한국음향학회지 제34권 제3호 pp. 247~255 (2015)
The Journal of the Acoustical Society of Korea Vol.34, No.3 (2015)
http://dx.doi.org/10.7776/ASK.2015.34.3.247

pISSN : 1225-4428
eISSN : 2287-3775

감정 인식을 통한 음악 검색 성능 분석

A Study on the Performance of Music Retrieval Based on the Emotion Recognition

서진수

(Jin Soo Seo)

강릉원주대학교 전자공학과

(Received February 2, 2015; accepted March 17, 2015)

초 록: 본 논문은 자동으로 분류된 음악 신호의 감정을 기반으로 하는 음악 검색 의 성능을 분석하였다. 음성, 영상 등의 다른 미디어 신호와 마찬가지로 음악은 인간에게 특정한 감정을 불러일으킬 수 있다. 이러한 감정은 사람들이 음악을 검색할 때 중요한 고려요소가 될 수 있다. 그렇지만 아직까지 음악의 감정을 직접 인식하여 음악 검색을 수행하고 성능을 검증한 경우는 거의 없었다. 본 논문에서는 음악 감정을 표현하는 주요한 세 축인 유발성, 활성, 긴장 과 기본 5대 감정인 행복, 슬픔, 위안, 분노, 불안의 정도를 구하고, 그 값들의 유사도를 기반으로 음악 검색을 수행하였다. 장르와 가수 데이터셋에서 실험을 수행하였다. 제안된 감정 기반 음악 검색 성능은 기존의 특징 기반 방법의 성능에 대비해서 최대 75 % 수준의 검색 정확도를 보였다. 또한 특징 기반 방법을 제안된 감정 기반 방법과 병합할 경우 최대 14 % 검색 성능 향상을 이룰 수 있었다.
핵심용어: 음악 검색, 음악 감정 인식, 음악 유사도

ABSTRACT: This paper presents a study on the performance of the music search based on the automatically recognized music-emotion labels. As in the other media data, such as speech, image, and video, a song can evoke certain emotions to the listeners. When people look for songs to listen, the emotions, evoked by songs, could be important points to consider. However; very little study has been done on the performance of the music-emotion labels to the music search. In this paper, we utilize the three axes of human music perception (valence, activity, tension) and the five basic emotion labels (happiness, sadness, tenderness, anger, fear) in measuring music similarity for music search. Experiments were conducted on both genre and singer datasets. The search accuracy of the proposed emotion-based music search was up to 75 % of that of the conventional feature-based music search. By combining the proposed emotion-based method with the feature-based method, we achieved up to 14 % improvement of search accuracy.
Keywords: Music retrieval, Music emotion recognition, Music similarity
PACS numbers: 43.75.Zz

I. 서 론

디지털 저장 장치 및 신호처리 기술의 발달로 방대한 양의 오디오 데이터들을 빠르고 신뢰성 있게 보호, 검색 및 관리할 수 있는 오디오 정보 처리 기술의 필요성이 커지고 있다. 대표적인 오디오 정보 처리 기술에는 음악 검색, 추천, 인식, 분류 등이 있다.[1] 일반적으로 정보 검색 시스템은 입출력의 형태, 검색 기준에 따라 분류할 수 있다. 오디오 정보 검색도 마찬가지로 시스템을 설계할 때, 입력 형태를 텍스트로 할 것인지 오디오 파일로 할 것인지 정하고, 출력 형태 또한 텍스트에서 오디오의 부분 또는 특징 등 다양할 수 있다. 검색 기준에는 의미론적인 특징이냐 신호 레벨의 특징을 사용하느냐에 따라 그 기준이 달라질 수 있으며, 유사한 다수의 결과를 제시

†**Corresponding author:** Jin Soo Seo (jsseo@gwnu.ac.kr)
Department of Electronic Engineering Gangneung-Wonju National University, 7 Jukhun-Gil, Gangneung 210-702, Republic of Korea.
(Tel: 82-33-640-2428, Fax: 82-33-656-0740)

할 지 정확하게 매치되는 하나만을 제공할 지도 또 다른 분류 요소이다. 발달된 정보처리 기술을 이용하여 대용량 디지털 미디어 데이터 아카이브를 이용하여 다양한 종류들의 서비스가 가능해 지고 있다. 예를 들어, 유사도를 이용한 오디오 정보 처리 서비스의 경우에도 그 유사도의 선택 기준에 따라서 다양하며, 핑거프린팅과 같이 입력 음악과 정확히 일치하는 아카이브상의 음악을 찾는 경우도 있고,[2,3] 장르 분류[4] 및 유사음악 검색[5,6]과 같이 특정한 성질을 공유하는 다수의 결과를 출력하는 경우도 있다. 본 논문에서는 특정한 성질을 공유하는 유사음악 검색에 대해서 다룬다.

유사 음악 검색 시스템에서는 두 입력 음악 간의 유사도를 구하는 부분이 핵심이 된다. 음악 유사도는 크게 두 가지 방법으로 비교할 수 있다. 협업 필터(collaborative filtering) 방법과 음악 특징 기반 방법이 있다. 협업 필터 방식은 입력 음악의 유사도를 많은 사람들의 음악 소비 경향 또는 기호를 바탕으로 두 음악 간의 거리를 구하게 된다. 예를 들어 음악의 경우 현재 사용자와 비슷한 음악 취향을 가진 사람들이 어떤 음악을 좋아했다면 사용자에게 추천하는 방식이다. 반면에 특징기반 음악 유사도 비교 방법은 음악 신호에서 직접 특징을 구해서 그 특징 간의 거리비교를 통해서 음악 유사도를 구하는 방법이다. 두 방법 모두 그 자체로 장단점이 있다. 협업 필터 방법은 기존에 소비 패턴을 가지고 있지 않은 새로운 노래에 적용할 수 없는 반면에 특징 기반 방법은 인간 지각적으로 의미있는 특징을 추출해야 하고 특징 간의 거리를 비교해야 하므로 계산량이 크게 요구되는 단점이 있다. 본 논문은 특징 기반 음악 검색 방법에 관한 연구이다.

음악 유사도 검색의 어려움은 유사도 판정의 근거가 주관적이고 정량적으로 표현하기 어렵기 때문이다. 지금까지 음악 유사도 검색 방법들은 주로 MFCC (Melfrequency Cepstral Coefficients) 등 음악의 음색 특징들에 기반 해서 구현되었다.[5,6] 초기의 특징 기반 유사도 검색은 주로 어떤 노래의 음악 신호 전체에서 얻은 MFCC 벡터들을 하나의 GMM(Gaussian Mixture Model) 또는 *k*-means 군집을 통해서 모델링 하였다. 이렇게 각 노래에서 얻은 특징 모델들간의 거리를 EMD(Earth-Mover's Distance)[5] 또는 KL divergence[6]를 통해서 구하였다. 음악 특징 군집화 및 군집간의 거리를 비교하는 방법은 군집을 구할 때 수렴성의 문제가 발생할 수 있고, 군집간의 거리를 구하는 것이 닫힌 해가 존재하지 않을 경우가 많고 일반적으로 계산량이 많은 단점이 있다. 따라서 최근 이러한 단점을 보완한 방법으로 UBM(Universal Background Model)[7]을 이용한 방법들[6,8,9]이 제안되었다. 일반적으로 UBM은 미리 수집된 다양한 음악 신호로부터 특징벡터들을 얻고 이를 GMM을 통해서 모델링한 것이다. 특히 GMM을 UBM으로 이용하고 SV(Super Vector)[10]개념을 적용한 방법들[8,9]이 음악검색에 성공적으로 적용되었다. 따라서 본 논문에서는 SV에 기반한 음악 검색기를 제안된 감정인식 특징 기반 방법의 성능과 비교하였다. 본 논문에서는 Fig. 1과 같이 음악이 불러일으키는 감정을 인식하고 그 감정 특징을 기반으로 음악들 간의 유사도를 구해서 음악을 검색하는 방법에 대해 연구하였다. 음성, 영상 등의 다른 미디어 신호와 마찬가지로 음악은 인간에게 특정한 감정을 불러일으킬 수 있다. 이러한 감정은 사람들이 음악을 검색할 때 중요한 고려요소이지만, 아직까지 음악의 감정을 인식하여 음악 검색을 수행하고 성능을 검증한 경우는 거의 없었다. 음악 감정[11-14]은 유발성, 활성, 긴장의 세 개의 축으로 이루어진 3차원 공간상에서 표현 가능하므로 특징 차수가 간결한 장점도 있다. 장르와 가수 데이터셋에서 실험을 수행하여, 제안된 감정 특징 기반 음악 검색 성능을 평가하였다. 또한 본 논문에서는 기존의 SV에 기반한 음악 검색기와 제안된 감정 특징 기반 검색기를 병합하는 방법에 대해서도 연구하였다. 기존 SV 방법은 음악 신호의 MFCC만을 이용하므로 음악의 음색 특징만을 활용한다. SV에 감정 특징을 추가할 경우 검색 성능 향상의 정도를 실험적으로 확인하였다.

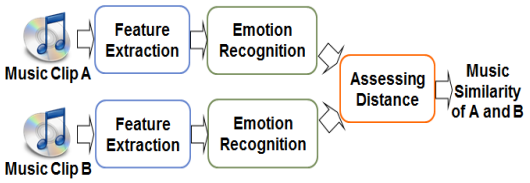

Fig. 1. An overview of the music-similarity computation based on music emotion.

본 논문은 음악 감정 특징 기반 유사도에 기반한 음악 검색에 관한 연구이다. II장에서 음악 감정 인식 방법을 소개하고, 감정 인식을 음악 유사도에 적용한다. III장에서 제안된 방법의 성능을 실험하고 결과를 분석한다.

II. 감정 인식 기반 음악 검색

2.1 음악 감정 인식

음악이 특정한 감정을 불러일으키거나 강화시킬 수 있는 것은 잘 알려져 있으며, 영화 배경음악이나 광고 등에 널리 사용되고 있다. 또한 특정한 기분 또는 분위기를 내기 위해 음악을 듣는 경우도 많이 있다. 따라서 음악 감정은 음악 정보 처리의 중요한 요소 중 하나이며, 최근 큰 관심을 받아 연구가 진행되어 왔다.[11-13] 음악 감정은 유발성(valence or pleasure-displeasure continuum)과 활성(activity or energetic arousal)의 두 축을 사용한 V-A 공간을 이용해서 모델링 해왔으나,[14] 이러한 2차원 모델이 음악 감정을 완벽히 표현하는 것이 불가능하다는 실험 결과[15]에 따라서 요즘은 기존의 유발성과 활성에 긴장(tension or tense arousal)을 추가해서 V-A-T 공간[16]을 주로 이용한다. Fig. 2에 주어진 바와 같이 인간의 여러 가지 감정들이 V-A-T 공간상에서 표현가능하다. 예를 들어 V값이 큰 감정으로는 행복, 기쁨, 만족 등이 있고, V값이 작은 감정들에는 연민, 좌절 등이 있다. A값이 큰 감정에는 놀라움, 경악 등이 있고, A값이 작은 감정에는 지겨움, 지침 등이 있다. T값은 V와 A에 비해서 감정 구분이 명확하지는 않으며 V-A 공간을 보완하는 역할을 한다.

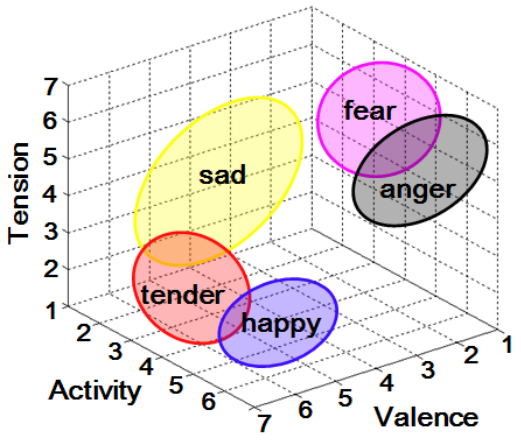

Fig. 2. The music emotion based on Valence-Activity-Tension space.[12]

음악 감정은 사람마다 느끼는 정도의 차이가 크고 주변 환경에 따라서 좌우될 수 있어서 자동으로 인식하는 것은 어려운 문제이다. 기존의 방법들은 대부분 음색, 리듬 등과 같은 기본적인 음악의 저수준 특징들을 이용하여 기존에 미리 태깅해둔 학습 음악 신호들의 V-A 또는 V-A-T 값을 회귀분석하는 방법을 사용한다. 기본적으로 V-A 또는 V-A-T 매핑을 이용하지만 어떠한 특징들을 사용하고 회귀분석 방식에 따라 감정 인식 성능이 결정된다. 그러나 실제 감정 인식기를 구현할 때 가장 문제가 되는 부분은 학습 감정 태깅 데이터의 수집[17,18]이다. 예시 감정들 중에서 하나를 선택하게 할 수도 있고, 직접 V-A 또는 V-A-T 공간상에서 위치를 정하도록 하는 방법도 있다. 통계학적인 오류를 줄이기 위해서는 최대한 많은 노래에 많은 수의 사람들이 감정 태깅을 수행해야 하므로 큰 어려움이 있다. 직접 실험자를 구하여 감정 태깅 데이터를 얻기 어려운 경우에는, 온라인 게임과 유사한 방식[19] 또는 아마존의 Turk 서비스[20] 등이 작은 비용으로 태깅 데이터 수집에 이용될 수 있으나 얻어진 데이터의 신뢰도가 낮은 문제가 있다.

본 논문에서는 감정 인식이 주요한 연구 목표가 아니고, 현존하는 감정 인식기의 결과값이 어느 정도의 신뢰도로 음악 검색에 활용될 수 있는가를 확인하는 것이다. 여타 다른 감정인식 방법도 적용할 수 있지만, 본 논문에서는 실험결과의 재현성을 위하여 공개 소프트웨어인 MIRtoolbox[21]에서 제공되는 감정인식기를 활용한다. MIRtoolbox 버전 1.3의 감정인식기를 사용했다. MIRtoolbox의 감정인식기는 음색, 리듬, 피치 등 MIRtoolbox 내의 29차수의 음악 특징들을 사용하고, 특징 차수 축소(dimension reduction) 후 정규화한 뒤에 미리 수집한 110개의 음악 클립에 대한 감정 태깅 데이터와 선형회귀분석을 하여 만들어졌다.

MIRtoolbox에서 제공하는 감정 인식기는 두 가지 종류의 출력을 제공한다. Fig. 2에 주어진 바와 같이 첫 번째는 입력 음악의 V-A-T 값, 두 번째는 기본 5대

감정인 행복(happiness), 슬픔(sadness), 위안(tenderness), 분노(anger), 불안(fear)일 기대치이다.[1] V-A-T와 각 기본 감정에 속할 기댓값은 모두 1에서 7사이의 값을 가지도록 되어 있다. 다만 학습에 사용되지 않은 입력 음악 신호에 대해서는 1과 7을 벗어나는 값이 출력되기도 한다. 본 논문에서는 고수준 음악 특징인 음악 감정인식 결과를 이용하여 음악 유사도 비교 방법을 제안하며, 실험을 통해서 그 성능을 검증한다.

2.2 음악 감정 인식 기반 음악 유사도

음악 유사도의 판정은 주관적이고 특정한 값으로 산정하기 어려운 특징이 있다. 따라서 기존 연구결과들은 주로 음악 신호의 MFCC 값 등의 저수준의 스펙트럼 특징들을 비교하여 음악 유사도로 활용하였다. 아직까지 좀 더 높은 수준의 음악 특징을 활용하여 유사도를 비교하는 연구는 많지 않다. 기존의 음악 태그를 이용하는 방법도 있었으나, 음악에 대한 태그를 직접 사람이 해야 하는 단점이 있다. 본 논문에서는 고수준 음악 특징으로 음악 감정 인식 결과를 활용하여 음악 비교를 수행하는 방법을 제안하고 성능을 검증하였다. 본 논문에서 감정기반 특징은 미리 학습된 감정인식기의 결과를 사용하므로, 매번 입력 음악의 감정을 수작업으로 태깅하지 않아도 된다. 또한 기존의 MFCC를 이용한 SV기반 음악 검색 방법[8,9]과는 다른 정보를 활용하므로 추가로 병합하여 사용할 수 있다.

본 논문에서는 각 음악 파일에서 얻은 감정인식 결과를 모아서 하나의 벡터로 보고 거리비교를 수행하고자 한다. 따라서 대표적인 벡터간 거리 비교 방법인 유클리디안 이나 코사인 거리 등을 사용하여 음악 검색을 위한 음악 유사도로 활용할 수 있다. 본 논문에서는 인지적으로 인간의 지각은 특정 문턱값 이상이 되면 무뎌지는 특성이 있다. 이를 활용하여 감정의 차이가 미리 정해둔 문턱값 T 이상일 경우 거리값을 T 로 정규화하는 거리비교 방법을 제안한다. 두 개의 음악 클립으로부터 얻은 D 차원 감정인식 결과 벡터를 각각 e_A와 e_B라고 할 때 정규화된 거리 $D_N(e_A, e_B)$는 아래와 같이 주어진다.

$$D_N(e_A,e_B)=\frac{1}{D}\sum_{m=1}^{D}\min\left(\frac{|e_A[m]-e_B[m]|^k}{T^k},1\right). \quad (1)$$

정규화된 거리값 D_N에서 T와 k 값은 미리 정하는 상수이며, 본 논문에서는 T와 k 모두 3의 값을 사용하였다. 제안된 감정거리 값을 일반적인 벡터 거리 방법인 유클리디안 이나 코사인 거리 등과 성능을 비교하였다. 실제적으로 제안된 거리비교 방법은 V-A-T의 세 감정 축 중 어느 한 축 방향의 차이가 음악 유사도에 큰 영향을 주는 것을 막는 효과가 있다. 또한 제안된 거리값은 0과 1사이로 정규화 되어 있어서 거리값 해석 및 다른 종류의 특징들과 병합하여 사용할 때 용이하다.

본 연구의 주요한 다른 방향은 감정기반 특징이 기존의 MFCC 등 음악 검색에 사용되고 있는 기존의 스펙트럼 기반 특징에 비해서 새로운 정보를 가지고 있는 지 확인하는 것이다. 이를 위해서 MFCC 기반의 SV를 이용한 음악 유사도와 제안된 음악 감정 기반 유사도를 병합하고 성능을 확인하였다. 본 논문에서는 기존 실험에서 우수한 음악 검색 성능을 보인 UBM 정규화 SV[9]를 사용했다. 공정한 병합을 수행하기 위해서는 먼저 SV 방법에 의한 거리값과 음악 감정 거리값을 각각 정규화 시켜야 한다. 그렇지 않을 경우 직접적으로 병합하는 것이 의미가 없게 된다. 본 논문에서는 SV s를 SV를 구하는 데 사용한 UBM의 평균값인 μ_{UBM}을 이용하여 다음과 같이 UCS (UBM-Centered Spherical) 정규화[9] 하였다.

$$s_{UCS}=\frac{s-\mu_{UBM}}{\|s-\mu_{UBM}\|}. \quad (2)$$

정규화된 SV는 정규화 하지 않은 경우에 비해서 더 좋은 음악 검색 성능을 보였다.[6,9] 본 논문에서는 따로 언급이 없으면 위와 같이 정규화된 SV를 사용한다. 정규화된 SV는 벡터 크기(vector norm)이 1이므로 유클리디안과 코사인 거리의 차이가 없다. 따라서 본 논문에서는 두 음악신호에서 얻어진 정규화된

1) 인간의 기본 감정은 anger, disgust, fear, happiness, sadness, surprise의 6개로 나누는 것이 일반적이나 MIRtoolbox에서는 음악 감정을 happiness, sadness, tenderness, anger, fear의 5개로 나눈다.

SV 들인 p 와 q 간의 거리 비교 방법으로 다음과 같이 0과 2사이의 값을 가지는 코사인 거리를 사용하였다.

$$D_S(p,q) = 1 - \frac{pq^T}{\sqrt{pp^T}\sqrt{qq^T}}. \quad (3)$$

감정값의 경우 제안된 거리 비교 방법인 D_N을 활용하므로 따로 감정인식 벡터를 정규화 하지 않았다. 두 음악 신호 A와 B의 SV를 각각 s_A와 s_B 감정기반 특징 벡터를 각각 e_A와 e_B로 나타낼 경우 병합 음악간 거리는 아래와 같이 나타낼 수 있다.

$$D_C(A,B) = w_S D_S(s_A, s_B) + w_E D_N(e_A, e_B). \quad (4)$$

실험에서는 SV 거리와 감정 기반 거리를 선형 결할 할 때, SV 거리에 대한 가중치인 w_S=1 로 할 때, 가장 좋은 성능을 주는 감정 기반 거리의 가중치 w_E를 실험적으로 결정하였다.

III. 실험 결과

본 장에서는 II장에서 살펴본 감정 인식을 통한 감정 특징들과 SV를 이용한 음악 검색 성능을 비교하고, 감정인식 방법과 SV 방법을 병합한 음악 검색기를 통한 성능 향상의 정도를 측정하였다. 음악 검색의 성능을 비교하는 것은 음악 유사도에 대한 ground truth가 존재하지 않으므로 상당히 어려운 문제이다.[22,23] 또한 음악 유사도는 주관적인 면이 많고 수치화하기 어렵다.[24] 객관적인 음악 유사도를 찾으려는 연구들이 있었으나, 일반적으로 널리 알고 있는 장르 등의 메타 레이블이 인간의 유사도 인지와 상당히 연관이 크다는 정도의 결과만 있어왔다.[25] 따라서 대부분의 선행 음악 검색 연구들[5,6,8,9]에서는 같은 장르 또는 가수의 음악들이 다른 장르 또는 가수의 음악들에 비해서 서로 인간 지각적으로 유사하다는 가정에 바탕을 두고 음악 검색 성능을 평가했다. 본 논문에서도 같은 가정을 통해서 제안된 감정 인식 기반 음악 검색 방법의 성능을 검증하도록 하겠다. 선행 연구와 같은 방식의 성능 실험을 함으로써, 선행 연구들의 결과와도 직접적으로 비교할 수 있는 장점도 있다.

본 논문에서는 장르와 가수 데이터셋 두 가지 데이터셋을 사용하였다. GTZAN 음악 데이터셋[4]은 blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock의 10개의 장르에 각각 100곡씩 30 s 길이의 1000개의 음악파일로 이루어져있다. 가수 데이터셋[6]은 남녀 각각 17명씩 총 34명의 가수 별로 20곡씩 680곡으로 이루어져있다. 데이터셋의 음악들로 특징 DB를 만들고, 데이터셋 각 음악을 질의 음악으로 만들어진 특징 DB에 대해 음악 검색을 수행하였다. 음악 검색 결과 질의 음악과 가장 가까운 5곡, 10곡, 20곡 중에서 질의 음악과 같은 장르 또는 가수의 음악이 몇 곡이나 포함되어 있는 지를 음악 검색의 정확도로 사용하였다.

감정인식은 MIRtoolbox에서 제공되는 함수를 사용하며 음악 신호에 대해 매 3 s마다 감정인식을 수행하여 V-A-T값과 5개 감정인식값을 1에서 7까지 크기로 구한다. 다만 MIRtoolbox에서 사용된 회귀분석의 한계로 인해서 7 이상의 값과 1 이하의 값이 나올 수 도 있으나 그대로 사용하여 거리 비교를 수행하였다. 입력 음악 클립에서 매 3 s마다 감정인식을 수행하여 얻어진 V-A-T값, 5개 감정인식값의 평균값을 그 음악의 특징 감정 벡터로 활용하였다.

감정기반 특징을 이용한 음악 검색 성능을 최근 제안되어 널리 사용되고 있는 SV에 기반한 방법[8-9]의 성능과 비교하였다. SV를 얻기 위해서 실험에 사용되는 음악 파일들을 모노로 바꾸고 22050 Hz로 샘플링 주파수를 맞춘 후, 512 길이의 해닝(Hanning) 윈도우를 overlap 없이 옮겨 가면서 적용하고 FFT를 가한다. 이렇게 주파수 도메인으로 신호를 변환해서 얻은 각 프레임의 스펙트럼으로부터 19차 MFCC를 각각 계산하였다. 1000곡의 다양한 음악을 이용하여 미리 학습한 GMM(mixture 개수는 12개)을 이용하여 각 음악 파일마다 228차원의 SV를 구하였다.

장르 데이터셋에 대한 실험결과는 Table 1에 제시되어 있다. 감정인식 결과로 V-A-T를 이용한 것과 기본 5대 감정일 기대치를 이용한 것, 그리고 둘 다를 이용한 것 세 가지 경우에 대해서 검색 성능을 실험하였다. II장에서 제시한 바와 같이 음악 감정 벡터 간의 거리는 Euclidean, Cosine, 제안된 D_N 거리를 사

Table 1. Average number of closest songs correctly retrieved with the criterion of the same genre using the GTZAN genre dataset. Five emotions refer to the happiness, sadness, tenderness, anger, and fear.

Types of Features	Number of Dimensions	Distance Measure	Average Number of Correctly-Retrieved Songs		
			Closest 5	Closest 10	Closest 20
V-A-T	3	Euclidean	1.734	3.349	6.430
		Cosine	1.553	3.075	6.008
		D_N	1.731	3.365	6.437
Five Emotions	5	Euclidean	1.971	3.649	6.647
		Cosine	1.921	3.620	6.680
		D_N	1.977	3.655	6.624
V-A-T & Five Emotions	8	Euclidean	2.111	3.928	7.212
		Cosine	2.110	3.950	7.237
		D_N	2.127	3.939	7.237
SV[9] (UCS, K=12)	228	Cosine (D_S)	2.922	5.243	9.195
Logan's Method[5]		EMD	2.743	4.801	8.384
Random Selection			0.5	1.0	2.0

Table 2. Average number of closest songs correctly retrieved with the criterion of the same singer using a singer dataset. Five emotions refer to the happiness, sadness, tenderness, anger, and fear.

Types of Features	Number of Dimensions	Distance Measure	Average Number of Correctly-Retrieved Songs		
			Closest 5	Closest 10	Closest 20
V-A-T	3	Euclidean	0.419	0.768	1.325
		Cosine	0.377	0.706	1.275
		D_N	0.418	0.750	1.312
Five Emotions	5	Euclidean	0.529	0.950	1.672
		Cosine	0.478	0.810	1.474
		D_N	0.479	0.885	1.609
V-A-T & Five Emotions	8	Euclidean	0.599	0.991	1.684
		Cosine	0.575	0.999	1.691
		D_N	0.577	0.959	1.672
SV[9] (UCS, K=12)	228	Cosine (D_S)	2.118	3.457	5.218
Logan's Method[5]		EMD	1.743	2.776	4.044
Random Selection			0.147	0.294	0.588

용하였고 성능을 비교하였다. 사용한 3가지 거리 비교 방법은 비슷한 성능을 보였다. V-A-T와 기본 5대 감정의 성능의 차이는 크지 않았으며, V-A-T와 기본 5대 감정을 같이 사용할 경우 약 10 % 정도 성능 향상을 보였다. 감정기반 특징은 기존 SV와는 최고 성능 기준으로 약 75 % 수준의 성능을 보였으며, 차수가 1/20 이하임을 고려하면 준수한 검색 성능이다. 가수 데이터셋에 대한 실험결과는 Table 2에 제시되어 있다. 장르 데이터셋과 비교할 때 감정인식을 통한 검색 정확도가 SV에 비해서 상당히 낮음을 알 수 있다. 이는 가수의 개인별 특징이 주로 음악 신호의 스펙트럼 상에 존재하며, 개인별 특징은 SV와 같은 스펙트럼 특징에 크게 의존하게 된다. 따라서 제안된 감정 인식 결과는 가수 기준 검색 정확도에서 SV에 비해서 크게 떨어지게 된다. 이는 감정기반 특징의 한계로써 이를 극복하기 위해서는 MFCC 같은 스펙트럼 특징을 병합하여 사용하는 것을 고려해 볼 수 있다.

감정기반 특징은 음악검색에 독립적으로 사용될 수도 있고, 다른 특징들과 병합하여 사용될 수 도 있다. 본 논문에서는 기존에 널리 사용되고 있는 SV 기반 음악 검색과 감정기반 특징을 병합하여 그 성능의 개선 정도를 확인하였다. 2.2장에 주어진 바와 같이 정규화된 거리값들인 D_N과 D_S를 선형결합한 Eq.(4)의 D_C 거리를 사용하였다. Fig. 3는 장르 데이터셋에 대해서, SV 거리값에 대한 가중치인 w_S=1 로 고정하

고 감정 거리값에 대한 가중치인 w_E를 가변시켜가면서 음악 검색 결과 중 가까운 10곡(closest 10)에서 장르가 일치하는 결과의 개수를 그렸다. 가중치 w_E의 값을 증가시킴에 따라 장르 일치도가 좋아지다가 w_E의 값이 1.5를 넘어서면 오히려 장르 일치도가 나빠짐을 알 수 있다. 장르 데이터셋에서 감정기반 특징을 기존의 SV에 추가함으로써 SV 대비 최대 14 % 내외의 음악 검색 성능 향상을 이룰 수 있음을 확인했다. 다만 V-A-T 와 SV를 병합하는 것과 V-A-T와 기본 5대 감정을 같이 사용하여 SV를 병합하는 것의 성능 차이는 그리 크지 않았다. 이는 음악 감정이 SV 에 대비해서 추가로 가지는 음악 정보가 V-A-T로 이루어지는 3차원 공간 상에서 잘 표현된다는 것을 의미한다. 즉, V-A-T에 기본 5대 감정을 추가하여도 추가 되는 음악 정보가 크지 않음을 의미한다. 가수 데이터

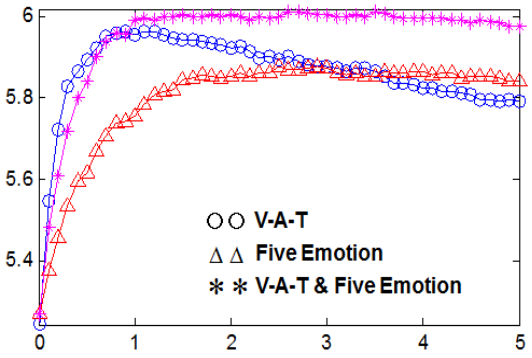

Fig. 3. Average number of correctly retrieved songs among the closest 10 songs with the criterion of the same genre versus w_E when the emotion features are combined with the SV.

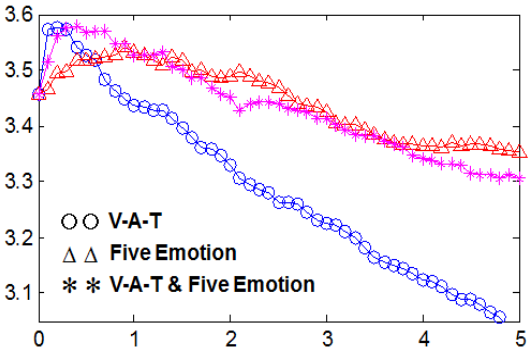

Fig. 4. Average number of correctly retrieved songs among the closest 10 songs with the criterion of the same singer versus w_E for N=10 when the emotion features are combined with the SV.

셋에 대해서 같은 실험을 수행하여, 역시 음악 검색 결과 중 가까운 10곡(closest 10)에 대한 가수 일치도를 Fig. 4에 나타내었다. 가수 일치도의 경우 감정기반 특징을 추가했을 때 장르 데이터셋에 비해서 상대적으로 일치도 향상의 정도는 작았다. 이는 Table 2의 결과와 같이 개별 가수의 특징이 음색에서 두드러지게 나타나므로 감정 특징을 추가하여도 가수 기준 음악 검색 성능 실험에서는 특별히 큰 성능향상을 가져올 수 없기 때문인 것으로 생각된다. 가수 데이터셋에서 감정기반 특징을 기존의 SV에 추가함으로써 SV 대비 최대 3 % 내외의 음악 검색 성능 향상이 있었다. Figs. 3과 4는 N이 10일 때의 실험 결과이며, N이 5와 20 일 경우에도 결과 값 그래프의 형태는 비슷하였다. N 값이 5일 때 장르 기준 검색 결과는 최대 3.35이고, 가수 기준 검색 결과는 최대 2.23 으로 SV 성능에 대비해서 각각 14.2 %, 5.7 % 성능 향상이 있었다. N값이 20일 때는 장르 기준 검색 결과는 최대 10.6이고, 가수 기준 검색 결과는 최대 5.38 로 SV 성능에 대비해서 각각 14.8 %, 3 % 성능 향상이 있었다.

본 논문에서 다른 음악 감정은 음악의 고수준 특성의 하나로써 작은 차수(3차 ~ 8차)로 나타낼 수 있다. 이러한 고수준 특징을 음악 검색에 활용하여, 특징의 차수가 1/20 이하임에도 기존 최대 성능 대비 약 75 % 정도의 검색 성능을 얻었다. 또한 저수준 음색 특징인 MFCC 기반의 SV 거리값과 감정기반 특징의 거리값을 선형 결합을 통해서 같이 사용할 경우 기존 방법들에 대비해서 검색 성능을 개선할 수 있음을 확인하였다. 본 논문에서는 SV 거리값과 감정기반 특징 거리값 간의 최적의 가중치를 실험적으로 구하였으나, 최적의 가중치를 이론적으로 구하기 위한 분석이 필요하며, 두 거리값에 따라 가중치를 적응적으로 가변시키거나 선형이 아닌 비선형 결합 방법에 대해서도 추후 연구가 필요하다.

IV. 결 론

본 논문에서는 자동으로 인식된 음악의 감정정보를 기반으로 음악검색을 수행하는 방법을 제안하고 성능을 분석하였다. 음악의 감정은 음악의 고수준 정보 중하나이며, 사용자들이 음악을 검색할 때 중

요한 고려 요소이다. 특히 음악의 감정은 3차원 공간 상에서 간결하게 표현되므로, 제안된 감정기반 음악 검색은 낮은 차수의 특징으로 DB 검색을 수행할 수 있는 장점이 있어 대용량 음악 아카이브 상에서 음악 검색을 수행할 때 DB 저장 비용 등의 측면에서 장점이 있다. 기존의 음색 기반 특징과 병합하여 사용될 경우 기존 음악 검색의 성능을 최대 14 % 정도 향상시킬 수 있음을 실험을 통해 확인하였다. 본 연구 결과를 바탕으로 앞으로 다른 고수준 음악 정보를 활용한 음악 검색에 대한 연구도 가능할 것으로 기대된다.

감사의 글

이 논문은 2013년도 강릉원주대학교 장기해외 파견연구 지원에 의하여 수행되었음.

References

1. M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-based music information retrieval: Current directions and future challenges,” Proc. *IEEE* **96**, 668-696 (2008).
2. P. Cano, E. Battle, T. Kalker, and J. Haitsma, “A review of audio fingerprinting,” J. VLSI Sig. Process. **41**, 271-84 (2005).
3. J. Seo, “A robust audio fingerprinting method based on segmentation boundaries” (in Korean), J. Acoust. Soc. Kr. **31**, 260-265 (2012).
4. G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Speech Audio Process. **10**, 293-302 (2002).
5. B. Logan and A. Salomon, “A music similarity function based on signal analysis,” in Proc. ICME-2001, 745-748 (2001).
6. J. Seo, “A music similarity function based on the centroid model,” IECIC Trans. Info. and Sys. **96**, 1573-1576 (2013).
7. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital. Sig. Process. **10**, 19-41 (2000).
8. C. Cao and M. Li, “Thinkit’s submissions for MIREX 2009 audio music classification and similarity tasks,” in Proc. ISMIR-2009 (2009).
9. C. Charbuillet, D. Tardieu, and G. Peeters, “GMM supervector for content based music similarity,” in Proc. DAFX-2011, 425-428 (2011).
10. W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Process. Lett. **13**, 308-311 (2006).
11. Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, “A regression approach to music emotion recognition,” IEEE Trans. Audio, Speech, Language Process. **16**, 448-457 (2008).
12. T. Eerola, O. Lartillot, and P. Toiviainen, “Prediction of multidimensional emotional ratings in music from audio using multivariate regression models,” in Proc. ISMIR-2009, 621-626 (2009).
13. M. Barthet, G. Fazekas, and M Sandler, “Music emotion recognition: from content-to context-based models,” From Sounds to Music and Emotions, 228-252 (2013).
14. J. A. Russell, “A circumplex model of affect,” J. pers. soc. psychol. **39**, 1161-1178 (1980).
15. E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet, “Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts,” Cognition & Emotion **19**, 1113-1139 (2005).
16. U. Schimmack and R. Reisenzein, “Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation,” Emotion **2**, 412-417 (2002).
17. J. Skowronek, M. McKinney, and S. van de Par, “A demonstrator for automatic music mood estimation,” in Proc. ISMIR-2007, 345-346 (2007).
18. X. Hu, M. Bay, and J. S. Downie, “Creating a simplified music mood classification ground-truth set,” in Proc. ISMIR-2007, 309-310 (2007).
19. Y. E. Kim, E. Schmidt, and L. Emelle, “Moodswing: A collaborative game for music mood label collection,” in Proc. ISMIR-2008, 231-236 (2008).
20. J. H. Lee and X. Hu, “Generating ground truth for music mood classification using mechanical turk,” in Proc. JCDL-2012, 129-138 (2012).
21. O. Lartillot and P. Toiviainen, “A Matlab toolbox for musical feature extraction from audio,” in Proc. Digital Audio Effects, 237-244 (2007).
22. W.-J. Yoon, K.-K. Lee, and K.-S. Park, “A Study on the Efficient Feature Vector Extraction for Music Information Retrieval System” (in Korean), J. Acoust. Soc. Kr. **23**, 532-539 (2004).
23. C. Park, M. Park, S. Kim, and H. Kim, “Music Identification Using Pitch Histogram and MFCC-VQ Dynamic Pattern” (in Korean), J. Acoust. Soc. Kr. **24**, 178-185 (2005).
24. J. Lee, “How similar is too similar?: Exploring users' perceptions of similarity in playlist evaluation,” in Proc. ISMIR-2011, 109-114 (2011).
25. A. Novello, M. M. F. McKinney, and A. Kohlrausch,

"Perceptual evaluation of inter-song similarity in western popular music," J. New Music Res. **40**, 1-26 (2011).

저자 약력

▸ 서 진 수 (Jin Soo Seo)

1998년 2월: KAIST 전기 및 전자공학과 공학사
2000년 2월: KAIST 전기 및 전자공학과 공학석사
2005년 2월: KAIST 전기 및 전자공학과 공학박사
2005년 3월 ~ 2006년 2월: KAIST 정보전자 연구소 연구원
2006년 3월 ~ 2008년 2월: 한국전자통신 연구원 디지털콘텐츠 연구단 선임연구원
2008년 3월 ~ 현재: 강릉원주대학교 전자공학과 조교수, 부교수