

# Searching Music in the Emotion Plane

Yi-Hsuan Yang and Homer H. Chen, National Taiwan University, Taiwan

affige@gmail.com, homer@cc.ee.ntu.edu.tw

Music plays an important role in human's history, even more so in the digital age. Never before has such a large collection of music been created and accessed daily by people. Because almost all music is created to convey emotion, music organization and retrieval by emotion is a meaningful way for accessing music information. The proliferation of tiny mobile devices and the like also calls for content-based retrieval of music through a small display space.

Music emotion recognition (MER) aims at recognizing the affective content of music signals. A typical approach is to categorize emotions into a number of classes (e.g., happy, angry, sad and relaxing) and apply machine learning techniques to train a classifier [1]–[3]. This approach, though widely adopted, faces the granularity issue in practice, because classifying emotions into only a handful of classes cannot meet the user demand for effective information access. Using a finer granularity for emotion description does not necessarily address the issue since language is inherently ambiguous, and the description for the same emotion varies from person to person.

Instead, we propose to view emotions from a dimensional perspective and define emotions in a 2-D plane in terms of arousal (how exciting or calming) and valence (how positive or negative), the two emotion dimensions found to be most fundamental by cognitive study [4]. In this way, MER becomes the prediction of the arousal and valence (AV) values of a song corresponding to a point in the emotion plane [5]–[8]. The granularity and ambiguity issues associated with emotion classes no longer exist since no categorical classes are needed. Moreover, because the 2-D emotion plane provides a simple means for user interface, novel emotion-based music organization, browsing, and retrieval can be easily created for mobile devices.

## 1. Emotion-Based Retrieval

The advantages of the emotion-based approach are that each music sample can be represented as a point in the emotion plane and that the similarity between music samples can be measured by Euclidean distance. As shown in Fig. 1, a user can retrieve music of a certain emotion by simply specifying a point in the emotion plane. The system then returns the music samples whose AV values are close to the point. A user can also generate an emotion-based playlist by drawing a trajectory in the emotion plane. This way, songs of various emotions corresponding to different points on the trajectory are added to the playlist and played back in order.

One can also couple other musical metadata such as artist name, genre, or lyrics with emotion to narrow down the search range. For example, one can specify an artist, and the system would display all songs of the artist in the emotion plane. It is also possible to playback music that matches the user's mood detected by using physiological, prosodic, or facial cues [9]. This retrieval paradigm is functionally powerful since people's criterion is often related to the emotion state at the moment of music selection [10].

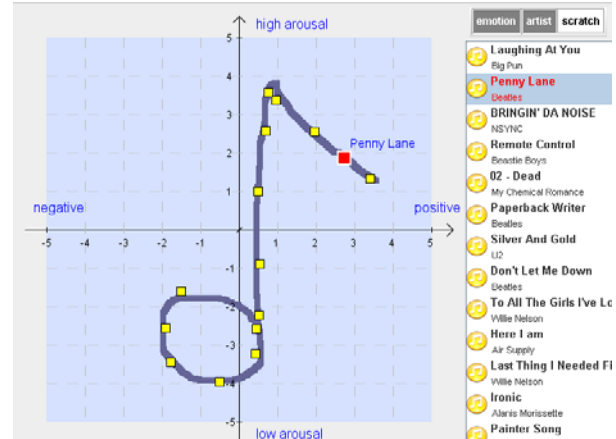


Fig. 1. With emotion-based music retrieval, a user can retrieve music of certain emotions by specifying a point or drawing a trajectory in the 2-D emotion plane [5], [6].

## 2. Emotion Recognition

MER can be formulated as a regression problem [5] by viewing arousal and valence as real values in  $[-1, 1]$ . Then a regression model can be trained to predict the AV values. More specifically, given  $N$  inputs  $(x_i, y_i)$ ,  $1 \leq i \leq N$ , where  $x_i$  is a feature vector of the  $i$ th input sample, and  $y_i$  is the real value to be predicted, a regression model (regressor)  $R(\cdot)$  is created by minimizing the mismatch (i.e., mean squared difference) between the predicted and the ground truth values. Many good regression algorithms, such as support vector regression (SVR) or Gaussian process regression [11], are readily available. In [5], two SVR models are trained for arousal and valence respectively. A schematic diagram of this MER system is shown in Fig. 2.

Usually, timbral, rhythmic, melodic, and harmonic features of music are extracted to represent the acoustic property of a song. Because of its ability to model auditory sensation based on psychoacoustic models, the computer program PsySound [12] is often employed for feature extraction. The use of mid-level features such as chord progression or genre metadata has also been explored [13], [14]. Many features such as loudness (loud/soft), tempo (fast/slow), and pitch (high/low) have been found relevant to arousal, but only few features are relevant to valence. Thus, valence recognition is more challenging than arousal recognition.

Typically a subjective test is conducted to collect the ground truth needed for model training. The subjects are asked to annotate the music pieces by rating their emotion perception of the music pieces using either the standard ordinal rating scale or the graphic rating scale [5], [15]. Because emotion perception is subjective, each music piece is annotated by multiple subjects and the ground truth is set to the average rating.

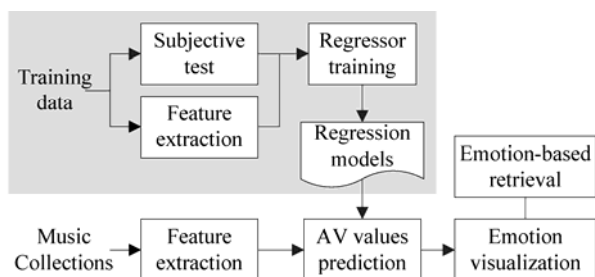


Fig. 2. The schematic diagram of a MER system [5].

### 3. Challenges

As MER is still in its infancy, there are many open issues. Some major issues and proposed solutions are discussed in this section.

#### 3.1 Subjectivity of Emotion Perception

Emotion perception is intrinsically under the influence of many factors such as cultural background, generation, sex, and personality. Developing a general retrieval model that performs equally well for everyone is a challenging task. This can be explained via Fig. 3, where each circle corresponds to the annotation of a song in the emotion plane by a subject. Obviously, simply assigning one emotion value to each song in a deterministic manner does not work well in practice because the emotion perception varies greatly from person to person.

The subjectivity issue can be addressed by *personalizing* the MER system [7], [15]. We can ask a user to annotate a small number of songs and use the annotations to train a personalized model. A two-stage personalization scheme is proposed in [7]. Two models are trained: one for predicting the general perception of a song, and the other for predicting the difference between the general perception and a user's individual perception. This is a simple personalization process because the music content and the individuality of the user are treated separately. To make it more sophisticated, one can take into account the demographic property, music preference, or listening context of the user in the process.

#### 3.2 Difficulty of Emotion Annotation

The emotion annotation process of MER requires the subjects to rate the emotion in a continuum. But it has been found that such rating imposes a heavy cognitive load to the subjects [8]. In addition, it is difficult to ensure a consistent rating scale between and within the subjects [16]. As a result, the quality of the ground truth varies, which in turn degrades the accuracy of MER.

To address this issue, ranking-based emotion annotation is proposed [8]. A subject is asked to compare the affective content of two songs and determine, for example, which song has a higher arousal value, instead of the exact emotion values. The rankings of music emotion are then converted to numerical values by a greedy algorithm [17]. Empirical evaluation shows that this scheme relieves the burden of emotion annotation on the subjects and enhances the quality of the ground truth. It is also possible to use an online game to harness the so-called *human computation* and make the annotation process more engaging [18].

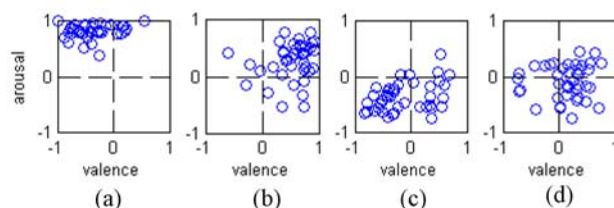


Fig. 3. Emotion annotations in the emotion plane for four songs: (a) Smells like teen spirit by Nirvana, (b) A whole new world by Peabo Bryson and Regina Belle, (c) The rose by Janis Joplin, and (d) Tell Laura I love her by Ritchie Valens. Each circle corresponds to the annotation of a song by a subject [7].

#### 3.3 Semantic Gap Between Audio Signal and Human Perception

The viability of an MER system largely lies in the accuracy of emotion recognition. However, due to the semantic gap between the object feature level and the human cognitive level of emotion perception, it is difficult to accurately compute the emotion values, especially the valence values. What intrinsic element of music, if any, causes a listener to create a specific emotional response is still far from well-understood. While mid-level audio features such as chord, rhythmic patterns, and instrumentation carry more semantic information, robust techniques for extracting such features need to be developed.

Available data for MER are not limited to the raw audio signal. Complementary to music signal, lyrics are semantically rich and have profound impact on human perception of music [19]. It is often easy for us to tell from the lyrics whether a song expresses sadness or happiness. Incorporating lyrics to MER is feasible because most popular songs sold in the market come with lyrics [20]. One can analyze lyrics using natural language processing to generate textual feature descriptions of music. It has been shown that using lyrics indeed improves valence recognition [21], [22].

### 4. Conclusion

The past decade has witnessed a growing interest in analyzing the affective content of music. In this article, we have described a new music retrieval paradigm that allows users to search music in the emotion plane. It opens up a new playground for advanced research on music emotion recognition and understanding.

#### Acknowledgments

This work was supported by the National Science Council of Taiwan under the contract number NSC 97-2221-E-002-111-MY3.

#### References

- [1] T. Li and M. Ogihara, "Detecting emotion in music," in *Proc. ISMIR*, 2003.
- [2] L. Lu et al., "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.
- [3] X. Hu et al., "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. ISMIR*, pp. 462–467, 2008.
- [4] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York, Oxford University Press, 1989.

- [5] Y.-H. Yang et al, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.
- [6] Y.-H. Yang et al, "Mr. Emo: Music retrieval in the emotion plane," in *Proc. ACM Multimedia*, pp. 1003–1004, 2008.
- [7] Y.-H. Yang et al, "Personalized music emotion retrieval," in *Proc. ACM SIGIR*, pp. 748–749, 2009.
- [8] Y.-H. Yang and H. H. Chen, "Music emotion ranking," in *Proc. ICASSP*, pp. 1657–1660, 2009.
- [9] T.-L. Wu et al, "Interactive content presenter based on expressed emotion and physiological feedback," in *Proc. ACM Multimedia*, pp. 1009–1010, 2008.
- [10] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*. Oxford: Oxford University Press, 2001.
- [11] A. Sen and M. Srivastava, *Regression Analysis: Theory, Methods, and Applications*. New York, Springer, 1990.
- [12] D. Cabrera, "PSYSOUND: A computer program for psychoacoustical analysis," in *Proc. Australian Acoustic Society Conf.*, pp. 47–54, 1999. <http://psysound.wikidot.com/>.
- [13] H.-T. Cheng et al, "Automatic chord recognition for music classification and retrieval," in *Proc. ICME*, pp. 1505–1508, 2008.
- [14] Y.-C. Lin et al, "Exploiting genre for music emotion classification," in *Proc. ICME*, pp. 618–621, 2009.
- [15] Y.-H. Yang et al, "Music emotion recognition: The role of individuality," in *Proc. ACM Int. Workshop on Human-Centered Multimedia*, pp. 13–21, 2007.
- [16] S. Ovadia, "Ratings and rankings: Reconsidering the structure of values and their measurement," *Int. J. Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.
- [17] W. W. Cohen et al, "Learning to order things," *J. Artificial Intelligence Research*, vol. 10, pp. 243–270, 1999.
- [18] Y. E. Kim et al, "Moodswings: A collaborative game for music mood label collection," in *Proc. ISMIR*, 2008.
- [19] S. Omar Ali et al, "Songs and emotions: Are lyrics and melodies equal partners," *Psychology of Music*, vol. 34, no. 4, pp. 511–534, 2006.
- [20] J. Fornäs, "The words of music," *Popular Music and Society*, vol. 26, no. 1, pp. 37–53, 2003.
- [21] Y.-H. Yang et al, "Toward multi-modal music emotion classification," in *Proc. PCM*, pp. 70–79, 2008.
- [22] C. Laurier et al, "Multimodal music mood classification using audio and lyrics," in *Proc. ICMLA*, pp. 1–6, 2008.



**Yi-Hsuan Yang** received the B.S degree in Electrical Engineering from National Taiwan University, Taiwan, in 2006. He is currently working toward the Ph.D. degree in the Graduate Institute of Communication Engineering, National Taiwan University. His research interests include multimedia information retrieval and analysis, machine learning, and affective computing. He has published over 20 technical papers in the above areas.

Mr. Yang is a Microsoft Research Asia Fellowship recipient 2008–2009.



**Homer H. Chen** (S'83-M'86-SM'01-F'03) received the Ph.D. degree in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign, Urbana.

Since August 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University, Taiwan, R.O.C., where he is Irving T. Ho Chair Professor. Prior to that, he held various R&D management and engineering positions with US companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island. He was a US delegate for ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now part of the MPEG-4 and JPEG-2000 standards. His professional interests lie in the broad area of multimedia signal processing and communications.

Dr. Chen is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology. He served as Associate Editor of IEEE Transactions on Image Processing from 1992 to 1994, Guest Editor of IEEE Transactions on Circuits and Systems for Video Technology in 1999, and an Associate Editorial of Pattern Recognition from 1989 to 1999.