대규모 음악 DB에서 방송 배경음악 식별을 위한 특징 추출 및 검색

*박지현 **김정현 ***김혜미
한국전자통신연구원
*juhyun@etri.re.kr

Feature Extraction and Search for Broadcasting Background Music Identification in A Large-Scale Music DB

*Jihyun Park **Junghyun Kim ***Hyemi Kim
Electronics and Telecommunications Research Institutes

요약

최근 방송 사용 음악에 대한 저작권료 배분의 투명성을 위하여 방송음악 식별 기술에 대한 관심이 커지고 있다. 음악 DNA라 부르는 음악의 신호적 특징을 이용하는 기존의 음악식별 기술이 존재하지만, 방송 배경음악의 특성으로 인해 방송 사용 음악 식별에 그대로 활용하기는 어렵다. 방송이나 영화에 사용되는 배경음악은 우리가 일상생활에서 주로 소비하는 가요나 팝과 같은 음악과 비교하여 그 수가 매우 많고, 하나의 음악 테마에 대하여 조금씩 다르게 편곡한 유사 음악들이 다수 존재한다. 즉, 방송 배경음악을 식별을 위해서는 유사한 음악이 많은 대규모 음악 DB를 대상으로 잡음이 섞여 있는 음악을 식별하여야 한다. 한편, 대부분의 음악 식별 시스템은 빠른 검색을 위하여 모든 데이터를 메모리에 올려두고 처리하는 방식으로 동작하는데, 대규모 음악 DB를 지원하기 위해서는 시스템 자원을 적게 사용하면서도 식별율이 떨어지지 않는 특징 추출 파라미터와 인덱싱 파라미터를 찾는 것이 중요하다. 본 논문에서는 이러한 요구사항을 만족하는 배경음악 특징의 추출방법과 이 특징을 효율적으로 검색할 수 있도록 하는 검색 구조를 기술한다.

1. 서론

방송음악 모니터링은 TV 또는 라디오 방송 콘텐츠에 삽입된 음악들을 자동분석하여 방송 내에서 사용된 음악들의 사용내역을 큐시트로 생성한 후 권리자에게 제공하는 기술 및 서비스이다. 음악을 식별하는 기술 중 대표적인 것이 음악의 신호적 특성을 분석하여 생성한 음악의 특징 정보를 이용하는 핑거프린팅 기술이다. 이는 어떤 사람을 식별할 때 지문과 같이 사람마다 고유하게 지니는 정보를 이용하는 것처럼 음악을 구별할 수 있는 고유 정보를 추출하여 사용하는 방법이다. 오디오 핑거프린팅 기술은 식별율이 매우 높기 때문에 음악 검색, 세컨드 스크린, 음성 인식 등 여러 콘텐츠 서비스에서 활용되고 있다.

하지만 방송음악 모니터링을 위한 식별 기술로는 아직 널리 활용되지 못하고 있다. 방송 콘텐츠에 삽입된 음악은 대사와 같은 다른 소리와 겹쳐져서 나오는 경우가 대부분이고, 분위기 고조와 같이 시청자에게 콘텐츠의 감정을 더 잘 전달할 수 있도록 하는 보조적인 역할을 하기 때문에 작은 소리로 합성되는 경우가 많다. 이처럼 방송이나 영화에서 사용되는 배경음악은 다양한 소리와 섞이는 경우가 대부분이어서 일반적인 오디오 핑거프린팅 기술로 정확하게 식별하는 것이 매우 어렵다. 기존의 오디오 핑거프린팅 기술도 어느정도의 잡음에 강인성을 가지고 있지만, 식별하고자 하는 음악이 전경음인 경우가 아닌, 음악보다 다른 소리가 더 강한 상태에서는 매우 낮은 식별율을 보인다.

방송이나 영화에 사용되는 배경음악은 OST 음악도 있지만 대부분은 극의 분위기에 맞는 연주음악인 라이브러리 음악이 사용된다. 라이브러리 음악은 우리가 일상생활에서 주로 소비하는 가요나 팝과 같은 대중음악과 비교하여 그 수가 매우 많고, 하나의 음악 테마에 대하여 조금씩 다르게 편곡한 유사 음악들이 다수 존재한다. 따라서 특징 DB 측면에서 보면, 방송 배경음악을 식별을 위해서는 유사한 음악이 많은 대규모 음악의 특징을 DB로 구축하고 이 DB에서 잡음이 섞여 있는 음악을 식별하여야 한다.

한편, 대부분의 음악 식별 시스템은 빠른 검색을 위하여 모든 데이터를 메모리에 올려두고 처리하는 방식으로 동작하는데, 대규모 음악 DB를 지원하기 위해서는 시스템 자원을 적게 사용하면서도 식별율이 떨어지지 않는 특징 추출 파라미터와 인덱싱 파라미터를 찾는 것이 중요하다. 본 논문에서는 이러한 요구사항을 만족하는 배경음악 특징의 추출방법과 이 특징을 효율적으로 검색할 수 있도록 하는 검색 구조를 기술한다.

2. 음악 특징 추출

2.1 방송 배경음악의 특성

방송 배경음악에는 주로 라이브러리 음악이 사용된다. 라이브러리 음악은 하나의 음악테마에 대하여 템포, 악기 등을 일부 요소만을 달리

본 연구는 문화체육관광부 및 한국저작권위원회의 2020년도 저작권보호및이용활성화연구개발 지원사업으로 수행되었음

한 다양한 버전의 유사한 음악들을 포함하는 경우가 많다. 실제 방송이나 영화 콘텐츠에 음악이 삽입되는 경우에는 드라마내 대사나 다른 주위 잡음과 섞이는 경우가 많고, 음악이 끊어졌다 이어지는 경우도 있다. 따라서, 음악 특징 추출 및 정합시 다음의 특성을 고려해야 한다.

- 유사한 음악이 다수 존재
- 하나의 악기만으로 연주하므로 일부 주파수 영역만 신호가 존재
- 신호사이에 묵음 구간이 존재하거나 신호가 약한 구간이 존재
- 다른 소리와 겹치는 오디오 왜곡이 강함

2.2 차분기반 이진 핑거프린트

대표적인 오디오 특징추출 방법은 오디오의 주파수 밴드별 에너지 값을 사용하는 방법과 오디오의 피크쌍을 사용하는 방법이 있다. 두가지 방법 모두 어느정도의 주위 소음에 강인성을 보이지만, 방송이나 영화처럼 주위음이 큰 경우는 사용할 수 없다. 일반적으로 이러한 주위 잡음이 심한 경우에는 에너지의 차분값을 사용하는 것이 더 좋은 성능을 보인다. 비교해야할 음악의 수가 매우 많은 경우에는 특징의 크기를 줄이는 것이 중요한 요소 중 하나이므로 특징값을 이진화하여 간결한 형태로 만드는 것이 좋다. 본 논문에서는 이러한 핑거프린트의 요구사항을 만족하기 위하여 차분기반 이진 핑거프린트를 사용한다.

음악 신호 n번째 프레임의 m번째 부밴드의 에너지를 En,m이라고 할 때, 아래와 같이 부밴드 에너지 값을 시간과 주파수 축 방향으로 차분한 특징값 Fn,m을 구한다[1].

$$F_{n,m} = (E_{n,m} - E_{n,m+1}) - (E_{n-1,m} - E_{n-1,m+1})$$

최종적으로는 특징값 Fn,m의 부호에 따라 0과 1로 이진화하여 이진 해시 특징값 Hn,m 을 생성한다. 특징 추출시 잡음에 대한 강인성을 위해서는 프레임의 간격을 작게 하는 것이 좋지만, 프레임의 간격을 작게 할수록 특징의 수가 많아지므로 검색시간 및 메모리 측면에서는 불리하다. 따라서 최적의 프레임 간격을 찾는 것이 중요하다. 마찬가지 이유로 특징 추출을 위한 부밴드 수는 특징의 크기에 영향을 미치므로 식별율을 유지하면서도 특징의 크기를 줄일 수 있는 부밴드 수를 찾는 것도 중요하다.

본 논문에서는 부밴드 수를 16, 프레임 간격을 256을 사용하였으며, 더 큰 부밴드 수와 더 작은 프레임 간격의 특징과 비교하였을 때 식별율이 거의 동일함을 확인하였다.

일부 라이브러리 음악은 한가지 악기만을 사용하기 때문에 부밴드를 나눠 에너지를 분석하면 소수의 부밴드에서만 의미있는 에너지 값이 나타나고 나머지 부밴드는 값이 거의 없는 현상이 발생한다. 이러한 음악들에 대해서는 검색시 값이 없는 부밴드들을 제외시킬 수 있도록 특징값을 부여함으로써 식별율 향상이 가능하다.

3. 특징 DB 검색

3.1 특징 DB 및 인덱싱

특징 DB는 음악의 이진 핑거프린값과 음악정보로 구성된다. 이진화된 특징값을 사용하므로 특징간 거리 계산방법은 해밍 거리를 사용한다. 해밍 거리는 빠른 계산이 가능하므로 전체적인 검색시간을 단축하는데도 도움이 된다.

인덱스는 전체 DB 중 거리값을 계산할 후보 위치를 결정하는데 사용하는 정보이다[2]. 대규모 DB의 전체 특징에 대하여 검색하면 검색 시간이 매우 많이 소요되기 때문에 각 특징값을 참조하는 인덱스 정보를 생성하고 이를 이용하여 검색 후보군을 결정하게 된다. 특징 추출시 잡음에 강인성을 위해 특징을 매우 촘촘히 추출하였으므로 인접한 특징은 동일한 인덱스를 가질 가능성이 매우 크다. 실험결과 아래와 같이 인덱스 생성시 특징의 간격(β)을 주는 것이 검색시간과 식별율에 도움이 되었다.

$$Index_{n,m} = H_{i,j} \ (i = j/8 \times \alpha, \ j = m \times 2)$$

잡음으로 인한 왜곡으로 질의 특징의 일부 비트값이 DB와 다를 가능성이 크다. 따라서 인덱스를 사용해 후보군을 찾을 때 인덱스의 일부 비트값을 반전한 값도 함께 사용함으로써 검색율을 높일 수 있다. 반전하는 비트 수와 검색시간은 반비례하므로 본 논문에서는 최대 4개의 비트를 반전하여 사용하였다.

3.2 검색 실험

핑거프린트와 검색방법의 성능 평가를 위해 58만곡의 음악 DB를 대상으로 약 1시간 길이 드라마인 가면 1회의 배경음악 매칭율을 실험하였다. 드라마 동영상을 1초 단위의 오디오 프레임으로 나누어 검색하였으며, 프레임 검색결과를 합쳐서 구간 검색결과를 생성하도록 하였다. 매칭율은 전체 정답 구간의 길이 대비 검색된 구간의 비율로 계산하였다. 아래 표와 같이 검색율과 검색시간의 우선순위에 따라 인덱스 비트 길이와 유사 인덱스 개수를 선택함으로써 전체적인 성능의 조절이 가능하였다.

표 1 인덱스 파라미터에 따른 검색 성능

인덱스 비트 길이	16			20			24		
유사 인덱스 갯수	0	4	16	0	4	16	0	4	16
인식률(%)	86.78	90.26	91.38	82.75	86.27	89.15	81.71	84.90	87.44
검색시간(초)	246	757	2079	29	66	241	3	8	27

4. 결론

본 논문에서는 방송이나 영화 내에서 다른 소리와 겹쳐진 상태의 배경음악을 검색하기 위한 특징 추출 방법과 검색 방법에 대하여 기술하였다. 제안한 방법은 대규모 음악 DB에서 빠른 검색이 가능할 수 있도록 특징을 간결하게 만들고 효율적인 특징 인덱싱 구조를 가지도록 하였다. 본 논문에서는 질의시 동영상 오디오를 그대로 사용하였는데, 음악을 분리하여 질의한다면 더 높은 식별율을 기대할 수 있다. 향후에는 대사 등 다른 잡음을 제거하여 음악을 분리한 다음 음악을 식별하는 방법에 대한 실험을 진행할 예정이다.

참고문헌

[1] 서진수, “파워 가중치를 이용한 오디오 핑거프린트 정합”, 한국음향학회지 제 18권 제6호, 2019.

[2] J. Haitsma, “A highly robust audio fingerprinting system”, Proc. ISMIR 2002, 2002.