크로마 레벨 표현, 동적 시간 왜곡, 꺾인 거리함수에 기반한 멜로디 사이의 유사도 개발

*장달원, 박성주, 장세진, 이석필
전자부품연구원
*dalwon@keti.re.kr

Development of melody similarity based on chroma representation, dynamic time warping, and hinge distance

Jang, Dalwon, Sung-Ju Park, Sei-Jin Jang, and Seok-Pil Lee
Korea Electronics Technology Institute

요약

이 논문에서는 쿼리-바이-싱잉/허밍 (Query-by-singing/humming, QbSH) 시스템 또는 커버 노래 인식 (cover song identification) 시스템에서 사용 가능한 멜로디 유사도를 제안한다. QbSH 또는 커버 노래 인식은 디지털 음악의 사용이 보편화되면서 음악 검색의 방법으로 많은 연구가 진행되어 오고 있다. 멜로디 유사도는 이런 시스템을 구현하는데 필수적인 요소이며, 두 개의 음악에서 멜로디가 추출되었다고 가정하고, 추출된 멜로디 사이의 유사한 정도를 수치로 표현한다. QbSh 시스템이나 커버 노래 인식 시스템은 멜로디 유사도에 기반하여 입력 노래와 유사한 노래를 데이터베이스에서 검색하는 작업을 수행한다. 이 논문에서 제안하는 멜로디 유사도 방식은 기존의 많이 연구되던 동적 시간 왜곡 (dynamic time warping, DTW) 방법과 크로마 표현 방법 (chroma representation)을 사용하였다. DTW방법은 비대칭적으로 사용하고 미디 노트 영역에서 표현된 멜로디 특징은 0이상 12 미만의 크로마 레벨로 표현하였다. 기존의 방법에서는 정수값을 많이 사용하였으나 이 논문에서는 실수값을 사용한다. DTW 에 사용하는 거리 함수를 기존에 사용하던 차이의 절대값 대신 꺾인 함수 형태를 사용함으로써 성능을 높였다. QbSH 시스템에서의 실험을 통해서 성능을 검증하였다. 본 논문에서는 10-12초 길이의 1000번의 쿼리 (Query)에 대해서 28시간 정도의 데이터베이스에서 실험한 결과, 순위 역의 평균 (Mean reciprocal rank, MRR) 값이 0.713을 보였다.

1. 서론

디지털 음악의 유통과 사용이 많아지면서 대용량의 음악 데이터베이스 (database, DB)에 저장된 음악 정보들을 효율적으로 정리하고 검색하는 기술에 대한 많은 수요가 있어왔다. 특히 최근 스마트폰과 스마트 TV의 보급은 음원 시장에 대한 쉬운 접근을 가능하게 하였고, 그에 따라서 디지털 음악의 구입이 활성화되고 있다. 그런 이유로 음악 정보 검색 (Music information retrieval, MIR) 분야가 많은 관심을 받고 있다 [1,2]. 이 분야에는 쿼리-바이-싱잉/허밍 (QbSH) [3-8], 장르 분류 (music genre classification)[9], 멜로디 표기 (melody transcription) [10], 오디오 인식 (audio identification) [11], 커버 노래 인식 (cover song identification) [12] 등의 연구분야가 존재한다. 이런 연구들은 일반적으로 음원 제작자 또는 음원 소비자가 만든 텍스트 기반 메타 데이터에 기반하여 이루어지던 기존 검색 방법을 뒤집는 것으로 자동적으로 메타 데이터를 생성하고 검색이 이루어질 수 있게 한다.

일반적으로 QbSH 시스템 또는 커버 노래 인식 시스템은 입력된 노래로부터 멜로디를 추출하는 부분과 추출한 멜로디로부터 멜로디 사이의 유사도를 계산하여 검색하는 부분으로 이루어진다. QbSH 시스템은 사용자의 허밍 또는 싱잉 입력에 대해서 그것에 맞는 노래를 데이터베이스에서 찾아주는 것이고 커버 노래 인식 시스템은 같은 노래를 다른 버전으로 부른 것을 찾아주는 것이다.

이 논문에서 제안하는 멜로디 유사도는 크로마 표현, 동적 시간 왜곡 (dynamic time warping, DTW), 그리고 꺾인 거리 함수에 기반을 두고 있다. 멜로디 유사도는 미디 노트 영역에서 표현된 두 개의 멜로디를 입력으로 받고, 두 가지의 유사한 정도를 수치로 표현한다. 입력된 두 개의 멜로디를 크로마 레벨로 표현하고[12], 그 값을 DTW를 사용해서 거리를 구한다. 거리를 구할 때 사용하는 거리 함수는 기존의 차이의 절대값이 아닌 특정값 이상의 거리가 나오면 그 특정값으로 거리를 설정하는 꺾인 거리 함수를 사용하였다 [5,6,8].

논문은 다음과 같이 구성된다. 2장에서는 제안하는 멜로디 유사도 방법에 대해서 설명하고, 3장에서 실험결과를 보이고 4장에서 결론을 보이면서 마친다.

2. 멜로디 유사도

본 논문에서는 DTW를 기반으로 한 거리를 멜로디 유사도의 척도로 사용한다. 즉, 거리가 큰 두 개의 멜로디는 유사도가 적은 것이고, 반대로 거리가 작은 두 개의 멜로디는 유사도가 큰 것이다. 두 멜로디 사이의 거리를 $d_M(.)$, 두 멜로디 사이의 거리를 DTW를 이용해서 구한 것을 $d_{DTW}(.)$, 두 개의 멜로디를 각각 $\vec{p}$, $\vec{q}$라고 했을 때 거리는 다음과 같이 표현할 수 있다.

$$d_M(\vec{p},\vec{q}) = \min_c d_{DTW}(\Psi(\vec{p}+c),\Psi(\vec{q}))$$

이 때, c는 보정 상수이며, 함수 $\Psi(.)$는 크로마 레벨로 표현하는 역할을 한다. 즉 두 멜로디 벡터에 대해서 음의 높낮이에 대한 보정을 c라는 보정 상수를 더해서 보정한 후에 크로마 레벨로 표현하고, 그것을 DTW를 이용해서 거리를 구하는 것이다. 하지만, 여기서 정확한 c 값을 해석적으로 구하지 못 하고 여러 후보값들을 사용하여 최저의 거리를 찾는 방법을 사용한다. 우리의 실험에서는 0이상 11이하의 정수값을 c의 후보값으로 사용하였다.

멜로디는 미디 노트 영역의 값을 가지고 있다고 가정한다. 일반적으로 미디 노트값을 정수값을 가지나 이 논문에서는 실수값도 인정한다. 즉, $\vec{p}$와 $\vec{q}$는 실수값을 가지는 시퀀스라고 생각할 수 있다.

DTW는 [6]에서와 같이 비대칭적으로 (Asymmetric) 구현되었다. DTW 패스를 결정하고, 패스에 따라 나온 거리 값을 $d_{DTW}(.)$ 값으로 결정한다. DTW를 이용해서 거리를 구하기 위해서는 각 시퀀스의 원소 사이의 거리가 정의되어 있어야 한다. 이는 일반적으로 절대값의 차이나 또는 그의 제곱 형태를 많이 사용하였다. 그러나 이 논문에서는 그것을 변형하여 아래와 같은 거리를 사용하였다. 두 개의 실수값 a, b가 있을 때 이 둘 사이의 거리는 다음과 같이 표현된다.

$$d^{(\lambda)}_{HINGE}(a,b) = \begin{cases} |a-b| & \text{if } |a-b| < \lambda \\ \lambda & otherwise \end{cases}$$

이는 두 값 차이의 절대값이 특정 상수보다 작을 시에는 그 값을 그대로 사용하고 큰 경우에는 특정 상수의 값으로 치환하는 것이다. 이 거리 함수는 꺾인 형태를 가지게 된다. 상수 λ값은 실험을 통해서 가장 좋은 값을 선정하여야 하며, 본 시스템에서는 $\lambda = 2$의 값을 사용하였다.

크로마 레벨로 표현하는 함수 $\Psi(.)$는 입력 시퀀스의 각 원소를 12로 나눈 나머지를 구해준다. 미니 노트 영역에서는 한 옥타브 올라갈 때마다 12만큼의 값이 상승하게 된다. 12로 나눈 나머지를 구하면 각 옥타브에서의 음의 높이를 구할 수 있다. 이 값은 0이상 12미만의 값을 가진다.

3. 결과

멜로디 유사도를 검증하기 위해서 QbSH 시스템에서 이를 실험하였다. 실험에 사용한 QbSH 시스템은 MIDI가 아닌 MP3와 같은 음악에 기반으로 하고 있으며, 신호로부터 64ms 프레임당 하나의 피치 값을 구하여 피치 시퀀스를 추출하였다. 사용자 입력으로부터 피치 추출은 시간과 주파수 조합영역에서의 선형 보간법을 이용해서 구해진다 [13]. 다음 음악으로부터 피치를 추출하는 방법은 평균 하모닉 구조 (average harmonic structure, AHS)에 기반을 두었다 [14,15].

실험에 사용한 DB에는 28시간 분량의 총 450곡의 노래가 들어가 있다. 여기에는 동요, 댄스, 팝, 트로트 등 다양한 장르의 노래가 들어가 있으며, 노래에 따라서 약 34초에서 396초 길이의 다양한 노래가 들어있다. 우리는 노래의 제목을 찾는 것을 목표로 하였으며, 오디오 인식에서와 같이 노래에서의 위치를 찾는 것은 목표로 하지 않았다 [11]. 테스트 데이터에는 총 1000개의 쿼리 입력이 들어가 있으며 각각은 약 10초에서 12초 정도의 길이를 가진다. 이 입력은 32명의 지원자에게서 모은 것이며 지원자는 14명의 여자와 18명의 남자로 구성된다. 1000개의 입력에는 298개의 허밍 입력과 702개의 싱잉 입력이 존재한다.

성능측정은 4가지 요소를 사용하였다. 먼저 1개의 노래제목을 결과로 냈을 때 정답이 검출되는 확률, 10개에서 검출되는 확률, 20개에서 검출되는 확률, 그리고 순위 역의 평균 (Mean reciprocal rank, MRR)을 사용하였다. MRR은 QbSH 시스템의 성능을 평가하기 위해서 많이 사용되는 지표이다.

MRR	Top1	Top 10	Top 20
0.713	0.668	0.808	0.843

표 1. QbSH 시스템에 사용하였을 때의 결과

최종적인 성능은 표 1. 에 나타난 것과 같다 0.713의 MRR 값을 얻할 수 있었는데, 이것은 [8] 논문에 나와 있는 427곡의 DB에서 0.578의 MRR 값보다 훨씬 우월한 것이다.

4. 결론

본 논문에서는 QbSH 시스템이나 커버 노래 인식 시스템에서 사용 가능한 멜로디 유사도를 제안하였다. DTW와 크로마 표현에 기반하여 멜로디 사이의 거리를 측정하고, 거리가 가까운 멜로디가 서로 유사한 것이라고 판단한다. 본 논문에서 제안한 멜로디 유사도를 450곡 데이터베이스를 사용하는 QbSH 시스템에 적용해 본 결과, 약 0.713의 MRR 값을 얻을 수 있었다.

6. 참고문헌

[1] J. S. Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, 37:295-340, 2003.

[2] R. Typke, F. Wiering, and R. Veltkamp, "A survey of music information retrieval systems," *Proc. ISMIR,* 2005, pp. 153-160.

[3] A. Ghias, J Logan, and D Chamberlin, "Query by humming:musical information retrieval in an audio database," *Proc. of ACM Multimedia,* 1995, pp. 231-236.

[4] J. -S. R. Jang and M. Y. Gao, "A query-by-singing system based on dynamic programming," *International Workshop on Intelligent Systems Resolution (the 8th Bellman Continuum),* Hsinchu, Taiwan, pp 85-89, Dec. 2000.

[5] J. -S. R. Jang and H.-R. Lee, "A general framework of progressive filtering and its application to query by singing/humming," *IEEE Trans. on Audio, Speech, and language Processing,* vol. 16, no. 2, pp 350-358, Feb., 2008

[6] Y. Zhu and D. Shasha, ""Warping indexes with envelope transforms for query by humming,"" *In Proc. ACM SIGMOD Int. Conf. on Management of Data*, pp. 181-192, 2003

[7] L. Wang, S. Huang, S. Hu, J. Liang, and B. Xu, "An effective and efficient method for query by humming system based on multi-similarity measurement fusion," *Proc. ICALIP*, 2008

[8] H. M. Yu, W. H. Tsai, and H. M. Wang, "A queryby-singing system for retrieving karaoke music,"

IEEE Trans. on multimedia, vol. 10, no. 8, 2008, pp. 1626-1637.
[9] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," *Proc. Int. Conf. Music Information Retrieval,* Bloomington, 2001, pp. 205-210.
[10] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Language Process.,* vol. 15, no. 4, pp. 1247.1256, Apr. 2007.
[11] D. Jang, C. D. Yoo, S. Lee, S. Kim, and T. Kalker, "Pairwise Boosted Audio Fingerprint," *IEEE Trans. Information Forensics and Security,* vol. 4, no. 4, pp. 995-1004, Dec. 2009.
[12] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," ICASSP 2007
[13] 김기출, 박성주, 이석필, 김무영, "선형 보간법을 이용한 시간과 주파수 조합영역에서의 피치 추정 방법," 전자공학회 논문지, 2010.
[14] 이세원, 윤제열, 심동규, 박성주, 이석필, 박호종, "예측 알고리즘을 이용한 다중 기본 주파수 측정 기술," 신호처리합동 학술대회 제22권 1호, 2009
[15] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio Speech Language Process.*, Vol. 16, No. 4, pp. 766-778, 2008.