Spring 2023

Prediction and Root-Cause Analysis for Smart Speaker Intentional Electromagnetic Interference Attacks

Tanner Fokkens
Missouri University of Science and Technology

Department:

PREDICTION AND ROOT-CAUSE ANALYSIS FOR SMART SPEAKER INTENTIONAL ELECTROMAGNETIC INTERFERENCE ATTACKS

by

TANNER FOKKENS

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING

2023

Approved by:

Chulsoon Hwang, Advisor
DongHyun Kim
Daryl G. Beetner

PUBLICATION THESIS OPTION

This thesis consists of the following two articles, formatted in the style used by the Missouri University of Science and Technology:

Paper I, found on pages 2–30, has been published in *2021 IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMCSI)*, August 2021.

Paper II, found on pages 31–46, is intended for submission to *2023 IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMCSI)*, July 2023.

ABSTRACT

Smart home and Internet of Things (IoT) devices have become ubiquitous in homes over the past decade. The smart speaker itself is often the device that interfaces all these devices together. Because of this, the smart speaker can become a point of attack for someone trying to exploit or hack into the smart home devices. In the past few years, it was discovered that smart speakers with microwave electromechanical system (MEMS) microphones are susceptible to intentional electromagnetic interference (I-EMI) attacks by modulating an audio command to a high-frequency carrier signal. This attack allows for command recognition for long-distances and smart speakers behind walls.

First, a method for modeling and understanding the smart speaker I-EMI attack is shown. This includes a method for finding the ideal attack angle, locating the region sensitive to the coupled EMI, and modeling the attack. Finally, using all these methods, a long distance (6-meter) attack is demonstrated using 6.3 Watts of power at the aggressor antenna.

Next, the effectiveness of using machine learning (ML) synthesized voice samples to control smart speaker devices through radiated intentional electromagnetic interference (I-EMI) in presented. Devices that are trained to only recognize a single person's voice or only execute certain commands from that person will not be as susceptible to the I-EMI attack. By training a neural network using samples of the target's voice, this security feature can be bypassed, increasing the feasibility of the attack.

ACKNOWLEDGEMENTS

I would like to express my gratitude and thanks to my advisor, Dr. Chulsoon Hwang for his exceptional guidance, mentorship, and patience. His support has been instrumental to my success as a graduate student.

Next, I would like to thank Dr. DongHyun (Bill) Kim, Dr. Daryl Beetner, Dr. Victor Khilkevich, and all the other graduate students at the EMCLAB for both mentorship and friendship. This is a wonderful lab and a fantastic group of people – It has been a privilege to be involved in such a group.

Additionally, I would like to thank Wei Huang of ESDEMC Technology for initially taking me on as an intern as a freshman in undergrad and introducing me to this exciting field. I will always be indebted to him for taking a risk on me so early in my time as a student and spurring my interest in high-speed signals and interference.

I also want to recognize the friends I made during my 4 years of undergraduate studies at Missouri S&T. These friendships, which I am certain will last a lifetime, are as treasured to me as the academic experience I gained during my degree. These same friends also motivated and pushed me to persue graduate studies. For this, I am forever grateful.

Finally, I am greatly thankful for my parents and family's support and love throughout my academic journey, and every other step in my life. Without their support, I am certain I would not have made it to where I am today.

TABLE OF CONTENTS

SECTION

LIST OF ILLUSTRATIONS

PAPER I Page

PAPER II

SECTION

1. INTRODUCTION

1.1. BACKGROUND

A smart speaker is a device that executes commands based on a user's voice. These devices started as simply allowing for basic conversational skills, but later full-function smart home devices began allowing the smart speakers to control functions like electric plugs, locks, or thermostats. Eventually, this same technology made its way to most smart phones and computers.

These devices utilize a different type of microphone from that which is conventionally seen, which is known as a Microelectromechanical systems (MEMS) microphone. This microphone contains an analog to digital (ADC) converter, amplifier, and lowpass filter all within the package of the microphone. Recently, it was discovered that this microphone is susceptible to an attack that allows for commands to be sent inaudibly through intentional electromagnetic interference (I-EMI). Understanding this attack is important for both mitigation, and consideration of its threat profile.

PAPER

I. PREDICTION AND ROOT-CAUSE ANALYSIS FOR SMART SPEAKER INTENTIONAL ELECTROMAGNETIC INTERFERENCE ATTACKS

Tanner Fokkens, Shengxuan Xia, Aaron Harmon, Chulsoon Hwang

ABSTRACT

This paper shows a method for modeling and understanding an inaudible intentional electromagnetic interference (IEMI) attack on smart speaker devices. This includes a method for finding the ideal attack angle, locating the region sensitive to the coupled EMI, and modeling the attack. In previous works, it was shown to be possible to send RF commands to a smart speaker and have these commands be interpreted as voice commands by the microphone. However, the attack still had some limited understanding in terms of the coupling path location and long-distance attack potential. Using the behavioral modeling methods shown in this paper, a longer attack distance is achieved (6 meters) with only 6.5 Watts of power.

1. INTRODUCTION

A smart speaker is a device that executes commands based on a user's voice. These devices started as simply allowing for basic conversational skills, but later full-function smart home devices began allowing the smart speakers to control functions like

electric plugs, locks, or thermostats. Eventually, this same technology made its way to most smart phones and computers.

In these smart-speaker devices, most utilize microphones with micro-electromechanical system (MEMS) technology [1]. MEMS microphones work like acoustic microphones, but they are active microphones that often have self-contained analog to digital converters and amplifiers. The primary benefit of these microphones is that they have lower power consumption and reduced footprints. These microphones receive an analog voltage signal from the vibrated membrane, which is then amplified and digitized all within the microphone.

This new type of microphone, while having marked benefits, also introduced unintended security implications. Ultrasounds attacks were the first to be discovered [2], followed by laser-based attacks [7]. These worked through having an ultrasound or laser signal carry a modulated audio command, which is then coupled onto the MEMS microphone circuit of the smart speaker. These two attack methods were not effective for the reason because these types of attacks worked through a line-of-sight attack mechanism.

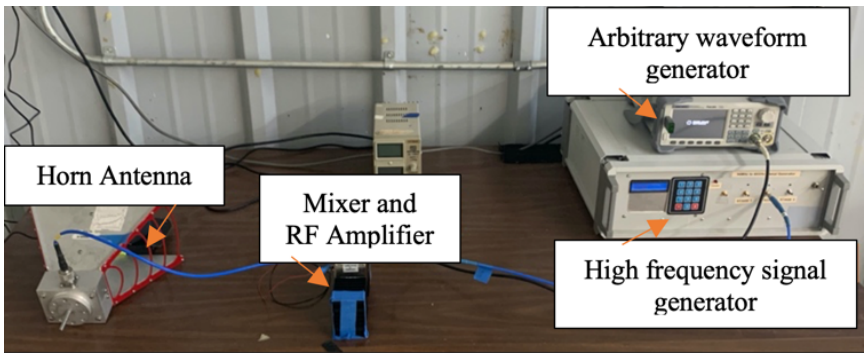
Recently, it was discovered that these commands can be sent through IEMI [1]. This attack is notably more effective than the ultrasound-based attacks in that the attack can be performed through walls. The study of Buzz noise proved that an audio coupling path exists through EMI from the nearby Wi-Fi antenna to the microphone [6]. Based on the understanding of Buzz noise mechanism, the IEMI was firstly demonstrated in [1]. This attack was shown to be possible by modulating the audio range attack signal with a

much higher (6 to 18 GHz) carrier signal that was radiated using a highly directional horn antenna.

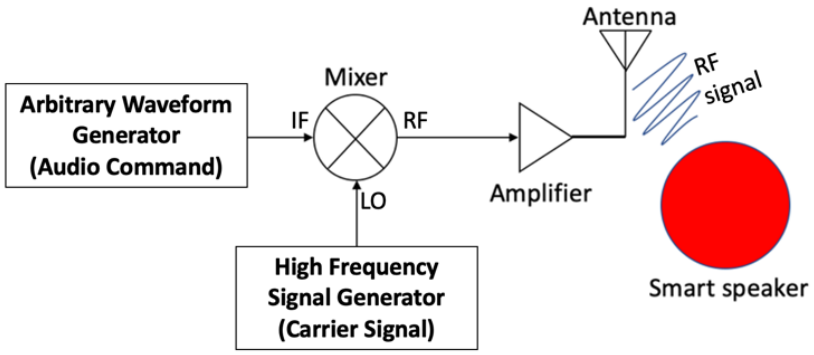
Additionally, it was shown that this attack used in conjunction with machine learning techniques can be used to circumvent a common feature in smart speakers and phones that allows for the device to learn the voice of the speaker and only wake up to that person's voice [10]. In the work shown in [1], a suspected coupling path was proposed, but there was no experimental verification for the suspected most sensitive point. Additionally, the attack range was more limited due to less understanding about ideal attack angles and antenna polarization relative to the smart speaker.

In this work, the previous understanding of the point on the smart speaker that is sensitive to the IEMI attack is re-evaluated and revised in Section 3. After the correct identification of the sensitive point, the sensitive point was validated by creating a simulated model with this sensitive point chosen as the driven port in Section 4. Through simulation and measurement, the most sensitive attack angles to the attack were correctly identified, validating the simulation model. Next, a method for predicting if the IEMI attack occurs for any combination of attack angle, input power at the aggressor antenna, and attack distance was shown and validated using this model in Section 5. These methods were used to greatly increase the attack distance, further showing the necessity for mitigation against this attack. Finally, in Section 6, some potential mitigation strategies against the attack are proposed based on the findings in this paper. These discoveries advance the understanding of the IEMI attack and the ability to model and predict the attack feasibility for any smart speaker. IEMI attack in this paper is described

as the unintended ability (by the designer) to activate a smart speaker's command recognition function through IEMI.



(a)



(b)

Figure 1. (a) The experimental test setup for the finding the most sensitive carrier frequency for the attack (b) Block diagram representation of the test setup.

2. IEMI ATTACK DESCRIPTION AND SIGNAL PROCESSING

The test setup for the IEMI attack is shown in Figure 1a and represented as a block diagram in Figure 1b. An arbitrary waveform generator was used to output the

audio range signal (an audio-range voice command), and a high frequency signal generator outputs the carrier that the audio signal was modulated with.

The audio commands were modulated to the high-frequency carrier by the mixer and then radiated from the attacking antenna to the microphone. The effectiveness of the attack is related to the carrier frequency. The optimal carrier signal was found by sweeping the modulating (carrier) frequency using the test setup shown in Figure 1a with everything else held constant (including the modulated audio). Then, the recorded attack audio amplitude was retrieved from the given device online cloud and compared to the modulating frequency to find the most effective frequency for the attack. The attack carrier frequency itself was found as a transfer function of the sent attack audio amplitude versus the received command amplitude (in the device's cloud service) in [1].

The modulated signal radiates out from the aggressor antenna and then the EM wave is received at the MEMS microphone. From here, the signal is coupled back to the internal amplifier in the MEMS microphone where the non-linearity is present that causes the signal to be demodulated back to the audible range, and then converted to digital in the analog-digital converter (ADC) of the microphone where it can be processed by the microprocessor.

Audio range amplifiers have linear amplification in the audible range when not driven to distortion. However, strong non-linearity can be observed in the non-audible range [15]. The mechanism of the non-linearity for the IEMI attack was found to be associated with the pre-amplifier that is self-contained in the MEMS microphone [4][7]. The output signal that results from amplifier nonlinearity Vis below (1):

$$S_{out} = AS_{in} + BS^2{}_{in} + \cdots DS^4{}_{in} + mS^n{}_{in}. \quad (1)$$

Where S_{out} is the signal that results from the non-linearity effect, S_{in} is the input signal in the IEMI attack, and the alphabetical coefficients *A*, *B*, and *D* represent the amplitude of the resulting coefficients, with each subsequent one decreasing in magnitude. Coefficients *m* and *n* represent infinite order coefficients for this series. Previous work has shown that each subsequent S_n term decreases in magnitude strongly with each iteration [3][6], so only the S^2 term from (1) needs to be considered for this attack. The S^2 term produces both a high and low frequency component. The lower frequency component is less than the cutoff frequency of the low-pass filter of the MEMS microphone after the nonlinearity occurs within the microphone, so only the audio range signal is maintained.

The square term of (1) necessarily produces harmonic distortion, which degrades the quality of voice injected into the system. To minimize the harmonic distortion associated with the demodulation process for the IEMI attack, effective processing method is proposed in (2) as below [1]:

$$S_{in} = \sqrt{Af(t) + A}. \tag{2}$$

A DC offset equal to the maximum amplitude A of the audio waveform is added under the square-root to avoid imaginary part created by the square root function. This DC offset is not an issue as the internal coupling of the mixer removes this component. A poor power supply rejection ratio (PSRR) of the internal amplifier at the attack frequencies is most likely the cause of the IEMI coupling from the outside of the microphone into the amplifier. The PSRR of an amplifier indicates how much of the

change in voltage on the power rail of an amplifier will be translated onto the output. It is well known that the PSRR of an amplifier degrades with frequency [16]. Thus, if a voltage is developing on the microphone can, which is the ground reference for the amplifier within the MEMS microphone, this variation caused by the IEMI attack will induce a voltage at the output of the internal amplifier through this poor PSRR. From here, the harmonic distortion of the internal amplifier will cause the demodulation described by equation (1). The anticipated coupling mechanism that results from the poor PSRR is shown in Figure 3.

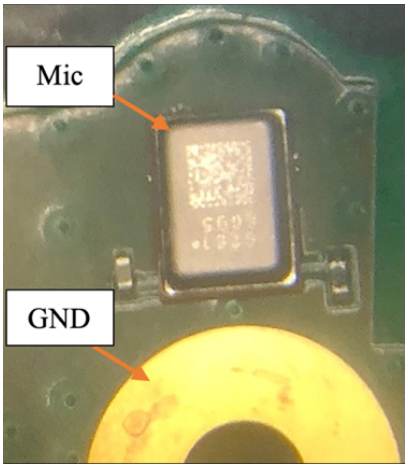


Figure 2. Microphone can and PCB return plane connections for the direct injection of the IEMI attack waveform on smart speaker 1.

3. IDENTIFICATION OF THE MOST SENSITIVE POINT

Previously in [1], the sensitive point of a smart speaker (which is referred to as smart speaker 1) was identified using near field scanning. Specifically, the measurement indicated that the capacitive volume sensor was sensitive at the same frequency at which

the attack was most effective. This speaker has large nearby metal structures close to the microphone in the enclosure that are thought to enhance the coupling.

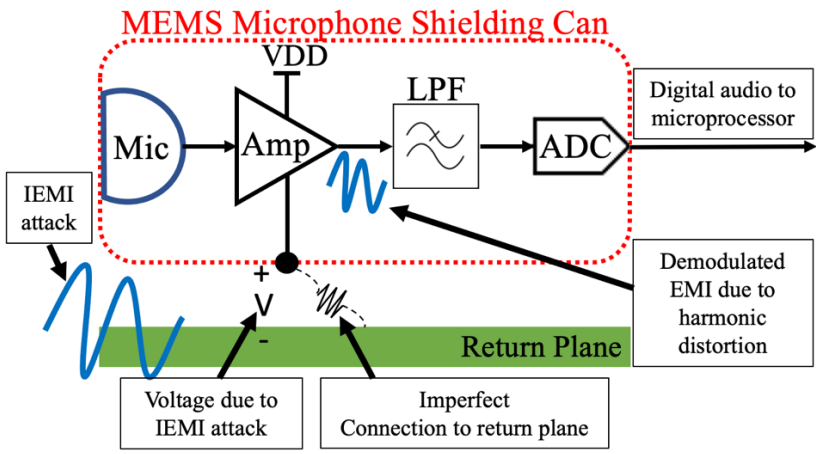


Figure 3. Anticipated coupling mechanism and block diagram for the smart speaker IEMI attack.

To verify this finding in this work, a coaxial cable with an SMA connector was directly soldered to the capacitive switch on the PCB. After this coax-SMA was soldered to the PCB, the setup shown in Figure 1b was used to directly inject the command known to wake the device into the smart speaker. The command was the simply the 'wake word' for the device with the processing described in (2). When injecting audio to the point that was predicted to be susceptible at the 5 GHz sensitive frequency, however, no audio was heard, and the device would not interpret any direct-injected commands.

Given that the physical injection of the command should be equivalent to receiving the waveform through a radiated mechanism, this indicates that the previous understanding of the most sensitive point for the microphone could be incorrect. Thus, alternative techniques were used to find the true most sensitive point.

To start, copper tape was added to various potential coupling points on the smart speaker. The rational for this was that the added copper tape could serve as a more effective coupling structure to increase the coupling interference from the antenna to the microphone. It was observed that the audio amplitude would increase when the copper tape was added to the top of the microphone can structure, but not when it was added to the previously determined sensitive point. Thus, the direct command injection was performed once again onto the top of the microphone can as shown in Figure 2.

After the modification, the played back audio had much higher amplitude during initial testing for a 5 GHz carrier signal that the injected audio was far louder than the received amplitude purely from the normal microphone operation. This result indicated that the true sensitive point was between the top can of the microphone and the PCB return plane of the smart speaker. However, this result only showed that the sensitive point was located on the top of the microphone for this smart speaker specifically.

To see if this sensitive point is the same for other speakers, a second smart speaker was found by a separate manufacturer that does not have any obvious vulnerability to the IEMI attack. Additionally, the original smart speaker (smart speaker 1) also had monopoles added to the microphone can structure, while removing any nearby metal structures that enhance coupling. From here, monopoles were cut to a quarter-wavelength of the carrier frequency so that they resonate at 5 GHz (which was identified as a sensitive carrier frequency on other speakers). These monopoles were soldered to the top of the microphone can as shown in Figure 4.

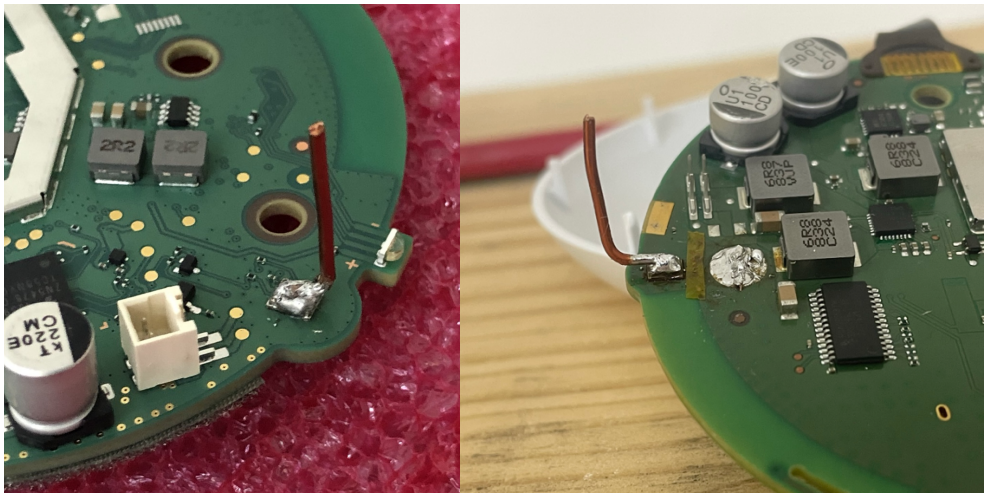


Figure 4. Adding monopole structures to the top of the two smart speaker's microphone can to try and cause the IEMI attack on a non-susceptible device.

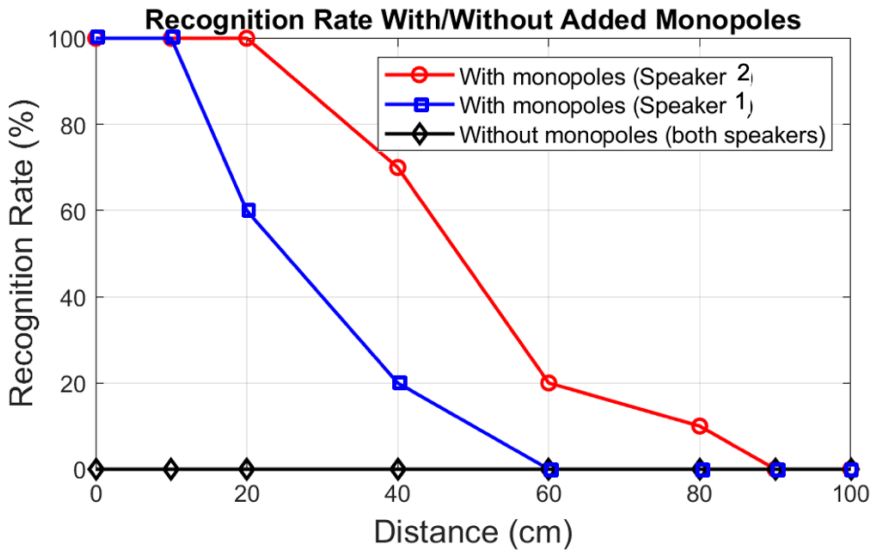


Figure 5. Result of adding monopoles to the previously unsusceptible device.

While these devices previously showed no recognition at all (indicated by diamonds in Figure 5), the structure with four monopoles added has command recognition out to 1 meter for the second smart speaker (indicated by circles), and 40 centimeters for the first smart speaker (indicated by squares). The videos relating to this

experiment are available in [8]. From this experiment of adding monopole structures to the previously unsusceptible smart speakers, it was clear to see that the microphone can structure itself is the point of sensitivity for the IEMI attack. Additionally, this experiment and the observations seen by adding copper tape show that coupling is enhanced in the smart speaker by metal structures that are nearby to the microphone can.

4. IDEAL ATTACK ANGLE ANALYSIS

4.1. REASONING AND EXPLANATION FOR ATTACK ANGLE PREDICTION

During testing of the smart speaker IEMI attack, it was observed experimentally that certain ‘attack angles’, meaning where the aggressor antenna was pointing in relation to the smart speaker, were more effective at causing the smart speaker IEMI attack. If these optimal angles for attacking could be determined through simulation techniques, it would save considerable time. The smart speaker has two microphones, so there are two susceptible points on the smart speaker based on the analysis in part III. To reduce complexity, one MEMS microphone was removed so the radiation pattern of one microphone can be measured.

The experimental setup for measuring the radiation pattern of this unintentional antenna is not straight-forward, as any augmentation of the most sensitive point, located on the can of the microphone, would change the radiation pattern of unintentional antenna. The chosen solution was to send the attack waveform with a single audio-range tone modulated to the sensitive frequency for this device, while sweeping the attack angles for this device. After radiating this attack waveform, the received audio amplitude

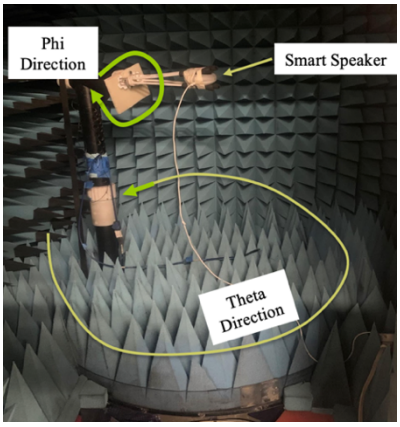
at the smart speaker can be plotted as a function of the attack angle to determine the most effective attack angles.

4.2. MEASURING SMART SPEAKER UNINTENTIONAL ANTENNA RADIATION PATTERN

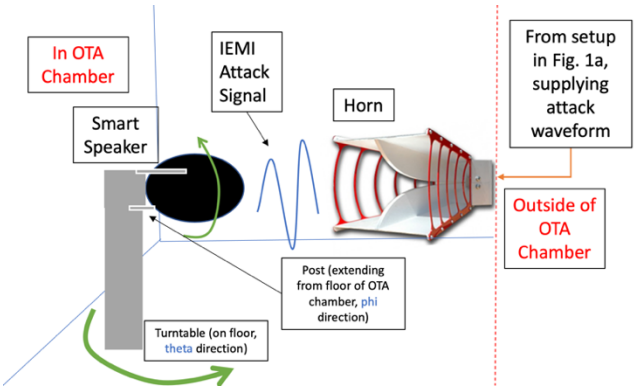
The smart speaker was placed in an over the air (OTA) chamber that can sweep both theta and phi angles 360 degrees. In this measurement, the turntable on the floor of the OTA chamber was chosen as the theta coordinate, while the part that spins the DUT itself was chosen as the phi coordinate.

The coordinate system was developed in relation to the electric field vector position of attack antenna (polarization). For this setup, the E-field vector of the horn antenna was aligned with the theta axis. Figure 6a shows the visual representation of this coordinate system with the arrow direction showing theta/phi spin directions for full 360-degree angle sweeping. In Figure 6b, a diagram representing the setup is shown for improved clarity. In Figure 6c, the coordinate system is drawn with reference to the smart speaker itself.

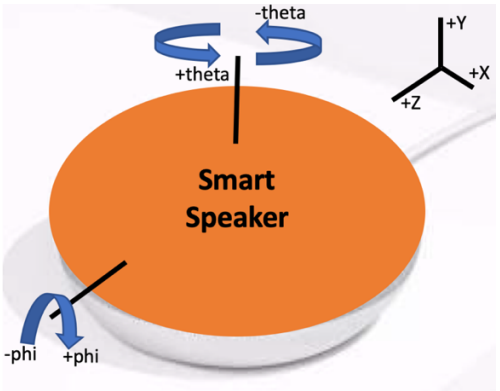
As in the simulation case, angle cuts for theta = 0° and 30° were measured. Data was recorded using an automated script that outputs a 1 kHz tone through the headphone connector of the computer. A 1 kHz tone was chosen because this tone is within the human speaking voice range, and command recognition was not the focus of the sensitive angle finding. The smart speaker is setup to continuously record audio during this test.



(a)



(b)



(c)

Figure 6. (a) Coordinate system for this radiation pattern measurement test (b) System diagram for the measurement of the smart speaker radiation pattern for the unintentional antenna within the smart speaker allowing for the IEMI attack (c) Coordinate system relative to the smart speaker itself.

4.3. MODELING SMART SPEAKER IN FULL-WAVE SIMULATION

The structure was first modeled in a full-wave simulation tool. Given that the internal stack-up of the smart speaker PCB is not known, only the large metal features of the smart speaker are modeled. This includes the metal shield that is placed over the microphone, the PCB ground, and PCB dielectric. Additionally, the Z11 of the loading for the sensitive point was included. This Z11 was measured using the same port connection shown in Figure 2. In Figure 7, the resulting model can be seen.

To get proper comparison results between the simulation and measurement, the coordinate systems between the measurement and simulation should be matched. Additionally, the polarization of the electric field vector in the simulation was chosen so it was aligned with the theta direction, as it was in the measurement case.

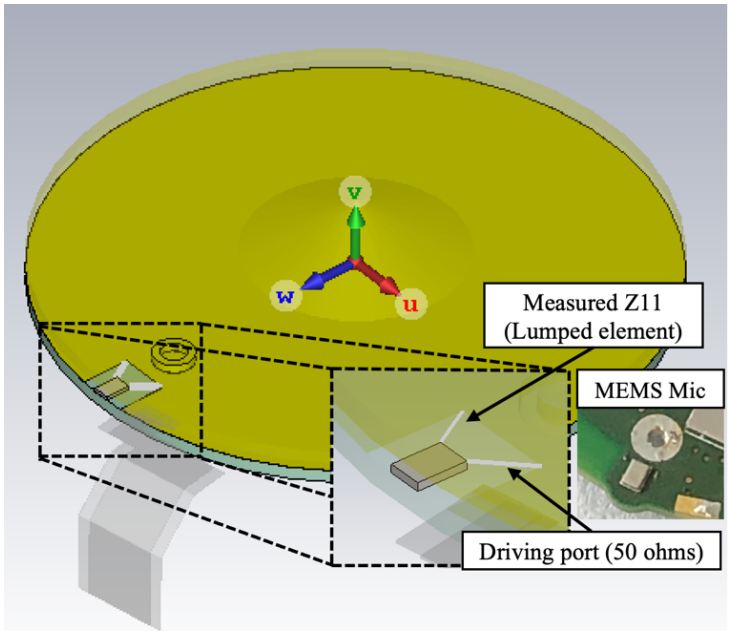


Figure 7. CST Model of the internal structure of the smart speaker.

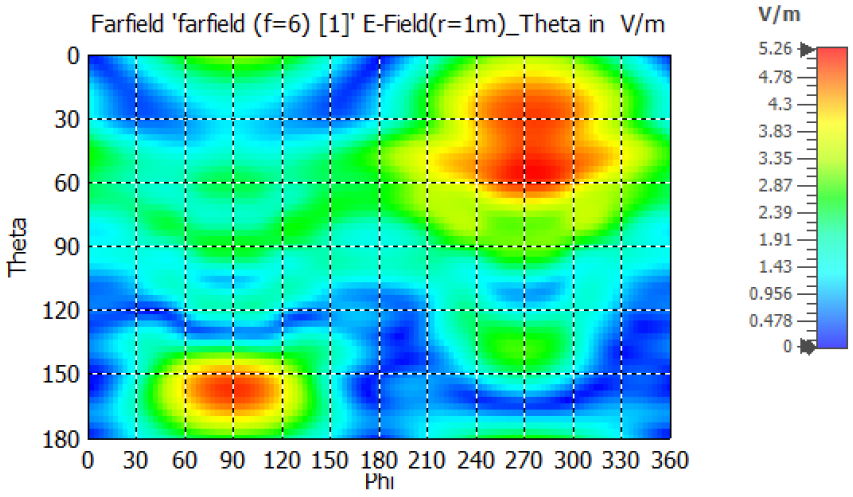


Figure 8. Directivity of the modeled smart speaker for electric field vector oriented in the theta direction (polarization).

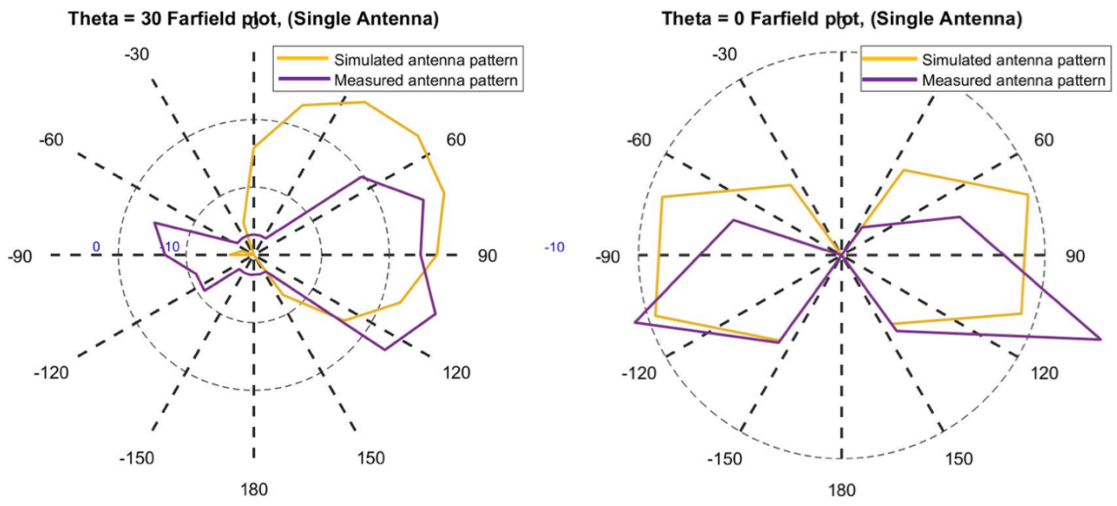
As seen in Figure 7, the driven port for the simulation was placed between the driving port and PCB return plane. Ideally, this Z11 should represent the imperfect connection between the microphone can and PCB return plane at higher frequencies that causes a noise voltage. This Z11 was measured at the sensitive point shown in Figure 2 and inserted into the model as a lumped element to add more physicality to the model. The port impedance for the driving port is selected as 50 ohms. Choosing 50-ohms as the port impedance will affect the unintentional antenna efficiency due to reflections, but not the pattern itself, which is the focus for the attack angle finding.

When examining the directivity pattern in Figure 8 (swept across theta and phi), areas of higher directivity are observed. At these combinations of theta/phi, the highest amplitude should be observed. The hotspots of high directivity are located where theta = 180 degrees/phi = 90 degrees, and theta = 30/ phi = 270 degrees.

4.4. COMPARISON BETWEEN SIMULATION AND MEASUREMENT

In this part, the directivity results between the measurement case and the simulation case are directly compared. The measured data was normalized to the simulation data because the only point of interest is where the amplitude was maximum, rather than the value of the amplitude itself.

These results are seen in Figure 9. From Figure 9a, there is a clear alignment between the simulation results and the measured results. The results were not perfectly aligned for this angle cut, but some variation is expected with this test because the measurement is carried out by simply plotting the received audio amplitudes and then normalizing them to the level of the simulation results.



(a) (b)

Figure 9 (a) Simulated versus Measured Directivity for the theta = 30 angle cut (in dB) (b) Simulated versus Measured Directivity for the theta = 0 angle cut (in dB)

For Figure 9b, there is better alignment. At phi=90 and phi=-90 degrees, the simulation and measurement results are relatively well aligned. These two results further support the finding that the sensitive point is located at the microphone can structure, and that the directivity for the unintentional antenna can be obtained without knowing anything about the PCB of the smart speaker itself.

5. PREDICTING IEMI ATTACK RECOGNITION AND LONG DISTANCE ATTACKING

In this Section, an example and procedure for predicting the success of the IEMI attack for any combination of attack angle, power, carrier frequency, or distance is described and tested. The end goal is to predict whether the IEMI attack successfully triggers command recognition for any smart speaker and attack situation without doing the IEMI attack for that scenario.

5.1. SIMULATION OF ELECTRIC FIELD FOR FIXED DISTANCE, POWER, AND ANGLE

For the simulation part, the model from Section 4 was reused. This simulation was instead used in-conjunction with a field probe placed at a distance/angle configuration that corresponded to an actual test with the IEMI attack with these same parameters for comparison and verification purposes. From here, the received electric field magnitude at the field probe in the simulation was plotted as a function of frequency. These results are seen in Figure 10.

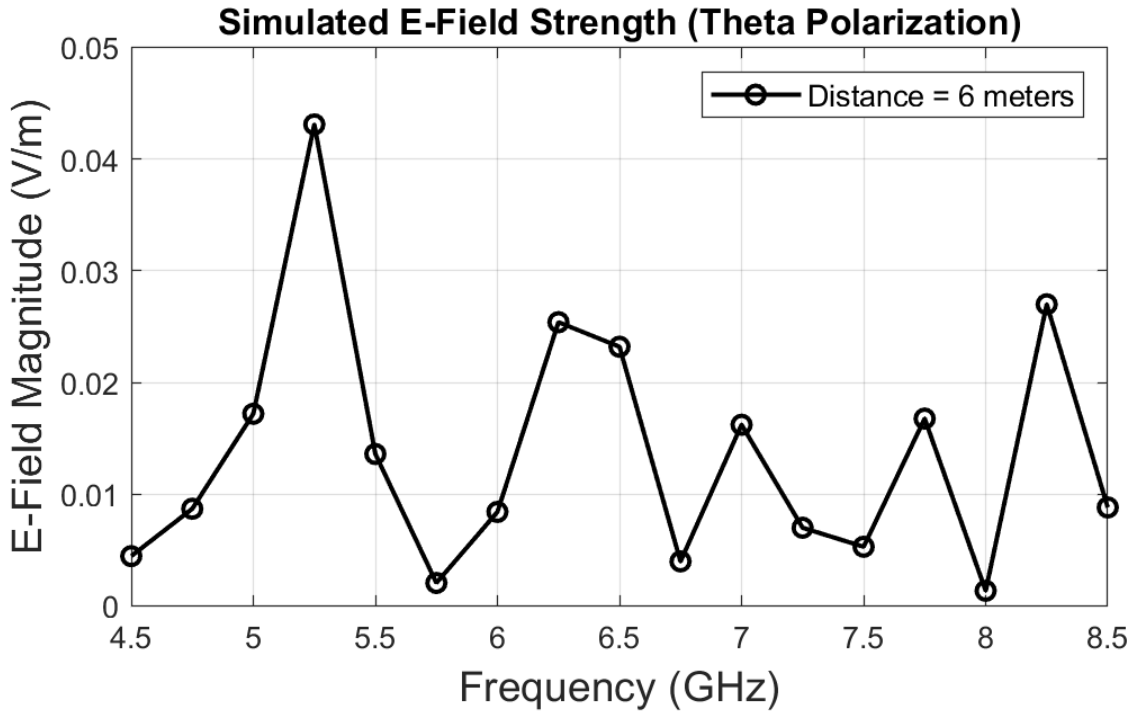


Figure 10. Simulated electric field intensity versus frequency for phi= 10 degrees, radius = 20 feet.

5.2. DERIVATION OF COUPLED VOLTAGE AT SENSITIVE POINT

Since the point that is sensitive to the IEMI attack was directly driven in the CST simulation, the reciprocity theorem can be utilized to get the coupled voltage at the sensitive point of the smart speaker.

The derivation starts with the definition for radiation intensity (U) in terms of theta angle (θ), phi angle (ϕ), the distance from the speaker to the aggressor antenna (r), impedance of free space (η), and the electric field values found in Figure 10 for the θ, ϕ and r ($E_{speaker}$) in (3) [14].

$$U(\theta,\phi) = r^2 \mathrm{x} \frac{{E_{speaker}}^2}{2\eta} \left(\frac{\mathrm{W}}{\Omega}\right) \tag{3}$$

The gain of an antenna is the basis for the next part of the derivation [12]. The radiation intensity (3) is substituted in for *U* in the equation below, where $P_{in,sim}$ is the total input power to the antenna in the simulation, and $G_{speaker}$ is the gain of the smart speaker structure (unitless):

$$G_{speaker} = 4\pi\ x\ \frac{U(\theta,\phi)}{P_{in,sim}} = \frac{4\pi\ x\ r^2 x \frac{{E_{speaker}}^2}{2\eta}}{P_{in,sim}}\ (unitless) \quad (4)$$

For the next part, the antenna factor equation [18] in terms of gain, where $Z_{in,sim}$ is the impedance of the driving port of the antenna structure in simulation, and λ is the wavelength of the aggressor carrier signal (across frequency) was used. $AF_{speaker}$ represents the antenna factor of the smart speaker's unintentional antenna, so the $G_{speaker}$ is used for this antenna factor:

$$AF_{speaker} = \frac{E-field}{Volt} = \sqrt{\frac{4\ x\ \pi\ x\ \eta}{\lambda^2\ x\ G_{speaker}\ x\ Z_{in,sim}}}\ \ (m^{-1}) \quad (5)$$

Then, (4) was substituted into equation (5) for the $G_{speaker}$ variable, and $AF_{speaker}$ is simplified:

$$AF_{speaker} = \sqrt{\frac{2\ x\ \eta^2\ x\ P_{in,sim}}{\lambda^2\ x\ r^2\ x\ Z_{in,sim}\ x\ {E_{speaker}}^2}}\ \ (m^{-1}) \quad (6)$$

From here, the equation for the E-field strength as a result of the aggressor horn antenna that sends the IEMI attack signal (in the physical test setup) is shown, where V_{horn} is the voltage at the horn antenna port, AF_{horn} is the known antenna factor of the horn antenna at the same distance as radius r, and E_{horn} is the electric field as a result of the aggressor antenna at radius r [13]:

$$E_{horn} = V_{horn}\ x\ AF_{horn}\ \ (\frac{V}{m}) \tag{7}$$

Finally, the coupled voltage at the sensitive point of the smart speaker is derived using the calculated electric field strength at the smart speaker ($V_{coupled}$):

$$V_{coupled}\,(r) = E_{horn}\ x\ \frac{1}{AF_{speaker}}\quad (\mathrm{V}) \tag{8}$$

5.3. MINIMUM NEEDED VOLTAGE FOR COMMAND RECOGNITION

Measuring the minimum needed voltage for command recognition versus frequency relationship for a smart speaker requires two connections onto the sensitive point: One for direct injection at the most sensitive point of the smart speaker, and another for measuring the resulting magnitude. For this measurement, at each carrier frequency, the IEMI attack waveform was injected at the sensitive point at an amplitude at the minimum required amplitude to cause the attack to occur. Then, at the measurement connection (connected to an oscilloscope), the voltage at this most sensitive point can be recorded and marked as a point for the minimum needed voltage for command recognition versus frequency relationship.

To find the minimum needed voltage for command recognition, a variable attenuator was adjusted, started at maximum attenuation, and lowered until the smart speaker wakes 100% of the time while the command known to wake the device was constantly sent. This setup is shown in Figure 11. To calculate the minimum voltage relation, the carrier frequency on the LO of the mixer in this setup was swept from 4.5 GHz to 8.5 GHz.

After measuring the minimum required voltage relation using the setup in Figure 11, the calculated voltage at most sensitive point of the smart speaker for a 6.5W input into a parabolic dish (aggressor antenna) with a known antenna factor, where the angles are theta=90 degrees, phi=10 degrees, and the distance was 20 feet was found using the equation in (8). The result of plotting both the minimum required voltage and predicted voltage at the most sensitive point at the same time is shown in Figure 12.

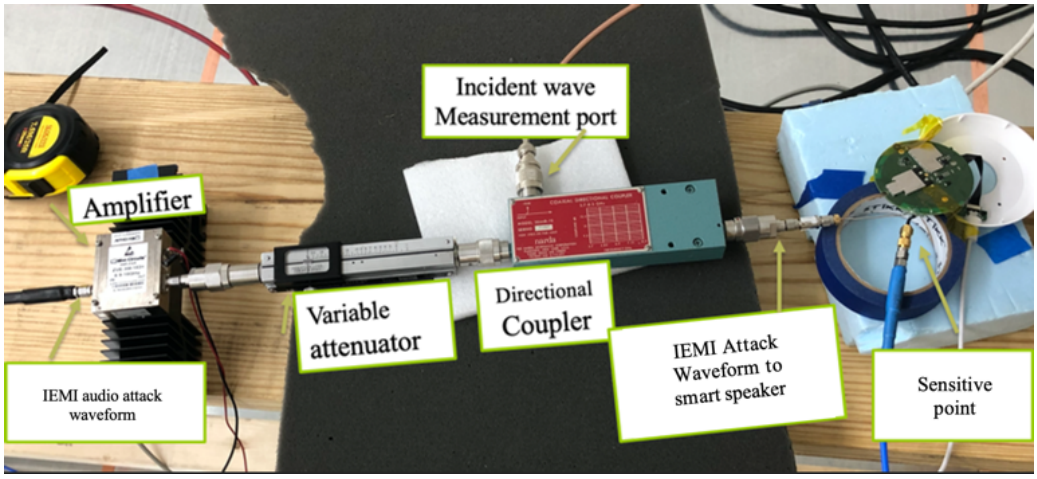


Figure 11. Setup for measuring the minimum voltage needed to cause the IEMI attack for a smart speaker.

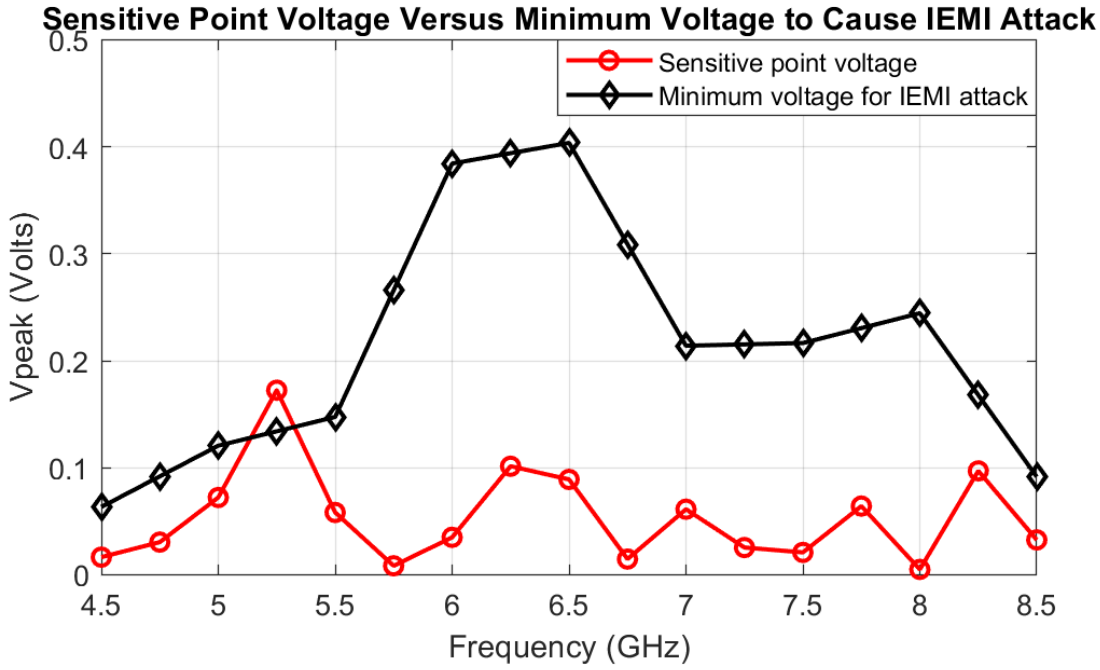


Figure 12. Minimum required voltage for command recognition versus predicted voltage at the sensitive point based on a given theta, phi, distance, and input power for a parabolic dish aggressor

In Figure 12, the calculated predicted voltage at the sensitive point exceeds the minimum voltage required for command recognition between 5 GHz and 5.5 GHz, meaning that the sensitivity to the attack is highest at these frequencies. The smart speaker was observed to wake at these frequencies previously when sweeping the attack waveform carrier frequency manually for the radiated attack.

For a clearer view for what frequency causes the IEMI attack to happen, the simulated voltage at the antenna was divided by the susceptibility curve and plotted again. For the plot in Figure 13, any carrier frequency of the curve that exceeds 1 will cause the IEMI attack. From this plot, it is clearer to see those frequencies between 5 and 5.5 GHz are the most sensitive frequencies predicted by this calculation.

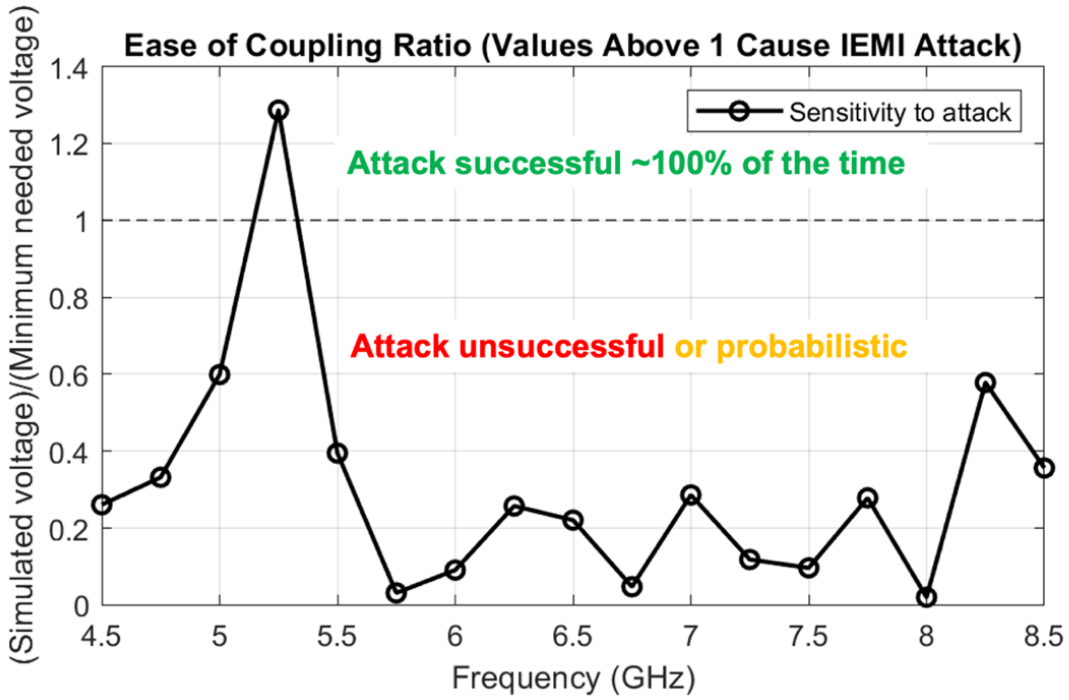


Figure 13. Predicted voltage at the sensitive point divided by minimum required voltage for command recognition. Values greater than 1 mean the IEMI attack is possible at the frequency.

This alignment with the expected most sensitive frequencies provides supporting evidence for this technique. However, the method can be further scrutinized by attacking with a 'real world' IEMI attack with the same parameters that were used for the CST simulation.

5.4. LONG DISTANCE ATTACKING

The IEMI attack was optimized to extend the range out to 6 meters by using the optimal attack angles shown in Section 4 of this work. Theta, in this case, was chosen to be 10 degrees. 30 degrees would be more optimal, but this was not possible with the available test space. This results in an angle that points towards the top of the smart

speaker. This choice was based on the observation that increasing theta from 0 to 30 degrees resulted in increased directivity of the attack for a phi angle equal to 90 degrees. Subsequently, 90 degrees (for the shown coordinate system) was chosen for phi. This 90-degree attack angle means physically that the E-field vector of the parabolic dish must be perpendicular to the smart speaker.

The parameters for the test are the same as the simulated CST scenario for finding the predicted voltage at the sensitive point in part C. 6.5 Watts was the input power to the parabolic dish, same angle positioning, and the smart speaker placed 6 meters away from the parabolic dish. This diagram is shown in Figure 14.

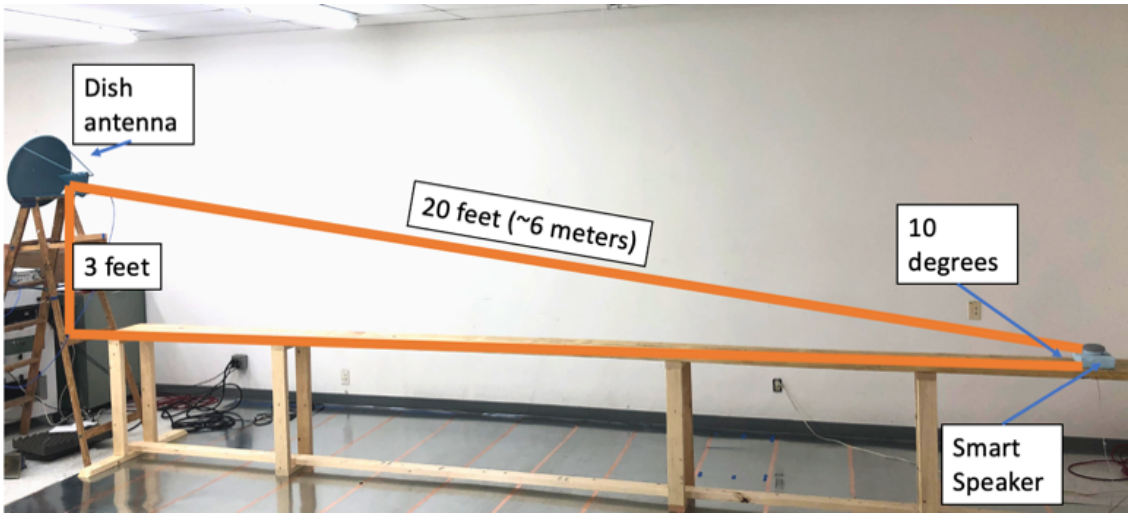


Figure 14. Distance diagram for the 20-foot attack verification

To help verify the results of the prediction shown in Figure 13, the carrier frequency was swept from 4.5 GHz to 8.5 GHz (measured every 500 MHz) for the setup shown in Figure 14. In Figure 15, the results of this prediction are shown, where the probability of successful attack recognition is indicated by the red bars.

This test shows the feasibility of the attack at ranges that exceed the distances (~1.5 meters) shown in the previous work [1]. Additionally, the smart speaker wakes at most sensitive frequency predicted by the calculation, and unsuccessful or probabilistic for the other frequencies in this tested range. Based on this validation, this method can predict whether the attack successfully causes command execution at this distance, angle, and power combination. A video was taken of the IEMI attack for the ~5 GHz carrier frequency case. The video of this test can be viewed at [9].

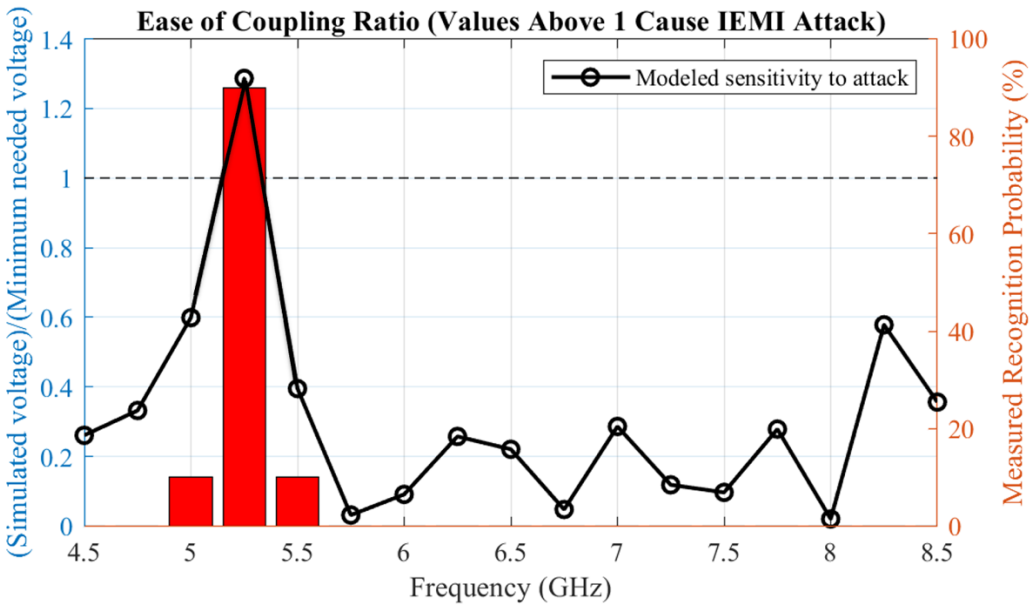


Figure 15. Predicted voltage at the sensitive point divided by minimum required voltage for command recognition. Command recognition rate for the equivalent test is super-imposed onto this ease of coupling ratio to validate the prediction.

6. CONCLUSION AND MITIGATION DISCUSSION

In this work, the methods for understanding the behavior of the smart speaker IEMI attack were presented and tested. A method for determining the sensitive point on

the smart speaker was shown after the previously found most sensitive point was found to be insensitive to the IEMI attack. After the correct sensitive point was identified, the most effective attack angle was determined through simulation methods and subsequently compared with the measured most sensitive angles for alignment. Next, by calculating the voltage at the most sensitive point, the effectiveness of the smart speaker IEMI attack was shown to be possible for a set distance, attack angle, and input power relative to the aggressor antenna. Finally, a non-simulation experiment was used for a long-distance attack to not only show the feasibility of longer-distance attacks, but also to verify the calculation and simulation correctly predicted that the smart speaker would be vulnerable to the IEMI attack for the attack parameters.

A test like IEC 61000-4-3 could be devised in the future to evaluate a smart speaker's immunity to the IEMI attack. In 61000-4-3, radiated immunity is evaluated by using sweeping the frequency at a broadband antenna that is pointed at the device under test (DUT). Using the antenna factor and input power to the device, the electric field strength at the DUT can be found, or through field probe measurements. In the test plan, an electric field value is targeted as a baseline for what the DUT should be able to pass 61000-4-3. To create test plan for smart speakers from the framework of IEC 61000-4-3, the measurement of the minimum voltage for command recognition shown in Section 5 part C of this work could be used. By measuring and averaging this minimum voltage across many smart speakers and converting to the equivalent electric field, a threshold for strong immunity to the IEMI attack could be established. From here, the 61000-4-3 test could be carried out for smart speakers by sweeping frequency into the antenna at this electric field and modulating the wake command with the swept frequency.

In the future, efforts should be made to mitigate this attack to prevent any security concerns associated with it. Based on the findings in Section 3, nearby metal structures located on or nearby the microphone structure can significantly enhance the effectiveness of the IEMI attack. To reduce this coupling, large floating metal structures, if they must be used, should be placed far away from the MEMS microphone to avoid direct capacitive coupling to the microphone.

Additionally, the high frequency connection from the microphone can structure to the PCB return plane of the smart speaker should be improved. Based on the prior measurement of the Z11 from the smart speaker to PCB return plane, the most sensitive frequencies were observed to happen where the impedance from the PCB return plane to microphone can structure was elevated. By improving this high frequency connection, the attack can be mitigated.

Finally, due the role of the internal amplifier's non-linearities in the attack, efforts should be taken to improve the PSRR of this internal amplifier within the microphone to reduce the amount of noise that is coupled onto the microphone can structure that ends up in the output of the amplifier. This would entail characterizing how much power rail noise is coupled onto the output of this internal amplifier during the design process.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under Grant No. IIP-1916535.

REFERENCES

[1] Z. Xu, R. Hua, J. Juang, S. Xia, J. Fan and C. Hwang, "Inaudible Attack on Smart Speakers With Intentional Electromagnetic Interference," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 5, pp. 2642-2650, May 2021.

[2] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., and Xu, W.,"Dolphinattack: Inaudible Voice Commands." *Proceedings of the 2017 ACM SIGSAC. ACM, 2017*

[3] Roy, N., Shen, S., Hassanieh, H., and Choudhury, R. R., "Inaudible Voice Commands: The Long-range Attack and Defense." *15th USENIX. USENIX, 2018.*

[4] J. Mao, S. Zhu, X.Dai, Q. Lin and J. Liu, "Watchdog: Detecting Ultrasonic-Based Inaudible Voice Attacks to Smart Home Systems," *IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8025-8035, Sept. 2020*

[5] J. Gago, J. Balcells, D. GonzÁlez, M. Lamich, J. Mon and A. Santolaria, "EMI Susceptibility Model of Signal Conditioning Circuits Based on Operational Amplifiers," *IEEE Transactions on Electromagnetic Compatibility,* vol. 49, no. 4, pp. 849-859, Nov. 2007.

[6] Y. Zhong, Q. Huang, T. Enomoto, S. Seto, K. Araki, and C. Hwang, "Measurement Based Characterization of Buzz Noise in Wireless Devices." *IEEE Int. Symp. On Electromagnetic Compatibility*, Long Beach, CA, pp. 134-138, 2018.

[7] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. "Light commands: Laser-based Audio Injection Attacks on Voice-Controllable Systems". *Proceedings of the 29th USENIX Conference on Security Symposium*. USENIX Association, USA, Article 148, 2631–2648.

[8] EMC-Laboratory. *Hardware-security/smart speaker I-EMI/videos/monopole-adding videos at main · EMC-Laboratory/Hardware-Security*. GitHub. Retrieved March 26, 2022, from https://github.com/EMC-Laboratory/Hardware-Security/tree/main/Smart%20Speaker%20I-EMI/Videos/Monopole-adding%20videos

[9] EMC-Laboratory. *Hardware-security/smart speaker I-EMI/videos/monopole-adding videos at main · EMC-Laboratory/Hardware-Security*. GitHub. Retrieved March 26, 2022, from https://github.com/EMC-Laboratory/Hardware-Security/blob/main/Smart%20Speaker%20I-EMI/Videos/6%20Meter%20attack.mp4

[10] T. Fokkens, Z. Xu, O. Hoseini Izadi and C. Hwang, "Machine Learning Voice Synthesis for Intention Electromagnetic Interference Injection in Smart Speaker Devices," *2021 IEEE International Joint EMC/SI/PI and EMC Europe Symposium*, 2021, pp. 673-677.

[11] John A. Polo, Tom G. Mackay, Akhlesh Lakhtakia, *Electromagnetic Surface Waves*, Elsevier, 2013, Pages 81-125.

[12] Dragan Poljak, Mario Cvetković, *Human Interaction with Electromagnetic Fields, Academic Press*, 2019, Pages 21-52.

[13] McLean, James, Robert Sutton, and Rob Hoffman. "Interpreting Antenna Performance Parameters for EMC Applications: Part 3: Antenna Factor." *TDK RF Solutions Inc* (2004).

[14] Balanis, Constantine A. *Antenna Theory: Analysis and Design*. 3rd ed. John Wiley, 2005, Pages 37-38.

[15] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. “BackDoor: Making Microphones Hear Inaudible Sounds”, *15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*.

[16] “Op Amp Power Supply Rejection Ratio (PSRR) and Supply Voltages”, Analog Devices, from https://www.analog.com/media/en/training-seminars/tutorials/mt-043.pdf

II. MACHINE LEARNING VOICE SYNTHESIS FOR INTENTION ELECTROMAGNETIC INTERFERENCE INJECTION IN SMART SPEAKER DEVICES

Tanner Fokkens, Zhifei Xu, Omid Hoseini Izadi, Chulsoon Hwang

ABSTRACT

This work presents the effectiveness of using machine learning (ML) synthesized voice samples to control smart speaker devices through radiated intentional electromagnetic interference (I-EMI). In previous works, the feasibility of using I- EMI to control smart speaker devices was shown. However, devices that are trained to only recognize a single person's voice or only execute certain commands from that person will not be as susceptible to this attack. By training a generative adversarial network (GAN) using samples of the target's voice, this security feature can be bypassed directly, increasing the feasibility of the attack.

1. INTRODUCTION

Smart speakers are Internet of things (IoT) devices that can actively listen for spoken word commands and then execute these commands by processing the voice information on a cloud server. This same technology is also utilized in cell phones, tablets, and most other internet connected devices.

Due to size and power constraints, micro-electromechanical system (MEMS) microphones are utilized in these devices [1]. In a conventional condenser microphone, the audio signal is generated by vibrating the diaphragm of the microphone to create a

time varying analog voltage. MEMS microphones work through similar principles; however, the analog voltage signal is internally amplified and converted to a digital signal using an analog to digital converter (ADC). From this point, the digital signal is sent to the microprocessor so the voice signal can be interpreted.

This microphone technology has provided convenience for new designs, but it has also left smart speakers and phones vulnerable to potential susceptibilities. The first of these susceptibilities allows for an attack that utilizes ultrasound to inject inaudible commands into these devices [2]-[8]. An ultrasound signal sent to the microphone can be demodulated internally in the MEMS microphone through an observed non- linearity. This attack has drawbacks, however. Since the attack works through ultrasound, the feasibility of through-wall attacks is limited due to sound dampening associated with building construction. This limits the ultrasound attack almost purely to short-range line of sight cases.

Given the possibility of the ultrasound attack, an intentional electromagnetic interference (I-EMI) attack using the non- linearity of microphones also showed promise [1]. One of the primary advantages of this I-EMI attack versus the ultrasound attack is that through-walls attacks are feasible due to the properties of the electromagnetic (EM) wave. This new method expands the use case of the known non-linearity in the MEMS microphone, increasing the attack feasibility. However, despite the improvements in feasibility that the I-EMI attack method makes possible, many of these smart speaker devices have a feature that will only recognize the owner's voice or only execute limited commands, rendering the attack useless for an attacker other than the owner themselves.

This was an issue that was brought up as a potential point of investigation in the previous work [1]. Previous papers related to the ultrasound attack also investigated possible circumvention methods for voice personalization, but they were only investigated from the standpoint of a smart speaker trained on the voice of a synthesizer, and then using this synthesizer to circumvent the voice personalization feature [8]. This work also differs from [14], which utilized the headphone wire connected to a smart phone to carry out the I-EMI attack. This method does not rely on the headphone cable or wiring since its primary coupling mechanism is the MEMS microphone. Because of this new method, the coupling can be carried out on devices that do not have headphone input connections.

This paper investigates circumventing the voice personalization feature in smart speakers trained on a human voice to expand the attack to a wider range of devices. The I- EMI method was used for execution of the attack. Machine learning methods were utilized to generate the wake commands for controlling the smart speakers by training the model with the owner's vocal samples. The recognition rate between the natural voice and artificial voice was compared to determine the effectiveness of the machine learning methods.

2. I-EMI ATTACK MECHANISM AND SET-UP

The I-EMI attack was motivated by audio interference from Wi-Fi transmissions called Buzz noise [13]. [13] proved that an audio coupling path exists through EMI to the microphone, and the I-EMI attack was firstly demonstrated in [1]. This attack was shown

to be possible by modulating the audio range attack signal with a much higher (6 to 18 GHz) carrier signal that was radiated using a highly directional horn antenna [1].

The carrier signal was found by sweeping the modulating (carrier) frequency using the test setup shown in Figure 5 with everything else held constant (including the modulated audio). Then, the recorded attack audio amplitude was retrieved from the given device online cloud and compared to the modulating frequency to find the most effective frequency for the attack. This carrier frequency transfer function is shown in Figure 1.

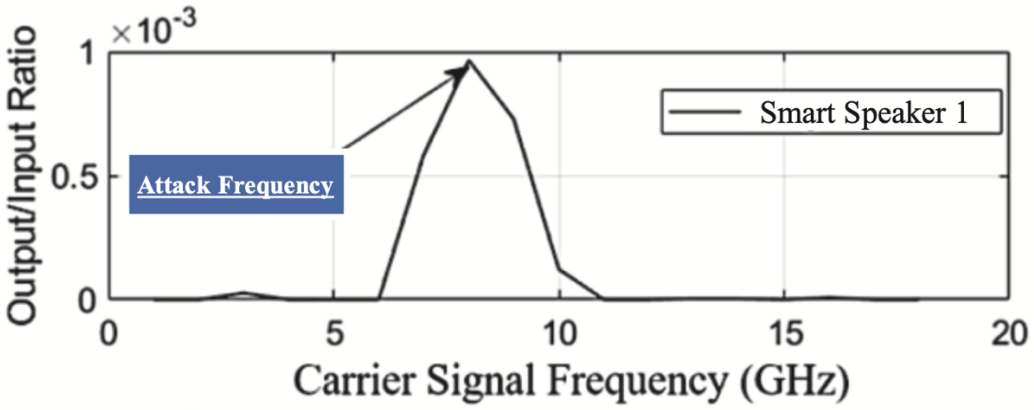


Figure 1. Carrier signal Frequency sweep for finding the highest susceptibility for smart speaker 1.

The modulated voice command is received by the traces of the system. From here, the signal is coupled back to the microphone where the non-linearity is present that causes the signal to be demodulated back to the audible range, and then converted to digital where it can be processed by the microprocessor. The anticipated coupling path for the I-EMI attack is shown in Figure 2.

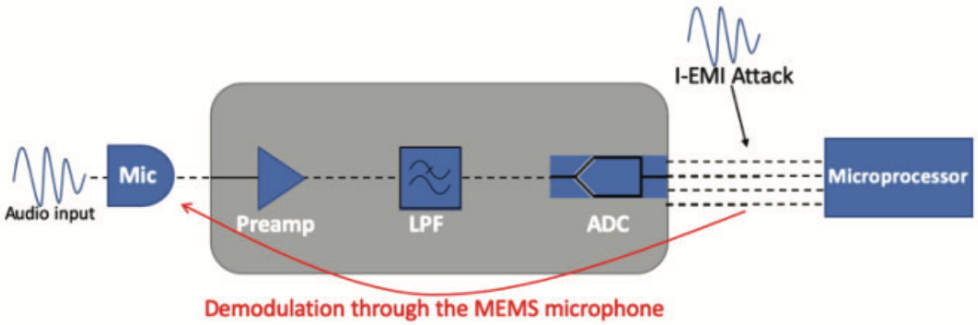


Figure 2. Diagram of anticipated coupling method reinterpreted from [1].

The mechanism of the non-linearity was found to be associated with the internal amplifier of the microphone [8][12]. The demodulated audio-range signal that results from the non-linearity is described by (1):

$$S_{out} = AS_{in} + BS^2{}_{in} + \cdots dS^4{}_{in} + mS^n{}_{in}, \quad (1)$$

Previous work has shown that each subsequent S^n term decreases in magnitude strongly with each iteration [3], so only the s^2 term from (1) needs to be considered for this attack. The s^2 term produces both a high and low frequency component. The lower frequency component is less than the cutoff frequency of the low-pass filter (LPF) of the MEMS microphone, so only the audio range signal is maintained as shown in Figure 3.

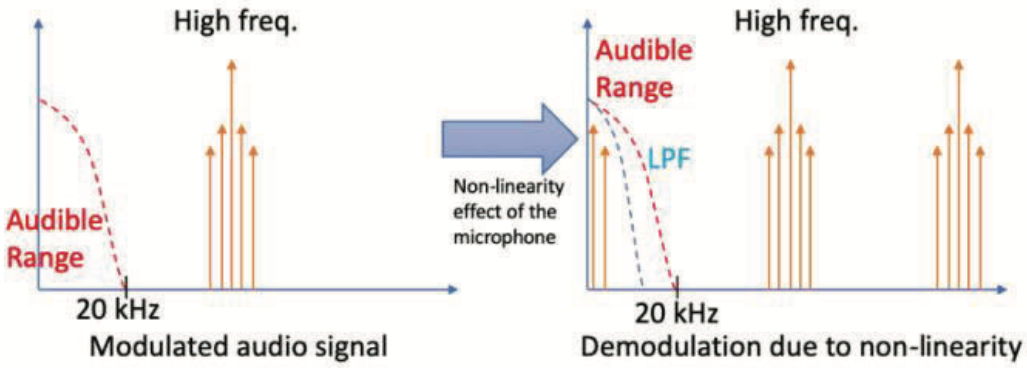


Figure 3. The demodulation of an audio range signal with a high frequency carrier. Re-interpreted from [1].

The square term necessarily produces harmonic distortion, which in result degrades the quality of voice injected into the system. To minimize the harmonic distortion associated with the demodulation process, effective processing method is proposed in (2) as below:

$$S_{in} = \sqrt{Af(t) + A}, \tag{2}$$

Since the square root function will result in an imaginary component, a DC offset equal to the maximum amplitude A of the audio waveform is added under the square-root. This DC offset will not be an issue as the internal inductive coupling of the mixer will remove this component.

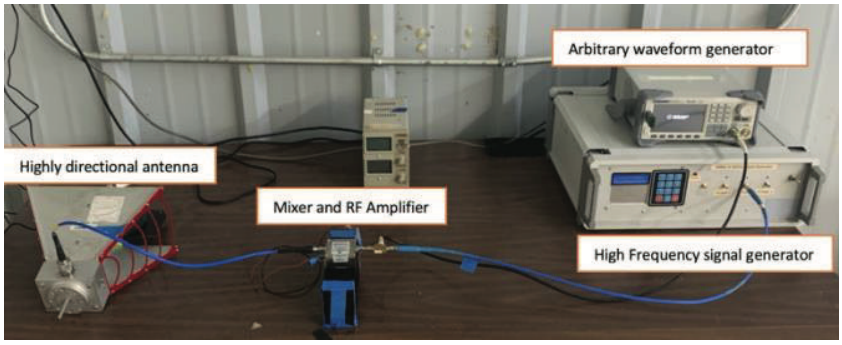


Figure 4. The experimental test setup for the artificial versus natural voice tests. The smart speaker was placed in front of the antenna.

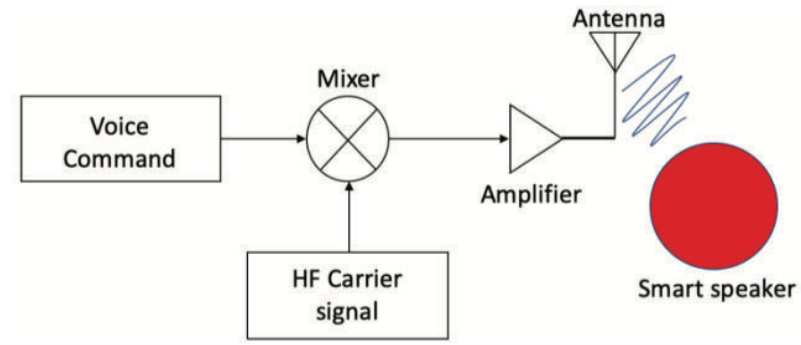


Figure 5. Block diagram representation of the test setup shown in Figure 4.

The test setup for this experiment is shown in Figure 4 and represented as a block diagram in Figure 5. For clarity, in this paper the input voice command will be generally called the “transmitted voice”, and the received command will be referred to as the “attack voice”. To observe a measurable difference in effectiveness for the artificially synthesized voice compared to the naturally recorded voice signal, the smart speaker was placed far enough away so that the recognition rate is less than 100%. This distance varied depending on the DUT. An arbitrary waveform generator was used to output the

audio range signal, while a high frequency signal generator outputs the carrier that the audio signal will be modulated with.

The transmitted voice was modulated with the carrier frequency that the smart speaker was determined to be most susceptible to using a mixer. In a mixer, there is a feedthrough component that will cause a portion of the carrier amplitude to be present in the output (RF) signal. The distortion will be further amplified by the RF amplifier connected to the RF port of the mixer, which will cause distortion in the demodulated waveform. This issue can be mitigated by finding a balance between the intermediate frequency (IF) and local oscillator (LO) that result in minimal distortion.

3. VOICE SYNTHESIS AND EXPERIMENTAL SETUP

To generate the vocal samples, an existing open-source synthesis method called Speaker voice to multi-speaker Text-to- Speech Synthesis (SV2TTS) was implemented [9]. The basic flow of the SV2TTS synthesis is described in Figure 6. This synthesis method is useful for the application of the EMI attack as can generate reasonably similar speech from only 5 seconds of audio.

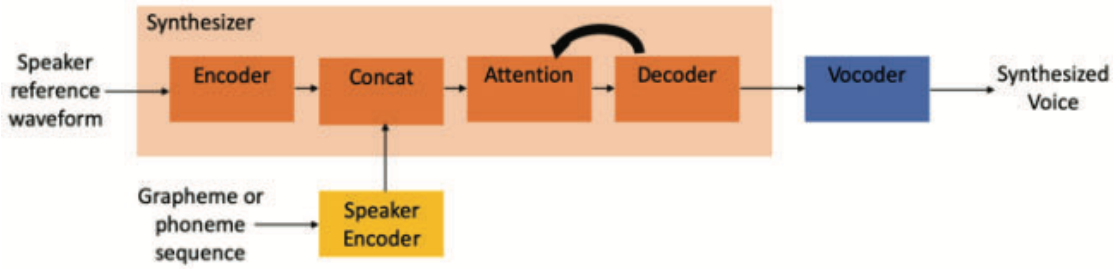


Figure 6. Diagram of SV2TTS method, reinterpreted from [9].

However, this circumstance is not ideal as greater voice samples will still result in greater audio similarity to the speaker that is being cloned. As such, the SV2TTS method was used with 50 different voice samples with a variety of text. Resemble.ai is an implementation of SV2TTS that allows for faster training of the model for this purpose [12]. This website was used for generating the voice samples.

Two voice samples are generated using the software. The samples for each include the ‘wake phrase’ for the device, and then the command “What time is it?”. It is known from released machine learning research that only the wake word is checked for ensuring a match to the user’s voice [10]. However, the generated command that results from SV2TTS is audibly the same to the human voice, so the entire phrase was generated at once.

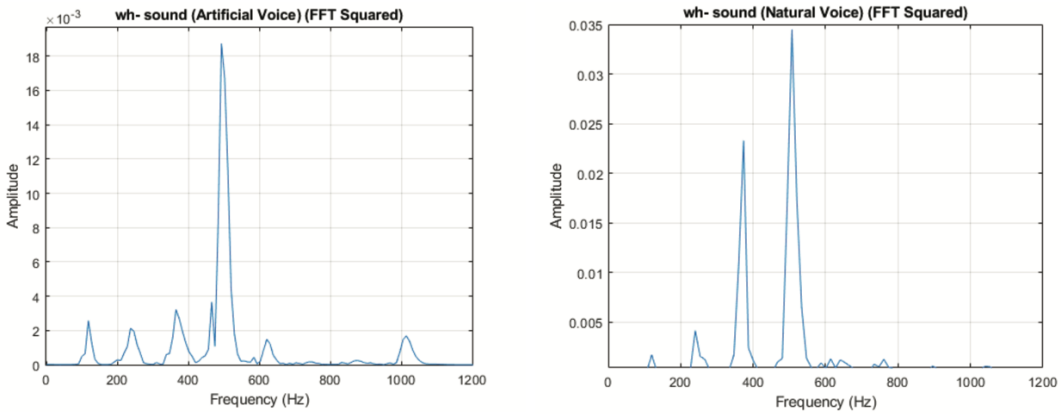


Figure 7. The voice spectra (a) artificial voice (b) natural voice (FFT Squared).

The artificial and natural voice samples for the sound “-wh” in “what” are examined in the frequency domain using the single-sided spectrum in Figure 7 (a) and (b). For ease of viewing, the FFT is squared as this is proportional to PSD (Power

spectral density). From pure observation, the highest amount of frequency content at approximately ~550 Hz appears in both the natural and artificial voice spectra. However, the ~390 Hz peak is not a similar amplitude when comparing the natural and artificial voice spectrums. This observation is significant since the personalized voice feature relies on the pronunciation of word beginnings and endings for recognition [8]. Initially, it seems that the frequency content of the artificially generated voice is a decent match to the voice the neutral network was trained to mimic.

The voice command itself is retrieved using one of two methods, depending on the device: Either a phone called is placed using VOIP (Voice over IP) and the audio is recorded form the call for further processing, or the recorded audio is directly extracted from the cloud service of the smart speaker device. Many smart speakers include a feature that record the input command for playback at a later date on a computer, so this feature can be used to hear the audio that was interpreted on the device's server. This is the idealized method because the audio that is sent to the cloud is what is processed for command interpretation.

The experiment itself was conducted on two different devices which will be identified as "Smart Speaker 1" and "Smart Speaker 2". These two devices originate from different manufacturers. The purpose of using two different smart speakers for this experiment was to see if there was a change in recognition when the personalized voice feature was used across different devices.

4. COMMAND RECOGNITION RESULTS

Smart speaker 1 was tested using a carrier frequency of 18 GHz (LO of the mixer). The audio waveforms are processed using the processing method shown in (2). The transmitted voice that was sent for this test was the wake word followed by ‘what time is it’. The artificial voice yielded a recognition rate of 60% for the EMI attack, while the natural voice had a recognition rate of 70% for the received attack voice. As mentioned in the earlier in the paper, this lesser recognition rate for the natural voice was purposefully chosen so that there would potentially be greater variance between the artificial and natural voice results. In order to reduce the recognition rate of the test, the distance between the horn antenna. Both tests consisted of 20 command send attempts. Recognition of the command was defined as the smart speaker waking up, and then executing the given command. If the smart speaker only woke up, but does not interpret the command, this was designated as a failure. This result is shown in Figure 8.
speaker 1 test. Smart speaker 2 required a different type of attack message to test the ability to circumvent the personalization feature. Since this smart speaker does not completely stop the user from sending commands to the device, but rather limits so-called ‘personalized results’, such as sending an email, the command can be changed to something that warrants a personalized result. In the case of this test, the commands that were sent using both the artificial and natural voice were ‘Send an email’, with the wake word proceeding the command. The results of this test are shown in Figure 9.

5. DISCUSSION

The results of this experiment show the use of artificially synthesized voice in the attack is not only feasible, but with a nearly identical or sometimes even better recognition rate compared to the natural voice. A possible explanation as to why the voice synthesis method was so effective at bypassing the personalized audio feature is that the synthesized audio was replicating the same vocal signatures that the smart speaker learns when learning the user's voice. Personalized voice systems work by using feature extraction to convert the incoming voice single into a "speaker vector" [10], which is then compared to the given model and scored based on its similarity. Since this speaker vector has a limited size of 442 parameters for each voice signal, it would make sense that the SV2TTS vocal synthesis overlaps with most of these parameters, which resulted in the command being recognized.

Although this method of circumvention is primarily relevant to smart speakers, it can be applied to any device susceptible to an I-EMI through the MEMS microphone. Given that the requirement for this I-EMI attack is a structure that couples the incoming EM wave to the MEMS microphone, nearly almost every telecommunication device that relies on the tonal qualities of the voice for identification is susceptible to this attack.

Additionally, the discovery of this additional attack method can have several potential consequences since it can be performed completely inaudibly. For example, in Android versions 5.0 through 7.1.2, it is possible to unlock the phone solely using a voice command, making it possible to send commands that send text messages and email. In conventional smart speaker devices that are connected to devices like smart plugs, it

would be possible to use artificial voice methods described in this paper to circumvent the personalized voice security features.

To protect against this sort of attack, existing machine learning models can be used to accurately identify fake speech compared to natural (real) speech. An example of this type of model is Resemblyzer, which is a voice encoder that provides a high-level representation of a vocal samples that can be subsequently analyzed with a pre-trained model. A similar idea could be implemented into the existing models of the smart speaker voice recognition to ignore voices that are predicted to be fake by the fake speech detection model.

Some downsides of implementing this sort of model into existing smart speakers is that there may be added issues involving misinterpreting real speech as fake speech, causing overall recognition rate to decrease. Additionally, since smart speakers only listen for the wake word when trained for voice personalization, this sample may be too short to determine if a voice is fake or not. A future work could examine the feasibility of protecting against this problem.

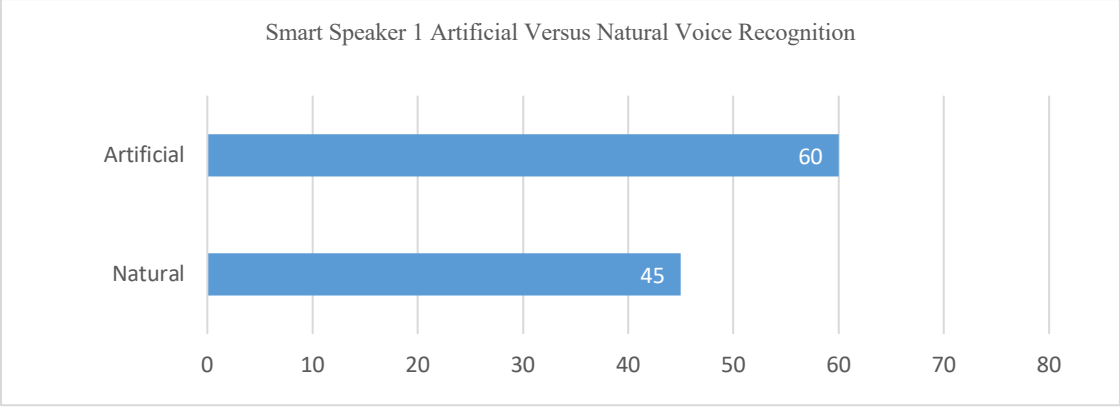


Figure 8. The recognition rate comparison between the natural voice and artificial voice for the first smart speaker.

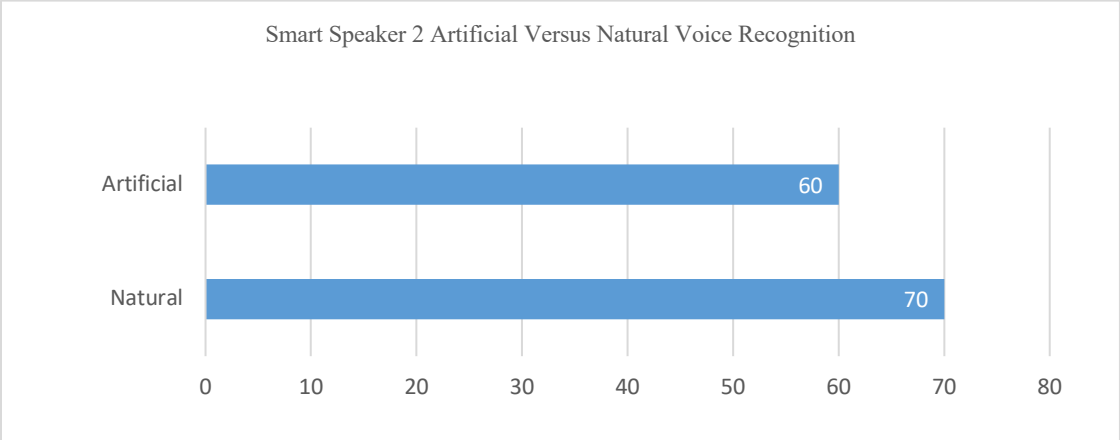


Figure 9. The recognition rate comparison between the natural voice and artificial voice for the second smart speaker.

Smart speaker 2 was tested with a carrier frequency of 8.5 GHz. The audio processing method was the same as the smart overall recognition rate to decrease. Additionally, since smart speakers only listen for the wake word when trained for voice personalization, this sample may be too short to determine if a voice is fake or not. A future work could examine the feasibility of protecting against this problem.

6. CONCLUSION

Artificial voice synthesis can circumvent personalized voice learning in smart speakers. Because the inaudible RF smart speaker attack would not be effective on devices with this personalization feature, this paper has shown the feasibility of cloning someone's voice by using arbitrary voice samples and open-source voice synthesis methods. Recognition of behalf of the device for the artificially generated samples appears to show minimal difference compared to samples containing the person's natural

voice. As a result, it can be concluded that this is an effective method for circumventing the voice personalization feature for the I-EMI attack for the devices that were tested.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under Grant No. IIP-1916535.

REFERENCES

[1] Z. Xu, R. Hua, J. Juang, S. Xia, J. Fan, C. Hwang, “Inaudible Attack on Smart Speakers With Intentional Electromagnetic Interference,” IEEE Trans. Microwave Theory and Techniques (accepted for publication).

[2] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., and Xu, W.,"Dolphinattack: Inaudible voice commands." Proceedings of the 2017 ACM SIGSAC. ACM, 2017

[3] Roy, N., Shen, S., Hassanieh, H., and Choudhury, R. R., "Inaudible voice commands: The long-range attack and defense." 15th USENIX. USENIX, 2018.

[4] Song, L., and Prateek M., "Poster: Inaudible voice commands." Proceedings of the 2017 ACM SIGSAC. ACM, 2017.

[5] Wang, Q., Ren, K., Zhou, M., Lei, T., Koutsonikolas, D., and Su, L.,"Messages behind the sound:real-time hiddenacoustic signal capture with smartphones." Proceedings of the 22nd ACM. ACM, 2016.

[6] Lijima, R., Minami, S., Zhou, Y., Takehisa, T., Takahashi, T., Oikawa, Y., & Mori, T., "Audio Hotspot Attack: An Attack onVoice Assistance Systems Using Directional Sound Beams and its Feasibility," IEEE Trans. Emerg. Topics Comput, 2019.

[7] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," IEEE Trans. Dependable Secure Comput, 2019.

[8] J. Mao, S. Zhu, X.Dai, Q. Lin and J. Liu, "Watchdog: Detecting Ultrasonic-Based Inaudible Voice Attacks to Smart Home Systems," IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8025-8035, Sept. 2020

[9] Ye Jia, et al. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis." (2019).

[10] “Personalized Hey Siri.” Apple Machine Learning Research, Apple Machine Learning Research, Apr. 2018, machinelearning.apple.com/research/personalized-hey-siri.

[11] “Cloned ~ Resemble AI.” Resemble AI, 25 June 2020, www.resemble.ai/cloned/.

[12] Javier Gago, Josep Balcells, David GonzÁlez, Manuel Lamich, Juan Mon, andAlfonso Santolaria. 2007. “EMI susceptibility model of signal conditioning circuitsbased on operational amplifiers,” IEEE Transactions on Electromagnetic Compatibility 49, 4 (2007), 849–859

[13] Y. Zhong, Q. Huang, T. Enomoto, S. Seto, K. Araki, and C. Hwang, “Measurement Based Characterization of Buzz Noise in Wireless Devices.” IEEE Int. Symp. On Electromagn. Compat, Long Beach, CA, pp. 134-138, 2018.

[14] C. Kasmi and J. Lopes Esteves, "IEMI Threats for Information Security: Remote Command Injection on Modern Smartphones," in IEEE Transactions on Electromagnetic Compatibility, vol. 57, no. 6, pp. 1752-1755, Dec. 2015, doi: 10.1109/TEMC.2015.2463089.

SECTION

2. CONCLUSIONS

The first paper showed a method for modeling and understanding the smart speaker I-EMI attack. This included a method for finding the ideal attack angle, locating the region sensitive to the coupled EMI, and modeling the attack. Finally, using all these methods, a long distance (6-meter) attack was demonstrated using 6.3 Watts of power at the aggressor antenna.

The second paper presented the effectiveness of using machine learning (ML) synthesized voice samples to control smart speaker devices through radiated intentional electromagnetic interference (I-EMI). Devices that are trained to only recognize a single person's voice or only execute certain commands from that person will not be as susceptible to the I-EMI attack. By training a neural network using samples of the target's voice, this security feature was bypassed, increasing the feasibility of the attack.

VITA

Tanner Fokkens received his B.S. degree in Electrical Engineering from the Missouri University of Science and Technology in Rolla, Missouri in 2021. He worked at ESDEMC Technology LLC from January 2018 to January 2021 during his undergraduate studies. From 2020 to 2022, he did research at the Missouri University of Science and Technology Electromagnetic Compatibiltity Laboratory. His research interests included electromagnetic interference and signal integrity. He received his Master of Science degree in Electrical Engineering from Missouri University of Science and Technology, USA, in May 2023.