

저작권 이슈 트렌드



COPYRIGHT ISSUE TREND



한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제90호(2026. 9-2)

- 인공지능 생성 음악 탐지 기술의 진화와 산업적 대응
- 영상 워터마킹을 활용한 콘텐츠 출처 확인과 진본 보호
- 속도 변조에도 성능 유지하는 AI 생성 음악 판별 기술



인공지능 생성 음악 탐지 기술의 진화와 산업적 대응

뉴스 브리프

AI로 만든 노래가 음원 시장에 유입되면서, 이를 사람이 만든 음악과 구분해 표시하려는 움직임이 확대되고 있다. 이에 빌보드 차트의 데이터를 제공하는 루미네이트를 비롯해 주요 스트리밍 사업자들이 AI 생성 음악에 표시를 붙이는 체계를 잇따라 도입하고 있다. 그러나 음악을 만들어 내는 생성 도구가 계속 늘어나고 있어, 처음 등장한 도구로 만든 음악까지 정확히 가려내는 일은 쉽지 않다. 특정 도구에 맞춰 학습한 기존 탐지 방식은 새로운 도구 앞에서 성능이 떨어지기 때문이다. 이는 창작자의 권리 보호와 음악 생태계의 신뢰 유지라는 과제와 맞닿아 있다. 본 보고서는 사람이 만든 음악의 특성만을 학습해 AI 생성 음악을 걸러내는 탐지 기술을 분석하고, 이 기술이 창작물 보호에 기여할 방향과 남은 과제를 전망한다.

뉴스 플러스

1. 서론: AI 생성 음악의 확산과 탐지 기술의 부상

• AI 생성 음악의 확산과 표시 체계의 도입

AI로 만든 노래가 음원 시장에 빠르게 퍼지면서, 이를 사람이 만든 음악과 구분해 이용자에게 알리려는 시도가 음원 서비스 곳곳에서 잇따르고 있다. 2026년 8월, 미국의 엔터테인먼트 데이터 분석 업체이자 빌보드(Billboard) 차트의 데이터를 제공해 온 루미네이트(Luminate)는 AI 생성 음악을 식별해 표시하는 체계를 발표하며, 연내에 고객과 파트너를 대상으로 적용하겠다고 밝혔다. 이 회사는 자사 플랫폼에서 AI 생성 음악으로 확인된 곡과 AI 아티스트에 'AI 생성(AI Generated)'이라는 표시를 붙이고, 잘못 표시된 경우 재검토를 신청할 수 있는 절차도 함께 두기로 했다. 스포티파이(Spotify)와 애플 뮤직(Apple Music) 역시 연내에 비슷한 표시 체계를 예고하면서, 이러한 움직임은 주요 음원 서비스 플랫폼 전반으로 번지고 있다.

루미네이트가 밝힌 표시 체계의 목적은, 늘어나는 AI 생성 음악을 사람이 만든 음악과 구분해 그 규모와



흐름을 측정하는 데 있다. 이 체계는 세 가지 정보를 함께 활용한다. 첫째는 스트리밍 데이터(streaming data)*와 메타데이터(metadata)**를 분석하는 자사의 독점 식별 시스템이다. 둘째는 음원 서비스에서 집계되는 재생 기록이다. 셋째는 생성 단계에서 AI 음악 기업이 제공하는 출처 정보다. 표시 대상은 완전히 AI로 만든 곡과 아티스트이며, 사람이 AI를 도구로 활용해 만든 음악은 이와 구분된다. 이 가운데 독점 식별 시스템은 재생 기록과 곡 정보를 살피는 것에 더해, 음악 신호 자체를 분석해 AI 생성 여부를 식별하는 방식을 포함하며, 뒤이어 살펴볼 탐지 기술도 이러한 음악 신호 분석에 해당한다.

* 스트리밍 데이터(streaming data): 재생 횟수, 재생 시간, 이용자 정보 등 음원, 영상 등 콘텐츠가 스트리밍되는 과정에서 발생하고 수집되는 데이터

** 메타데이터(metadata): 어떤 데이터를 설명하기 위해 부여하는 데이터. 생성일, 작성자, 형식, 제목처럼 대상의 속성과 내용을 나타내는 정보 등이 포함됨

• 사람이 만든 음악만으로 학습하는 탐지 기술의 등장

음악 신호를 분석해 AI 생성 여부를 가려내는 과제는, 그동안 특정 생성 도구가 남기는 고유한 흔적을 학습해 그 도구의 산출물을 찾아내는 방식으로 주로 다뤄져 왔다. 이러한 방식은 학습에 사용한 도구로 만든 음악에는 높은 정확도를 보이지만, 학습 단계에서 접하지 못한 새로운 도구로 만든 음악에서는 정확도가 떨어지는 특성이 있어, 여러 도구가 뒤섞이는 실제 환경에서는 활용에 제약이 따른다. 음악을 만들어 내는 생성 도구가 계속 늘어나는 상황에서, 도구마다 별도의 탐지기를 만드는 방법은 현실적으로 한계가 있다. 이에 학습에 쓰지 않은 도구로 만든 음악까지 가려내는, 생성 도구에 얽매이지 않는 탐지 방식을 구현하는 일이 과제로 떠올랐다.

이러한 배경에서 제안된 MusicDET는, AI 생성 음악의 사례를 학습에 전혀 쓰지 않고 오직 사람이 만든 음악만으로 학습하며, 학습 대상에 없던 새로운 도구의 산출물까지 다루는 제로샷(zero-shot)* 탐지 기술이다. 이 기술은 사람이 만든 음악이 지닌 소리의 분포를 미리 학습해 두고, 새로 들어온 음악에 그 분포를 기준으로 점수를 매긴 뒤, 분포에서 벗어난 정도가 큰 음악을 AI 생성 음악으로 가려내는 원리를 따른다. 전문 연주자가 사람이 만든 음악을 오랫동안 접하며 진짜 음악의 소리를 분별하는 것과 같은 접근에서 착안한 방식이다. 사람이 만든 음악만을 학습에 쓰기 때문에, 새로운 도구가 등장하더라도 별도의 재학습 없이 대응할 수 있다는 점이 기존 방식과 구별되는 지점이다.

* 제로샷(zero-shot): 학습 단계에서 특정 대상의 예시를 전혀 사용하지 않고도 그 대상을 처리해 내는 방식

II. 본론: 사람이 만든 음악의 분포 학습을 통한 탐지

• 제로샷 탐지와 생성기 비의존 접근의 등장 배경

MusicDET는 사람이 만든 음악만으로 학습한 상태에서 AI 생성 음악을 가려내는 것이 핵심이다. 따라서 학습에 쓰이는 음악에는 AI 생성 음악이 한 곡도 들어가지 않는다. 성능은 교차 생성기 평가(cross-generator evaluation)로 확인하는데, 이는 학습과 개발 과정에서 접하지 않은 하나 이상의 도구로 만든 음악을 탐지 대상에 포함해 그 음악까지 제대로 가려내는지를 측정하는 방식이다. 곧 학습에 쓰지 않은 도구로 만든 음악에서도 제대로 작동하는지를 확인하는 절차다.

이 기술의 접근은 이상 탐지(anomaly detection)에 해당하는데, 이는 정상으로 규정한 데이터의 분포를 학습한 뒤 그 분포에서 벗어나는 데이터를 찾아내는 방식이다. 여기서는 사람이 만든 음악을 정상으로 두고, 그 분포에서 벗어나는 음악을 AI 생성 음악으로 가려낸다. AI 생성 음악을 학습에 전혀 쓰지 않고도 사람이 만든 음악이 아닌 것을 가려내는 구조이기 때문에, 학습에 없던 새로운 도구로 만든 음악이 들어와도 그 도구의 흔적을 미리 배워 둘 필요가 없다는 점이 이 접근의 특징이다. 도구마다 탐지기를 따로 학습시키지 않아도 되므로, 새로운 도구가 이어서 나오는 환경에서 탐지 체계를 유지하기에 알맞다.

이 방식이 성립하는 근거는, 사람이 만든 음악과 AI 생성 음악이 소리의 구조에서 서로 다른 특징을 보인다는 관찰에서 나오며, 이는 여러 곡을 비교했을 때 일관되게 확인된다. 사람이 만든 음악은 시간과 주파수에 걸쳐 짜임새 있는 구조를 보이는 반면, AI 생성 음악은 상대적으로 불규칙한 패턴을 나타내는 경향이 있다. 다만 사람이 만든 음악은 장르, 악기, 제작 방식이 제각각이어서 그 분포가 여러 갈래로 복잡하게 나타나고, 이 때문에 분포 전체를 하나의 방식으로 뭉뚱그려 다루면 학습이 불안정해지고 탐지 정확도도 함께 떨어질 수 있다. 이처럼 복잡하게 나타나는 분포를 안정적으로 다루는 것이, 이어지는 탐지 방법 설계의 출발점이 된다.

• 에너지 스펙트럼과 정규화 흐름의 작동 원리

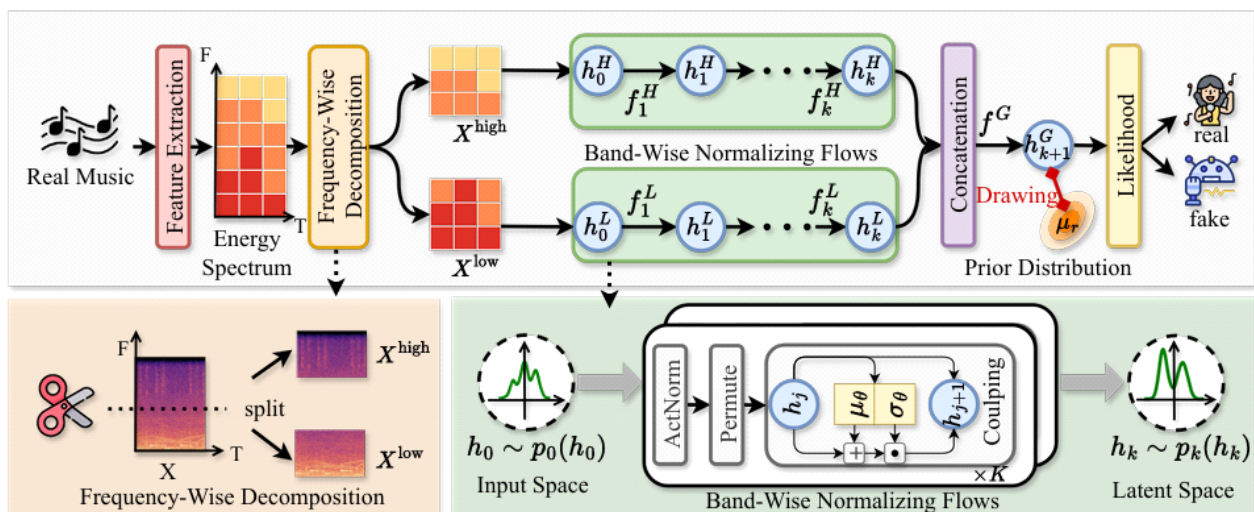
MusicDET는 먼저 음악 신호를 시간과 주파수에 따라 에너지의 크기로 나타낸 표현인 에너지 스펙트럼(energy spectrum)으로 변환시킨다. 이는 어느 시점에 어느 높낮이의 소리가 얼마나 강하게 담겼는지를 하나의 표현으로 정리한 것이다. 이를 위해 음악을 짧은 구간으로 나누어 주파수 성분을 분석하고, 그 결과를 처리해 에너지 특성을 뽑아낸다. 이 표현에는 화음의 짜임, 리듬의 규칙성, 음색의 결처럼 음악을 이루는 여러 특성이 함께 담겨 있어, 사람이 만든 음악과 AI 생성 음악의 차이를 소리 자체의 짜임에서 곧바로 살필 수 있게 해 주는 표현으로 쓰이고 있다.

에너지 스펙트럼의 분포를 학습하는 데에는 정규화 흐름(normalizing flow)이 쓰이는데, 이는 여러 단계의 변환을 거쳐 복잡한 데이터의 분포를 단순한 분포로 옮기는 방법이다. 여러 갈래로 흩어져 있는 사람이 만든 음악의 소리 특성을, 정규화 흐름은 일정한 규칙에 따라 다루기 쉬운 형태로 바꾸어 학습한다. 이때 각 변환은 되돌릴 수 있게 설계되어 단순한 형태에서 원래 분포로도 옮겨 갈 수 있으며, 덕분에 어떤 음악이 학습한 분포에 얼마나 들어맞는지를 하나의 값으로 계산할 수 있다. 이러한 성질은 학습한 분포에서 벗어나는 음악을 찾아내는 데 쓰기 좋아, 정상에서 벗어난 데이터를 가려내는 여러 분야에서 활용되어 왔다.

음악은 주파수의 높낮이에 따라 특성이 다른데, 낮은 음역은 리듬과 기본 음높이를, 높은 음역은 음색과 미세한 소리의 변화를 담는다. 이 때문에 MusicDET는 에너지 스펙트럼을 주파수 높낮이에 따라 여러 구간으로 나누는데, 이를 주파수 대역 분해(frequency-wise decomposition)라 한다. 곧 낮은 음역과 높은 음역을 따로 떼어, 각 구간의 특성을 나누어 다루는 것이다. 나뉜 각 구간은 저마다 별도의 정규화 흐름으로 학습되어, 해당 음역이 지닌 규칙을 따로 담아낸다. 전체 스펙트럼을 하나의 흐름으로 한꺼번에 다루면 서로 다른 음역의 특성이 뒤섞여 학습이 불안정해질 수 있는데, 이렇게 음역을 구간별로 나누어 다루면 각 구간의 규칙을 놓치지 않으면서 분포를 더 안정적으로 학습할 수 있다.

각 구간에서 학습한 결과는 다시 하나로 합쳐져 전역 흐름을 거치는데, 이는 음역 사이의 연관성을 함께 담기 위한 단계다. 이 과정을 거치면, 최종적으로 새 음원이 사람이 만든 음악의 분포에 얼마나 부합하는지를 나타내는 우도(likelihood)를 얻을 수 있다. 우도가 높으면 사람이 만든 음악의 분포에 가깝고, 낮으면 그 분포에서 벗어난 것으로 보아 AI 생성 음악으로 가려낸다. 이렇게 소리 자체를 분석해 사람이 만든 음악과 AI 생성 음악을 구분하는 것은, 어떤 음원이 사람의 창작물인지 확인하고 표시하며 관리하는 절차에 활용될 수 있는 기술적 근거로 이어지고, 창작자 보호를 뒷받침하는 수단이 된다.

[그림 1] MusicDET의 AI 생성 음악 탐지 과정



출처: Chaolei Han 외 2인, "MusicDET: Zero-Shot AI-Generated Music Detection", arXiv, 2026.05.18., <https://arxiv.org/pdf/2605.18072>

• 클래스 조건부 확장과 탐지 성능의 확인

MusicDET는 사람이 만든 음악만으로 학습하지만, AI 생성 음악을 함께 쓸 수 있는 경우에는 클래스 조건부(class-conditional) 방식으로 확장된다. 클래스 조건부란 데이터를 종류별로 구분해, 각 종류의 분포를 따로 상정하고 학습하는 방식이다. 여기서는 사람이 만든 음악과 AI 생성 음악을 각각 다른 중심을 가진 분포로 두고, 두 종류의 음악을 모두 학습에 사용해 각 중심을 정한다. 다만 변환 자체는 두 종류가 공유하고 종류 정보는 분포의 중심에만 반영되기 때문에, 이 방식도 둘을 갈라놓는 분류기가 아니라 분포에서 벗어났는지를 확인하는 탐지기로 작동한다. 실제 탐지 단계에서는 사람이 만든 음악의 분포만을 기준으로 우도를 계산하며, AI 생성 음악은 그 중심에서 떨어진 영역에 놓여 우도가 낮게 나타난다.

성능은 여러 데이터셋에서 동일오류율(Equal Error Rate, 이하 EER)로 측정되며, EER은 놓치는 오류와 잘못 짚는 오류가 같아지는 지점의 오류율로 값이 낮을수록 성능이 좋다. 학습에 쓰지 않은 도구로 만든 음악을 대상으로 한 평가에서, 특정 도구의 흔적에 치우친 기존 방식들은 성능이 떨어진 반면 MusicDET는 낮은 EER을 유지했고, AI 생성 음악을 함께 학습한 클래스 조건부 방식에서는 EER이 더 낮아졌다. 게다가 학습에 필요한 규모가 작아, 정확도와 처리 효율을 함께 확보한 것으로 평가된다. 다만 음정 변경, 잡음 추가, 재압축과 같은 후처리를 거친 음악에서는 EER이 올라가는데, 이러한 후처리가 탐지의 근거가 되는 소리의 특성을 흐트러뜨리기 때문이다. 곧 창작자가 흔히 적용하는 후처리만으로도 탐지가 약해질 수 있어, 실제 환경에서의 견고성 확보가 과제로 지적된다.

III. 결론 및 전망: 탐지 기술의 고도화와 창작 생태계의 균형

• 탐지 기술의 발전 방향과 남은 과제

MusicDET는 특정 생성 도구의 흔적을 쫓는 대신, 사람이 만든 음악의 분포를 학습해 거기서 벗어나는 음악을 가려내는 방식을 택한다. 이는 새로운 도구가 계속 등장하는 환경에서도 도구마다 탐지기를 새로 만들 필요가 줄어든다는 것을 의미한다. 결국 AI 생성 음악 탐지의 초점이, 개별 도구를 사후 추적하는 방식에서 사람이 만든 음악 자체를 학습하는 쪽으로 옮겨 간다. 이러한 점에서 소리 자체를 분석하는 탐지 기술은, 음원의 창작물 여부를 표시하고 관리하려는 산업의 요구를 기술적으로 뒷받침하는 수단이 된다.

다만 이 기술은 음정 변경이나 재압축 같은 후처리를 거친 음악에서 탐지가 약해지는 한계를 안고 있다. 따라서 실제 환경에서 쓰려면, 다양한 후처리를 거친 음악에서도 견고하게 작동하도록 기술을 보완하는 작업이 필요하다. 동시에 탐지 결과만으로 창작물 여부를 단정하기보다, 생성 단계에서 출처를 기록하는 절차나 사람의 검토와 함께 쓰는 방향이 요구된다. 이를 통해 탐지 기술과 표시 체계가 서로를 보완하며, 사람이 만든 음악과 AI 생성 음악이 공존하는 환경에서 창작자의 권리를 지키는 토대가 마련될 수 있다.



참고문헌

- Kristin Robinson, “Luminate to Add Labels to AI Music Later This Year”, Billboard, 2026.08.25., <https://www.billboard.com/pro/luminate-to-label-ai-music-later-this-year/>
- InsideRadio, “Luminate Lays Out Plan For Labeling AI Music”, 2026.08.26., https://www.insideradio.com/free/luminate-lays-out-plan-for-labeling-ai-music/article_57db3344-697a-4d5f-9f22-f294ee7d6bae.html
- Chaolei Han 외 2인, “MusicDET: Zero-Shot AI-Generated Music Detection”, arXiv, 2026.05.18., <https://arxiv.org/pdf/2605.18072>

기술용어

순번	용어	설명
1	스트리밍 데이터 (streaming data)	재생 횟수, 재생 시간, 이용자 정보 등 음원, 영상 등 콘텐츠가 스트리밍되는 과정에서 발생하고 수집되는 데이터
2	메타데이터 (metadata)	어떤 데이터를 설명하기 위해 부여하는 데이터. 생성일, 작성자, 형식, 제목처럼 대상의 속성과 내용을 나타내는 정보 등이 포함됨
3	제로샷 (zero-shot)	학습 단계에서 특정 대상의 예시를 전혀 사용하지 않고도 그 대상을 처리해 내는 방식



영상 워터마킹을 활용한 콘텐츠 출처 확인과 진본 보호

뉴스 브리프

영상 콘텐츠가 온라인으로 빠르게 퍼지면서, 미공개 영상이 유출되어 소셜미디어와 영상 공유 플랫폼으로 확산되는 사례가 이어지고 있다. 최근 미국의 게임 개발사 록스타 게임스가 개발한 그랜드 테프트 오토 6의 미출시 영상이 유출되어 삼시간에 널리 퍼졌고, 권리자는 유출물을 찾아내고 그 출처를 확인하려는 대응에 나섰다. 이 과정에서 재유통을 거친 영상에서 원본의 출처와 유출 경로를 어떻게 확인할 것인가가 쟁점으로 떠오른다. 영상에 눈에 보이지 않는 정보를 심어 재인코딩과 편집을 거쳐도 이를 복원하는 워터마킹 기술이 그 해법으로 주목받고 있다. 본 보고서는 이러한 영상 워터마킹의 작동 원리를 단계적으로 살펴보고, 콘텐츠 보호와 침해에 맞닿는 지점을 분석하며, 앞으로 필요한 제도와 기술적 보완 방향을 전망한다.

뉴스 플러스

1. 서론: 영상 유출 확산과 출처 확인 문제

• 게임 영상 유출과 온라인 확산의 파장

2026년 8월 미국의 게임 개발사 록스타 게임스(Rockstar Games)가 준비해 온 그랜드 테프트 오토 6(Grand Theft Auto VI)의 미공개 플레이 영상이 정식 공개 전에 유출되어 온라인으로 빠르게 번졌다. 유출된 영상은 소셜미디어와 영상 공유 플랫폼을 거치며 짧은 시간에 수많은 이용자에게 노출됐고, 권리자는 유출물 삭제와 출처 확인을 위한 대응에 나섰다. 권리자는 유출물 삭제 조치에 그치지 않고, 최초 유출 출처와 유출 경로를 함께 확인하려는 대응에 본격적으로 나섰다. 그러나 유출된 영상은 재인코딩과 편집을 거치며 원본과 달라지고, 여러 계정과 플랫폼으로 흩어진 뒤라 출처를 특정하는 일은 간단하지 않다.

이 사례는 한 번의 유출 사건에 그치지 않고, 디지털 영상의 유통 환경 자체가 안고 있는 구조적 취약함을 그대로 드러낸다는 점에서 주목된다. 영상은 복제와 재전송이 쉬운 데다, 압축이나 화면 녹화, 부분 편집을 거치면 겉모습은 조금씩 달라지지만 담긴 내용은 그대로 남아 빠르게 번져 나간다. 이 때문에 권리자는 어떤 영상이 자신의 원본에서 나온 것인지, 또 그 영상이 어떤 경로로 흘러나왔는지를 유출이 벌어진 뒤에 일일이 가려내야 하는 상황에 놓이게 된다. 그러나 변형을 거친 영상이나 클립(clip) 등의 짧은 영상까지 잡아내기는 쉽지 않아, 영상 자체에 출처를 확인할 정보를 미리 심어 두는 방식이 대안으로 떠오른다.

• 영상 워터마킹 기술의 등장 배경과 의미

영상에서 출처를 확인하는 방법은 접근 방식에 따라 크게 두 갈래로 나눌 수 있다. 하나는 원본에서 특징값을 미리 뽑아 두었다가 온라인에 떠도는 영상과 대조하는 방식이고, 다른 하나는 영상 안에 식별 정보를 직접 심어 두는 워터마킹(watermarking) 방식이다. 특징값 기반 방식은 원본 데이터베이스와의 대조가 필요하며, 변형 정도에 따라 성능이 달라질 수 있는 반면, 워터마킹은 영상 자체가 정보를 지니고 다니므로 재유통 상황에서도 출처를 되짚을 근거가 된다. 다만 영상은 압축과 재인코딩, 화면 녹화, 프레임 편집을 수시로 겪기 때문에, 심어 둔 정보가 이런 변형을 견디지 못하면 워터마킹은 제 역할을 하기 어렵다.

이러한 배경에서 영상 워터마킹 연구는 다양한 변형을 겪어도 심어 둔 정보가 지워지지 않는 강건성을 높이는 방향으로 발전해 왔다. 초기의 워터마킹은 영상 전체에 정보를 고르게 심는 방식이 많았으나, 압축이나 편집이 조금만 가해져도 정보가 함께 지워지는 한계가 있었다. 2026년 공개된 플로우마크 (Flow Mark)는 영상에서 워터마크를 심기에 유리한 영역을 찾아 정보를 넣는 방식으로, 압축·재인코딩, 기하 변형, 프레임 편집 등을 거친 뒤에도 심어 둔 정보를 복원한다. 유출된 영상이 여러 플랫폼을 떠도는 현실을 감안하면, 이처럼 재유통으로 인한 변형을 견디는 능력은 출처 확인의 실효성을 좌우하는 요소로 평가된다.

II. 본론: 플로우마크 영상 워터마킹의 작동 원리

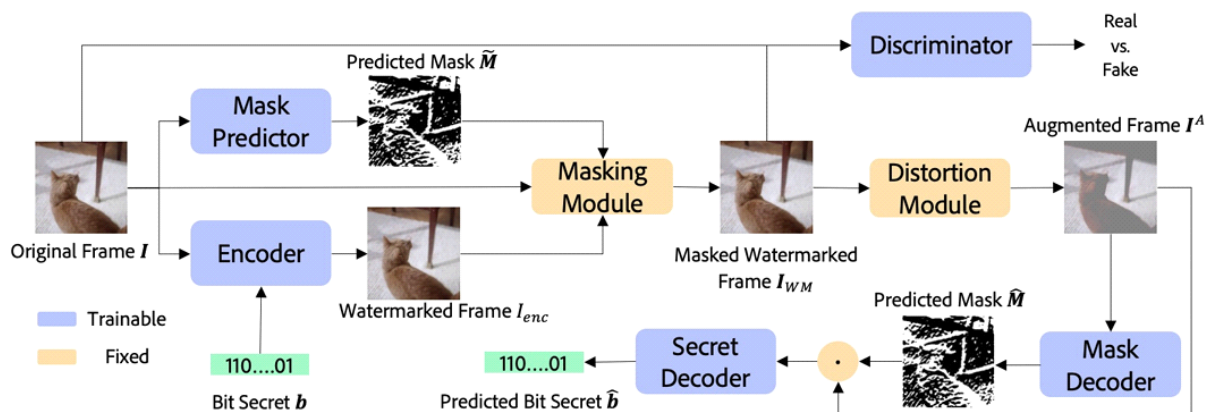
• 마스크 예측을 통한 삽입 영역 선정

마스크(mask)는 영상의 각 화면에서 워터마크를 넣을 영역과 넣지 않을 영역을 서로 구분하여 화면 위에 표시해 둔 하나의 지도를 뜻한다. 쉽게 말하면, 마스크는 한 장면 안에서 정보를 심어도 되는 자리와 손대지 말아야 할 자리를 화면 위에 나누어 표시해 둔 것이다. 기존의 영역 방식 워터마킹은 이 마스크를 사람이 영상마다 직접 지정해 주어야 했기 때문에, 다뤄야 할 영상의 수가 늘어날수록 마스크를 준비하는 데 드는 작업과 시간의 부담도 그만큼 함께 늘어나, 대량의 영상에 적용하기에는 한계가 뚜렷했다. 플로우마크는 마스크 예측망이라는 별도의 신경망을 두어, 각 영상에서 워터마크를 넣기에 적합한 영역을 사람의 개입 없이 프레임마다 스스로 찾아 마스크의 형태로 만들어 낸다.

마스크 예측망은 영상의 움직임과 화면 내용을 프레임 단위로 함께 살펴, 워터마크를 넣어도 눈에 띄지 않으면서 여러 변형을 거쳐도 정보가 잘 남는 영역을 골라낸다. 이렇게 영상의 내용과 움직임에 맞추어 워터마크의 삽입 위치를 장면마다 서로 다르게 정하는 방식을 콘텐츠 적응형 마스킹이라 한다. 즉 색과 밝기가 거의 변하지 않는 정지된 배경보다, 움직이는 물체처럼 워터마크가 주변 화면에 자연스럽게 묻어 들어갈 수 있는 부분을 장면마다 하나씩 실제로 골라내어 그 자리에 워터마크를 잘게 나누어 심어 넣는다는 뜻이다. 플로우마크는 삽입 위치를 영상의 움직임에 맞추어 정함으로써, 프레임이 넘어갈 때 워터마크가 미세하게 깜빡이는 현상을 줄이고 화질 저하도 억제한다.

삽입 위치를 자동으로 정하는 방식은 사람이 일일이 손대지 않고도 많은 분량의 영상에 워터마크를 일관되게 심는 작업을 처리할 수 있게 해 준다. 워터마크가 눈에 띄지 않으면서도 원래의 화질을 거의 해치지 않아야, 관리자가 정식으로 공개하는 영상에 워터마크를 남겨 두더라도 이용자의 시청 경험을 떨어뜨리지 않는다. 특히 삽입 위치가 영상의 움직임에 맞추어 화면 안에 자연스럽게 자리 잡을수록, 압축이나 편집을 거치며 영상이 재유통되더라도 심어 둔 정보가 함께 사라지지 않고 남을 여지가 커진다.

[그림 1] 플로우마크의 워터마크 삽입과 복원 과정



출처: Vishal Asnani 외 2명, "FlowMark: Mask-Guided Video Watermarking", arXiv, 2026.07.06., <https://arxiv.org/pdf/2607.05261>

• 잡음 증강 학습과 워터마크의 삽입·복원 원리

플로우마크는 정보를 삽입하는 인코더(encoder)와 그 정보를 다시 읽어 내는 디코더(decoder)라는 두 개의 신경망을 하나의 짝으로 두고 함께 사용한다. 인코더는 마스크가 가리키는 영역에 감추려는 정보를 눈에 띄지 않는 신호로 바꾸어, 원본 영상 위에 티가 나지 않게 더해 넣는 역할을 맡는다. 디코더는 이렇게 워터마크가 심어진 영상을 받아, 화면 곳곳에 섞여 든 신호를 다시 읽어 내어 처음에 심어 둔 정보를 복원해 낸다. 두 신경망은 처음부터 끝까지 하나로 이어진 채 함께 학습되기 때문에, 인코더가 정보를 심는 방식과 디코더가 그것을 읽어 내는 방식이 서로 맞물리면서 복원 과정의 전체적인 정확도도 학습을 거듭할수록 점차 높아지게 된다.



잡음 증강 학습(noise-augmented training)은 학습 과정에서 영상에 여러 종류의 왜곡을 일부러 가한 뒤에도 정보를 복원하도록 신경망을 반복해서 훈련하는 방법이다. 쉽게 말하면, 압축이나 화면 녹화, 부분 편집처럼 영상이 실제로 유통되는 동안 흔히 생길 법한 여러 변형을 학습 단계에서 미리 겪게 하여, 그런 상황이 닥쳐도 심어 둔 정보를 끝까지 놓치지 않도록 대비시키는 것이다. 이 과정을 여러 번 되풀이하면 디코더는 변형으로 크게 흐트러진 신호에서도 원래 심어 둔 정보를 되찾아 내는 힘을 기르게 된다.

• 재인코딩과 시간축 편집에 대한 강건성

플로우마크가 견디도록 설계된 변형은 크게 두 갈래로 나눌 수 있는데, 하나는 화면의 겉모습 자체를 바꾸어 놓는 변형이고 다른 하나는 영상이 시간에 따라 흘러가는 화면들의 순서와 전체적인 길이를 바꾸어 놓는 시간축 편집이다. 화면의 모습을 바꾸는 변형에는 파일 크기를 줄이는 압축과, 화면을 자르거나 돌리고 그 크기를 바꾸어 놓는 기하 변형 등이 든다. 압축과 기하 변형은 정지된 이미지에서도 흔히 나타나는 변형으로, 플로우마크는 영상에 삽입한 신호가 이런 변형에 함께 흐트러지거나 지워지지 않도록 처리하여 정보를 지켜 낸다.

시간의 순서를 바꾸는 편집은 영상을 이루는 낱장 화면인 프레임의 순서나 개수에 손을 대는 방식으로, 정지된 이미지에는 없고 영상에서만 나타나는 변형이다. 여기에는 프레임의 앞뒤 순서를 맞바꾸는 교체, 프레임 사이에 새 프레임을 끼워 넣는 삽입, 일부 프레임을 들어내는 삭제, 그리고 프레임 수를 다시 맞추는 리샘플링(resampling)*과 보간(interpolation)**이 포함된다. 이런 시간축 편집은 영상을 한 장 한 장의 화면으로만 따로 떼어 다루는 이미지 워터마킹 방식으로는 온전히 대응하기 어려운 지점에 해당한다. 플로우마크는 워터마크의 삽입 위치를 영상의 움직임에 맞추어 정하기 때문에, 프레임의 순서와 개수가 달라지더라도 남아 있는 화면에서 심어 둔 정보를 다시 읽어 낼 수 있다.

실제 유통 과정에서 가장 흔하게 일어나는 변형은, 영상 공유 플랫폼이 이용자가 업로드한 영상을 자체 규격에 맞추어 새로 다시 압축하는 재인코딩이다. 플로우마크는 유튜브(YouTube)와 페이스북(Facebook) 같은 플랫폼의 재인코딩을 거친 영상에서도 심어 둔 정보를 되찾아 낸다. 이는 유출 영상이 여러 플랫폼을 오가며 여러 차례 거듭 재인코딩되는 상황에서도, 처음에 심어 둔 정보가 영상의 화면과 함께 지워져 사라지는 일 없이 본래의 형태를 유지한 채 고스란히 그대로 남아 있게 된다는 것을 뜻한다.

* 리샘플링(resampling): 영상의 초당 프레임 수를 다른 값으로 바꾸기 위해 프레임을 다시 뽑아내는 처리

** 보간(interpolation): 프레임 수나 화면 크기를 바꿀 때 없는 화면이나 값을 앞뒤 프레임을 참고해 중간값으로 계산하여 채워 넣는 방식

III. 결론 : 영상 출처 확인 기술의 방향

• 출처 확인 기술의 시사점과 향후 과제

결국 플로우마크가 보여 주는 것은, 영상 속 워터마크가 여러 단계의 재유통 과정을 거치고 난 뒤에도 영상 속에 그대로 남아 있을 수 있다는 점이다. 마스크 예측으로 삽입 위치를 고르고, 잡음 증강 학습으로 여러 변형에 미리 대비하며, 재인코딩과 시간축 편집을 거친 뒤에도 심어 둔 정보를 다시 되찾아 내는 일련의 흐름이 서로 맞물려 하나의 과정으로 자연스럽게 이어진다. 이는 유출된 영상에서 원본과의 연결 고리를 되짚어 확인할 실마리가, 영상 바깥에 따로 마련한 별도의 장치가 아니라 영상 그 자체의 안에 온전히 담길 수 있음을 의미한다. 이러한 점에서 영상 워터마킹은 콘텐츠가 어디에서 비롯되었는지를 영상에 붙여 다니는 정보의 형태로 밝히는 수단에 한층 더 가까워진다.

다만 이러한 기술이 실제 산업 현장에서 널리 쓰이기 위해서는, 복원 성능과는 별개로 함께 풀어야 할 과제가 여전히 남아 있다. 워터마크를 심고 읽어 내는 방식이 서비스마다 제각각이면 한 곳에서 심은 표시를 다른 곳에서 확인하기 어려워지므로, 삽입과 검증을 함께 아우르는 공통의 규격을 업계 차원에서 미리 마련해 두는 일이 무엇보다 먼저 필요하다. 또한 워터마크를 일부러 지우거나 무력화하려는 시도가 이어질 수 있어, 이러한 우회 시도에 맞서는 기술적 보완이 꾸준히 뒤따라야 한다. 따라서 영상 워터마킹은 그 자체로 완결된 해법이라기보다, 제도와 운영, 그리고 다른 여러 식별 수단과 함께 맞물려 작동할 때 비로소 제 역할을 넓혀 갈 것으로 보인다.

참고문헌

- Ernesto Van der Sar, "Take-Two Expands GTA 6 Leak Hunt With DMCA Subpoenas to X and YouTube", TorrentFreak, 2026.08.23., <https://torrentfreak.com/take-two-expands-gta-6-leak-hunt-with-dmca-subpoenas/>
- Vishal Asnani 외 2명, "FlowMark: Mask-Guided Video Watermarking", arXiv, 2026.07.06., <https://arxiv.org/pdf/2607.05261>

기술용어

순번	용어	설명
1	리샘플링 (resampling)	영상의 초당 프레임 수를 다른 값으로 바꾸기 위해 프레임을 다시 뽑아내는 처리
2	보간 (interpolation)	프레임 수나 화면 크기를 바꿀 때 없는 화면이나 값을 앞뒤 프레임을 참고해 중간값으로 계산하여 채워 넣는 방식

속도 변조에도 성능 유지하는 AI 생성 음악 판별 기술

뉴스 브리프

애플 뮤직은 2026년 내에 생성형 AI로 만든 음원에 AI 라벨을 붙이겠다고 밝혔다. 다만 플랫폼이 직접 음원을 탐지하는 것이 아니라, 음반사와 유통사가 제출 단계에서 라벨을 붙이도록 하는 방식이다. 자율 신고에 의존하는 표시 체계는 신고 누락이나 의도적 회피를 걸러내기 어려워, 표시가 붙지 않은 음원이 그대로 유통될 수 있다. 이를 보완하는 수단으로 음원의 신호를 분석해 AI 생성 여부를 판별하는 기술이 함께 논의되고 있으나, 재생 속도를 조금만 바꾸어도 판별 정확도가 떨어지는 취약점이 확인됐다. 본 보고서는 속도 변조에도 판별 성능을 유지하도록 설계된 판별 방식의 작동 원리를 살펴보고, 표시 체계와 판별 기술이 함께 작동하기 위한 조건을 분석한다.

뉴스 플러스

1. 서론: 음악 표시 체계 확산과 판별 기술의 필요성

• AI 음원 라벨 요건과 자율 신고 방식의 한계

구독형 음악 스트리밍 서비스 애플 뮤직(Apple Music)은 생성형 AI로 만들어진 음원에 '메이드 위드 AI(Made With AI)' 라벨을 붙이겠다는 방침을 밝혔다¹⁾. 라벨의 근거가 되는 정보는 애플이 지난 3월부터 운영해 온 'AI 투명성 태그(AI Transparency Tags)'다. 태그는 음반사와 유통사가 음원을 제출할 때 AI 사용을 애플에 알리기 위한 것으로, 애플과 콘텐츠 제공자 사이에서만 활용됐다. 이번 방침은 그 태그 정보를 이용자가 볼 수 있는 라벨로 바꾸는 것으로, 애플은 콘텐츠의 상당 부분이 AI로 만들어진 경우 태그를 반드시 기재할 것을 요구한다.

1) Tim Ingham, "Apple Music to slap AI labels on Suno tracks – but pushes the onus onto record companies", Music Business Worldwide, 2026.08.20., <https://www.musicbusinessworldwide.com/apple-music-to-slap-on-suno/>

주목할 점은 태그를 기재하는 책임이 플랫폼이 아니라 음반사와 유통사에 있다는 사실이다. 애플 뮤직은 라벨을 붙이기 위해 음원을 직접 탐지하지 않으며, 콘텐츠 제공자가 제작 방식을 가장 잘 알고 있다는 전제 하에 자율 신고 방식을 도입한다. 반면 다른 방식을 택한 음악 스트리밍 플랫폼도 있다. 스포티파이(Spotify)는 AI로 생성된 가상 아티스트 프로필에 ‘AI 페르소나(AI Persona)’ 배지(badge)를 표시하기로 했으며, 디저(Deezer)는 자체 기술을 사용하여 AI가 제작한 것으로 의심되는 음원을 탐지하고 이를 명확히 표기한다. AI 음악 생성 플랫폼 수노(Suno) 역시 오디오 워터마킹과 지문 인식 도입을 예고했다. 이러한 점에서 표시 체계가 실효를 갖추려면 신고 여부와 무관하게 음원 자체에서 생성 여부를 확인할 수 있는 기술적 수단이 함께 필요하다.

• 아티팩트 탐지 기술의 등장 배경과 취약점

AI 음악 판별 기술은 생성 모델이 음원을 만들 때 신호에 남기는 규칙적인 특징을 근거로 삼는다. 생성 과정에서 의도치 않게 남는 이러한 신호 특징을 아티팩트(artifact)라 하며, 특정 주파수 대역에 반복적으로 나타나는 에너지 봉우리(energy peaks)가 대표적이다. 이 아티팩트는 학습으로 얻어진 가중치가 아니라 생성 모델의 신경망(neural network) 설계에 의해 결정된다는 점에서 모델마다 고유한 표식처럼 작동한다. 스트리밍 플랫폼과 유통사가 대규모 음원을 점검해야 하는 상황에서 이 방식은 연산 부담이 적어 빠르게 작동한다는 이점을 지니며, 판별 근거를 사람이 확인할 수 있다는 점에서도 활용 가치가 있다.

문제는 간단한 후처리로 아티팩트가 다른 주파수로 밀려날 수 있다는 데 있다. 재생 속도를 바꾸거나 표본화 주파수(sampling frequency)*를 조정하면 아티팩트가 나타나는 위치가 이동하고, 고정된 주파수 대역을 살펴보도록 설계된 판별기는 그 대상을 놓치게 된다. 실제로 입력 음원의 표본화 주파수를 22.05킬로헤르츠(kHz)로 낮추는 조작만으로도 상당수 AI 음원의 판정 결과가 뒤집혔으며, 음높이 변조에서는 특정 판별기의 정확도가 1~3% 수준까지 떨어진 사례가 보고됐다. 속도 변조는 일부 음악 장르에서 창작 수단으로 쓰이고 짧은 영상에 음악을 맞추는 과정에서도 널리 사용되는 만큼, 무료 편집 도구만으로도 판별을 우회할 수 있게 된다. 이는 아티팩트가 이동하더라도 판별에 영향을 주지 않는 표현 방식이 필요함을 의미한다.

* 표본화 주파수(sampling frequency): 연속적인 아날로그 신호에서 1초 동안 추출하는 표본의 횟수

II. 본론: 속도 변조를 건디는 AI 음악 판별 방식

• 전치 합성곱 층이 남기는 스펙트럼 아티팩트의 규칙

대부분의 AI 음악 생성기는 압축된 내부 표현을 실제 소리 파형으로 확대하는 과정을 거친다. 이 확대 과정에는 전치 합성곱(transposed convolution) 층이 쓰이며, 신경망 오디오 디코더(audio decoder)*에서 표본 수를 늘리는 표준 요소로 사용된다. 전치 합성곱은 입력 특징맵의 해상도를 높이는 업샘플링(upsam-

pling) 과정에서 표본 사이에 값을 채워 넣고 필터 연산을 적용하며, 이 과정의 구조가 스펙트럼 아티팩트를 만들 수 있다. 그 결과 원래 음악에는 존재하지 않는 에너지 봉우리가 특정 주파수에 규칙적으로 나타난다.

봉우리가 어느 주파수에 나타나는지는 신경망 설계를 통해 예측할 수 있다. 전치 합성곱 층 하나만 놓고 보면, 아티팩트는 일정한 간격으로 나란히 늘어선 봉우리 묶음을 이룬다. 간격은 그 층이 내부에서 다루는 표본화 주파수로 정해지고, 봉우리의 개수는 그 층이 표본을 몇 배로 늘리는지에 따라 달라진다. 여러 층이 쌓이면 각 층이 만든 묶음이 서로 결합해, 층별 내부 표본화 주파수를 배수만큼 반영해 더할 수 있는 모든 위치마다 봉우리가 놓인 집합이 만들어진다. 이 집합의 위치는 학습에 사용한 음악이나 학습으로 얻은 가중치와 무관하게 신경망 설계만으로 결정된다. 실제로 같은 설계를 서로 다른 데이터로 학습시키거나 초기값을 바꿔 학습시킨 실험에서도 봉우리 위치는 동일하게 나타났다.

이러한 규칙성은 판별 기술이 성립하는 전제가 된다. 아티팩트 위치가 생성 모델의 설계만으로 정해지므로, 음원의 스펙트럼에서 예측 가능한 위치의 봉우리를 찾아내면 해당 음원이 AI로 만들어졌는지 판단할 수 있다. 실제로 시간 축을 따라 스펙트럼을 평균 낸 뒤 완만한 기준선을 빼내 봉우리만 남기고, 판별에 유용한 5~16kHz 대역에 한정해 분석하자, 이전 연구의 AI 생성 음원 대상 평가 조건에서 분류 정확도가 99%를 넘었다. 논문은 모델이 학습한 이 아티팩트 배열 표현을 페이크프린트(fakeprint)라 부르며, 음원마다 추출한 페이크프린트를 판별의 입력으로 사용한다. 이는 음악의 내용이 아니라 생성 과정이 남긴 신호 특징이 판별의 근거가 됨을 의미한다.

* 오디오 디코더(audio decoder): 압축되거나 부호화된 디지털 오디오 신호를 다시 사람이 들을 수 있는 아날로그 신호나 원본 디지털 신호로 변환해 주는 장치나 소프트웨어

• 속도 변조가 아티팩트 위치를 이동시키는 원리

속도 변조는 음원의 표본을 다시 뽑아내 재생 시간을 늘리거나 줄이는 조작이다. 이때 원본을 몇 배로 재생하는지를 나타내는 값이 배속 값으로, 1.0이면 원본 그대로이고 그보다 작으면 느리게, 크면 빠르게 재생된다. 배속 값이 정해지면 시간 축은 그 값만큼 늘어나거나 줄어들고, 주파수 축은 그 역수만큼 함께 조정된다. 따라서 원래 8kHz에 있던 아티팩트는 배속을 낮추면 더 낮은 주파수로, 배속을 올리면 더 높은 주파수로 옮겨간다. 이때 페이크프린트의 봉우리 배열 자체가 사라지는 것은 아니며, 봉우리 사이의 비율 관계를 유지한 채 전체가 이동한다.

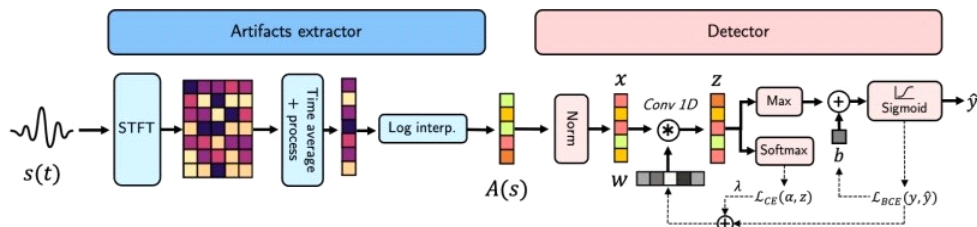
기존 판별기가 속도 변조에 무너지는 까닭은 주파수를 일정한 간격으로 나눈 축에서 아티팩트를 찾기 때문이다. 이 축에서는 주파수 전체에 배속의 역수가 곱해져 봉우리 배열이 원래 형태를 유지하지 못한다. 그 결과 판별기가 학습한 봉우리 배열과 실제 음원의 배열 사이에 불일치가 발생하게 되고, 예측 가능한 위치에 봉우리가 없다는 판정이 내려진다. 이 불일치는 음원의 신호 품질과 무관하게 발생하므로, 속도 변조는 음질 손실 없이 판별을 회피하는 수단이 될 수 있다. 이때 아티팩트가 사라진 것은 아니어서, 이를 어느 축 위에서 표현하는지가 판별 성능을 좌우하는 조건이 된다.

• 로그-주파수 축 변환과 교차상관 필터의 작동

논문이 제시한 판별 방식은 아티팩트를 로그-주파수 축(log-frequency axis) 위에서 표현하는 데서 출발한다. 로그-주파수 축은 주파수가 두 배가 될 때마다 축 위에서 같은 거리만큼 이동하도록 눈금을 매긴 축이다. 이 축에서는 주파수 전체에 배속의 역수를 곱하는 조작이 축 위의 덧셈으로 바뀌어, 페이크프린트가 형태를 유지한 채 한 방향으로 통째로 밀려나기만 한다. 판별이 어려웠던 문제가 봉우리 배열의 위치 이동을 무시하는 문제로 단순해지는 것이다.

판별은 두 단계로 나뉜다. 선행 단계인 아티팩트 추출기는 음원에 단시간 푸리에 변환(Short-Time Fourier Transform, STFT)*을 적용해 시간과 주파수에 따른 스펙트럼을 얻은 뒤, 시간 축으로 평균을 내 하나의 스펙트럼으로 압축한다. 이어서 이 스펙트럼을 옥타브당 1,920개 눈금의 로그-주파수 축으로 옮기고, 완만한 기준선을 빼내 그보다 높은 봉우리만 남긴 페이크프린트를 만든다. 후속 단계인 판별기는 페이크프린트의 극단값을 잘라내고 음원마다 다른 값의 크기를 하나의 기준으로 맞춘 뒤, 학습으로 얻은 하나의 필터와 교차상관(cross-correlation)을 계산한다. 교차상관은 두 배열을 조금씩 밀어 가며 겹치는 정도를 모든 이동량에 대해 계산하는 연산으로, 배열이 밀려 있어도 가장 잘 맞아떨어지는 지점을 찾아낸다.

[그림 1] AI 음악 판별 방식의 작동 원리



출처: Emile Eugelay 외 5명, "IMPROVED ROBUSTNESS IN AI-GENERATED MUSIC DETECTION", arXiv, 2026.07.31., <https://arxiv.org/abs/2607.27454>

교차상관 결과에서 가장 큰 값만 골라내면, 그 값이 어느 자리에서 나왔든 판정에는 차이가 없다. 페이크프린트가 몇 눈금 밀려 있으면 교차상관 결과도 같은 만큼 밀리기 때문에, 그 최대값은 변하지 않는다. 이 최대값에 학습 가능한 편향을 더한 뒤 시그모이드(sigmoid)** 함수를 통과시키면 AI 생성 여부에 대한 점수가 나온다. 학습에는 판정 오차를 줄이는 항과 교차상관의 최대값이 실제 배속에 해당하는 자리에 놓이도록 하는 항을 결합한 손실 함수가 쓰이며, 후자는 필터의 양쪽 끝 부분까지 학습이 골고루 미치게 해, 끝부분 값이 0으로 줄어드는 현상을 막는다. 학습을 마친 필터는 해당 생성 모델에서 나타나는 아티팩트 배열을 담게 되어, 음원을 대조할 기준 페이크프린트로 기능한다.

* 단시간 푸리에 변환(Short-Time Fourier Transform, STFT): 음원을 짧은 구간으로 나누어 각 구간의 주파수 성분을 계산해 시간과 주파수에 따른 스펙트럼을 얻는 변환

** 시그모이드(sigmoid): 머신 러닝과 통계학에서 주로 사용되는 S형 함수로, 보통 0과 1 사이의 값을 출력하며, 형태가 S자 모양을 닮아 시그모이드라 불림

• 판별 성능과 속도 변조 추정의 활용 범위

성능은 수노 v5, 수노 v3.5, 우디오(Udio) v120에서 생성된 음원 5,000곡과 실제 음악 5,000곡을 같은 비율로 구성한 자료를 사용해 검증됐다. 변조되지 않은 음원에서는 학습 가능한 값이 수천 개에 불과한 설계로도 기존 판별기와 대등한 수준을 보였다. 속도 변조가 적용된 음원에서는 기존 방식의 F1 점수(F1 Score)*가 수노 v5 기준 0.67 수준으로 내려간 반면, 이 방식은 0.986을 기록했다. 변조 음원을 학습에 전혀 넣지 않은 상태에서도 모든 자료에서 곡선 아래 면적(Area Under the Curve, AUC)**이 0.93을 넘어, 학습이 아니라 설계 자체로 변조에 대응했음을 보여준다.

이 방식은 판별에 그치지 않고 음원에 적용된 배속 값까지 찾을 수 있어, 사후 확인 수단으로 활용될 여지가 있다. 교차상관 최대값의 위치가 페이크프린트가 밀려난 거리에 대응하므로, 그 위치에서 배속 값을 역산할 수 있다. 수노 v5에서는 평균 절대 오차가 0.0004 수준으로 나타났다. 이는 어떤 음원이 AI로 만들어졌는지뿐 아니라 어떤 조작을 거쳐 유통됐는지까지 근거를 남길 수 있음을 의미한다. 다만 적용 범위에는 제약이 있다. 음높이 변조에서는 시간 늘림이 함께 작용해 아티팩트의 세기가 약해지므로 판정 점수가 낮아지고, 평활화(equalization), 동적 범위 압축(dynamic range compression), 코덱 압축(codec compression) 같은 다른 조작은 검증되지 않았다. 또 학습 대상이 수노 v5, 수노 v3.5, 우디오 v120 등 세 개의 생성 모델에 한정되므로 다른 생성기에는 그대로 적용되지 않아 생성기별 재학습이 필요하다.

* F1 점수(F1 Score): 판별기가 AI 음원을 놓치지 않는 정도와 잘못 지목하지 않는 정도를 함께 반영한 성능 지표로, 1에 가까울수록 두 정도가 함께 높음

** 곡선 아래 면적(Area Under the Curve, AUC): 판정 기준값을 바꿔가며 측정한 판별 성능을 하나의 값으로 요약한 지표

III. 결론: 표시 체계와 판별 기술의 보완 과제

• 표시 체계 신뢰도를 좌우하는 판별 기술의 과제

애플 뮤직의 라벨 도입은 AI 생성 음원에 대한 정보를 이용자에게 전달하는 통로를 넓혀준다. 그러나 태그 기재가 음반사와 유통사에 맡긴다는 점에서 그 실효는 신고의 정확성에 달려 있다. 결국 표시 체계가 작동하려면 신고에 의존하지 않고 음원 자체에서 생성 여부를 확인하는 수단이 함께 갖춰져야 한다. 아티팩트를 근거로 삼는 판별 기술은 그 수단이 될 수 있으나, 속도 변조 한 번으로 판정이 뒤집히는 취약점이 확인된 상태였다. 그러나 아티팩트를 로그-주파수 축에서 표현하고 교차상관으로 대조해 주파수 이동의 영향을 상쇄하면서, 조작을 거친 음원에서도 판별 성능이 유지될 수 있는 조건이 마련됐다. 이러한 점에서 주목할 부분은 판별 정확도의 상승 그 자체가 아니라, 우회 수단이 나올 때마다 하나씩 학습으로 따라잡는 대신, 설계 단계에서 우회를 차단했다는 데 있다.

다만 남은 과제가 적지 않다. 검증되지 않은 조작 방식이 여럿 남아 있고 학습 대상 생성기도 세 개에 한정돼 있어, 생성 모델이 늘어나는 속도에 맞춘 재학습 체계가 필요하다. 연산 부담이 적은 설계 덕분에 재학습 비용이 낮다는 점은 이 과제를 다루는 데 유리한 조건으로 평가된다. 아울러 사람이 만든 음원을 AI 생성물로 잘못 판정하는 경우에는 음악가에게 회복하기 어려운 손실이 발생할 수 있으므로, 판별 점수는 자동 제재의 유일한 근거가 아니라 사람의 검토를 전제로 한 보조 신호로 사용할 필요가 있다. AI를 부분적으로 활용한 음악 제작이 늘어나면서 사람이 만든 음악과 AI 생성 음악의 경계도 흐려지고 있어, 어느 수준의 AI 이용까지 표시 대상으로 삼을지에 대한 기준도 정립되어야 한다. 따라서 산업계의 대응은 판별 기술의 정확도 확보와 표시 기준의 구체화, 오판정에 대한 이의 절차 마련이 함께 진행되는 방향으로 나아갈 것으로 보인다.

참고문헌

- Tim Ingham, "Apple Music to slap AI labels on Suno tracks - but pushes the onus onto record companies", Music Business Worldwide, 2026.08.20., <https://www.musicbusinessworldwide.com/apple-music-to-slap-on-suno/>
- Emile Eugelay 외 5명, "IMPROVED ROBUSTNESS IN AI-GENERATED MUSIC DETECTION", arXiv, 2026.07.31., <https://arxiv.org/pdf/2607.27454>

기술용어

순번	용어	설명
1	표본화 주파수 (sampling frequency)	연속적인 아날로그 신호에서 1초 동안 추출하는 표본의 횟수
2	오디오 디코더 (audio decoder)	압축되거나 부호화된 디지털 오디오 신호를 다시 사람이 들을 수 있는 아날로그 신호나 원본 디지털 신호로 변환해 주는 장치나 소프트웨어
3	단시간 푸리에 변환 (Short-Time Fourier Transform, STFT)	음원을 짧은 구간으로 나누어 각 구간의 주파수 성분을 계산해 시간과 주파수에 따른 스펙트럼을 얻는 변환
4	시그모이드 (sigmoid)	머신 러닝과 통계학에서 주로 사용되는 S형 함수로, 보통 0과 1사이의 값을 출력하며, 그래프가 S자 형태를 보여 시그모이드라 불림
5	F1 점수 (F1 Score)	판별기가 AI 음원을 놓치지 않는 정도와 잘못 지목하지 않는 정도를 함께 반영한 성능 지표로, 1에 가까울수록 두 정도가 함께 높음
6	곡선 아래 면적 (Area Under the Curve, AUC)	판정 기준값을 바꿔가며 측정한 판별 성능을 하나의 값으로 요약한 지표