ASR 시스템에 대한 오디오 적대적 공격 연구 동향 분석

김나현 [1], 이연준 [2]
[1] 한양대학교 ERICA 컴퓨터학부 학부생 [2] 한양대학교 컴퓨터공학과 교수
Ksknh7@hanyang.ac.kr, yeonjoonlee@hanyang.ac.kr

A Survey on Adversarial Attacks Against ASR Systems

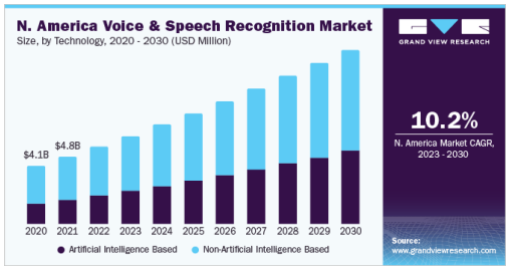
Na Hyun Kim, Yeon Joon Lee[2]
[1]Dept. of Computer Science, Han-Yang University ERICA
[2]Dept. of Computer Science, Han-Yang University

요 약

오디오 적대적 공격 연구는 최근 몇 년 동안 빠르게 발전해 왔다. 이전에는 음성 신호를 직접 수정하거나 추가하여 공격을 수행하는 방법이 일반적이었지만 최근에는 딥러닝 모델을 이용한 적대적 공격 기술이 주목을 받고 있다. 이러한 적대적 공격은 현재 다양한 분야에 널리 쓰이는 ASR 시스템에 심각한 보안 위협이 될 수 있다. 이에 본 논문에서는 현재까지의 음성신호 적대적 공격 기술과 방어기술의 연구 흐름을 분석하여 더욱 강건한 ASR 시스템을 구축하는 데 기여하고자 한다.

1. 서론

ASR(자동음성인식) 시스템은 인간의 음성 언어를 컴퓨터가 이해할 수 있는 텍스트 형태로 변환해주는 기술로, 스마트 스피커를 통해 전화를 걸거나 음성 명령으로 각종 기기들을 동작하는 등 다양한 분야에서 활용되고 있다. Grand View Research 에서 발표한 음성 인식 시장 규모, 점유율 및 동향 보고서에 따르면, 전 세계 음성 인식 시장 규모는 2022 년 171.7 억 달러로 평가되었으며 2023 년부터 2030 년까지 연평균증가율(CAGR) 14.9%로 성장할 것으로 예상된다[1].



(그림 1) 음성 인식 시장 규모 연평균 증가율 예상치

음성인식시장의 규모가 커짐에 따라 ASR 시스템을 대상으로 한 공격기법이 많이 등장했다. 특히, 최근 활발히 연구되고 있는 오디오 적대적 공격은 오늘날 널리 사용되는 ASR 시스템의 안전성과 신뢰성에 큰 위협이 될 수 있어 더욱 연구가 필요한 분야이다. 오디오 적대적 공격 기법은 음성 입력에 작은 조작을 가하여 인간이 듣기에는 정상적인 음성으로 인식되지만 음성 인식 모델은 잘못 해석하도록 만들어 음성 인식 정확도를 크게 저하시키는 공격 방법이다. 예를 들어, 공격자는 원래 음성인 " Close the window"에 섭동을 추가하여 ASR 시스템이 “Open the window” 로 인식하는 오류를 발생시킬 수 있다. 최근 ASR 시스템에 대한 더욱 강건하고 은밀한 공격과 그에 대한 방어 기법 연구들이 많이 진행되고 있다. 본 논문에서는 이러한 연구의 최근 흐름에 대해 살펴보고 앞으로의 연구 방향을 제시하고자 한다.

2. 공격기술

오디오 적대적 공격은 주로 원래 음성신호에 섭동(Perturbation)을 추가하는 방식으로 이뤄지며, 음성인식 모델의 오인식을 유발하도록 하는 대한 적절한 섭동은 FGSM, PGD 등과 같은 최적화 알고리즘을 사용하여 찾는다. 이는 이미지 적대적 공격에서 사용하던 기법을 변형하여 적용한 것으로, 대상 모델의 그래디언트 정보를 이용해 섭동을 생성하고, 생성된 섭동이 대상 모델의 손실 함수를 최대화하도록 하여 해당 입력에 대해 잘못된 결과를 유도하는 방법이다. 초기 오디오 적대적 공격 연구들은 단순히 이미지 분야에서 사용되던 섭동 생성기법을 적용한 연구들이 대부분이었지만 추후 이러한 섭동이 사람들의 눈에 띄지 않고 자연스럽게 들릴 수 있도록 하는 연구들이 진행되었다. 악의적으로 만들어진 적대적 음성 명령을 노래에 숨겨 사람들이 섭동을 눈치채지 못하면서 ASR 시스템을 효과적으로 제어할 수 있도록 하는[2] 연구가 그 예이다. 이후 연구는 더욱 은밀한 공격을 위해 인간의 눈에 띄지 않는 최적의 섭동을 찾아내는 데에

초점을 맞춘다. 한 연구는 이를 위해 회귀 분석으로 인간 인식 프로세스를 역 설계하여 이 회귀 모델을 통해 인간의 인식 편차 예측 값을 최소화하는 섭동을 찾는 방식[3]을 제안한다. 또한, 기존의 적대적 공격 방식과 달리 섭동 생성과 같은 음성 데이터에 대한 수정을 필요로 하지 않는 오디오 Replaying 공격도 제안되었다. 단순히 공격 대상자의 음성 명령을 녹음한 후, 이을 다시 재생하여 공격하는 간단한 방식으로 더 쉽게 사용될 수 있다는 장점이 있다. 쉬운 공격 방법에 비해 탐지하기 쉽지 않아 Replaying 탐지에 관련한 많은 연구[4, 7]들이 진행되었다.

3. 방어기술

오디오 적대적 공격에 대한 방어 기법 연구는 이미지 도메인에서 진행된 선행 연구들을 바탕으로 한다. 대표적으로 적대적 섭동을 걷어 내기 위한 입력 변환 기반 방어 기법이 있다. JPEG 압축, 양자화 등과 같은 이미지 변환을 사용하여 섭동을 무력화 시킴으로써 양성 이미지를 복구하는 기존의 이미지 적대적 공격 방어 기법을 차용하여 원본 신호와 변환을 거친 신호의 ASR transcription 의 차이를 분석하여 적대적 입력을 탐지하는 연구[5]가 진행된 바 있다. 이는 원본 오디오는 변환에 강건한 반면 적대적 예제는 작은 변환에도 쉽게 깨지는 특성을 이용한 것이다. 이미지 도메인의 영향을 받은 두번째 흐름으로, 모델 학습 중에 오디오 적대적 예제들을 생성해 이를 학습 데이터셋에 포함시켜 Adversarial Robustness 를 달성하는 적대적 학습(Adversarial Training)을 오디오 도메인에 적용한 연구[6]이다. 최근 연구들은 기존의 적대적 학습을 보완하여 섭동에 더욱 강건한 음성인식 모델을 만들고자 노력하고 있다. 또한, 녹음된 음성을 재생하여 공격하는 Replaying 기법을 탐지하는 기법에 관한 연구들도 꾸준히 발표되고 있다. 재생된 음성의 고주파 대역에서 나타나는 특징을 사용하거나[4], 실제 인간 음성과 스피커로 재생되는 음성 간의 Spectral power 와 Decay pattern 의 차이를 이용하여 공격을 탐지[7]하는 등의 Liveness detection 연구가 있다. 한편, ASR 시스템에 대한 적대적 예제의 존재는 불가피 하다고 보고 이러한 적대적 예제가 인간이 듣기에 자연스럽지 않도록 만들고자 하는 노력도 있다[8]. ASR 시스템에 Psychoacoustic filtering 과 Band-pass filtering 을 적용하여 인간의 청각 시스템과 더욱 비슷하게 수정하여 인간이 잘 인식하지 못하는 은밀한 적대적 예제가 생성될 수 없도록 하는 것이다. 이러한 연구는 적대적 예제는 Non-robust 한 Feature 에 기인한다고 주장한 이전 연구[9]의 아이디어를 오디오 적대적 사례에 적용한 것이다.

4. 결론

본 논문에서 ASR 시스템에 대한 오디오 적대적 공격의 연구 흐름과 이에 대한 최근의 방어 기법에 대하여 살펴보았다. 오디오 적대적 공격 분야는 이미지 도메인에 비해 아직 밝혀지지 않은 공격의 여지가 비교적 많으며 방어 기법에 대한 연구 또한 소수에 불과하다. 또한, ASR 시스템의 보급률은 점진적으로 더욱 증가할 것이기 때문에 앞으로도 해당 분야의 연구가 활발히 진행될 것으로 예상된다. 특히 공격 기법 측면에서는 인간의 인식 시스템을 고려하여 더욱 은밀한 섭동을 생성하여 공격을 정교화 하고자 하는 연구들과, 발견되지 않은 ASR 시스템의 취약점을 찾거나 기존의 탐지 기법들을 우회하여 새로운 공격벡터를 제시하는 연구들이 지속될 것으로 보인다. 향후 방어 기법 연구로는 기존의 공격 벡터를 제거 및 탐지하는 연구와 병행하여 ASR 시스템을 인간의 인식 시스템과 비슷하게 디자인하여 은밀한 적대적 예제 생성을 저지하는 새로운 흐름의 연구들이 필요할 것으로 보인다.

사사

본 연구는 2023 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음(2018-0-00192)

참고문헌

[1] “Voice And Speech Recognition Market Size, Share & Trends Report ”, Grand View Research Apr 2021

[2] Xuejing Yuan et al. “CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition” USENIX Security Aug 2018

[3] Rui Duan et al. “Perception-Aware Attack: Creating Adversarial Music via Reverse-Engineering Human Perception” ACM CCS Jul 2022

[4] Marcin Witkowski et al. “Audio Replay Attack Detection Using High-Frequency Features” Stockholm, Sweden INTERSPEECH Aug 2017

[5] Shehzeen Hussain et al. “WaveGuard: Understanding and Mitigating Audio Adversarial Examples” USENIX Security 2021

[6] Sining Sun et al. “Adversarial regularization for attention based end-to-end robust speech recognition” IEEE/ACM Transactions on Audio, Speech, and Language Processing Aug 2019

[7] Muhammad Ejaz Ahmed et al. “Void: A fast and light voice liveness detection system” USENIX Security 2020

[8] Thorsten Eisenhofer et al. “DOMPTEUR: Taming Audio Adversarial Examples” USENIX Security 2021

[9] Andrew Ilyas et al. “Adversarial Examples Are Not Bugs, They Are Features.” In Advances in Neural Information Processing Systems (NeurIPS) 2019