내용 기반 음악 유사 구간 검색 시스템

김현우*, 한병준**, 김철환*, 이교구*
*서울대학교 융합과학기술대학원 디지털정보융합학과
**고려대학교 전기전자전파공학부
{kimbellw, monologist, kglee}@snu.ac.kr *
hbj1147@korea.ac.kr **

A Content-based Music Similarity Retrieval System

Hyunwoo Kim*, Byeong-jun Han**, Cheol-Hwan Kim*, Kyogu Lee*
*Department of Digital Contents Convergence, Seoul National University.
**School of Electrical Engineering, Korea University.

요 약

본 연구에서는 음악 데이터 베이스에서 노래의 특정 구간과 가장 유사한 구간을 검색하는 시스템을 제안한다. 제안된 시스템에서는 음악을 다차원 시계열 데이터로 간주하고, 음악의 조성 차이 및 템포(tempo) 차이를 고려한 음악의 유사도 계산 방법을 사용한다. 유사도 계산의 전처리 단계에서 조성 차이를 보정하고, 비트(beat)를 검출하며, 추출된 크로마그램(chromagram)을 검출된 비트와 동기화 하여 평균한다. 이후, 동적 시간 왜곡(DTW; dynamic time warping)을 사용하여 두 구간 사이의 유사도를 계산한 후 계산된 유사도 순서로 정렬된 검색 결과를 출력한다. 사용자는 제안된 시스템을 사용하여 선택 구간 유사도 검색과 자동 유사도 검색 결과로 도출된 구간 쌍을 검토하여 유사 구간을 보다 쉽게 찾을 수 있다.

1. 서론

최근 디지털 음원 위주로 음악 콘텐트 시장이 재편되고 다양한 형태의 음원을 다양한 모바일 장치를 통해 어디서든 쉽게 접근 할 수 있게 되었다. 이에, 음악 콘텐트 산업과 콘텐트의 규모가 폭발적으로 성장하고 있다. 한편, 음악 콘텐트 저작권자가 저작권을 수호하기 위해 음악의 유사성을 효율적으로 계산하고 검색 할 수 있는 시스템이 필요하다.

기존 검색 시스템에서 사용되는 방법은 일반적으로 음원 제작자 또는 음악 청취자가 작성한 텍스트 기반(text-annotated) 메타 데이터에 의존한다. 이러한 방법은 메타 데이터를 생성하기 위한 수작업을 필요로 한다. 그러나, 이러한 수작업 과정에서 여러 사람에 의해 생성되는 메타 데이터는 일관성(consistency)이 결여될 가능성이 높다.

이에, 본 연구에서는 음악 콘텐트 간 유사도(similarity)를 계산하는 방법을 제안한다. 제안된 방법은 비트 검출, 크로마그램 추출 및 조성 변화 보정의 전처리 단계를 거친다. 이후, 동적 시간 교정법(DTW; dynamic time warping)을 사용하여 비교 대상 곡들의 특정 구간 간의 유사도를 계산한다. 본 연구에서는 실험을 통해 음악 유사구간 검색의 가능성을 보인다.

2. 관련 연구

지금까지 음악 정보 검색(MIR; music information retrieval) 분야의 다양한 문제를 해결하기 위한 다양한 음원 유사도 계산 방법이 제안되었다. 과거의 유사도 연구는 음원의 형태에 따라 심볼릭 음악 또는 오디오 음악을 대상으로 하는 유사도 검색 알고리즘으로 분류할 수 있으며, 유사도 연구가 적용된 연구 분야로 리메이크 곡 식별 문제가 있다.

2.1 심볼릭 음악 유사도 검색 알고리즘

심볼릭 음악(symbolic music)은 문자 또는 악보 등으로 표현된 음악 콘텐트를 일컫는다. MIDI 파일은 심볼릭 음악의 대표적인 예이다. 심볼릭 음악 유사도 검색을 위한 다양한 입력 방법을 사용한 음원 검색 방법이 제안되었다. 모종식 등[1]은 음악을 일정한 선율을 따르는 음들의 배합으로 정의하고, 음고(pitch) 및 음의 길이 정보를 사용하여 두 곡 간의 유사도를 계산하는 알고리즘을 제안하였다. 한편, 박정일 등[2]은 제안한 정합 및 이동 변환을 지원하는 유사 모델에서 멜로디 특징의 인덱싱 기법을 사용하여 표절 음악을 감지하는 기법을 보였다.

2.2 오디오 음악 유사도 검색 알고리즘

Kurth 와 Müller 는 오디오 일치(audio matching) 문제를 정의하고 해결하려 시도하였다[3]. 오디오 일치 문제는 음악 데이터베이스에서 주어진 짧은 질의 음원 클립으로 음악적으로 일치하는 음원 부분을 검색하는 문제이다. Kurth 와 Müller 는 [3]에서 CENS 특징, 대각 일치(diagonal matching), 그리고 특징 코드북(feature codebook) 등을 제안하였으며, 퍼지(fuzzy) 방법론에

기반하여 음원 간 유사도를 정의하고 음원 데이터베이스에서의 검색 문제를 해결하고자 하였다.

기존 연구에서는 음원의 유사도 계산을 위해 음원으로부터 추출된 MFCC 와 같은 특성을 가우시안 혼합 모델(GMM)화 하여 Kullback-Leibler(KL) divergence 등과 비교하여 왔다. 하지만 Jensen 은 연구에서는 기존의 GMM 및 KL divergence 를 사용할 경우, 여러 악기가 조합된 음원 신호에서 볼륨이 낮은 악기 신호를 반영하지 못하며, 같은 음악에 대한 서로 다른 악기가 조합된 경우 유사도의 신뢰도가 떨어짐을 실험을 통하여 밝혀냈다[4].

최근의 West 및 Cox 는 기존의 음원 유사도 계산 방법에서 벗어나, 음악 장르 분류기를 훈련하고, 이를 사용하여 음악 유사도를 계산하는 방법을 제안하였다[5]. 또한 Casey 와 Slaney 는 음원들 간의 유사한 구간의 교차 정도를 계산하는 응용으로 일정 길이의 연속된 오디오 특성값을 shingle 로 정의하고 LSH(Locality Sensitive Hashing)를 이용하여 확장성을 고려한 유사한 이웃을 검색하는 방법이 제안되었다[6].

2.3 리메이크 곡 식별 문제

한편, 최근 리메이크 곡 식별(cover song identification) 문제는 음악정보검색 연구자 커뮤니티에서 Music Information Retrieval Evaluation eXchange (MIREX)의 한 과제가 될 정도로 주목 받는 이슈가 되고 있다.. Serrà *et al.*[7]은 이 문제를 해결하기 위해 크로마 이산 유사도(chroma binary similarity)를 정의하고 DTW 에 국소 정렬(local alignment)을 사용하였다. Ellis 등[8]은 동적 계획법(Dynamic Programming)을 이용한 비트 추적 방법 및 DTW 와 크로마그램을 이용한 방법을 제안하였다.

3. 내용 기반 유사 음악 검색 시스템

3.1 시스템 구성

제안된 시스템에서는 음원 비교를 위해 다음과 같은 순서로 전처리 과정을 거친 후 유사도를 계산한다.

먼저, 음악의 화음 특성을 반영하는 특성인 크로마그램(chromagram)을 추출한다. 다음으로, 추출된 크로마그램을 기반으로 조성이 서로 다른 곡을 조옮김(key transposition)하여 조성을 일치시킨다. , DTW 로 음원의 거리를 측정한다. 서로 다른 템포를 가진 음원 간 거리를 정확하게 계산하기 위하여 고정 시간 크기로 추출된 크로마그램의 값을 비트 단위로 재보간(reinterpolation)한다[7]. 이후, 중복된 결과를 제거하여 유사도 순으로 결과를 출력한다.

3.2 크로마그램 및 DTW 를 이용한 거리 계산

3.2.1 크로마그램(chromagram)

크로마그램은 멜로디 및 전반적인 반주의 화음을 표현하는 특성값이다[9]. 화음 인식, 리메이크송 인식 문제[8][16], 음악 구조 분석[10], 음악 분류 문제[12] 등에 다양하게 사용되고 있으며 음색 및 악기 등에 영향을 적게 받는 특성값으로 알려져 있다. 또한 음악 분류 문제[11] 에서 FPH(a folded PH)로 불리는

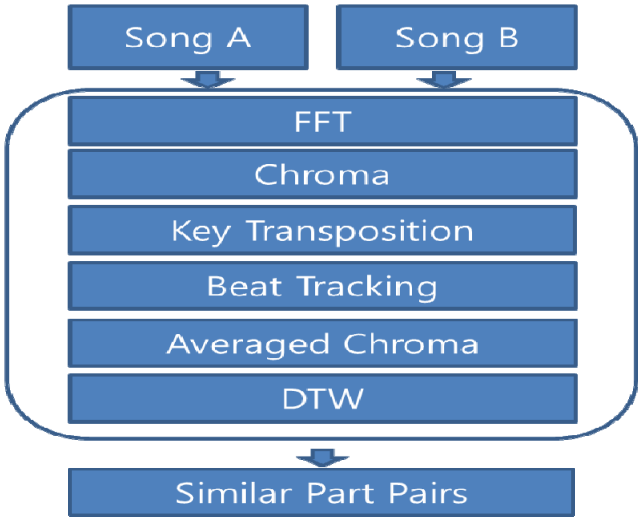

그림 1. 시스템 구성도

특성값을 사용하였는데 크로마그램과 유사한 형태이다. 크로마그램은 주파수 영역의 히스토그램을 음의 높이로 표현한 뒤, 12개의 클래스로 사상시킨 형태로 조성을 표현하는 적절한 특성값이다.

3.2.2 동적 시간 왜곡 알고리즘

동적 시간 왜곡 알고리즘(Dynamic Time Warping)은 시계열 데이터의 유사도 계산 및 인덱싱, 분류 등에 사용되는 방법으로 음성인식 분야에서 오래 전부터 사용되어 왔다[13]. 동적 시간 왜곡 알고리즘은 Berndt 와 Clifford 에 의해서 데이터마이닝 커뮤니티에 소개되었고[14] 음성 인식, 이미지 분류, 음악의 유사도 계산 등에 다양한 분야에 널리 쓰이는 방식으로 최적 경로 계산을 위해 동적 계획법(Dynamic Programming)을 이용한다. 본 논문에서는 두 비교 구간의 거리를 코사인 함수를 이용하여 유사도 행렬을 만든 후 동적 시간 왜곡 알고리즘의 최적 경로의 비용으로 계산한다.

3.3 실험방법

유사도 구간 검색 시스템은 선택 구간 유사도 검색과 자동 유사도 검색이 가능하다. 선택구간 검색은 우선 질의 곡에서 특정 구간을 선택한다. 선택 구간 유사도 검색 시스템은 선택된 구간의 비트 수를 계산하여 비트 수 크기의 구간으로 대상 곡에서 검색한다. 검색 결과를 유사도 순으로 정렬한 후 중첩이 있는 구간을 제거하고 결과를 출력한다

자동 유사도 검색은 구간의 길이 및 한 비트당 크로마그램의 개 수를 지정한 후 지정된 길이의 모든 쌍의 유사도를 계산한다. 유사도 순으로 정렬하고 중첩된 구간을 제거하고 결과를 출력한다.

3.4 조성 및 템포를 고려한 전처리

3.4.1 조옮김

유사한 곡이 다양한 조성으로 연주될 된다. 하지만 조성을 일치 시키지 않을 경우 크로마그램 사이의 코사인 거리를 이용할 경우 유사성을 적절하게 판단하

기 어렵다. 조성을 일치시키기 위하여 최적 조 옮김 (Optimal Transposition Index) 방법이 제안되었다[7]. $\vec{h}_A, \vec{h}_B$ 는 각 곡의 크로마그램의 평균을 구한 조성 프로파일 벡터이다. N_H 는 크로마그램의 차원수를 의미한다. $circhift_R(\vec{h}, id)$ 는 $\vec{h}$ 벡터를 id 만큼 오른쪽으로 회전시키는 함수이다. $corr$ 는 두 벡터 사이의 상관계수를 구하는 함수를 의미한다. 식 1 은 두 벡터 사이의 거리를 상관계수를 이용하였으며 이것은 논문[3]의 최적 조옮김과 유사하다.

$$OTI(\vec{h}_A, \vec{h}_B) = \underset{0 \le id < N_H - 1}{argmax} \{corr(\vec{h}_A, circhift_R(\vec{h}_B, id))\}$$

식 1. 최적 조옮김

위와 같은 조옮김 방법은 간단하고 효과적이다. 하지만, 여전히 조성 불일치 문제는 존재 할 수 있다. 예를 들면 CNBlue 의 '외톨이야' 경우 곡의 중간에 조성이 변하는 조바꿈이 있다. 이와 같은 경우 곡 전체의 조성 프로파일에 영향을 미치고 적절하게 조성을 일치시키는 것에 실패할 수 있다. 이를 해결하기 위해 곡의 일부분 만을 이용하여 최적 조옮김을 할 수 있다. 본 논문에서는 조바꿈이 있는 경우 곡의 일부분 만을 이용하여 조옮김을 하였다.

3.4.2 템포를 고려한 비트 동기화 크로마그램

음원의 구간의 유사도를 측정하기 위하여 사용하는 DTW 알고리즘은 서로 다른 길이의 유사도를 측정하는데 적절한 거리 측정 기준이 될 수 없다[15]. 또한 비트 추출 없이 고정 길이로 비교할 때 템포가 크게 차이가 날 경우 비교를 원하는 구간 이외의 구간까지 비교하게 되어, 유사도 측정에 나쁜 영향을 미치게 된다. 이를 해결하기 위해 정확한 템포 추출과 비트 추적(Beat Tracking)을 하고 통해 고정 시간 길이 마다 추출된 크로마그램을 비트당 일정 개수의 크로마그램을 갖도록 재보간을 한다. 이를 통해서 서로 다른 템포의 곡의 유사도 비교가 가능하다. 템포가 다른 곡을 비트 추적을 통해 각 비트 마다 크로마그램을 평균하여 리메이크 곡 인식(cover song identification)의 성능을 향상시켰다[8].

3.5 데이터 셋

곡 번호	곡명	가수	조성	템포 (bpm)	측정된 템포 (bpm)
1	외톨이야	C.N.Blue	Dm, E♭m	105	104.90
2	파랑새	Ynot	Dm	102	102.04
3	Heart Breaker	G-Dragon	Cm	135	135.14
4	Right Round	Flo Rida	Am	125	125
5	If I could fly	Joe Satriani	Bm	130	130
6	Frances limon	Los Enanitos Verdes	Em	130	126
7	가질 수 없는 너	마야	D♭	80	80.21
8	가질 수 없는 너	뱅크	B♭	80	76.92
9	Maria	김아중	B	160	159.57
10	Maria	Blondie	A	160	159.57

표 1. 데이터 셋 세부사항

4. 실험결과

질 의			검색 결과 상위 4			유사도
곡번호	시작(초)	끝(초)	곡번호	시작(초)	끝(초)	
1	57	75	2	160.816	179.028	O
			2	57.284	75.516	O
			2	179.62	197.856	O
			2	0	17.864	O
3	14	28	4	29.956	45.316	O
			4	7.344	22.756	X
			4	46.76	62.12	O
			4	106.28	121.636	O
5	49	63	6	50.944	65.704	O
			6	34.756	49.512	O
			6	294.252	308.992	O
			6	187.072	201.864	O
7	71	93	8	107.8760	130.4440	O
			8	234.8760	257.4800	X
			8	52.5640	75.1240	X
			8	166.3000	188.8960	X
9	43	67	10	47.616	71.66	O
			10	170.852	194.904	X
			10	218.928	242.976	O
			10	108.856	132.904	O

표 2. 선택구간 유사도 검색 결과 상위 4

선택 구간 검색 결과를 분석하면 질의 구간과 연관성이 있는 유사한 구간이 검색되었다. 하지만 "가질 수 없는 너"의 경우 비트 추적의 결과가 지역적으로 실패하였다. 또한 자동 유사도 검색 결과는 상위 다섯개의 결과가 연관성이 있었다. 하지만 "If I could fly" 와 "Frances Limon"는 검색에 결과가 연관성이 없었다. 서로 다른 코드에 같은 멜로디를 연주하는 부분이 질의로 입력된 유사한 구간이다. 이 구간은 각 곡이 Bm 의 Phrygian scale 과 Em 의 Aeolian scale 을 사용하여 서로 다른 조성 및 반주에서 같은 음의 멜로디를 연주한다. 이런 점이 다른 음원과 차이를 보인다.

곡 번호 (A,B)	A		B		유사도
	시작 (초)	끝 (초)	시작 (초)	끝 (초)	
1,2	111.056	128.76	74.928	93.164	O
	56.76	74.476	74.928	93.164	O
	6.484	24.192	181.384	199.604	O
	9.332	27.048	134.344	152.592	X
	111.056	128.76	0	17.864	O
3,4	8.272	22.04	25.156	40.036	O
	8.272	22.04	118.76	133.64	X
	8.272	22.04	56.36	71.24	X
	19.82	33.596	139.396	154.28	O
	75.82	89.596	110.116	125	O
5,6	4.284	18.624	7.156	21.904	X
	3.416	17.7	166.168	180.912	X
	302.472	316.78	15.244	30	X
	75.392	89.704	10.008	24.764	X
	4.776	19.088	98.08	112.84	O
7,8	200.644	223.888	220.08	244.232	O
	170.644	193.896	64.232	88.376	O
	92.644	115.896	220.08	244.232	O
	62.64	85.892	64.232	88.376	O
	170.644	193.896	188.896	213.06	O
9,10	54.044	65.676	106.984	118.624	O
	90.04	101.676	106.984	118.624	O
	141.8	153.424	47.616	59.256	O
	52.176	63.808	57	68.632	O
	129.8	141.424	170.852	182.504	O

표 3. 자동 유사도 검색 결과 상위 5

5. 결론 및 향후 과제

커버송 인식 문제를 해결하기 위하여 다양한 방법이 제안되었고 성능 또한 향상 되어 왔다. 유사 구간 검색문제는 커버송 인식 문제와 유사하지만 다소 차이가 있다. 지역적 유사성 검색이 필요하다. 본 논문의 시스템은 음악의 표절 또는 특정 구간의 참신성 등을 판단하는 참고 자료로 쓰일 수 있다. 또한 무수히 많은 음원들이 발매되는 상황에서 유사한 구간이 있는 음원들을 찾고 검색하는 것은 막대한 시간과 비용이 요구된다. 이런 비용은 유사 구간 검색 시스템의 지속적인 연구를 통해 줄일 수 있을 것이다. 또한 음반 제작 과정에서 유사 음원들을 쉽게 검색하여 제작 후 표절 시비 또는 저작권 문제로 생기는 손실을 줄일 수 있을 것이다. 이를 위해서는 유사 구간 검색을 대용량 데이터에서 가능하게 하기 위한 검색 방법의 계산 복잡도를 줄이는 연구가 필요하다. 또한 화성 이외의 다양한 특성 멜로디, 음색 등을 고려해 음원들의 유사도를 다양한 각도에서 구하는 것에 대한 연구 또한 필요하다.

참고문헌

[1] 모종식, 김소영, 구경이, 한창호, 김유성, “선율의 음높이와 리듬 정보를 사용한 음악의 유사도 계산 알고리즘,” 한국정보처리학회 논문지, 7 월 12 호, pp.3762-3774, 2000 년 12 월.

[2] 박정일, 김상욱, “음악 데이터베이스를 이용한 음악 표절 감지 시스템 개발,” 멀티미디어 학회지, 8 권 12 호, pp.1-8, 2005 년 1 월.

[3] Frank Kurth and Meinard Müller, “Efficient index-based audio matching,” IEEE Trans. On ASLP, vol.16, no.2, Feb. 2008.

[4] Jesper Højvang Jensen, Mads Græsbøll Christensen, Daniel P. W. Ellis, and Søren Holdt Jensen, “Quantaitive analysis of a common audio similarity measure,” IEEE Trans. On ASLP, vol.17, no.4, pp.693-703, May. 2009.

[5] Kris West and Stephen Cox, “Incorporating cultural representations of features into audio music similarity estimation,” IEEE Transansactions on Audio, Speech, and Language Processing, vol.18, no.3, pp.625-637, Mar. 2010.

[6] M. Casey and M. Slaney, “Song intersection by approximate nearest neighbor search,” in Proc. ISMIR, Victoria, BC, Canada, pp.144–149, 2006.

[7] Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra, “Chroma biniary similarity and local alignment applied to cover song identification,” IEEE Trans. On ASLP, vol.16, no.6, pp.1138-1151, Aug. 2008.

[8] D.P.W. Ellis and G.E. Poliner, "Identifying Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking," IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.

[9] M. A. Bartsch and G. H. Wakefield, “To catch a chorus: Using chroma-based representations for audio thumbnailing,” In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, New York, Oct. 2001.

[10] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, “Content-based music structure analysis with applications to music semantics understanding,” In Proc. ACM MultiMedia, pages 112–119, New York NY, 2004.

[11] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Transactions on speech and audio processing, vol. 10, no. 5, pp. 293-302, 2002.

[12] D. Ellis, “Classifying music audio with timbral and chroma features,” Dins Proc. ISMIR, 2007.

[13] Sakoe, H. & Chiba, S., “ Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-26, 1978.

[14] Berndt, D. & Clifford, J. “Using dynamic time warping to find patterns in time series,” AAAI-94 Workshop on Knowledge Discovery in Databases. Seattle, Washington, 1994.

[15] C.A. Ratanamahatana and E. Keogh, “Everything you know about dynamic time warping is wrong,” Third Workshop on Mining Temporal and Sequential Data, Citeseer, 2004.

[16] K. Lee, "Identifying cover songs from audio using harmonic representation," extended abstract submitted to Music Information Retrieval eXchange task.