이동 평균 필터를 적용한 음악 세그멘테이션 및 요약

김길연, 오영환
한국과학기술원 전자전산학과 전산학전공

Moving Average Filter for Automatic Music Segmentation & Summarization

Kilyoun Kim, Yung-hwan Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology
{kykim, yhoh}@vilab.kaist.ac.kr

Abstract

Music is now digitally produced and distributed via internet and we face a huge amount of music day by day. A music summarization technology has been studied in order to help people concentrate on the most impressive section of the song andone can skim a song as listening the climax(chorus, refrain) only. Recent studies try to find the climax section using various methods such as finding diagonal line segment or kernel based segmentation. All these methods fail to capture the inherent structure of music due to polyphonic and noisy nature of music. In this paper, after applying moving average filter to time domain of MFCC/chroma feature, we achieved a remarkable result to capture the music structure.

I. 서론

음악을 디지털로 변환하여 인터넷을 통해 유통하면서 접할 수 있는 음악의 양이 방대해지고 있다. WWW상에서 스트리밍 을 통해 음악을 듣고 이를 저장하는 방법이 보편화 되었으며, 메모리 가격의 하락으로 MP3 Player에 담기는 음악의 양도 기하급수적으로 늘어나고 있다. 이에 사용자는 음악을 구입할지 결정하기 위해 전체 음악을 들어보지 않고도 그 곡의 요지를 파악하는 것이 필요하다.

음악 요약은 음악의 전체 부분 중에서 그 음악의 주제나 가장 특징적인 부분, 즉 클라이맥스(climax), 코러스(chorus) 또는 후렴(refrain)을 자동으로 추출하는 것이다. 이를 통해 사용자는 짧은 시간 내에 음악의 구조와 주제를 파악할 수 있으며, 음악 검색이나 휴대폰의 벨소리 및 통화대기음 등에도 이용될 수 있다.

본 논문에서는 MFCC 또는 Chroma 자질을 시간 영역에서 이동평균필터(moving average filter)로 스무딩하여 내재된 음악의 구조를 명확히 분석하는 방법을 제안한다. 이를 통해 기존에는 드러나지 않던 음악의 구조를 명확하게 분석할 수 있음을 확인할 수 있었다. 논문의 알고리즘은 윈도우 미디어 플레이어와 연동하여 실제 시스템으로 구현하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구와 각 연구의 문제점을 살펴보고, 3장에서 제안하는 방법이 어떻게 음악을 자동 구분하고 요약 성능을 향상시킬 수 있는지 살펴본다. 4장에서는 실제 음악을 대상으로 한 실험결과와 함께 구현된 MP3 Player를 소개하고, 5장에서 결론을 맺는다.

II. 관련 연구

음악 요약의 연구는 크게 두 가지 방향으로 구분할 수 있다. 첫째는 음악을 분석하는 자질(Feature)에 관한 것으로 MFCC, Chroma, Spectral Flux 등 주파수 영역에서의 분석 기법을 포함한다. 둘째는, 자질을 통해 분석된 음악에서 클라이맥스 부분을 자동으로 추출하는 방법으로 세그먼테이션 및 클러스터링에 관한 것이다.

첫 번째로 음악의 분석하는 자질에 관한 연구는 주로 음성인식에 사용되는 MFCC를 사용하는 것과 12음계에 해당하는 주파수 성분을 합쳐 자질로 사용하는 Chroma를 사용하는 것으로 나눌 수 있다. 음악분석에

는 Chroma 자질의 성능이 MFCC보다 좋음이 보고되어 있다[1]. 또한, 음성인식에는 사용되지 않으나 음악 분석에는 유용한 피처에 대한 연구도 있었다. 신호 크기(Signal Magnitude), 스펙트럴 중심(Spectral Centroid), 스펙트럴 롤오프 지점(Spectral Rolloff Point), 스펙트럴 유동률(Spectral Flux) 등이 사용된다[2][3]. 덧붙여 대부분의 디지털 음악은 PCM이 아닌 MP3로 압축되어 있으므로, MP3 파일에서 위의 자질을 자동으로 추출하는 방법에 대한 연구도 있었다[4]. 본 논문에서는 자질에 상관없이 시간영역에서 이동평균 필터를 사용하면 음악 구조를 명확히 분석할 수 있음을 보인다.

두 번째로 추출된 자질을 바탕으로 음악의 구조를 자동으로 분석하는 방법은 기계학습에 의한 방법과 비확률적 방법으로 유사도 행렬을 기준으로 하는 것으로 나눌 수 있다. 학습에 의한 방법은 간주와 보컬, 코러스를 나눈 후 이를 HMM, SVM을 통해 학습한 후 새로운 곡에 적용하여 음악을 세그먼트 한다[3]. 그러나 이 방법은 학습을 자동화하기가 힘들며 음악의 장르에 따른 차이로 정확도가 떨어지는 문제가 있다.

이런 문제를 해결하기 위해 비확률적 방법인 유사도 행렬로부터 세그먼트를 나누는 방법이 주로 사용된다[5]. [그림1-(a)]은 "싸이-환희" 노래에 대해 프레임별 유사도를 나타낸 것이다. 유사도는 cos distance (-1<D<1)로 색깔이 밝을수록 유사도가 높다. 유사도 행렬에서 반복되는 구간은 유사도 행렬 상에서 대각선에 평행하게 흰색 선으로 나타나게 된다. 코러스는 반복된다는 특징을 이용하여 흰색 선을 음악 요약으로 추출하는 방법에 대한 연구들이 있었다[1][6].

그러나 대부분의 음악에서 전반부와 후반부의 코러스는 그 길이 및 음정이 달라지는 경우가 많아 대각의 흰색선이 명확히 드러나지 않는다는 문제가 있다[그림1-(a)]. 이를 해결하기 위해 유사도행렬에 커널을 적용하여 Novelty Score로 일단 음악을 각 섹션으로 구분하고, 그 후 SVD와 클러스터링을 통해 수렴할 때까지 세그먼트를 합쳐서 가장 대표적인 구간으로 제시하는 방법이 제안되었다[5][7]. 그러나 현대의 음악은 여러 개의 악기, 보컬 등이 혼재되어 있어 유사도 행렬에서 각 소절을 정확히 세그먼트하기 힘들다. [그림1-(c)]는 유사도 행렬에 커널을 적용한 결과로 Novelty Score를 도시한 것이다. 그림에서 보듯이 노이즈 때문에 세그먼트가 너무 세분화되어 어떤 값을 기준으로 잘라야 할지 명확하지 않다. 이를 해결하기 위해 k-Means clustering을 통해 세그먼트를 클러스터링 하려는 시도가 있었으나[7], 그 결과가 정확하지 않다.

이에 본 논문에서는 추출된 자질에 시간 영역에서 이동평균필터를 적용하여 스무딩 효과로 음악의 내재된 구조를 명확하게 분석하는 방법을 제안한다.

III. 이동평균필터를 적용한 음악 분석

3.1 특징 추출 - 크로마(Chroma)

크로마의 사전적 의미는 광도, 채도를 뜻하지만, 음악에서는 옥타브에 불변하는 음의 속성을 칭한다. 피치 클래스(Pitch class)라고도 표현한다[1]. 예를 들어 옥타브가 3의 솔(G3)와 옥타브 4의 솔(G4)는 음 높이는 다르지만 같은 피치 클래스로 묶을 수 있다. [그림1]에서 옥타브는 "Tone Height"로 표현되고 각 음은 "Chroma"로 표현된다. 정리하면 [그림1]의 피아노 건반에서 모든 옥타브에 대해 각 음 [C, C#, D, D#, E, F, F#, G, G#, A, A#, B]의 파워 스펙트럼을 합하면 12차원의 벡터를 얻을 수 있으며 이를 크로마그램(Chromagram)이라 부른다.

본 논문에서는 16bit/16kHz로 샘플링된 음악에 대해 3~7까지의 6개의 옥타브(130Hz ~ 8kHz)에서 80ms의 프레임 사이즈로 크로마 벡터를 추출하였다.

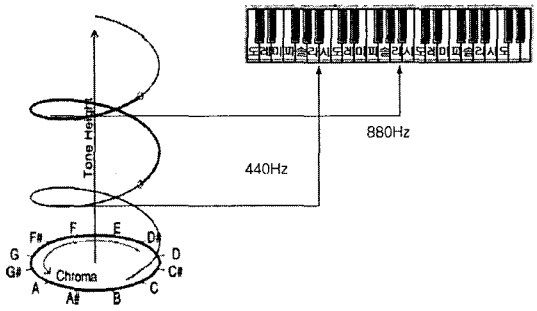

그림1. 크로마의 나선형 구조

3.2 이동평균필터의 효과

음악에서 크로마를 추출하면 한 프레임(80ms)은 12차 벡터의 열로 표현된다. k번째 프레임에 대한 크로마 벡터를 V_k라 하면 이동평균필터는 다음과 같이 정의된다. N은 윈도우의 크기이다.

$$\overline{V_k} = \frac{1}{N} \sum_{t=k-N+1}^{k} V_k \qquad (1)$$

이와 같이 적용된 프레임에 대해 유사도 행렬을 구한다. 프레임간의 유사도 행렬(Simliarity Matirx/SM)은 cos distance로 정의하였다.

[그림1-(b)]는 "싸이-환희" 음악에 대해 이동평균필터(Moving Average Filter)를 적용했을 때의 유사도

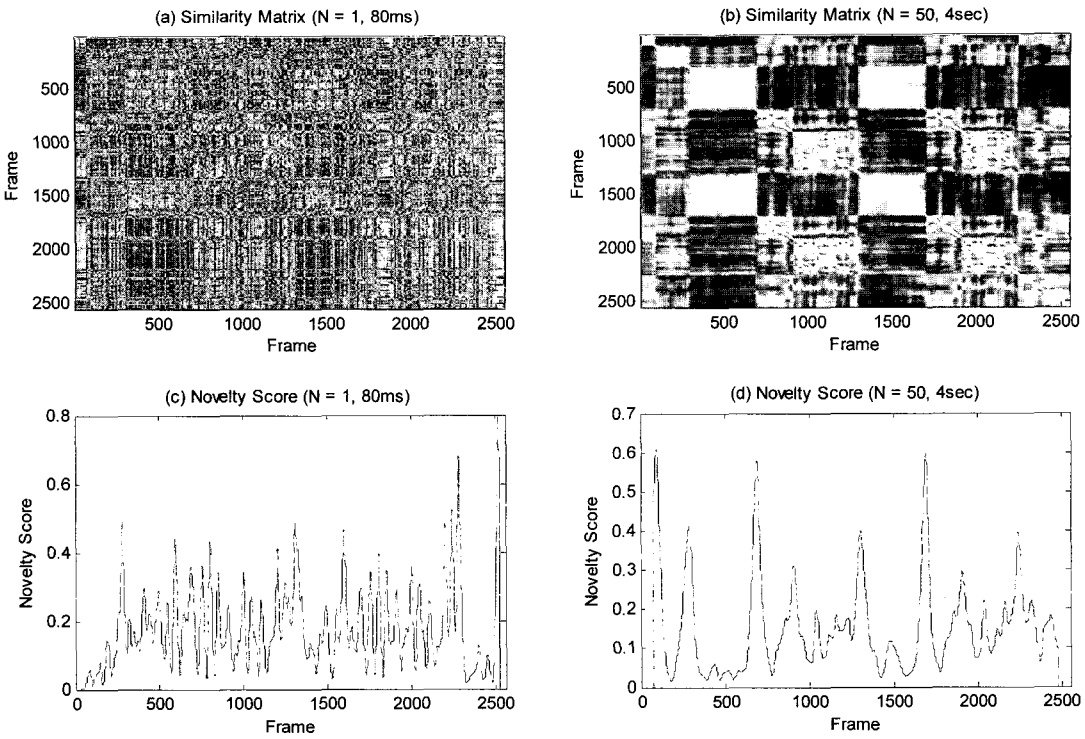

그림2. Similarity Matrix & Novelty Score.
(a),(c): 이전방법 (b),(d): 제안한 방법 (음악의 구조가 명확하게 드러남)

행렬이다. "싸이-환희" 음악은 최신 곡의 특성상 매우 시끄러운 드럼과 보컬, 랩, 변형된 전자음이 뒤섞여 있다. 이에 기존 방식의 [그림1-(a)] 에서는 각 프레임 간의 유사도가 명확히 드러나지 않으며 노이즈가 많이 섞여 있다. 그러나 (a)와 (b)를 비교해서 보면 이동평균 필터를 통해 시간영역에서 자질을 스무딩했을 때 그 전에 드러나지 않던 음악의 구조가 명확해지는 것을 확인할 수 있다. 이는 크로마 벡터 뿐만 아니라 MFCC를 사용했을 때도 마찬가지의 효과를 보인다.

3.2 음악 세그멘테이션

[그림1]의 유사도 행렬을 보면 프레임간 유사도에 따라 각 소절이 명확히 구분됨을 눈으로 확인할 수 있다. 음악 세그멘테이션을 위해 Foote가 제안한 Kernel correlation을 통해 Novelty Score(이하 NS)를 구한다 [5]. Foote의 방법은 유사도 행렬의 대각선을 따라 체크보드 행렬(커널)을 곱해 나가는 것이다. 이렇게 하면 변화가 큰 부분에서 NS의 값이 커진다. [그림1-(c)(d)]는 계산된 NS 값을 그래프로 표현한 것이다.

[그림1]에서 (c)와 (d)를 비교해 보면 이동평균필터의 효과를 다시 한 번 확인할 수 있다. (c)에서는 노이즈 때문에 각 섹션의 구분이 너무 세분화되어 명확하지 않지만, (d)에서는 NS의 피크들이 확연히 드러난다. 기존 연구에서는 성능향성을 위해 각 곡마다 다른 커널 사이즈를 사용했으나, 본 논문에서는 8초의 커널크기를 고정하여 사용하였다.

3.3 음악 요약 - 세그먼트 클러스터링

3.2절에서 구분된 음악의 세그먼트에서 각 세그먼트 간의 유사도를 분석하여 다음과 같은 과정을 통해 작은 세그먼트를 합쳐 가장 대표적인 구간을 자동으로 찾아낸다.

① 맨 처음과 끝의 세그먼트는 제외하고, 맨 처음과 끝의 세그먼트와 Kullback-Leibler Distance가 Threshold 이상이면 해당 세그먼트를 제외한다.

$$KL(N0,N1)=\frac{1}{2}(\log(\frac{\det\Sigma_0}{\det\Sigma_1})+\mathrm{tr}(\Sigma_0^{-1}\Sigma_1)+(\mu_1-\mu_0)^T\Sigma_0^{-1}(\mu_1-\mu_0)-N)$$

② 남은 세그먼트를 대상으로 Hierarchical Clustering을 수행한다.

③ 2개의 클러스터로 구분되었을 때 유사도가 가장 높은 세그먼트의 그룹을 클라이맥스 반복 그룹으로 추출하고, 이중 가장 긴 세그먼트를 클라이맥스로 추출한다.

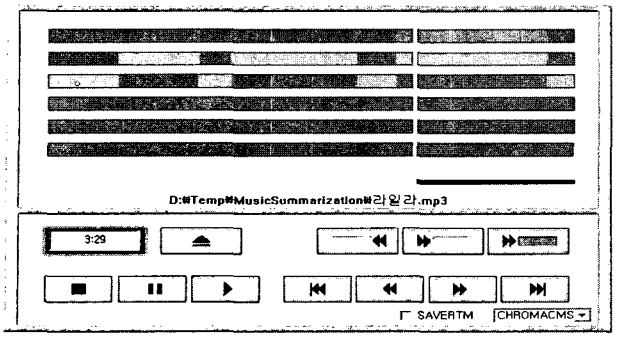

그림3. 구현된 클라이맥스 플레이어

Ⅳ. 구현 및 실험결과

[그림2]는 실제 MP3를 각 세그먼트로 구분하고 그 중 가장 대표적인 구간을 클라이맥스로 제시하는 프로그램의 구현 결과이다. 기존의 Play, Stop, Pause 버튼에 Next Section, Prev Section, Next Chorus 버튼이 추가되었다. 버튼의 윗부분에는 알고리즘에 의해 구분된 세그먼트와 클러스터를 각 라인에 구분하며, 맨 위에 빨간색으로 최종적으로 추출된 클라이맥스를 표시하였다. 이를 통해 전체 음악을 듣지 않고 클라이맥스만으로 해당 곡이 어떤 곡인지 바로 알 수 있으며, 세그멘트의 결과로 간주, 보컬, 브릿지 등 음악의 구조가 변하는 부분을 바로 찾아갈 수 있다.

실험을 위해서 2006년 4월의 인기 가요 1위~10위에 대해 음악 요약을 수행하였다. 휴대폰 통화대기음[1]으로 선정된 부분을 해당 음악을 가장 대표하는 정답으로 보고, 본 논문에서 제안한 방법으로 선정된 클라이맥스를 정답과 비교하였다. 통화대기음은 40초로 길이가 고정되어 있으므로 끝 부분의 오차는 별도로 비교하지 않고 시작부분의 오차를 측정하였다.

[표1]에 10곡에 대한 음악 요약 실험결과를 정리한다. 추출된 클라이맥스는 모두 정답보다 선행하며, 최대오차가 40초를 넘지 않는다. 따라서 40초를 기준으로 했을 때 본 논문의 알고리즘을 통해 클라이맥스의 시작부를 100% 추출할 수 있었으며, 시간 정확도면에서는 평균 6.8초의 오차를 보인다. 오차의 이유는 대부분의 음악에서 클라이맥스 바로 전의 도입부에서도 반복되는 구간이 등장하여 클라이맥스와 함께 클러스터링 되기 때문이다.

Ⅴ. 결론

본 논문에서는 이동평균필터를 음악 분석에 적용하였다. 기존의 방법들은 보컬과 여러 악기가 혼합되어 전

표1. 음악 요약 실험 결과

	통화연결음	제안방법	오차
1.SG워너비-내사람	1:28	1:27	1초
2.김종국-편지	1:33	1:30	3초
3.현영-누나의꿈	1:29	1:24	5초
4.백지영-사랑안해	1:13	0:46	27초
5.엠투엠-미라클	1:15	1:10	5초
6.바이브-그남자그여자	2:24	2:14	10초
7.SG워너비-사랑했어요	2:02	2:01	1초
8.이한철-슈퍼스타	1:28	1:26	2초
9.서영은-웃는거야	1:10	1:05	5초
10.임정희-사랑아가지마	1:37	1:28	9초

주파수 영역에서 잡음의 형태로 나타나는 음악의 특성을 제대로 모델링하지 못했으나 제안한 알고리즘을 사용했을 때 내재된 음악의 구조를 보다 명확히 분석할 수 있음을 확인할 수 있었다. 기존의 방법과 비교 실험하여 성능이 향상된 결과를 제시하며, 실제 MP3 Player와 연동하여 기능을 구현한 데모 시스템을 함께 시연한다. 향후에는 음성 구간과 악기 구간의 구분, 악구간의 경계 구분 등에 좀더 세밀한 자질을 추가하여 성능 향상에 대한 실험이 필요하다. 구현된 시스템은 WWW 스트리밍 서비스의 미리듣기에 적용하거나 벨소리 및 통화연결음을 자동으로 찾아주는 등의 다양한 응용에 사용될 수 있다.

참고문헌

[1] Mark A.Bartsch, "Audio thumbnailing of popular music Using chroma-based Representations", IEEE Transactions on Multimedia, 7(1), 2005.

[2] C. Xu, Y. Zhu, Q. Tian, "Automatic Music Summarization based on Temporal, Spectral, Cepstral Features," *ICME*, 2002

[3] 오승은, "음악 구조를 이용한 MP3 형식의 대중 가요의 요약," MS. Thesis, KAIST, 2005

[4] X. Shao, CS Xu, Y. Wang and M. Kankanhalli, "Automatic Music Summarization in Compressed Domain," ICASSP, 2004.

[5] Cooper, M., and Foote, J. "Automatic Music Summarization via Similarity Analysis", In Proc. ISMIR. 2002.

[6] M. Goto, "A chorus-section detecting method for musical audio signals," Proc. ICASSP 2003.

[7] 고서영, "MP3 음악 요약의 성능 향상을 위한 효과적인 세그먼트 구성," MS. Thesis, KAIST, 2006

1) www.ez-i.co.kr 의 통화연결음. 2006년 4월 28일 순위