Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 15, No. 3, September 2015, pp. 159-165
http://dx.doi.org/10.5391/IJFIS.2015.15.3.159

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Ranking Tag Pairs for Music Recommendation Using Acoustic Similarity

Jaesung Lee and Dae-Won Kim
School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea

ljfis

Abstract

The need for the recognition of music emotion has become apparent in many music information retrieval applications. In addition to the large pool of techniques that have already been developed in machine learning and data mining, various emerging applications have led to a wealth of newly proposed techniques. In the music information retrieval community, many studies and applications have concentrated on tag-based music recommendation. The limitation of music emotion tags is the ambiguity caused by a single music tag covering too many subcategories. To overcome this, multiple tags can be used simultaneously to specify music clips more precisely. In this paper, we propose a novel technique to rank the proper tag combinations based on the acoustic similarity of music clips.

Keywords: Music emotion annotation, Acoustic feature extraction, Music emotion recognition

1. Introduction

Recently, the number of music clips encountered in daily life has grown rapidly with social network services for music, giving popularity to music categorization based on music tags such as genre and theme. In addition to conventional tags, many people often categorize music by mood or expressed emotions, ranging from happiness to sadness. This is why music emotion recognition (MER) has gained popularity and has been applied to music information retrieval for improving the effectiveness of several applications that search for and recommend music [1]. The goal of MER is to identify the intended or perceived music emotion of a given music piece [2–8]. The process of MER can be described as 1) annotating music emotions, 2) extracting acoustic features, and 3) recognizing music emotions [9]. One aim of this paper is to discuss the issues related to each process and their effects on MER performance by reviewing different MER techniques.

However, the limitation of music tags is the ambiguity that comes from the fact that a single music tag covers too many subcategories [4]. To overcome this, users may use multiple tags simultaneously to specify their target music clips more precisely. Because this also increases the number of possible tags to be considered by users, the recommendation system may allow or suggest a proper subset of tag combinations for users. In this paper, we propose a novel technique to rank the proper tag combinations based on acoustic similarity of music clips.

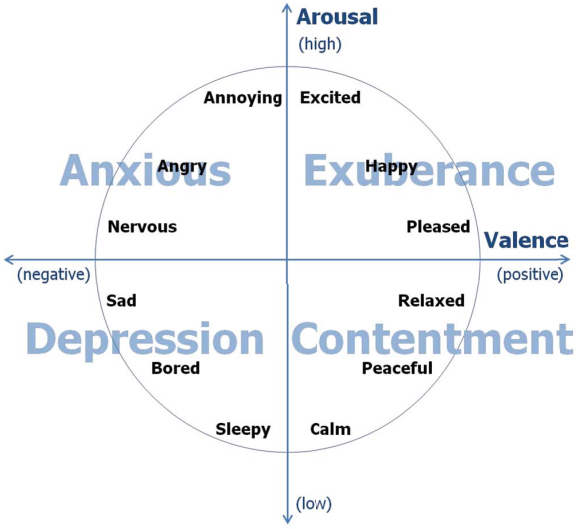
2. Music Emotion Recognition

2.1 Annotating Music Emotion

Received: Aug. 18 2015
Revised : Sep. 22, 2015
Accepted: Sep. 24, 2015

Correspondence to: Dae-Won Kim
(dwkim@cau.ac.kr)

(a)



(b)

Figure 1. Two common emotion representation schemes: (a) Hevner's adjective list, (b) Thayer's emotional circumflex model.

One of the most important issues of MER is how to describe music emotions. In early psychological research, Hevner proposed an adjective checklist that was divided into eight mood clusters to represent music emotions [3]. This study was performed with ordinary people from several cultural backgrounds based on adjectives that were selected by subjects. As shown in Figure 1(a), Hevner summarized the words used by the subjects to an impressive 66 adjectives, and these words were further extended [10]. Because of its intuitive representation, there are a series of studies using these emotional adjectives to represent music emotions [3].

Instead of describing music emotions as a set of adjectives, Thayer proposed a circumflex model of music emotion by adapting a basic motion model to music using a two-dimensional energy-stress emotion model [4]; this dimensional approach indicates two underlying stimuli involved in music emotion responses (Figure 1(b)). Based on the level of stress and energy, Thayer's emotional model (TEM) divides music emotions into four clusters (each in one quadrant): contentment, depression, exuberance, and anxiety. Although a considerable number of works have been conducted based on TEM [11, 12], the limitation of TEM is that it is not intuitive to untrained subjects. To represent music emotions using this model, the subject must understand the role of each axis. Thus, most research using TEM has used a small number of musical experts with substantial knowledge about musicology and psychology. As a result, it may not be scalable to the real-world music corpus.

In MER research, music emotions are often divided into two categories: intended emotion and personalized emotion [1]. Intended emotion (IE) is usually described by the performer or songwriter, primarily in the way they express their feelings. When a listener hears those music clips, they may evoke certain feelings based on cultural agreement and personal experience, namely, personalized emotion (PE). Although most current MER research focuses on IE, a few studies have attempted to directly attack the difficulty involving PE; these studies have attempted to solve or avoid the subjectivity problem associated with cultural agreement or modeling the individuality of listeners [1, 13]. However, this also reduces the coverage of the MER system; for example, the MER system may specialize into western classical music [5]. Because of the subjective nature of music emotion, it is difficult to obtain a general response to the same music clip [13]. Typically, the responses of music emotion are obtained by two approaches: one is annotation by music experts, and the other is annotation by crowdsourcing [14]. Annotation by experts refers to emotional responses obtained by experts who have been trained in musicology. Therefore, it is difficult to recruit a large number of such experts. In contrast, annotation by crowdsourcing refers to emotional response collected from social tags, resulting in difficult quality control of the gathered responses.

2.2 Extracting Acoustic Features

Different emotional expressions are usually associated with different patterns of acoustic music signals [2]. It has been recog-

nized that some acoustic features extracted from music signals are typically relevant to music emotions: dynamics, timbre, harmony, register, rhythm, and articulation [9]. It is noteworthy that although the exact words and symbols are different, such as rhythm versus tempo and intensity versus sound level, the meanings of these concepts are very similar. Here, we briefly explain the basic principle of how to extract those features. Timbral features are based on a spectrogram with statistically divided stationary frames. Some features are developed for timbre analysis in current literature, such as Mel-frequency cepstral coefficients, short-time energy, and zero-crossing rate [16]. Moreover, feature extraction methods specialized in music signal analysis have also been proposed [3]. In addition to signal-level extractions, high-level musical properties such as rhythm, harmony, and articulation may contribute to expressing music emotions. These musical properties are extracted by using acoustic feature recognizers that analyze music signals based on musical theories [2, 12, 17]. For example, harmony features are extracted from a viewpoint of spectrogram roughness, harmonic change, key clarity, and majorness. To extract these acoustic features, readers may use acoustic feature extraction tools such as MIRToolbox [16]. These tools have demonstrated effectiveness in several MER studies [4, 5, 8].

Several signal transforming techniques have been proposed to handle acoustic waveform inputs of music clips. Although individual acoustic features have been tested and are shown to have emotion representation power, it is known that optimal MER performance can be achieved by using them in combination. Therefore, the issues of how to select the optimal acoustic feature subset should be addressed carefully because too many acoustic features can also degrade the performance of MER [20].

2.3 Recognizing Music Emotion

In the psychological domain, two words, mood and emotion, have different meanings [6]; emotion refers to a strong response of a relatively short duration, whereas mood indicates a long-term response. Most MER research has assumed that if a music clip is segmented in statistically stable frames, then this music clip expresses a unique emotion [19]. For each music segment in each frame, the MER system was trained to detect the emotion type in each segment. In contrast, an MER system is able to assume that music emotion is continuously changing according to time [10]. This approach expresses the emotional content of a music clip as a function of time-varying musical features [5].

A second issue that should be considered in MER systems is the ambiguous nature of music emotion. It is easily understood that people may use multiple adjectives to describe the emotion of a music clip. To solve this problem, MER systems employed fuzzy or multi-label schemes [2, 8]. However, it can be claimed that even when multiple adjectives are allowed to describe music emotion, the categorical taxonomy of emotion is still inherently ambiguous [4]. For example, the first quadrant of TEM contains emotional adjectives such as excited, happy, and pleased, which are different in nature. This ambiguity confuses the subjects in the subjective test and confuses the users when retrieving a music piece according to their emotional states. An alternative is to view the emotion plane as a continuous space and recognize each point of the plane as an emotional state.

The last issue originated from the prediction difficulty for each stimulus of music emotion. For example, an MER system may take the form of a hierarchical structure [6]; according to empirical results, arousal (or energy) in TEM is more computationally tractable and can be estimated using simple amplitude-based acoustic features. In the experiments, a hierarchical MER system outperformed its non-hierarchical variant. Moreover, support vector machines (SVMs) were employed to detect complicated relations between the distribution of music features and the music emotion [20]. Based on the empirical tests, they reported that arousal of music emotion was predicted quite accurately (95%), whereas the overall prediction accuracy was degraded owing to the difficulty of predicting valence. Moreover, it has been found generally that valence is much more difficult to predict than arousal [13]; based on the comparison between group-wise experiments and individual experiments, it has been observed that the difference for valence is larger than that for arousal.

3. Tag Combinations for Recommendation

To describe our proposed method for discovering proper tag combinations, we first introduce some mathematical definitions. Let $x \in X$ be a music clip in the music corpus $X \subset R^d$ represented by d acoustic features. For each music clip x, there is a set of tags $y \in Y$ and $Y = \{y_1, \ldots, y_q\}$, where q is the maximum number of possible tags. Therefore, a music clip x_i and annotated tags y_i can be represented as (x_i, y_i). In the proposed method, the tag combinations are filtered by two conditions: the number of music clips covered by a corresponding tag combination and the utility of the tag combination. Let $S = \{(x, y)|y = R\}$ be a set of music clips annotated by tags

$R \subseteq Y$. The number of music clips $N(R)$ covered by R is then denoted as [21],

$$N(R) = |S| \tag{1}$$

where $|\cdot|$ represents the cardinality of the given set. After all possible tag combinations are identified, tag combinations with higher $N(\cdot)$ value are selected and used for suggestions. However, it is computationally inefficient to consider all possible tag combinations because there can be $2^q - 1$ possible combinations of tags. As the number of tags in a combination grows, the number of music clips covered by that tag combination decreases. Therefore, the effectiveness of music recommendation by that tag combination will decrease because the tag combinations cover an overly specific case of combination. In this paper, we actually examined tag combinations that are composed of no more than two music tags and then selected tag combinations that cover more than 10% of music clips in the music database.

Because our goal for combining tags is to specify music clips in more detail, the acoustic similarity of music clips annotated by multiple tags should be larger than that of music clips annotated by each tag. In this paper, we propose a new ranking method for tag combinations based on the acoustic similarity of music clips. Let $D(R)$ be the acoustic dissimilarity of music clips, denoted as:

$$D(R) = \frac{1}{|S|} \sum_{x \in S} \sqrt{\left(S - \bar{S}\right)^2} \tag{2}$$

where $\bar{S}$ is the centroid of music clips in S. The utility of the tag pair–i.e., the benefit of combining the tags–can then be denoted as follows:

$$U(R) = \frac{D(R)}{\min\left(D(r_1), D(r_2)\right)} \tag{3}$$

where r_i is the ith tag in R. The utility of a tag pair evaluates the specification power of a given tag pair; i.e., the value of $U(R)$ decreases as the acoustic similarity of music clips annotated by R increases. After filtering tag pairs by corresponding $N(R)$ values, the proposed method outputs ranked tag pairs based on $U(R)$.

Table 1. Top five tag pairs with highest utility value

Rank	Tag pair	$D(r_i)$	$D(R)$	$N(R)$	$U(R)$
1	T01	6.92	6.15	51	0.89
	T02	6.90			
2	T03	7.14	5.78	51	0.90
	T04	6.41			
3	T05	7.30	6.61	54	0.91
	T06	7.58			
4	T07	7.63	6.62	55	0.91
	T08	7.29			
5	T09	7.29	6.28	58	0.91
	T10	6.89			

T01, Not-Emotion-Calming-Soothing; T02, Usage-Driving; T03, Not-Emotion-Tender-Soft; T04, Genre-Classic-Rock; T05, Song-Catchy-Memorable; T06, Not-Song-Changing-Energy-Level; T07, Not-Emotion-Light-Playful; T08, Song-Heavy-Beat; T09, Not-Emotion-Touching-Loving; T10, Instrument-Electric-Guitar.

Table 2. Top five tag pairs with highest utility value

Rank	Tag pair	$D(r_i)$	$D(R)$	$N(R)$	$U(R)$
889	T11	7.76	8.58	61	1.11
	T12	8.47			
890	T13	7.26	8.02	65	1.11
	T14	7.64			
891	T15	7.84	8.52	57	1.11
	T16	7.67			
892	T17	7.76	7.50	59	1.11
	T18	6.75			
893	T19	8.00	8.54	89	1.11
	T20	7.67			

T11, Not-Emotion-Angry-Aggressive; T12, Not-Emotion-Powerful-Strong; T13, Not-Emotion-Bizarre-Weird; T14, Not-Emotion-Happy; T15, Emotion-Tender-Soft; T16, Not-Song-Very-Danceable; T17, Not-Emotion-Angry-Aggressive; T18, Genre-Rock; T19, Emotion-Calming-Soothing; T20, Not-Song-Very-Danceable.

4. Results

To verify the effectiveness of the proposed method, we employed a CAL500 dataset that is well known in the music information retrieval community [22]; the CAL500 dataset is composed of 502 music clips, 68 acoustic features, and 174 corresponding tags that are annotated by 66 undergraduate students of California University in the USA. Among 15,000 possible tag pairs, 14,000 tag pairs were filtered at the first step because they covered less than 10% of music clips. Table 1 shows the top five tag pairs selected by our ranking method.

Table 1 contains six columns; each column represents the

Table 3. Summarization of experimental setup and music emotion recognition results

Ref.	ET	TE	MD	NA	AF	NM	TVE	Detect	Perf.
[2]	AL	IE	Pop	-	Tempo, Articulation	353	Static	Fuzzy	67%
[3]	AL	IE	Jazz	2	Timber, MFCC, DWCH	235	Static	Multi	83%
[6]	EM	IE	Classic	3	Intensity, Timbre, Rhythm	800	Dynamic	Single	86%
[5]	EM	IE	Western	35	Dynamics, Pitch, Timbre, Harmony, Tempo, Texture	6	Dynami	Real	78% (A), 22% (V)
[19]	EM	IE	Classic, Pop, Jazz, Hip-Hop, Punk	3	Timbre, Tempo	800	Static	Single	91%
[7]	EM	IE	12 Major genres	12	Tempo, Rhythm, Tonal	1059	Static	Multi	85%
[13]	EM	PE	English Pop	40	Timbre, Tonal, MFCC, Rhythm, Pitch	60	Static	Real	72% (A), 19% (V)
[8]	EM	IE	Classic, Reggae, Rock, Pop, Hip-Hop, Techno, Jazz	30	Rhythm, Timbre	593	Static	Multi	82%
[4]	EM	PE	Western, Chinese, Japanese	253	Timbre, Tonal, Structure, MFCC, Rhythm, Pitch	195	Static	Real	58% (A), 28% (V)
[12]	EM	IE	Sound Track	116	Timbre, Harmony, Register, Rhythm, Articulation, Structure	110	Static	Real	85% (A), 72% (V), 79% (T)

ET, emotional taxonomy; AL, adjective list; EM, emotional state model; TE, types of emotion; IE, intended emotion; PE, personalized emotion; MD, music domain covered by this research; NA: number of annotators participating in annotation of given music clips; AF, acoustic features extracted from given music clips; MFCC, Mel-frequency cepstral coefficient; DWCH, Daubechies wavelet coefficients histogram; NM, number of music clips considered; TVE, time-varying emotion considered; Detect, how to consider music emotion; Perf., recognition performance; A, arousal; V, valence; T, tender.

rank of each tag pair, the name of each tag in the tag pair, the acoustic dissimilarity value of each tag, the acoustic dissimilarity value of the tag pair, the number of music clips, and the utility value of the tag pair. For example, the best tag pair is composed of (Not-Emotion-Calming-Soothing, Usage-Driving) because it reduces the acoustic dissimilarity of music clips by approximately 11% compared to that of music clips annotated with (Usage-Driving) alone.

Table 2 shows the characteristics of five low-ranked tag pairs. Because $U(R)$ of each tag pair exceeds 1, the experimental results indicate that the acoustic dissimilarity of music clips by given music tags is increased by approximately 11%. For example, the acoustic dissimilarity of music clips based on (Not-Emotion-Angry-Aggressive, Not-Emotion-Powerful-Strong) is higher than that of music clips based on (Not-Emotion-Angry-Aggressive). Thus, the experimental results indicate that there is a set of music tag pairs that may not be helpful for specifying desired music clips in terms of acoustic similarity, and they can be effectively filtered by our proposed ranking technique for tag pairs.

5. Conclusion

In this paper, many techniques involved in MER were presented. Despite the subject nature of music emotion and ambiguity in the emotional state, there were many favorable achievements in this domain. As shown in Table 3, a common effort of MER research was invested to solve the ambiguity in music emotions using music emotion models and detecting real valued emotional states. Moreover, to enlarge the applicability of MER systems, research tends to recognize a wide range of music. To recognize a wide range of music, researchers have focused on more relevant and novel acoustic features and incorporated more features into their research.

Although music emotion research has shown how to solve the intrinsic problems in the MER domain, such as subjectivity of music emotion, time-varying emotion, and ambiguous problems, there are some limitations due to unsolved problems. Extensive and fruitful efforts have been made in recent years in the disambiguation of music emotions by emotion state modeling. In general, we observed that many researchers have successfully used TEM. This model eliminates the ambiguity in music emotion by sacrificing the intrinsic use of adjectives. For example, if we attempt to model emotion through the dimensional model, then the problem of how to exactly measure the quantity of valence or arousal arises. Thus, there still needs to be an advancement of the emotional model or the development of a new emotional model for describing and annotating music emotion. In this case, a hybrid model that combines adjectives on emotional space may be an attractive solution.

Moreover, we proposed a new method of ranking music tag combinations for music categorization. Experimental results demonstrated that our proposed method is able to rank the proper tag combinations based on acoustic similarity of music clips and filter tag combinations which leads to acoustically inconsistent music clips.

Acknowledgments

This research is supported by the Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2015.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] Y. Yang and H. Chen, "Ranking-Based Emotion Recognition for Music Organization and Retrieval," *IEEE Transactions on Audio, Speech,and Language Processing*, vol. 19, no. 4, pp. 762-774, Aug. 2010.

[2] Y. Feng, Y. Zhuang, and Y. Pan, "Music Information Retrieval by Detecting Mood via Computational Media Aesthetics," in *Proceedings of the IEEE/WIC International Conference on Web Intelligence*, Halifax, 2003, pp. 235-241.

[3] T. Li and M. Ogihara, "Toward Intelligent Music Information Retrieval," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 564-574, Jun. 2006.

[4] Y. Yang, Y. Lin, Y. Su, and H. Chen, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448-457, Feb. 2008.

[5] M. Korhonen, D. Clausi, and M. Jernigan, "Modeling Emotional Content of Music Using System Identication," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 3, pp. 588-599, Jun. 2006.

[6] L. Lu, D. Liu, and H. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5-18, Jan. 2006.

[7] J. Skowronek, M. McKinney, and S. Van De Par, "A Demonstrator for Automatic Music Mood Estimation," in *Proceedings of International Conference on Music Information Retrieval*, Vienna, 2007, pp. 345-346.

[8] K. Tsoumakas, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label Classication of Music into Emotions," in *Proceedings of International Conference on Music Information Retrieval*, Philadelphia, 2008, pp. 325-330.

[9] Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Trunbull, "Music Emotion Recognition: A State of the art review," in *Proceedings of International Conference on Music Information Retrieval*, Utrecht, 2010, pp. 255-266.

[10] E. Schmidt, D. Turnbull, and Y. Kim, "Feature Selection for Content-Based, Time-Varying Musical Emotion Regression," in *Proceedings of International Conference on Music Information Retrieval*, Utrecht, 2010, pp. 267–274.

[11] X. Zhu, Y. Shi, H. Kim, and K. Eom, "An Integrated Music Recommendation System," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 917-925, Aug. 2006.

[12] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of Multidimensional Emotional Ratings in Music from Audio using Multivariate Regression Models," In *Proceedings of International Conference on Music Information Retrieval*, Kobe, 2009, pp. 621-626.

[13] Y. Yang, Y. Su, Y. Lin, and H. Chen, "Music Emotion Recognition: The Role of Individuality," in *Proceedings of the International Workshop on Human-centered Multimedia*, Augsburg, 2007, pp. 13-22.

[14] M. Soleymani, M. Caro, and E. Schmidt, C. Sha, and Y. Yang, "1000 Songs for Emotional Analysis of Music," In *Proceedings of ACM International Workshop on Crowdsourcing for Multimedia*, Barcelona, 2013, pp. 1-6.

[15] K. Bischoff, C. Firan, W. Nejdl, and R. Paiu, "How Do You Feel about Dancing Queen? Deriving Mood and Theme Annotations from User Tags," In *Proceedings of ACM/IEEE-CS Joint Conference on Digital Libraries*, Austin, 2009, pp. 285-294.

[16] O. Lartillot, and P. Toiviainen, "MIR in MATLAB (II): A toolbox for musical feature extraction from audio," in *Proceedings of International Conference on Music Information Retrieval*, Vienna, 2007, pp. 237-244.

[17] M. Ruxanda, B. Chua, A. Nanopoulos, C. Jensen, "Emotion-based Music Retrieval on a Well-reduced Audio Feature Space," In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, 2009, pp. 181-184.

[18] B. Zhu and K. Zhang, "Music Emotion Recognition System Based on Improved GA-BP," In *Proceedings of International Conference on Computer Design and Applications*, Qinhuangdao, 2010, pp. 409-412.

[19] M. Wang, N. Zhang, H. Zhu, "User-adaptive Music Emotion Recognition," In *Proceedings of International Conference on Signal Processing*, Beijing, 2004, pp. 1352-1355.

[20] B. Rocha, R. Panda, and R. Paiva, "Music Emotion Recognition: The Importance of Melodic Features," In *Proceedings of International Workshop on Machine Learning and Music*, Prague, 2013, pp. 1-4.

[21] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," In *Proceedings of International Conference on Very Large Data Bases*, Santiago, 1994, pp.487-499.

[22] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 467-476, Feb. 2008.

Jaesung Lee received the M.S. and Ph.D in computer science from Chung-Ang University, Korea in 2009 and 2013 respectively. He participates in Post-Doc. course in the School of Computer Science and Engineering, Chung-Ang University in Seoul. His Research interest includes biomedical informatics and affective computing. In theoretical domain, he also studies classification, feature selection, and multi-label learning with information theory.
E-mail: jslee.cau@gmail.com

Dae-Won Kim is currently a professor in the School of Computer Science and Engineering, Chung-Ang University in Seoul, Korea. Prior to coming to Chung-Ang University, he did his Post-Doc., Ph.D., M.S. at KAIST, and the B.S. at Kyungpook National University, Korea. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.
E-mail: dwkim@cau.ac.kr