선형 근사를 통한 MP3 음악 요약의 성능 향상

고서영, 박정식, 오영환
한국과학기술원 전자전산학과 전산학전공

Improvement of MP3-Based Music Summarization Using Linear Regression

Seoyoung Koh, Jeongsik Park,Yung-hwan Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology
{sykoh, dionpark, yhoh}@bulsai.kaist.ac.kr

Abstract

Music Summarization is to extract the representative section of a song such as chorus or motif. In previous work, the length of music summarization was fixed, and the threshold to determine the chorus section was so sensitive that the tuning was needed. Also, the rapid change of rhythm or variation of sound effects make the chorus extraction errors. We suggest the linear regression for extracting the changeable length and for minimizing the effects of threshold variation. The experimental result shows that proposed method outperforms conventional one.

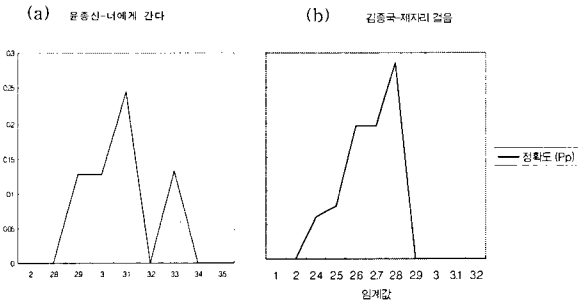
Ⅰ. 서론

최근 인터넷 상의 멀티미디어 콘텐츠가 급격히 증가하고 있다. 현재 일반적으로 쓰이는 멀티미디어 콘텐츠의 검색 방법은 텍스트로 멀티미디어 콘텐츠의 제목이나 설명을 질의하는 것이다. 그러나 멀티미디어 콘텐츠의 양과 사용자 편의성을 고려하면, 사용자가 직접 오디오나 비디오 등의 스트리밍 데이터를 검색하는 내용 기반 검색 시스템이 필요하다. 음악 데이터 역시 대부분이 디지털 형태이며, 특히 MP3(MPEG-1 Layer 3) 형태의 음악이 대중적으로 사용되고 있다. [1]

음악 요약(music summarization)이란 특정 음악의 간략한 표현으로써, 사용자에게 전체 음악의 특징 등을 효율적으로 전달하여 검색과 검색 결과의 피드백 및 결정에 도움을 주는 것을 목적으로 한다. 음악 요약은 주로 곡에서 반복적으로 나타나는 주제 선율이나 후렴구 부분(Chorus)을 추출하는 방법을 사용한다. 음악 심리학적으로 청자는 곡에서 반복적으로 나타나는 부분을 가장 인상적으로 인식하기 때문이다. [2]

본 논문에서는 사용자가 좀 더 자연스럽게 들을 수 있는 음악 요약을 위해, 선형 근사 방법을 MP3 음악 요약에 적용하여 성능을 향상시키고자 한다. 음악에서 반복되는 부분이 유사도 분석에서 선형으로 나타나는 성질을 이용하여, 미리 지정된 임계값 이하의 프레임들을 대상으로 선형 근사를 하여, 후렴구를 보다 원곡에 적합하게 추출한다. 또, 기존 방법은 곡마다 다른 임계값으로 최적화 시켜주어야 했는데, 본 논문에서는 위의 방법을 통해 임계값에 따른 요약 결과에 미치는 변화를 최소화시켜, 임계값에 민감하지 않은 안정적 결과를 얻어내고자 한다.

본 논문은 총 5장으로 구성된다. 2장에서는 관련 연구를 개괄하고, 3장에서는 본 논문에서 사용한 MP3



<그림1> 임계값 변화에 따른 요약 결과

음악 요약에 대해 다룬다. 4장에서 실험 및 결과를 제시하고, 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

음악 요약은 후렴구 추출 방식에 따라 크게 고정 길이 방식과 가변 길이 방식으로 나눌 수 있다.

고정 길이 방식은 사전에 정해진 길이 만큼 후렴구를 추출하는 방식이다. 주로 클러스터링[4][5]이나 유사도 매트릭스[3] 등을 사용한 악곡 분석을 이용한다.

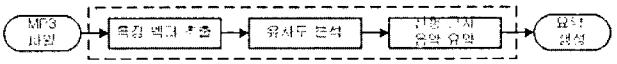
가변 길이 방식은 악곡의 구조를 분석을 한 후, 분석 결과에서 후렴구(Chorus)만을 추출하는 방식으로 SVM, GMM 등 모델 기반 방법을 이용한다.[1][6][7]

가변 길이 방식은 곡의 후렴구를 비교적 정확하게 추출할 수 있다는 장점이 있으나, 곡의 정형화된 구조를 미리 지정해야 하므로, 사전에 정의된 이 구조를 벗어나는 곡에 대해서는 정확한 분석이 어렵다.

이에 비해 고정 길이 방식은 곡의 형식이 정형화되어 있지 않더라도, 반복되는 구간을 추출할 수 있다. 이 방식의 단점은 임계값에 지나치게 민감하게 반응하여, 각각의 곡에 대해 임계값이 조절되어야 한다는 것이다. <그림1>은 고정 길이 방식을 이용할 때, 임계값의 변화에 따른 요약 결과를 보여 준다. 참고 문헌 [4][5]의 방법을 이용하여 가요를 요약했다. <그림1>의 (a)는 임계값이 3.1 일때 최상의 요약 결과를 보여 주나, (b)는 임계값이 2.8 일 때 결과가 가장 좋다. 곡마다 서로 다른 임계값을 조절하는 작업은 음악 요약 자동화에 장애가 된다.

기존에는 주로 PCM이나 MIDI 형태의 데이터를 대상으로 음악 요약이 연구되어 왔으나, 시장의 요구에 따라 MP3 음악 요약 연구도 활발해 졌다. MP3와 같이 압축된 오디오는, PCM 샘플(WAV 파일)로 변환하여 음악 요약을 생성해야 하나, 디코딩을 위해 많은 시간적 부담이 따르므로, 디코딩 과정이 없는 요약 기능이 요구된다[5][8].

본 연구에서는 형식을 알지 못하는 곡에 대해서 선형 근사를 통해 후렴구를 최대한 추출해내는 가변 길이의 요약을 생성하면서, 임계값에 민감하게 반응하지 않도록 자동적으로 보정하는 방법을 제안한다.



<그림 2> 선형 근사 음악 요약 시스템 구성도

Ⅲ. 선형근사를 이용한 MP3 음악의 요약

<그림2>는 본 논문의 음악 요약 시스템 구성도이다.

MP3 파일을 입력으로 받아, 디코딩 과정 없이 바로 특징 벡터를 추출한다. 이를 이용해 각 벡터 간 유사도 분석을 한 후, 유사도가 높은 벡터들에 대해 선형 근사를 적용해 음악 요약을 생성 한다.

1.특징 벡터 추출

본 논문에서는 MP3에서 PCM 샘플로 디코딩하지 않고, MP3자체에서 스펙트럼 포락(Spectral Envelope), 스펙트럼 중심(Spectral Centroid), 10차 MFCC를 추출한다. 특징 벡터 분석은 MP3 구조의 기본 단위인 한 개의 granule(약 13ms)에 대해 행해진다[5].

각각의 특징 파라미터들은 식(1)과 같은 형태의 특징 벡터로 실험에 이용된다. 이 특징 벡터를 프레임이라고 본 논문에서는 칭한다.

$$V_i = (MAE_i, VAE_i, MSC_i, VSC_i, MFCC_i) \quad (1)$$
$$i = 1,2...N$$

스펙트럼 포락과 스펙트럼 중심은 실제 실험에서 30 granule의 평균(MAE, MSC)과 분산(VAE, VSC)을 구해 이용한다. MFCC는 30 granule에서의 각 MDCT 계수(Modified Discrete Cosine Transform)의 평균을 통해 구해진다.

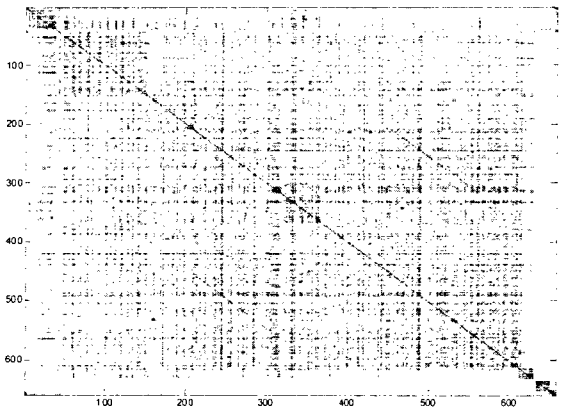
2. 프레임간 유사도 분석

유사도 분석을 위해 식(1)의 각 프레임 간의 Mahalanobis 거리를 식(2)와 같이 구한다.

$$D_M(V_i, V_j) = [V_i - V_j]R^{-1}[V_i - V_j]', \quad i \neq j \quad (2)$$

V_i, V_j는 (1)에서 구한 각 프레임의 특징 벡터이고, R은 공분산(covariance)이다.

두 프레임이 서로 유사할수록, 거리 D_M의 값은 작아지고, 유사도는 높아진다. <그림3>은 윤종신의 "너에게 간다"를 대상으로 유사도 분석한 후 D_M을 나타낸 그림이다. <그림3>의 x축,y축은 음악의 각 프레임이며, 그래프 상의 한 점은 f_x,f_y의 유사도를 나타낸다. 이 유사도 분석 매트릭스는 대각선으로 대칭이다.



<그림3> 윤종신 "너에게 간다" 의 유사도 분석

대각선 부분은 자기 자신과의 유사도를 나타낸 것이며 Mahalanobis 거리(D_M)는 0이다. D_M값이 작을수록 그림에서는 진하게 표시된다. 그림에서 중앙의 대각선에 평행하게 선들이 보이는데, 이 선이 반복구간을 의미한다. 아래쪽의 선이 약 (220,420)부터 (300,500)에 걸쳐 있다. 이는 220번째 프레임부터 300번째 프레임까지의 구간이 420번째 프레임부터 500번째 프레임 구간과 유사하며, 두 구간이 반복 구간임을 뜻한다.

3. 선형 근사를 이용한 음악 요약

<그림3>의 유사도 분석에서 보이듯이, 반복구간을 의미하는 선들은 중앙 대각선에 평행하다. 프레임 f_i (V_i)와 프레임 f_j(V_j)가 반복 구간에 포함되고, 이들의 유사도가 높다면, 프레임 $f_{i+1}(V_{i+1}),f_{j+1}(V_{j+1})$도 반복 구간에 포함되기 때문에,,이 선들의 기울기는 1이다. 이에 근거해, 다음과 같이 후렴구 구간을 찾는다. w값은 근사를 위한 버퍼값으로 요약 구간이 중간에 끊기지 않도록 해준다. $D_M(f_i, f_j)$가 임계값보다 큰 경우에도, 사전에 지정된 w값에 따라 음악 요약 구간을 계속 찾아나갈 수 있으므로, 반복 구간에서 음악의 효과음, 가수의 발성 변화 등의 음향적 특성의 사소한 차이에 의해 발생하는 끊김 현상에 대해 유연하게 대처할 수 있다. w값은 곡의 중간 요약 결과에 따라 과정 ② 에서와 같이 유동적으로 증가 혹은 감소한다. 따라서, 중간 요약 결과가 길다면, w도 커지고, 중간 요약 결과가 짧다면 w도 작아진다.

또, 임계값이 최적화되지 않더라도 , w의 값이 완충 역할을 하므로, 안정된 결과를 얻을 수 있다. 기존 방법으로는 <그림1>과 같이 임계값에 따라 성능의 차이가 크게 발생하여 임계값에 대한 최적화가 필요하다. 그러나, ②의 방법을 통해 임계값의 변화에 대해 안정된 결과값을 얻을 수 있다.

① $j = i + a$인 f_i, f_j에 대해,
while ($D_M(f_i, f_j) <$ 임계값)
$i{+}{+}, j = i + a$
end

이 때, $a,i,j \in [1,N]$ 이고, N은 특징 벡터의 전체 개수이다. $D_M(f_i, f_j)$)는 (2)의 Mahalanobis 거리로 구한다.

② ①의 과정 중 ,만약, $D_M(f_i, f_j)$ 가 임계값보다 크다면, 사전에 지정된 w값을 하나 감소시킨다. 그렇지 않다면, w값을 하나 증가시킨다. 만약 w값이 0이 된다면, 해당 과정을 멈추고, 다음 패턴으로 넘어간다.

Ⅳ. 실험 및 결과

1.실험 환경

실험을 위해 10 개의 MP3 파일을 이용하였다. 사용된 MP3 파일의 규격은 비트율(bit rate)은 128bps이고, 샘플링률(sampling rate)는 44.1kHz 이며, 모노 채널이다. 음악은 대중 가요 및 팝송을 대상으로 하였다.

본 논문에서는 고정 길이 방식 중 하나인 클러스터링 방법과 비교실험을 수행하였다.[4][5] 실험을 위한 곡은 구조를 정하지 않고 무작위로 선정하였다.

2. 품질 평가 기준

실험을 통해 추출된 후렴구와 수동으로 추출한 정확한 후렴구(ground truth)를 비교하는 것을 품질 평가 기준으로 하였다. 비교 대상으로 쓰인 후렴구(ground truth)는 가요 및 팝송 악보에서 후렴구(Chorus)로 표시된 부분을 실제 파일에서 시간을 측정했다.

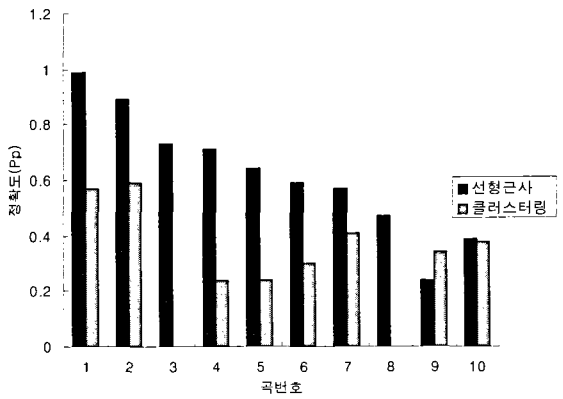
품질 평가 기준은 정확도 P_p(precision rate)로 하였다.[1][3]

$$P_p = \frac{|x \cap z_i|}{|z_i|} \qquad i = \arg max_j |x \cap z_j| \quad (3)$$

x는 실험을 통해 추출된 후렴구이며, z_i는 악보를 기준으로 추출한 후렴구(ground truth) 구간 중 x와 가

장 많이 중첩되는 구간이다.

3. 실험 결과



<표1> 선형 근사 적용 후 성능 향상

<표1>은 선형 근서 적용후 성능 향상을 나타낸 표이다. 정확도가 1에 근접할 수록 성능이 좋다. 기존 방법인 클러스터링 실험에서는 곡에 알맞게 임계값을 각각 최적화 시켜 주었다. 선형 근사에서는 일률적으로 3.3의 임계값을 주고 실험한 결과이다.

성능이 많이 향상된 곡3, 곡8의 경우, 효과음이 많이 사용된 경우로, 기존 클러스터링 방법에서 정확도가 0이었지만, 선형 근사 사용후 정확도가 각각 0.73, 0.47로 많이 향상되었다. 곡4와 곡5의 경우도, 창법에 변화가 심하거나, 효과음이 많이 사용된 경우로, 성능 향상이 많이 되었다. 곡9의 경우 성능이 오히려 하락했다. 곡9의 경우, 클러스터링 방법에서는 최적화된 임계값으로 2.7을 사용했으나, 실험에서는 3.3의 고정된 임계값을 적용했기 때문으로 추정된다. 일반적인 음악의 경우 임계값의 범위가 3~3.3 의 범위에서 결정되나 곡 9 의 경우 특별한 경우에 속한다.

전체적으로, 선형 근사를 사용했을 경우, 정확도가 평균 0.307에서 평균 0.622로 향상되었다.

V. 결론

본 논문에서는 MP3 음악 요약의 성능 향상을 위하여 선형 근사 방법을 제안하였다. 기존의 클러스터링을 이용한 고정 길이 방식은, 효과음이나 가수의 창법에 따라 정확도(Precision rate)가 안 좋게 나올 수 있었으나, 본 논문에서 제안한 방법을 통해 정확도가 0.307에서 0.622 로 향상되었다.

기존 가변 길이 방식은 곡의 형식을 알아야만 적용 가능하였으나, 본 논문의 방법은 곡의 형식과 관계없이 각 곡마다 적합한 후렴구의 길이로 음악 요약이 생성된다. 이와 더불어, 하나의 임계값을 이용해 여러 곡을 동시에 비교적 안정적으로 요약할 수 있는 장점이 있다.

참고문헌

[1] 오승은, "음악 구조를 이용한 MP3 형식의 대중 가요의 요약", 한국과학기술원 전산학 전공 석사학위 논문, 2005

[2] 박유미, "음악심리학의 이해", 음악춘추사.,

[3] Mark A.Bartsch, Gregory H.Wakefield , "Audio Thumbnailing of Popular Music Using Chroma-Based Representations",IEEE Transactions on Multimedia, Vol. 7, No. 1 ,Feb 2005

[4] Changsheng Xu, Yongwei Zhu, Qi Tian, "Automatic music summarization based on Temporal, Spectral and Cepstral features", ICME, 2002

[5] Xi Shao, Changsheng Xu, Ye Wang, Mohan S Kankanhalli,"Automatic Music Summarization In Compressed Domain", ICASSP 2004

[6] Changsheng Xu, Namunu C.Maddage, Xi Shao, Fang Cao, Qi Tian,"Music Genre Classification Using Support Vector Machines",ICASSP,2003

[7] Guodong Guo, Stan Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector Machines", IEEE Transactions on Neural Networks, 2003

[8] Chih-Chin Liu, Pang-Chia Yao , "Automatic Summarization of MP3 Music Objects", ICASSP, 2004