J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

스마트 스피커 대상 가청 주파수 대역을 활용한 적대적 명령어 공격 방법 제안

Proposal of Hostile Command Attack Method Using Audible Frequency Band for Smart Speaker

박 태 준[1] 문 종 섭[1*]
Tae-jun Park Jongsub Moon

요 약

최근 스마트 스피커의 기능이 다양해지면서 스마트 스피커의 보급률이 증가하고 있다. 보급이 증가함에 따라 스마트 스피커에 대해 비정상적인 행위를 발생시키는 기법이 제안되고 있으며 여러 가지 공격 중 Voice Controllable System(VCS)에 대해 비정상적인 행위를 발생시키는 DolphinAttack은 초음파($f>20kHz$)를 이용하여 사용자의 인식 없이 VCS를 제3자가 제어하는 방법이다. 하지만 기존의 제어 방법은 초음파 대역을 사용하기 때문에 초음파 신호를 출력할 수 있는 초음파 스피커나 초음파 전용 장비의 설치가 필요했다. 본 논문에서는 추가적인 장비, 즉, 초음파 장비의 설치 없이 사람의 가청 주파수 대역이지만 노화에 의해 듣기 힘든 주파수($18k$~$20kHz$)로 변조된 음성신호를 출력하여, 스마트 스피커를 제어하는 방법을 제안한다. 스마트 스피커의 경우 마이크가 내장되어 있어, 변조된 음성신호를 수신할 수 있다. 본 논문에서 제안한 방법으로 수행한 결과, 가청 대역임에도 불구하고 사람은 음성명령을 인식하지 못하였으며, 스마트 스피커에 대해 82~96%의 확률로 제어가 가능했다.

☞ 주제어 : 돌핀어택, 스마트 스피커, 음성 명령, 화자 인증

ABSTRACT

Recently, the functions of smart speakers have diversified, and the penetration rate of smart speakers is increasing. As it becomes more widespread, various techniques have been proposed to cause anomalous behavior against smart speakers. Dolphin Attack, which causes anomalous behavior against the Voice Controllable System (VCS) during various attacks, is a representative method. With this method, a third party controls VCS using ultrasonic band ($f>20kHz$) without the user's recognition. However, since the method uses the ultrasonic band, it is necessary to install an ultrasonic speaker or an ultrasonic dedicated device which is capable of outputting an ultrasonic signal. In this paper, a smart speaker is controlled by generating an audio signal modulated at a frequency (18 to 20) which is difficult for a person to hear although it is in the human audible frequency band without installing an additional device, that is, an ultrasonic device. As a result with the method proposed in this paper, while humans could not recognize voice commands even in the audible band, it was possible to control the smart speaker with a probability of 82 to 96%.

☞ keyword : DolphinAttack, Smart speaker, Voice Controllable System, Speaker recognition

1. 서 론

최근 스마트 기기들은 다양한 분야에 보급되고 있으며 스마트 기기들이 하나의 네트워크로 연결되어 동작하고 있다. 그중 스마트 스피커는 내장된 마이크를 통하여 사람의 음성을 받아들여, 음성을 디지털 정보로 변환 후에, 디지털 정보를 사용하여 네트워크에 연결되어 있는 장비를 제어하고 IR 센서를 이용하여 다양한 장비를 제어할 수 있다. 스마트 스피커는 마이크가 내장되어 있기 때문에 사람의 음성을 이용하여 장비를 제어하는 Voice Controllable System(VCS)의 기능이 존재하며, 다른 장비를 제어할 수 있기 때문에 스마트 스피커를 대상으로 비정상 행위를 발생시킬 수 있는 다양한 방법이 제안되고 있다[1-6]. 여기서 비정상 행위의 정의는 실제 사용자 외의 인가되지 않은 제3자가 장비를 제어하는 것을 말한다.

VCS를 지원하는 장비에 비정상 행위를 일으키는 방법 중 하나인 DolphinAttack[7-8]은 초음파 신호($f>20kHz$)

1 Division of Information Security, School of Cybersecurity Korea University, Seoul, 02841, Republic of Korea
* Corresponding author (jsmoon@korea.ac.kr)
[Received 21 June 2022, Reviewed 6 July 2022(R1 3 August 2022), Accepted 11 August 2022]

를 이용하여 VCS에 비정상 행위를 일으킨다. 스마트 스피커의 VCS에 비정상 행위를 일으키는 방법은 레이저를 사용하여 Micro Electro Mechanical System(MEMS)에 신호를 보내[2] 비정상 행위를 발생시키거나, 사용자가 VCS를 활성화시키고, 음성명령을 통해 동작을 시킬 때 별도의 신호를 출력하여 사용자가 원하는 동작과는 다르게 동작하게 하는[4] 등 다양한 기법이 제안되고 있으나 비정상 행위를 발생시키기 위해서는 전문적인 장비가 필요하고, 다수의 VCS에 대해 비정상 행위의 발생은 제한되어 본 논문에서는 초음파를 이용하여 비정상 행위를 발생시키는 DolphinAttack을 선택하여 비교하였다. VCS를 지원하는 장비들은 음성을 통해 장비를 제어하기 때문에 마이크가 내장되어 있다. VCS를 지원하는 대표적인 장비는 스마트폰, 스마트 스피커 등이 존재한다. 스마트 스피커 또한 VCS를 지원하기 때문에 마이크가 내장되어 있어 사람의 음성을 통한 제어가 가능하여, 비정상 행위의 발생이 가능하다. 이와 같이 VCS를 이용한 비정상 행위가 가능한 이유는 스마트 스피커와 같이 VCS를 지원하는 장비들은 마이크가 내장되어 있고, VCS를 사용하기 위해 항상 마이크가 실행 중이기 때문이다. 또한 기존에 많이 사용되던 Electret Condenser Microphone(ECM)이 아닌 Micro Electro Mechanical System(MEMS) 마이크를 사용하여 음성신호를 수신하기 때문이다. MEMS 마이크는 ECM 보다 크기가 작아, 높은 주파수에 민감하게 반응한다[9, 10]. 높은 주파수는 파형이 작기 때문에 크기가 작은 MEMS 마이크를 떨리게 하고, MEMS 마이크는 높은 주파수에 의해 떨린 신호를 정상신호로 수신하게 된다. 높은 주파수에 의해 떨린 신호를 정상신호로 수신하기 때문에 DolphinAttack과 같은 비정상 행위를 발생시키는 환경에 노출되어 있다. 하지만, DolphinAttack의 경우 사람의 가청 주파수(20~20,000Hz)이상의 초음파 대역을 이용한 방법으로 초음파 신호를 출력하여 MEMS 마이크가 장착된 장비에 대해 비정상 행위를 일으킨다. 일반적인 스피커는 초음파 대역을 지원하지 않기 때문에 DolphinAttack은 초음파 장비의 설치 없이 비정상 행위를 발생시킬 수 없다.

본 논문에서는 가청 주파수 범위를 이용하여 스마트 스피커에 대해 비정상 행위를 발생시키기는 방법을 제안한다. 가청 주파수를 사용하기 때문에 초음파 스피커의 설치 없이 Google home[11], Amazon echo[12], Samsung home mini[13], Kakao hexa[14]와 같은 스마트 스피커에 대해, 사람이 인식하지 못하는 상태에서 비정상 행위를 가능하게 하는 명령어 생성이 가능함을 보인다. 또한, 화자 인증, 활성 명령어 변경 이후 비정상 행위의 실행률을 측정하여 H/W의 추가 없이 본 논문에서 제시하는 비정상 행위 발생에 대한 방어가 가능한지 알아본다.

본 논문의 기여는 다음과 같다.

1. 사람의 가청 주파수 대역에도 불구하고 대부분의 사람이 듣지 못하는 주파수 대역(18k~20kHz)을 이용하여 스마트 스피커를 사용한 기기들에 대해 비정상 행위가 가능함을 보인다.
2. 추가적인 장비의 설치 필요 없이 스마트 스피커에 대한 비정상 행위가 가능함을 보여 공격 벡터(Attack vector)가 증가한다.
3. 추가적인 하드웨어 없이 화자 인증 설정, 활성 명령어(“Hey kakao 등”)의 변경으로 일정부분 방어가 가능함을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대한 설명을 하였으며, 3장에서는 제안 방법을 설명하였으며, 4장에서는 제안하는 기법의 실험 결과를 제시한다. 마지막으로 5장에서는 결론을 맺는다.

2. 관련 연구

본 장에서는 VCS가 무엇이며, 대표적인 마이크의 작동 방식인 ECM, MEMS 마이크의 주파수 대역(20~20,000Hz)을 어떻게 인식하는지 알아본다. 또한 20kHz 이상의 초음파 대역은 Low Pass Filter(LPF)에 의해 차단됨에도 불구하고 초음파 대역을 녹음이 가능한 MEMS 마이크의 특징을 설명한다. 이어서 MEMS 마이크의 특징을 이용하여 비정상 행위를 발생시키는 DolphinAttack의 장비 구성 및 절차를 알아보고, 마지막으로 사람의 연령대에 따른 최대 가청 주파수[15]를 알아본다.

2.1 Voice Controllable System

VCS는 사용자의 음성명령을 통해 장비를 제어하는 기능이다. VCS의 실행단계는 음성 입력, 입력 음성 인식, 인식된 명령 실행의 단계로 진행된다. 사용자가 “Hey kakao 등”과 같은 활성 명령어를 발음하면, MEMS 마이크를 통해 음성신호가 입력되고 입력된 음성신호가, 명령어로 인식이 된다. 이 인식된 명령어를 통하여, VCS가 활성화된다. VCS가 활성화된 후 “오늘 날씨 어때 등”과 같은 실행 명령어를 인식하여 장비를 제어한다. VCS를 지원하는 장비는 스마트폰, 스마트 스피커, 웨어러블 디바이스,

자동차 등이 존재하며 실행 명령어를 통해 VCS를 지원하는 장비들에 설치된 응용프로그램을 실행하거나, 전화를 걸 수 있는 등 다양한 기능을 지원한다. VCS는 사용자가 언제 호출하는지에 대한 특징이 없어 마이크가 활성화되어있는 상태로 대기하기 때문에 정상, 비정상적인 음성을 항상 입력받는 상태이다. 이러한 특징으로 인해 비정상 음성도 입력을 받게 되는 취약점이 발생하게 된다.

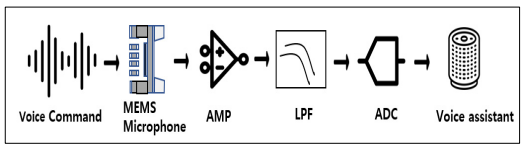
2.2 마이크

음성을 녹음하는 시스템은 마이크를 입력장치로 사용하며 마이크의 대표적인 동작 방식은 아래와 같다.

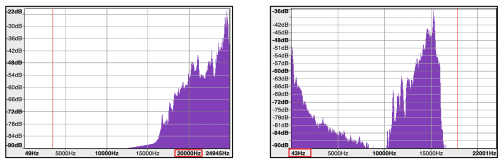
1. Electret Condenser Microphones(ECM)
2. Micro Electro Mechanical Systems(MEMS)

MEMS 마이크는 소비전력도 적고 크기도 기존 ECM보다 작다. 이러한 이유로 소형화가 필요한 장비들에 많이 사용되고 있으며, 특히 그중에서도 스마트폰, 스마트 스피커와 같은 모바일 기기에 많이 사용되고 있다. ECM, MEMS 마이크의 작동 방식은 유사하다. 먼저 음파가 공기를 통해 공기압을 전달하게 되면 공기압이 캐패시터의 상태를 변화시키고 이를 통해 AC 신호가 생성된다. 이렇게 변환된 전기 신호 중 사람의 가청 주파수 대역(20~20,000Hz)에 해당되는 주파수 대역은 녹음되고 그 이상의 주파수 대역은 LPF를 통해 필터링 된다. 세부적인 음성신호의 변환 과정은 그림 1과 같다.

ECM, MEMS 마이크의 동작 방식이 유사함에도 불구하고 MEMS 마이크는 ECM에 비해 2k~20kHz 주파수 대역 전 구간에서 강한 신호를 수신한다. 이는 MEMS microphone의 크기가 ECM의 크기보다 작아 짧은 파장을 갖는 높은 주파수 대역에 의해 떨리기 때문이다. 따라서 MEMS 마이크는 ECM보다 크기가 작기 때문에 MEMS 마이크를 통해 높은 주파수 신호가 수신되면 높은 주파수가 MEMS 마이크를 떨리게 만들고 떨린 신호를 수신하게 된다. 이런 특징으로 인해 MEMS 마이크가 탑재된 장비 중 VCS를 지원하는 장비들은



(그림 1) 신호 처리 단계
(Figure 1) Signal processing phase



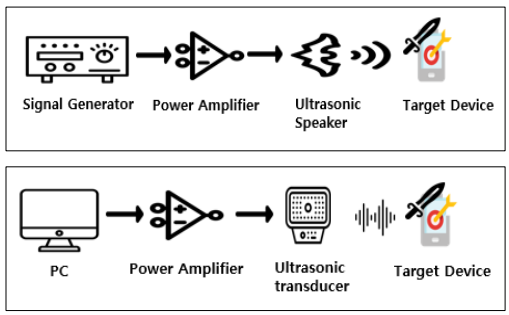
(a) (b)

(그림 2) (a) 초음파 신호, (b) 녹음된 초음파 신호
(Figure 2) (a) Ultrasonic signal, (b) Recorded ultrasonic signal

DolphinAttack과 같은 환경에 노출되어 있다.

2.3 DolphinAttack

DolphinAttack은 초음파 신호에 의해 MEMS 마이크가 떨리게 되고 이를 정상적인 신호로 인식한다. 그림 2의 (a)의 신호와 같이 초음파 신호를 초음파 스피커를 통해 출력을 하게 되면, 초음파 신호에 의해 MEMS 마이크가 떨리게 되고 떨린 신호를 정상적인 신호로 인식하게 된다. 그림 2의 (a)의 신호를 출력하였지만 실제로 MEMS 마이크가 인식한 신호는 (a)의 파형이 아닌, (b)의 파형으로 인식한다. 결국 초음파 대역의 신호를 정상적인 신호로 인식하고 VCS를 통해 장비를 제어할 수 있다. DolphinAttack은 초음파 대역을 이용하여 진행된다. 그림 3의 위의 그림은 초음파 대역을 발생시킬 수 있는 신호 발생기를 이용해 음성명령 신호를 발생시킨 뒤, 발생된 신호를 Power Amplifier를 통해 증폭시킨다. 증폭시킨 신호는 초음파 대역이기 때문에 초음파 센서나, 초음파 스피커를 연결하여 초음파 신호를 출력하여 VCS에 비정상 행위를 발생시킨다. 그림 3의 아래 그림은 PC, 스마트폰 등을 통해 신호를 출력할 수 있는 장비에서 초음파 대역으로 변조한 음성 명령을 실행하여 Power Amplifier로 전



(그림 3) 돌핀어택 구조
(Figure 3) DolphinAttack structure

송하고 받은 신호를 증폭 후, 초음파 대역의 출력이 가능한 transducer를 통해 초음파 음원을 출력하여 VCS를 지원하는 장비에 비정상 행위를 발생시킨다.

2.4 사람의 가청 주파수

사람의 가청 주파수 대역은 20~20,000Hz로 알려져 있다. 하지만 노화나 소음이 있는 환경에 노출되는 시간에 따라 가청 주파수 대역 중 높은 주파수부터 듣지 못하게 되는데 표 1에서 보는 것처럼 각 연령대별 들을 수 있는 최대 주파수[15]는 나이가 많아질수록 낮아진다. 10대의 경우 평균 최대 18,000Hz까지 들을 수 있으며 노화에 따라 점점 최대 가청 주파수 대역이 줄어듦을 볼 수 있다.

기존에 제시된 DolphinAttack의 경우에는 초음파를 이용하여 비정상 행위를 발생하기 때문에 DolphinAttack에 필요한 초음파 스피커가 추가로 필요하였으나 사람의 가청범위(20~20,000Hz)를 이용한다면 추가적인 장비의 필요 없이 일반적인 스피커를 이용하여 스마트 스피커에 비정상 행위를 발생시킬 수 있다.

(표 1) 연령별 평균 최대 가청 주파수
(Table 1) Average maximum audible frequency by age

Age	Maximum Frequency
10s	18,000 Hz
20s	17,000 Hz
30s	16,000 Hz
40s	14,000 Hz
50s	12,000 Hz

3. 제안하는 기법

3.1 가청 주파수를 활용한 비정상 행위 발생

본 논문에서는 사람의 가청 주파수 대역(20~20,000 Hz) 중 대부분의 사람들이 듣지 못하는 주파수 대역(18k~20kHz)을 활용한 명령어 전송 기법을 제안한다.

3.2 위협 모델

공격자의 목표는 장비 소유자의 인식 없이 스마트 스피커에 음성명령을 전달하여 장비 소유자가 아닌 제3자에 의해 스마트 스피커의 VCS를 실행하여 스마트 스피커를 작동시키는 것이다. 이 방법의 장점은 다음과 같다.

1. 스마트 스피커에 어떠한 물리적 접근을 하지 않고 제어한다. 즉, 공격자가 직접 악성코드 설치, 스마트 스피커의 설정 변경과 같은 행위는 할 수 없다.
2. 음성명령이 스마트 스피커에 전달될 때 사람이 인지하지 못한다. 사람의 가청 주파수 대역을 이용하지만 어떤 음성명령이 전달되었는지 사람은 알지 못한다.
3. 추가적인 장비의 설치가 필요하지 않다. 즉, 음성명령을 전달하기 위해 신호 증폭기, 초음파 스피커와 같은 장비의 설치가 필요 없다.

일반적인 스피커를 이용하여 스마트 스피커를 제어하기 때문에 아파트나, 가정에 설치되어 있는 스피커를 이용하여 스마트 스피커를 제어할 수 있다. 따라서 발생 가능한 공격 시나리오는 아래와 같다.

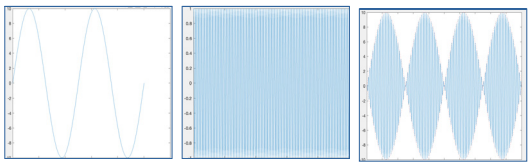
1. 아파트의 방송 시스템에 접근한 뒤 변조된 음성명령 파일이 저장된 장치(스마트폰, 노트북 등)를 방송 시스템에 연결, 아파트에 설치된 방송용 스피커를 이용하여 가정에 설치된 스마트 스피커를 동작 시킬 수 있다.
2. 시청자가 많은 인기 유튜버, 트위치 방송의 기능 중 하나인 영상 재생 donation으로 변조된 음원을 재생함으로써 시청자의 가정에 설치된 스마트 스피커를 동작시킬 수 있다.

3.3 가청 주파수 대역 음성명령 변환 방법

본 논문에서는 기존에 제시된 초음파 대역을 활용한 공격이 아닌 음성명령을 사람의 가청 주파수(18k~20kHz)로 변조하여 스마트 스피커를 제어한다. 먼저, 음성명령이 녹음된 음원을 사람의 가청 주파수(18k~20kHz)로 변경한다. 변경 방법은 AM Modulation을 사용하며 대표적인 Double Side Band(DSB)의 수식은 아래와 같다.

$$V_{au}(t) = C(t) \cdot \cos(2\pi f_{au} t) + \cos(2\pi f_{au} t) \quad (1)$$

수식 (1)에서 “Hey google, 오늘 날씨 어때?”를 녹음한 음원은 $C(t)$이고 비정상 행위의 발생을 위해 변조할 주파수 대역은 $18kHz < f_{au} < 20kHz$이며, t는 시간이다. 따라서 $C(t)$는 시간에 따른 입력신호이고, $V_{au}(t)$는 시간에 따른 출력신호이다. f_{au}는 반송파(Carrier)로



(a) $C(t)$ (b) f_{au} (c) $V_{au}(t)$

(그림 4) 변조 과정
(Figure 4) Modulation phase

C(t)의 정보를 반송파(f_{au})에 실어 보내게 된다. 즉, 그림 4의 (a)는 녹음된 음원의 파형으로 가정할때, (b)와 같은 반송파를 이용하여 AM modulation을 진행하면, (c)와 같은 파형이 나오게 된다. 이를 통해 DSB 변조가 된다. 본 논문에서 제시한 방법은 가청 대역을 이용하여 비정상 행위를 발생시키기 때문에 반송파보다 낮은 주파수도 같이 출력되는 DSB 보다 반송파 주파수 이상을 출력할 수 있는 Single Side Band(SSB)를 이용한다. SSB를 사용하기 위해선 기존 $C(t)$를 Hilbert[16]로 변환한 신호 $C'(t)$를 $\sin(2\pi f_{au}t)$와 곱하고 이를 수식 (1)에서 빼주면 반송파 보다 높은 주파수만 출력 가능하다. 이 방법은 수식(2)에 해당된다.

$$V_{au}(t) = C(t)\cos(2\pi f_{au}t) - C'(t)\sin(2\pi f_{au}t) + \cos(2\pi f_{au}t) \quad (2)$$

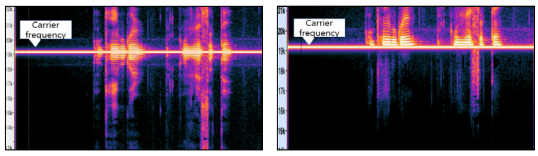
그림 5의 (a)는 수식 (1)을 이용하여 변조한 파형(DSB)이며 (b)는 수식 (2)를 이용하여 변호한 파형(SSB)이다. 그림 5에서 보는 것처럼 수식 (2)를 사용하였을 때 반송파(f_{au}) 이상의 주파수 대역으로 변조된 것을 확인할 수 있다.

4. 실 험

본 장에서는 3장에서 제안한 방법에 대한 성능을 세부적으로 측정하고, 측정 결과를 제시한다.

4.1 실험 환경

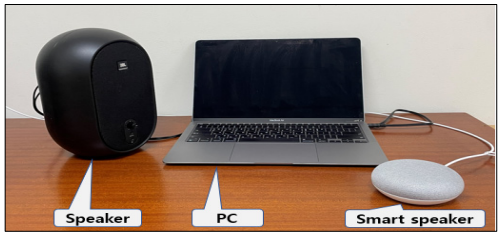
본 논문은 일반적인 스피커가 설치되어 있는 환경으로 실험 환경을 그림 6과 같이 구축하였고, 실험에 사용된 장비는 표 2의 장비를 이용한다. Macbook Air를 통해 변조된 음성을 JBL professional(스피커)으로 출력하였다. 변조된 음성명령을 통해 비정상 행위를 발생시킬 스마트 스피커는 Kakao mini hexa, Samsung home mini, Google home mini , Amazon Alexa를 대상으로 실험을 진행하였고. 화자 인증에는 iPhone 12 mini, Galaxy A20(Bixby 별도 설치)을 이용하여 실험하였다. 사용된 스피커는 사람의 가청범위 대역인 20~20,000Hz를 지원하고 그 이상의 초음파 대역은 지원하지 않는다. 따라서 본 논문에서 실험을 진행한 가청 주파수 대역(18k~20kHz)은 문제없이 출력 가능하다. 개인의 억양, 사투리 등과 같은 특징으로 인한 인식률에 영향을 최소화하기 위하여 명령어의 녹음은 Papago의 Text to Speech(TTS)를 이용하였다. TTS를 통해 녹음된 음성명령은 사람의 가청 주파수 대역(18k~20kHz)으로 변조하였고, 음성의 출력은 MacBook Air의 최대출력의 70%, 스피커의 최대출력의 70%로 설정하여 실험을 진행하였다. TTS로 녹음한 음성이 정상적으로 본 논문에서 제시한 주파수 대역(18k~20kHz)으로 정상적으로 변경이 되었는지 그림 7을 통해 확인할 수 있다.



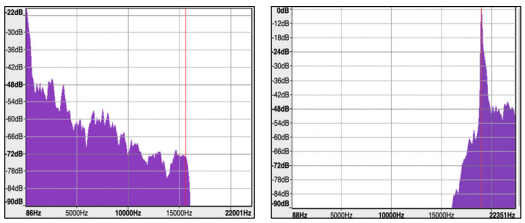
(a) Double side band (b) Single side band

(그림 5) 변조 방법
(Figure 5) Modulation methods

표 3에서 $f(Hz)$는 최대 거리에서 스마트 스피커에 대해 비정상 행위의 실행률이 제일 높은 주파수 대역을 나타내고, Active는 “Hey kakao 등” 과 같은 활성 명령어가 VCS를 활성화시키는 최대 거리를, Recog.는 “오늘 날씨 어때? 등” 과같이 동작을 실행하는 실행 명령어를 스마트 스피커가 인식하여 실행되는 최대 거리를 나타낸다.



(그림 6) 실험 환경
(Figure 6) Evaluation setup



(a) Before modulateion (b) After modulation

(그림 7) 변조 결과

(Figure 7) Result of Modulation

4.2 스마트 스피커에서 실행 가능한 명령어

스마트 스피커에서 사용하는 기능 중 많이 사용하는 기능[17]을 실행하기 위한 명령어를 본 논문에서 제시하는 주파수 대역($18k$~$20kHz$)으로 변조하여 스마트 스피커에서 비정상 행위가 발생하는지 확인하였고, 명령어마다 10회 진행하였으며 비정상 행위의 실행 가능한 명령어 목록은 표 4와 같다.

4.3 스마트 스피커 대상 비정상 행위 실험 결과

실험은 각 스마트 스피커마다 표 3에서의 주파수 대역, 거리로 설정하여 진행하였다. 표 5의 Human은 청각에 문제가 없는 20 대 10명, 30 대 10명을 대상으로 측정하였으며, Command recog.는 변조한 음성명령이 무엇인지 알아내는지에 대한 것을 표현하였고, Recog. noise의 경우 잡음과 같은 불필요한 소리를 느끼는지 표현하였다. Attack success rate는 장비마다 소음이 없는 환경 Noiseless(30~33 dB), 소음이 있는 환경 Noise(60~65dB)에서 "오늘 날씨

(표 2) 실험 장비

(Table 2) Laboratory equipment

Device	Model	Release year
Lab top	MacBook Air M1 Processor	2021
Speaker	JBL professional 1 series 104 compact (f : 20~20,000Hz)	2017
Smart speaker	Kakao mini hexa	2020
	Samsumg home mini	2019
	Google home mini	2017
	Amazon Echo dot 2th	2016
Smart phone	iPhone 12 mini	2020
	Galaxy A20	2019

(표 3) 실험 환경(주파수, 거리)

(Table 3) Test environment(frequency, distance)

Model	$f(Hz)$	Max dist.(cm)	
		Active	Recog.
Kakao Mini Hexa	19,250	66	54
Galaxy home mini	19,750	84	74
Google home mini	19,000	114	104
Amazon Echo dot	18,750	73	68

8어때?"의 실행 명령어를 스마트 스피커마다 50회 진행하였고 스마트 스피커의 비정상 행위의 실행률을 표현하였다. Active는 "Hey kakao 등"과 같이 활성 명령이 스마트 스피커의 VCS를 활성화시키는 것을, Recog.는 "오늘 날씨 어때?"와 같은 실행 명령이 스마트 스피커가 인식하는 것을 표현한다.

표 5에서 보는 것처럼 20대의 경우 $18kHz$ 이상의 주파수 대역에서는 2명이 잡음을 인지했으나 어떤 음성명령이 들렸는지 인식하지 못하였으며, 30대의 경우 $18kHz$ 이상의 주파수에서 잡음을 1명만 인지했고, 20·30대 모두 어떤 음성명령이 실행되었는지 인식하지 못했다. 비정상 행위의 실행률은 스마트 스피커는 조용한 환경(30~33dB)에서 82~96%의 실행률을 보였고, 소음이 있는 경우(60~65dB)는 8~30%의 실행률을 보였다. 기존 DolphinAttack의 경우 스마트폰을 대상으로 80%의 실행률을 보였으며, 스마트 워치는 55~65dB에서 100%, 65~75dB에서 80%, 75~85dB에서는 30%의 실행률을 보였다. 본 논문에서 제시한 방법을 통해 추가적인 초음파 장비의 설치 없이 기존에 제시되었던 DolphinAttack과 같은 수준

(표 4) 스마트 스피커에서 실행 가능한 명령어

(Table 4) Commands executable on smart speakers

Command \ Model	Google home mini	Amazon Echo dot	Galaxy home mini	Kakao mini hexa
Play music	O	O	O	O
Search information	O	O	O	O
Turn off TV	O	O	O	O
Set an alarm	O	O	O	O
How's the weather	O	O	O	O
Stop video	O	O	O	O
Whait time is it?	O	O	O	O
Call 010-1234-5678	O	O	O	O

(표 5) 변조된 음성명령을 통한 스마트 스피커 실험 결과
(Table 5) Comparison Results for smart speaker via modulation voice command

Model	Human				Attack success rate			
	Command recog.		Recog. noise		Active		Recog.	
	20s	30s	20s	30s	Noiseless (30~33dB)	Noise (60~65dB)	Noiseless (30~33dB)	Noise (60~65dB)
Kakao Mini Hexa	0%	0%	20%	10%	82%	14%	82%	8%
Galaxy home mini	0%	0%	0%	0%	86%	54%	86%	24%
Google home mini	0%	0%	20%	10%	96%	54%	96%	30%
Amazon Echo dot	0%	0%	20%	10%	94%	32%	94%	26%

으로 스마트 스피커에 대해 비정상 행위의 발생이 가능할 수 있음을 증명하였으며 추가적인 초음파 장비가 필요하지 않아 기존에 설치된 스피커를 통해 다수의 스마트 스피커에 비정상 행위의 발생이 가능함을 보였다.

4.4 설정 변경을 통한 비정상 행위 방어방법 실험

화자 인증[18-20]은 사람마다 목소리 주파수 대역이 다름을 특징으로 화자의 음성을 장비에 등록한 뒤 등록된 목소리가 아닌 경우 응답하지 않는 기능이다. 화자 인증을 지원하는 대표적인 기기는 Apple의 Siri, Samsung의 Bixby, Google의 Voice Macth가 있다. 표 6에서 Siri, Bixby, Voice Match를 TTS로 화자 인증으로 등록하였고, 화자 인증을 등록한 TTS의 음성명령이 아닌 사람의 음성을 통한 비정상 행위의 발생에 대한 실험은 Human voice, 화자 인증을 등록한 TTS를 이용한 비정상 행위의 발생에 대한 실험은 Text To Speech이며 표 6에서의 실행률을 볼 수 있다.

음성 활성 명령의 변경은 “Hey Kakao 등”과 같이 스마트 스피커의 VCS 기능을 활성화시키기 위한 활성 명령어를 변경하는 것을 말한다. 활성 명령이 다를 경우 VCS는 활성화되지 않는다. VCS를 지원하는 장비들 대부분이 활성 명령을 사용자가 원하는 활성 명령으로 변경할 수 있는 기능을 제공하지 않기 때문에 활성 명령의 변경이 가능한 Kakao mini hexa에 대해서만 실험을 진행하였으며, 활성 명령어를 “Hey Kakao”에서 “라이언”으로 변경했다. 결과는 표 7과 같다. 표 7에서 Acitve는 활성 명령을 통해 VCS가 활성화되는지, Recog.는 VCS의 활성화 이후 실행 명령을 통해 스마트 스피커가 실행 명령을 인식하는지에 대한 실험을 나타낸다. 활성 명령을 통해 VCS가 활성화되지 않았기 때문에 실행 명령도 인식하지 못하였다. 따라서 화자 인증 설정, 활성 명령어의 변경을 통해 스마트 스피커의 비정상 행위의 발생에 대한 방어가 가능할 것으로 보인다.

(표 6) 화자 인증이 설정된 장비의 실행률
(Table 6) Execution rate of devices for which speaker recognition is set

Device / Speaker type	Apple Siri (TTS)	Galaxy Bixby (TTS)	Google Voice Match (TTS)
Human voice (18k~20kHz)	10%	8%	10%
Text To Speech (18k~20kHz)	92%	90%	94%

(표 7) 호출 명령어 변경 시 인식률
(Table 7) Recognition rate when calling command is changed

Model	Active	Recog.
Kakao mini hexa	0%	0%

5. 결 론

본 논문에서는 사람의 가청 범위 중 18k~20kHz 대역을 이용하여 기존의 DolphinAttack과 동일하게 스마트 스피커에 대해 비정상 행위를 발생시킬 수 있음을 입증하였다. 또한 기존 DolphinAttack의 경우 초음파 대역을 지

원하는 스피커나 초음파 센서와 같이 공격을 위한 추가적인 장비가 필요하였으나, 본 논문에서는 제시하는 방법을 이용할 경우 일반적인 환경에서 사용하는 스피커를 통해 비정상 행위를 발생시킬 수 있어 초음파 장비의 설치 필요 없이 다수의 스마트 스피커를 대상으로 비정상 행위를 발생시킬 수 있음을 보였다. 따라서 본 논문에서 제시한 공격 시나리오를 대응하기 위해서는 스마트 스피커의 환경설정을 기본값으로 사용하는 것이 아닌, 활성 명령어가 변경 가능한 스마트 스피커의 경우 사용자가 주로 사용하는 활성 명령으로 변경하고 화자 인증을 지원하는 스마트 스피커는 화자 인증을 설정하여 사용하게 된다면 본 논문에서 제시한 공격에 대한 대응이 가능할 것으로 보인다. 본 논문의 제한사항은 다음과 같다. 조용한 환경(30~33dB)의 환경에서는 82~96%의 실행률을 보였으나, 노래가 틀어져 있는 환경(60~65dB)에서는 8~30%로 실행률이 떨어진다. 또한 화자 인증을 적용했을 경우 8~10%의 실행률로 떨어지는 제한사항이 존재한다. 향후 연구에서는 활성 명령을 변경하지 못하거나, 화자 인증을 지원하지 않는 장비에 대한 공격을 방지하기 위한 연구를 진행하고자 한다.

참고문헌(References)

[1] Z. Wu, S. Gao, E. S. Cling and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, pp. 1-5, 2014. https://doi.org/10.1109/APSIPA.2014.7041636.

[2] Sugawara, T., Cyr, B., Rampazzi, S., Genkin, D., & Fu, K., "Light Commands:{Laser-Based} Audio Injection Attacks on {Voice-Controllable} Systems," USENIX Security Symposium, pp. 2631-2648, 2020. https://www.usenix.org/conference/usenixsecurity20/presentation/sugawara.

[3] 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, pp. 673-677, 2021. https://doi.org/10.1109/EMC/SI/PI/EMCEurope52599.2021.9559146.

[4] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk and H. Zhu, "Voiceprint Mimicry Attack Towards Speaker Verification System in Smart Home," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, pp. 377-386, 2020. https://doi.org/10.1109/INFOCOM41043.2020.9155483.

[5] S. Godwin, B. Glendenning and K. Gagneja, "Future Security of Smart Speaker and IoT Smart Home Devices," 2019 Fifth Conference on Mobile and Secure Services (MobiSecServ), pp. 1-6, 2019. https://doi.org/10.1109/MOBISECSERV.2019.8686545.

[6] H. Chung, M. Iorga, J. Voas and S. Lee, "Alexa, Can I Trust You?", in Computer, vol. 50, no. 9, pp. 100-104, 2017. https://doi.org/10.1109/MC.2017.3571053.

[7] Zhang, Guoming, et al. "Dolphinattack: Inaudible voice commands." Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp.103-117 2017. https://doi.org/10.1145/3133956.3134052.

[8] Roy, Nirupam, et al. "Inaudible Voice Commands: The {Long-Range} Attack and Defense." 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18), pp. 547-560, 2018.

[9] R. N. Dean et al., "A Characterization of the Performance of a MEMS Gyroscope in Acoustically Harsh Environments," in IEEE Transactions on Industrial Electronics, vol. 58, no. 7, pp. 2591-2596, 2011. https://doi.org/10.1109/TIE.2010.2070772.

[10] Z. Wang, Q. Zou, Q. Song and J. Tao, "The era of silicon MEMS microphone and look beyond," 2015 Transducers - 2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS), pp. 375-378, 2015. https://doi.org/10.1109/TRANSDUCERS.2015.7180939.

[11] Google Home. https://store.google.com/product/google_home.

[12] Amazon Alexa. https://developer.amazon.com/en-US/alexa.

[13] GalaxyHome. https://www.samsung.com/sec/ai-speaker/galaxy-home-mini/.

[14] Kakao mini hexa. https://kakao.ai/product/minihexa.

[15] Sakamoto, M., Sugasawa, M., Kaga, K., & Kamio, T., Average thresholds in the 8 to 20 kHz range as a function of age. Scandinavian audiology, vol. 27, no. 3, pp. 189-192, 1998. https://doi.org/10.1080/010503998422728.

[16] Johansson, M. The Hilbert transform. Masters Thesis. Växjö University, 1999. http://www.fuchs-braun.com/media/d9140c7b3d5004fbffff8007ffffff0.pdf.

[17] Bentley, Frank, et al. "Understanding the long-term use of smart speaker assistants." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 3, pp. 1-24, 2018. https://doi.org/10.1145/3264901.

[18] S. Debnath, B. Soni, U. Baruah and D. K. Sah, "Text-dependent speaker verification system: A review," 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), pp. 1-7, 2015. https://doi.org/10.1109/ISCO.2015.7282386.

[19] NIST. National institute of standards and technology speaker recognition evaluation. https://www.nist.gov/itl/iad/mig/speaker-recognition.

[20] M. Lindasalwa, B. Mumtaj and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques", Journal of Computing, vol. 2, no. 3, pp. 138-143, 2010. https://doi.org/10.48550/arXiv.1003.4083.

◐ 저 자 소 개 ◑

박 태 준(Tae-jun Park)
2014년 대진대학교 컴퓨터공학과(공학사)
2021년~현재 고려대학교 정보보호대학원(석사과정)
관심분야 : 정보보호, 시스템 보안, IoT 보안
E-mail : ptj9787@gmail.com

문 종 섭(Jongsub Moon)
1981년 서울대학교 계산통계학과(학사)
1983년 서울대학교 계산통계학과(석사)
1991년 Illinois Institute of Technology 전산학과(박사)
1993년 3월~현재 고려대학교 전자 및 정보공학부 교수
2001년 2월~현재 고려대학교 정보보호대학원 겸임교수
관심분야 : 정보보호, 운영체제, 침입탐지
E-mail : jsmoon@korea.ac.kr