



저작권 이슈 브리프

SUMMARY

산업/기업

기술

산업 F-Secure 사례로 본 멀티모달 딥페이크 탐지 기술의 소비자 보안 서비스 통합

▶ 생성형 AI 기술의 보편화로 이미지, 영상, 음성을 결합한 멀티모달 사기 콘텐츠가 확산되면서, 의심 링크나 메시지 차단 중심의 기존 소비자 보안 관리 방식이 구조적 한계에 직면하고 있다. 이에 따라 사기 탐지의 대상이 텍스트와 링크에서 AI 생성 멀티모달 콘텐츠로 확장되는 움직임이 나타나고 있다. 최근 핀란드의 사이버 보안 기업인 F-Secure와 이탈리아의 딥페이크 탐지 기술 스타트업 IdentifAI의 협업은 탐지 기술을 서비스 제공자 단계에서 직접 내재화하는 방식으로, 보안 관리의 트렌드가 변화하는 사례로 볼 수 있다. 이러한 사례는 일부 보안 서비스 영역에서의 시도를 넘어, 소비자 대상 디지털 서비스 전반에서 딥페이크 탐지 기능이 보안 서비스의 기본 구성 요소로 통합되는 조짐으로 읽힐 수 있다.

산업 오디오 핑거프린팅 기술 사례로 본 AI 음악의 학습 데이터 검증

▶ AI 음악 생성 기업이 학습 데이터의 세부 구성을 공개하지 않는 관행 속에서, 권리자가 자사 음원의 학습 활용 여부를 사전에 확인하기 어려운 구조가 지속되어 왔다. 미국음반산업협회가 2024년 6월 수노를 상대로 제기한 저작권 침해 소송에서, 원고 측 전문가는 2025년 11월부터 약 2주간 수노 학습 데이터의 핑거프린트를 오디오 매직의 데이터베이스와 대조했다. 권리자 측은 검증 결과를 근거로 소송 대상을 확대하는 수정 소장 제출을 신청했다. 본 사례는 오디오 핑거프린팅이 비공개 학습 데이터에 대한 사후 검증 수단으로 적용되며, 기술의 활용 범위가 분쟁 과정의 증거 확보 영역으로 확대되는 양상을 보여준다.

산업 AI 생성 음악 확산에 따른 음원 관리 체계의 변화

▶ 생성형 AI의 확산으로 AI 생성 음악이 급증하면서 부정 스트리밍에 따른 저작권 사용료 왜곡과 학습 단계의 저작권 문제가 새로운 관리 과제로 부상하고 있다. 이에 주요 음악 스트리밍 플랫폼은 AI 생성 음악을 식별하고 표시하는 한편, 추천과 저작권 사용료 정산 방식을 차별화하는 관리체계를 도입하고 있다. 대표적으로 프랑스 음악 플랫폼 디저는 탐지, 표시, 정산을 분리한 운영 방식을 선도적으로 적용하고 있다. 향후 AI 생성 음악의 식별과 표시, 저작권 사용료 정산을 연계하는 관리체계는 플랫폼별 대응을 넘어 산업 공동의 표준과 인프라로 확대될 가능성이 제기된다.



저작권 이슈 브리프

SUMMARY

산업/기업

기술

산업 AI 생성 음원의 대량 유입에 따른 노출 제한과 정산 배제 중심의 관리 방식

▶ 2026년 월드컵을 계기로 AI 생성 음원이 단기간에 대량 유입되면서, 스트리밍 플랫폼의 탐지 및 표시 체계와 수익 정산 관리가 주요 쟁점으로 부상하고 있다. 디저는 AI 생성 여부를 업로드 단계에서 탐지해 표시하고, 해당 음원의 자동 추천과 재생목록 노출을 제한하는 방식으로 플랫폼 내 유통 범위를 관리하고 있다. 또한 로열티 수취를 목적으로 재생 수를 인위적으로 부풀린 부정 재생을 정산 대상에서 제외함으로써, AI 생성 음원의 대량 등록 유인을 낮추는 역할을 한다. 이러한 흐름은 애플 뮤직 등 타 플랫폼의 표시 기준 마련으로도 이어지고 있으며, 향후 쟁점의 범위는 AI 생성 여부 표시를 넘어 표시된 AI 생성 음원을 추천 노출, 수익 정산, 권리 관리 체계에서 어떻게 처리할 것인지로 확대될 전망이다.

산업 AI 합성 콘텐츠 대응을 위한 외부 위협 방어 플랫폼의 탐지 기능 통합

▶ AI를 활용해 기업 임원이나 브랜드를 사칭하는 딥페이크 공격이 피싱, 투자 사기, 계정 탈취 등 후속 위협으로 이어지면서, 기업 보안 대응은 콘텐츠 유포 차단을 넘어 AI 합성 여부를 신속히 검증하는 방향으로 확대되고 있다. 이에 제로폭스와 리얼리티 디펜더는 2026년 6월 2일 딥페이크 탐지 기능을 외부 위협 방어 플랫폼에 통합하는 협력을 발표하였다. 이번 통합은 사칭 계정이나 악성 도메인 탐지와 함께 음성, 영상, 이미지의 합성 여부를 자동 판정해, 위협 발견부터 검증 및 차단까지 단일 대응 절차로 처리하도록 한 점이 핵심이다. 다만 이는 차단 판단의 근거를 보강하는 검증 계층이며, AI 합성 콘텐츠 확산에 따른 신뢰 문제를 단독으로 해소하는 수단은 아니라는 점에 유의할 필요가 있다.

기술 주간 기술 동향

▶ 생성형 AI의 음성 복제 기술이 고도화되면서 짧은 음성만으로도 특정인의 목소리를 정교하게 모방할 수 있게 되었고, 금융 사기와 부정 결제에 악용될 위험도 커지고 있다. 기존 음성 워터마킹은 파형이나 스펙트로그램에 신호를 삽입해 일반적인 압축과 잡음에는 견디지만, 신경 코덱이나 보코더가 음성을 재합성하면 워터마크가 약화되는 한계가 있다. DuraMark는 이러한 문제를 해결하기 위해 음질 지속시간의 짝수/홀수 상태에 비트를 대응시켜 정보 수준에서 워터마크를 삽입한다. 지속시간 제어형 LLM 기반 TTS 모델과 지속시간 추출기를 결합해 삽입과 탐지를 수행하며, 재합성 공격에서도 기존 신호 수준 방식보다 안정적인 성능을 보인다. 다만 현재 검증 범위가 중국어와 일정 길이 이상의 발화에 집중되어 있어 다른 언어와 짧은 음성에 대한 추가 검증이 필요하다.



저작권 이슈 브리프

SUMMARY

산업/기업

기술

F-Secure 사례로 본 멀티모달 딥페이크 탐지 기술의 소비자 보안 서비스 통합

다양해지는 딥페이크 사기 수단과 기존 보안 방식의 한계

• 멀티모달 딥페이크 확산에 따른 소비자 피해 양상

- 생성형 AI 기술의 보편화로 복제 음성, AI 합성 영상, 딥페이크(deepfake) 이미지 등 멀티모달(multi-modal)* 사기 수단이 다양화되면서, 기존의 단순한 피싱 시도보다 훨씬 정교하고 개인화된 형태의 사기가 가능해지고 있음
- 미국 FBI의 조사에 따르면 미국 내 AI 기반 사기로 인한 피해액은 2025년 기준 약 9억 달러 (원화 약 1조 3,530억 원)¹⁾에 달하는 것으로 추정되며, 이는 딥페이크 악용 사기가 정부와 산업계 차원의 대응이 필요한 수준에 이르렀음을 보여줌
- 핀란드의 사이버 보안 기업 F-Secure의 자체 조사에 따르면 전체 소비자의 56% 이상이 매월 사기 시도의 표적이 되고 있으며, 금전적 피해 규모는 불과 지난 1년 만에 2배 증가한 것으로 나타남²⁾

* 멀티모달 (Multimodal): 이미지, 영상, 음성 등 두 가지 이상의 서로 다른 형태의 데이터를 동시에 처리하거나 분석하는 방식

• 링크·메시지 차단 중심 보안 구조의 대응 공백

- 기존 소비자 보안 서비스는 의심스러운 링크나 메시지를 탐지하고 차단하는 방식을 중심으로 설계되어 왔으며, AI로 생성된 이미지, 영상, 음성 콘텐츠를 분석하고 검증하는 기능은 포함하지 않는 구조임
- 사기 콘텐츠의 형태가 텍스트 기반 피싱에서 음성 복제 및 AI 합성 영상으로 이동하면서, 기존 보안 체계의 탐지 대상과 실제 위협의 형태 사이에 구조적 간극이 발생하고 있음
- 이에 따라 서비스 제공자들이 딥페이크 탐지 기능을 기존 보안 서비스에 통합하거나 추가해야 한다는 요구가 나타나고 있으며, 실시간 콘텐츠 검증이 보안 서비스의 핵심 기능으로 부상함

실시간 멀티모달 탐지 기술의 구조와 서비스 내재화 방식

• IdentifAI 멀티모달 탐지 기술의 작동 방식과 탐지 전용 접근의 특징

- 이탈리아의 딥페이크 탐지 전문 스타트업 IdentifAI는 이미지, 영상, 음성을 실시간으로 분석해 AI가 생성하거나 조작한 콘텐츠를 식별하는 멀티모달 탐지 기술을 개발하고 있음
- IdentifAI는 2025년 7월 시리즈 A 투자 라운드에서 500만 유로(원화 약 88억 원)³⁾를 조달했으며, 이는 딥페이크 탐지 전문 기업에 대한 시장의 관심이 높아지고 있음을 보여줌

1) 1달러=1,503.2원(KEB 하나은행 매매기준율 적용, 2026.06.01.)

2) F-Secure, "F-Secure and IdentifAI Take Aim at AI-Generated Scams with Deepfake Protection Launch", 2026.06.16., <https://www.f-secure.com/en/partners/newsroom/f-secure-and-identifai-launch-deepfake-protection>

3) 1유로=1,752.05원(KEB 하나은행 매매기준율 적용, 2026.06.01.)

- IdentifAI는 AI 콘텐츠 탐지에만 집중하며 생성형 모델을 자체적으로 개발하지 않는 원칙을 취하고 있어, 탐지와 생성을 함께 다루는 사업자에게 발생할 수 있는 이해충돌 구조를 배제한 접근 방식을 택하고 있음

• F-Secure 내장형 보안 서비스 공급 체계를 통한 서비스 제공자 단계 통합 구조

- F-Secure와 IdentifAI는 2026년 6월 협업을 공식화했으며, IdentifAI가 개발한 멀티모달 탐지 기술을 F-Secure의 사기 방지 통합 솔루션(Scam Protection Suite)에 딥페이크 탐지 기능으로 추가하는 방식으로 통합을 완료함
- F-Secure의 내장형 보안 서비스 공급 체계는 200개 이상의 서비스 제공자 파트너를 통해 공급되며, 파트너사는 별도 인프라 구축 없이 딥페이크 탐지 기능을 자사 서비스에 통합할 수 있음
- 이 방식은 딥페이크 탐지를 기업 전용 전문 도구로 제공하는 기존 방식과 달리, 소비자가 일상적으로 이용하는 보안 서비스 안에 해당 기능을 직접 내재화한다는 점에서 공급 경로의 변화를 보여줌

• 딥페이크 탐지만을 전문으로 하는 기업 모델의 구조적 특성과 생성 모델 고도화에 따른 대응 과제

- 딥페이크 탐지만을 전문으로 하는 기업은 시장에서 극소수에 불과하며, 대부분의 사업자는 탐지 기능을 포괄적인 통합 보안 솔루션의 일부로 제공하는 방식을 취하고 있음
- 탐지 전용 접근은 새로운 AI 생성 모델이 시장에 등장할 때마다 탐지 기술도 함께 지속적으로 업데이트되어야 하는 과제를 안고 있으며, 생성 모델의 고도화 속도가 탐지 기술의 갱신 속도를 앞설 경우 탐지 공백이 발생할 수 있음
- IdentifAI는 탐지 기능만을 전문으로 하는 구조가 새로운 생성 모델에 대응하는 속도를 유지하는 데 유리하다는 입장이나, 실제 탐지 정확도와 오탐률에 대한 구체적인 수치는 공개되지 않은 상태임

소비자 보안 서비스에서 딥페이크 탐지 기능 통합의 확산 가능성과 과제

• 기업 전용 탐지 기술의 소비자 보안 서비스 통합 가능성

- F-Secure와 IdentifAI의 협업은 기업 환경을 중심으로 제공되던 딥페이크 탐지 기술이 소비자 대상 보안 서비스에 통합된 사례로, 탐지 기술의 적용 범위가 일반 소비자 영역으로 확장됨을 시사함
- 딥페이크 탐지를 별도 전문 도구가 아닌 기존 보안 서비스의 구성 요소로 통합하는 방식은, 소비자가 별도의 설치나 전환 없이 탐지 기능을 이용할 수 있다는 점에서 실질적인 접근성을 높이는 구조임

• 기술적 탐지와 법적·제도적 대응 체계의 병행 필요성

- AI 생성 콘텐츠를 기술적으로 탐지하는 것과 해당 콘텐츠를 이용한 사기 행위에 법적 책임을 묻는 것은 별개의 절차이며, 탐지 기술의 도입만으로 딥페이크 악용 사기에 따른 법적·제도적 문제가 해소되는 것은 아님
- 탐지 결과의 오탐이나 미탐이 발생했을 때 서비스 제공자와 탐지 기술 공급자 중 책임 분배 기준이 아직 명확히 정립되어 있지 않아, 제도적 논의가 필요한 상황임



참고문헌

- F-Secure, “F-Secure and IdentifAI Take Aim at AI-Generated Scams with Deepfake Protection Launch”, 2026.06.16., <https://www.f-secure.com/en/partners/newsroom/f-secure-and-identifai-launch-deepfake-protection>
- Joel McConvey, “IdentifAI locks in on deepfakes with F-Secure integration”, BioMetric, 2026.06.18., <https://www.biometricupdate.com/202606/identifai-locks-in-on-deepfakes-with-f-secure-integration>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

오디오 핑거프린팅 기술 사례로 본 AI 음악의 학습 데이터 검증

AI 음악 학습 데이터 비공개 구조와 권리 침해 입증의 한계

• AI 학습 데이터 비공개 관행과 권리자의 확인 부담

- AI 음악 생성 기업이 학습 데이터의 세부 구성을 외부에 공개하지 않는 관행을 유지함에 따라, 권리자는 자사 음원의 학습 활용 여부를 사전에 확인할 수 없고 침해 입증 부담이 권리자 측에 집중되는 구조가 형성됨
- 기존 분쟁에서는 권리자가 AI 산출물과 자사 음원의 유사도에 기반해 침해 가능성을 제기해 왔으나, 해당 방식은 학습 데이터 내 실제 포함 여부를 외부에서 확인할 수단이 되지 못하는 한계가 존재함

• 소송을 통해 드러난 학습 데이터 검증 쟁점

- 미국음반산업협회(Recording Industry Association of America, 이하 RIAA)는 2024년 6월 매사추세츠 연방법원(U.S. District Court for the District of Massachusetts)에 AI 음악 생성 기업 수노(Suno)를 상대로 저작권 침해 소송을 제기했으며¹⁾, 이는 글로벌 메이저 음반사 3곳을 대리한 산업 단체 차원의 대응에 해당함
- 초기 쟁점은 권리자 측이 특정 프롬프트를 활용해 수노 출력물에서 자사 음원과 유사한 결과물을 식별한 560건의 저작물에 한정되어 있었으며, 이는 학습 데이터에 대한 검증을 거치지 않은 일부 표본에 해당함
- 수노가 학습 데이터의 구체적 구성을 공개하지 않으면서 권리자 측의 침해 입증 범위가 제한되었고, 이에 따라 법적 공방의 쟁점은 개별 출력물의 유사성 판단에서 학습 데이터 전체의 복제 규모 및 위법성 입증으로 확대됨

[사례] 오디오 핑거프린팅을 활용한 학습 데이터 검증

• 학습 데이터 핑거프린트 생성 절차

- 원고 측이 지정한 전문가는 2025년 11월 3일부터 수노의 외부 법률 사무소 내 보안 구역에서 학습 데이터에 접근했으며, 약 2주간 학습 데이터 내 오디오 파일의 디지털 핑거프린트(fingerprint)*를 생성함
- 생성된 핑거프린트는 오디오 매직(Audible Magic)**의 콘텐츠 인식 기술에 기반한 레퍼런스 데이터베이스와 대조되었으며, 이는 학습 데이터 원본을 외부로 반출하지 않은 상태에서 저작권 음원의 포함 여부를 식별하는 절차에 해당함

* 핑거프린트(fingerprint): 콘텐츠 식별을 위해 음원·영상·이미지의 고유한 특징을 추출해 만든 디지털 식별값

** 오디오 매직(Audible Magic): 음원·영상 콘텐츠 식별과 권리 관리를 지원하는 미국의 디지털 저작권 기술 기업

1) Richard L. Wells, "AI Music Copyright Lawsuit: Suno Discovery Shows Millions of Songs, July Ruling Nears",

TechTimes, 2026.06.16., <https://www.techtimes.com/articles/318471/20260616/ai-music-copyright-lawsuit-suno-discovery-shows-millions-songs-july-ruling-nears.htm>

[표1] 수노 학습 데이터 검증 절차 일정

단계	시점	주요 내용
접근 시작	2025.11.03	수노 외부 법률 사무소 보안 구역에서 학습 데이터 접근 개시
핑거프린트 생성	약 2주간	학습 데이터 내 전체 오디오 파일의 디지털 핑거프린트 생성
매칭 분석 완료	2026.01.02	오디블 매직 레퍼런스 데이터베이스와의 대조 분석 완료
결과 전달	2026.01.15	권리자 측에 최종 매칭 결과 전달
수정 소장 신청	2026.05.21	소송 대상을 확대하는 수정 소장 제출 신청

출처: 참고문헌 종합하여 재구성

• 오디오 핑거프린팅의 식별 원리와 기술적 의미

- 오디오 핑거프린팅은 음원의 파형, 주파수 분포, 시간대별 음향 특징 등에서 고유한 신호 패턴을 추출해 디지털 식별자로 변환하는 기술로, 변환된 식별자는 기존 음원 데이터베이스와의 매칭 기준으로 활용됨
- 피치* 변형이나 템포 조정과 같은 변형이 가해진 음원에 대해서도 일정 범위 내에서 식별이 가능하므로, 단순한 파일의 동일성 확인을 넘어 변형·재가공된 음원의 포함 여부까지 검증할 수 있는 특성을 보유함
- 해당 기술은 유통 단계의 저작권 관리 및 음원 식별에 주로 활용되었으나, 이번 사례에서는 학습 데이터 내 저작권 음원 포함 여부를 검증하는 용도로 적용되며 활용 범위가 확대됨

* 피치(pitch): 사람이 인식하는 소리의 높낮이로, 음성이나 음악에서 보컬의 음높이, 멜로디, 음색 특성을 구분하는 데 활용되는 요소

• 검증 결과의 활용과 학습 데이터 규모 공개 쟁점

- 권리자 측인 RIAA는 검증 결과를 근거로 2026년 5월 소송 대상을 61,026건으로 확대하는 수정 소장 제출을 신청했으며, 해당 건수가 전체 매칭 결과의 일부에 해당함을 명시함²⁾
- 반면 수노는 학습 데이터에 포함된 전체 오디오 파일 수가 공개될 경우 경쟁사가 모델 학습 규모를 추정해 상업적 피해가 발생할 수 있다고 주장하며, 해당 수치의 비공개를 법원에 신청함³⁾
- 이에 권리자 측은 학습 데이터 규모가 미국 저작권법상 공정이용(fair use)* 판단의 핵심 요소인 복제의 본질과 범위에 직결된다고 반박했으며, 법원의 공개 결정 여부가 소송의 주요 쟁점으로 부각됨

* 공정이용(fair use): 미국 저작권법 제107조에 규정된 원칙으로, 비평·논평·교육·변형적 창작 등 일정한 목적에 한해 저작권자의 허락 없이 저작물을 제한적으로 사용할 수 있도록 허용함

시사점: 학습 데이터 분쟁에서 기술 검증의 역할 확대

• 기술의 사후 검증 및 증거 확보 수단화

- 이번 사례는 오디오 핑거프린팅이 비공개 학습 데이터에 대한 사후 검증 수단으로 작용하였으며, 기술의 활용 범위가 유통 단계의 저작권 관리에서 법적 단계의 증거 확보 영역으로 확대되는 양상을 보임

2) Richard L. Wells, "AI Music Copyright Lawsuit: Suno Discovery Shows Millions of Songs, July Ruling Nears",

TechTimes, 2026.06.16., <https://www.techtimes.com/articles/318471/20260616/ai-music-copyright-lawsuit-suno-discovery-shows-millions-songs-july-ruling-nears.htm>

3) Murray Stassen, "UMG and Sony urge court to reject Suno's bid to seal the size of its AI training data, arguing the public has a right to see the scale of its copying", Music Business Worldwide, 2026.06.09., <https://www.musicbusinessworldwide.com/umg-and-sony-urge-court-to-reject-sunos-bid-to-seal-the-size-of-its-ai-training-data-arguing-the-public-has-a-right-to-see-the-scale-of-its-copying/>

- 또한 권리자 측은 학습 데이터 원본을 외부에서 직접 확인하지 못하는 구조에서도 기술적 매칭 절차를 통해 침해 사실을 특정할 수 있는 수단을 확보했으며, 이는 권리자 측의 입증 부담을 완화할 수 있는 가능성을 보여준 것으로 분석됨
- 향후 AI 음악 기업에는 학습 데이터의 출처, 구성 및 이용 근거를 기록, 관리할 수 있는 내부 관리체계 마련 요구가 확대될 가능성이 있으며, 분쟁 발생 시 외부 검증 절차에 어떻게 대응할 것인지가 주요 리스크 관리 과제로 부상할 수 있음

참고문헌

- Murray Stassen, "UMG and Sony urge court to reject Suno's bid to seal the size of its AI training data, arguing the public has a right to see the scale of its copying", Music Business Worldwide, 2026.06.09., <https://www.musicbusinessworldwide.com/umg-and-sony-urge-court-to-reject-sunos-bid-to-seal-the-size-of-its-ai-training-data-arguing-the-public-has-a-right-to-see-the-scale-of-its-copying/>
- Richard L. Wells, "AI Music Copyright Lawsuit: Suno Discovery Shows Millions of Songs, July Ruling Nears", TechTimes, 2026.06.16., <https://www.techtimes.com/articles/318471/20260616/ai-music-copyright-lawsuit-suno-discovery-shows-millions-songs-july-ruling-nears.htm>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

AI 생성 음악 확산에 따른 음원 관리 체계의 변화

AI 생성 음악의 확산과 음원 관리 방식의 한계

• 대량 유입과 부정 스트리밍에 따른 저작권 사용료 왜곡

- AI 음악 생성 도구가 빠르게 확산되면서 스트리밍 플랫폼에 유입되는 음원이 크게 늘어, 2026년 4월 기준 글로벌 스트리밍 플랫폼인 디저(Deezer)에서는 하루 약 75,000곡, 신규 업로드의 약 44%가 AI 생성 음악으로 집계됨¹⁾
- 이러한 AI 생성 음악의 실제 청취 비중은 전체 스트림의 3% 안팎에 그치지만 업로드의 상당수가 재생 횟수를 인위적으로 늘리는 부정 스트리밍(stream manipulation)*을 목적으로 하며, 일부 플랫폼에서는 해당 트랙 스트림의 최대 85%가 이에 해당되는 것으로 확인됨
- 스트리밍 저작권 사용료는 한정된 재원을 재생 횟수에 따라 배분하는 구조이므로 부정 스트리밍으로 발생한 재생은 실제 창작자에게 돌아갈 사용료를 감소시키며, 2026년 국제음반산업협회(International Federation of the Phonographic Industry)도 이를 산업 전반의 공동 과제로 제시함²⁾
- 또한 AI가 생성한 음악이 인간이 창작한 음악과 구분하기 어려울 만큼 정교해지면서, 청취자의 97%가 이를 구분하는데 어려움을 겪으며, 80%가 명확한 표시를 원한다는 조사 결과가 나오는 등 AI 생성 음악임을 표시해야 할 필요성이 커짐³⁾

* 부정 스트리밍(stream manipulation): 봇이나 자동화된 계정을 이용해 재생 횟수를 인위적으로 부풀려 로열티를 부정하게 수취하는 행위

• 사후 적발의 한계와 학습 및 유통의 이중 쟁점

- 기존의 음원 관리는 업로드 이후 정산 단계에서 부정행위를 가려내는 사후 대응에 의존해 온 반면, 부정 스트리밍 행위자가 AI 도구로 대량 생성한 트랙을 각각 낮은 빈도로 재생하는 방식으로 기존 탐지 기준을 회피할 수 있어 사후 관리의 한계가 드러남
- 또한 AI 생성 음악은 권리자의 동의 없이 학습된 모델을 통해 생성된 산출물일 수 있어, 유통 단계의 부정행위와 별개로 학습 단계에서의 저작권 문제가 함께 논의되고 있음
- 이에 따라 플랫폼은 학습 단계의 저작권 문제와 유통 단계의 저작권 사용료 왜곡이라는 두 가지 부담을 동시에 안게 되며, 사후 대응만으로는 이를 충분히 관리하기 어렵다는 인식이 확산됨

1) Jesper Wendel, "AI-generated tracks now represent 44% of all new uploaded music", Deezer Newsroom, 2026.04.20.,

<https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>

2) IFPI, "Global Music Report 2026: Global recorded music revenues grow 6.4% as record companies drive innovation", IFPI, 2026.03.18.,

<https://www.ifpi.org/global-music-report-2026-global-recorded-music-revenues-grow-6-4-as-record-companies-drive-innovation/>

3) Jesper Wendel, "AI-generated tracks now represent 44% of all new uploaded music", Deezer Newsroom, 2026.04.20.,

<https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>

유입 단계로 이동하는 식별 및 표시 중심 관리 구조

• AI 생성 음악에 대한 디저의 음원 관리 방식

- 디저는 2025년 6월 업계 최초로 AI 생성 음악에 별도의 라벨을 붙여 표시하기 시작했으며, 자체 탐지 기술의 식별 정확도는 99.8% 수준으로 2025년 한 해 약 1,340만 곡을 AI 생성 음악으로 식별 및 표시함⁴⁾
- 식별된 AI 생성 음악은 알고리즘 추천과 편집 목록에서 자동으로 제외되는데 이는 실제 창작자의 노출 기회와 저작권 사용료를 보호하기 위한 조치로, 음원을 삭제하지 않으면서도 추천 단계에서 분리하는 방식임
- 또한 부정 스트리밍으로 확인된 AI 생성 음악은 저작권 사용료 정산에서 제외되어 실제 창작자의 보상을 보호하는 장치로 운영됨
- 디저는 AI 생성 음악을 플랫폼에서 삭제하지 않고 라벨과 함께 유지해 이용자의 선택권은 보장하되, 추천과 정산에서만 분리하는 단계적 방식을 취하고 있으며 공급 정책 변경이나 콘텐츠 삭제는 신중히 검토 중임

• AI 생성 음악에 대한 플랫폼 간 접근 차이

- 디저와 달리 일부 플랫폼은 AI 생성 음악 탐지나 제외보다 창작자의 자율 신고와 표시에 무게를 두고 있음
- 대표적으로 스포티파이(Spotify)는 2025년 9월 음원 메타데이터에 AI 사용 여부를 표기하도록 한 산업 표준 DDEX(Digital Data Exchange)*를 채택하고 스팸 필터를 도입함
- 애플뮤직(Apple Music)은 2026년 5월 자체 탐지 기술과 함께 청취자에게 보이는 투명성 태그를 도입했으며⁵⁾, 신규 업로드의 33% 이상이 AI 생성 음악으로 파악되는 가운데 부정 업로드를 약 60% 줄였다고 밝힘⁶⁾
- 이처럼 AI 생성 음악을 식별해 표시한다는 점은 공통되나, 디저가 추천과 정산 단계에서의 관리에 적극적인 반면, 다른 플랫폼은 표시와 자율 신고를 우선하는 등 관리 방식에서 차이가 나타남

* DDEX(Digital Data Exchange): 음원의 메타데이터와 크레딧 정보를 표준화된 형식으로 주고받기 위한 국제 규격으로, AI 사용 여부 등을 통일된 방식으로 기재하도록 함

[표1] 주요 음악 스트리밍 플랫폼의 AI 생성 음악 대응 비교

플랫폼	식별 방식	이용자 표시(적용 시점)	추천 및 정산 처리
디저(Deezer)	자체 기술 탐지 (정확도 99.8%)	AI 라벨 (2025.06~)	추천·편집 목록 제외, 부정 스트림 정산 제외
스포티파이(Spotify)	자율 신고 기반 (DDEX 메타데이터)	메타데이터 내 AI 표기 (2025.09~)	스팸 필터 적용
애플뮤직(Apple Music)	자체 탐지 + 자율 태그	투명성 태그 (2026.05~)	부정 업로드 차단

출처: 참고문헌 종합하여 재구성

4) Jesper Wendel, "Deezer launches AI music detector for playlists on all major streaming platforms", Deezer Newsroom, 2026.06.11., <https://newsroom-deezer.com/2026/06/check-ai-generated-music-in-playlists-with-deezer-detector/>

5) Daniel Brooks, "AI Music Detection Stack 2026: How Spotify, Apple, Deezer, and YouTube Actually Catch AI Tracks", Chartlex, 2026.05.12., <https://www.chartlex.com/blog/business/ai-music-detection-stack-2026>

6) Rowan Davies, "Every label in the world is delivering AI": Apple Music executive says over a third of uploads are '100% AI' as it clamps down on AI fraud", TechRadar, 2026.04.27., <https://www.techradar.com/audio/apple-music/every-label-in-the-world-is-delivering-ai-apple-music-executive-says-over-a-third-of-uploads-are-100-percent-ai-as-it-clamps-down-on-ai-fraud>

시사점: 식별 및 정산 인프라의 산업 공동화 가능성

• 탐지 및 표시 기능의 산업 공동 인프라화 가능성

- 디저는 2026년 1월부터 자체 탐지 기술을 다른 스트리밍 플랫폼과 권리단체, 유통사에 라이선스 형태로 제공하기 시작했으며, 이는 개별 플랫폼의 대응이 산업 공동 인프라로 확장될 수 있음을 보여줌
- 탐지와 표시가 일부 플랫폼에만 적용되면 식별되지 않은 음원이 다른 경로로 유입될 수 있어, 산업 전반이 공유하는 공동 식별 및 표시 체계와 통일된 표시 규격을 마련할 필요성이 함께 제기됨
- 다만 탐지 기술이 특정 사업자를 중심으로 제공될 경우 식별 기준과 투명성을 둘러싼 의존 문제가 생길 수 있어, 중립적인 표준화와 검증을 뒷받침하는 거버넌스 마련이 요구됨

• 자율 신고 중심 관리의 한계와 제도적 보완 과제

- 자율 신고와 표시에 의존하는 방식은 신고가 누락되면 AI 생성 음악을 식별하기 어렵고, 기술적 탐지와 병행되지 않으면 이용자에게 충분한 정보가 제공되지 않는다는 한계가 있음
- 이에 자율 신고에만 의존하기보다 기술적 탐지를 병행하고, 표시 의무와 정산 투명성을 제도적으로 뒷받침해 이용자가 신뢰할 수 있는 표시 체계를 갖출 필요성이 커짐
- 결국 음원의 식별과 표시, 사용료 정산 분리가 개별 플랫폼의 자율 조치를 넘어 권리자 보상과 정보 공개를 연계하는 산업 공동 관리체제로 발전할 수 있을지가 향후 과제로 남음

참고문헌

- Jesper Wendel, "AI-generated tracks now represent 44% of all new uploaded music", Deezer Newsroom, 2026.04.20., <https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>
- IFPI, "Global Music Report 2026: Global recorded music revenues grow 6.4% as record companies drive innovation", IFPI, 2026.03.18., <https://www.ifpi.org/global-music-report-2026-global-recorded-music-revenues-grow-6-4-as-record-companies-drive-innovation/>
- Jesper Wendel, "Deezer launches AI music detector for playlists on all major streaming platforms", Deezer Newsroom, 2026.06.11., <https://newsroom-deezer.com/2026/06/check-ai-generated-music-in-playlists-with-deezer-detector/>
- Daniel Brooks, "AI Music Detection Stack 2026: How Spotify, Apple, Deezer, and YouTube Actually Catch AI Tracks", Chartlex, 2026.05.12., <https://www.chartlex.com/blog/business/ai-music-detection-stack-2026>
- Rowan Davies, "'Every label in the world is delivering AI': Apple Music executive says over a third of uploads are '100% AI' as it clamps down on AI fraud", TechRadar, 2026.04.27., <https://www.techradar.com/audio/apple-music/every-label-in-the-world-is-delivering-ai-apple-music-executive-says-over-a-third-of-uploads-are-100-percent-ai-as-it-clamps-down-on-ai-fraud>



AI 생성 음원의 대량 유입에 따른 노출 제한과 정산 배제 중심의 관리 방식

AI 생성 음원의 대량 유입과 그에 따른 창작자 수익 영향 우려

• 글로벌 스포츠 이벤트를 계기로 한 AI 생성 음원 유입 확대

- 2026년 월드컵 개막 전부터 대회를 소재로 한 응원가와 앨범이 스트리밍 플랫폼에서 빠르게 늘어난 것으로 나타남
- 프랑스 스트리밍 서비스 디저(Deezer)에 따르면 'World Cup 2026'을 제목으로 단 음원은 270곡 이상 등록됐으며, 이 중 70% 이상이 AI 생성 음원으로 확인됨¹⁾
- 특히 글로벌 스포츠 이벤트에서는 검색과 재생 수요가 단기간에 집중되는 만큼, 이를 겨냥한 음원 제작과 등록이 늘어나기 쉬운 환경이 형성됨
- 생성형 AI는 이러한 이벤트성 수요에 맞춰 음원을 빠르게 제작하고 업로드할 수 있게 하면서 관련 음원의 대량 유입을 촉진한 것으로 풀이됨

[표1] 월드컵 음원 명칭별 및 국가별 AI 생성 음원 탐지 비율

음원 명칭	등록 곡 수	AI 생성 탐지 비율
World Cup 2026	270곡 이상	70% 이상
FIFA World Cup 2026	150곡 이상	65% 이상
Coupe du Monde 2026 (프랑스)	70곡 이상	86%
Copa do Mundo 2026 (브라질)	180곡 이상	71%

출처: Jesper Wendel, "Unofficial 2026 World Cup anthems flood music streaming - 70% are tagged as AI on Deezer", Deezer Newsroom, 2026.06.08., <https://newsroom-deezer.com/2026/06/world-cup-2026-ai-generated-music-deezer/>

• AI 생성 음원 유입에 따른 부정 재생과 창작자 수익 영향 우려

- 이처럼 AI 생성 음원이 대량으로 유입되고 있으나, 실제 재생 비중은 전체 재생의 1~3% 수준에 그쳐 전체 스트리밍 이용에서 차지하는 비중은 아직 제한적임
- 그러나 AI 생성 음원 재생의 최대 85%가 로열티 수취를 목적으로 재생 수를 인위적으로 부풀린 부정 재생으로 탐지되면서, 플랫폼의 정산 결과가 부정 재생에 의해 왜곡될 가능성이 제기됨¹⁾
- 부정 재생이 정산에 반영될 경우 한정된 로열티 재원이 실제 이용자 청취에 기반하지 않은 음원에도 배분되어, 창작자에게 돌아갈 수익이 줄어들 우려가 있음

1) Jesper Wendel, "Unofficial 2026 World Cup anthems flood music streaming - 70% are tagged as AI on Deezer", Deezer Newsroom, 2026.06.08., <https://newsroom-deezer.com/2026/06/world-cup-2026-ai-generated-music-deezer/>

사후 삭제 중심 대응에서 탐지, 표시, 수익 배제 관리로 전환

• 업로드 단계의 탐지와 표시를 통한 노출 관리

- 디저는 AI 생성 음원과 부정 재생 문제에 대응해, 음원 업로드 단계에서 AI 생성 여부를 탐지하고 이후 노출 범위를 조정하는 관리 방식을 도입함
- 음원이 전적으로 AI로 생성된 것으로 탐지된 경우에는 ‘AI 생성’ 표시를 부착해, 이용자가 청취 전에 해당 음원의 제작 방식을 확인할 수 있도록 함
- AI 생성 표시가 붙은 음원은 자동 추천 목록뿐 아니라 플랫폼이 직접 선별한 재생목록에서도 제외되어, 일반 이용자에게 노출될 가능성이 낮아짐
- 이 방식은 추천과 배치 경로를 제한해 대량 유입된 AI 생성 음원의 확산 속도를 낮추는 데 초점을 둔 대응으로, 음원을 삭제하지 않고도 플랫폼 내 유통 범위를 관리하는 수단으로 기능함

• 정산 단계에서의 부정 재생 배제

- 노출 제한과 함께, 수익 정산 부문에서 디저는 부정 재생으로 판정된 재생을 로열티 산정 대상에서 제외함
- 이러한 정산 배제는 부정 재생을 통한 로열티 수취를 제한함으로써, 부당 수익 창출을 목적으로 AI 생성 음원을 대량 등록하려는 유인을 낮추는 역할을 함
- 이에 따라 디저의 AI 생성 음원 관리 방식은 AI 생성 여부 탐지와 표시, 추천 및 재생목록 노출 제한, 부정 재생의 로열티 정산 제외를 결합하는 형태로 구체화되고 있음

플랫폼 투명성 중심 관리 방식의 확산

• 타 플랫폼으로 확산되는 AI 생성 음원 표시 움직임

- 디저가 대형 스트리밍 서비스 중 가장 앞서 AI 생성 음원을 자동으로 탐지하고 표시하는 방식을 운영하는 가운데, 다른 플랫폼에서도 AI 생성 음원 표시 기준을 마련하려는 움직임이 나타나고 있음
- 예컨대, 애플 뮤직(Apple Music)은 음원 등록 과정에서 입력되는 제목, 창작자, 제작 방식 등 등록 정보를 활용해 AI로 제작됐거나 AI의 도움을 받은 음원을 구분해 표시하는 방식을 도입함
- 이와 같이 현재 플랫폼의 AI 생성 음원 표시 방식은 음원 자체를 분석하는 자동 탐지와 창작자 및 유통사가 제출한 등록 정보를 기준으로 한 표시가 병행되는 흐름으로 형성되고 있음

• AI 생성 음원 표시 수요 확대와 수익 배분 과제

- AI 생성 음원 표시 확대는 이용자의 정보 요구와도 맞물려 있음. 디저가 2025년에 약 9천 명을 대상으로 진행한 조사에서 응답자의 80%는 AI 생성 음원이 명확히 표시되어야 한다고 답함²⁾
- 다만 AI 생성 음원 표시가 확대되더라도, AI 생성 음원과 인간 창작 음원에 로열티를 어떤 기준으로 배분할 것인지는 별도의 과제로 남아 있음
- 따라서 향후 쟁점은 AI 생성 여부를 표시하는 단계를 넘어, 표시된 AI 생성 음원을 추천 노출, 수익 정산, 권리 관리 등 체계에서 어떻게 처리할 것인지로 확대될 전망이다

2) Jesper Wendel, "Unofficial 2026 World Cup anthems flood music streaming – 70% are tagged as AI on Deezer", Deezer Newsroom, 2026.06.08., <https://newsroom-deezer.com/2026/06/world-cup-2026-ai-generated-music-deezer/>

참고문헌

- Jesper Wendel, “Unofficial 2026 World Cup anthems flood music streaming - 70% are tagged as AI on Deezer”, Deezer Newsroom, 2026.06.08., <https://newsroom-deezer.com/2026/06/world-cup-2026-ai-generated-music-deezer/>
- Johanna Campbell, “Deezer says most unofficial World Cup 2026 songs are AI-generated”, Route Note Blog, 2026.06.11., <https://routenote.com/blog/deezer-ai-generated-world-cup-2026-songs/>
- Lachlan Williams, “Most Unofficial World Cup Songs Are Now Made by AI, Deezer Finds”, The Rio Times, 2026.06.15., <https://www.riotimesonline.com/deezer-ai-generated-world-cup-songs-streaming-june-2026/>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

AI 합성 콘텐츠 대응을 위한 외부 위협 방어 플랫폼의 탐지 기능 통합

딥페이크 사칭과 AI 합성 콘텐츠 활용에 따른 진위 검증의 한계

• 생성형 AI를 결합한 기업 대상 딥페이크 공격의 확산

- 최근 AI로 생성하거나 조작한 음성, 영상, 이미지를 동원하는 딥페이크 공격이 늘면서, AI 합성 콘텐츠가 사칭과 사기에 활용되어 한층 더 정교한 기만 수단으로 쓰이고 있음
- 특히 특정 기업의 임원 발언을 가짜 영상으로 위조하거나 브랜드 명의를 도용한 AI 합성 콘텐츠로 신뢰를 훼손하는 사례가 나타나고 있으며, 피싱, 투자 사기 및 계정 탈취 등 후속 공격으로 이어지는 양상도 늘고 있음
- 이에 따라 기업 보안의 초점은 의심 콘텐츠의 유포 차단에 그치지 않고, 해당 콘텐츠가 실제 촬영된 것인지 AI로 합성된 것인지에 대한 명확한 진위 검증으로 옮겨가고 있음

• 딥페이크 탐지와 외부 위협 방어 플랫폼의 결합 사례 대두

- 기존 외부 위협 방어(external threat protection)* 플랫폼은 소셜 미디어, 웹·도메인 영역, 다크웹(dark web) 등 기업 외부 환경에서 발생하는 사칭 계정, 악성 도메인 및 브랜드 침해 요소를 탐지하고 차단하는 데 강점이 있었으나, 의심 콘텐츠의 AI 합성 여부를 직접 검증하는 기능은 제한적이었음
- 이러한 한계를 보완하기 위해 외부 위협 방어 기업 제로폭스(ZeroFox)와 AI 합성 콘텐츠 탐지 전문 기업 리얼리티 디펜더(Reality Defender)는 2026년 6월 2일 딥페이크 탐지 기능을 제로폭스 플랫폼에 통합하는 협력을 발표함
- 양사는 외부 위협 탐지 및 차단 역량과 AI 합성 콘텐츠 탐지 기술을 결합해, 기존에는 별도로 수행되던 콘텐츠 진위 검증을 외부 위협 방어 절차 안에서 함께 처리하는 구조를 제시함

* 외부 위협 방어(external threat protection): 외부 사이버 공격자가 기업의 네트워크나 시스템에 접근하기 전에 악성코드, 피싱, 취약점 노출 등의 위협을 사전에 탐지하고 차단하는 보안 전략

탐지 기능을 대응 절차에 편입한 통합 대응 체계 형성

• 위협 발견부터 차단까지 단일 흐름으로의 일원화

- 이번 통합의 핵심은 외부 위협을 탐지하고 차단하는 기존 대응 절차 안에 딥페이크 탐지 기능까지 포함해, 사칭 계정이나 악성 도메인의 탐지 및 차단에 더해 의심 콘텐츠의 AI 합성 여부까지 함께 검증할 수 있도록 한 데 있음
- 이에 따라 분석가는 기존 작업 환경을 벗어나지 않고 합성 여부 판정 결과를 확인할 수 있으며, 위협 탐지부터 검증과 차단까지를 하나의 대응 절차 안에서 처리할 수 있음

- 그 결과 외부 위협 방어는 위협 인프라의 탐지 및 차단에 머물지 않고, 의심 콘텐츠가 실제 자료인지 AI 합성 콘텐츠인지까지 함께 검증하는 통합 대응 체계로 확장됨

• 무인 자동 판정을 통한 검증 단계의 신속성 확보

- 리얼리티 디펜더의 합성 콘텐츠 탐지 API*인 리얼API(RealAPI)는 음성, 영상, 이미지를 입력받아 AI 합성 여부를 자동으로 판정하므로, 기존 외부 위협 방어 플랫폼 안에서도 별도 도구 전환 없이 콘텐츠 진위 검증 결과를 활용할 수 있음
- 또한 입력된 콘텐츠를 사람이 직접 확인하지 않아도 자동 분석과 판정이 가능해, 위협 탐지 이후 검증 단계에서 발생하는 대기 시간을 줄이고 차단 여부를 더 신속하게 판단할 수 있음
- 여기에 합성 콘텐츠 탐지에 특화된 전용 AI 모델이 음성의 비정상적 패턴과 영상, 이미지의 조작 흔적을 분석해, 콘텐츠가 AI로 합성되었는지 판정할 수 있는 근거를 제공함

* API(Application Programming Interface): 플랫폼이나 서비스가 보유한 기능·데이터를 외부 애플리케이션과 연동할 수 있도록 제공하는 접속 규칙 및 인터페이스

차단 중심 대응에서 검증 선행 구조로의 이행

• 탐지 기술 내재화의 업계 확산 전망과 적용 범위의 한계

- 이번 사례는 콘텐츠 유포 이후 합성 여부를 별도로 검증하던 기존의 사후 검증 방식에서 벗어나, 외부 위협을 차단하기 전에 선제적으로 콘텐츠의 진위를 판별할 수 있음을 보여줌
- 딥페이크 위협은 사칭 계정이나 악성 도메인 차단만으로 대응하기 어려운 만큼, 탐지 도구를 개별적으로 운용하기보다 기존 보안 대응 절차 안에서 합성 여부를 함께 검증하는 방식이 현실적 대응 방향으로 부상하고 있음
- 특히 외부 위협 방어 플랫폼이 딥페이크 탐지를 정식 기능으로 통합한 비교적 초기 사례라는 점에서, 탐지 기술과 대응 역량을 결합하려는 시도가 보안 업계 전반으로 확산될 가능성이 있음
- 다만 이러한 내재화는 대응 시간을 줄이고 차단 판단의 근거를 보강하는 검증 계층에 해당한다는 점, 그리고 AI 합성 콘텐츠 확산으로 발생하는 근본적인 신뢰 문제를 단독으로 해소하는 수단은 아니라는 점을 함께 고려할 필요가 있음

참고문헌

- ZeroFox, Inc., “ZeroFox and Reality Defender Partner to Close the Synthetic Media Gap in External Threat Protection”, GlobeNewswire, 2026.06.02., <https://www.globenewswire.com/news-release/2026/06/02/3305057/0/en/zerofox-and-reality-defender-partner-to-close-the-synthetic-media-gap-in-external-threat-protection.html>
- Ben Colman, “ZeroFox Adds Reality Defender’s Deepfake Detection for External Threat Protection”, Reality Defender, 2026.06.02., <https://www.realitydefender.com/insights/zerofox-adds-reality-defenders-deepfake-detection-for-external-threat-protection>
- Joel R. McConvey, “Reality Defender integrates deepfake detection into ZeroFox platform”, Biometric Update, 2026.06.03., <https://www.biometricupdate.com/202606/reality-defender-integrates-deepfake-detection-into-zerofox-platform>



저작권 이슈 브리프

SUMMARY

산업/기업

기술

주간 기술 동향

음성 길이 변조를 활용한 워터마킹 기술, DuraMark

• AI 음성 복제 시대, 딥페이크 사기에 맞서는 음성 워터마킹 기술의 부상

생성형 AI 기술의 발전으로 단 몇 초의 음성 샘플만으로도 특정인의 음성을 정교하게 복제하는 것이 가능해졌다. 최근, 이러한 AI 음성 복제 기술은 금융 사기에 악용되고 있으며, JP모건의 조사에 따르면 전체 응답자의 52%가 딥페이크 공격을 경험한 것으로 나타났다. 해킹이나 방화벽 침해 없이도 AI가 생성한 음성만으로 신뢰받는 인물을 사칭해 부정 결제를 승인하도록 유도하는 것이 현실이 되었다. 이러한 위협은 이미 특정 집단만의 문제가 아닌, 개인과 조직 모두를 아우르는 광범위한 보안 과제로 자리 잡았다.

특히 금융 현장에서의 피해가 심각해지는 추세다. 음성 기반 본인 인증은 한때 신뢰할 수 있는 보안 수단으로 여겨졌으나, 음성 복제 기술 앞에서 음성 인증 단독으로는 더 이상 안전하지 않다. 조사에 따르면 금융기관의 60%는 딥페이크 사기에 대응하기 위한 포렌식 도구를 갖춘 전단 대응 계획을 마련하지 못한 상태이며, 결제가 실행되기 전 지급 지시의 정당성을 검증할 수 있는 체계가 미비한 곳이 대부분이다. 결제 처리 속도가 빨라질수록 부정 결제가 실행된 이후 자금을 회수할 수 있는 시간은 더욱 줄어든다. 기술적 대응이 위협의 속도를 따라잡지 못하는 상황에서, 음성 인증에 의존하는 기존 보안 체계는 구조적 취약점을 노출하고 있다.

이에 대응하기 위한 기술로 음성 워터마킹이 주목받고 있으나, 기존 방식은 구조적 한계를 안고 있다. 현재 주류를 이루는 신호 수준 워터마킹은 음성 파형이나 스펙트로그램에 식별 신호를 삽입하는 방식으로, 신경망 기반 음성 코덱이나 보코더가 음성을 재합성하면 탐지 성능이 저하될 수 있다. 재합성 공격은 음성의 내용과 의미, 운율 등 정보 수준의 요소는 유지하면서 신호의 세부 정보를 다시 구성하므로, 신호에 의존하는 워터마크가 함께 약화되거나 소실될 수 있다.

이러한 배경에서 신호가 아닌 정보 수준에서 워터마크를 삽입하는 새로운 접근이 제안되었다. 본 보고서에서 다룰 DuraMark는 대규모 언어모델 기반 음성 합성 모델을 활용하여 음절의 지속시간을 조절하는 방식으로 워터마크를 삽입하는 기술이다. 음절 지속시간은 재합성 과정에서도 비교적 잘 유지되는 언어적 속성이기 때문에, 기존의 신호 수준 워터마크보다 재합성 공격에 강한 특징을 갖는다. 지속시간 제어형 음성 합성 모델과 음절 단위 지속시간 추출기를 결합해 워터마크의 삽입과 탐지를 하나의 프레임워크로 구성한 것이 핵심이다.

[사례] 음절 지속시간에 워터마크를 삽입하는 DuraMark

• 기존 신호 수준 음성 워터마킹의 한계

- 기존 음성 워터마킹은 파형이나 스펙트로그램(spectrogram)*에 식별 신호를 삽입하는 신호 수준 방식이 중심이며, 오디오실(AudioSeal), 팀브르(Timbre), 웨이브마크(WavMark) 등이 대표적임
- 신호 수준 워터마크(signal-level watermarks)는 잡음 추가, 압축, 저역 통과 필터링 등 일반적인 변형에는 비교적 강하지만, 신경 오디오 코덱**이나 보코더***가 음성을 재합성하면 탐지 성능이 크게 저하될 수 있음
- 재합성 공격(generative attack)은 발화 내용과 의미, 운율 같은 정보 수준의 요소는 보존하면서 음성 파형의 세부 구조를 새로 생성하므로, 신호에 삽입된 워터마크가 불필요한 세부 정보로 처리되어 약화되거나 소실됨
- 기존 정보 수준 워터마킹(information-level watermarking)은 음성 구간의 음높이를 후처리하여 정보를 삽입하지만, 인위적인 음높이 조절이 자연스러운 운율을 훼손할 수 있어 음질과 공격 강건성을 함께 확보하기 어려움

* 스펙트로그램(spectrogram): 음성 신호를 시간과 주파수 축으로 변환하여, 시간대별로 어떤 주파수 성분이 얼마나 강하게 나타나는지 시각적으로 표현한 음향 특징 자료

** 신경 오디오 코덱(neural audio codec): 음성 신호를 신경망 기반의 압축 표현으로 변환한 뒤 다시 음성 파형으로 복원하는 기술

*** 보코더(vocoder): 멜 스펙트로그램과 같은 음성의 특징이나 주파수 정보를 입력받아 사람이 들을 수 있는 음성 파형을 생성하는 기술

• 지속 시간 기반 워터마크 생성 구조

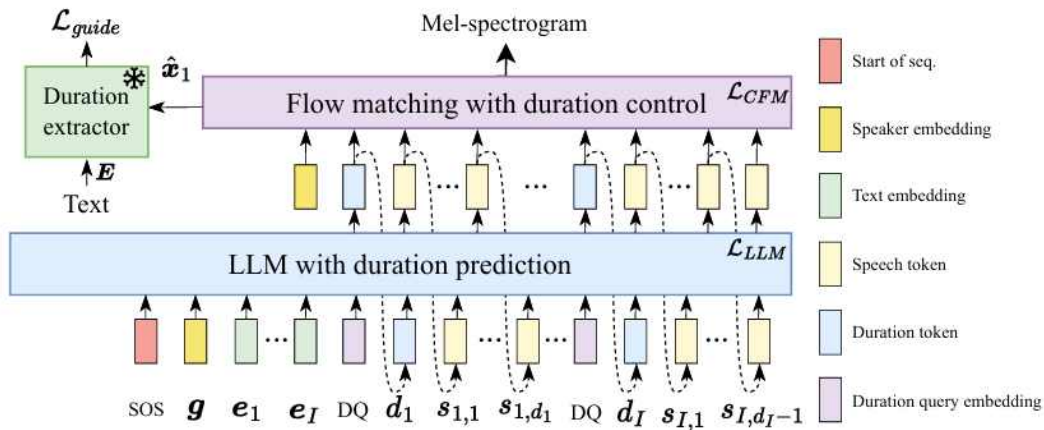
- DuraMark는 대규모 언어모델(large language model, 이하 LLM) 기반 텍스트 음성 변환(text-to-speech) 모델에 음절별 지속시간 예측 기능을 추가하여, 각 음절이 몇 개의 음성 프레임으로 구성될지 명시적으로 제어하는 구조임
- LLM은 화자 임베딩과 음절 단위 텍스트 임베딩을 입력받고, 각 음절마다 지속시간 질의 임베딩(duration query embedding)*을 통해 지속시간 토큰을 먼저 예측한 뒤 해당 길이에 맞는 음성 토큰을 순차적으로 생성함
- 한 음절에서 생성된 지속시간 토큰과 음성 토큰은 다음 음절을 예측하는 조건으로 다시 입력되며, 이러한 자기회귀 방식(autoressive method)으로 문장 전체의 음절 길이와 발화 흐름을 연속적으로 구성함
- 생성된 지속시간 토큰과 음성 토큰은 흐름 정합(flow matching)** 디코더에 입력되며, 디코더는 화자 임베딩, 음성 토큰, 음절별 지속시간을 조건으로 실제 음성 합성에 필요한 멜 스펙트로그램(mel-spectrogram)***을 생성함

* 지속 시간 질의 임베딩(duration query embedding): LLM에 특정 음절의 지속 시간 토큰을 예측하도록 지시하는 입력 표현

** 흐름 정합(flow matching): 잡음 분포에서 목표 음향 데이터로 이동하는 연속적인 변화 경로를 학습하여 멜 스펙트로그램을 생성하는 방식

*** 멜 스펙트로그램(mel-spectrogram): 음성 신호를 사람이 소리를 인식하는 방식에 가까운 주파수 축으로 변환해, 시간에 따른 음향 특징을 시각적으로 나타낸 자료

[그림 1] DuraMark의 지속시간 제어형 음성 합성 구조



출처: Zhenwei Mou 외 6인, "DuraMark: Duration-Embedded Watermarking in LLM-based TTS", arXiv, 2026.06.13., <https://arxiv.org/pdf/2606.15264>

• 짝수/홀수 상태를 이용한 워터마크 삽입

- DuraMark는 입력 문장의 음절 수와 동일한 길이의 이진 워터마크 배열을 생성하고, 각 음절의 지속시간을 하나의 비트 전달 단위로 활용해 문장 전체에 워터마크 정보를 순차적으로 분산 삽입하는 방식으로 작동함
- 각 음절의 지속시간이 짝수 프레임이면 비트 0, 홀수 프레임이면 비트 1로 대응시키며, 예측된 지속시간의 상태가 목표 워터마크 비트와 일치하면 기존 값을 그대로 유지함
- 예측된 지속시간이 목표 비트와 일치하지 않으면 프레임 수를 1 늘린 값과 1 줄인 값을 후보로 비교하고, LLM의 지속시간 확률 분포에서 더 높은 값을 선택해 발화 리듬과 자연스러움의 변형을 최소화함
- 편집된 지속시간에 맞춰 음성 토큰을 생성한 뒤 흐름 정합 디코더에 함께 입력하면, 음절별 홀짝 배열에 워터마크 정보가 반영되면서도 원래 화자의 특성과 발화 내용이 안정적으로 유지된 합성 음성이 생성됨

• 지속시간 추출과 상관 점수 기반 워터마크 탐지

- DuraMark는 검증 대상 음성과 해당 전사문(transcription)을 지속시간 추출기에 함께 입력하고, 각 음성 프레임이 어느 음절에 속하는지 확률로 추정한 뒤 동일 음절에 배정된 프레임 확률을 모두 합산하여 음절별 지속시간 배열을 추정함
- 추출된 지속시간은 짝수에 가까울수록 -1, 홀수에 가까울수록 1이 되도록 코사인 함수를 이용해 변환하며, 원래 워터마크 비트도 0은 -1, 1은 1로 바꾸어 두 배열을 동일한 값 범위에서 직접 비교할 수 있도록 정규화함
- 변환된 두 배열의 대응 원소를 곱해 평균한 상관 점수(correlation score)를 계산하고, 이 값이 사전에 설정한 임계값을 넘으면 워터마크가 존재하는 것으로, 넘지 못하면 존재하지 않는 음성으로 분류함

• 탐지 조건별 성능과 재합성 공격 강건성

- DuraMark는 비교적 짧은 발화에서도 안정적인 탐지 성능을 보였으며, 발화 길이가 늘어날수록 더 많은 음절에 워터마크 비트를 분산할 수 있어 개별 음절의 추출 오차가 전체 판정에 미치는 영향이 줄어드는 경향이 확인됨

- 원본 전사문을 사용하는 정보 제공형 탐지(informed detection)*는 음절 경계를 정확하게 파악할 수 있어 블라인드 탐지(blind detection)**보다 전반적으로 우수했으며, 블라인드 탐지도 자동 음성 인식 결과만으로 높은 탐지 성능을 유지함
- 기존 신호 수준 워터마킹은 일부 신경 오디오 코덱이나 보코더를 거치면 탐지 성능이 크게 낮아지는 반면, DuraMark는 대부분의 재합성 조건에서 상대적으로 안정적인 성능을 유지함
- 압축, 잡음 추가, 음성 향상과 같은 일반적인 변형에서는 기존 기술도 비교적 안정적으로 작동했으나, 생성 모델이 음성 파형을 새로 구성하는 공격에서는 DuraMark의 성능 우위가 더 뚜렷하게 확인됨
- 이러한 강건성은 워터마크를 파형의 세부 신호가 아니라 재합성 이후에도 비교적 유지되는 음절 지속시간에 삽입한 결과이며, 정보 수준 특징을 활용하는 방식의 기술적 효과가 확인됨

* 정보 제공형 탐지(informed detection): 원본 전사문을 함께 입력해 음성과 텍스트의 음절 경계를 맞춘 뒤, 각 음절에 대응하는 프레임 수를 계산하여 지속시간을 추출하는 방식

** 블라인드 탐지(blind detection): 원본 전사문 없이 자동 음성 인식 결과를 이용해 음절 배열을 구성하고, 각 음절에 대응하는 프레임 수를 계산하여 지속시간을 추출하는 방식

결론 및 시사점

• 기술적 의의와 향후 발전 가능성

- DuraMark는 음성 파형이 아닌 음절 지속시간에 워터마크를 삽입하여, 신경 코덱과 보코더를 이용한 재합성 공격에서도 기존 신호 수준 방식보다 안정적인 탐지 성능을 확보함
- 음질과 발화 자연스러움을 유지하면서 원본 전사문이 없는 환경에서도 탐지가 가능하므로, 향후 음성 합성 서비스와 온라인 유통 환경의 생성 음성 식별 기술로 활용될 가능성이 있음

• 현재 기술의 한계점

- 현재 성능 검증은 표준 중국어(Mandarin Chinese)로 녹음된 일정 길이 이상의 발화를 중심으로 진행되어, 음절 구조가 다른 언어와 짧은 음성에서도 동일한 탐지 성능이 유지되는지는 확인되지 않음
- 탐지 과정에서 원본 전사문이나 자동 음성 인식 결과가 필요하므로 전사 오류의 영향을 받을 수 있으며, 발화 속도 조절처럼 음절 지속시간을 직접 변경하는 공격에 대한 추가 검증도 필요함

참고문헌

- Zhenwei Mou 외 6인, "DuraMark: Duration-Embedded Watermarking in LLM-based TTS", arXiv, 2026.06.13., <https://arxiv.org/pdf/2606.15264>
- J.P. Morgan, "Deepfake fraud is evolving fast—here's how to protect your payments", 2026.06.22., <https://www.jpmorgan.com/insights/payments/security-trust/deepfake-fraud-prevention-strategies>