Biweekly Report
통권 제92호(2026. 10-2)

www.copyright.or.kr

저작권 이슈 트렌드



COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

본 보고서는 EC21R&C(컨설팅사)에서 작성하였고, 국내외 저작권 기술·산업 동향을 조사한 자료로 한국저작권위원회 의견이 반영되어 있지 않습니다.

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제92호(2026. 10-2)

- 생성형 인공지능 텍스트에 새기는 엔트로피 보정 워터마크
- 하이브리드 음악의 AI 생성 스템을 찾는 스템 비의존 탐지
- 오디오 언어 모델로 진화하는 위조 인지 화자 검증

생성형 인공지능 텍스트에 새기는 엔트로피 보정 워터마크

뉴스 브리프

생성형 AI가 작성한 글이 빠르게 늘면서, 텍스트의 출처를 기계가 식별할 수 있게 하는 기술이 주목받고 있다. 오픈AI는 유럽연합 인공지능법의 투명성 요건에 대응해 유럽연합 내 챗GPT 출력에 보이지 않는 워터마크를 넣기로 했다. 이 워터마크는 텍스트그레인이라는 기술로 모델의 단어 선택에 통계적 신호를 더하는 방식이다. 핵심 쟁점은 출력 품질과 응답 다양성을 유지하면서 탐지 가능한 신호를 남길 수 있는가, 그리고 편집된 텍스트에서도 그 신호가 유지되는가에 있다. 본 보고서는 텍스트그레인이 워터마크를 생성하고 탐지하는 원리를 살펴보고, 이 기술이 AI 산출물 식별과 어떻게 연결되는지를 분석한다. 나아가 텍스트 출처 식별 기술의 확산에 필요한 과제와 전망을 짚어본다.

뉴스 플러스

Ⅰ. 서론: AI 생성 텍스트 표시 의무화와 기술적 배경

- **챗GPT 텍스트 워터마크 도입과 출처 식별의 한계**

미국의 인공지능 개발 기업 오픈AI(OpenAI)는 2026년 10월 유럽연합 인공지능법(EU AI Act)*에 대응해 유럽연합 지역에서 제공되는 챗GPT(ChatGPT)와 코딩 보조 도구인 코덱스(Codex) 텍스트 출력에 보이지 않는 워터마크(watermark)**를 넣는다고 발표했다. 이 법의 투명성 규정은 생성형 AI 제공자에게 AI가 생성한 텍스트를 기계가 읽을 수 있는 방식으로 식별할 수 있게 하도록 요구한다. 이 워터마크는 눈에 보이는 기호가 아니라 모델이 어떤 단어를 고를지를 미세하게 조정해 남기는 통계적 신호로, 사람의 눈으로는 알아볼 수 없지만 탐지기는 찾아낼 수 있다. 신호가 문장을 이루는 단어 자체에 담기므로, 텍스트를 복사해 다른 문서에 붙여 넣어도 워터마크가 함께 이동한다. 오픈AI는 유럽연합 외 지역에는 워터마크를 기본 적용하지 않고, API를 이용하는 전 세계 개발자가 일부 모델에서 선택적으로 활성화할 수 있도록 했다.

다만 텍스트는 다시 쓰거나 편집하기 쉬워, 단어 일부만 바꿔도 워터마크 신호가 약해질 수 있다는 한계를 안고 있다. 오픈AI의 자체 평가에서도 단어 일부를 동의어로 바꾸면 워터마크 탐지율이 떨어졌으며, 바꾼 단어의 비율이 높아질수록 탐지율은 더 낮아지는 경향을 보였다. 길이가 짧은 텍스트, 수학 풀이처럼 단어 선택의 폭이 좁은 텍스트, 번역을 거친 텍스트 등에서도 워터마크 탐지가 어렵다. 또한 워터마크는 오픈AI의 시스템이 텍스트의 일부를 생성하거나 처리했다는 점을 나타낼 수 있을 뿐, 사람의 편집과 창의성이 얼마나 반영되었는지, 텍스트의 소유자가 누구인지 등은 판단하지 않는다.

* 유럽연합 인공지능법(EU AI Act): 유럽연합이 인공지능 시스템에 관한 의무를 정한 규범으로, 생성형 AI 제공자가 AI 생성 콘텐츠를 기계가 읽을 수 있는 방식으로 표시하도록 하는 투명성 규정을 담은 법령

** 워터마크(watermark): 콘텐츠에 식별 신호를 삽입해 출처나 AI 생성 여부를 확인할 수 있게 하는 기술

• 샘플링 다양성을 보존하는 비편향 워터마크의 필요성

대규모 언어 모델(Large Language Model, 이하 LLM)은 단어 또는 단어의 일부인 토큰(token)을 하나씩 이어 붙이는 방식으로 문장을 만든다. 이때 LLM은 매 단계 앞선 문맥을 바탕으로 후보 토큰마다 선택될 확률을 계산하고, 그 확률에 따라 하나를 뽑는 샘플링(sampling)을 거쳐 다음 토큰을 정한다. 텍스트 워터마크는 이 샘플링 과정에 비밀키(secret key)*로 만든 무작위 값을 개입시켜 토큰 선택에 통계적 신호를 남긴다. 탐지기는 같은 비밀키로 무작위 값을 다시 계산한 뒤, 텍스트에 담긴 토큰이 그 값과 연관되어 있는지를 검사한다. 비편향 워터마크(unbiased watermark)는 비밀키에 따라 토큰 확률을 바꾸되 그 변화가 특정 토큰 쪽으로 치우치지 않아, 평균적으로는 워터마크가 없을 때와 같은 확률을 유지하는 방식이다.

그러나 비편향 워터마크라도 실제 서비스처럼 하나의 비밀키를 계속 쓰는 환경에서는 샘플링이 지닌 무작위성이 줄어들 수 있다. 일부 기존 방식은 비밀키와 문맥이 같으면 매번 같은 토큰을 고르기 때문에, 동일한 질문에 동일한 응답이 반복될 수 있다. 오픈AI가 대학 연구진과 함께 공개한 텍스트그레인(textGrain)은 워터마크 신호를 얻는 대가로 줄어드는 엔트로피(entropy)**를 일종의 예산으로 정해 두고, 그 범위 안에서 다음 토큰을 고르는 방식을 제시한다. 이를 통해 탐지 가능한 신호를 남기면서도 샘플링 다양성을 얼마나 유지할지를 조절할 수 있다는 점에서, 텍스트 워터마크의 실용성을 높이려는 시도로 볼 수 있다.

* 비밀키(secret key): 워터마크를 넣는 쪽과 탐지하는 쪽만 공유하는 값으로, 토큰 선택에 쓰이는 무작위 값을 똑같이 다시 만들어 내는 데 필요한 정보

** 엔트로피(entropy): 어떤 결과가 나올지 예측하기 어려운 정도, 즉 무작위성의 크기를 나타내는 정보 이론의 척도

II. 본론: 텍스트그레인의 워터마크 생성과 탐지 원리

• 비밀키로 생성한 무작위성과 토큰 선택의 결합

다음 토큰 예측(Next Token Prediction, 이하 NTP) 분포는 앞선 텍스트가 주어졌을 때, 모델이 사용할 수 있는 전체 토큰의 집합인 어휘에서 각 토큰이 다음 자리에 올 확률을 모아 놓은 것이다. 다시 말해 NTP 분포는 LLM이 다음 토큰을 정하기 직전에 계산해 두는, 후보 토큰별 선택 확률의 목록이다. 텍스트그레인은 토큰을 생성하는 자리마다 비밀키와 직전 토큰 몇 개로 이루어진 문맥 창(context window)을 입력으로 삼아 무작위 값을 만들며, 비밀키와 문맥 창이 같으면 언제나 같은 값이 그대로 다시 만들어진다. 워터마크는 이 무작위 값과 실제로 선택되는 토큰 사이에 통계적 연관이 생기도록 하는 방식으로 텍스트에 삽입된다.

텍스트그레인은 이 연관을 커플링(coupling)* 형태로 설계해, 무작위 값에 따라 토큰 확률은 달라지더라도 모든 무작위 값에 대해 평균을 내면 NTP 분포와 같아지도록 한다. 이 조건을 갖춘 커플링이 곧 비편향 워터마크이며, 워터마크를 넣어도 모델의 평균적인 출력이 달라지지 않으므로 출력 품질을 유지해야 하는 서비스에서 도입을 검토할 수 있는 바탕이 된다. 토큰과 무작위 값이 서로 무관한 커플링은 워터마크 없이 생성하는 경우와 같아서, 탐지기가 찾아낼 신호가 남지 않는다. 반대로 무작위 값이 토큰을 완전히 결정하면 탐지기가 찾아낼 신호는 가장 강해지지만, 비밀키와 문맥 창이 같을 때마다 언제나 같은 토큰만 선택되어 샘플링 무작위성이 모두 사라지고 같은 질문에도 매번 같은 응답이 나오게 된다.

텍스트그레인은 커플링이 두 극단 사이 어디에 놓이는지를 엔트로피 감소량으로 나타내며, 이는 무작위 값이 정해졌을 때 줄어드는 엔트로피의 평균이다. 엔트로피 감소량은 토큰과 무작위 값 사이의 상호정보량(mutual information)**과 같고, 두 값이 서로 무관한 상태에서 커플링이 벗어난 정도와도 일치한다. 이는 토큰과 무작위 값 사이의 연관을 키울수록 샘플링 다양성이 줄어들며, 연관의 크기와 줄어든 엔트로피가 같은 값으로 나타난다는 것을 의미한다. 탐지기는 텍스트에 남아 있는 이 연관을 확인하는데, 무작위 값은 비밀키 없이는 다시 만들 수 없으므로 워터마크 여부의 검증도 비밀키를 가진 쪽에서만 이루어질 수 있다. 이러한 점에서 커플링은 AI 산출물의 출처를 토큰 선택 자체에 새기는 장치로 작동하며, 별도의 표시 없이도 생성 여부를 확인할 근거를 제공한다.

* 커플링(coupling): 두 대상이 함께 나타날 확률을 정한 분포로, 각 대상을 따로 떼어 보면 원래의 분포가 그대로 유지되도록 만든 확률 분포

** 상호정보량(mutual information): 두 값 가운데 하나를 알게 되었을 때 다른 하나에 대한 불확실성이 줄어드는 정도를 나타내는 정보 이론의 척도

• 엔트로피 예산으로 조절하는 워터마크 신호 강도

텍스트그레인은 토큰과 무작위 값 사이의 커플링을 정할 때 최적 수송(Optimal Transport, 이하 OT)이라는 계산 방법을 쓴다. OT는 두 분포를 짝지을 때 각 분포의 확률은 그대로 지키면서, 짝마다 매긴 비용의 합이 가장 작아지도록 짝을 정하는 방법이다. 텍스트그레인은 토큰과 무작위 값의 짝마다 비밀키로

비용을 매겨 두고 OT로 비용이 작은 짝이 더 자주 선택되도록 하는데, 이렇게 자주 선택되는 짝이 곧 탐지기가 찾는 워터마크 신호가 된다. 다만 비용이 작은 짝만 고르면 무작위 값이 토큰을 완전히 결정하는 극단에 가까워지므로, 두 값이 무관한 상태에서 커플링이 벗어난 정도인 쿨백-라이블러 발산(Kullback-Leibler divergence, 이하 KL 발산)*을 함께 제한한다.

KL 발산은 앞서 설명한 엔트로피 감소량과 같은 값이므로, KL 발산을 제한하는 것은 샘플링 다양성이 얼마나 줄어들지를 미리 정해 두는 것과 같다. 텍스트그레인은 이 한도를 엔트로피 예산(entropy budget)이라 부르며, 원래 엔트로피의 어느 정도까지 줄어들어도 되는지를 비율로 정한다. 예산을 크게 잡으면 신호가 강해져 AI 산출물을 식별하기 쉬워지지만 샘플링 다양성은 줄어들고, 예산을 작게 잡으면 그 반대가 된다. 서비스 운영자는 여러 답안이 필요한지, AI 산출물 식별이 더 중요한지에 따라 예산을 정할 수 있으며, 탐지기는 생성할 때 쓴 예산을 몰라도 같은 절차로 신호를 찾을 수 있어 서비스마다 예산을 달리 정하기도 쉬워진다. 다만 예산은 평균적인 엔트로피 감소량만 제한할 뿐 개별 비밀키마다 엔트로피 감소량이 예산 안에 든다고 보장하지는 않으며, 실제 계산 결과도 목표 예산과 어긋날 수 있다.

* 쿨백-라이블러 발산(Kullback-Leibler divergence): 두 확률 분포가 서로 얼마나 다른지를 나타내는 정보 이론의 척도로, 두 분포가 같으면 0이 되고 차이가 클수록 커지는 값

• 어휘 블록 분할과 최적 수송을 통한 토큰 생성

텍스트그레인은 토큰을 하나 고를 때마다 먼저 비밀키와 문맥 창을 이용해 어휘를 서로 겹치지 않는 여러 개의 블록으로 나눈다. 그리고 블록마다 그 안에 든 토큰들의 확률을 모두 더해 블록 확률을 구하는데, 이후의 OT 계산은 어휘에 속한 토큰 하나하나가 아니라 이 블록 확률을 대상으로 이루어진다. 어휘 전체를 직접 다루지 않으므로 계산량이 줄어들고, 블록을 나누는 방식은 비밀키와 문맥 창으로만 정해지므로 탐지기도 NTP 분포를 알지 못한 채 텍스트와 비밀키만으로 같은 블록을 그대로 다시 만들 수 있다. 워터마크 신호는 어떤 블록이 선택되는지에만 담기고, 블록 안에서 어떤 토큰을 고를지는 원래의 NTP 분포를 그대로 따른다.

OT 계산의 결과는 여러 개의 열로 정리되는데, 각 열은 블록 확률을 저마다 다르게 조정해 둔 후보 목록이며 모든 열을 평균하면 원래의 블록 확률과 같아진다. 토큰을 고를 때는 먼저 비밀키와 문맥 창으로 열 하나를 정하고, 그 열에 적힌 확률에 따라 블록 하나를 뽑는다. 마지막으로 뽑힌 블록 안에서 원래의 NTP 분포에 따라 토큰 하나를 고르면, 그 토큰이 다음 자리에 들어간다. 어떤 열이 정해질지는 비밀키를 모르면 알 수 없지만 모든 열을 평균해 보면 원래 확률이 유지되므로, 출력 품질을 지키면서도 비밀키를 가진 쪽만 알아볼 수 있는 AI 산출물의 신호가 텍스트 곳곳에 남게 된다.

블록 단위 계산은 토큰마다 OT를 푸는 부담을 줄여, 많은 사용자가 이용하는 서비스에서도 워터마크를 적용하기가 쉬워질 것으로 보인다. 또한 텍스트그레인은 생성 속도를 높이는 추측 샘플링(speculative sampling)*과 함께 쓸 수 있으며, 초안 모델과 목표 모델이 같은 비밀키를 쓰면 목표 모델이 직접 생성했을 때와 같은 워터마크가 그대로 유지된다. 이는 서비스 속도를 높이는 기법을 적용하더라도 탐지기가 같은 방식으로 신호를 찾아 워터마크 여부를 확인할 수 있다는 것을 의미한다. 다만 실제로 속도가 얼마나 빨라지는지는 초안 모델이 제안한 토큰이 얼마나 받아들여지는지와, 두 모델 모두에 워터마크를 넣는 계산 비용에 따라 달라진다.

* 추측 샘플링(speculative sampling): 작은 초안 모델이 여러 토큰을 먼저 제안하고 큰 목표 모델이 이를 한 번에 검증해 받아들일지 정하는 방식으로, LLM의 생성 속도를 높이기 위한 기법

• 텍스트와 비밀키만으로 수행하는 워터마크 탐지

탐지기는 생성할 때와 같은 비밀키와 문맥 창 규칙을 이용해, 텍스트의 각 자리에서 블록 분할, 비용, 선택된 열을 다시 만든다. 이어서 실제로 텍스트에 쓰인 토큰이 어느 블록에 속하는지 확인하면 그 블록과 선택된 열의 짝에 매겨진 비용을 알 수 있으며, 탐지기는 이를 자리별 탐지 점수로 변환해 비용이 낮은 조합에 더 높은 점수를 부여한다. 워터마크가 있는 텍스트는 생성 과정에서 비용이 작은 짝이 자주 선택되었으므로 점수가 높게 나오고, 워터마크가 없는 텍스트의 점수는 우연에 따라 흩어진다. 이 과정에는 생성에 쓰인 모델이나 커플링, 엔트로피 예산이 필요하지 않으므로, 모델을 다시 실행하지 않고 텍스트와 비밀키만으로 탐지를 수행할 수 있다.

탐지기는 자리별 점수를 모두 더한 뒤, 워터마크가 없는 텍스트라면 이 합이 어느 범위에 놓이는지를 미리 계산해 둔 기준과 비교한다. 합이 기준을 넘으면 워터마크가 있다고 판정하는데, 이 기준은 워터마크가 없는 텍스트를 워터마크가 있다고 잘못 판정하는 비율인 위양성률(false positive rate)을 원하는 수준으로 맞추도록 미리 정해 둔 값이다. 같은 문맥 창이 반복되면 같은 무작위 값이 다시 쓰여 위양성률 계산이 맞지 않게 될 수 있으므로, 문맥 창이 처음 나타난 자리만 점수에 포함한다. 이러한 판정 방식은 텍스트가 다른 곳으로 옮겨지더라도 비밀키를 가진 쪽이 생성 여부를 확인할 수 있게 해, AI 산출물 출처 식별의 기술적 근거가 된다. 다만 위양성률 계산은 텍스트와 비밀키가 서로 무관하다는 가정에 기대고 있어, 하나의 비밀키를 계속 쓰는 실제 서비스에서는 오류율을 따로 검증해야 한다.

III. 결론 및 전망: AI 텍스트 출처 식별의 의의와 과제

• 텍스트 출처 식별 기술의 확산과 보완 과제

결국 텍스트그레인은 AI 산출물의 출처를 텍스트의 토큰 선택 자체에 새기고, 비밀키를 가진 쪽이 텍스트만으로 이를 확인할 수 있게 하는 기술이다. 특히 엔트로피 예산으로 워터마크 신호의 세기와 샘플링 다양성 사이의 균형을 정할 수 있다는 점은, 출력 품질을 유지해야 하는 서비스에 워터마크를 도입할 여지를 넓힌다. 이는 AI 산출물을 표시하는 방식에, 텍스트 안에 통계적 신호를 남겨 기계가 식별하도록 하는 방식이 더해지고 있음을 의미한다. 이러한 점에서 텍스트 워터마크는 해당 시스템이 생성한 텍스트인지를 기술적으로 확인하는 수단이 되지만, 탐지 결과가 곧 사람의 창작 기여나 텍스트의 권리 관계를 보여 주는 것은 아니므로 기술의 역할과 한계를 구분해 이해할 필요가 있다.

다만 단어 일부를 바꾸거나 번역을 거치면 신호가 약해지고, 짧은 텍스트에서는 탐지가 어렵다는 한계가 남아 있다. 따라서 편집을 거친 텍스트에서도 신호를 찾을 수 있도록 탐지 기술을 보완하고, 하나의 문서 안에 사람이 작성한 부분과 AI가 생성한 부분이 섞인 텍스트에서 워터마크가 담긴 구간을 찾아내는 기술도 함께 발전할 필요가 있다. 제도 측면에서는 탐지기에 접근할 수 있는 주체와 탐지 결과의 활용 범위를 정하는 기준이 마련되어야 하며, 탐지 결과를 판단의 유일한 근거로 삼지 않도록 하는 원칙도 필요하다. 산업 측면에서는 여러 AI 기업의 워터마크를 함께 확인할 수 있는 공통 기준이 갖춰질 때, 텍스트 출처 식별이 콘텐츠 유통 전반으로 확산될 수 있을 것으로 보인다. 이를 통해 텍스트 워터마크는 AI 산출물과 사람의 창작물이 함께 유통되는 환경에서 출처를 확인하는 기술적 수단으로 자리 잡을 가능성이 있다.

참고문헌

- Aditya Mehta, “OpenAI will start watermarking ChatGPT’s text in the EU”, TechCrunch, 2026.10.05., https://techcrunch.com/2026/10/05/openai-will-start-watermarking-chatgpts-text-in-the-eu/
- OpenAI, “EU 텍스트 출처 규정에 대한 OpenAI의 접근 방식”, 2026.10.05., https://openai.com/ko-KR/index/eu-text-provenance/
- Xiang Li 외 8명, “textGrain: Entropy-Calibrated Watermarking for Language Model Text”, OpenAI, 2026.10.05., https://cdn.openai.com/pdf/e9508624-d767-41b6-a26d-e34ca798ada6/textgrain-entropy-calibrated-watermarking-for-language-model-text.pdf

기술용어

순번	용어	설명
1	**유럽연합 인공지능법** (EU AI Act)	유럽연합이 인공지능 시스템에 관한 의무를 정한 규범으로, 생성형 AI 제공자가 AI 생성 콘텐츠를 기계가 읽을 수 있는 방식으로 표시하도록 하는 투명성 규정을 담은 법령
2	**워터마크** (watermark)	콘텐츠에 식별 신호를 삽입해 출처나 AI 생성 여부를 확인할 수 있게 하는 기술
3	**비밀키** (secret key)	워터마크를 넣는 쪽과 탐지하는 쪽만 공유하는 값으로, 토큰 선택에 쓰이는 무작위 값을 똑같이 다시 만들어 내는 데 필요한 정보
4	**엔트로피** (entropy)	어떤 결과가 나올지 예측하기 어려운 정도, 즉 무작위성의 크기를 나타내는 정보 이론의 척도
5	**커플링** (coupling)	두 대상이 함께 나타날 확률을 정한 분포로, 각 대상을 따로 떼어 보면 원래의 분포가 그대로 유지되도록 만든 확률 분포
6	**상호정보량** (mutual information)	두 값 가운데 하나를 알게 되었을 때 다른 하나에 대한 불확실성이 줄어드는 정도를 나타내는 정보 이론의 척도
7	**쿨백-라이블러 발산** (Kullback-Leibler divergence)	두 확률 분포가 서로 얼마나 다른지를 나타내는 정보 이론의 척도로, 두 분포가 같으면 0이 되고 차이가 클수록 커지는 값
8	**추측 샘플링** (speculative sampling)	작은 초안 모델이 여러 토큰을 먼저 제안하고 큰 목표 모델이 이를 한 번에 검증해 받아들일지 정하는 방식으로, LLM의 생성 속도를 높이기 위한 기법

하이브리드 음악의 AI 생성 스템을 찾는 스템 비의존 탐지

뉴스 브리프

고음질 음악 스트리밍 플랫폼 코부즈가 AI 생성 음악으로 확인된 앨범에 태그를 표시하기 시작했다. 코부즈는 AI 생성 음악 재생의 60%를 부정 스트리밍으로 판정해 저작권료 정산에서 제외하고 있다. 그러나 사람의 연주와 AI 생성 스템이 섞인 하이브리드 음악은 곡 전체를 판정하는 기존 이진 탐지로는 가려내기 어렵고, 가려내더라도 어느 스템이 AI로 생성되었는지까지는 구별하지 못한다는 한계가 있다. 이 보고서는 시간과 주파수 대역별 생성 확률을 나타낸 인스펙트로그램과 스템의 에너지 정보를 결합해, 악기 종류와 관계없이 하나의 모델로 AI 생성 스템을 판별하는 스템 비의존 탐지 기술의 작동 원리를 살펴본다. 아울러 음원 분리 품질이라는 과제와 함께 AI 음악 표시 체계와 저작권료 정산 방식이 나아갈 방향을 전망한다.

뉴스 플러스

Ⅰ. 서론: 스트리밍 플랫폼의 AI 음악 표시와 과제

- **코부즈의 AI 생성 음악 태그 도입과 정산 제외**

2026년 9월 프랑스 고음질 음악 스트리밍 플랫폼 코부즈(Qobuz)는 AI 생성 음악으로 확인된 콘텐츠에 태그를 표시하기 시작했다. 태그는 앨범 단위로 붙고, AI 생성 음악이 담긴 앨범을 이용자가 앱 화면에서 바로 알아볼 수 있게 한다. 코부즈는 이와 함께 AI 생성 음악을 편집진이 직접 고르는 신규 앨범 소개, 주간 추천 앨범, 플레이리스트와 알고리즘 추천에서 모두 제외하고, 재생 기록이 전혀 없는 AI 생성 앨범의 3분의 1 이상을 검색 결과에서 내렸다. 검색 결과에서 내려간 앨범에 담긴 곡은 400만 곡에 이르며, 이 앨범들은 제목을 직접 검색할 때만 찾을 수 있다[1].

1) Mandy Dalugdug, "Qobuz starts tagging AI-generated music in its app – and says 60% of streams on AI tracks are fraudulent", Music Business Worldwide, 2026.09.28., https://www.musicbusinessworldwide.com/qobuz-starts-tagging-ai-generated-music-in-its-app-and-says-60-of-streams-on-ai-tracks-are-fraudulent/

이러한 조치는 저작권료 정산과도 이어진다. 코부즈가 확인한 AI 생성 음악의 재생 가운데 60%는 자체 탐지 체계에서 부정 스트리밍으로 판정되어 권리자에게 지급되는 저작권료 정산에서 제외된다. AI 생성 음악이 코부즈 전체 재생에서 차지하는 비중은 0.38%로, 1~3% 수준으로 알려진 경쟁사 디저(Deezer)보다 낮다. 디저에는 2026년 6월 하루 약 9만 곡의 완전 AI 생성 음악이 유입됐으며, 유입이 가장 많았던 날에는 새로 들어온 음악의 절반을 넘었다. 이처럼 AI 생성 음악의 유입이 늘면서 타이달(TIDAL), 애플 뮤직(Apple Music), 스포티파이(Spotify) 등 다른 음악 스트리밍 플랫폼도 AI 생성 음악 표시 제도를 도입하고 있다. 다만 코부즈는 사람의 개입이 일부 섞인 음악까지 확실하게 가려낼 방법은 아직 없다고 밝혔고, 이는 이용자를 위한 표시와 권리자에게 돌아가는 저작권료 정산의 정확도가 결국 탐지 기술의 수준에 달려 있음을 보여준다.

• 하이브리드 음악의 확산과 이진 탐지의 한계

생성형 음악 모델이 널리 퍼지면서 누구나 AI 생성 음악을 손쉽게 만들고 공유할 수 있게 되었고, 생성형 음악 기업과 대형 음반사 간 제휴도 이러한 흐름을 가속하고 있다. 창작자들이 생성형 도구를 제작 과정의 일부로 받아들이면서, 사람이 연주한 소리와 AI가 만든 스템(stem)을 섞은 곡도 늘고 있다. 스템은 한 곡을 이루는 보컬, 드럼, 베이스 등의 소리를 악기별로 나누어 따로 저장한 개별 음원 파일을 말한다. 사람과 AI가 함께 만든 이러한 음악을 하이브리드 음악(hybrid music)이라 부른다.

기존 AI 음악 탐지 모델은 곡 전체가 생성되었는지 아닌지를 둘 중 하나로 가르는 이진 탐지(binary detection) 방식으로 작동하며, 완전 AI 생성 음악에서는 99%가 넘는 정확도를 보인다. 그러나 생성된 소리가 사람의 연주와 섞이면 탐지 모델이 의존하는 생성 흔적이 상대적으로 약해져 판정이 어려워진다. 하이브리드 음악을 AI 생성 음악으로 맞게 판정하더라도, 곡 안의 어느 스템이 AI 생성 스템인지는 알려주지 못한다는 한계도 남는다. 이에 디저의 사내 R&D 조직인 디저 연구소(Deezer Research)와 미국 뉴욕대학교(New York University) 연구진은 하이브리드 음악에서 AI 생성 스템을 악기 종류와 관계없이 찾는 스템 비의존(stem-agnostic) 탐지 방식을 제안했다.

II. 본론: 스템 단위 AI 생성 음악 탐지의 작동 원리

• 신경 오디오 코덱으로 구성한 하이브리드 학습 데이터

기존 탐지 모델은 생성 모델의 디코더(decoder)가 소리를 만들어 내는 과정에서 남기는 생성 흔적을 찾아, 해당 곡 전체가 AI 생성 음악인지 아닌지를 판정한다. 디코더는 생성 모델 내부에서 압축된 형태로 다루던 정보를 사람이 들을 수 있는 실제 소리 파형으로 되돌리는 역할을 한다. 이 과정에서 압축된 정보를 늘려 원래 길이로 복원할 때 쓰이는 전치 합성곱(transposed convolution)** 연산은 주파수 스펙트럼에 일정한 간격으로 솟은 정점을 남기며, 탐지 모델은 이 정점을 판정의 신호로 학습한다.

연구진은 하이브리드 음악을 학습시킬 공개 데이터셋(dataset)이 아직 없다는 문제를 신경 오디오 코덱(neural audio codec)을 활용하는 방식으로 풀었다. 신경 오디오 코덱은 인공신경망(artificial neural network)을 이용해 소리를 작은 용량으로 압축했다가 디코더로 다시 복원하는 기술로, 연구진은 그중 인코덱(EnCodec)을 사용했다. 인코덱의 디코더는 소리를 복원하는 과정에서 생성 모델의 디코더와 비슷한 생성 흔적을 남기므로, 같은 스템을 인코덱에 통과시키기 전과 후로 짝지으면 연주 내용은 같고 생성 흔적의 유무만 다른 학습 자료를 만들 수 있다. 이를 통해 탐지 모델이 연주 내용이나 곡의 분위기 차이가 아니라 생성 흔적 자체를 근거로 판정하도록 학습 과정을 유도할 수 있으며, 다른 변수의 영향도 줄일 수 있다. 이 연구의 AI 생성 스템은 실제 생성 서비스로 만든 스템이 아니라, 인코덱에 통과시켜 생성 흔적을 남긴 스템을 말한다.

탐지 모델의 학습에는 음원 분리 연구에 널리 쓰이는 공개 데이터셋 MUSDB18-HQ에 담긴 150곡이 사용되었다. 각 곡은 보컬, 드럼, 베이스, 그 밖의 악기로 구성된 4개 스템을 가지며, 연구진은 스템마다 실제 스템과 AI 생성 스템 중 하나를 고르는 16가지 조합을 모두 만들었다. 학습에 없던 악기에서의 성능을 확인하기 위해, 4개 이상의 다양한 스템으로 분리된 곡을 담은 음원 분리용 공개 데이터셋 MoisesDB도 함께 활용했다. MoisesDB의 240곡은 곡마다 들어 있는 악기 구성이 달라, 연구진은 이를 보컬, 드럼, 베이스, 기타, 피아노, 그 밖의 악기라는 6개 스템으로 정리하고 곡마다 생성할 스템을 무작위로 골라 평가용 데이터셋을 구성했다.

* 전치 합성곱(transposed convolution): 작은 크기의 데이터를 더 큰 크기로 늘리면서 그 사이 값을 채워 넣는 인공신경망 연산으로, 늘리는 간격이 일정해 결과물에 규칙적인 무늬가 남을 수 있음

• 대역 분할 회귀와 인스펙트로그램의 생성 원리

연구진은 데이터셋을 확보한 뒤 탐지의 첫 단계로 인스펙트로그램(inspectrogram)을 도입했다. 이는 곡의 시간대와 주파수 대역마다 생성 흔적이 있을 확률을 계산해 표시한 2차원 자료다. 가로축은 시간, 세로축은 주파수 대역이며, 각 칸에는 생성된 소리가 섞여 있을 확률이 0에서 1 사이 값으로 담긴다. 그 값이 1에 가까울수록 그 칸의 생성 흔적이 뚜렷하다. 인스펙트로그램을 만들기 위해 연구진은 먼저 완전 AI 생성 음악과 실제 음악을 구별하도록 탐지 모델 하나를 따로 학습시켰다. 학습 자료는 음악 분석 연구를 위해 공개된 데이터셋인 프리 뮤직 아카이브(Free Music Archive, FMA)의 음악을 인코덱에 통과시키기 전과 후로 짝지어 마련했다.

이 탐지 모델은 곡 전체를 한 번에 판정하지 않고, 곡을 짧은 시간 구간과 좁은 주파수 대역으로 잘게 나누어 각각 따로 판정한다는 점에서 기존 이진 탐지와 다르다. 먼저 곡을 5초 길이의 구간으로 자르되 0.25초씩 옮겨 가며 겹치게 잘라, 시간의 흐름을 촘촘하게 따라간다. 다음으로 각 구간에서 페이크프린트(fakeprint)를 뽑는데, 이는 생성 흔적이 나타나는 주파수 성분을 정리한 특징값이다. 생성된 소리에 나타나는 주기적인 정점이 잘 드러나도록, 각 구간의 소리를 주파수별로 살펴 일정한

형식의 숫자 묶음으로 요약한 것이다. 이어 0~16kHz 범위를 500Hz 폭의 주파수 대역 32개로 나누고 주파수 대역마다 독립된 분류기를 두어 생성 흔적이 있을 확률을 계산하는데, 이 분류기 묶음을 대역 분할 회귀(Band-Split Regression, BS-Reg)라 부른다.

연구진은 구간 길이와 주파수 대역 폭을 여러 조합으로 시험했고, 구간이 길고 주파수 대역이 넓을수록 정확도는 상승하지만 시간과 주파수를 세밀히 나누는 정도는 떨어졌다. 최종적으로 고른 5초 구간과 500Hz 폭은 정확도와 세밀함 간의 균형을 가장 잘 맞춘 설정으로 평가되었다. BS-Reg는 3kHz 이상에서 성능이 안정적이었으나, 저주파에서는 소리 정보가 빽빽하게 몰려 있어 오탐이 늘고 탐지율이 떨어졌다. 이렇게 만든 인스펙트로그램은 완전 AI 생성 음악과 실제 음악만으로 학습한 탐지 모델에서 나왔는데도, 하이브리드 음악에서 생성된 소리가 있는 위치를 시간과 주파수 대역 단위로 보여준다. 다만 인스펙트로그램에는 잘못된 판정값이 일부 섞여 있어 이것만으로는 생성된 소리가 어느 스템에서 왔는지 알 수 없다.

• 위너 필터 에너지 마스크와 스템 비의존 CNN

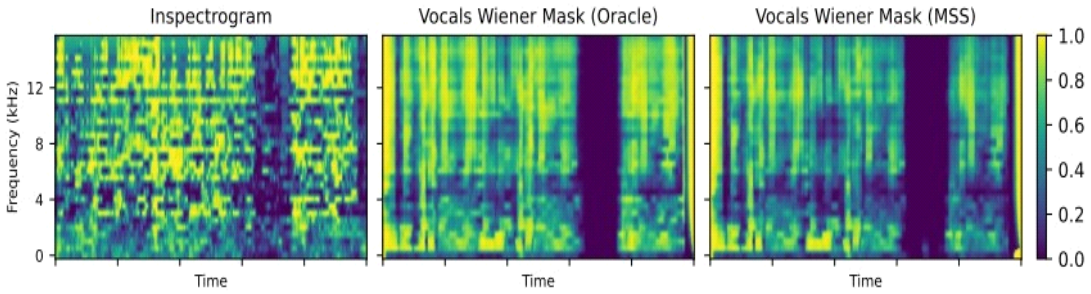
인스펙트로그램이 보여주는 생성 흔적의 위치를 특정 스템과 이어 주는 역할은, 위너 필터(Wiener filter)*의 원리로 계산한 스템별 에너지 마스크(energy mask)가 맡는다. 에너지 마스크는 혼합 음원 전체 에너지 가운데 특정 스템의 에너지가 차지하는 비율을 시간과 주파수 대역별로 0에서 1 사이 값으로 나타낸 자료다. 연구진은 스템의 에너지가 우세한 주파수 대역일수록 AI 생성 스템을 찾기 쉽다는 선행 연구의 관찰을 바탕으로, 에너지 마스크와 인스펙트로그램에서 높은 값이 함께 나타나는 곳이 많으면 그 스템이 AI 생성 스템일 가능성이 크다고 가정했다. 에너지 마스크도 인스펙트로그램과 같은 5초 구간과 32개 주파수 대역 단위로 계산해, 두 자료를 칸 단위로 비교해 볼 수 있게 했다.

에너지 마스크는 녹음 단계에서 따로 저장된 스템을 그대로 쓰는 정답 스템(ground-truth stem) 조건과, 완성된 혼합 음원에서 스템을 분리해 쓰는 분리 스템(separated stem) 조건에서 각각 따로 계산되었다. 음원 분리는 여러 악기가 함께 섞인 혼합 음원에서 특정 악기의 소리만 따로 떼어 내 별도의 음원으로 만드는 기술이다. 스트리밍 플랫폼에는 스템이 아닌 완성된 곡만 들어오므로 분리 스템 조건이 실제 운영 환경에 가까우며, 연구진은 이 조건에서 음원 분리 모델 HTDemucs의 4개 스템 버전과 기타 및 피아노까지 나누는 6개 스템 버전을 함께 사용했다.

판별은 합성곱 신경망(Convolutional Neural Network, CNN)이 맡는데, CNN은 격자 형태로 배열된 자료에서 이웃한 칸들의 패턴을 단계적으로 읽어 내는 인공신경망이다. CNN은 인스펙트로그램과 에너지 마스크를 겹친 두 겹의 입력을 약 20초 단위로 받아 대상 스템이 AI 생성 스템인지 판정하고, 곡 전체의 판정값은 구간별 판정값의 평균으로 구한다. 이 CNN은 악기 종류를 따로 학습하지 않고 에너지 마스크로 대상 스템을 지정하므로, 스템마다 전용 탐지 모델을 따로 두던 기존 방식과 달리 에너지 마스크만 주어지면 하나의 모델로 혼합 음원 속 어떤 스템이든 판정할 수 있다.

* 위너 필터(Wiener filter): 여러 신호가 섞인 소리에서 원하는 신호가 차지하는 비중을 계산해, 그 비율만큼 해당 신호를 남기도록 설계된 필터를 말함

[그림 1] 하이브리드 음악의 인스펙트로그램(왼쪽)과 에너지 마스크(가운데, 오른쪽) 비교



출처: Richa Namballa 외 2명, "A Stem-Agnostic Approach to Hybrid AI Music Detection", arXiv, 2026.09.22., https://arxiv.org/pdf/2609.26956v1

• **음원 분리 품질 및 베이스와 피아노 탐지의 한계**

연구진은 보컬, 드럼, 그 밖의 악기 스템으로 CNN을 학습시켰으며, 하나의 모델이 정답 스템 조건과 분리 스템 조건 모두에서 세 스템을 높은 정확도로 판별했다. 정답 스템으로 학습한 모델은 분리 과정에서 소리가 일그러진 분리 스템 조건에서도 성능을 유지했고, 학습에 쓰지 않은 MoisesDB에서도 보컬과 드럼의 탐지 성능이 이어졌다. 학습 단계에서 그 밖의 악기에 묶여 있어 따로 학습되지 않았던 기타와 피아노도 정답 스템 조건에서는 비교적 정확하게 판별되었다. 이러한 결과는 하이브리드 음악 안에서 정확히 어느 악기가 AI 생성 스템인지 특정할 수 있다는 점에서, 태그 표시와 저작권료 정산을 곡 전체가 아니라 해당 AI 생성 스템의 범위에 맞춰 조정하는 기술적 근거로 활용될 가능성이 있다.

반면 베이스는 분리 스템 조건에서 탐지 성능이 무작위로 고르는 수준까지 떨어져, 연구진은 베이스를 최종 판별 대상에서 제외했다. HTDemucs는 베이스 에너지가 몰린 0~500Hz 주파수 대역을 보존하는 데 집중하면서 그 위의 배음(harmonics)*과 음이 시작되는 순간의 소리를 걸러 내는데, AI 생성 스템을 가려낼 정보는 바로 이 높은 주파수 대역에 담겨 있었다. 그 결과 베이스의 에너지 마스크가 인스펙트로그램에서 판정이 가장 불안정한 저주파 주파수 대역에 몰렸고, 모델은 대부분의 베이스를 AI 생성 스템으로 판정하는 쪽으로 치우쳤다. 피아노도 분리가 어려운 악기여서 분리 스템 조건에서 오탐이 늘었으며, 대상 스템의 소리가 약하거나 없는 구간에서는 판정값이 0.5 부근에 머물렀다. 실제 스템이 AI 생성 스템으로 잘못 판정되는 것을 방지하기 위해, 탐지 결과를 쓰기 전에 스템의 에너지와 분리 품질을 확인하는 후처리가 필요하다.

* 배음(harmonics): 악기가 한 음을 낼 때 기본음과 함께 그 정수배 주파수에서 울리는 소리로, 같은 음이라도 악기마다 음색이 달라지는 원인이 됨

III. 결론: 스템 단위 탐지와 AI 음악 표시의 정교화

• 음원 분리 고도화와 코덱 확장 대응 과제

결국 하이브리드 음악의 확산은 AI 생성 음악을 곡 단위로 가려내던 탐지 방식이 스템 단위로 옮겨 가야 함을 보여준다. 스템 비의존 탐지 방식은 인스펙트로그램으로 생성 흔적의 위치를 찾고 에너지 마스크로 그 위치를 특정 스템과 연결해, 어느 스템이 AI 생성 스템인지를 하나의 모델로 판별한다. 이를 통해 스트리밍 플랫폼은 실제 스템과 AI 생성 스템을 나누어 표시하고, 부정 스트리밍 차단과 저작권료 정산의 근거를 지금보다 세밀하게 마련할 수 있으며, 창작자는 AI를 일부 활용한 작업도 그 범위를 드러낼 수 있을 것으로 보인다. 이는 AI 음악 표시가 AI 생성 여부라는 단순한 구분을 넘어, 곡을 구성하는 어느 스템에 AI 생성 요소가 포함되었는지 알려주는 방향으로 넓어질 수 있음을 의미한다.

다만 이 기술의 성능은 음원 분리 품질에 묶여 있어, 베이스와 피아노처럼 분리가 어려운 스템에서는 아직 판정을 신뢰하기 어렵다. 이를 보완하려면 더 정교한 음원 분리 모델을 도입하거나, 인스펙트로그램과 같은 크기의 에너지 마스크를 바로 계산하는 가벼운 전용 분리 모델을 개발해야 한다. 연구진은 새로운 코덱이나 생성 서비스가 나와도 가벼운 BS-Reg만 다시 학습하고 CNN은 그대로 쓸 수 있을 것으로 추정해, 생성 모델이 바뀔 때마다 탐지 체계 전체를 다시 만들 부담은 줄어들 수 있다. 그러나 이 추정은 하나의 신경 오디오 코덱으로 만든 AI 생성 스템으로만 실험한 결과에 근거하므로, 실제 상업 생성 서비스로 만든 스템에서도 같은 성능이 나오는지는 추가 검증이 필요하다.

참고문헌

- Mandy Dalugdug, “Qobuz starts tagging AI-generated music in its app – and says 60% of streams on AI tracks are fraudulent”, Music Business Worldwide, 2026.09.28., https://www.musicbusinessworldwide.com/qobuz-starts-tagging-ai-generated-music-in-its-app-and-says-60-of-streams-on-ai-tracks-are-fraudulent/
- Richa Namballa 외 2명, “A Stem-Agnostic Approach to Hybrid AI Music Detection”, arXiv, 2026.09.22., https://arxiv.org/pdf/2609.26956v1

기술용어

순번	용어	설명
1	전치 합성곱 (transposed convolution)	작은 크기의 데이터를 더 큰 크기로 늘리면서 그 사이 값을 채워 넣는 인공신경망 연산으로, 늘리는 간격이 일정해 결과물에 규칙적인 무늬가 남을 수 있음
2	위너 필터 (Wiener filter)	여러 신호가 섞인 소리에서 원하는 신호가 차지하는 비중을 계산해, 그 비율만큼 해당 신호를 남기도록 설계된 필터를 말함
3	배음 (harmonics)	악기가 한 음을 낼 때 기본음과 함께 그 정수배 주파수에서 울리는 소리로, 같은 음이라도 악기마다 음색이 달라지는 원인이 됨

오디오 언어 모델로 진화하는 위조 인지 화자 검증

뉴스 브리프

음성 합성과 음성 복제 기술이 발전하면서, 적은 비용으로도 특정인의 목소리와 닮은 합성 음성을 만들 수 있게 되었다. 이에 따라 기업의 고객 통화 현장을 중심으로 합성 음성을 실시간으로 탐지하고, 음성 지문으로 등록된 발신자를 식별하는 서비스가 잇따라 등장하고 있다. 그러나 기존 방식은 합성 음성 여부만 가리거나, 화자 검증과 위조 탐지를 별도 모델로 처리한 뒤 점수를 사후에 합치는 데 머물러 왔다. 본 보고서는 두 판정을 대규모 오디오 언어 모델 하나로 통합하는 위조 인지 화자 검증 기술의 작동 원리와 성과를 살펴본다. 아울러 판정 근거를 문장으로 제시하는 추론 학습의 가능성과 한계를 짚고, 목소리 무단 복제에 대응하는 기술적 보호의 방향과 과제를 전망한다.

뉴스 플러스

Ⅰ. 서론: 목소리 신원 확인을 위협하는 합성 음성의 확산

- **음성 지문 식별과 합성 음성 탐지의 결합 움직임**

생성형 AI를 활용한 사기 수법이 사회공학 공격(social engineering attack)*과 결합해 목소리로 신원을 판단하던 관행이 흔들리자, 신원 확인 업계는 일회성 확인을 넘어 대화 내내 신원을 점검하는 방식으로 옮겨 가고 있다. 영국의 통신 기업 원컴(Onecom)이 최근 공개한 차세대 음성 플랫폼 복스(Vox)는 이러한 흐름을 보여주는 사례로, 합성 음성 통화 탐지와 신원 확인, 녹음과 텍스트 변환 기능을 하나로 묶었다. 특히 새 버전에는 합성 음성을 실시간으로 탐지하는 기술과, 목소리의 고유한 특징을 담은 음성 지문으로 등록된 발신자를 자동 식별하는 기술이 함께 추가되었다. 누가 말하는지를 확인하는 일과 그 목소리가 실제 사람의 것인지를 가리는 일이 한 시스템 안에서 동시에 이루어지기 시작한 셈이다.

미국의 음성 보안 기업 핀드롭(Pindrop)은 기계 발신 통화를 약 2초 만에 판별하는 봇스토퍼(Bot Stopper)를 내놓았으며, 이 도구는 5,000개 이상의 AI 음성을 등록한 데이터베이스를 활용해 어떤 AI 음성인지까지 구분한다. 한 의료 기관에서는 한 달 동안 6,668건에 달하던 기계 발신 통화가 이 도구 도입 4개월 만에 94.3% 줄어든 것으로 나타났다. 익숙한 목소리라도 AI가 정교하게 흉내 낼 수 있게 되면서, 이제는 음성 통화만으로는 상대가 정말 본인인지 믿기 어려운 상황에 이르렀다. 그러나 두 판정을 한 시스템에 함께 담는 것과, 합성 음성의 가능성까지 고려해 하나의 일관된 신원 판단을 내리는 것은 별개의 문제로 남아 있다.

* 사회공학 공격(social engineering attack) : 기술적 취약점 대신 사람의 신뢰나 심리적 허점을 이용해 상대를 속이고, 정보나 금전, 접근 권한을 탈취하는 공격 방식

• 화자 검증과 위조 탐지를 하나로 묶는 접근의 등장

음성 합성과 음성 복제 기술이 발전하면서, 실제 사람의 목소리와 구별하기 어려운 합성 음성을 만드는 데 드는 비용이 낮아졌다. 이는 목소리로 본인을 확인하는 음성 인증과 원격 신원 확인 서비스에 직접적인 위협이 되며, 공격자는 자연스러우면서도 특정인의 목소리로 들리는 합성 음성을 만들어 그 사람 행세를 할 수 있다. 지금까지의 대응은 주로 들어온 음성이 합성 음성인지 실제 음성인지를 가리는 위조 탐지에 집중되어 왔다. 그러나 위조 탐지 모델은 학습 과정에서 접하지 못한 합성 기법이나 녹음 환경, 화자, 후처리 방식이 등장하면 성능이 떨어진다는 한계를 안고 있다.

이 한계를 보완하는 방법으로 주목받는 것이 위조 인지 화자 검증(spoofing-aware speaker verification, 이하 SASV)이다. 이는 미리 등록해 둔 등록 음성과 새로 들어온 검증 음성을 비교해 본인의 실제 음성만 받아들이고, 다른 사람의 음성과 합성 음성은 거부하는 방식이다. 다만 기존 연구는 화자 검증(speaker verification) 모델과 위조 탐지 모델을 따로 만든 뒤 점수를 사후에 합치는 방식에 머물러 왔다. 최근에는 소리를 듣고 지시에 따라 답을 문장으로 내놓는 대규모 오디오 언어 모델(large audio language model, 이하 LALM)이 등장해, 두 판정을 하나의 모델에서 다룰 가능성이 열리고 있다. 본 보고서는 LALM을 SASV에 체계적으로 적용해, 판정 성능과 판정 근거 제시의 가능성, 남은 과제를 평가했다.

II. 본론: 대규모 오디오 언어 모델의 화자 검증 원리

• 등록 음성과 검증 음성을 비교하는 세 갈래 판정

SASV는 신원을 주장하는 사람의 등록 음성과 확인 대상인 검증 음성 한 쌍을 입력받아, 세 가지 결과 중 하나로 답하는 문제로 정의된다. 쉽게 말해 검증 음성이 본인의 실제 음성인지, 다른 사람의 음성인지, 아니면 누구의 목소리를 흉내 냈든 기계가 만든 합성 음성인지를 한 번에 가리는 것이다. 판정은 등록 음성을 미리 받아 두었다가 검증 음성이 들어오면, 두 음성의 특징을 나란히 비교해 같은 사람인지와 실제

음성인지를 함께 따지는 순서로 진행된다. 이때 합성 음성은 화자가 누구인지와 상관없이 거부 대상으로 묶이므로, 특정인의 목소리를 무단 복제한 합성 음성이라도 본인의 실제 음성으로 통과할 수 없다.

기존 SASV 시스템은 대부분 화자 검증 모델과 위조 탐지 모델을 각각 학습시킨 뒤, 점수를 사후에 합쳐 최종 판정을 내리는 방식을 택해 왔다. 점수 결합은 화자 검증 모델이 합성 음성에 속는 정도를 줄이는 데 효과를 보였고, 이후에는 두 모델의 특징을 합치거나 판정을 여러 단계로 나누는 등 결합 방식이 점차 정교해졌다. 그러나 이 방식에서는 최종 성능이 두 개별 모델의 성능에 좌우되며, 시스템은 점수만 내놓을 뿐 판정의 근거는 드러내지 않는다. 이와 달리 LALM을 활용한 통합 방식은 두 목표를 따로 학습시킨 뒤 합치지 않고, 하나의 모델이 등록 음성과 검증 음성을 함께 듣고 판정 전체를 직접 학습한다. 이때 모델이 학습하는 것은 검증 음성을 본인의 실제 음성, 타인의 음성, 합성 음성 중 하나로 가르는 세 갈래 판정이다.

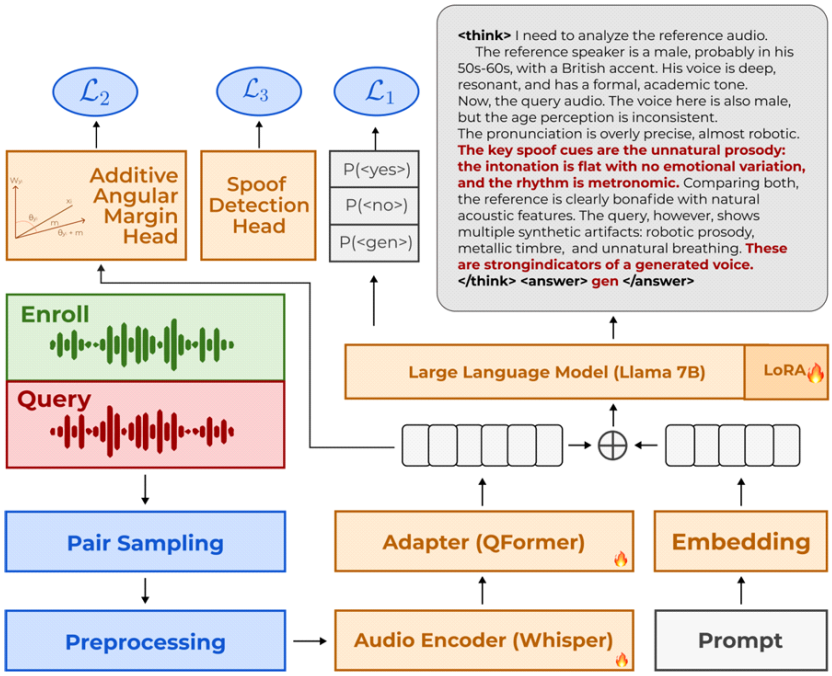
LALM은 음성을 비롯한 소리를 입력받아 이해하고, 사용자의 지시에 따라 그 결과를 문장으로 출력하는 인공지능 모델이다. 즉 소리를 듣고 질문에 답하듯 판단 결과와 그 이유를 글로 설명할 수 있다는 점에서, 점수만 내놓는 기존 모델과 차이가 있다. 판정 이유를 문장으로 남길 수 있다는 특성은, 판정 결과를 사후에 검토하고 오류를 따져야 하는 신원 확인 분야에서 LALM을 활용할 근거가 된다. 다만 별도의 조정 없이 화자 검증이나 위조 탐지에 쓰인 LALM은 낮은 성능을 보였고, 과제에 맞춘 추가 학습을 거친 뒤에도 기존 모델을 일관되게 앞서지는 못했다.

• 저랭크 적응을 통한 오디오 언어 모델의 과제 정렬

SASV에 적용된 두 LALM은 원래 음성을 글로 옮기거나 소리의 내용을 설명하도록 학습된 모델로, 목소리의 주인과 진위를 가리는 일에는 맞춰져 있지 않다. 이 간극을 메우는 방법이 저랭크 적응(low-rank adaptation, 이하 LoRA)으로, 모델 전체를 다시 학습시키지 않고 일부에 소량의 학습 요소를 더해 그 부분만 새로 학습시키는 방식이다. 다시 말해 모델이 이미 익힌 소리에 대한 지식은 그대로 두고, 새 과제에 필요한 부분만 적은 비용으로 고쳐 쓰는 방법이다. 두 모델은 음성을 받아들이는 방식에서 차이를 보이는데, 살몬(SALMONN)은 등록 음성과 검증 음성을 하나로 이어 붙여 입력받고, 큐윈 오디오(Qwen-Audio)는 두 음성을 따로 입력받는다.

조정을 거치기 전 두 모델의 정확도는 세 가지 결과 중 하나를 무작위로 고르는 것과 크게 다르지 않은 수준에 머물렀다. 반면 LoRA를 적용한 뒤에는 두 모델 모두 정확도가 높아져, 열 번 중 여덟 번 이상 올바른 판정을 내리는 수준에 이르렀다. 이는 일반적인 소리 학습만으로는 SASV 능력이 저절로 생기지 않지만, 과제에 맞춘 조정을 거치면 LALM도 실제 활용을 검토할 만한 수준에 이를 수 있음을 의미한다. 학습 과정에서는 두 음성에 서로 다른 잡음과 울림을 섞어, 녹음 환경이 달라도 같은 사람인지와 합성 여부를 가려내도록 했다. 이러한 학습 방식은 목소리를 무단 복제한 합성 음성이 유통 과정에서 음질 변화를 거치더라도, 판정이 흔들리지 않도록 하는 장치로 해석된다.

[그림 1] 살몬 기반 SASV 모델의 판정 구조



출처: Sofya Savelyeva 외 5명, "Large Audio Language Models for Spoofing-Aware Speaker Verification", arXiv, 2026.07.16., https://arxiv.org/pdf/2607.14753

• **화자 구별과 위조 탐지의 균형을 맞추는 학습 전략**

SASV는 본인과 다른 사람을 가려내는 화자 구별과, 실제 음성과 합성 음성을 가려내는 위조 탐지라는 두 목표를 동시에 만족해야 한다. 그런데 세 갈래 판정만 그대로 학습시키면 위조 탐지는 높은 정확도를 보이는 반면, 화자 구별은 상대적으로 뒤처지는 불균형이 나타난다. 이를 보완하기 위해 세 갈래 판정 외에, 화자 구별과 위조 탐지를 각각 따로 학습시키는 두 개의 보조 판정 장치가 추가되었다. 화자 구별 쪽에는 가산 각도 마진(additive angular margin) 학습이 쓰이는데, 이는 같은 사람의 목소리 특징은 서로 가깝게 모으고 다른 사람의 목소리 특징은 일정 간격 이상 떨어뜨리도록 학습시키는 방식이다. 결과적으로 목소리가 비슷한 두 사람이라도, 모델이 목소리 특징을 비교할 때는 서로 분명히 구분되는 대상으로 다루게 된다.

가산 각도 마진 학습을 더하자 화자 구별 정확도는 높아졌지만, 위조 탐지 정확도는 이전의 4분의 1 수준 아래로 떨어지는 상충 관계가 드러났다. 이에 가산 각도 마진 학습은 실제 음성에만 적용되는데, 합성 음성까지 사람별로 묶으면 합성 음성이 해당 화자의 목소리 특징 쪽으로 끌려가 위조 탐지가 더 어려워지기 때문이다. 여기에 실제 음성과 합성 음성만을 구분하는 위조 탐지 전용 판정 장치를 더하자, 화자 구별 정확도를 처음보다 높은 수준으로 유지하면서 떨어졌던 위조 탐지 성능도 상당 부분 회복되었다. 두 보조 장치가 각자 맡은 신호에 집중하도록 역할을 나눈 것이 두 목표의 균형을 되찾은 요인으로 볼 수 있다.

마지막으로 판정이 까다로운 사례의 학습 비중을 높이는 방식이 더해졌는데, 대상은 목소리가 달라 보이는 같은 사람, 목소리가 비슷한 다른 사람, 그리고 등록 화자와 가장 닮은 합성 음성이다. 특히 등록 화자와 가장 닮은 합성 음성을 반복해 학습시키는 과정은, 특정인의 목소리를 정교하게 복제한 합성 음성을 가려내는 능력과 직결된다. 이 단계까지 거친 모델은 약 89%의 정확도로, 같은 조건에서 비교한 기존 점수 결합 시스템보다 높은 정확도와 낮은 오류 비용(detection cost function)*을 기록했다. 다만 동일 오류율(equal error rate)** 지표에서는 기존 시스템이 여전히 앞섰는데, 이는 LALM이 판정 결과를 답하도록 학습되었을 뿐 판정의 확신 정도를 정밀한 점수로 나타내도록 학습되지는 않았기 때문으로 해석된다.

* 오류 비용(detection cost function): 본인을 잘못 거부하는 경우와 다른 사람이나 합성 음성을 잘못 받아들이는 경우에 각각 비용을 매겨 합산한 평가 지표로, 값이 낮을수록 실제 서비스에서 발생하는 손실이 적다는 의미

** 동일 오류율(equal error rate): 본인을 잘못 거부하는 비율과 다른 사람이나 합성 음성을 잘못 받아들이는 비율이 같아지는 지점의 오류율로, 값이 낮을수록 두 종류의 오류를 고르게 줄였다는 의미

• 판정 근거를 설명하는 추론 학습과 그 한계

추론 학습은 모델이 최종 판정을 내리기 전에, 그 판정에 이른 이유를 문장으로 먼저 서술하도록 학습시키는 방식이다. 쉽게 말해 같은 사람이라거나 합성 음성이라는 답만 내놓지 않고, 목소리의 높낮이나 말의 리듬처럼 판단의 근거가 된 소리의 특징을 함께 답하게 하는 것이다. 학습에 쓰인 추론 문장은 더 큰 추론 모델이 자동으로 작성했으며, 같은 음성 쌍에 대해 세 번 작성한 결과 중 두 번 이상 판정이 맞고 근거가 서로 일치하는 것만 남겼다. 여기에 화자의 나이와 성별, 감정, 말의 높낮이 같은 특징을 별도 분석 도구로 추출해 더함으로써, 추론 문장이 실제 소리에서 확인할 수 있는 근거를 담도록 했다. 이후에는 정답을 맞히고 정해진 형식을 지킬 때 보상을 주는 강화학습을 더해, 화자 구별과 위조 탐지 사이의 판정 균형을 다듬었다.

같은 양의 데이터로 비교하면, 추론 학습을 거친 모델은 판정만 학습한 모델보다 전체 정확도가 다소 낮았지만 기존 점수 결합 시스템보다는 높은 정확도를 보였다. 다만 LALM은 역시 환각(hallucination) 현상을 보이기 쉬워, 근거를 따라 답에 이르기보다 먼저 짐작한 답에 근거를 끼워 맞추기도 한다. 이 때문에 신뢰할 만한 추론 데이터를 충분히 확보하기 어려웠고, 생성된 근거의 타당성에 대한 사람의 검증과 기존 시스템보다 많은 연산 자원이 필요하다는 점도 과제로 남아 있다. 그럼에도 판정마다 근거가 문장으로 남는다는 점은, 특정인의 목소리를 복제한 합성 음성을 두고 판정이 왜 그렇게 내려졌는지를 사후에 검토할 단서를 제공한다는 데 의의가 있다.

III. 결론: AI 생성 인물 판별 기술의 확산 방향과 과제

• 판별 결과의 설명 가능성과 남은 과제

결국 AI 생성 인물을 둘러싼 표시 체계는 아티스트의 자진 공개와 플랫폼의 검토만으로는 넓은 범위를 감당하기 어렵다. 다중 에이전트 포렌식 추론 프레임워크는 영상 자체에 남은 질감, 조명, 움직임, 물리적 타당성의 흔적을 나눠 읽어 이러한 한계를 보완하는 방향을 제시한다. 이를 통해 합성 여부의 판단이 학습해 둔 기법에 갇히지 않고, 처음 등장한 생성 방식에도 이어질 수 있으며, 새로운 합성 기법이 나올 때마다 표시 체계가 뒤따라야 하는 부담이 줄어들 수 있다.

따라서 앞으로의 과제는 이 기술의 정확도와 적용 범위를 함께 넓히고, 판별 결과를 어떤 절차에 연결할지를 정하는 데 있다. 분포 외 평가에서 기록된 정확도는 70%에 미치지 못해, 진짜 영상을 합성물로 잘못 판단하거나 그 반대의 오판이 일어날 여지가 남아 있다. 기술적으로는 새로운 생성기가 등장할 때마다 관찰 기준을 갱신하는 절차, 설명의 정확성을 함께 평가하는 방법을 마련할 필요가 있다. 산업적으로는 플랫폼의 표시 체계, 창작자의 자진 공개, 기술적 판별 결과를 어떻게 맞물리게 할지가 과제로 떠오른다. 판별 결과는 침해를 단정하는 근거가 아니라 검토를 돕는 단서로 다루는 신중함이 함께 요구되며, 기술과 제도가 보조를 맞출 때 AI 생성 인물을 둘러싼 불확실성도 점차 줄어들 것으로 전망된다.

참고문헌

• Joel R. McConvey, “Deepfakes, social engineering push identity checks into live conversations”, Biometric Update, 2026.10.09., https://www.biometricupdate.com/202610/deepfakes-social-engineering-push-identity-checks-into-live-conversations
• Sofya Savelyeva 외 5명, “Large Audio Language Models for Spoofing-Aware Speaker Verification”, arXiv, 2026.07.16., https://arxiv.org/pdf/2607.14753

기술용어

순번	용어	설명
1	**사회공학 공격** (social engineering attack)	기술적 취약점 대신 사람의 신뢰나 심리적 허점을 이용해 상대를 속이고, 정보나 금전, 접근 권한을 탈취하는 공격 방식
2	**오류 비용** (detection cost function)	본인을 잘못 거부하는 경우와 다른 사람이나 합성 음성을 잘못 받아들이는 경우에 각각 비용을 매겨 합산한 평가 지표로, 값이 낮을수록 실제 서비스에서 발생하는 손실이 적다는 의미
3	**동일 오류율** (equal error rate)	본인을 잘못 거부하는 비율과 다른 사람이나 합성 음성을 잘못 받아들이는 비율이 같아지는 지점의 오류율로, 값이 낮을수록 두 종류의 오류를 고르게 줄였다는 의미