적대적 공격을 활용한 클라우드 기반 음성 인식 서비스의 개인 정보 보호 강화

강하람†, 윤상운‡, 강경태*

†한양대학교 인공지능융합학과, ‡한양대학교 컴퓨터공학과, *한양대학교 인공지능학과

{†hrkang, ‡swyun, *ktkang}@hanyang.ac.kr

Enhancing Cloud Speech Recongnition Service Using Adverserial Attacks

Haram Kang†, Sangwoon Yun‡, Kyungtae Kang*

† Dept. of Applied Artificial Intelligence, Hanyang University

‡ Dept. of Computer Science and Engineering, Hanyang University

* Dept. of Aritificial Intelligence, Hanyang University

요 약

클라우드 기반의 음성 인식 서비스에서는 화자 식별이 필요하지 않는 기능을 사용해도 음성 데이터를 수집하는 과정에서 화자의 목소리 정보로 인하여 신상 정보 또는 개인 정보 노출로 인한 프라이버시 문제를 야기할 수 있다. 이러한 문제를 해결하기 위해, 본 논문에서는 적대적 공격 기법을 활용한 새로운 개인 정보 보호 강화 방법을 제안한다. 이 방법은 입력 음성 인식 신호에 가우시안 노이즈를 적대적 예제를 추가함으써 화자 인식 정확도는 낮추고, 음성의 전사 정확도는 유지한다. wav2vec 2.0 모델을 활용한 음성 인식과 xvector를 이용한 화자 인식 실험을 통해, 개인 정보 보호 및 음성 전사 정확도 사이의 trade-off를 평가한다. 실험 결과, 20%의 비율을 가진 노이즈를 활용하였을 때, 음성 인식은 0.28의 WER을 유지되면서 화자 인식의 정확도는 4.5%로 급격하게 감소하는 것을 확인하였다.

Ⅰ. 서 론

클라우드 기반의 음성 인식 서비스는 사용자의 편의성과 접근성을 크게 향상시키고, 일상생활과 산업 전반에 걸쳐 다양한 형태로 적용되고 있다. [1] 하지만, 이런 서비스는 원격 서버로의 음성 정보 전송 및 처리 과정에서, 사용자의 개인 정보가 노출될 위험이 상존한다. 음성 데이터는 사용자의 신원 및 민감한 정보를 포함하고 있으며, 이는 데이터 유출 시 사용자의 프라이버시 침해로 직결될 수 있다.

본 논문에서는 별도의 화자 식별이 필요치 않은 기술에서 적대적 공격을 활용하여 음성 데이터 내의 화자 음성 정보를 난독화하여 사용자의 개인 정보 보호를 강화할 수 있는 새로운 접근법을 제안한다. 가우시안 노이즈를 활용한 적대적 변형을 통해 원본 데이터의 음성 전사 정확도는 기존의 수준을 유지하면서, 화자 정보 추출을 어렵게 하여 개인 정보 보호 수준은 향상시키는 것을 목표로 한다.

Ⅱ. 배경 지식

i. 적대적 공격(Adversarial Aattack)

적대적 공격은 머신러닝 알고리즘이 내재하고 있는 취약점을 활용하여 머신러닝 엔진이 스스로 잘못된 판단을 하도록 유도하는 방식의 공격 [2]을 의미한다. 본 연구에서는 가우시안 노이즈를 사용한 적대적 공격을 통해 화자 인식 모델이 입력 음성 데이터의 화자를 오분류하도록 설계하였다.

ii. wav2vec 2.0

wav2vec 2.0 [3]은 2020년 페이스북에이 개발한 최신 딥러닝 모델로, 음성 신호의 파형 데이터를 직접 학습하여 음성 인식 성능을 향상시킨다. CNN 기반의 특징 추출기와 변환기 기반의 컨텍스트 네트워크로 구성되어 있어, 음성 데이터의 의미적 특징을 추출할 수 있다. 해당 모델은 53,000시간 분량의 라벨이 없는 데이터를 사전 학습하고, 이후 특정 음성 인식 작업에 맞게 미세 조정을 할 수 있다. 본 연구에서는 음성 인식 모델을 만들기 위해 wav2vec 2.0 모델을 사용하고, 미세 조정을 통해 성능을 개선하였다.

iii. X-vector

X-vector [4]는 음성의 고유한 특성을 임베딩 벡터로 변환하여 화자 식별에 사용되는 모델이다. DNN을 활용하여 화자의 음성으로부터 중요한 정보를 벡터 형태로 추출하고, 이러한 특징 벡터를 이용해 화자를 식별한다. 본 실험에서는 라벨링된 음성 데이터와 X-vector를 사용하여 화자 인식 모델을 생성하였다.

Ⅲ. 실 험

본 실험의 목적은 적대적 공격을 활용하여, 입력 음성 신호에 가우시안 노이즈를 추가하여 화자 정보를 의도적으로 난독화하는 것과 동시에, 음성 전사 정확도를 유지하는 방안을 탐색하는 것이다. 이를 위해, 원본 음성 데이터와 가우시안 노이즈가 추가된 음성 데이터를 비교 분석하였다. 이런 과정에서 노이즈가 화자 인식 정확도에 어떤 영향력을 미치는지, 그리고 음성 전사 정확도에는 어떠한 변화가 있는지 측정한다.

데이터는 공개 오디오 데이터셋인 LibriSpeech train-clean-360 [5]을 사용하였다. 해당 데이터는 영어 오디오북을 녹음한 360시간 분량의 음성 데이터셋으로, 화자 및 전사 정보를 포함하고 있다. 총 104,010개의 음성 파일로 구성되어 있으며, 이 중, 92,664개의 훈련 데이터로 설정하였으며, 그 외 11,350개의 파일은 평가 데이터로 설정하여 훈련 및 평가 데이터를 9:1로 분할하였다. 이후 기존 평가 데이터에 가우시안 노이즈를 10%의 증가시켜 총 11개 수준의 노이즈 평가 데이터를 준비하였다.

---

† ‡ 바이오인공지능융합전공

* 교신 저자

표 1. 노이즈 비율에 따른 화자 인식 및 음성 인식 평가 결과

노이즈 비율(%) / 평가 모델 (성능 지표)	0	10	20	30	40	50	60	70	80	90	100
음성 인식 (WER)	0.16	0.20	0.28	0.41	0.56	0.70	0.80	0.88	0.93	0.96	0.98
화자 인식 (%)	91.7	10.6	4.50	2.50	1.60	1.20	1.10	0.68	0.51	0.50	0.45

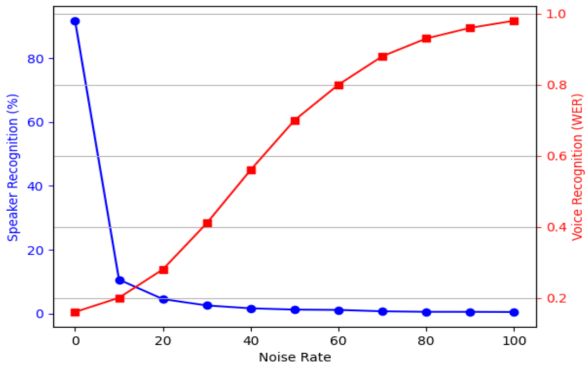

그림 1. 노이즈 비율에 따른 화자 인식 정확도와 음성 인식 WER

우선, 화자 인식 실험에서는 데이터 전처리 과정을 통해 각 음성 데이터에 대응하는 화자 ID를 매핑하였다. 데이터 전처리 후, X-vector 추출기를 이용해 총 921명의 화자의 음성 데이터를 학습시켰다. 학습 과정에서는 Adam 최적화 알고리즘을 사용하였으며, 손실 함수는 다중 클래스 분류에 적합한 CrossEntropyLoss을 적용하였다.

음성 인식 실험은 wav2vec 2.0 모델을 사용하였으며, 미세 조정을 하기 위해 Transformers [6] 라이브러리를 활용하였다. 음성 데이터와 해당 전사 데이터를 맵핑한 뒤, Dataset 클래스를 활용하여 훈련 및 평가 단계에서 데이터를 효율적으로 활용할 수 있도록 하였다. 또한, wav2vec2-base 모델을 기반으로, wav2vec2ForCTC 클래스를 활용하여 미세 조정을 하였다. 최적화 함수와 손실함수는 각각 AdamW와 CTCLoss가 사용되었다.

IV. 실험 결과

본 연구에서는 노이즈 비율을 10%씩 증가시킨 노이즈 데이터셋을 활용하여 미세 조정된 모델의 성능을 평가하였다. 화자 인식 모델은 훈련된 모델에 의해 예측된 레이블과 실제 정답 레이블 일치하는 레이블의 개수를 통해 정확도를 측정하여 성능을 평가하으며 음성 인식 모델은 전사된 텍스트와 원본 텍스트 간의 차이를 수치화한 WER을 사용하여 성능을 평가하였다.표 1과 그림 1은 각각 노이즈 비율에 따른 화자 인식의 정확도와 음성 인식의 WER을 표와 그래프로 나타낸 것이다.

실험 결과, 음성 인식 모델에 20% 비율의 노이즈를 추가했을 때, 노이즈가 없는 상태의 WER인 0.16 대비 0.28의 안정적인 WER을 유지하는 것을 확인할 수 있었다. 반면, 화자 인식 모델은 소량의 노이즈에 의해 정확도가 10.6%로 급격히 감소하는 것으로 나타났다.

V. 고 찰

현재 연구에서는 가우시안 노이즈를 활용하여 적대적 예제를 생성하였으나, 사용된 노이즈의 케이스와 종류가 한정적이었다. 이에 따라, 향후 연구에서는 가우시안 노이즈의 분포와 평균을 조절하여 더욱 다양한 경우의 적대적 예제를 생성할 계획이다. 또한, 가우시안 노이즈 외에도 다른 유형의 노이즈를 적용하여 연구의 범위를 확장할 예정이다.

본 연구는 모델의 평가 과정에서 입력 음성 신호에 적대적 공격의 효과를 검증하였다. 추후 연구에서는 훈련 과정에서 적대적 훈련을 적용하여 효과적인 개인 정보 보호 방안을 개발하는 데 초점을 맞출 것이다.

또한, 이번 실험에서는 화자 인식의 정확성을 개인 정보 보호 수준을 가늠하는 척도로 삼았다. 앞으로의 연구에서는 보다 광범위한 개인 정보 보호 지표를 마련하여 적대적 공격의 보편적 활용 가능성을 제안함으로써 클라우드 기반 음성 인식 서비스에서의 개인 정보 보호의 새로운 가능성을 탐구해 나갈 것이다.

VI. 결 론

본 논문에서는 클라우드 음성 인식 서비스의 개인 정보 보호를 강화하기 위해 적대적 공격 기법을 활용하는 새로운 접근법을 소개하였다. 이 접근법에서 가우시안 노이즈를 음성 데이터에 추가함으로써, 화자 인식에 부정적인 영향을 주면서 음성 전사 정확도는 유지하는 방안을 모색했다. 실험을 통해 화자 인식 정확도는 노이즈의 비율에 관계없이 노이즈가 추가된 것만으로도 91.7%의 정확도에서 10.6%로 크게 감소하는 것을 확인하였다. 반면에, 음성 인식 모델은 20% 비율의 노이즈까지는 노이즈가 없는 데이터의 WER과 유사한 WER을 유지하는 것을 확인하였다. 이러한 결과는, 특정 비율의 노이즈를 활용한 적대적 공격이 음성 인식 기능을 저해하지 않으면서 화자 인식의 정확도를 낮추어, 개인정보 보호에 효과적으로 활용될 수 있음을 시사한다. 이는 음성 인식 기술이 개인 정보 보호와 성능의 균형을 위한 새로운 접근법을 제공할 수 있다.

ACKNOWLEDGMENT

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155885, 인공지능융합혁신인재양성(한양대학교 ERICA))

참 고 문 헌

[1] KISA, "Analysis of personal information processing and user protection scheme in speech recognition based service," 2019. [Online]. Available: https://www.kisa.or.kr/201/form?postSeq=12013.

[2] 최창, 홍인표, “인공지능 보안과 적대적 공격,” *정보과학회지*, 제 41권 제 1호, pp. 30-37, 1월 2023.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems.*, vol. 33, pp. 12449-12460, 2020.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329-5333, 2018.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," OpenSLR, Apr. 2015. [Online]. Available: https://www.openslr.org/12.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi *et al.*, "HuggingFace's Transformers: State-of-the-Art Natural Language Processing," Jul 2020, arXiv:1910.03771v5