오디오-시각 모달리티의 비대칭성을 고려한 방향성 기반 딥페이크 탐지

이준희[01], 김정욱[1†]
경희대학교 인공지능학과[1], 경희대학교 컴퓨터공학과[1†]
jhlee39@khu.ac.kr[1],ju.kim@khu.ac.kr[1†]

요 약

최근 오디오와 비주얼 정보를 함께 활용한 딥페이크 탐지 연구는 두 모달리티 간의 동기화 관계를 정량화하여 위조 여부를 판별하는 데 초점을 맞추어왔다. 이러한 접근은 주로 오디오와 비주얼을 대칭적인 관계로 가정하고, 시간 축상에서의 일대일 대응성을 강화하는 방식으로 학습이 이루어진다. 그러나 실제로 두 모달리티는 본질적으로 비대칭적인 특성을 가진다. 하나의 오디오는 특정한 입모양에만 대응하지만, 동일한 입모양이 여러 오디오 표현을 가질 수 있다. 기존 방법들은 이러한 오디오-시각 모달리티의 비대칭성을 고려하지 않아 학습 안정성과 일반화 성능이 저하되는 한계를 보인다. 따라서 본 연구는 이러한 문제를 해결하기 위해 오디오를 기준으로 시각 정보를 판별하는 방향성 기반 단방향 매핑 학습을 제안한다. 제안된 방법은 오디오 특성을 중심으로 해당 시각 정보가 실제인지 위조인지를 판별하도록 학습하며, 반대로 비주얼 정보를 통해 오디오를 추론하는 비효율적인 상호 의존 구조는 제거하였다. 또한 오디오 내부의 위조 단서를 정교하게 포착하기 위해 생성 분포에 기반한 재구성 오차 기반 잔차 학습을 도입하였다. 이를 통해 합성 오디오와 실제 오디오 간의 미세한 통계적 차이를 효과적으로 학습하도록 하였다. 추가로, 오디오 단서만으로도 위조 탐지가 가능하도록 학습하기 위해 화자 변환(voice conversion) 모델에 동일 화자를 reference로 입력하여 자기 화자 변환(self-conversion)을 수행한 동기화 유지 고난도 데이터를 구축하였다. 이를 통해 기존 모델들이 양방향 의존 관계에 빠지던 문제를 완화하고, 오디오-시각 모달리티의 비대칭성을 반영한 보다 강건한 탐지 모델을 구현하였다.

1. 서 론

최근 생성 모델의 급속한 발전으로 인해 오디오와 비주얼 정보를 동시에 조작한 딥페이크 콘텐츠가 현실적인 수준으로 정교해지고 있다. 특히 영상 합성과 음성 변환 기술의 결합은 인간의 인지로는 진위를 구분하기 어려운 수준의 위조물을 만들어내고 있으며, 이는 개인의 명예 훼손, 허위 정보 확산, 사회적 신뢰 붕괴 등 심각한 문제로 이어지고 있다. 이러한 위험에 대응하기 위해 오디오-비주얼 기반 딥페이크 탐지 연구가 활발히 이루어지고 있으며, 최근에는 두 모달리티 간의 시간적 일관성, 즉 동기화 관계를 활용하는 방식이 주된 연구 흐름으로 자리잡고 있다. 이러한 접근은 발화 시점, 입 모양, 음성 강도 등 시간 축상의 일치를 검증함으로써 단일 모달리티 기반 탐지보다 높은 정확도를 달성해왔다.

그러나 기존의 오디오-비주얼 딥페이크 탐지 모델들은 두 모달리티가 서로 완전히 대응된다고 가정하는 대칭적 구조에 기반한다는 한계를 가진다. 대부분의 방법은 오디오와 비주얼 피처를 동일한 수준에서 연결하고, 시간적으로 일대일 대응되도록 학습한다. 하지만 실제 발화 상황에서 두 모달리티는 본질적으로 비대칭적인 특성을 가진다. 하나의 오디오는 특정한 입모양에 대응하지만, 동일한 입모양이 여러 발화 내용과 음향적 표현을 가질 수 있다. 즉, 오디오에서 시각 정보를 판별하는 것은 합리적인 방향이지만, 시각 피처만으로 오디오를 예측하려는 시도는 본질적으로 불안정하다. 이러한 비대칭성을 고려하지 않은 대칭적 학습은 모호한 상관관계를 내재화하여 학습 과정의 안정성을 떨어뜨리고, 특히 데이터셋이 달라질 경우 성능이 급격히 저하되는 일반화 문제를 야기한다.

본 연구는 이러한 한계를 해결하기 위해 오디오-시각 모달리티 간의 비대칭성을 반영한 방향성 기반 단방향 매핑 학습 구조를 제안한다. 제안된 방법은 오디오를 기준으로 시각 정보를 판별하는 방향성을 갖추며, 오디오 피처를 통해 해당 시각 피처가 실제(real)인지 위조(fake)인지를 구분하도록 학습한다. 반대로, 시각 피처를 입력으로 하여 오디오를 복원하거나 예측하는 과정은 제거하였다. 이는 오디오가 발화의 시간적 패턴과 화자 특성에 대한 직접적인 정보를 제공하는 반면, 시각 피처는 상대적으로 모호하거나 다의적인 정보를 포함하기 때문이다. 이러한 단방향 학습 구조는 학습 과정의 혼선을 줄이고, 오디오 중심의 위조 단서 학습을 강화하여 더 안정적이고 해석 가능한 탐지 모델을 구현한다.

또한 본 연구는 오디오 모달리티 내부의 위조 흔적을 정량적으로 포착하기 위해 재구성 오차 기반 잔차 학습을 도입하였다. 이는 생성 모델의 확률적 복원 특성을 활용하여 합성 오디오와 실제 오디오 간의 분포 차이를 학습하

는 방식이다. 구체적으로, 디퓨전 모델을 통해 입력 오디오를 복원한 뒤 원본과의 차이를 계산하여 잔차(residual) 형태로 표현하고, 이를 탐지 네트워크의 학습 신호로 활용한다. 이러한 접근은 단순한 이진 분류 기반 학습보다 위조 오디오의 생성 분포와 실제 신호 간의 통계적 불일치를 더욱 명확히 반영할 수 있다. 결과적으로 모델은 발화 내용이나 화자의 개별적 차이에 영향을 받지 않고, 합성 과정에서 발생하는 미세한 왜곡과 잡음을 안정적으로 학습할 수 있다.

더 나아가, 본 연구에서는 오디오 단서만으로도 위조 여부를 정밀하게 판별할 수 있도록 화자 자기변환 기반의 정제 오디오 데이터셋을 새롭게 구축하였다. 기존 FakeAVCeleb 데이터셋의 RealVideo-FakeAudio 구성을 기반으로, 동일 화자의 상이한 발화를 참조하여 음색만을 변환하는 방식으로 합성 오디오를 재구성하였다. 이를 통해 원본 영상의 시간적 구조와 입 모양의 동기화를 유지하면서도, 오디오적으로는 위조된 특성을 가진 고난도 데이터를 확보하였다. 이러한 데이터는 모델이 오디오 단서만으로도 위조를 인식하도록 학습할 수 있게 하며, 기존 접근 방식에서 발생하던 양방향 의존 문제를 완화시킨다.

결과적으로 본 연구는 기존의 오디오-비주얼 딥페이크 탐지에서 간과되어 온 모달리티 비대칭성 문제를 재조명하고, 이를 해결하기 위한 새로운 학습 패러다임을 제시한다. 제안된 방향성 기반 단방향 매핑 구조와 잔차 학습 전략은 학습 안정성과 일반화 성능을 동시에 향상시키며, 오디오 중심의 신호 단서만으로도 정밀한 탐지가 가능함을 실험적으로 검증하였다. 이러한 결과는 오디오-시각 모달리티 간의 관계를 단순한 정합의 문제로 보던 기존 관점을 넘어, 각 모달리티의 구조적 특성을 반영한 방향성 학습이 딥페이크 탐지의 새로운 방향으로 발전할 수 있음을 보여준다.

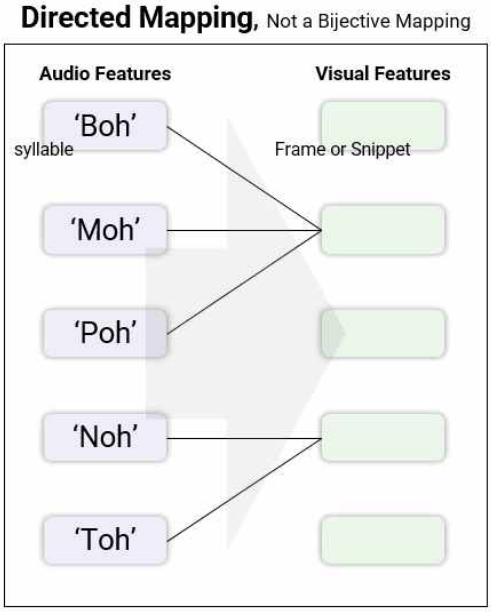


그림 1

2. 관련 연구

2.1 오디오-비주얼 딥페이크 탐지의 동향

오디오-비주얼 딥페이크 탐지 연구는 초기에 비디오 단서나 오디오 단서를 개별적으로 분석하는 단일 모달리티 접근에서 출발하였다. 비디오 기반 방법들은 얼굴 합성이나 프레임 단위의 왜곡, 눈 깜빡임 주기, 미세한 표정 전이 등 시각적 불연속성을 포착하는 데 집중하였다. 반면, 오디오 기반 연구들은 화자 음색, 포먼트 분포, 스펙트럼 왜곡 등의 음향적 이상치를 분석하여 위조된 음성을 식별하였다. 그러나 실제 딥페이크 영상은 두 모달리티가 함께 조작되기 때문에, 단일 모달리티만으로는 위조 여부를 완전하게 판단하기 어렵다는 한계가 존재했다.

이후 연구의 흐름은 두 모달리티를 통합적으로 고려하는 멀티모달 탐지(multimodal detection) 로 확장되었다. 이러한 접근은 오디오와 비주얼 간의 동기화 정도를 정량화하거나, 두 모달리티의 표현 공간을 통합하는 방식으로 발전하였다. 일반적으로 오디오 신호에서 추출한 스펙트로그램 피처와 영상 프레임에서 얻은 얼굴 피처를 시간적으로 정렬하고, 이들 간의 대응 관계를 학습하는 구조가 사용된다. 이를 통해 실제 발화에서는 높은 상관관계를, 위조 영상에서는 낮은 상관관계를 나타내도록 모델을 훈련한다. 이러한 멀티모달 기반 탐지 기법은 동기화 불일치를 주요 단서로 삼아 위조 여부를 판별하며, 기존 단일 모달리티 접근에 비해 현저히 높은 성능을 보여 왔다.

2.2 Consistency 기반 접근의 한계

최근의 오디오-비주얼 딥페이크 탐지 연구들은 대부분 두 모달리티 간의 동기화 일관성(consistency) 을 핵심 단서로 삼는다. 이들은 오디오와 비주얼 신호가 시간적으로 정합되는 정도를 계산하거나, 서로의 특징을 복원·예측하는 방식으로 학습된다. 예를 들어, 일부 연구에서는 오디오와 시각 피처를 동일한 표현 공간으로 사상시켜 거리 기반의 대조 학습을 수행하며, 또 다른 연구에서는 한 모달리티를 입력으로 다른 모달리티를 복원하도록 설계해 상호 의존성을 극대화한다. 이러한 접근은 실제 발화에서 두 모달리티가 밀접하게 연결된다는 점을 활용해 높은 탐지 성능을 얻지만, 구조적으로 양방향 대칭성을 가정한다는 한계를 가진다.

문제는 실제 오디오-시각 관계가 이러한 대칭적 구조를 따르지 않는다는 데 있다. 하나의 오디오는 특정한 입모양에 대응하지만, 동일한 입모양이 여러 발화 내용과 다양한 음성 특성을 가질 수 있다. 즉, 오디오→시각 매핑은 단일 대응 관계를 가지지만, 시각→오디오 매핑은 다대일 관계

를 형성한다. 이러한 비대칭성을 고려하지 않은 채 상호 복원이나 일대일 대응 학습을 수행하면, 모델은 모호한 상관관계를 내재화하고 학습 안정성이 저하된다. 특히 서로 다른 데이터셋 간에서는 동기화 패턴이나 화자 특성이 달라지기 때문에, 대칭적 학습 구조를 가진 모델은 일반화 능력이 크게 떨어지는 경향을 보인다.

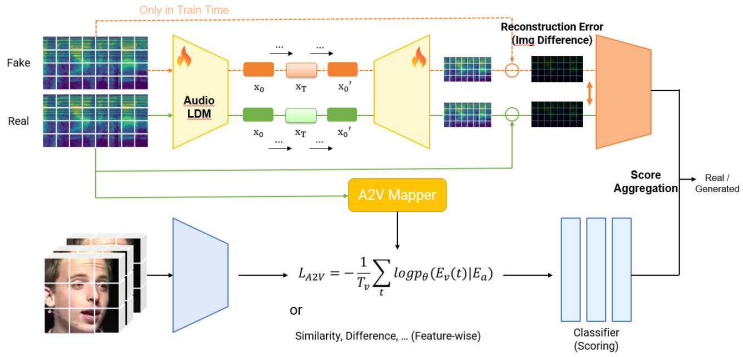
결국 이러한 consistency 기반 접근은 두 모달리티를 동일한 수준에서 대응시키는 데 집중한 나머지, 실제 발화 구조가 가진 비대칭적 특성과 방향성 정보를 충분히 반영하지 못한다. 본 연구는 이러한 구조적 한계를 보완하기 위해, 오디오를 기준으로 시각 정보를 판별하는 단방향 매핑 학습을 새롭게 제안한다.

3. 제안 방법

본 연구는 오디오-시각 모달리티의 비대칭성을 반영하여, 오디오를 중심으로 위조 여부를 판별하는 방향성 기반 탐지 프레임워크를 제안한다. 제안된 모델은 오디오 복원 오차를 활용한 잔차 표현 학습과 오디오-시각 간 단방향 매핑 검증 과정을 결합하여, 오디오 단서만으로도 안정적이고 일반화 가능한 탐지를 수행한다. 전체 프레임워크는 크게 (1) 오디오 잔차(feature residual) 학습 모듈과 (2) 오디오-시각 매핑 검증(A2V Mapping) 모듈로 구성되며, 두 모듈의 결합을 통해 최종 탐지 점수를 산출한다.

학습 과정에서는 Real Video-Real Audio와 이에 대응하는 Real Video-Fake Audio의 쌍(pair data)을 입력으로 사용한다. 두 데이터는 동일한 비디오 프레임을 공유하므로, 시각 피처는 하나의 인코더를 통해 추출한다. 오디오의 경우 각 멜스펙트로그램 프레임을 diffusion 기반 오디오 복원 모델(AudioLDM)에 입력하여 복원된 신호를 얻고, 원본 멜 프레임과의 차이를 계산해 잔차 멜 프레임(residual mel frame)을 구성한다. 이 잔차는 합성 과정에서 발생한 미세한 통계적 왜곡을 반영하며, 이를 오디오 아티팩트 인코더(audio artifact encoder)를 통해 임베딩 공간에서 실제(real) 오디오 피처와 명확히 구분되도록 학습한다.

한편, 오디오-비주얼 매퍼(A2V Mapper)는 오디오 프레임 정보로부터 대응되는 시각 정보를 예측하고, 그 결과를 실제 영상 프레임 피처와 비교하여 일치도를 계산한다. 이 과정은 오디오→비주얼 방향으로만 수행되며, 시각→오디오로의 복원은 포함하지 않는다. 즉, 오디오가 시각 정보를 설명할 수 있는지를 판단하는 단방향 매핑 구조를 구현한 것이다. 마지막으로, 오디오 잔차 학습 모듈과 A2V 매핑 모듈에서 산출된 피처 일치 정보를 통합하여 최종 탐지 점수를 계산하고, 이를 통해 입력 샘플이 실제(real)인지 위조(fake)인지를 판별한다.



3.1 비대칭 학습 구조

기존의 오디오-비주얼 딥페이크 탐지 모델은 두 모달리티가 시간 축상에서 일대일로 대응한다고 가정하고, 오디오와 시각 피처를 상호 복원하거나 동일한 표현 공간으로 정렬하도록 학습된다. 그러나 실제 발화 환경에서는 이러한 대칭 가정이 성립하지 않는다. 하나의 오디오는 특정한 입모양과 타이밍에 대응하지만, 동일한 시각적 발화가 여러 음향적 표현을 가질 수 있기 때문이다. 이처럼 오디오-시각 간 관계는 구조적으로 비대칭적이며, 따라서 양방향 복원 또는 대칭적 정합 학습은 학습 혼선을 유발하고 모델의 일반화 성능을 저하시킨다.

이를 해결하기 위해 본 연구에서는 방향성 기반 단방향 매핑 구조를 설계하였다. 제안된 구조는 오디오를 기준으로 대응하는 시각 정보의 타당성을 검증하는 방향만을 학습하며, 시각 정보를 이용해 오디오를 복원하거나 예측하는 역방향 경로는 제거한다. 이때 오디오 인코더(Audio Encoder)는 입력 오디오의 시간-주파수적 특징을 추출하고, 시각 인코더(Visual Encoder)는 해당 프레임의 얼굴 영역에서 표현된 발화 동작의 특징을 추출한다. 이후 오디오-비주얼 매퍼(A2V Mapper)가 오디오 피처를 기반으로 시각 피처를 예측하고, 그 결과를 실제 시각 피처와 비교하여 매핑 일치도를 계산한다.

이 구조의 핵심은 오디오가 시각 피처를 "설명할 수 있는가"에 초점을 맞추는 것이다. 실제(real) 샘플의 경우 오디오와 시각 정보 간의 시간적, 발화적 일관성이 높기 때문에 A2V 매퍼가 생성한 피처가 실제 시각 피처와 높은 유사도를 갖는다. 반면, 위조(fake) 샘플의 경우 음성 변조나 합성 과정에서 화자의 타이밍, 발화 세기, 포먼트 분포 등이 변형되어 오디오-시각 간의 일관성이 무너진다. 모델은 이러한 불일치 패턴을 학습하여, 오디오 단서만으로도 위조 신호를 안정적으로 구분할 수 있다.

이와 같이 단방향 매핑 구조를 채택함으로써, 모델은 비대칭적 모달리티 관계를 반영하고 불필요한 상호 복원 경로로 인한 학습 불안정성을 제거할 수 있다. 결과적으로 제안된 구조는 오디오 중심의 강건한 피처 학습을 가능하게 하며, 데이터셋 변화나 화자·도메인 차이에 대한 일반화 성능을 크게 향상시킨다.

3.2 오디오 재구성 기반 잔차 학습

딥페이크 오디오는 표면적으로는 자연스럽게 들리더라도, 생성 모델이 만들어내는 확률적 잡음과 미세한 스펙트럼 왜곡을 포함한다. 이러한 위조 신호의 분포적 차이를 명확히 포착하기 위해 본 연구에서는 재구성 오차 기반 잔차 학습(residual learning) 방식을 도입하였다. 이 방법은 원래 이미지 위조 탐지에서 제안된 확률적 복원 원리를 오디오 도메인으로 확장한 것으로, 오디오 신호의 생성 분포와 실제 신호의 차이를 잔차 형태로 표현하여 학습한다.

구체적으로, 입력 오디오의 멜스펙트로그램을 diffusion 기반 오디오 복원 모델인 AudioLDM에 입력하여 복원된 멜 프레임을 얻는다. 이후 원본 멜 프레임과 복원된 멜 프레임 간의 차이를 계산하여 잔차 멜 프레임(residual mel frame) 을 생성한다. 이 잔차는 오디오의 스펙트럼 밀도, 잡음 분포, 위상 정보 등에서 발생하는 합성 특유의 불연속성을 반영하며, 실제(real) 오디오보다 위조(fake) 오디오에서 상대적으로 더 큰 오차를 나타낸다.

이 잔차 표현은 오디오 아티팩트 인코더(audio artifact encoder) 를 통해 임베딩 공간으로 투영된다. 학습 과정에서 모델은 실제 오디오 피처와 잔차 피처 간의 거리를 최대화하도록 학습되며, 이는 위조 과정에서 발생한 통계적 왜곡을 분리해내는 역할을 한다. 결과적으로 모델은 단순한 분류 경계가 아닌, 생성 분포와 실제 분포 간의 구조적 차이를 내재적으로 학습하게 된다.

이러한 잔차 학습은 두 가지 장점을 가진다. 첫째, 오디오 내용(발화 문장, 화자 특성 등)과 무관하게 합성 과정의 흔적만을 추출하므로 데이터의 다양성에 덜 의존한다. 둘째, 오디오의 주파수 분포 및 노이즈 패턴에 대한 민감도를 높여, 단순한 시각-청각 정합 기반 모델보다 위조 탐지의 안정성과 일반화 성능을 크게 향상시킨다.

4. 데이터셋

본 연구에서는 제안된 오디오 중심 탐지 모델의 학습과 평가를 위해 기존 FakeAVCeleb 데이터셋을 기반으로 한 정제된 RealVideo-FakeAudio(Refined RVFA) 데이터를 새롭게 구축하였다. 기존 FakeAVCeleb의 RVFA 구성은 실영상(Real Video)에 다른 화자의 합성 음성을 결합하여 생성된 위조 샘플로, 기본적인 모달리티 불일치를 포함하지만 합성 음성의 품질이 낮아 실제 발화 패턴과는 다소 동떨어져 있었다. 이러한 특성은 모델이 쉽게 동기화 불일치를 학습하게 만들어, 오디오 자체의 위조 단서보다는 단순한 싱크 오류에 의존하는 결과를 초래하였다.

이를 개선하기 위해, 본 연구에서는 화자 자기변환(voice conversion) 기반의 정밀 합성 절차를 도입하였다. 구체적으로, FakeAVCeleb의 RealVideo-RealAudio 쌍을 기준으로 하여 VoxCeleb2 데이터셋에서 동일 화자의 다른 발화를 참조 오디오로 선택하였다. 이후 Seed-VC 모델을 이용해 원본 오디오의 시간적 구조(발화 타이밍, 길이, 강도 등)는 유지한 채 음색만을 참조 화자의 특성으로 변환하였다. 이렇게 생성된 오디오는 발화 시점과 입모양이 원본과 정확히 일치하므로 시각적으로는 자연스럽지만, 음향적으로는 새로운 화자 특성을 가진 합성(fake) 음성이다.

이 정제된 Refined RVFA 데이터셋은 다음과 같은 특성을 갖는다.

1. 동기화 품질이 높음: 원본 비디오의 시간 구조를 그대로 보존하여, 기존 RVFA보다 훨씬 정밀한 싱크를 유지한다.
2. 합성 난이도 증가: 시각 단서만으로는 위조 여부를 구분하기 어려우며, 오디오 신호 내부의 생성 흔적을 분석해야 탐지가 가능하다.
3. 강건성 평가에 적합: 기존 모델이 동기화 단서에 과의존하는지를 검증할 수 있는 실험적 환경을 제공한다.

실험에서는 본 연구의 모델과 기존의 오디오-비주얼 동기화 기반 탐지 모델들을 동일한 학습 프로토콜 하에 비교하였다. 그 결과, Refined RVFA 환경에서는 대부분의 기존 모델이 명확한 성능 저하를 보인 반면, 제안된 오디오 중심 잔차 학습 모델은 높은 탐지 정확도를 유지하였다. 이는 본 연구의 접근이 단순한 모달리티 정합이 아닌, 오디오 생성 분포의 내재적 불일치를 학습했음을 보여준다.

4.1 분석

Seen Domain (VoxCeleb2)

Partially Success

Seen Domain (VoxCeleb2)

Fail

Uneen Domain (YouTube)

Fail