Temporal 특성을 이용한 내용기반 음악 정보 검색

박철의*, 박만수*, 김성탁*, 김회린*, 강경옥**
*한국정보통신대학교 공학부, {pce337, mansoo, stkim, hrkim}@icu.ac.kr
**한국전자통신연구원 3D 미디어 연구팀, kokang@etri.re.kr

Content-based music retrieval using temporal characteristics

Chuleui Park*, Mansoo Park*, Sungtak Kim, Hoi-Rin Kim*, Kyeongok Kang**
*School of Engineering, ICU , {pce337,mansoo,stkim, hrkim}@icu.ac.kr
**3D media team, ETRI, kokang@etri.re.kr

요약

본 논문에서는 내용 기반 음악 정보 검색에 음악의 temporal 특징을 이용한 검색 방법을 제안한다. 방송 환경에 적용하기 위해 검색 범위를 드라마나 영화의 배경 음악으로 사용되는 OST 앨범으로 제한하였다 오디오의 특징 벡터로써 MFCC(Mel Frequency Cepstral Coefficient)를 사용하였으며 이 특징 벡터를 이용하여 VQ(Vector Quantization)로 부호화한 codeword로 오디오 신호의 시변 특성을 표현한다. 본 논문에서는 제안한 음악의 temporal 특성을 반영한 codeword-sequence를 이용하는 방법을 pitch-histogram을 기반으로 하는 방법 및 MFCC codeword-histogram을 기반으로 하는 방법과 비교하고 성능 개선을 보여주었다.

1. 서론

인터넷 기술의 발달로 많은 멀티미디어 데이터들을 우리는 쉽게 접하고 있다. 특히 음악 데이터들은 MP3와 대중화로 인하여 엄청난 수요와 공급이 이루어 지고 있다. 그러나, 대부분의 음악 포털 사이트나 방송국에서는 이러한 수많은 데이터들을 관리하고 운영하는 일 대부분이 아직까지는 텍스트를 기반으로 한 수작업으로 이루어 지고 있다. 따라서 이러한 단점을 보완하기 위하여 컨텐츠의 내용 기반의 특징을 이용하여 음악을 검색하고 관리하는 기술에 대해 다양한 연구가 진행되고 있다.

음악은 음의 높이, 길이, 빠르기와 같은 특성을 가지고 있다. 이러한 음악의 멜로디 특성을 잘 표현하는 것이 음악 검색에 있어서 중요한 부분이다. 보통 단음으로 이루어진 곡의 경우는 이러한 특성을 쉽게 파악할 수 있지만 실제 대부분의 경우 여러 악기와 음성으로 구성된 다중 음으로 표현되기 때문에 멜로디의 특성을 표현하기가 어렵다. 따라서, 이러한 멜로디의 특성을 나타내기 위해서 여러 기술들이 이용되고 있다. 예를 들면, 음의 비트정보, pitch 정보나 MFCC, FFT 계수, low-level audio feature를 특징 벡터로 이용하여 음악의 멜로디 패턴을 나타낸다. 이전의 연구에서는 이러한 특징 벡터의 histogram을 이용하여 멜로디의 정적 패턴을 표현하고 패턴매칭을 수행하여 음악 검색을 수행하거나 특징벡터의 조합[1]을 이용하여 표현하고 NN(Neural Network)[2], SVM(Support Vector Machine)[3], LDA(Linear Discriminant Analysis)[1]과 같은 classifier에 적용하여 음악을 검색을 수행하였다.

본 논문에서는 오디오 쿼리를 기반으로 하고 있으며 음악의 다중 음의 특성을 표현하기 위해 MFCC 특징벡터를 사용하였고 VQ코드화를 통하여 codeword sequence로써 멜로디를 표현하였다. 즉, 기존의 정적 패턴인 histogram으로 멜로디를 표현한 방식과 다르게

음악의 temporal 특성을 반영함으로써 음악의 시간적 변화에 대한 정보를 이용하였다.

2. OST 검색 시스템

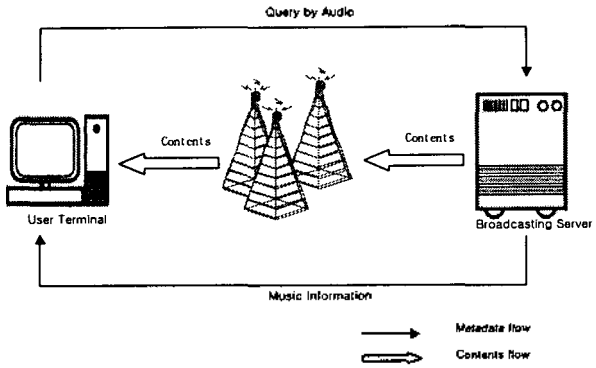

그림 1. 오디오 검색을 이용한 음악 정보제공 서비스.

그림1은 오디오 쿼리를 이용한 음악 정보 제공 시스템을 나타내고 있다. 시청자가 검색을 요청할 경우 컨텐츠에서 배경음악에 해당하는 오디오 신호 일부를 추출하여 오디오 검색을 수행하고 그 결과에 해당하는 음악 정보를 시청자에게 제공한다. 메타 데이터에 의해 양방향 전송이 가능하기 때문에 TV 단말에서 검색요청을 위한 오디오 쿼리와 검색결과에 해당하는 배경음악 정보는 메타 데이터 형태로 전송된다.

3. Baseline 시스템

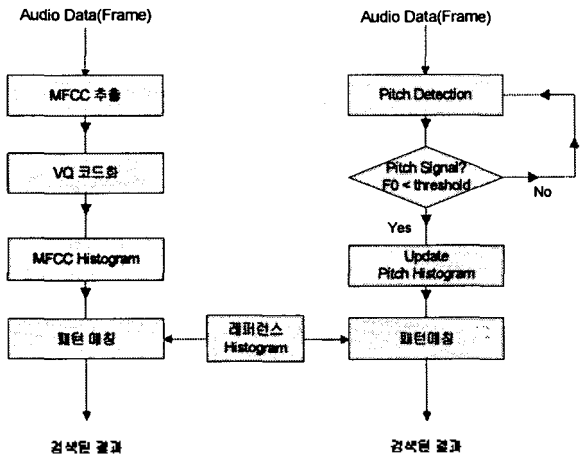

그림 2. 음악 검색을 위한 baseline 시스템
(a)MFCC histogram 을 이용한 음악 검색
(b)Pitch histogram 을 이용한 음악 검색

그림2-(a)는 MFCC를 기반으로 한 histogram 방법이고 그림2-(b)는 pitch를 기반으로 한 histogram 방법[4]이다.

MFCC histogram을 이용한 방식에서는 각 오디오 클립에 대해 40ms의 프레임 크기로 overlapping 없이 데이터의 샘플링 주파수와 그에 따른 주파수의 해상도를 고려하여 64차의 MFCC를 추출하였다. 훈련 데이터에 대해 위의 특징벡터 추출과정을 거친 후 VQ를 통해 OST 전체 곡 수와 음의 변화 정도를 고려하여 512개의 codeword로 이루어진 하나의 codebook을 얻는다. 그런 다음 각 훈련 데이터의 각 곡에 대해 1초의 hop size로 8초 길이의 클립들을 추출한다. 각 클립의 MFCC codeword histogram을 구한 후 이를 해당 곡의 레퍼런스 템플릿으로 사용한다. 테스트 클립에 대해서도 위와 같은 과정을 거쳐 histogram을 구한다. 그런 다음 레퍼런스 histogram과 테스트 클립의 histogram을 패턴 매칭하여 검색 결과를 얻는다. 그리고 Pitch histogram을 이용한 방식[4]은 pitch frequency bin에 대해서 각 클립의 histogram을 구한 다음 레퍼런스와 테스트 클립에 대해 패턴매칭을 하여 결과를 얻게 된다.

4. Temporal 특성을 이용한 시스템

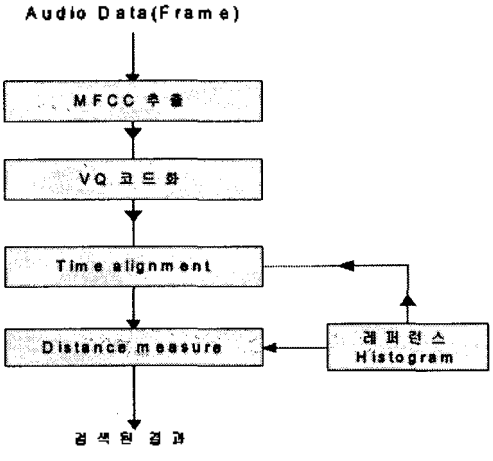

그림 3. Temporal 특성을 이용한 음악 검색

그림3에서 보듯이 temporal 특성을 이용한 검색방법은 크게 3가지의 과정으로 이루어져 있다. 즉,

특징벡터 추출 부분과 VQ 코드화를 통한 indexing부분, 그리고 패턴 매칭을 통한 검색과정으로 이루어져 있다.

4.1 특징벡터 추출

특징벡터 추출은 위의 baseline 시스템의 과정과 동일하다.

4.2 VQ 코드화를 통한 sequence indexing

위의 baseline 시스템과 같은 방식으로 생성된 codebook을 이용하여 훈련 데이터의 각 클립에 대해 프레임마다 해당되는 codeword index로 코드화한다. 테스트 클립에 대해서도 위의 과정을 거치면 비교할 두 클립의 *indexed codeword sequence*를 생성할 수 있다.

4.3 패턴 매칭

패턴 매칭은 위에서 추출한 길이가 같은 두 클립의 codeword sequence 간의 거리를 측정한다. 두 클립은 OST내의 각 곡들에 대해 1초의 hop size로 추출된 8초 길이의 클립들의 집합으로 이루어진 CD 음질의 레퍼런스 템플릿과 원하지 않는 신호가 섞인 테스트 클립 간의 패턴을 비교해야 하기 때문에 잡음에 의해 멜로디가 왜곡되는 문제가 발생한다. 또한, 두 클립의 동기 차이로 인해 거리가 왜곡되는 문제를 고려해야 한다. 따라서 본 논문에서는 이러한 두 가지 문제점을 적절히 보상할 수 있는 방법을 제안한다.

4.3.1 Time alignment

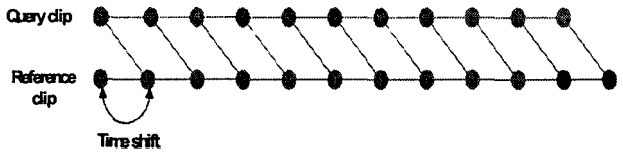

그림4. Shift를 통한 time alignment

비교하고자 하는 두 클립의 패턴 매칭에서 비슷한 멜로디를 포함하고 있다 할지라도 거리를 측정할 때 동기가 맞지 않으면 그에 따른 거리 값의 왜곡으로 인하여 잘못된 결과가 발생 할 수 있다. 따라서 위의 그림4의 경우와 같이 테스트 쿼리의 첫 프레임과 레퍼런스 클립의 전체 프레임 중 처음 N개의 프레임들을 비교하여 그 중 가장 거리가 가까운 프레임까지 레퍼런스 클립을 쉬프트 시킨다. 이렇게 함으로써 테스트 클립과 레퍼런스 클립간의 동기 차이를 어느 정도 보상 할 수 있다. (본 실험에서는 N을 15로 정의하였다.)

4.3.2 Smoothing을 이용한 거리 측정

$$WD(q,r) = \frac{1}{w}\sum_{n=1}^{N}\sqrt{(q_n - r_n)^2} \quad (1)$$

$$S = \underset{r \in R}{\operatorname{argmin}}\{WD(q,r)\} \quad (2)$$

WD : 가중된 거리값 N : 한 클립의 전체 프레임수
w : 같은 codeword 인덱스를 가지는 프레임수
q_n : 테스트 클립의 n번째 프레임의 codeword 벡터
r_n : 레퍼런스 클립의 n번째 프레임의 codeword 벡터
R : 레퍼런스 클립 전체의 집합
S : 검색된 결과

위의 time alignment 과정에서 동기를 맞춘 두 클립에 대해 ED(Euclidean 거리)방식을 단순히 적용하는 대신에 식(1)에서처럼 기존의 ED로 프레임 별 측정한 거리를 같은 codeword index를 가지는 프레임 수로 나누어주는 modified ED식을 적용한다. 따라서 잡음을 포함하고 있는 테스트 패턴와 레퍼런스 패턴를 비교할 때 전체 클립의 프레임에 대해 정확히 일치하지 않더라도 전체 프레임 중 구간구간 일치하는 부분이 발생하면 그 만큼의 가중치를 주어 나누어 줌에 따라 잡음에 따른 성능저하를 어느 정도 보상할 수 있었다.

5. 실험 결과

본 논문에서 사용한 실험 데이터는 시스템의 실제 환경에서의 성능 평가를 위해 드라마 '다모' 14부작 과 '옥탑방고양이' 16부작의 비디오 파일로부터 테스트 오디오 쿼리를 추출하였고 14곡으로 구성된 '다모' OST 앨범(CD)과 20곡으로 구성된 '옥탑방고양이' OST 앨범(CD)을 레퍼런스 템플릿으로 구성하여 성능을 평가하였다. 이 경우 테스트 오디오 쿼리와 레퍼런스는 모두 44.1kHz, 16bit, 스테레오 데이터로 구성되어 있다. 비디오로부터 추출된 테스트 오디오 쿼리의 경우 배우들의 대사와 배경 잡음이 포함되어 있어 음질이 깨끗한 음질의 레퍼런스와 다르다. 테스트 클립은

8초의 길이의 '다모' 3,613개와 '옥탑방고양이' 5,659개의 오디오 클립들로 구성되었다.

표1. 거리 측정 방법에 따른 성능변화

Contents / 거리척도	'다모'	'옥탑방고양이'
ED	66.7%	65.2%
Time alignment+ Modified ED	87.4%	90.4%

표1에서는 패턴 매칭 방법에 따른 '다모'와 '옥탑방고양이' DB에 대해 성능 평가한 결과를 나타내고 있다. 결과에서 보듯 패턴 매칭 하기 전에 전처리로써 time alignment를 수행하고, 거리 측정시 가중치를 주는 modified ED를 사용하는 방법이 전처리 없이 기존의 ED방식만 사용한 방법보다 우수한 성능을 나타내었다. 즉, time alignment를 함으로써 비교할 두 대상의 동기 문제를 보상하였고 또한 modified ED를 사용하여 테스트 오디오 쿼리에 포함된 잡음에 의해 발생하는 문제도 어느 정도 해결해 주고 있다.

표2. 시스템의 종류에 따른 성능변화

Contemts / 시스템종류	'다모'	'옥탑방고양이'
Pitch histogram	83.1%	81.3%
MFCC histogram	79.2%	83.0%
MFCC temporal	87.4%	90.4%

표2에서는 baseline 시스템과 본 논문에서 제안한 방법을 비교할 결과를 나타내고 있다. 결과에서 보듯이 음악의 일정부분에서 반복되는 확률적 특성을 이용하는 pitch histogram방식과 MFCC histogram을 이용한 방식은 비슷한 성능을 나타내고 있다. 이러한 방식은 음악에 있어서 중요한 시간적 변화 정보를 사용하지 않고 단지 음악의 정적 패턴만 이용하기 때문에 음악의 멜로디를 표현할 때 정확성이 떨어질 수 있다. 따라서, 본 논문에서 제안한 음악의 시간적 변화를 반영함으로써 멜로디 표현의 정확성을 높여 준 MFCC temporal 방법이 위의 두 방법보다 우수한 성능을 보여 주었다.

6. 결론

음악 검색을 방송 환경에서 적용하기 위해서는 멜로디를 효율적으로 표현할 수 있는 방법, 그리고 패턴 매칭시 대사나 배경 잡음에서 오는 문제점과 비교할 두 음악패턴 간의 동기의 불일치에서 오는 문제를 해결해야 한다. 본 논문에서는 temporal 특성을 반영한 codeword sequence로써 멜로디를 표현하고 검색방법으로 time alignment와 modified ED방법을 제안함으로써 위에서 언급한 문제점을 해결할 수 있는 가능성을 제시하였다. 그러나 아직도 심한 잡음이나 동기 차이가 클 때는 성능저하가 발생한다. 그러므로 잡음에 의해 특성이 변하지 않으면서 음악의 패턴을 표현할 수 있는 방법에 대한 연구가 필요하다. 또한 패턴 매칭 시 DTW와 같은 좀 더 정밀한 거리 측정에 대한 연구가 필요할 것이다.

참고문헌

1. Esmaili, S., Krishnan, S., Raahemifar, K., "Content based audio classification and retrieval using joint time-frequency analysis", Acoustics, Speech, and Signal Processing, 2004. Proceedings., vol. 5 , pp. v-665-8, May 2004.
2. Yibin Zhang, Jie Zhou, "A Study On Content-Based Music Classification", 2003. Proceedings. Seventh International Symposium on Signal Processing and Its Applications, vol. 2 , pp. 113-116, July 2003.
3. L. Lu, H.Zhang, and S. Li, "Content-based audio classification and segmentation by using support vector machines", ACM Multimedia Systems Journal 8, vol. 8, no. 6, pp. 482-492, March 2003.
4. 박만수, 박철희, 김회린, 강경옥, "Pitch 히스토그램을 이용한 내용기반 음악 정보 검색", 방송공학회논문지, 제9권, 제1호, pp. 2-8, 3월, 2004.