattern, when available.
- Protect the service account linked to the device with a strong password and 2FA.
- Disable unused services, such as music streaming services.
- Do not turn off automatic update functions on the device.
- Don't use the voice assistant to remember private information such as passwords.
- Use a WPA2 encrypted Wi-Fi network and not an open hotspot at home.
- Create a guest Wi-Fi network for guests and unsecure IoT devices.
- Pay attention to notification emails, especially ones about new orders for goods.

But researchers give us their own ideas. For example, researches of SurfingAttack think Locking the device and turning off the personal results feature on the lock screen can be one way to protect against SurfingAttack. Note that only pattern, PIN and password screen lock can withstand SurfingAttack, swipe screen lock cannot.

Table 3. Tested devices with minimum activation power and maximum distance achievable at the given power of 5 mW and 60 mW. A 110 m long hallway was used for 5 mW tests while a 50 m long hallway was used for tests at 60 mW [3].

Device	Backend	Category	Authentication	Minimum power [mW]*	Max distance at 60 mW [m]**	Max distance at 5 mW [m]***
Google Home	Google Assistant	Speaker	No	0.5	50+	110+
Google Home Mini	Google Assistant	Speaker	No	16	20	–
Google Nest Cam IQ	Google Assistant	Camera	No	9	50	–
Echo Plus 1st Generation	Alexa	Speaker	No	2.4	50+	110+
Echo Plus 2nd Generation	Alexa	Speaker	No	2.9	50+	50
Echo	Alexa	Speaker	No	25	50+	–
Echo Dot 2nd Generation	Alexa	Speaker	No	7	50+	–
Echo Dot 3rd Generation	Alexa	Speaker	No	9	50+	–
Echo Show 5	Alexa	Speaker	No	17	50+	–
Echo Spot	Alexa	Speaker	No	29	50+	–
Facebook Portal Mini	Alexa	Speaker	No	1	50+	40
(Front Mic)	Portal	Speaker	No	6	40	–
Facebook Portal Mini	Alexa	Streamer	No	13	20	–
(Front Mic)§	Alexa	Thermostat	No	1.7	50+	70
Fire Cube TV						
EcoBee 4						
iPhoneXR (Front Mic)	Siri	Phone	Yes	21	10	–
iPad 6 Gen	Siri	Tablet	Yes	27	20	–
Samsung Galaxy S9	Google Assistant	Phone	Yes	60	5	–
(Bottom Mic)	Google Assistant	Phone	Yes	46	5	–
Google Pix 2 (Bottom Mic)						

*at 30 cm distance, **Data limited to a 50 m long corridor, ***Data limited to a 110 m long corridor, §Data generated using only the first 3 commands.

Table 4. Maximum attack distance on different tables (attack power is less than 1.5 W). The width of Aluminum metal table is 910 cm, the width of metal table is 95 cm, and the width of glass table is 85 cm (A – Activation, R – Recognition) [4].

Device	Max attack distance (cm)							
	Aluminum metal sheet (0.3 mm)		Steel metal sheet (0.8 mm)		Glass (2.54 mm)		MDF (5 mm)	
	A	R	A	R	A	R	A	R
Xiaomi Mi 5	910+	910+	95+	95+	85+	85+	50	47
Google Pixel	910+	910+	95+	95+	85+	85+	45	42
Samsung Galaxy S7	910+	910+	95+	95+	85+	85+	48	N/A

According to laser attack researchers the fundamental solution to prevent Light Commands requires a redesign of the microphone, which seems to require a large cost.

8 Conclusion

To summarize, I wanted to show the difference between some attacks. Each attack, while unique, has its own pros and cons. Manufacturers are already considering measures to protect voice-controlled devices. For example, detection of processing traces in the received signal in order to change its frequency can help from ultrasonic attacks. It would be nice to teach all smart devices to recognize the owner by voice - however, Google, which has already tested these measures in practice on its assistant, honestly warns that this protection can be bypassed using voice recording, and with proper acting skills, the timbre and manner a person's speech can be faked.

Regarding laser attack, as I noted in Sect. 7, it is the opinion of the researchers that a fundamental solution to prevent light commands requires a redesign of the exactly MEMS microphones, which appears to be expensive.

Acknowledgment. This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(grant number: NRF-2016R1D1A1B01011908).

References

1. Zhang, R., Chen, X., Wen, S., Zheng, X., Ding, Y.: Using AI to attack VA: a stealthy spyware against voice assistances in smart phones. IEEE Access **7**, 153542–153554 (2019)
2. Schönherr, L., Kohls, K., Zeiler, S., Holz, T., Kolossa, D.: Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding, 16 August 2018

3. Sugawara, T., Cyr, B., Rampazzi, S., Genkin, D., Fu, K.: Light commands: laser-based audio injection attacks on voice-controllable systems (2020)
4. Yan, Q., Liu, K., Zhou, Q., Guo, H., Zhang, N.: SurfingAttack: interactive hidden attack on voice assistants using ultrasonic guided waves, January 2020
5. Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., Xu, W.: DolphinAtack: inaudible voice commands, 31 August 2017
6. Roy, N., Shen, S., Hassanieh, H., Choudhury, R.R.: Inaudible voice commands: the long-range attack and defense, 9–11 April 2018
7. Zhou, M., Qin, Z., Lin, X., Hu, S., Wang, Q., Ren, K.: Hidden voice commands: attacks and defenses on the VCS of autonomous driving cars, October 2019
8. Gong, Y., Poellabauer, C.: An overview of vulnerabilities of voice controlled systems, 24 March 2018
9. Yan, C., Zhang, G., Ji, X., Zhang, T., Zhang, T., Xu, W.: The feasibility of injecting inaudible voice commands to voice assistants, 19 March 2019
10. Gong, Y., Poellabauer, C.: Protecting voice controlled systems using sound source identification based on acoustic cues, 16 November 2018