MPEG-7 오디오 하위 서술자를 이용한 음악 검색 방법에 관한 연구

*박만수, *박철의, *김회린, **강경옥
*한국정보통신대학교 공학부
**한국전자통신연구원 방송미디어연구부
E-mail: mansoo@icu.ac.kr

A Study on the Music Retrieval System using MPEG-7 Audio Low-Level Descriptors

*Mansoo Park, *Chuleui Park, *Hoi-Rin Kim, **Kyeongok Kang
*School of Engineering, ICU
**Electronics and Telecommunications Research Institute

요 약

본 논문에서는 MPEG-7에 정의된 오디오 서술자를 이용한 오디오 특징을 기반으로 한 음악 검색 알고리즘을 제안한다. 특히 timbral 특징들은 음색 구분을 용이하게 할 수 있어 음악 검색뿐만 아니라 음악 장르 분류 또는 Query by humming에 이용 될 수 있다. 이러한 연구를 통하여 오디오 신호의 대표적인 특성을 표현 할 수 있는 특징벡터를 구성 할 수 있다면 추후에 멀티모달 시스템을 이용한 검색 알고리즘에도 오디오 특징으로 이용 될 수 있을 것이다. 본 논문에서는 방송 시스템에 적용 할 수 있도록 검색 범위를 특정 컨텐츠의 O.S.T 앨범으로 제한하였다. 즉, 사용자가 임의로 선택한 부분적인 오디오 클립만을 이용하여 그 컨텐츠 전체의 O.S.T 앨범 내에서 음악을 검색할 수 있도록 하였다. 오디오 특징벡터를 구성하기 위한 MPEG-7 오디오 서술자의 조합 방법을 제안하고 distance 또는 ratio 계산 방식을 통해 성능 향상을 추구하였다. 또한 reference 음악의 템플릿 구성 방식의 변화를 통해 성능 향상을 추구하였다. Classifier로 k-NN 방식을 사용하여 성능 평가를 수행한 결과 timbral spectral feature들의 비율을 이용한 IFCR(Intra-Feature Component Ratio) 방식이 Euclidean distance 방식보다 우수한 성능을 보였다.

* Keyword: MPEG-7 Audio Descriptor, Timbral Spectral, Intra-Feature Component Ratio

I. 서 론

디지털 방송 서비스가 활성화 되면서 시청자에게 다양한 기능을 제공할 필요성이 제시 되고 있고 현재 지능형 TV 서비스를 위한 SmarTV와 같은 미래 지향적인 디지털 TV 시스템에 관한 연구가 활발히 진행되고 있다. 그 중에서 시청자가 원하는 정보를 효율적으로 검색할 수 있는 기능이 중요시 되고 있다. 컨텐츠의 모든 정보가 메타데이터에 서술되어 있다면 시청자에게 정보제공을 손쉽게 할 수 있다. 하지만 모든 정보를 메타데이터에 서술하기 위해서는 대부분을 수작업으로 작성해야 한다. 이러한 단점을 보완하기 위해 컨텐츠 내용 기반의 특징을 이용하여 검색을 수행하게 된다. 이 경우 검색에 필요한 템플릿 구성 및 메타데이터를 사람의 수작업 대신 자동으로 생성하게 된다. 예를 들어, 오디오 분야에서는 오디오 신호의 특징만을 이용한 음악 장르 분류[1]-[2], Query by humming[3], 그리고 음표 검출[4]과 같은 연구가 진행되고 있고, 비디오 특징과 더불어 오디오 특징의 조합을 이용한 멀티모달 시스템에 관한 연구도 활발히 진행되고 있다. 또한 이러한 검색 기능이 가능하도록 다양한 특징

에 관한 연구도 진행되고 있다. 특히 MPEG-7에서는 방송 컨텐츠의 내용 및 특징을 서술할 수 있도록 다양한 서술자들을 정의하고 있다. MPEG-7 서술자는 컨텐츠 내용 기반의 특징을 추출할 수 있기 때문에 사용자가 원하는 정보를 자동으로 검색하는데 효율적으로 이용될 수 있다.

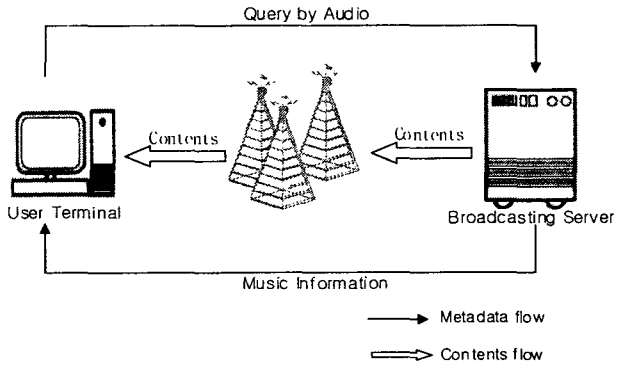

그림 1. 오디오 검색을 이용한 배경음악 정보제공 서비스

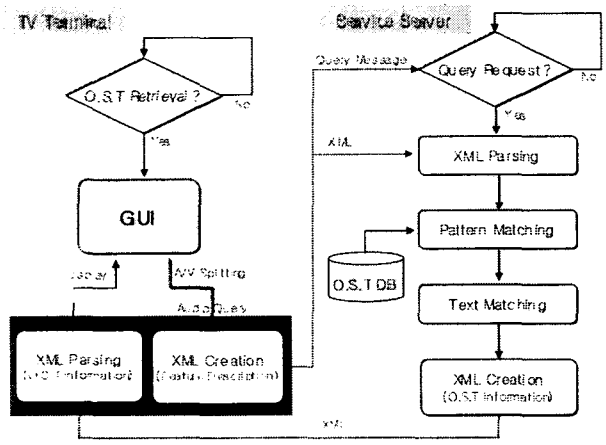

그림 2. O.S.T 검색 시스템 구조

그림 1은 오디오 쿼리를 이용한 음악 정보 제공 서비스의 한 예를 나타낸다. 시청자가 현재 방송 중인 컨텐츠의 배경음악의 정보를 원할 경우 O.S.T 앨범 내에서 그 배경음악의 정보를 시청자에게 제공하는 서비스이다. 즉, 시청자가 검색을 요청 할 경우 컨텐츠에서 배경음악에 해당하는 오디오 신호 일부를 추출하여 오디오 검색을 수행하고 그 결과에 해당하는 음악 정보를 시청자에게 제공한다. 이러한 서비스를 제공하기 위한 시스템 구조를 그림 2에서 나타내고 있다. 메타데이터에 의해 양방향 전송이 가능하기 때문에 TV 단말에서 검색요청을 위한 오디오 쿼리와 검색결과에 해당하는 배경음악 정보는 메타데이터 형태로 MPEG-7 표준화에 맞는 XML로 전송된다. 본 논문에서는 방송 서비스에 적용할 수 있는 오디오 검색 시스템을 구성하기 위해 MPEG-7 오디오 서술자[5]의 조합을 통해 오디오 신호의 특징벡터를 추출 하였다.

II. 특징벡터 구성

2.1 MPEG-7 오디오 하위 서술자

MPEG-7 오디오 하위 서술자[5]는 오디오 신호의 다양한 특징들을 표현할 수 있다. 그림 3은 MPEG-7 오디오 하위 서술자의 framework[6]을 나타낸다. 그림에서 나타나듯이 오디오 신호의 기본 파라미터 부터 음악이나 악기의 음색을 나타낼 수 있는 timbral spectral 파라미터 까지 다양하다. 총 18개의 temporal 및 spectral 서술자들을 포함하고 있고 의미적으로 8개 group로 구분할 수 있다. 여기에서 'D'는 서술자(Descriptor)를 의미한다.

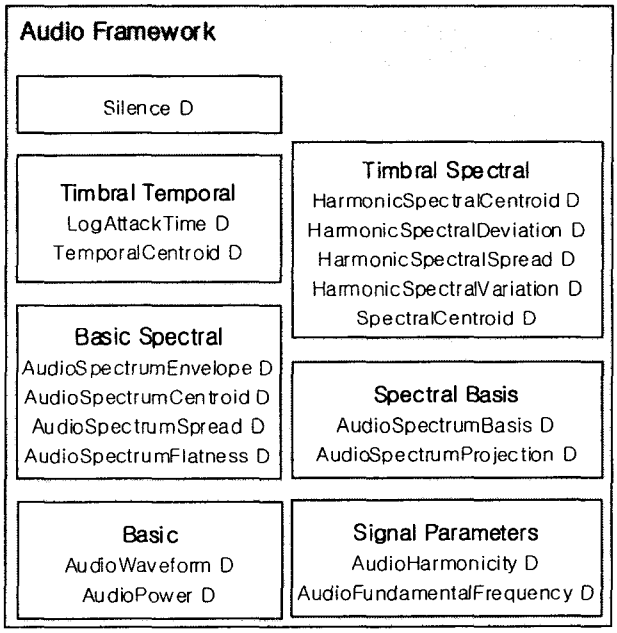

그림 3. MPEG-7 오디오 하위 서술자 framework

2.2 음악 검색을 위한 특징벡터 구성

음악 검색 서비스는 여러 장르가 섞여 있고 음악 파일 일부 클립만을 쿼리로 사용하기 때문에 클래스의 구분이 명확하지 않고 각각의 음악을 표현하기가 어렵다. 따라서 본 논문에서는 서로 다른 음악의 특성을 표현하기 위해 음색을 기반으로 하는 서술자를 이용하여 특징벡터를 구

성하였다. 그림 4는 timbral spectral[4] 특징들을 나타낸다. 파워스펙트럼의 중심 값을 나타내는 SpectralCentroid는 식 (1)과 같다. 그 외의 timbral spectral 특징은 각 프레임에 해당하는 오디오 신호의 기본 주파수와 하모닉 피크 검출을 통해 구할 수 있다. HarmonicSpectralCentroid는 식 (2)와 같이 하모닉 피크 값의 중심이 되는 주파수를 의미하고 HarmonicSpectralDeviation은 식 (3)과 같이 로그단위의 하모닉 피크의 편차 값을 나타낸다. 그리고 HarmcnicSpecralSpread는 식 (4)와 같이 하모닉 피크의 표준편차를 의미하고 HarmonicSpectralVariation은 식 (5)와 같이 시간적으로 인접한 프레임간의 variation 값을 나타낸다.

$$ISC(i) = \frac{\sum_{k=1}^{N} f(k) \cdot S(i,k)}{\sum_{k=1}^{N} S(i,k)} \quad (1)$$

$$IHSC(i) = \frac{\sum_{h=1}^{H} f(i,h) \cdot A(i,h)}{\sum_{h=1}^{H} A(i,h)} \quad (2)$$

$$IHSD(i) = \frac{\sum_{h=1}^{H} |\log_{10}(A(i,h)) - \log_{10}(SE(i,h))|}{\sum_{h=1}^{H} \log_{10}(A(i,h))} \quad (3)$$

$$IHSS(i) = \frac{\sqrt{\dfrac{\sum_{h=1}^{H} A^2(i,h) \cdot [f(i,h) - IHSC(i)]^2}{\sum_{h=1}^{H} A^2(i,h)}}}{IHSC(i)} \quad (4)$$

$$IHSV(i) = 1 - \frac{\sum_{h=1}^{H} A(i-1,h) \cdot A(i,h)}{\sqrt{\sum_{h=1}^{H} A^2(i-1,h)} \cdot \sqrt{\sum_{h=1}^{H} A^2(i,h)}} \quad (5)$$

여기에서 i 는 프레임, N 은 파워스펙트럼 사이즈, $S(i,k)$ 는 i 번째 프레임의 k 번째 파워스펙트럼 계수, 그리고 $f(k)$ 는 k 번째 파워스펙트럼 계수의 주파수를 나타낸다. $A(i,h)$ 는 i 번째 프레임의 h 번째 하모닉 피크 값을 의미한다. $SE(i,h)$ 는 i 번째 프레임의 h 번째 하모닉 피크 주변의 스펙트럼 envelope 값을 나타낸다.

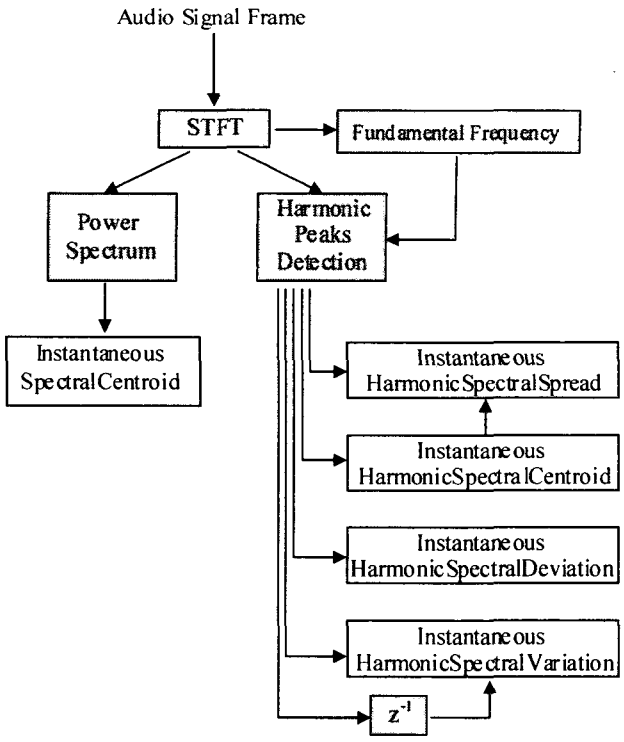

그림 4. 오디오 특징벡터 구성

III. Distance 계산

패턴 매칭을 하기 위한 distance 계산 방식인 Euclidean distance 방법은 특징벡터 성분간의 scale 차이로 인해 scale이 큰 특징 성분의 영향을 많이 받는다. 이로 인한 문제를 해결하기 위해 본 논문에서는 IFCR(Intra-Feature Component Ratio) 방식을 제안한다. IFCR은 식 (6)과 같이 특징벡터의 성분 간의 비율의 곱을 나타내어 이 값이 '1'에 가까울수록 같은 클래스에 근접한다.

$$IFCR = \prod_{n=1}^{D} \frac{\min[o_n, y_n]}{\max[o_n, y_n]} \quad (6)$$

여기에서 D는 특징벡터의 차원, o_n 는 입력(오디오 쿼리) 특징벡터의 n 번째 성분, 그리고 y_n 는 템플릿(레퍼런스 패턴) 특징벡터의 n 번째 성분을 나타낸다. 레퍼런스 템플릿(패턴)을 생성하기 위해 각 음악에 대해 texture window 단위로 특징벡터의 평균을 적용하였다.

IV. 실험 및 결과

본 논문에서 사용한 오디오 특징의 성능을 알아보기 위해 classifier로는 잘 알려진 k-NN rule을 사용하였다.

본 논문에서 제안한 음악 검색 시스템의 성능을 평가하기 위해 영화 '코요테 어글리'의 총 12곡으로 구성된 O.S.T 앨범을 사용하였다. 입력에 해당하는 오디오 쿼리는 5초에서 10초 사이의 임의의 구간을 선택하여 각 음악당 30개의 오디오 클립을 추출하여 사용하였고 레퍼런스 템플릿의 생성을 위해 10초와 15초 texture window를 5초 간격으로 overlap을 적용하였다. 표 1은 템플릿 생성을 위한 texture window의 길이에 따른 성능을 나타낸다. 오디오 쿼리의 길이가 5초에서 10초까지 가변 길이 이므로 texture window가 10초인 경우의 성능이 우수하게 나타난다.

표 1. Texture window 길이에 따른 성능변화

Distance	Texture Window	
	10 sec	15 sec
Euclidean	64.2 %	60.1 %
IFCR	81.9 %	75.8 %

표 2는 k-NN classifier의 k 값에 의한 성능 변화를 나타낸다. 이 경우 texture window 길이는 10초를 사용하였고 distance 계산은 성능이 우수한 IFCR 방식을 사용하였다. k-NN을 적용할 경우 k 값이 작을수록 우수한 성능을 보이고 있다. O.S.T의 경우 유사한 음색의 음악들이 존재하기 때문에 본 논문에서 사용한 레퍼런스 템플릿 방식의 경우는 k 값이 증가할수록 성능이 저하된다.

표 2. k-NN의 성능 변화 (Texture Window: 10 sec)

k-NN	IFCR
1	81.9 %
3	76.9 %
5	75.3 %
9	72.8 %

V. 결 론

본 논문은 오디오 신호의 일부 클립만을 사용하는 오디오 검색 알고리즘을 제안하였고 템플릿의 구성시 texture window 길이와 distance 계산 방식의 차이에 대한 실험을 수행 하였다. 표 1과 표 2에 나타나듯이 본 논문에서 제안한 IFRC 방식이 우수한 성능을 나타내었고 texture window 길이는 실제 입력 오디오 쿼리 길이에 근접할 경우 우수한 성능을 보였다.

본 논문의 결과를 기반으로 향후에 오디오 검색에 좀 더 적합한 특징 및 MPEG-7 오디오 서술자와의 조합을 통한 특징벡터 구성에 관한 연구를 지속할 것이다. 또한 오디오 레퍼런스 템플릿 구성을 다양한 VQ(Vector Quantization) 방식에 적용하고 오디오 검색에 적합한 classification 방식을 토대로 성능향상을 추구할 것이다.

Acknowledgement

본 논문은 한국전자통신연구원 "지능형 방송서비스 핵심기술 개발"에 관한 공동연구과제 수행의 일환으로 얻어진 연구결과 입니다.

Reference

[1] Yibin Zhang, Jie Zhou, "A Study On Content-Based Music Classification," *IEEE Proc. 7th International Symposium on Signal Processing and Its Applications*, vol. 2 , pp. 113-116, July, 2003.

[2] Tong Zhang, C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. On Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, May 2001.

[3] Lie Lu, Hong You, H. J. Zhang, "A New Approach to Query by Humming in Music Retrieval," ICME 2001, Aug. 2001.

[4] K. Kashino, H. Murase, " A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*, vol. 27, pp. 337-349, 1999.

[5] *Information Technology - Multimedia Content Description Interface - Part 4: Audio*, ISO/IEC FDIS 15938-4.

[6] *Overview of the MPEG-7 Standard (version 6.0)*, ISO/IEC JTC1/SC29/WG11/N4509.