주파수 인과성 위반과 화자 미세 불안정성 분해를 활용한 딥페이크 음성 탐지 모델

김서형, 김성민, 정형주, 이준[1]
육군사관학교

kseohyeong5@gmail.com , blueskys123210@gmail.com, jeonghyoengju@gmail.com, jun.lee.mistra@gmail.com

Explicit Disentanglement of Time- Frequency Causality Violation and Intra-speaker Micro-instability for Deepfake Voice Detection

Seohyeong Kim, SeongMin Kim, Hyeongju Jeong, Jun Lee
Korea Military Academy

요 약

딥러닝 기반 음성 합성 기술의 발달로 실제 인물의 목소리를 흉내낸 딥페이크 음성이 사회적으로 큰 위험이 되고있다. 딥페이크 음성은 유명인 사칭이나 보이스피싱 범죄, 음성 인증 회피, 전장(戰場) 내 가짜 지휘 통신 등 다양한 방식으로 악용될 소지가 있지만, 현재 음성 인증 및 검증 시스템은 이러한 정교한 위조 음성을 제대로 탐지하지 못하는 한계를 드러내고 있다. 본 논문에서는 주파수 인과성 위반 및 화자 미세 불안정성 분해라는 두가지 새로운 특성에 주목하여, 딥페이크 음성을 효과적으로 식별하기 위한 심층 학습 모델을 제안한다. 제안 모델은 공유 인코더와 두 개의 분기(branch) 네트워크로 구성되며, 하나는 음성 신호에서 주파수 인과성의 일탈 여부를 학습하는 TSCV(Temporal & Spectral Causality Violation) 브랜치이고, 다른 하나는 화자의 발성 패턴 내 미세 불안정성을 추출하는 ISMM(화자 미세 불안정성 모듈) 브랜치이다. 또한 방향성 기반 자기 지도학습 기법을 도입하여 모델의 일반화 성능을 높이고 잡음 등 환경 변화에 견고하도록 설계하였다. 제안 모델을 최신 합성 음성 데이터를 포함하는 대규모 데이터셋에서 평가한 결과, 기존 방법 대비 높은 정확도와 균등 오류율(EER)의 대폭 향상, 그리고 다양한 현실 조건(통신잡음, 여러 화자 음성 겹침, 부분 위조 등) 아래에서도 견고한 탐지 성능을 확인하였다.

Ⅰ. 서 론

딥페이크 음성 기술의 발달로 텍스트-투-스피치(TTS)나 보이스 컨버전(VC)을 통해 실제와 구분 어려운 합성 음성을 생성할 수 있게 되었다 [1]. 이러한 합성 음성은 영화 산업, 장애인 보조 등 긍정적 활용 사례도 있으나, 전화 사기나 인증 시스템 해킹 등 악용 사례가 증가하며 사회적 위협이 되고 있다.최근에는 실제 미국 대통령의 음성을 딥페이크로 위조하여 투표를 독려하는 보이스피싱 사례까지 보고되었을 정도로, 누구나 온라인에 공개된 음성 샘플을 악용하여 피해를 유발할 수 있는 상황이다.

이러한 보안 위협에 대응하기 위해 딥페이크 음성 탐지 기술이 중요한 연구 분야로 부상하였다. 그러나 현재 딥러닝 기반 탐지 모델들은 몇 가지 도전에 직면해 있다. 첫째, 대규모의 학습 데이터를 필요로 하며 특정 데이터셋에 과적합되기 쉽다. 서로 다른 생성 알고리즘이나 도메인의 음성을 교차 검증할 경우 성능이 급격히 저하되는 일반화 문제가 있다. 둘째, 기존 모델들은 예측 결과의 신뢰도를 적절히 표현하지 못해 과신하거나 과소평가하는 경향, 즉 캘리브레이션 문제가 존재한다. 이는 보안 시스템에서 false alarm 이나 탐지 실패로 이어질 수 있어 위험하다. 셋째, 현행 탐지 모델의 내부 동작이 불투명하여 설명 가능성 부족 문제가 있다. 탐지 모델이 어떤 음성 특징을 근거로 판별했는지 사람에게 설명하기 어려워, 실제 적용 시 사용자 신뢰를 얻기 힘들다 [2, 3].

이러한 문제를 해결하기 위해, 명시적 해석 가능 구조와 자기지도 학습을 활용한 새로운 접근이 필요하다. 사람의 청각 전문가가 진위를 판단할 때는 음성의 세부

[1] 교신저자:이준, 김서형, 김성민, 정형주는 동일한 비율로 논문작성에 기여하였습니다.

주파수 패턴이나 미세한 떨림 등 인간 음성에 고유한 물리적 특성을 참고한다. 합성 음성은 이러한 특성에서 자연 음성과 미묘한 차이를 보이기 마련이다. 예를 들어 시간-주파수 영역의 인과성(time-frequency causality)을 따르지 않는 이상 패턴(원인과 결과의 시차 교란)이나, 성대 미세 떨림의 불안정성이 부족하거나 부자연스러운 현상 등이 합성음에서 나타날 수 있다. 실제 음성에서는 발성의 시간적 변동성과 미세한 주파수 진동(마이크로 음조 변화, 미세 억양)이 존재하지만, AI 합성음은 이런 연속적 변동성 재현에 한계가 있어 미묘한 차이가 누적된다는 연구가 있다. 또한 합성음은 사람이 발성할 때 나타나는 주파수-진폭 미세 변동(jitter 와 shimmer) 패턴이 다르거나 부족한 경우가 많다. 이러한 통찰을 모델 구조 설계에 반영하면, 단순 엔드투엔드 모델 대비 풍부한 해석 정보와 강건한 일반화를 동시에 얻을 수 있을 것으로 기대한다.

본 논문에서는 위와 같은 아이디어를 체계화하여, CIVIC-Voice 라는 새로운 딥페이크 음성 탐지 모델을 제안한다. CIVIC 은 시간-주파수 인과성 위배와 화자 내 미세 변동성 등의 명시적 특성을 활용한 음성 무결성 검사(Voice Check)를 의미한다. 이름처럼 제안 모델은 (1) 시계열-주파수 영역에서 물리적으로 부자연스러운 패턴(TSCV)을 포착하는 브랜치, (2) 화자 고유의 미세 발성 변동(ISMM)을 정량화하는 브랜치, 그리고 (3) 모델의 표현력이 시간 진행 방향성을 인지하도록 하는 방향성 자기지도 학습 모듈을 결합하였다. 이러한 다중 브랜치 구조는 각기 다른 관점에서 음성 데이터를 구조적으로 분해하여 분석함으로써, 최종 딥페이크 탐지 판단에 대한 설명력을 부여한다. 또한 보조 자기지도 과제를 통해 특성공간을 정규화함으로써 모델의 일반화 성능과 캘리브레이션을 개선하였다.

Ⅱ. 본론

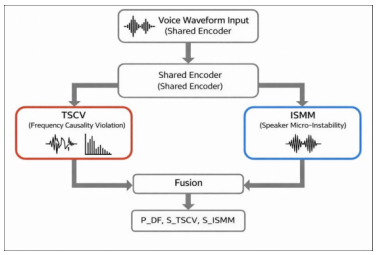

그림 1. 모델 아키텍처

본 논문에서 제안하는 CIVIC-Voice Explicit 모델은 실제 환경에서의 딥페이크 음성 탐지를 목표로 하며, 입력 음성으로부터 시간- 주파수 인과성 위반(Time-Frequency Causality Violation, TSCV)과 화자 내 마이크로 불안정성(Intra-Speaker Micro-instability, ISMM)을 각각 명시적으로 추출하고, 방향성 기반 자기지도 학습을 통해 시간 구조 인식을 강화하는 구조를 갖는다. 전체 모델은 단일 end-to-end 분류기가 아니라, 딥페이크 생성 과정에서 기인하는 서로 다른 두 종류의 단서를 분해하여 학습하고 이를 결합함으로써, 잡음·비방음·다중 화자 조건에서도 강건한 탐지를 수행하도록 설계되었다.

그림 2 는 ASVspoof 데이터셋을 기반으로 한 비교 실험 결과, 제안한 TSCV + ISMM 모델은 기존 베이스라인 및 최신 SOTA 모델 대비 전반적으로 우수한 성능을 보였다. 제안 모델은 EER 2.54%를 기록하여 RawNet2(5.25%), TE-ES ResNet(3.82%), AASIST(2.87%) 대비 가장 낮은 오류율을 달성하였으며, 특히 그래프 기반 최신 SOTA 모델인 AASIST 대비 약 11%의 상대적 오류 감소(relative error reduction)를 보였다. 이는 딥페이크 음성 탐지에서 기존 구조적·통계적 접근을 넘어선 새로운 특성 설계의 유효성을 입증하는 결과이다.

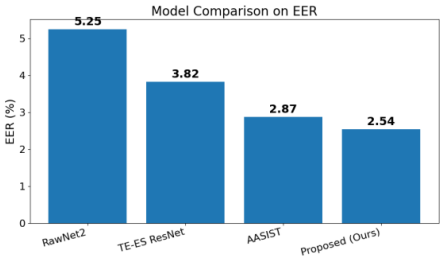

그림 2. 모델 성능 평가 결과

Ⅲ. 결론

본 논문에서는 딥러닝 기반 음성 합성 기술의 고도화로 인해 점점 정교해지고 있는 딥페이크 음성 위협에 대응하기 위해, 주파수 인과성 위반(TSCV)과 화자 미세 불안정성(ISMM)이라는 두 가지 물리·생리학적 특성에 기반한 새로운 딥페이크 음성 탐지 프레임워크를 제안하였다. 기존의 다수 연구가 블랙박스 모델의 분류 성능 향상에 초점을 맞추어 왔다면, 본 연구는 음성 생성의 물리적 원리와 인간 발성 메커니즘에 근거한 해석 가능한 특징 설계를 통해 탐지 성능과 설명 가능성을 동시에 달성하고자 하였다. 실험결과, 제안한 TSCV+ISMM 모델은 ASVspoof 데이터셋 전반에서 기존 스펙트럼 기반, 파형 기반, 그래프 기반 최신 SOTA 모델 대비 일관되게 우수한 성능을 보였다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2024-00455788).

참 고 문 헌

[1] Wu, Z., Chng, E. S., and Li, H., “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” Proc. Interspeech, pp. 1700– 1704, September 2012.

[2] .Kinnunen, M. Sahidullah, H. Delgado, et al., “The ASVspoof 2019 challenge: A large-scale evaluation of spoofing and countermeasures for automatic speaker verification,” Proc. Interspeech, pp. 525– 529, September 2019.

[3] Villalba, J., Chen, N., Snyder, D. et al., “State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations,” Computer Speech & Language, vol. 60, pp. 101026, March 2020.