Directionality-aware Audio-Visual Deepfake Detection Considering Cross-modal Asymmetry

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

요 약

기존 오디오-비주얼 딥페이크 탐지는 모달리티 간 동기화 불일치에 주로 의존하며, 오디오와 비주얼을 대칭적으로 정렬하거나 상호 복원해 일대일 대응을 강제해왔다. 그러나 실제 발화에서 두 모달리티 관계는 비대칭적이며, 특히 비주얼만으로 오디오를 예측하는 방향은 본질적으로 불안정하다. 본 연구는 이를 해결하기 위해 오디오를 기준으로 비주얼의 정합성을 검증하는 방향성 기반 단방향 프레임워크를 제안한다. 먼저 오디오 단일 모달리티 representation learning으로 생성 흔적에 강건한 오디오 인코더를 학습하고, 이후 오디오 조건으로 시점별 비주얼 표현을 예측해 실제 비주얼 표현과의 불일치를 탐지 신호로 사용한다. 또한 오디오 단서와 모달리티 정합 단서를 신뢰도 기반으로 융합해 샘플별로 더 강한 단서에 가중치를 부여한다. 실험 결과 오디오만으로도 높은 탐지 성능을 보였고, 멀티모달 결합 시 성능이 추가로 향상되어, 동기화 단서가 약한 고난도 조건에서도 제안한 방향성 모델링이 효과적임을 확인하였다.

1. 서 론

최근 생성 모델의 고도화로 인해 오디오와 비주얼 정보를 동시에 조작한 딥페이크 콘텐츠가 실제와 구분하기 어려운 수준으로 정교해지고 있다. 특히 얼굴 합성과 음성 변환 기술의 결합은 발화 동기화와 화자 유사성을 동시에 만족시키며, 개인 사칭, 허위 정보 확산, 사회적 신뢰 훼손 등 다양한 보안·사회적 위험을 증폭시킨다. 이러한 배경에서 오디오-비주얼 딥페이크 탐지는 단일 모달리티 기반 탐지의 한계를 보완하기 위한 핵심 과제로 부상했으며, 공개 데이터셋과 벤치마크의 확장 또한 빠르게 이루어지고 있다. 예를 들어 FakeAVCeleb은 실세계 대화 영상에서 수집된 화자·인종·성별 다양성을 고려한 오디오-비주얼 딥페이크 데이터셋으로 널리 사용되고 있다[1].

그럼에도 다수의 기존 접근은 오디오와 비주얼을 대칭적(bijective) 관계로 가정한 채, 두 모달리티를 동일한 수준에서 정렬하거나 상호 복원하도록 학습하는 경향이 있다. 그러나 실제 발화에서 두 모달리티는 구조적으로 비대칭적이다. 오디오는 음소·억양·강세·발화 속도 등 발화의 시간적 구조와 음향적 세부 정보를 직접 포함하는 반면, 비주얼은 제한된 입술·턱·표정 움직임을 통해 발화를 간접적으로 반영한다. 이로 인해 오디오로부터 비주얼을 생성하거나 동기화를 맞추는 방향은 비교적 안정적으로 구현될 수 있지만, 비주얼만으로 오디오를 복원하거나 생성하는 방향은 본질적으로 불안정해지기 쉽다. 실제로 오디오 조건으로 입 모양을 맞추는 대표적 접근인 Wav2Lip은 다양한 화자·환경에서도 높은 립싱크 품질을 보이며, 오디오→비주얼 방향의 강한 조건성을 보여준다[2]. 반면 비주얼→오디오 방향은 동일한 입 모양이 여러 가능한 음향 실현을 갖는 다대일 모호성을 내재하며, 이 모호성은 학습 안정성과 일반화 성능을 저해할 수 있고, 오디오→비주얼 정렬 관점에서 강한 귀납 편향을 형성한다[3].

본 연구는 이러한 비대칭성을 명시적으로 반영하여, 오디오를 조건으로 비주얼 정합을 검증하는 방향성 기반 탐지 프레임워크를 제안한다. 제안하는 방법은 두 가지 핵심 문제의 재정의와 그에 대응하는 해결책으로 구성된다. 첫째는 오디오-비주얼 비대칭성을 무시한 대칭적 결합 학습으로 인해 모델이 모호한 상관관계에 의존하고 도메인 변화에 취약해지는 문제이다. 이에 대한 해결 방안으로, 오디오 특징을 조건으로 비디오 내부의 시간별 시각 특징을 예측·정렬하는 Directional Audio-Visual Modeling Framework를 설계한다. 두 번째 문제는 기존 RVFA류 데이터가 오디오-비주얼 동기화 불일치에 지나치게 노출되어, 모델이 오디오 내부의 위조 흔적보다는 단순한 싱크 오류에 의존하도록 유도되는 데이터셋 편향 문제이다. 이에 대한 해결책으로, 고동기화 조건에서 음색 변환만 수행하는 Highly Synchronized Voice Conversion Dataset을 구성하여, 모델이 싱크 단서가 약한 환경에서도 오디오 위조 흔적과 방향성 정합 단서를 학습하도록 한다.

제안 프레임워크는 두 단계 학습 전략을 따른다. 먼저 오디오 단일 모달리티에서 디퓨전 기반 음성 변환/합성 과정이 남기는 미세 흔적에 강건한 표현을 학습하는 오디오 전용 인코더를 구축한다. 이후 메인 학습에서는 MARLIN 기반 비주얼 인코더를 사용해 비디오의 시간별 시각 특징과 비디오 전역 특징을 추출하고, 오디오 특징을 조건으로 전역 시각 특징에서 시간별 시각 특징을 예측하는 A2V mapper를 학습한다. 이때 정합이 정상인 경우 예측 시각 특징과 실제 시각 특징의 차이는 작아지고, 오디오 또는 비디오가 조작된 경우 차이가 커지도록 유도한다. 또한 오디오 전용 판별과 A2V 정합 판별 중 어느 쪽이 더 신뢰

할 만한 단서를 제공하는지 샘플 단위로 동적으로 가중하는 confidence 기반 융합을 통해 최종 판별 점수를 산출한다. 전체 프레임워크 개요는 그림 3에 제시한다. 또한 두 모달리티 관계가 bijective가 아님을 추상화한 예시는 그림 1에, 비주얼→오디오 방향의 불안정성을 실증적으로 보이기 위한 AV-HuBERT 기반 분석은 그림 2에 제시한다.

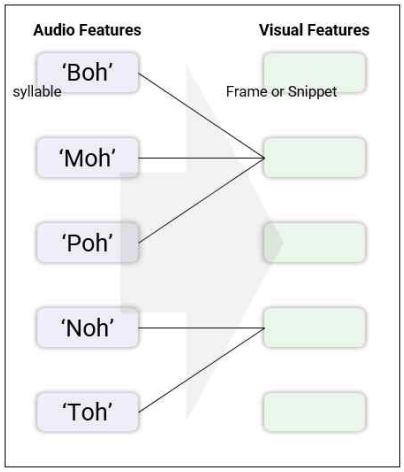

그림 1: Cross-modal Asymmetry

2. 관련 연구

2.1 오디오-비주얼 딥페이크 탐지

오디오-비주얼 딥페이크 탐지는 초기에는 비디오 기반 또는 오디오 기반 단일 모달리티 탐지에서 출발했으나, 실제 딥페이크가 두 모달리티를 동시에 조작하는 사례가 증가함에 따라 멀티모달 탐지로 확장되었다. 비디오 기반 탐지는 얼굴 합성에서 발생하는 프레임 단위 왜곡, 정체성 불일치, 미세 텍스처 이상 등을 포착하는 데 집중하고, 오디오 기반 탐지는 스펙트럼 왜곡, 화자 유사성 과적합, 생성 잡음 패턴 등을 단서로 사용한다. 멀티모달 탐지는 두 모달리티 간의 시간적 정합성과 상호 일관성을 활용하여 단일 모달리티의 취약점을 보완하며, 특히 립싱크 불일치 또는 음성-입모양 상관 약화를 핵심 단서로 삼는 접근이 널리 사용된다.

2.2 Consistency 기반 접근의 한계

동기화 기반 접근은 오디오와 비주얼의 시간 정렬을 직접 학습하거나, 두 모달리티를 공통 표현 공간에서 정렬하여 거리 기반으로 일치도를 측정한다. 이 계열의 방법들은 학습 신호가 명확하고 성능이 높다는 장점이 있으나, 두 가지 구조적 한계를 가진다. 첫째, 많은 방법이 오디오와 비주얼이 일대일로 대응된다는 가정 하에 상호 복원 또는 대칭 정렬을 수행하며, 이 과정에서 비대칭적 관계에서 기인하는 모호성이 학습에 그대로 유입될 수 있다. 둘째, 데이터셋이 제공하는 불일치 패턴이 지나치게 명확한 경우, 모델은 오디오 내부의 위조 흔적이나 비디오 내부의 합성 흔적을 학습하기보다 단순한 싱크 오류에 과의존하게 된다. 이러한 과의존은 동기화가 높은 고난도 위조(예: 화자 음색만 변환되고 발화 타이밍은 유지되는 경우)에서 급격한 성능 저하로 이어질 수 있다.

2.3 오디오-비주얼 비대칭성의 실증적 근거

오디오-비주얼의 비대칭성은 생성 방향을 비교함으로써 직관적으로 확인할 수 있다. 오디오를 조건으로 비디오의 입 모양을 맞추는 방향은 실제로 매우 높은 품질로 구현 가능하며, FakeAVCeleb의 FVRA 유형은 실오디오를 유지한 채 립싱크 기반으로 비디오를 조작함으로써 시각적으로 자연스러운 정합을 보이는 사례에 해당한다. 반대로 비주얼을 조건으로 오디오를 생성하거나 복원하는 방향은 동일한 입 모양이 다수의 가능한 음향 실현을 갖는 모호성 때문에 불안정해지기 쉽고, FakeAVCeleb의 RVFA 유형은 실비디오를 유지하면서 오디오가 교체/합성되는 과정에서 동기화 및 발화 일관성이 무너지는 사례가 빈번하다. 이 비대칭성을 보다 체계적으로 검증하기 위해, 본 연구에서는 오디오-비주얼 자기지도 학습 모델인 AV-HuBERT를 이용해 비주얼 스피치 인식(visual speech recognition) 다운스트림 평가를 수행하였다[5]. 그 결과, 비주얼을 정보를 입력으로 하여 발화 내용을 맞추는 케이스는 학습 도메인 내부에서도 거의 없어, 예측 불안정이 발생하는 사례가 관찰되었다. 해당 정성 결과는 그림 2에 제시하며, 이를 통해 비주얼→오디오 복원은 탐지 프레임워크 내부의 핵심 학습 경로로 채택하기에 부적절하다는 결론을 뒷받침한다.

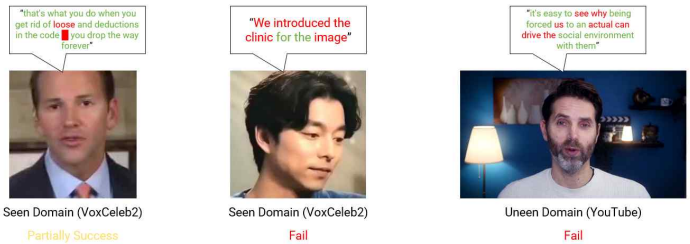

그림 2: Vision-to-Audio의 불안정성 검증

3. 제안 방법

3.1 오디오 전용 표현 학습: 디퓨전 흔적에 강건한 오디오 인코더

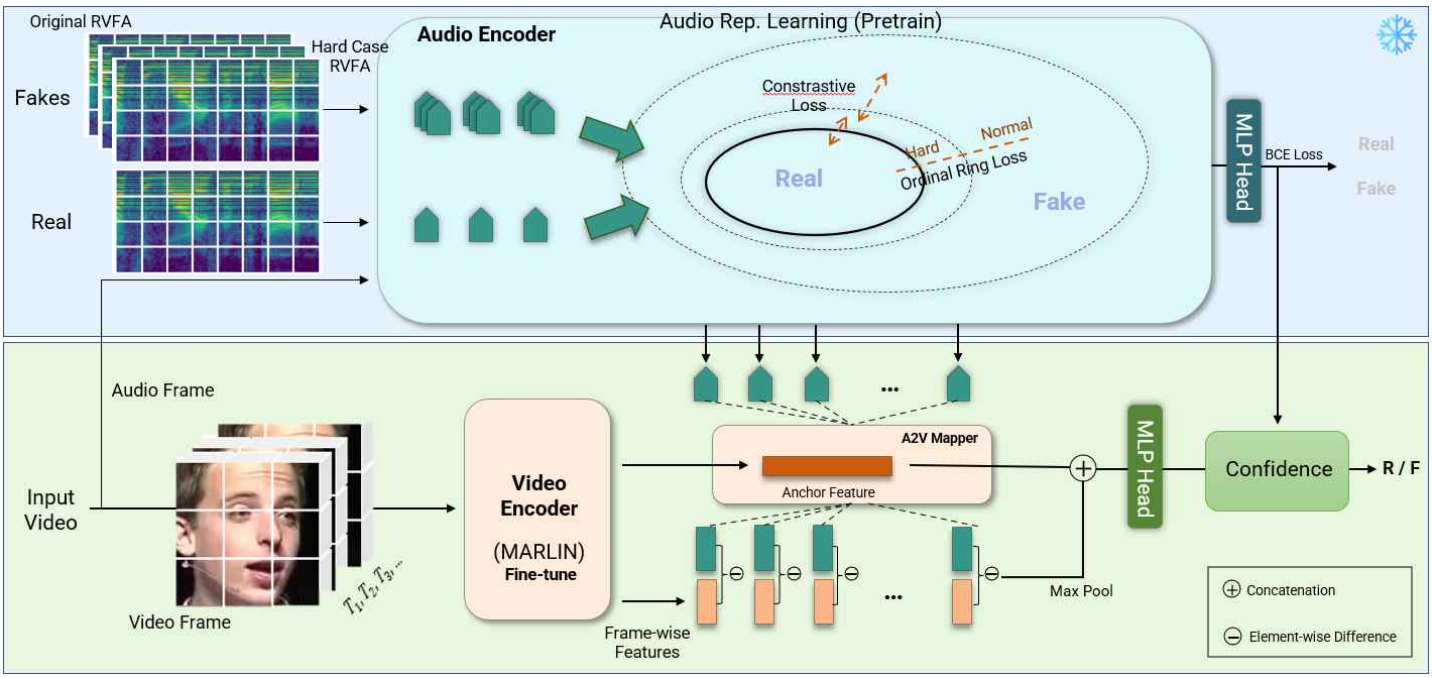

그림 3: 전체 프레임워크

첫 단계에서는 오디오 모달리티 내부에서 위조 흔적을 안정적으로 포착하는 표현을 학습한다. 목표는 발화 내용이나 화자 정체성 자체에 과적합하지 않고, 음성 변환/합성 과정에서 발생하는 미세한 생성 흔적을 분리하는 것이다. 이를 위해 멜 스펙트로그램을 입력으로 받아 시간 축을 패치 단위 토큰으로 분해한 뒤, 트랜스포머 인코더로 전역 오디오 특징과 시간별 토큰을 추출한다. 학습 신호는 (1) 클래스 조건의 대조 학습과 (2) 이진 위조 판별을 결합하여 구성한다. 특히 지도 대조 학습은 같은 클래스(예: real, 그리고 난이도별 fake) 내 샘플을 군집화하고 서로 다른 클래스 간 간격을 확보하는 데 효과적이며, 오디오 내용 다양성에 대한 강건성을 높이는 데 유리하다[8]. 또한 본 연구는 real을 중심으로 난이도(하드 네거티브)에 따라 거리가 점진적으로 증가하도록 유도하는 순서 제약을 추가함으로써, OOD 관점에서 더 해석 가능한 거리 구조를 학습하도록 한다.

3.2 방향성 기반 A2V 정합 모델링

두 번째 단계에서는 오디오를 조건으로 비디오의 시간별 시각 특징이 얼마나 일관되게 설명되는지를 검증한다. 입력 비디오는 MARLIN 기반 비주얼 인코더로 처리하여 (1) 시간/공간 토큰으로 구성된 시퀀스 시각 특징과 (2) 비디오 전역을 대표하는 단일 전역 특징(anchor)을 동시에 추출한다[4]. 입력 오디오는 3.1에서 학습된 오디오 인코더를 통해 전역 오디오 특징과 시간 토큰을 추출하며, 시간 토큰은 비주얼 시퀀스 길이에 맞추어 보간해 정렬한다.

A2V mapper는 오디오 토큰 시퀀스와 전역 시각 특징(anchor)을 함께 입력받아, 각 시간 인덱스에서 기대되는 시각 특징을 예측한다. 핵심은 전역 시각 특징을 조건으로 사용함으로써 정체성·배경·촬영 조건 등 비디오의 전역 정보를 유지한 채, 오디오가 제공하는 발화 단서에 따라 시간별 시각 변화를 복원하도록 유도하는 것이다. 이후 예측된 시각 시퀀스와 비주얼 인코더가 추출한 실제 시각 시퀀스의 차이를 계산하여, 불일치 시퀀스 특징을 얻는다. 정상 샘플의 경우 이 차이는 작아지고, 오디오 또는 비디오가 조작된 경우 차이가 커지도록 학습된다.

여기서 비디오가 조작된 FVRA/FVFA 유형의 처리는 다음과 같이 해석된다. 오디오가 실오디오이더라도 비디오가 조작되면, 비주얼 인코더가 추출한 실제 시각 시퀀스는 조작 흔적과 함께 변형된 분포를 갖게 된다. A2V mapper는 오디오 조건과 전역 시각 특징을 바탕으로 발화에 필요한 시간별 시각 패턴을 예측하지만, 조작된 시각 시퀀스는 그 예측과 구조적으로 어긋나게 된다. 결과적으로 FVRA/FVFA에서는 예측-실측 시각 차이가 커지는 방향이 자연스럽게 탐지 단서로 기능한다. 이는 RVFA처럼 오디오가 조작된 경우에도 동일하게 성립하며, 조작된 오디오는 발화 미세 패턴을 변형시켜 예측 시각 시퀀스를 흔들고, 실측 시각 시퀀스와의 일관성을 약화시킨다.

3.3 전역 시각 특징의 결합과 confidence 기반 융합

A2V 불일치 시퀀스는 시간 축으로 풀링하여 고정 길이 특징으로 요약된다. 본 연구는 이 요약 특징에 더해 비디오

전역 특징(anchor)을 함께 사용한다. 직관적으로, A2V 불일치 특징은 발화 동역학의 정합 오류에 민감한 반면, 전역 특징은 정체성 일관성, 전역 텍스처 및 조작 흔적 등 비디오 자체의 전역 단서를 포함한다. 따라서 두 특징을 결합하면, 동기화가 높게 유지되더라도 비디오 조작이 존재하는 경우(FVRA/FVFA)까지 포괄하는 판별 신호를 강화할 수 있다. 구체적으로, A2V 불일치 풀링 결과와 anchor를 연결(concatenate)한 뒤 MLP 분류기를 통해 A2V 기반 판별 점수를 산출한다.

최종적으로 오디오 전용 판별과 A2V 기반 판별을 confidence 모듈로 융합한다. confidence 모듈의 목적은 샘플별로 어떤 브랜치가 더 강한 위조 단서를 제공하는지 추정하여 가중치를 조절하는 것이다. 이를 위해 confidence 모듈은 (1) 고정된 오디오 인코더가 추출한 전역 오디오 특징과 (2) A2V 불일치-전역 시각 결합 특징을 입력으로 받아 두 브랜치에 대한 가중치를 출력한다. 이 가중치는 브랜치별 판별 확률을 선형 결합하는 데 사용되며, 한 브랜치가 강하게 위조를 지시하는 경우 해당 브랜치에 더 큰 비중이 부여되도록 학습된다. 본 설계에서 confidence 모듈은 MLP 분류기 이전이 아니라, 브랜치별 점수가 산출된 이후 확률 수준에서 결합을 수행하도록 구성한다. 이는 (a) 각 브랜치가 독립적으로 최적의 판별 경계를 학습하도록 보장하고, (b) confidence 모듈이 브랜치 내부 특징을 다시 뒤틀기보다 브랜치 신뢰도 조절이라는 본래 목적에 집중하도록 하기 위함이다.

3.4 손실 함수

본 연구는 (1) 오디오 전용 representation learning 단계와 (2) 멀티모달 A2V 정합 학습 단계에서 서로 다른 목적함수를 사용한다. 아래에서는 각 단계의 손실을 간결히 정리한다.

3.4.1 Audio-only (Stage 1) 손실

오디오 전용 인코더는 지도 대조 학습으로 클래스 구조를 형성하고, auxiliary 이진 판별로 위조 여부를 직접 학습하며, 추가로 난이도(ordinal) 구조가 거리 공간에서 점진적으로 벌어지도록 제약한다. 최종 손실은 다음과 같다.

$$\mathcal{L}_{\text{audio}} = \lambda_{\text{sup}}\,\mathcal{L}_{\text{supcon}} + \lambda_{\text{aux}}\,\mathcal{L}_{\text{bce}} + \lambda_{\text{ord}}\,\mathcal{L}_{\text{ord}}$$

(1) Supervised Contrastive Loss

배치 내 임베딩 z_i(정규화), 라벨 y_i에 대해:

$$\mathcal{L}_{\text{supcon}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(z_i^{\top}z_p/\tau\right)}{\sum_{a\neq i}\exp\left(z_i^{\top}z_a/\tau\right)}$$

(2) Auxiliary Binary Classification

오디오 특징 f_i로부터 얻은 로짓 s_i와 이진 라벨 $b_i\in\{0,1\}$에 대한 BCE.

(3) Ordinal Ring Constraint (거리 그라데이션)

Real 프로토타입 p_{real}에 대한 거리 $d_i = \| z_i - p_{real} \|_2$를 정의하고, 레벨 k와 k+1의 평균 거리가 margin m_{ord}만큼 증가하도록 제약한다.

$$\mathcal{L}_{\text{ord}} = \frac{1}{K}\sum_{k=0}^{K-1}\max\left(0,\ m_{\text{ord}} - \left(\bar{d}_{k+1} - \bar{d}_k\right)\right)$$

3.4.2 A2V + Fusion (Stage 2) 손실

멀티모달 학습에서는 (i) 최종 위조 판별을 위한 분류 손실과 (ii) 정상 샘플에서 A2V 예측 시각 특징이 실제 시각 특징에 가까워지도록 하는 정합 손실을 결합한다. 최종 손실은 다음과 같다.

$$\mathcal{L}_{\text{av}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}}$$

(1) Final Classification Loss (video-level)

Confidence 기반 융합으로 얻은 최종 확률 p_{fuse}와 비디오 라벨 $y\in\{0,1\}$에 대한 BCE.

(2) Alignment Loss (Directional A2V consistency)

비주얼 시퀀스 특징 v_t와 A2V mapper가 예측한 특징 $\hat{v}_t$의 평균 불일치 정도를

$$d = \frac{1}{S}\sum_{t=1}^{S}\|v_t - \hat{v}_t\|_1$$

로 두고, 정렬 라벨 $y_{align}\in\{0,1\}$을 오디오와 비디오가 모두 정상(RVRA)인 경우 1, 그 외 0으로 정의한다. 그러면 정합 손실은

$$\mathcal{L}_{\text{align}} = y_{\text{align}}\,d + (1 - y_{\text{align}})\max\left(0,\ m_{\text{align}} - d\right)$$

이다. 여기서 m_{align}은 조작 샘플에서 최소 불일치 하한을 유도하는 margin이다.

요약하면, Stage 1은 오디오 내부의 생성 흔적을 분리하기 위한 표현 공간(군집 + 거리 그라데이션)을 형성하는 손실을 사용하고, Stage 2는 최종 판별(BCE)과 방향성 정합(align)을 결합하여 정상 샘플에서 A2V 예측이 실제 시각 특징과 일치하도록 유도한다.

4. 데이터셋: 고동기화 Voice Conversion 기반 Refined RVFA 구축

기존 RVFA 구성은 실비디오 위에 타 화자 오디오를 결합하거나 합성 오디오를 삽입하는 방식이 많아, 입모양-발화 타이밍 불일치가 쉽게 발생한다. 이 경우 모델이 오디오 내부의 생성 흔적보다는 싱크 오류에 의존하는 편향이 생기기 쉽다. 이를 완화하기 위해 본 연구는 고동기화 조건에서 음색만 변환되는 Refined RVFA-VC 및 Refined RVFA-SVC를 구축한다.

데이터 소스는 다음과 같다. 기본 실비디오-실오디오 쌍은 FakeAVCeleb의 RVRA에서 선택한다. 각 샘플에 대해 source 오디오는 해당 비디오에 포함된 원본 발화이며, reference 오디오는 외부 화자 데이터에서 동일 화자의 발화를 선택한다. 동일 화자 발화 선택을 위해, 대규모 공개 화자 데이터셋인 VoxCeleb2를 사용한다[6]. 이후 Seed-VC를 이용해 source 발화의 언어 내용과 시간 구조(발화 길이, 타이밍, 리듬)는 유지하면서, reference 발화로부터 추출된 음색 특성을 반영하도록 오디오를 변환한다[7]. Seed-VC는 제로샷 음성 변환을 목표로 하며, 참조 발화의 전체 컨텍스트를 활용해 화자 음색을 정밀하게 반영하도록 설계된 diffusion transformer 기반 프레임워크이다.

본 연구는 참조 발화의 선택 난이도에 따라 두 가지 변형을 만든다. 첫째, Refined RVFA-VC는 동일 화자의 다른 발화를 reference로 사용한다. 이 경우 비디오의 입모양 및 발화 타이밍은 유지되지만, 음색이 변환되며 변환 과정의 미세 흔적이 포함된다. 둘째, Refined RVFA-SVC는 동일 화자의 동일 발화를 reference로 사용한다. 이 설정은 화자 음색 변화 요인이 최소화되는 대신, 변환 과정이 남기는 생성 흔적 자체를 더 직접적으로 드러내는 고난도 케이스를 제공한다. 결과적으로 두 데이터 모두 립싱크 품질이 높은 상태에서 오디오 위조 단서가 주로 남도록 구성되며, 모델이 단순 동기화 오류가 아닌 오디오 생성 흔적과 방향성 정합 단서를 학습하도록 유도한다.

데이터 규모는 500명의 화자를 기준으로 구성한다. 각 화자에 대해 (1) RVRA 원본 1개, (2) RVFA-VC 1개, (3) RVFA-SVC 1개를 생성하여, 추가적으로 1000개의 고동기화 fake-audio 샘플을 구축한다. 이 데이터셋은 기존 RVFA 대비 동기화 편향을 줄이면서도, 오디오 내부 위조 흔적과 오디오 조건 기반 정합 검증의 효과를 평가할 수 있는 실험 환경을 제공한다.

5. 실험 결과

본 절에서는 제안한 방향성 기반 프레임워크의 효과를 네 가지 설정에서 평가한다. 모든 실험은 평균정밀도(AP)를 주요 지표로 보고하며, 비교 기준으로는 대표적인 오디오-비주얼 동기화 기반 탐지기인 AVFF를 사용한다. 실험 결과는 표 1부터 표 4까지에 각각 정리한다.

5.1 오디오 브랜치 단독 성능

Category Splits		Score	
Train	**Test**	**AUC**	**AP**
RVFA	RVFA	96.70	97.35
RVFA-VC	RVFA-VC	98.21	98.51
RVFA + RVFA-VC	RVFA	94.77	95.51
RVFA + RVFA-VC	RVFA-VC	96.38	96.84
RVFA + RVFA-VC	RVFA + RVFA-VC	94.72	95.37

표 1

첫 번째 실험은 오디오 전용 인코더가 RVFA(기존 FakeAVCeleb)와 RVFA-VC(본 연구 구축)처럼 서로 다른 조작 분포를 포함하더라도, 실제 오디오와 위조 오디오를 안정적으로 분리할 수 있는지 확인하기 위한 것이다. 이를 위해 학습·평가 데이터를 RVFA, RVFA-VC, 그리고 두 데이터의 혼합으로 다양하게 조합하여 교차 실험을 수행하였다(표 1). 표 1의 결과는 특정 조합 간 성능 우열을 비교하기보다, 다양한 조합에서도 real과 fake가 높은 수준으로 구분된다는 점에 초점을 둔다. 이는 오디오 브랜치가 단순 동기화 오류가 아니라 오디오 내부의 생성 흔적을 반영하는 판별 표현을 학습했으며, 설계한 audio-only loss가 서로 다른 위조 분포에서도 유효한 discriminative capability를 제공함을 시사한다.

5.2 Cross-manipulation 평가

Method	Modality	RVFA		FVRA-WL		FVFA		AVG-FV	
		AP	AUC	AP	AUC	AP	AUC	AP	AUC
Xception	V	-	-	88.2	88.3	83.6	84.3	84.8	85.3
LipForensics	V	-	-	**97.8**	97.7	86.7	88.9	89.4	91.1
FTCN	V	-	-	96.2	97.4	89.3	90.3	92.3	93.1
RealForensics	V	-	-	88.8	93.0	97.5	98.5	95.3	97.1
AV-DFD	AV	74.9	73.3	97.0	97.4	86.1	85.0	88.8	88.1
AVAD (LRS3)	AV	70.7	80.5	91.1	93.0	91.3	92.7	91.3	92.8
AVFF	AV	93.3	**92.4**	94.8	**98.2**	**99.8**	**99.9**	**98.5**	**99.5**
ours	AV	**94.4**	91.8	95.1	97.9	96.8	97.0	95.5	96.5

표 2

두 번째 실험은 FakeAVCeleb에서 일반적으로 수행되는

cross-manipulation 프로토콜을 따른다. 이는 학습에서 노출되지 않은 조작 유형으로의 일반화 능력을 측정하기 위해, 조작 방식(예: 특정 합성/변환 계열)을 교차하여 학습·평가를 구성하는 방식이다. 표 2에 정리한 결과에서 제안 모델은 AVFF보다는 다소 낮은 성능을 보이지만, 전반적으로 높은 탐지 성능을 안정적으로 유지한다. 이는 동기화 기반 단서가 약화되거나 조작 방식이 달라지는 조건에서도, 오디오 전용 흔적과 방향성 정합(A2V 불일치) 단서가 보조적으로 작동하여 성능 하락을 완화하기 때문으로 해석할 수 있다.

5.3 In-domain 평가

Method	Modality	ACC	AUC
Xception	V	67.9	70.5
LipForensics	V	80.1	82.4
FTCN	V	64.9	84.0
CViT	V	69.7	71.8
RealForensics	V	89.9	94.6
Emotions Don't Lie	AV	78.1	79.8
MDS	AV	82.8	86.5
AVFakeNet	AV	78.4	83.4
VFD	AV	81.5	86.1
AVoiD-DF	AV	83.7	89.2
AVFF	AV	**98.6**	**99.1**
ours	AV	97.9	98.3

표 3

세 번째 실험은 동일한 조작 분포 내에서 학습과 평가를 수행하는 in-domain 설정이다. 이 환경은 모델이 훈련 데이터와 유사한 분포에서 얼마나 높은 상한 성능을 달성할 수 있는지, 그리고 멀티모달 결합이 충분히 활용되는지를 점검하는 목적을 가진다. 표 3에서 제시하듯, 제안 모델은 AVFF 대비 낮은 수치를 보이는 항목이 존재하지만 전반적으로 높은 AP를 유지하며, 특히 오디오 기반 흔적과 A2V 정합 단서가 동시에 유효한 구간에서 안정적인 판별력을 확인할 수 있다. 이는 본 연구가 대칭적 상호 복원에 의존하지 않고도 멀티모달 결합을 통해 유의미한 성능을 확보할 수 있음을 보여준다.

5.4 새 데이터 기반 학습 및 기존 RVFA로의 확장 평가

마지막 실험은 본 연구에서 구축한 고동기화 음성 변환 데이터(Refined RVFA-VC 및 Refined RVFA-SVC) 환경에서 제안 모델의 효과를 검증하기 위한 핵심 실험이다. 학습은 새 데이터로 수행하고, 평가는 동일 설정에서 모델

Method	Modality	Modules	Performance (AP)
AVFF	A+V	Full Model	87.56
ours	A	Audio-Only	85.32
ours	**A+V**	**Audio-Only + A2V**	**91.96**

표 4

구성에 따른 ablation을 함께 진행하였다. 표 4는 (1) 기존 멀티모달 기준선(AVFF, A+V), (2) 제안 모델의 오디오 단독(A, Audio-Only), (3) 제안 모델의 멀티모달(A+V, Audio-Only + A2V) 결과를 비교한다. 표 4에서 보이듯이 오디오만 사용해도 85.32 AP로 높은 성능을 보이지만, A2V 모듈을 결합한 멀티모달 구성에서 91.96 AP로 성능이 크게 향상된다. 이는 고동기화 조건에서 단순 동기화 오류 단서가 약해진 상황에서도, 오디오 내부의 생성 흔적과 방향성 기반 A2V 정합 단서가 상보적으로 작동하며 탐지 성능을 끌어올릴 수 있음을 보여준다.

정리하면, 표 1은 오디오 브랜치 단독 성능을, 표 2는 cross-manipulation 결과를, 표 3은 in-domain 결과를, 표 4는 새 데이터 학습 후 기존 RVFA 및 새 데이터에 대한 종합 평가를 각각 정리한다.

6. 결론 및 향후 연구 방향

본 연구는 오디오-비주얼 딥페이크 탐지에서 널리 사용되어 온 동기화 기반 접근이 두 모달리티를 대칭적 관계로 가정한다는 점에 주목하고, 실제 발화 구조가 갖는 비대칭성을 반영한 방향성 기반 탐지 프레임워크를 제안하였다. 기존 방식은 오디오와 비주얼을 동일한 수준에서 정렬하거나 상호 복원하도록 학습하는 경우가 많았으며, 이 과정에서 비주얼만으로 오디오를 복원하려는 경로가 내재적으로 불안정하다는 한계를 갖는다. 이에 본 연구는 비주얼→오디오 복원 경로를 배제하고, 오디오를 조건으로 비주얼 정합을 검증하는 단방향 모델링을 통해 학습 안정성과 일반화 가능성을 동시에 확보하고자 했다.

제안한 방법은 두 단계 학습 전략으로 구성된다. 먼저 오디오 단일 모달리티 representation learning을 통해 디퓨전 기반 음성 변환 과정에서 남는 생성 흔적에 강건한 오디오 전용 인코더를 학습하였다. 이후 멀티모달 단계에서는 MARLIN 기반 비디오 인코더가 추출한 전역 시각 특징을 앵커로 두고, 오디오 특징을 조건으로 시간별 시각 특징을 생성하는 A2V mapper를 학습하였다. 예측 시각 특징과 실제 시각 특징의 불일치가 정상 샘플에서는 감소하고 조작 샘플에서는 증가하도록 유도함으로써, 단순 싱

크 오류가 아닌 방향성 정합 실패를 탐지 신호로 활용하였다. 또한 오디오-only 단서와 A2V 기반 단서의 상대적 신뢰도를 샘플 단위로 조절하는 confidence 기반 융합을 통해, 한쪽 단서가 강한 경우 해당 브랜치에 더 큰 비중이 실리도록 설계하였다.

실험에서는 오디오 브랜치 단독 평가에서 85.32 AP를 달성하여 멀티모달 기준선과 유사한 수준의 탐지력을 확인했으며, cross-manipulation 및 in-domain 설정에서도 높은 성능을 유지하였다. 특히 본 연구가 구축한 고동기화 음성 변환 데이터로 학습한 뒤 기존 RVFA 및 새 데이터로 평가한 설정에서 두 모달리티를 사용한 제안 모델이 91.96 AP로 가장 높은 성능을 기록하였다. 이는 동기화 오류 단서가 약화된 환경에서도 방향성 정합과 오디오 생성 흔적을 결합하는 설계가 효과적으로 작동함을 뒷받침한다. 향후에는 더 다양한 비디오 조작 유형과 도메인 변화에 대한 확장 평가, confidence 융합의 calibration 강화, 그리고 고동기화 데이터의 규모·다양성 확대를 통해 일반화 성능을 추가로 검증할 필요가 있다.

참고문헌

[1] Hasam Khalid, Shahroz Tariq, Minha Kim, Simon S. Woo. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. arXiv:2108.05080, 2021.

[2] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C. V. Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild. arXiv:2008.10010, 2020.

[3] Joon Son Chung, Andrew Zisserman. Out of Time: Automated Lip Sync in the Wild. 2016.

[4] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, Munawar Hayat. MARLIN: Masked Autoencoder for facial video Representation LearnINg. CVPR, 2023.

[5] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. arXiv:2201.02184, 2022.

[6] Joon Son Chung, Arsha Nagrani, Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. arXiv:1806.05622, 2018.

[7] Songting Liu. Zero-shot Voice Conversion with Diffusion Transformers. arXiv:2411.09943, 2024.

[8] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan. Supervised Contrastive Learning. NeurIPS, 2020. arXiv:2004.11362.