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

화자식별 기반의 AI 음성인식 서비스에 대한 사이버 위협 분석☆

Cyber Threats Analysis of AI Voice Recognition-based Services with Automatic Speaker Verification

홍 천 호[1] 조 영 호[1*]
Chunho Hong Youngho Cho

요 약

음성인식(ASR: Automatic Speech Recognition)은 사람의 말소리를 음성 신호로 분석하고, 문자열로 자동 변화하여 이해하는 기술이다. 초기 음성인식 기술은 하나의 단어를 인식하는 것을 시작으로 두 개 이상의 단어로 구성된 문장을 인식하는 수준까지 진화하였다. 실시간 음성 대화에 있어 높은 인식률은 자연스러운 정보전달의 편리성을 극대화하여 그 적용 범위를 확장하고 있다. 반면에, 음성인식 기술의 활발한 적용에 따라 관련된 사이버 공격과 위협에 대한 우려 역시 증가하고 있다. 기존 연구를 살펴보면, 자동화자 식별(ASV: Automatic Speaker Verification) 기법의 고안과 정확성 향상 등 기술 발전 자체에 관한 연구는 활발히 이루어지고 있으나, 실생활에 적용되고 있는 음성인식 서비스의 자동화자 식별 기술에 대한 사이버 공격 및 위협에 관한 분석연구는 다양하고 깊이 있게 수행되지 않고 있다. 본 연구에서는 자동화자 식별 기술을 갖춘 AI 음성인식 서비스를 대상으로 음성 주파수와 음성속도를 조작하여 음성인증을 우회하는 사이버 공격 모델을 제안하고, 상용 스마트폰의 자동화자 식별 체계를 대상으로 실제 실험을 통해 사이버 위협을 분석한다. 이를 통해 관련 사이버 위협의 심각성을 알리고 효과적인 대응 방안에 관한 연구 관심을 높이고자 한다.

☞ 주제어 : 자동화자 식별, AI 음성인식 서비스, 음성분석 및 변환, 사이버 위협

ABSTRACT

Automatic Speech Recognition(ASR) is a technology that analyzes human speech sound into speech signals and then automatically converts them into character strings that can be understandable by human. Speech recognition technology has evolved from the basic level of recognizing a single word to the advanced level of recognizing sentences consisting of multiple words. In real-time voice conversation, the high recognition rate improves the convenience of natural information delivery and expands the scope of voice-based applications. On the other hand, with the active application of speech recognition technology, concerns about related cyber attacks and threats are also increasing. According to the existing studies, researches on the technology development itself, such as the design of the Automatic Speaker Verification(ASV) technique and improvement of accuracy, are being actively conducted. However, there are not many analysis studies of attacks and threats in depth and variety. In this study, we propose a cyber attack model that bypasses voice authentication by simply manipulating voice frequency and voice speed for AI voice recognition service equipped with automated identification technology and analyze cyber threats by conducting extensive experiments on the automated identification system of commercial smartphones. Through this, we intend to inform the seriousness of the related cyber threats and raise interests in research on effective countermeasures.

☞ keyword : Voice Recognition, AI Voice Recognition Speaker, Automatic Speaker Verification, Voice Conversion

1. 서 론

2008년 4월 미국의 마블 시네마틱 유니버스(Marvel Cinematic Universe)에서 제작한 영화 '아이언맨(Iron Man)'은 첨단기술이 집적된 새로운 영웅의 등장을 소개했다. 극 중에 등장하는 주인공의 인공지능(AI: Artificial Intelligence) 기반 개인비서 '자비스(Javis)'는 음성인식(ASR: Automatic Speech Recognition) 서비스로 단순한 대화뿐만 아닌 주변에서 일어나는 상황을 분석하고 최적의 해결책을 제공하는 모습을 보여줬다. 이는 영화의 첨단기술과 음성인식 서비스가 향후 우리의 일상에 영화와 같이 많은 변화를 줄 수 있다는 기대감을 주었다.

그로부터 13년이 지난 현재 우리는 4차 산업혁명 시대

[1] Department of Defense Science(Computer Engineering and Cyberwarfare Major), Korea National Defense University, Nonsan-si Chungcheongnam-do, 33021, Korea.
* Corresponding author (younho@kndu.ac.kr)
[Received 21 August 2021, Reviewed 1 October 2021, Accepted 15 November 2021]
☆ 본 논문은 2021년 KSII 춘계 학술대회에서 발표한 논문인 "음성합성 기술을 활용한 음성인식 서비스에 대한 사이버 위협 분석 연구"를 확장한 것이다.

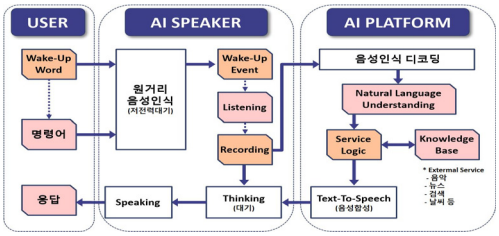
를 맞이하여 인공지능 기반의 음성인식 서비스를 노인 인구 증가에 따른 복지 문제의 대안으로 제시되기도 하며, 음성인식 기능을 탑재한 차세대 전투기 F-35 등 국방 분야에도 적용하여 사용되고 있다[1, 2]. 가정에서도 KT의 '기가지니(GiGA Genie)', 카카오의 '카카오미니(Kakao Mini)', 네이버의 '클로바(Clova)' 등 음성인식 기반의 다양한 AI 스피커(AI Speaker)를 사용하고 있다. 과학기술정보통신부는 2021년 3월 '2020 인터넷 이용 실태조사 결과'를 발표하였다. 대한민국 국민 만 3세 이상 5,097만 명 중 약 46,818,750명(91.9%)이 인터넷을 이용하고 있으며, 전체 AI 음성인식 서비스 이용률은 28.5%가 이용하고 있어 '19년 대비 3.3%가 증가한 수치였다. 비대면 흐름과 함께 인공지능 음성인식 서비스 이용이 증가한 것으로 나타났다[3]. 더욱이 팬대믹 이후 재택근무 및 원격수업으로 집에서 일상의 대부분 시간을 보내게 되므로, 음성인식 서비스에 관한 관심과 구매가 상승하게 되었고 사물인터넷(IoT: Internet of Things)과 연계한 '스마트홈(Smart Home)' 구성도 주목을 받고 있다. 이는 음성인식 서비스와의 소통과정이 인간에게 있어 친숙한 정보전달의 방법으로 별도의 학습이나 훈련 없이도 사용이 가능하다는 점과 이동 중에도 음성으로 기기를 사용할 수 있다는 편리함을 가지고 있다. 또한, 음성을 통해 신원확인, 심리, 건강 상태 등을 파악할 수 있어 개인별 서비스 제공이 가능하며, 입력 속도가 타자보다 빨라 고속 또는 실시간 정보처리가 가능하다는 장점으로 음성인식 서비스의 적용 범위가 확장되고 있다. 대표적으로 스마트폰 및 스마트 TV와 자동차 등 다양한 가전제품에 적용되고 있다.

하지만 음성인식 서비스에 대한 보안 이슈도 지속해서 지적되고 있다. 사람의 음성만으로 개인정보 및 사물을 제어할 수 있는 음성인식 서비스 기기 자체의 보안 취약점과 기업으로부터 발생하는 개인 정보 유출 및 프라이버시 침해 등의 문제들이 대두됐으며, 사람의 말소리를 흉내 내는 성대모사, 기계로 말소리를 재연하는 음성합성 및 음성 변환에 대한 위험성이 연구되고 있다[4, 5]. 반면에, 상용화된 스마트폰의 음성인식 서비스와 자동화자 식별에 대한 사이버 위협과 취약성에 관한 연구는 깊이 있게 수행되고 있지 못하는 실정이다. 2021년 6월 한국갤럽에서 전국 만 18세 이상 1,003명을 대상으로 '2012-2021 스마트폰 사용률 & 브랜드'를 조사하였으며, 2021년 6월 기준 스마트폰 사용자가 94.7%에 달한다[6]. 즉, 94.7%의 스마트폰 사용자는 스마트폰에서 제공하는 음성인식 서비스를 사용해 보았거나, 기능을 인식하고 있음을 가정해 볼 수 있다. 이러한 시점에서 상용화된 스마트폰 음성인식 서비스 중 자동화자 식별(ASV: Automatic Speaker Verification)에 대한 사이버 위협 분석 연구가 중요하다고 판단된다. 따라서, 기존 연구에서 소스 음성과 대상 음성 사이의 개별 음소를 모핑하여 ASV 기반의 음성인식 서비스를 공격하는 기존 공격모델을 본 논문에서는 음성 주파수 및 음성속도를 공격 대상의 음성데이터와 같게 하는 단순한 조작만으로 ASV 기반 AI 음성인식 서비스의 음성인증을 우회할 수 있는 공격모델을 제안한다. 이것은 기존 연구보다 단순한 조작만으로 수행할 수 있다는 장점이 있다. 또한, 실생활에 적용되고 있는 스마트폰 음성인식 서비스를 대상으로 실험하고, 공격의 가능성 및 사이버 공격의 위협에 대한 경각심을 높이고자 한다.

이후 논문의 구성은 다음과 같다. 2장에서는 배경지식 및 관련 연구를 살펴보고, 본 연구에서의 문제를 기술한다. 3장에서는 ASV 기반 음성인식 서비스에 대한 공격모델을 제안한다. 4장에서는 실험 결과를 제시하며, 5장에서는 향후 연구 방향과 함께 결론을 맺는다.

2. 배경지식 및 관련 연구

2.1 AI 스피커에서의 음성인식 및 처리



(그림 1) AI 스피커의 전체구조 및 흐름도
(Figure 1) AI Speaker's overall structure and flowchart

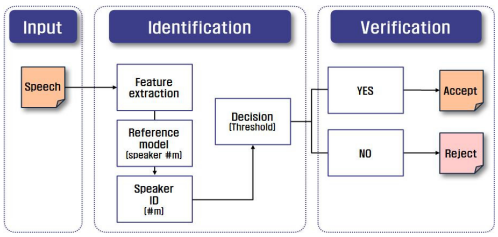
음성인식은 음성발화자와 음성인식 기기 사이의 통신 과정으로 주로 사람의 음성을 통한 정보전달 방식으로 특별한 훈련 없이도 사용자가 활용할 수 있는 인터페이스 기술이다. 일차원적 입력 수단으로 여겨지던 음성인식은 컴퓨팅 기술의 발전과 데이터의 증가로 자연어 처리 능력을 포함한 기계와 사람 간의 자연스러운 대화가 가능해짐에 따라 사용자가 이동 및 작업 중이거나 몸이 불편한 장애인 또는 노약자여도 음성만으로 제어가 가능한 기술로 편리함을 느끼는 사용자가 증가할 것이다[7].

위의 그림 1은 AI 스피커의 전체구조 및 흐름을 나타낸 것이다. 사용자가 발화하게 되면 AI 스피커가 먼저 반

응하게 된다. AI 스피커는 입출력 인터페이스 장치로 사용자의 음성을 입력받아 서버로 전달하기 이전까지 전처리 과정은 AI 스피커 내부에서 처리하며, 그 후 음성인식 디코딩 과정부터 자연어 이해(NLU: Natural Language Understanding) 및 음성합성(TTS: Text to Speech)은 클라우드 상에서 처리한다. 음성합성이 끝난 결과가 AI 스피커를 통해 출력된다[8].

2.2 자동화자 식별(ASV) 체계

자동화자 식별(ASV: Automatic Speaker Verification) 기술은 특정 화자의 음성이 누구의 음성인지 자동으로 찾아주는 기술을 말하며, ASV의 핵심인 화자 인식(speaker recognition)은 일반적으로 화자 식별(identification)과 검증(verification)의 2단계로 이루어진다[9]. 최근 ASV 기술 기반의 화자 식별 방식을 생체인증 방법으로 사용하는 스마트폰과 같은 상용 서비스 장치들이 증가하고 있다.

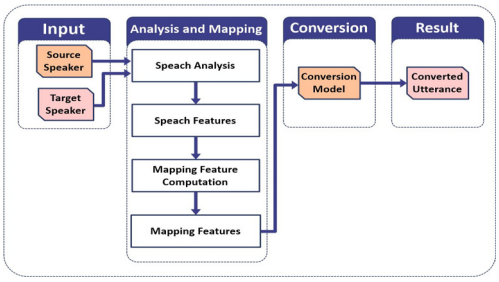


(그림 2) 자동화자 식별 체계
(Figure 2) Automatic Speaker Verification system

그림 2는 자동화자 식별 체계의 구조와 동작을 설명한 것이다. 음성 신호(input)가 입력되면 음성 신호의 특징을 추출한 후, 기존에 등록된 화자 음성 정보들(Reference model)과 비교하여 가장 유사한 화자 정보(speaker id)를 식별한다. 이후, 음성 신호와 식별된 화자 정보를 다시 비교하여 임곗값을 넘을 경우, 식별된 화자로 결정(accept)하고 그렇지 않을 때 거부(reject)한다[10].

2.3 음성 변환 및 합성

음성 변환은 화자 A(source speaker)의 음성을 화자 B(target speaker)의 음성으로 변환하는 것을 말한다. 이를 위해, 화자 B의 음성 특징들을 추출하고 추출한 특징을 활용하여 화자 A가 말한 내용의 변경 없이 화자 A의 음성 스펙트럼이나 운율의 특징을 수정한다. 이러한 과정을 통해, 화자 A의 음성이 화자 B의 음성과 거의 유사하게 변환되며, 이러한 이유로 음성 변환을 음성합성(Voice Synthesis)이라고도 한다[11].



(그림 3) 일반적인 음성 변환 체계의 동작 절차
(Figure 3) Operational Procedures of General Voice Conversion Systems

그림 3은 일반적인 음성 변환 체계의 동작 단계를 설명한다. 음성 변환 과정을 공격자의 관점에서 설명하면, 입력 단계(Input)에서는 공격자(source speaker)의 음성과 공격 대상(target speaker)의 음성을 함께 입력하고, 분석 및 맵핑(Analysis and Mapping) 단계에서는 두 개의 음성 정보는 시간 축으로 정렬하여 유사한 단어끼리 맵핑한다. 변환 단계(Conversion)에서는 정렬 맵핑된 공격자의 음성 신호를 공격 대상의 음성 신호와 유사한 형태로 스펙트럼 또는 속도 등의 조작 작업을 통해 변환 시킨다. 이러한 변환 과정을 기계학습(machine learning)을 통해 공격자는 공격 대상의 음성 신호를 원하는 대로 합성해낼 수 있게 된다[12].

2.4 관련 연구

음성인식 시스템과 서비스에 대한 사이버 위협에 관한 기존 연구들은 음성인식 장치 자체의 보안 취약성, 클라우드 기반 음성인식 시스템의 정보처리 과정에서 발생하는 개인정보 유출 및 프라이버시 침해, 그리고 관련 취약점을 예방할 수 있는 대응 기법에 관련되어 수행되었으며 대표적인 연구로는 다음과 같다.

Dibya Mukhopadhyay 등[13]은 상용 Voice-morphing tool을 사용하여 음성 변환 공격을 제안하고 ASV 알고리즘 및 인간 검증에 대한 타당성을 평가하는 방법으로 공격 가능성을 증명하였다. 하지만, 머신러닝을 활용한 음성합성보다는 적은 양의 음성데이터를 사용하지만, 훈련세트를 이용한 모델링 후 공격하는 절차가 필요했다.

Henry Turner 등[14]은 소스 음성과 공격 대상 음성 사

이의 개별 음소를 모핑하여 변환된 음성데이터를 이용하여 현실적인 스푸핑 공격에 대한 ASV의 취약점을 나타낸 공격 방법을 설명하였다. 하지만 음소 변경을 위해서는 언어의 음소 수와 온라인에서 사용 가능한 음성의 데이터세트, 공격 대상의 오디오 샘플에 대한 지식이 필요하다.

Nicholas Carlini 등[15]은 사람이 이해하거나 해석할 수 없는 음성으로 음성인식 시스템을 공격했을 때 음성인식 시스템이 작동하는 모습을 보여주면서 공격 방법을 설명하였다. 하지만, ASV 등을 활용한 사용자 인증 체계가 없는 음성인식 시스템에 대한 공격 실험으로 제한되었다.

Ji-seop Lee 등[16]은 AI 스피커의 보안성 평가하기 위해 STRIDE, LINDDUN 등 위협 모델링을 적용하여 클라우드 기반 음성인식 시스템의 정보처리 간 노출되는 정보와 사용자와 서버 간의 송·수신되는 패킷을 공격자가 가로채는 공격의 가능성을 증명하였으며, 인증서 및 공개키를 적용한다는 한정된 대응 방법을 제시하였다.

Massimiliano Todisco 등[5]은 ASVspoof 2019에서 스프핑 위협으로부터 자동화자 식별 체계를 보호하기 위한 대응책을 마련하기 위한 연구였으며, 스푸핑 시나리오와 합성, 변환 및 재생 음성이라는 세 가지 주요 스푸핑 공격에 관한 결과 및 대응 방법을 제시했다.

이외에도, 비밀번호나 핀(PIN)과 같은 인증과정이 생략된 AI 스피커 대상으로 레이저를 이용한 공격과 사람은 인지할 수 없는 음역의 소리를 이용한 음성인식 서비스를 공격하는 등 음성인식 서비스의 보안 취약성을 보여주는 여러 연구가 있었다[17, 18].

2.5 기존 연구의 제한사항

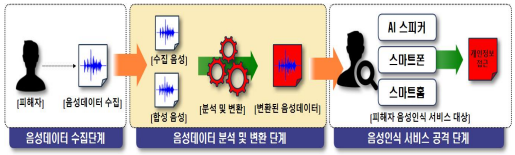
음성인식 서비스에 관한 기존 연구에서 보여준 AI 스피커 자체의 보안 취약성과 클라우드 기반의 서비스에서 발생할 수 있는 개인정보 노출 및 송·수신 패킷을 가로채는 방식은 전문가의 경험 및 특정 분야의 취약점을 식별하고자 하는 이론적 접근 방법이었다. 하지만, 딥러닝을 이용한 데이터 분석과 이미지 처리 그리고 음성합성 등은 전문가가 아니더라도 이제는 쉽게 구현할 수 있으며, 상용화된 프로그램도 대중화되고 있다. 이는 사용자 인증이 없는 기존의 음성인식 서비스와 인증기능이 있더라도 보안 취약성을 가지고 있는 음성인식 서비스에 대한 공격이 다방면적으로 가능하다는 것을 의미한다.

이러한 이유로, 본 연구에서는 기존 연구와 달리, ASV 체계가 포함된 음성인식 서비스를 대상으로 한 공격모델을 제시하고, 실험을 통해 제안 공격모델 기반의 공격기법들이 수행 가능함으로 보이는 연구를 수행함으로써 관련 사이버 공격 위협에 대한 경각심을 높이기 위함이다.

3. ASV 기반의 음성인식 서비스에 대한 공격 모델 제안

본 장에서는 ASV 기반의 음성인식 서비스에 대해 음성합성 기술을 악용한 공격모델을 제안하고 공격 수행 절차를 설명한다. 공격자는 ASV를 우회하기 위해 공격 대상(피해자)의 음성수집과 분석을 기본적인 단계로 포함한다.

제안 공격모델은 피해자의 음성데이터와 합성된 음성데이터를 분석하고, 음성데이터를 변환 후 공격 대상자의 음성인식 시스템을 공격하는 모델로 총 3단계로 구성된다(그림 4 참고).



(그림 4) 음성분석 및 합성 프로그램을 이용한 공격모델
(Figure 4) Attack Model using Voice Analysis and Synthesis Software

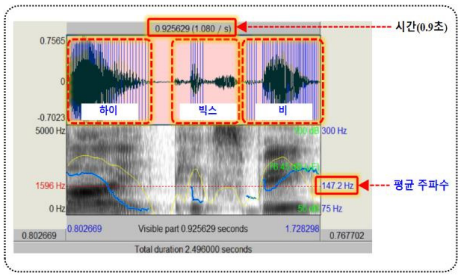
- **(1단계) 음성데이터 수집단계** : 그림 4의 공격모델과 같이, 공격자는 공격 대상자(피해자)의 음성데이터를 녹음 등의 방법으로 수집한다.
- **(2단계) 음성데이터 분석 및 변환 단계** : 공격 대상의 음성데이터를 분석하여 스펙트럼 및 운율 등 음성의 특징을 추출한다. 이후 무료 음성합성 S/W인 Prosody를 이용하여 기본으로 제공되는 성우(16명) 중 공격 대상의 성별과 나이를 고려하여 성우를 선택하고, 음성 스펙트럼 및 운율 등 유사한 데이터 값을 갖도록 변환한다. 본 연구에서는 음성분석을 위해 무료 S/W인 Praat을 이용한다.
- **(3단계) 음성인식 서비스 공격 단계** : 변환된 음성데이터를 재생하는 방식으로 음성인식 서비스를 공격한다.

4. 실험 결과 및 분석

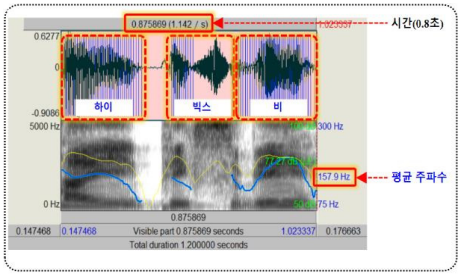
본 실험의 목적은 ASV가 포함된 음성인식 서비스를 대상으로 음성인증 및 명령어 입력을 통한 개인정보 접근 가능 여부를 실험을 통해 보이는 것이다. 이를 위해, 3장에서 제안한 공격모델을 무료 S/W를 활용하여 구현하고, 스마트폰 음성인식 서비스를 대상으로 실험을 수행했다. 실험 장비는 Intel i7 10th, RAM 16GB LG Gram 노트북을 활용하였고, 실험에 활용한 스마트폰은 우리나라 스마트폰 점유율 63%의 A사와 20%의 B사의 스마트폰 음성인식 개인비서를 대상으로 하였다.

음성분석과 합성을 위해 Praat[19]과 Prosody[20]를 사용하였다. Praat은 간단한 분석 및 전문적인 처리까지 가능한 음성분석 프로그램이다. Prosody는 다양한 성우와 성우의 다양한 감정, 디테일한 톤 조절까지 가능한 음성합성프로그램이다. 실험 단계별 수행 결과는 다음과 같다.

- **음성데이터 수집단계**: 공격 대상의 음성데이터 수집을 위해, 스마트폰의 녹음기능을 이용하여 A사와 B사의 스마트폰 개인비서의 Wake-Up Word(예: 하이 빅스비)를 저자의 음성(공격 대상)으로 녹음하였다.



[수집된 음성데이터]



[합성된 음성데이터]

(그림 5) 수집된 음성데이터의 분석 결과
(Figure 5) Voice Data Analysis Results

- **음성데이터 분석 및 변환 단계**: 수집된 음성데이터는 Praat의 분석 기능을 이용하여 그림 5과 같이 음성의 스펙트럼 및 평균 주파수와 음성속도 등 특징을 추출하였다.

한 문장에 대한 수집된 음성데이터의 평균 주파수는 147.2Hz였으며, 음성속도는 0.8초였다. 다음으로, 음성데이터 조작을 위해 Prosody의 음성합성 기능을 이용하였다. Prosody에서 제공되는 인공지능 성우는 16명으로 나이대, 성별 및 성우별 특징으로 구분되어 있다. 기본적으로 제공되는 인공지능 성우를 선택할 시 성별과 나이를 고려 30대 남성 그리고 공격 대상의 음색과 유사한 인공지능 성우를 선택했다. 이후, 합성된 음성의 평균 주파수를 확인하였다. 확인을 위해 A사와 B사의 Wake-Up Word를 기본적으로 생성하였고, 이렇게 생성된 음성의 평균 주파수는 157.9Hz로 공격 대상의 평균 주파수와 -10.7Hz 차이가 났다. 주파수의 차이는 사람마다 다르다. 성대의 길이와 두께, 탄력도에 따라 공기와 부딪힐 때 생기는 진동수가 각기 다르게 형성되고 1초 동안 진동하는 횟수 단위인 Hz에 차이가 생기는 것이다. 진동수에 따라 소리의 높낮이가 달라지기 때문에 음성의 높낮이를 조정하므로 평균 주파수 Hz를 조작할 수 있다.

[① Prosody: 음성합성 프로그램] [② 공격대상자와 같은 성별, 연령 선택]
[③ 음성의 속도 및 높낮이 설정] [④ 텍스트 입력: Wake-up Word입력]

(그림 6) Prosody를 이용한 음성 변환
(Figure 6) Voice conversion using Prosody

마지막으로, 공격 대상의 정보에 맞춰 음성데이터의 평균 주파수와 음성속도를 조작하였다. 그림 6에서와 같이, Prosody에서는 음성속도를 1% 단위로 올렸을 때 0.03초의 시간이 증가하고, 음성의 높낮이는 1을 낮췄을 때 주파수는 2Hz씩 낮춰진다.

• **음성인식 서비스 공격 단계**: 변환된 음성을 노트북 음악 재생플레이어를 이용하여 스마트폰 음성인식 서비스에 Wake-up Word를 입력한다.

(그림 7) 스마트폰 화자인식 및 프로그램 실행
(Figure 7) Smartphone Speaker Recognition and Program Execution

그림 7은 변환된 음성으로 스마트폰 화자인식 및 프로그램을 실행하는 단계를 명령어 입력, 화자인식, 프로그램 실행 등 3 단계로 보여준다. 공격 결과는 (표 1)과 같이 A사와 B사 스마트폰의 자동화자 식별 시스템이 화자를 정상 소유자로 인식하였다. 인증 성공 이후에 공격자는 변환되지 않은 음성을 사용하여 자유롭게 공격할 수 있다. 실험과정에서 발견한 특이한 사항으로, B사의 스마트폰 자동화자 식별 체계는 음성합성 간 음성의 속도 및 높낮이를 조절하지 않고도 공격 대상자와 유사한 음색을 가지고 있는 성우의 Wake-up Word를 사용하였을 때 음성인증이 되었고 정상적으로 서비스에 접근이 되어 더욱 심각한 것을 확인하였다.

(표 1) 음성인식 서비스 대상 공격 결과
(Table 1) Attack results for automatic speech recognition service

구 분	A사	B사
화자인식	소유자로 인식	소유자로 인식 (음성 변환 없이도 인식 가능)
잠금화면 (명령어 입력)	입력 가능 (명령어: 통화목록 보여줘)	입력 불가능
앱 실행	A사, B사 모두 최초 화자인식 후 변환된 음성이 아니어도 명령어 입력 및 앱 실행 가능	

본 연구에서는 스마트폰에서 지원하는 음성인식 서비스 중 개인비서를 대상으로 음성분석프로그램 및 음성합성 프로그램을 이용한 공격과 음성분석프로그램만을 이용한 공격모델로 제안했다. 이는 음성합성 및 변환을 이용하여 똑같은 음성을 만들어서 공격하는 방법이 아닌 음성의 주파수 및 속도만을 조작하여 ASV를 포함한 음성인식 서비스를 공격하는 공격기법의 가용성을 증명하였다. 또한 이를 통해 최초 음성인증 이후 입력되는 음성명령은 음성의 변조 및 조작이 필요 없이 모든 음성이 명령어로 입력될 수 있음을 확인하였다. 이는 음성인증 이후 음성으로 쉽게 개인정보(통화목록 및 개인 건강관리 앱)에 접근할 수 있음을 확인한 것이다.

향후 연구계획은 다음과 같다. 우선, 본 연구에서는 공격모델을 구현할 때 ASV 우회를 위한 여러 설정값을 찾는 과정을 수동으로 진행하였는데, 수기로 설정값을 조작하는 단순한 방안 대신 소스 음성과 공격 대상의 음성 데이터를 입력할 시 자동으로 음성데이터 값을 분석하고 주파수 및 속도를 조절할 수 있게 구성할 계획이다. 또한 음성 사이의 개별 음소를 모핑하여 ASV 기반 음성인식 서비스를 공격하는 기존 연구와 상호성능 검사를 통해 공격 단계별 절차 그리고 정확도를 분석하여 성능을 향상할 예정이다. 이외에도 같은 성별에서 변환된 음성뿐만 아니라 다른 성별의 음성 변환을 통한 음성인식 서비스에 대한 공격 가능성을 연구할 예정이다. 마지막으로, 기존에 연구했던 음성합성과 음성 변환을 이용한 음성인식 서비스 대상 공격기법에 대한 대응 기술을 연구할 예정이다.

참고문헌(Reference)

[1] JungWon Kim, YouJin Song, YongjunSung, SejungMarina Choi, “AI Speaker for the Elderly : Functional and Emotional Evaluation of AI Speaker,” Journal of Media Economics & Culture 18(4), pp.7-35, 2020.
https://doi.org/10.21328/JMEC.2020.11.18.4.7

[2] Seongwoo Kim, Chulsu Shin, BongGyu Kim, “A Study on Fighter Airplane’s Voice Command Recognition System Design and Verification Environment,” The Korean Society for Aeronautical & Space Sciences, pp.327-331, 2012.
https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02085538

[3] Ministry of Science and ICT, 「2020 Internet Use Survey Results, 2021.

[4] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis and Sébastien Marcel, “On the vulnerability of speaker verification to realistic voice spoofing,” IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), pp.1-6, 2015.
https://doi.org/10.1109/BTAS.2015.7358783

[5] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Hector Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, “ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection,” arXiv preprint arXiv:1904.05441v2, 2019.
https://arxiv.org/abs/1904.05441

[6] Gallup Korea, “2012-2021 Smartphone Utilization & Brands,” 2021

[7] Suji Baek, Youngjae Lee, “Game Interface based on Voice Recognition for Smartphone,” Korean Institute of Information Technology, Proceedings of KIIT summer Technology, pp.454-458, 2012.
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01881015

[8] Hongsu Yoon, “AI Speaker Trends,” The Korean Institute of Electrical Engineers, Vol.68, No.10, pp.16-21, 2019.

[9] Kyunhwha Kim, Buungmin So, Hajin Yu, “Forensic Automatic Speaker Identification System for Korean Speakers,” Phonetics and Speech Sciences, Vol.4, No.3, pp.95-101, 2012.
https://doi.org/10.13064/KSSS.2012.4.3.095

[10] Ravika Naike, “An Overview of Automatic Speaker Verification System,” Advances in Intelligent Systems and Computing, Vol.673, pp.603-610, 2017.
https://doi.org/10.1007/978-981-10-7245-1_59

[11] Yeongtae Hwang, Hyemin Cho, Hongsun Yang, Dongok. Won, Insoo Oh, and Seongwhan Lee, “Melspectrogram augmentation for sequence to sequence voice conversion,” arXiv preprint arXiv:2001.01401, 2020.
https://arxiv.org/abs/2001.01401

[12] Seyed Hamidreza Mohammadi, Alexamder Kain, “An Overview of Voice Conversion Systems,” Speech Communication, Vol.88, pp.65-82, 2017.
https://doi.org/10.1016/j.specom.2017.01.008

[13] Dibya Mukhopadhyay, Maliheh Shirvanian, Nitesh Saxena, “All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines,”ESORICS 2015: Computer Security - ESORICS 2015, pp 599-621, 2015.
https://doi.org/10.1007/978-3-319-24177-7_30

[14] Henry Turner, Giulio Lovisotto, and Ivan Martinovic, “Attacking Speaker Recognition Systems with Phoneme Morphing,” ESORICS 2019: Computer Security – ESORICS 2019, pp 471-492, 2019.
https://doi.org/10.1007/978-3-030-29959-0_23

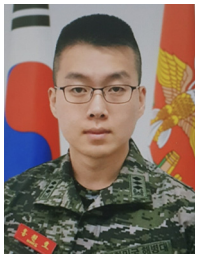
[15] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, MicahSherr, Clay Shields, David Wagner and Wenchao Zhou, “Hidden Voice Commands,” 25th USENIX Security Symposium, 2016.
https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini

[16] Ji-seop Lee, Soo-young Kang, Seung-joo Kim, “Study on the AI Speaker Security Evaluations and Countermeasure,” Journal of the Korea Institute of Information Security & Cryptology, Vol.28, No.6, 2018.
https://doi.org/10.13089/JKIISC.2018.28.6.1523

[17] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury, “Inaudible Voice Commands: The Long-Range Attack and Defense,” 15th USENIX

Symposium on Networked Systems Design and Implementation, 2018.
https://www.usenix.org/conference/nsdi18/presentation/roy
[18] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu, "Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems," 29th USENIX Security Symposium, 2020.
https://www.usenix.org/conference/usenixsecurity20/presentation/sugawara
[19] Byeongon Yang, "Theory and Substance of Speech Alalysis using PRRAT," Mansoo Publishing Company, 2010.
[20] Prosody, https://www.prosody-tts.com/

◐ 저 자 소 개 ◑



홍 천 호(Cheonho Hong)
2010년 침례신학대학교 졸업 (학사)
현재 국방대학교 국방관리대학원 컴퓨터공학/사이버전협동전공 석사과정
관심분야 : 네트워크 보안, 디지털포렌식 등
E-mail : marinehongs@mnd.go.kr

조 영 호(Youngho Cho)
1998년 공군사관학교 졸업 (학사)
2006년 연세대학교 졸업 (공학석사)
2013년 University of Maryland, College Park, USA 졸업 (공학박사)
현재 국방대학교 국방관리대학원 국방과학학과 컴퓨터공학/사이버전협동전공 주임교수
관심분야 : 네트워크 보안, 스테가노그래피 봇넷, 신뢰 메커니즘, 블록체인, 디지털 포렌식, AI 보안 등
E-mail : youngho@kndu.ac.kr