음원 분리 기술을 활용한 음악 식별 성능 향상

박지현, 김혜미, 김정현
한국전자통신연구원
juhyun@etri.re.kr, miya0404@etri.re.kr, bonobono@etri.re.kr

Music Identification Accuracy Enhancement through Audio Source Separation

Park Ji-hyun, Kim Jung-hyun, Kim Hye-mi
Electronics and Telecommunications Research Institute

요 약

최근 음악 생성 AI 기술의 발전으로 기존 음악의 반주에 AI 가창 음성을 믹싱하여 제작한 AI 커버곡이 빠르게 확산되고 있다. 또한 오디션, 경연, 라이브 공연 등 방송 음악 프로그램에서 실연 형태의 커버곡 활용이 증가함에 따라, 동일 반주에 다양한 가창이 결합된 콘텐츠가 늘어나고 있다. 이러한 콘텐츠가 방송에 사용될 경우, 정확한 저작권료 산정을 위해 사용된 음악을 정밀하게 식별할 수 있는 기술의 필요성이 더욱 커지고 있다. 이와 같은 환경에서는 대사와 배경음악이 혼합된 방송 콘텐츠에서 음악을 정확히 식별하거나, 동일 반주를 사용하는 서로 다른 가수의 음원을 구분하는 기술의 성능을 높이는 것이 매우 중요하다. 본 연구에서는 음원 분리(Audio Source Separation) 기술을 적용하여 입력 음원을 대사와 음악 또는 보컬과 반주 성분으로 분리한 뒤, 각 성분의 특징을 개별적으로 추출·활용함으로써 음악 식별 성능을 향상시키는 방법을 제안한다. 실험 결과, 음원 분리 기술을 활용할 경우 음악 식별 성능이 유의미하게 향상됨을 확인하였다.

Ⅰ. 서 론

최근 음악 생성 인공지능(AI) 기술의 발전으로, 기존 음악의 반주에 AI 가창 음성을 합성·믹싱하여 제작한 AI 커버곡이 빠르게 확산되고 있다. 이러한 AI 커버곡은 SNS, 스트리밍 서비스, 영상 플랫폼 등을 통해 대중적으로 소비되고 있으며, 원곡의 반주를 활용하면서도 다양한 가창 스타일을 반영할 수 있다는 점에서 새로운 음악 소비·창작 형태로 자리잡고 있다. 이와 함께 오디션 프로그램, 경연, 라이브 공연 등 방송 음악 프로그램에서도 실연 형태의 커버곡 활용이 꾸준히 증가하고 있다. 이러한 콘텐츠에서는 동일한 반주에 서로 다른 가수의 보컬이 결합된 형태가 빈번하게 등장하며, 특히 방송 편집 과정에서 대사·해설·현장음 등과 배경음악이 혼합되어 송출되는 경우가 많다.

방송이나 온라인 플랫폼에서 사용되는 음악을 정확하게 식별하는 것은 저작권 보호와 정산, 사용 이력 관리, 음원 추적 등 다양한 측면에서 필수적이다. 특히 동일한 반주를 기반으로 여러 가창 버전이 존재하거나, 대사와 음악이 혼합된 상태로 콘텐츠가 서비스되는 경우에는 기존의 원본 음원만을 사용한 음악 식별 방식만으로는 정확한 식별이 어려운 기술적 한계가 있다.

이러한 문제를 해결하기 위해서는 음원 분리(Audio Source Separation) 기술을 통해 입력 오디오 신호를 보컬과 반주 신호 또는 대사와 음악 신호로 분리하고, 각 신호의 특징을 독립적으로 추출하여 식별 성능을 높이는 접근이 필요하다. 음원 분리는 혼합 신호로 인해 손상된 음악적 특징을 복원하거나, 동일 반주 상황에서 음악 식별 성능을 저하시키는 가창자의 목소리를 제거하고 음악 반주 만을 이용하여 음악 식별을 처리함으로써 성능을 향상시킬 수 있다는 점에서 효과적인 전처리 기술로 주목받고 있다.

본 연구에서는 방송 및 커버곡 환경에서 음악 식별 정확도를 향상시키기 위해 음원 분리 기술을 음악 식별 과정에 통합하는 방법을 제안한다. 제안 방식은 대사가 혼합된 방송 음원이나 동일 반주 기반의 다양한 커버곡에 대해 보다 정밀한 음악 식별을 가능하게 하는데, 대사와 음악이 포함된 방송 데이터셋을 대상으로 한 실험을 통해 그 효과를 검증하였다.

Ⅱ. 음원 분리 기술

음악 신호는 보컬, 반주, 대사, 주변 잡음 등 다양한 성분이 혼합되어 있으며, 이를 효과적으로 분리하는 것은 음악 식별, 가창 분석, 음원 리마스터링 등 다양한 응용 분야에서 필수적인 전처리 과정으로 활용된다. 기존의 음악 분리 연구는 주로 드럼(drums), 베이스(bass), 보컬(vocals), 기타(other)의 4개 스템(stem)으로 음원을 분리하는 방향으로 진행되어 왔다. 이러한 방식은 음악 제작이나 리믹싱 등에는 적합하지만, 방송 콘텐츠 내에서의 음악 식별 문제에 직접 적용하기에는 적합하지 않다.

본 연구에서 대상으로 하는 방송 음악 식별이나 커버곡 식별에서의 성능을 향상하기 위해서는 기존의 4가지 스템 분리와는 다른 두 가지 형태의 음원 분리가 필요하다. 첫째, 방송 콘텐츠에서는 대사와 음악이 혼합된 경우가 많으므로, 음악과 대사를 분리하는 기술이 필요하다. 이때 음악에는 보컬 신호도 포함된 상태로 분리되어야 한다. 둘째, 동일한 반주에 다른 보컬이 결합된 AI 커버곡이나 실연 음악을 구분하기 위해서는 반주와 보컬을 분리하여 각각의 특징을 독립적으로 추출하는 방식이 필요하다. 이러한 분리 방식은 기존 음원 분리 연구의 목적과는 구분되는 음악 식별에 사용되는 기술적 접근이라 할 수 있다.

기존 음원 분리 기술 중 LaSAFT[1]는 Transformer 구조에 local attention을 적용하여 시간-주파수 영역의 세밀한 정보를 효과적으로 추출함으로써, 특히 보컬과 반주의 분리에 좋은 성능을 보이는 기술이다. 최근 발표된 SCNet[2]은 주파수 대역별 특성을 고려한 차별적 처리 방식을 적용한 것이 특징이다. 주파수 스펙트럼을 여러 서브밴드로 나누어, 정보량이 상대적으로 적은 중·고주파 대역에는 높은 압축률을 적용하여 효율적으로 처리하고, 정보가 풍부한 저주파 대역에는 더 많은 파라미터를 집

중적으로 할당함으로써 세밀한 특징을 정밀하게 학습한다. 이를 통해 중요도가 낮은 부분에는 계산 자원을 최소화하고, 중요한 세부 정보에 집중하는 처리 방식을 구현하였다. 이러한 구조적 최적화 덕분에 SCNet은 기존 모델들에 비해 크기가 작고 추론 속도 또한 빠르다는 장점을 가진다.

Ⅲ 음악 식별 기술

음악 식별을 위한 대표적인 오디오 특징추출 방법은 오디오의 주파수 밴드별 에너지값을 사용하는 방법과 오디오의 피크쌍을 사용하는 방법이 있다. 대규모 음원을 다루는 음원 특징은 대규모 데이터에서 통계적으로 다른 음원 특징과 구분되는 특성이 가장 중요한데, 일반적으로 주위 잡음이 심한 경우에는 에너지의 차분값을 사용하는 것이 더 좋은 성능을 보인다.

본 논문에서는 한국저작권위원회의 필터링 성능 검증에서 99.9% 이상의 식별율로 성능이 검증된 [3]의 핑거프린트 추출 방법을 사용하였다.

음악 식별 기술에서 일반적으로 사용되는 프레임 단위 검색 방식은 각 프레임별로 정밀한 비교계산을 수행하므로 정확도는 높지만 검색 시간이 오래 걸리는 단점이 있다. 인덱싱 구조를 활용하면 이러한 검색 속도의 문제를 해결할 수 있다.

음악 식별 전체 과정은 그림 1과 같다. 대사와 같은 잡음이나 커버곡 식별 성능을 높이기 위해 질의 오디오에 대해 음악 구간 검출과 음악-대사 신호 분리 후 음악 특징을 추출하여 DB에서 특징값을 검색하여 최종 식별 결과를 구하는 과정을 거친다.

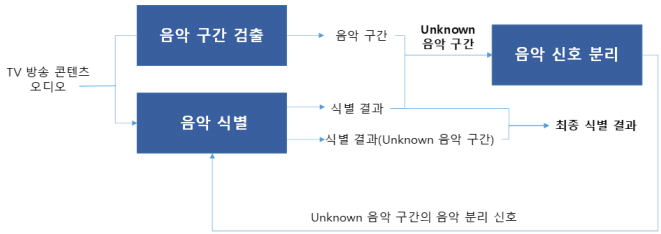

그림 1. 음악 식별 과정

Ⅳ. 성능 실험

1) 방송 배경음악 식별 실험

음악과 출연자의 대사가 중첩되어 있는 드라마와 예능과 같은 방송 콘텐츠에서의 음악 식별에서 음악 분리 전처리에 의한 정확도 개선 효과를 실험하였다. 실험 데이터는 12초 길이의 음악 데이터에 방송 콘텐츠에서 추출한 대사 오디오를 MNR(Music-to-Noise Ratio) 0dB 수준으로 합성하여 총 1,823개, 총 재생길이는 6시간의 데이터를 생성하여 사용하였다. 성능 비교를 위하여 기존 공개 음원 분리 모델 중 가장 좋은 성능을 보이는 것 중 하나인 SCNet과 음악-대사 분리 성능과 음악-대사 분리를 목적으로 LaSAFT 구조 기반으로 자체적으로 재학습한 LaSAFT-MS를 활용하였다. 실험에 사용한 1,823개의 데이터셋에 대하여 음악 분리후 SDR 측정값은 표 1에 보이는 바와 같이 SCNet 9.319, LaSAFT-MS 10.105로, 음악-대사 분리에서는 LaSAFT-MS 모델이 강점을 보인다.

표 1 음악-대사 분리 성능(SDR)

분리 모델	음악	대사
SCNet	9.319	7.038
LASAFT-MS	10.105	9.372

표 2는 테스트 데이터셋에 대하여 위에서 설명한 음악 식별 기술을 이용한 특징간 거리값을 비교한 결과이다. 표에서 보는 바와 같이 믹싱된 음원을 그대로 사용하는 것보다 음원 분리 후 식별하면 식별률이 향상됨을 알 수 있다. 그리고 음악-대사 분리에 특화된 모델을 사용한 경우 일반적인 음원 분리 모델보다 좋은 식별 성능을 보였다.

표 2 음원 분리에 따른 음악 특징간 거리값 비교

테스트 데이터	특징간 거리값
음악-대사 믹싱 오디오 (분리전)	0.280
SCNet 분리 음악	0.225
LASAFT-MS 분리 음악	0.188

2) 커버곡 식별 실험

일반적으로 커버곡은 원곡과 유사한 반주를 사용하여 다른 가창으로 부른 경우에도 단순한 템포나 조성 변경시 기존의 음악 특징으로는 식별율이 높지 않은 경우가 많다. 하지만 음악 식별 기술은 커버곡 식별 기술과 비교하면 매우 빠르게 수행할 수 있으므로, 대규모 음원 DB에서는 커버곡 식별 기술 적용전 음악식별 기술을 적용한다면 전체적인 식별 시간을 크게 줄일 수 있다. 표 3은 반주 분리후 음악 식별을 수행한 경우의 식별율 변화를 실험한 결과이다. 실험 데이터는 K-POP 음원 100곡에 대해 각 2개씩의 커버곡을 더해서 1세트로 구성하였으며 100세트, 총 300곡의 음원 데이터로 실험하였다. 실험 결과 표 3에서 보이는 바와 같이 원본 음원 DB를 대상으로 커버곡을 질의한 경우와 비교하여 원본 음원과 반주 음원의 특징을 모두 DB로 포함시키고, 커버곡 반주를 질의함으로써 식별율을 높일 수 있었다.

표 3 커버곡 식별 실험 결과

DB 구성	질의 음원	식별율
원본 음원	커버곡	66.5%
원본 음원 + 반주 음원	커버곡 반주	76.0%

Ⅴ. 결론

본 논문에서는 방송 및 AI 커버곡 환경에서의 음악 식별 정확도를 향상하기 위해 음원 분리 기술을 음악 식별 과정에 통합하는 방법을 제안하였다. 실험 결과, 방송 콘텐츠에서의 배경음악 식별에서는 음원 분리 전처리 적용 시 식별 특징 거리값이 감소하였으며, 특히 음악 - 대사 분리에 특화된 모델을 적용했을 때 기존 모델 대비 우수한 성능을 확인하였다. 또한 커버곡 식별 실험에서도 반주 분리를 통해 원본 음원과 반주 음원의 특징을 모두 DB에 포함하고, 커버곡의 반주를 질의로 사용함으로써 기존 방식보다 높은 식별률을 달성하였다. 이러한 결과를 통해 음원 분리 기술이 방송이나 커버곡과 같은 실제 환경에서 음악 식별의 정확도를 높일 수 있는 효과적인 전처리 기술임을 확인하였다.

ACKNOWLEDGMENT

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2025년도 문화체육관광 연구개발 사업으로 수행되었음 (연구개발과제명 : AI 생성 및 딥페이크 음악의 저작권 검증을 위한 핵심 기술 개발, 연구개발과제번호 : RS-2025-02216483)

참 고 문 헌

[1] Choi, W., “LaSAFT: Latent Source Attentive Frequency Transformation for Conditioned Source Separation”. ICASSP 2021, Apr 2021.

[2] Weinan Tong, “SCNet: Sparse Compression Network for Music Source Separation”. ICASSP 2024, Apr 2024

[3] 박지현, “대규모 음악 DB에서 방송 배경음악 식별을 위한 특징 추출 및 검색”, 2020년 한국방송·미디어공학회 하계학술대회, July 2020.