(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0094097
(43) 공개일자 2023년06월27일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G10L 25/30* (2013.01)
(52) CPC특허분류
G06N 3/088 (2023.01)
G10L 25/30 (2013.01)
(21) 출원번호 10-2022-0025778
(22) 출원일자 2022년02월28일
심사청구일자 2022년02월28일
(30) 우선권주장
1020210182992 2021년12월20일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
윤지원
서울특별시 용산구 한강대로43길 8, 102동 1401호
최형준
부산광역시 동래구 충렬대로202번길 70, 226호 (수안동, 성림@)
(74) 대리인
김등용

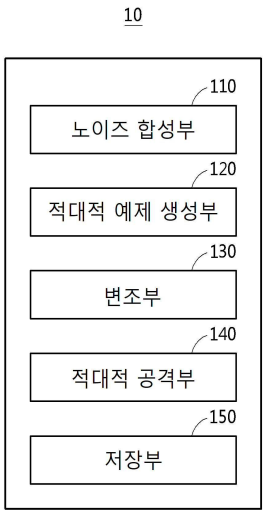
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **환경 노이즈를 이용한 적대적 예제 생성 장치 및 방법**

(57) 요 약

적대적 예제 생성 장치, 적대적 공격 장치, 및 적대적 공격 방법이 개시된다. 상기 적대적 공격 장치는 원본 음성 파일과 노이즈 파일을 결합하여 노이즈가 부가된 음성 파일을 생성하는 노이즈 합성부, 소정의 적대적 예제 생성 알고리즘을 이용하여 상기 노이즈가 부가된 음성 파일로부터 적대적 예제를 생성하는 적대적 예제 생성부, 상기 적대적 예제에 해당하는 음성 파일에 소정의 변조(Modulation) 기법을 적용하여 변조된 음성 신호를 생성하는 변조부, 및 상기 변조된 음성 신호를 송신하는 적대적 공격부를 포함한다.

대 표 도 - 도6



이 발명을 지원한 국가연구개발사업

과제고유번호	1711126336
과제번호	2020-0-00913-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보보호핵심원천기술개발(R&D,정보화)
연구과제명	무선 은닉채널 위험성 검증 연구
기 여 율	1/1
과제수행기관명	고려대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

원본 음성 파일과 노이즈 파일을 결합하여 노이즈가 부가된 음성 파일을 생성하는 노이즈 합성부;

소정의 적대적 예제 생성 알고리즘을 이용하여 상기 노이즈가 부가된 음성 파일로부터 적대적 예제를 생성하는 적대적 예제 생성부;

상기 적대적 예제에 해당하는 음성 파일에 소정의 변조(Modulation) 기법을 적용하여 변조된 음성 신호를 생성하는 변조부; 및

상기 변조된 음성 신호를 송신하는 적대적 공격부를 포함하는 적대적 공격 장치.

청구항 2

제1항에 있어서,

상기 노이즈 합성부는 상기 원본 음성 파일과 상기 노이즈 파일의 재생 시간이 같도록 상기 원본 음성 파일과 상기 노이즈 파일 중 길이기 긴 음성 파일의 길이를 줄이는,

적대적 공격 장치.

청구항 3

제1항에 있어서,

상기 노이즈 파일은 운행 중인 차량 내에서 발생하는 노이즈를 녹음한 음성 파일인,

적대적 공격 장치.

청구항 4

제1항에 있어서,

상기 변조부는 상기 변조된 음성 신호의 중심 주파수가 공격 대상인 모바일 장치가 위치한 차량에서 수신 중인 라디오 신호의 중심 주파수와 동일하도록 주파수 변조(Frequency Modulation)를 수행하는,

적대적 공격 장치.

청구항 5

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 적대적 공격 방법에 있어서,

원본 음성 파일과 노이즈 파일을 결합하여 노이즈가 부가된 음성 파일을 생성하는 단계;

소정의 적대적 예제 생성 알고리즘을 이용하여 상기 노이즈가 부가된 음성 파일로부터 적대적 예제를 생성하는 단계;

상기 적대적 예제에 해당하는 음성 파일에 소정의 변조(Modulation) 기법을 적용하여 변조된 음성 신호를 생성하는 단계; 및

상기 변조된 음성 신호를 송신하는 단계를 포함하는 적대적 공격 방법.

청구항 6

제5항에 있어서,

상기 노이즈가 부가된 음성 파일을 생성하는 단계는 상기 원본 음성 파일과 상기 노이즈 파일의 재생 시간이 같도록 상기 원본 음성 파일과 상기 노이즈 파일 중 길이기 긴 음성 파일의 길이를 줄이는 단계를 포함하는,

적대적 공격 방법.

청구항 7

제5항에 있어서,

상기 노이즈 파일은 운행 중인 차량 내에서 발생하는 노이즈를 녹음한 음성 파일인,

적대적 공격 방법.

청구항 8

제5항에 있어서,

상기 변조된 음성 신호를 생성하는 단계는 상기 변조된 음성 신호의 중심 주파수가 공격 대상인 모바일 장치가 위치한 차량에서 수신 중인 라디오 신호의 중심 주파수와 동일하도록 주파수 변조(Frequency Modulation)를 수행하는 단계를 포함하는,

적대적 공격 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 적대적 예제 생성 방법에 관한 것으로, 특히 환경 노이즈를 원본 음성 파일에 합성한 후에 적대적 예제를 생성하는 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근에, 아이폰(iPhone)의 시리(Siri)와 갤럭시(Galaxy)의 빅스비(Bixby)와 같은 다양한 자동 음성 인식(Automatic Speech Recognition, ASR) 시스템이 소개되었다. 2019년에, 시장 조사 기관인 eMarketer는 미국 내의 AI 스피커 사용자는 2017년 8,000만명에서 2020년 1억 1,770만명으로 증가하고, 2021년에는 1억 2,270만명으로 증가할 것으로 예상한 바 있다.

[0003] 따라서, 사용자의 편의를 위한 다양한 기능들이 추가되고 있다. 몇몇 예에서, 메시지를 전송하는 기능, 전화를 거는 기능, 및 가정 내의 기기(예컨대, 가전 제품, 에어컨, 보일러 등)를 제어하는 기능 뿐만 아니라 차고 문이나 현관 문을 여는 기능이 제공되고 있다. 이러한 기능들은 사용자에게 편리함을 주지만, 해킹이 된다면 매우 위험할 수 있다.

[0004] ASR 시스템을 대상으로 하는 공격의 중요한 점은 희생자(victim)의 인지가 없는 상태에서 진행된다는 것이다. 만약 희생자가 공격을 인지한다면, 희생자는 방어를 하거나 공격에 대항하는 행동을 취할 것이다. 이에 따라, 적대적 공격(adversarial attack)의 개념이 출현하였다. 적대적 공격은 사람들이 보거나 들을 수 있는 노멀 데이터(normal data)이지만, 적대적 공격은 ASR 시스템이 완전히 다른 판단을 하도록 만든다. 따라서, 희생자가 이를 알아채는 것은 매우 어렵다.

[0005] 따라서, 본 발명에서는 희생자가 눈치채지 못하는 시나리오를 제안한다. 제안하는 시나리오는 많은 사람들이 차를 탈 때 라디오를 듣는다는 사실에 기초한다. 자동차는 보통 운행을 한다. 라디오의 주파수는 미미하게 매치되지 않고, 이러한 경우에, 라디오는 다른 채널로 변경될 수 있다. 게다가, 터널과 같이 라디오 신호의 전파가 원활하게 공급되지 않는 장소를 통과할 때에는 채널이 변경될 수 있다. 이러한 현상들은 종종 발생하며, 따라서 운전자는 보통 큰 신경을 기울이지는 않는다. 이러한 현상을 이용하면 차량의 라디오로 적대적 예제를 전송할 수 있다. 그리고, 운전자가 알지 못하는 상태에서 이를 운전자의 휴대 전화나 내비게이션의 ASR 시스템을 대상으로 하는 공격에 이용하여, 차고의 문을 개방하거나 경로를 변경할 수 있다. 또한, 본 발명에서는 스피커와 마이크의 거리가 멀 때 공격이 실패하는 현상을 해결하는 방법을 제안한다.

발명의 내용

해결하려는 과제

[0006] 본 발명이 이루고자 하는 기술적인 과제는 환경 노이즈를 이용한 적대적 예세 생성 장치 및 방법과 이를 이용한

적대적 공격 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0007] 본 발명의 일 실시예에 따른 적대적 공격 장치는 원본 음성 파일과 노이즈 파일을 결합하여 노이즈가 부가된 음성 파일을 생성하는 노이즈 합성부, 소정의 적대적 예제 생성 알고리즘을 이용하여 상기 노이즈가 부가된 음성 파일로부터 적대적 예제를 생성하는 적대적 예제 생성부, 상기 적대적 예제에 해당하는 음성 파일을 소정의 변조(Modulation) 기법을 이용하여 변조된 음성 신호를 생성하는 변조부, 및 상기 변조된 음성 신호를 송신하는 적대적 공격부를 포함한다.

[0008] 또한, 본 발명의 일 실시예에 따른 적대적 공격 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되고, 원본 음성 파일과 노이즈 파일을 결합하여 노이즈가 부가된 음성 파일을 생성하는 단계, 소정의 적대적 예제 생성 알고리즘을 이용하여 상기 노이즈가 부가된 음성 파일로부터 적대적 예제를 생성하는 단계, 상기 적대적 예제에 해당하는 음성 파일을 소정의 변조(Modulation) 기법을 이용하여 변조된 음성 신호를 생성하는 단계, 및 상기 변조된 음성 신호를 송신하는 단계를 포함한다.

발명의 효과

[0009] 본 발명의 실시예에 따른 적대적 예제 생성 장치 및 방법과 이를 이용한 적대적 공격 장치 및 방법에 의할 경우, 환경 노이즈를 적대적 공격을 하기 전에 인자로 넣어서 환경 노이즈를 학습시킴으로써, 스피커와 마이크의 거리가 멀어졌을 경우와 주변에 소음이 있는 환경에서도 적대적 공격이 정상적으로 작동하는 효과가 있다.

[0010] 또한, 이를 통해 적대적 공격에 대한 시스템의 취약점을 파악하고 이에 기초하여 적대적 공격에 강건한 시스템을 개발할 수 있다.

도면의 간단한 설명

[0011] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명에서 제안하는 공격 시나리오를 설명하기 위한 개념도이다.

도 2는 환경 노이즈를 이용하여 적대적 예제를 생성하는 과정을 설명하기 위한 도면이다.

도 3은 Sugar, bad guy, 및 합성 신호의 크기를 비교한 그래프이다.

도 4는 거리에 따른 정확도 비교 결과를 도시한 그래프이다.

도 5는 환경 노이즈와 white Gaussian noise 를 이용한 적대적 예제의 정확도를 비교한 그래프를 도시한다.

도 6은 본 발명의 일 실시예에 따른 적대적 예제 생성 장치의 기능 블럭도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0013] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0014] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0015] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어"

있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0016] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0017] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0018] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0019] 우선, 본 발명의 배경이 되는 기술에 대해 설명한다.

[0020] 자동 음성 인식(Automatic Speech Recognition, ASR)

[0021] 자동 음성 인식(ASR)은 인간의 음성을 인식하고 이를 텍스트 형식으로 표현하는 기술을 의미한다. 대표적인 ASR 시스템은 애플(Apple)의 시리(Siri), 삼성(Samsung)의 빅스비(Bixby), 구글 어시스턴트(Google Assistant), IBM Watson Speech to Text, 그리고 Mozilla DeepSpeech 등이 있다. 본 명세서에서 기술된 실험에서는 ASR 시스템으로 DeepSpeech를 선택하였다.

[0022] 음성 적대적 예제(Audio Adversarial Examples)

[0023] DeepSpeech를 대상으로 하는 적대적 공격 알고리즘(adversarial attack algorithm)이 제안된 바 있다(Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE Security and Privacy Workshops (SPW), pages 1-7. IEEE, 2018.). 이러한 공격은 Nicholas Carlini 등의 공격(Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In 25th {USENIX} Security Symposium ({USENIX} Security 16), pages 513-530, 2016.)을 발전시킨 것이다.

[0024] Over-the-Air 적대적 공격(Over-the-Air Adversarial Attacks)

[0025] 생성되 오디오 적대적 예제(autio adversarial examples)을 스피커와 마이크를 통하여 전송하기 이해서는 반드시 공기를 통과하여 전파되어야 한다. 그러나, 아무런 액션을 취하지 않고 적대적 예제를 스피커를 통해 플레이하고 이를 마이크로 수신한다면, 공격은 적절히 성공하지 못할 것이다. 이는 over-the-air 기술이 중요한 이유이다. RIRs(Room Impulse Responses)를 이용하는 over-the-air 기술이 있다(Lea Sch¨onherr, Thorsten Eisenhofer, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems. In Annual Computer Security Applications Conference, pages 843-855, 2020.). 또한, Metamorph를 이용하여 over-the-air를 극복하거나(Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In Network and Distributed Systems Security (NDSS) Symposium, 2020.) 공격 알고리즘 자체를 강하게 변화시킴으로써 over-the-air를 극복하는 방법(Mohammad Esmaeilpour, Patrick Cardinal, and Alessandro Lameiras Koerich. Towards robust speech-to-text adversarial attack. arXiv preprint arXiv:2103.08095, 2021.)이 있다.

[0026] 이하에서는, 본 발명에서 제안하는 기법들에 대하여 설명한다.

[0027] 공격 시나리오(Attack Scenario)

[0028] 공격자(attacker)는 희생자가 듣고 있는 라디오 주파수(radio frequency)에서 더 강한 신호(a stronger signal)를 송신함으로써 신호를 하이잭(hijack)할 수 있다. 신호는 음악(music)이거나 단순히 덜컹거리는 소리일 수 있다. 중요한 점은 신호가 숨겨진 사운드(hidden sounds)를 포함하는 적대적 예제라는 것이다. 음악이 갑자기 라디오에서 플레이되거나 다른 소리가 플레이된다면 희생자는 눈치챌 수 있으나, 덜컹거리는 소리가 나는 경우에는 신호가 좋지 않다고 생각할 수 있기 때문에 음악 소리보다는 단순히 덜컹거리는 소리가 더 좋은 효과를 낼 수도 있다.

[0029] 공격 시나리오는 희생자가 듣고 있는 라디오를 알고 있는지 여부에 의존한다. 희생자가 듣고 있는 라디오를 공격자가 알고 있다면 공격자는 주파수를 공격할 수 있다. 그러나, 희생자가 듣고 있는 라디오를 공격자가 알고 있지 않다면 공격자는 모든 알려진 라디오 주파수를 공격함으로써 공격을 시도할 수 있다. 공격자는 희생자의 차량을 추적하는 동안에 신호를 송신함으로써 공격을 수행하거나 주차장 내에 있는 희생자에게 은밀하게 신호를 송신함으로써 공격을 수행할 수 있다. 이러한 공격 시나리오는 전통적인 공격 시나리오가 갖는 문제를 해결할 수 있다. 희생자가 YouTube 링크를 클릭함으로써 YouTube를 시청해야 하는 기존 공격 시나리오와는 다르게, 본 발명에서 제안하는 공격 시나리오는 희생자의 어떠한 행동도 요구하지 않는다. 즉, 희생자는 라디오를 듣고 있는 것만으로도 충분하다. 공격 시나리오에 대한 개략적인 설명은 도 1에 도시되어 있다.

[0030] FM 노이즈에 대한 저항(Tolerance to FM Noise)

[0031] 라디오 송수신기나 수신기는 회로들과 안테나들로 구성되어 있기 때문에, 생성된 라디오 신호(radio wave)는 노이즈에 의해 영향을 받을 것으로 예상할 수 있다. 추가적으로, 다른 환경적인 노이즈(environmental noise)가 복조 과정(demodulation process) 동안에 신호에 영향을 줄 수 있다. 적대적 예제는 적은 노이즈를 추가함으로써 생성되기 때문에, 적대적 예제는 노이즈에 매우 민감할 수 있다. 따라서, 적대적 예제가 노이즈에 대한 저항을 갖는지 여부를 알기 위한 간단한 실험을 진행하였다. 하나의 컴퓨터로 적대적 예제를 포함하는 FM 라디오 신호(FM radio waves)를 송신하고 다른 컴퓨터로 라디오 신호를 수신/복조하는 실험을 진행하였다. 실험을 진행한 이후에, 주파수 복조된 신호가 ASR 시스템에 의해 인식되는지 여부를 관찰하였다.

[0032] 라디오 방송국의 신호보다 더 강한 FM 신호(Winning the Signal of a Radio Station)

[0033] 많은 사람들이 자동차 안에서 라디오를 듣는다. 제안하는 공격 시나리오가 성공하기 위해서는, 희생자는 방송되는 라디오 신호가 아닌 공격자가 송신하는 보이스(voice)를 듣게 만들 필요가 있다. 이를 위해, 라디오 방송국(radio station) 보다 더 강한 FM 신호를 송신할 것을 제안한다. 이러한 방법에 따라, 희생자는 오직 공격자가 송신한 FM 신호만을 수신할 것이다. 이러한 현상을 캡쳐 효과(capture effect)라 불린다. 두 개의 라디오 신호를 수신하는 FM 라디오 시스템에서, 하나의 신호가 다른 신호보다 더 강하다면, 다른 신호는 제거될 것이다. 주파수 변조된 신호들 $V_{br}(t)$, $V_{adv}(t)$과 두 신호들의 합 $V_{all}(t)$은 다음과 같이 표현된다. 여기서, $V_{br}(t)$은 라디오 방송 시스템 신호, 즉 원래의 라디오 방송 신호를 의미하고, $V_{adv}(t)$는 공격 시스템의 신호, 즉 적대적 공격이 포함된 신호를 의미한다.

[0034] [수학식 1]

[0035] $$V_{br}(t) = A_{br}\cos\left(2\pi f_c t + 2\pi k_f \int x_{br}(t)dt\right)$$

[0036] [수학식 2]

[0037] $$V_{adv}(t) = A_{adv}\cos\left(2\pi f_c t + 2\pi k_{adv} \int x_{adv}(t)dt\right)$$

[0038] [수학식 3]

[0039] $$V_{all}(t) = V_{br}(t) + V_{adv}(t)$$

[0040] 상기 수학식에서, A와 f_c는 공격 대상인 방송 신호의 크기(amplitude)와 중심 주파수(center frequency)를 의미하고, x(t)는 라디오 송수신기(radio transceiver)로부터의 음파(sound wave)를 의미한다. 공격 시나리오에서, $V_{all}(t)$는 희생자 차량의 라디오가 수신하는 신호를 의미한다. 따라서, A_{br} 보다 큰 A_{adv}를 송신한다면, 차량의 라디오에서 복조되는 신호는 오직 타겟이 된 오디오 적대적 예제 $x_{adv}(t)$ 뿐이다.

[0041] 다른 노이즈에 대한 저항(Tolerance to Other Noise)

[0042] 레코딩 과정에서 몇몇 노이즈들이 섞일 수 있다. 컨텐츠에 대한 실험과 결과는 Hiromu Yakura 등의 연구(Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. arXiv preprint arXiv:1810.11793, 2018.)에서 찾아볼 수 있다. Hiromu Yakura 등은 Band-pass filter, Impulse Response, 및 White Noise(또는 White Gaussian noise)의 세가지 방법을 이용하였다.

[0043] 해당 방법들은 화이트 노이즈(white noise)와 같은 노이즈에는 효과적일 수 있다. 그러나, 이러한 방법은 이용하기가 어렵고 의도한 만큼 잘 작동하지 않는다. 따라서, 환경 노이즈(environmental noise)를 제안한다. 주변 환경의 소리를 녹음하여 환경 노이즈를 획득할 수 있다. 이는 매우 간단한 방법이다. 제안하는 방법의 정확성은 몇몇의 실험을 통해 증명될 수 있다.

[0044] 첫번째 실험은 white Gaussian noise에 의해 생성된 적대적 예제, 제안하는 환경 노이즈를 이용하여 생성된 적대적 예제, 및 두 가지 방법을 모두 사용하여 생성된 적대적 예제를 비교하는 것이다. 도 2에는, 환경 노이즈를 이용하여 적대적 예제를 생성하는 과정이 도시되어 있다.

[0045] 두번째 실험은 스피커와 마이크 사이의 거리에 따른 정확도의 변화 정도를 관찰하는 것이다.

[0046] 세번째 실험은 자동차가 운행 중임을 가정한다. 운행 중인 자동차의 소리를 녹음한 환경 노이즈를 이용하여 생성된 적대적 예제와 Gaussian white noise를 이용하여 생성된 적대적 예제가 운행중인 차량에서 플레이되는 상황을 가정하고 실험이 진행되었다.

[0047] 모든 방송 FM에 대한 공격(Attack to All Broadcast FM)

[0048] 공격 시나리오에서, 희생자는 특정인이거나 불특정 다수인이 될 수 있다. 대부분의 경우, 희생자가 듣고 있는 라디오를 알 수는 없다. 따라서, 희생자와는 무관하게 공격을 성공하기 위해서는, 특정 FM 주파수를 설정할 필요가 있다. 라디오 방송국의 모든 주파수를 대상으로 공격을 시도할 수 있다. 대한민국에서, 모든 주파수는 88Hz와 100Hz 사이에 모여 있다. 따라서, 위 모든 주파수를 대상으로 공격을 수행할 수 있다.

[0049] 이하에서는, 본 발명에 의한 실험의 평가에 대해 기술한다.

[0050] 공격 준비(Attack Preparation)

[0051] 제안하는 전체적인 실험 모델은 DeepSpeech(Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.)이다. 그리고, 오디오 적대적 공격(Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE Security and Privacy Workshops (SPW), pages 1-7. IEEE, 2018.)을 이용하였다. 소스 코드는 웹페이지(Audio adversarial examples. https://github.com/carlini/audio_adversarial_examples.)에 공개되어 있다. HackRF One(Hackrf one. https://greatscottgadgets.com/hackrf.)이 사운드를 전송하기 위해 이용되었다. 녹음은 iPhone 11의 기본 녹음 앱을 이용하여 진행되었다. iPhone 의 기본 포멧은 m4a extension 이다. 따라서, FFmpeg(Ffmpeg. https://www.ffmpeg.org.)를 이용하여 wav extension으로 변환할 필요가 있다. 적대적 예제는 16000 Hz이고, 이는 48000 Hz로 전송되었으며, 이를 변환하기 위해 sox를 이용하였다.

[0052] DeepSpeech에 대한 공격(Attack DeepSpeech)

[0053] 제안하는 공격 시나리오를 성공시키기 위해, 첫번째로 해야할 일은 생성된 신호가 라디오 방송국의 신호를 이길 수 있는지 여부(즉, 더 강한 신호인지 여부)를 체크하는 것이다. Billie Eilish의 bad guy를 30dB로 송신하고 이를 잘 들을 수 있는지 체크하였다. 그리고, Maroon 5의 Sugar를 60dB로 송신할 때, bad guy는 잘 들리지 않고 Sugar 만 들을 수 있었다. 이를 보다 정확하게 알아보기 위하여, 각 상황이 녹음되었고 신호(waves)는 비교되었다. 이는 도 3에 도시되어 있다. Sugar와 모두(both)를 비교하는 파형을 살펴보면, bad guy는 완전히 보이지 않으며 작은 노이즈만을 가지고 있음을 볼 수 있다. 이러한 노이즈는 적대적 공격에 중요한 영향을 미치지는 않는다.

[0054] FM 라디오의 필드 크기(field strength)는 60이기 때문에, 적대적 공격을 위해서는 이보다 더 강한 크기를 갖는 신호를 이용하여야 한다.

[0055] 적대적 예제가 FM을 통해 생성된 노이즈에 저항적인지 여부를 살펴본다. 첫째, 적대적 예제를 특정 주파수로 송신한다. 송신 신호는 캡쳐되고 다른 컴퓨터에 의해 저장된다. DeepSpeech에 의해 인식되었고, 이는 우리가 원하

는 바와 일치함을 보았다.

[0056] 환경 노이즈가 white Gaussian noise 보다 정확한 것으로 관찰되었기 때문에, 거리에 따른 정확성의 변화를 실험한다. cute와 환경 노이즈를 이용하여 생성된 적대적 예제와 단순히 cute만을 이용하여 생성된 적대적 예제를 이용하여 실험이 진행되었다. 스피커와 휴대폰의 거리를 넓히면서 실험을 진행하였다. 공격 시나리오에서, 희생자의 다바이스는 모바일 폰일 것이며, 따라서 실험은 모바일 폰에 내장된 마이크를 이용하여 진행되었다. 각각의 적대적 예제를 녹음한 후 녹음된 적대적 예제를 DeepSpeech에 플레이하고, DeepSpeech에 의해 인식된 문장과 공격을 시도했던 문장에 대하여 환경 노이즈에 대한 저항을 비교하였다. 실험 결과는 도 4의 그래프에 요약되어 있다. 거리가 가까우면 환경 노이즈를 학습한 결과가 약간 덜 정확하지만, 거리가 멀어짐에 따라 환경 노이즈를 학습한 결과가 더 나음을 알 수 있다.

[0057] 다음으로, 자동차가 운행 중임을 가정한 실험이 진행되었다. 보다 실제와 같은 상황을 연출하기 위하여, glitch, rattling sound가 이용되었다. 사용된 환경 노이즈는 차량이 운행 중에 차량 내부에서 녹음된 것이다. 실험 결과는 도 5에 도시되어 있다.

[0058] 도 6은 본 발명의 일 실시예에 따른 적대적 예제 생성 장치의 기능 블럭도이다.

[0059] 도 6을 참조하면, 적대적 예제 생성 장치(10)는 노이즈 합성부(110)와 적대적 예제 생성부(12)를 포함한다. 실시예에 따라, 적대적 예제 생성 장치(10)는 변조부(130), 적대적 공격부(140), 및 저장부(150) 중 적어도 하나 이상을 포함할 수 있다. 또한, 적대적 예제 생성 장치(10)가 변조부(130)와 적대적 공격부(140)을 포함할 경우, 적대적 예제 생성 장치(10)는 적대적 공격 장치로 명명될 수도 있다.

[0060] 적대적 예제 생성 장치(10)는 적어도 프로세서 및/또는 메모리를 포함하는 컴퓨팅 장치로 구현될 수 있다. 컴퓨팅 장치는 PC(Personal Computer), 랩탑 컴퓨터, 태블릿 PC, 서버 등을 포함할 수 있다.

[0061] 노이즈 합성부(110)는 원본 음성 파일과 노이즈 파일을 합성하여 노이즈가 부가된(합성된) 음성 파일을 생성할 수 있다. 원본 음성 파일은 소정의 콘텐츠 파일로써, 음악 파일 등을 의미할 수 있으나, 본 발명이 이에 제한되는 것은 아니다. 원본 음성 파일에 포함된 콘텐츠의 내용이 중요한 것은 아니다. 또한, 노이즈 파일은 미리 녹음되어 저장부(150)에 저장되어 있을 수 있다.

[0062] 이를 위해, 노이즈 합성부(110)는 저장부(150)에 저장되어 있는 복수의 원본 음성 파일들 중 어느 하나와 저장부(150)에 저장되어 있는 복수의 노이즈 파일들 중 어느 하나를 선택할 수 있다. 또한, 노이즈 합성부(110)는 두 음성 파일(원본 음성 파일, 노이즈 파일)을 결합하기 전에 두 파일의 길이(시간 길이, 즉 재생 시간을 의미할 수 있음)가 동일하도록 두 음성 파일 중 길이가 긴 음성 파일의 길이를 축소시킬 수도 있다.

[0063] 적대적 예제 생성부(120) 소정의 적대적 예제 생성 알고리즘을 이용하여 노이즈가 합성된 음성 파일로부터 적대적 예제를 생성할 수 있다. 실 실시예에 의하면, 적대적 예제 생성부(120)는 노이즈가 합성된 음성 파일에 대해 적대적 공격을 실시하여 적대적 예제를 생성할 수 있다. 적대적 예제는 Nicholas Carlini 등의 기법(Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In 25th {USENIX} Security Symposium ({USENIX} Security 16), pages 513-530, 2016., 또는 Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE Security and Privacy Workshops (SPW), pages 1-7. IEEE, 2018.)이 이용될 수 있다. 그러나, 본 발명의 권리범위가 이에 제한되는 것은 아니며, 적대적 예제는 기존의 널리 알려진 기법들 중 적어도 하나를 이용하여 생성될 수 있다.

[0064] 변조부(130)는 적대적 예제에 소정의 변조(Modulation) 기법을 적용하여 변조된 음성 신호를 생성할 수 있다. 예컨대, 변조부(130)는 FM 라디오 신호를 전송하기 위해 사용되는 FM(Frequency Modulation, 주파수 변조) 기법을 이용하여 조정의 주파수(또는 소정의 주심 주파수)를 갖는 변조된 음성 신호를 생성할 수 있다.

[0065] 또한, 변조부(130)는 변조된 음성 신호의 크기(세기)가 미리 정해진 값보다 큰 값을 갖거나, 라디오 방송 시스템 상에서 전송되는 라디오 신호의 크기(세기) 보다 큰 값을 갖도록 크기를 조정할 수 있다.

[0066] 적대적 공격부(140)는 변조부(130)에 의해 생성된 변조된 음성 신호를 송신할 수 있다. 예컨대, 변조된 음성 신호의 타겟은 차량의 네비게이션 시스템, 차량의 운전자나 탑승자가 소지하고 있는 휴대용 단말기(스마트폰, 태블릿 PC 등)가 될 수 있다.

[0067] 저장부(150)에는 적대적 예제 생성 장치(10) 또는 적대적 공격 장치가 적대적 예제를 생성하는 과정 또는 적대적 공격을 수행하기 위해 필요한 프로그램, 적대적 예제를 생성하는 과정 및/또는 적대적 공격 과정 중에 일시

적으로 또는 비일시적으로 생성되는 데이터가 저장될 수 있다. 또한, 저장부(140)에는 원본 음성 파일, 노이즈 파일, 노이즈 합성부(110)에 의해 생성된 노이즈가 부가된 음성 파일, 적대적 예제 생성부(120)에 의해 생성된 적대적 예제, 변조부(130)에 의해 변조된 음성 파일 등이 저장될 수 있다.

[0068] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0069] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0070] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0071] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0072] 10 : 적대적 예제 생성 장치

110 : 노이즈 합성부

120 : 적대적 예제 생성부

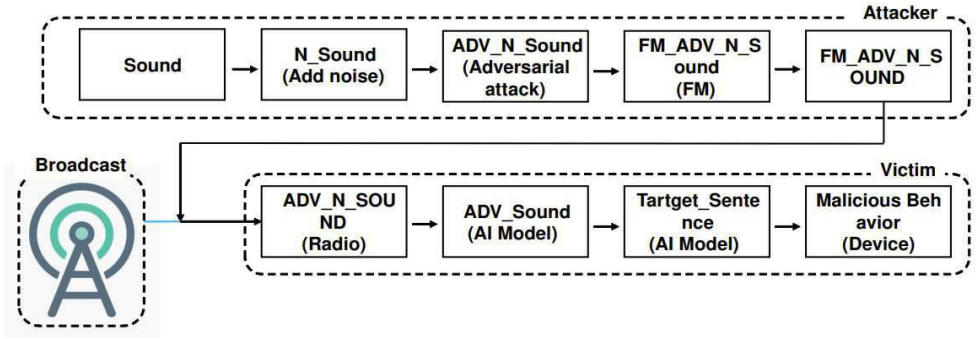
130 : 변조부

140 : 적대적 공격부

150 : 저장부

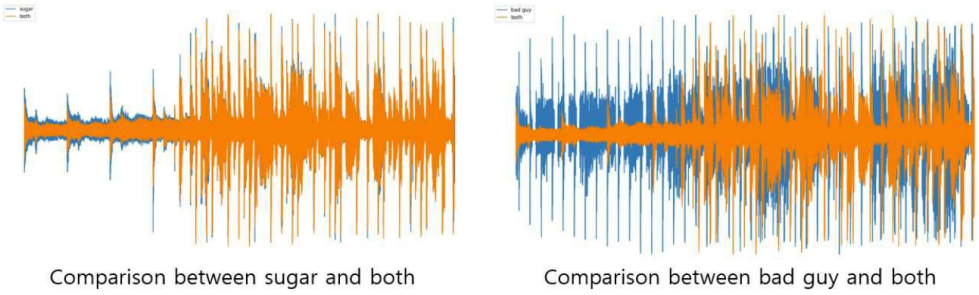
도면

도면1

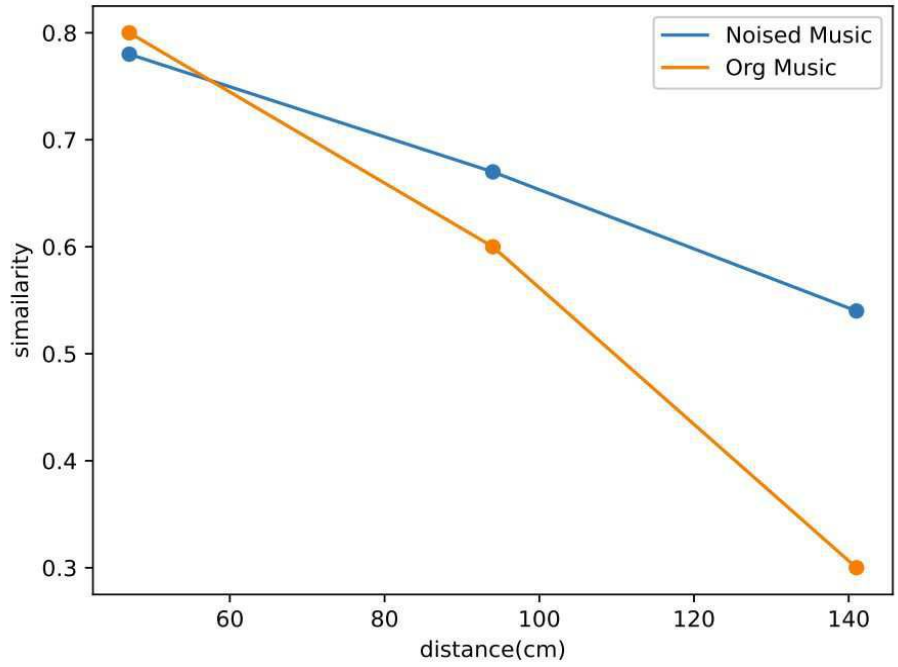


도면2

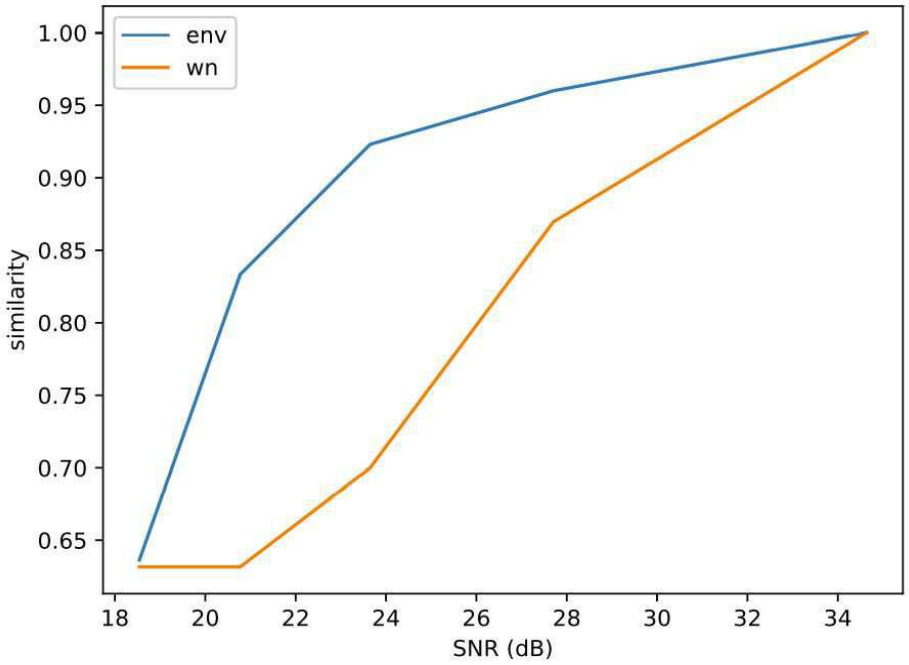
도면3



도면4


Noised Music
Org Music
simailarity
distance(cm)
0.8
0.7
0.6
0.5
0.4
0.3
60
80
100
120
140

도면5


env
wn
similarity
SNR (dB)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
18
20
22
24
26
28
30
32
34

도면6