AV 모델을 이용한 음악 분위기 자동 분류[1)]

문창배*, 김현수*, 송민균*, 김병만*
*금오공과대학교 소프트웨어공학과
e-mail:moonyeses@naver.com

Automatic Classification of Music Moods Based on the AV Model

Chang Bae Moon*, Hyun Soo Kim*, Min Kyun Song*, Byeong Man Kim*
*Dept. of Computer and Software Engineering. Kumoh National Institute of Technology

요 약

본 논문에서 구조 분석 기법을 이용하여 음악을 구간들로 나누고, 그 구간 중 대표 구간들을 자동으로 설정 후 다수의 사용자에게 그 구간들에 대한 분위기 값을 입력받은 후 이 값들을 바탕으로 구간의 대표 분위기를 결정하는 방법을 제안하였다. 또한, 이렇게 결정된 대표 분위기와 그 구간의 음악적 특징들을 이용하여 음악 분위기 판별 실험을 하였다. 이를 위해 음원의 분위기를 211명에게 수집하였고, 음원에서 특징들을 결정계수를 이용하여 특징의 수를 줄인 후 신경망을 이용하여 학습 및 판별을 하였다. 그리고 Leave-one-out 교차 검증을 통하여 성능 분석을 하였다. 실험결과, 3,000번 학습 시 은닉층 17개를 이용하였을 때 66%의 판별율을 보였다.

1. 서론

분위기 추출에 관한 초창기의 연구들 [1, 2, 3] 은 일반적인 기계 학습/판별 방법을 사용하였으나, 이러한 방법은 음악을 하나의 분위기로 판단하기 때문에 정확성이 떨어지는 문제가 있고, 또한 개인의 주관적인 느낌과 이질감을 반영하지 못하는 문제가 있었다. 단일 분위기로 판정하기 때문에 발생하는 불확실성을 해결하기 위해 [4] 의 연구에서는 퍼지 기반의 학습/분류 방법을 사용하였으나, 이 또한 개인에게 느껴지는 음악의 분위기에 대한 주관적 성향을 해결하기에는 한계가 있었다.

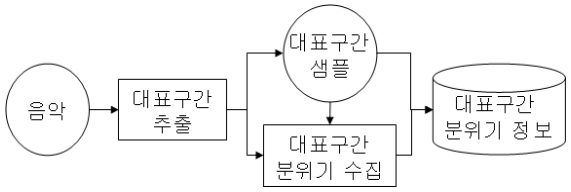
또한, 하나의 음악은 전제 내용이 동일한 분위기 특성을 유지하기 보다는 중간 중간에 다른 분위기로 변화하며, 음률의 변화 또한 다양하다. 따라서 음악의 분위기를 탐지하기 위해서는 전체 음악을 의미 있는 몇 개의 부분으로 나누고 각 부분들에 대하여 독립적인 분위기를 탐지하는 기법이 필요하다. 하지만, 기존 연구들에서는 이러한 특성을 고려하지 않고 각 연구의 필요성에 따라 음악의 일부분을 전문가의 수작업을 통해 잘라내어 사용하거나 임의로 설정된 구간(예를 들어 음악의 시작 후 30초 구간부터 30초 길이의 구간)을 사용하였다. 이러한 방법들은 새로이 출판되는 음악에 적용시키기에 무리가 있으며 변화가 많은 음악의 특성상 정확도가 떨어지는 단점이 있다.

이러한 문제점을 해결하기 위해 [5]에서는 수동이 아닌 자동으로 음악 자체의 내용을 바탕으로 한 구조 분석기법을 통하여 음악을 의미 있는 구간들로 나누고, 각 구간들의 독립적인 분위기를 탐지하는 방법을 제안 하였다. 하지만 [5]의 연구는 피 실험자의 수가 적어 대중적인 성향 보다는 개인적 성향에 대한 비중이 높다고 할 수 있다.

본 논문에서 구조 분석 기법을 이용하여 음악을 구간들로 나누고, 그 구간 중 대표 구간들을 자동으로 설정 후 다수의 사용자에게 그 구간들에 대한 분위기 값을 입력받은 후 이 값들을 바탕으로 구간의 대표 분위기를 결정하는 방법을 제안하였다.

2. 대표구간 추출 및 음원 분위기 수집

음원의 분위기와 분위기의 색상을 수집하기 위해 본 논문에서는 (그림 1)과 같이 음원을 분석하여 대표구간을 파악하고, 대표구간에서 추출한 음원을 웹을 이용하여 피 실험자에게 제공하여 피 실험자에게 음원의 분위기를 입력받았다. 웹을 통하여 입력받은 분위기를 최종적으로 데이터베이스에 저장하여 본 논문에서 분석하기 위한 자료로 구축하였다.

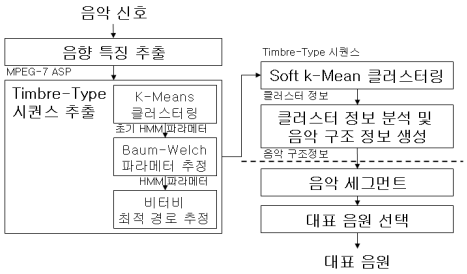


(그림 1) 자료 수집 방법

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2010-0021097)

2.1 음악의 대표구간 추출

대표 구간을 추출하기 위해 (그림 2)와 같이 음악의 구조 분석을 통한 세그먼트를 실행하였다. 대표구간 추출 방법은 음악 구조 정보를 추출하고, 분석된 구조정보를 이용하여 음악을 분리한 후 분리된 음원 중 음악의 에너지 값이 가장 큰 위치를 선택하여 대표음원을 추출하였다. 음악의 구조 분석은 상태열 기반 [6]의 유사 구간 클러스터링 방법을 사용하였다. [6]의 유사 구간 클러스터링 방법은 (그림 2)에서 점선부분까지로 음악 특징 벡터 추출, Timbre-Type 시퀀스 추출, Timbre-Type Soft k-Means 클러스터링 방법을 통하여 음악의 구조정보를 파악한다.



(그림 2) 대표음원 추출 알고리즘 구조도

기존 상태열 기반 [6]의 음악 구조 분석 연구에서는 1차 음향 특징 추출을 위한 특징 추출 프레임의 길이 결정 방법으로 비트 탐색 알고리즘을 통하여 8개의 비트에 해당하는 길이를 프레임 윈도우의 홉사이즈로 사용하였으나 본 연구에서는 1.2s의 길이와 300ms의 홉사이즈를 가진 고정된 프레임을 사용하여 1차 음향 특징을 추출하였다.

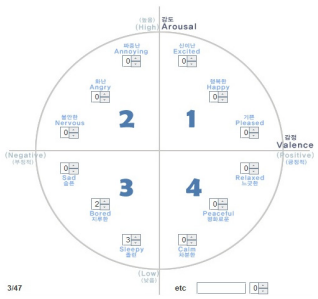
1차 음향 특징은 각 프레임 별로 MPEG-7의 1/8 옥타브의 해상도를 가지는 ASE(Audio Spectrum Envelop)를 추출 한 후, PCA(Principal Component Analysis) 알고리즘을 통하여 상위 20 프로젝션(ASP(Audio Spectrum Projection)를 계산하여 사용한다. 하지만 본 논문에서는 PCA알고리즘에 의해 정규화 된 ASP 값을 사용할 경우는 각 프레임의 에너지 차이에 대한 정보가 사라지기 때문에 각 프레임별 파워 스펙트럼 값의 L2-Norm을 구하여 총 21차의 멜로디와 에너지 모두를 표현하는 음향 특징을 추출하였다.

프레임별 음향 특징들이 추출되면 가상코드 (또는 Timbre-Type) 시퀀스를 추출하기 위하여 HMM을 사용하였다. 가상코드 시퀀스가 추출된 후 가상코드 시퀀스 정보를 이용하여 음악의 특징적 구간(Segment)으로 나누기 위해 [6]에서 사용했던 히스토그램 기반 Soft k-Means 클러스터링 방법을 사용하였다. 즉 M 개의 세그먼트로 나누기 위해 우선 가상코드를 W의 크기를 가지는 윈도우를 한 스텝씩 이동시키며 윈도우 내의 시퀀스를 사용한 데이터 히스토그램을 생성하고, 각 데이터 히스토그램이 속해 있는 세그먼트 라벨을 할당하여 세그먼트 시퀀스를 생성하였다.

본 논문에서는 유사구간을 획득한 후 유사구간의 시작부부터 12초 단위로 음악을 분리시키고, 분리된 12초 단위의 음원들 중 음악의 도입부에서 1개와 종결부에서 1개를 선택하고, 음원들의 에너지를 계산하여 에너지가 가장 큰 샘플을 1개 선택하였다. 즉 음악당 최대 3개의 샘플이 선택되지만 도입부나 종결부에 에너지가 최대인 경우 음악당 2개의 샘플이 선택된다.

2.2 음원의 분위기 및 분위기 색상 수집 방법

음원의 분위기와 분위기에 따른 색상을 수집하기 위한 환경은 10 ~ 17시, 한쪽 벽면이 어두운 유리(동쪽)인 실내에서 3일간에 걸쳐 데이터를 수집하였다. 또한 헤드셋을 제공하여 외부에서 발생되는 소음을 방지하였고, 실험에 사용된 음원의 재생 시간은 각 12초 이다. 분위기 수집에 참여한 인원은 총 211명이고, 제공된 음원샘플은 음악 101곡에서 추출된 총 281개로 189명에게는 41개의 음원샘플을 랜덤하게 제공하고, 13명에게는 전체 음원샘플을 제공하였다. 음원의 분위기 설문시 제공된 웹 UI는 (그림 3)과 같이 확장된 Theyer의 2차원 모델[7]을 이용하였다.



(그림 3) 설문 UI

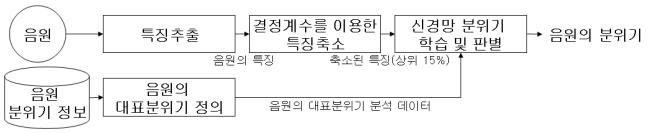
음원의 분위기 설문시 피 실험자에게 최대 3개 분위기를 입력 가능케 하였는데, 한 분위기에 최대 5점 부여가능하며, 여러분위기 입력시 총합이 5점 넘지 않도록 하였다. 또한 분위기 입력 시 제약 조건은 Angry와 Peaceful 같이 극과 극인 분위기들은 동시에 선택할 수 없도록 하였고, (그림 3)에서 etc는 입력자가 12개의 분위기 이외의 분위기를 느낄 수 있어 12개의 분위기 이외의 분위기를 느낀 경우 그 분위기를 입력할 수 있도록 하였다. 하지만 본 논문에서는 분석 시 사용하지 않았다.

웹을 이용하여 수집한 음악 분위기 수는 각 음원별 약 31건이고, 동아리 회원 13명의 데이터를 더하면 각 음원별 평균 44건이다.

3. 음원의 분위기 판별

음원의 분우기 판별을 (그림 4)와 같이 음원의 특징을 추출하고, 음원의 특징을 추출하였을 때 391개의 특징이 추출되어 특징의 수를 줄이기 위한 결정계수를 계산하여 특징을 축소한다. 축소된 특징과 음원의 대표 분위기를 이용하여 신경망에 학습을 시키고, 새로운 음원이 입력되면

음원의 분위기를 판별하는 구조이다.



(그림 4) 음원의 분위기 학습 및 판별 방법

3.1 음원의 대표분위기 정의

한 음원에 대하여 여러 사용자가 다른 분위기를 지정할 수 있기 때문에 대표 분위기를 파악할 필요가 있다. 이를 위해 먼저 각 분위기에 대한 사용자들의 평가치의 합을 식 1과 같이 계산하였다.

$$ed_i^s = \sum_{u=1}^{n} data_{ui}^s, \quad \begin{cases} i = 1,2,...,12 \\ n : \text{피실험자 수} \end{cases} \qquad (1)$$

여기서, ed_i^s는 음원 s에 대한 i번째 분위기에 대한 피실험자들의 평가치 합이며, i는 분위기 인덱스로 1~12까지의 범위를 가지며, 1번부터 반시계 방향으로 pleased, happy, excited, ..., relaxed에 대응된다. $data_{ui}^s$는 음원 s에 대한 피실험자 u의 i번째 분위기에 대한 평가치를 의미한다.

모든 분위기에 대한 ed_i^s를 구한 후, 이 값을 AV 공간상의 좌표 값으로 변환하였다. 변환은 각 ed_i^s에 대해 원점 (0, 0)을 기점으로 하며 거리 ed_i^s만큼의 위치에 표기한 후 그 점의 x, y 좌표 값을 구하였다. 즉, 식 2와 3을 이용하여 x 값과 y 값을 구하였다.

$$x_i = \sin(f\theta_i) ed_i, \quad i = 1,2,3,...,11,12 \quad (2)$$
$$y_i = \cos(f\theta_i) ed_i, \quad i = 1,2,3,...,11,12 \quad (3)$$

여기서, $f\theta_i = f\theta_{i-1} + 30,\ 2 \le i \le 12$이며 $f\theta_1$는 15도이다

12개의 분위기에 대한 ed_i^s의 값을 AV 좌표로 변환 후에는 이들의 중심좌표를 구한 후 이 중심좌표와 각 분위기 축과의 각도를 구하여 가장 근접한 분위기를 대표 분위기로 결정한다. 즉, 식 4와 5를 이용하여 12개의 AV 벡터의 중심 벡터$(\bar{x}, \bar{y})$를 구한 후 이 벡터의 각도 $c\theta$를 $atan(\bar{x}/\bar{y})$에 대입하여 구하고 식 6을 이용하여 가정 근접한 분위기의 인덱스 NI를 구한다.

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i \quad (4) \qquad \bar{y} = \frac{1}{12}\sum_{i=1}^{12} y_i \quad (5)$$

$$IN = \arg min\{d\theta_1, d\theta_2, d\theta_3, ..., d\theta_{11}, d\theta_{12}\}$$
$$where, d\theta_i = |f\theta_i - c\theta|, \ i = 1,2,3,...,11,12 \quad (6)$$

본 논문에서 획득한 음원 분위기의 수가 고루 분포되어 있지 않기 때문에 분위기를 3개씩 클러스터링 하여 AV 모델을 기준으로 4사분면으로 데이터를 구축하였고, 렌덤 선택방법을 이용하여 164개의 음원데이터를 사용하였다.

3.2 음원의 특징추출 및 차원 축소

음원 특징은 [8]를 이용하여 추출하였고, 이 특징들은 대분류로 Dynamics, Fluctuation, Rhythm, Spectral, Timbre, Tonal 이고, 중분류로 MFCC, Tempo, Chromagram, Rolloff등을 포함한 28개이다. 이 28개의 특징들 각각에 대해 Mean, Std, Slope, PeriodFreq, PeriodAmp, PeriodEntropy등의 특징들을 획득하게 된다. 또한 MFCC 와 기타 특징들의 경우 특징 벡터로 구성되어 최종적으로 본 논문에서 추출한 특징들의 수는 391개의 특징들이 존재한다.

[8]를 이용하여 획득한 391개의 특징들을 모두 사용할 경우 잡음 특징들 때문에 오히려 역효과가 날 수 있다. 따라서 본 논문에서는 각 특징들에 대하여 결정계수를 계산하고, 결정계수의 상위 15%를 이용하여 분별력이 좋은 특징들을 추출 하였다[9]. 결정 계수를 구하는 식은 식 7과 같다.

$$r^2 = \left(\frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}} \right)^2 \qquad (7)$$

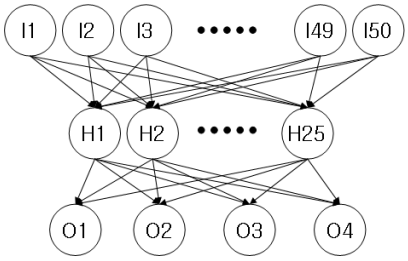
여기서 r^2이 결정 계수를 의미하고, x는 특징의 순번, y는 특징 값 이다.

x의 구성은 pleased, happy, excited, ..., relaxed와 같이 12개의 분위기 단어이지만 결정 계수를 계산하기 위해 pleased부터 (그림 3)의 반시계 방향으로 각 분위기에 1~12까지 맵핑하여 계산 하였다. 이렇게 획득한 특징 계수를 이용하여 분별력 좋은 특징들을 추출 하였다.

분별력 좋은 특징은 dynamics에서 1개, rhythm의 attack에서 2개, spectral에서 36개, timbre에서 3개, tonal에서 8개로 총 50개의 특징으로 축소된다. 그 중 spectral의 경우 18개의 특징이 MFCC에서 선택 되었다.

3.3 신경망을 이용한 음원 분위기 학습 및 판별

본 논문에서 사용한 신경망 구조는 (그림 5)와 같다. 신경망의 각 층의 구성은 입력층 50개, 출력층 4개로 구성하고, 은닉층의 수에 따라 성능적 차가 발생하여 은닉층을 가변적으로 구성시켰다. 즉, 본 논문의 실험에서 은닉층을 2개 부터 25개 까지 성능을 판별하고, 가장 좋은 은닉층을 선택하였다.



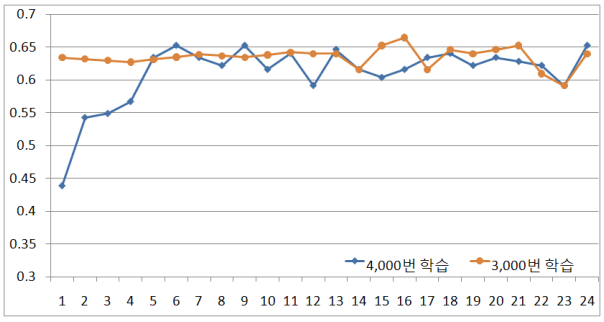
(그림 5) 본 논문에서 사용한 신경망 구조

입력층의 입력은 3.2에서 획득한 데이터를 사용하였고, 출력층은 4개로 구성하여 좌측부터 AV모델의 1사분면, 2사분면, 3사분면, 4사분면을 대응시켰다. 즉 O1, O2, O3, O4의 값이 (1, 0, 0, 0)은 AV모델의 1사분면을 의미하고, (0, 1, 0, 0)은 2사분면, (0, 0, 1, 0)은 3사분면, (0, 0, 0, 1)은 4사분면을 의미한다.

4. 실험 및 성능분석

본 논문에서는 3.3과 같이 설정된 신경망을 이용하여 Leave-one-out 교차 검증방법을 이용하여 성능을 실험하였다. Leave-one-out 교차 검증은 하나의 데이터를 테스트로 사용하고, 나머지 데이터를 학습으로 사용하는 방법으로 본 논문에서는 각 은닉층에 대하여 164번의 실험을 실시하였다. 즉, 2개의 은닉층 부터 25개의 은닉층 까지 각 164번의 실험을 실시하여 총 4,100번의 실험을 실시하였다. 또한 신경망 학습 수를 3,000번과 4,000번을 실시하였다.

실험 결과는 (그림 6)과 같고, 3,000번 학습시 은닉층 17개에서 66%로 3,000번 학습시 가장 좋은 성능을 가지고, 4000번 학습시 은닉층 7, 10, 25개에서 65%로 가장 좋은 성능을 가진다.



(그림 6) 신경망 판별 성능

5. 결론

본 논문에서는 음원의 분위기를 파악하기 위해 특징을 추출하였고, 추출된 특징 중에서 분별력이 좋은 특징을 추출하기 위해 결정계수를 계산하였다. 결정계수에 의해 획득한 특징을 신경망을 통하여 실험하였고, Leave-one-out 교차 검증방법을 이용하여 성능분석 하였다. 실험 결과, 3,000번 학습시 66%를, 4,000번 학습시 65%의 성능을 보였다.

향후, 판별 성능을 높이기 위하여 대표 구간 선택방법, 대표 분위기 선택방법 그리고 특징 축소방법 등에 대한 보다 세밀한 연구가 필요하다. 이와 함께 분위기 클레스를 세분화한 경우의 성능분석 실험도 필요하다.

참고문헌

[1] T. Li and M. Ogihara, "Detectiong Emotion in Music, " Proc. of the International Symposium on Music Information Retrieval(ISMIR), pp.239-240, Washington D.C., USA, 2003

[2] L. Lu, D. Liu and H. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14, pp5-18, 2006

[3] Y.Feng, Y.Zhang and Y.Pan, "Popular Music Retrieval by Detecting Mood, " Proc. of ACM SIGIR 2003, pp 375-376, 2003

[4] Y.H. Yang, C.C. Liu and H.H. Chen, "Music Emotion Classification: a Fuzzy Approach," Proc. of ACM Multimedia 2006(ACM MM'06), pp.81-84, Santa Barbara, CA, USA, 2006

[5] Jong In Lee, Dong-Gyu Yeo, Byeong Man kim, Hae-Yeoun Lee, "Automatic Music Mood Detection through Musical Structure Analysis", International Conference on Computer Science and its Application CSA 2009, pp. 510-515, 2009

[6] Levy, M., Sandier, M. and Casey, M., "Extraction of High-Level Musical Structure From Audio Data and Its Application to Thumbnail Generation", Proc. of ICASSP'06, Vol. 5, pp. 13-16, Toulouse, France, May 2006.

[7] R. E. Thayer, The Biopsychology of Mood and Arousal, New York, Oxford University Press, 1989.

[8] Olivier Lartillot, "MIRtoolbox 1.2.4", Finnish Centre of Excellence in Interdisciplinary Music Research, March, 18th, 2010

[9] 김종완, 김희재, 김병만, "퍼지추론과 신경망을 사용한 유즈넷 뉴스그룹 결정", 한국퍼지 및 지능시스템학회 2004년도 춘계학술대회 학술발표논문집 제14권 제1호, 2004.4