Music Emotion Classification: A Fuzzy Approach

Yi-Hsuan Yang, Chia-Chu Liu, and Homer H. Chen
Graduate Institute of Communication Engineering, National Taiwan University
1 Roosevelt Rd. Sec.4, Taipei, 10617, Taiwan R.O.C.
886-2-3366-3549
b91901109@ntu.edu.tw, b90901034@ntu.edu.tw, homer@cc.ee.ntu.edu.tw

ABSTRACT
Due to the subjective nature of human perception, classification of the emotion of music is a challenging problem. Simply assigning an emotion class to a song segment in a deterministic way does not work well because not all people share the same feeling for a song. In this paper, we consider a different approach to music emotion classification. For each music segment, the approach determines how likely the song segment belongs to an emotion class. Two fuzzy classifiers are adopted to provide the measurement of the emotion strength. The measurement is also found useful for tracking the variation of music emotions in a song. Results are shown to illustrate the effectiveness of the approach.

Categories and Subject Descriptors
H.5.5 [**Sound and Music Computing**]: *systems*

General Terms
Algorithms, Performance, Experimentation, Human Factors.

Keywords
Music emotion strength, Fuzzy vector, Fuzzy k-NN (FKNN), Fuzzy Nearest-Mean (FNM), Model generator (MG), Emotion classifier (EC), Music emotion variation detection (MEVD).

1. INTRODUCTION
Music is important to our daily life. The influence of music becomes more profound as we enter the digital world. As the music databases grow, more efficient organization and search methods are needed. Music classification by perceived emotion is one of the most important research topics, for it is content-based and functionally more powerful.

Due to the subjective nature of human perception, classification of the emotion of music is a challenging problem. Listening mood, environment, personality, age, cultural background etc, can influence the emotion perception. Because of these factors, classification methods that simply assign one emotion class to each song in a deterministic manner do not perform well in practice [1], [2], [3].

The subjective nature of emotion perception suggests that fuzzy logic is a more appropriate mathematical tool for emotion detection [4]. We employ two fuzzy classifiers in our music classification system to measure the strength of an emotion class in association with the song under classification. Based on the measurement, people can know how likely a song segment belongs to an emotion class and use it to track the variation of emotions in a song. To our best knowledge, this paper represents one of the first attempts that take the subjective nature of human perception into consideration for music emotion classification.

The paper is organized as follows. In Section 2, we introduce the taxonomy of emotion used in our work. Section 3 gives an overview of the classification system. Experimental results are given in Section 4, and conclusions and extensions in Section 5.

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010

2. TAXONOMY
In our system, we adopt Thayer's model [5] for the description of emotions. As shown in Figure 1, the 2D emotion space (2DES) is divided into 4 quadrants, and different emotions are placed on the plane in such a way that each emotion (a point in 2DES) can be represented by a 2x1 vector. This results in a valence-arousal plane. The right (left) side of the plane refers to the positive (negative) emotion, whereas the upper (lower) side refers to the energetic (silent) emotion. To be consistent with the 2DES model, we define 4 emotion classes, each corresponding to a quadrant.

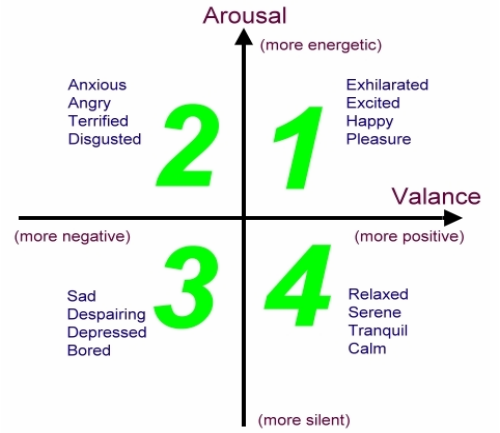

Figure 1. Thayer's model of mood.

3. SYSTEM OVERVIEW
The proposed system can be divided into two parts: model generator (MG) and emotion classifier (EC). The MG generates a model according to the features of the *training samples*, while the EC applies the resulting model to classify the *input samples*. Block diagrams are given in Figures 2 and 3, while details are described in the following sub-sections.

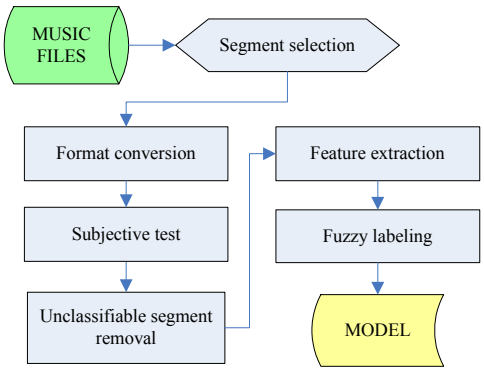

Figure 2. Block diagram of model generator.

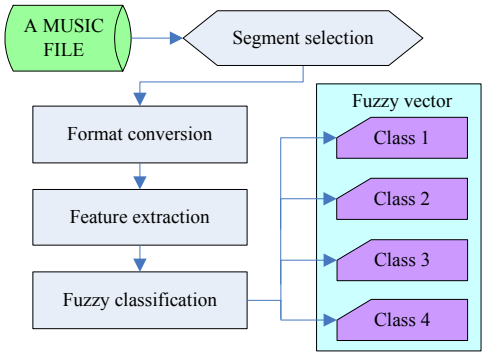

Figure 3. Block diagram of emotion classifier.

Table 1. Distribution of segments.

	Class 1	*Class 2*	*Class 3*	*Class 4*
# segment	49	48	49	49
Total	195			

3.1 Pre-processing

In MG, we first collect 243 popular songs from Western, Chinese, and Japanese albums and choose a 25 second segment with strong emotion from each song as the training samples. Next, the subjects are asked to classify the songs by their opinions. If less than half of the subjects have the same emotion (class 1, 2, 3, or 4) for a song segment, the segment is considered emotion-weak and thus removed. 195 segments are retained, each labeled with a class voted by the subjects (decision by majority), see Table 1.

After converting these segments to 22,050 Hz, 16 bit, mono channel PCM WAV format, we use PsySound2 [6] to extract music features and choose 15 features as recommended in [7].

3.2 Fuzzy Classifiers

Compared to traditional classifiers which can only assign one class to the input sample, fuzzy classifiers assign a "fuzzy vector" that indicates the relative strength of each class. For example, $(0.1\ 0.0\ 0.8\ 0.1)^t$ represents a fuzzy vector with the strongest emotion strength for class 3, while $(0.1\ 0.4\ 0.4\ 0.1)^t$ shows an ambiguity between class 2 and 3. The ambiguity that fuzzy vectors carry is very important since the music emotion is intrinsically subjective.

The two fuzzy classifiers adopted in our work are described next.

3.2.1 *Fuzzy k-NN classifier (FKNN)*

The k-nearest neighbor (k-NN) classifier is commonly used in pattern recognition. An input sample is assigned to the class that is represented by the majority of the *k*-nearest neighbors. However, once an input sample is assigned to a class, there is no indication of its strength of membership in that class.

Fuzzy k-NN classifier [8], a combination of fuzzy logic and k-NN classifier, is designed to solve the above problem. It contains two steps: *fuzzy labeling* that computes the fuzzy vectors of the training samples (done in MG), and *fuzzy classification* that computes the fuzzy vectors of the input samples (done in EC).

In fuzzy classification, we assign a fuzzy membership μ_{uc} for an input sample x_u to each class c as a linear combination of the fuzzy vectors of *k*-nearest training samples:

$$\mu_{uc} = \frac{\sum_{i=1}^{k} w_i \mu_{ic}}{\sum_{i=1}^{k} w_i}, \tag{1}$$

where μ_{ic} is the fuzzy membership of a training sample x_i in class c, x_i is one of the *k*-nearest samples, and w_i is the weight inversely proportional to the distance d_{iu} between x_i and x_u:

$$w_i = d_{iu}^{-2}. \tag{2}$$

With Eq. (1), we get the $C \times 1$ fuzzy vector μ_u indicating music emotion strength ($C = 4$ in our system) of the input sample:

$$\mu_u = \{\mu_{u1}, ..., \mu_{uc}, ..., \mu_{uC}\}^t, \tag{3}$$

$$\sum_{c=1}^{C} \mu_{uc} = 1. \tag{4}$$

In fuzzy labeling we compute μ_i, the fuzzy vector of the training sample. Several methods have been developed ([8], [9]) and can be generalized as:

$$\mu_{ic} = \begin{cases} \beta + (n_c / K) * (1-\beta), & \text{if } c = v. \\ (n_c / K) * (1-\beta), & \text{otherwise.} \end{cases} \tag{5}$$

where v is the voted class of x_i, n_c is the number of samples that belong to class c in the *K*-nearest training samples of x_i, and β is a bias parameter indicating how v takes part in the labeling process ($\beta \in [0,1]$). When β=1, this is the crisp labeling that assigns each training sample full membership in the voted class v. When β=0, the memberships are assigned according to the *K*-nearest neighbors (*K* may be different from the *k* used in EC).

3.2.2 *Fuzzy Nearest-Mean classifier (FNM)*

For the fuzzy Nearest-Mean classifier, we need to calculate the mean of each feature of the classes in MG by:

$$\mu(c, f) = \frac{1}{N_c} \sum_{n=1}^{N_c} F_{c,f,n}, \tag{6}$$

where $\mu(c,f)$ is the mean of *f*th feature (f=1,2,…,15) in class c (c=1,2,3,4), $F_{c,f,n}$ is the value of the *f*th feature of the *n*th segment in class c, and N_c is the total number of segments in class c.

In EC, we compute the sum of the squared error (SSE) between the features of x and the mean of each class. The class whose

mean has the minimum SSE is the class to which x is assigned; that is,

$$d_{xc} = \sum_{f=1}^{15} (x_f - \mu(c,f))^2, \quad (7)$$

$$C(x) = \{y \mid \min(d_{xc}), c \in \{1,2,3,4\}\}, \quad (8)$$

where $C(x)$ denotes the predicted class of x, d_{xc} is the SSE between x and the mean of class c, and x_f is the value of the fth feature of x.

The fuzzy vector of the input sample is obtained by computing the inverse of the distance,

$$\mu_{uc} = \frac{d_{xc}^{-n}}{\sum_{c=1}^{C} d_{xi}^{-n}}, \quad (9)$$

where n, the degree of fuzziness [10], is empirically chosen.

3.3 Emotion Classification (EC)

After preprocessing an input sample, we compute its fuzzy vector using the model we generated in MG. The fuzzy vector is computed using Eq. (1) in FKNN and Eq. (8) in FNM. The maximum element in the vector is chosen as the final decision of classification. In the case of equal music emotion strength in two or more classes, the class of the nearest sample is chosen.

3.4 Feature Selection

To improve the classification accuracy, feature selection techniques can be applied to remove weak features. We adopt the stepwise backward selection method [11]. It begins with all 15 features and then greedily removes the worst feature sequentially until no more accuracy improvement can be obtained. The method we adopt to evaluate the classification accuracy is the 10 fold cross-validation technique. In this technique, 90% of the segments are randomly selected as training samples to generate the model. The remaining 10% are used for testing. The above process is repeated 50 times before the average accuracy is computed.

3.5 Music Emotion Variation Detection (MEVD)

Music emotion varies within a song. We develop a music emotion variation detection scheme (MEVD) to track the variation.

We segment the entire song every 10 second, with 1/3 overlapping between segments to increase correlation (see Figure 4), and classify the segments sequentially. Then we use the following equations to translate the resulting fuzzy vectors into valence and arousal:

$$\text{valence of } x_u = \mu_{1u} + \mu_{4u} - \mu_{2u} - \mu_{3u}, \quad (10)$$

$$\text{arousal of } x_u = \mu_{1u} + \mu_{2u} - \mu_{3u} - \mu_{4u}, \quad (11)$$

where μ_{yu}, y=1,2,3,4, are defined in Eq. (1).

After obtaining the valence and arousal values, we can plot them on the 2D plane to see the distribution of emotions. Another more convenient way to see the music emotion variation is to track the change of valence and arousal values separately. More discussion is given in Section 4.2.

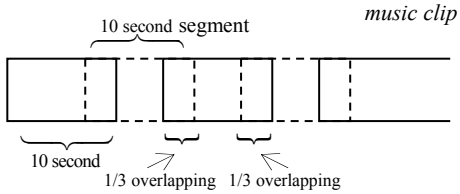

Figure 4. Segmentation of music clip in MEVD.

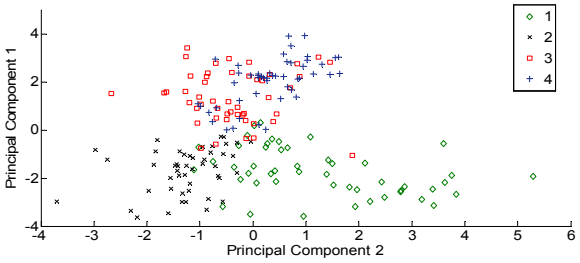

Figure 5. Scatter plot of first two principal components.

Table 2. Cross-validation results of FKNN using different β.

β	*1→1*	*2→2*	*3→3*	*4→4*	*average*
0.0	46%	92%	39%	66%	60.60%
0.25	51%	93%	51%	68%	65.86%
0.51	54%	93%	59%	66%	67.95%
0.75	56%	92%	61%	64%	68.22%
1.0	57%	89%	61%	63%	67.39%

Table 3. Cross-validation results of FNM.

	1→1	*2→2*	*3→3*	*4→4*	*average*
FNM	71%	88%	61%	66%	71.34%

Table 4. FKNN (β=0.75) results, after feature selection.

	Class 1	Class 2	Class 3	Class 4
Class 1	53.36%	41.63%	3.57%	1.42%
Class 2	2.08%	94.16%	3.43%	0.31%
Class 3	3.16%	9.89%	59.28%	25.91%
Class 4	0.1%	3.77%	21.12%	75%
Overall	70.88%			

Table 5. FNM results, after feature selection.

	Class 1	Class 2	Class 3	Class 4
Class 1	74.32%	18.24%	7.42%	0%
Class 2	4.83%	94.45%	0.7%	0%
Class 3	2.04%	6.12%	72.81%	19.02%
Class 4	0%	0%	28.24%	71.75%
Overall	78.33%			

4. EXPERIMENTAL RESULTS

4.1 Classification Accuracy

Table 2 shows the cross-validation results of FKNN using different β values in Eq. (5). The K used in MG and the k used in EC are both empirically decided and are set to 11. We can see that the highest accuracy 68.22% is achieved with β=0.75.

Results of FNM are shown in Table 3. We can see that this classifier has better accuracy (71.34%) than FKNN, especially in classifying class 1 samples. By examining the scatter plot of the first two principal components of the dataset (see Figure 5), we find that the distribution of class 1 samples is rather sparse, which causes FKNN to end up with a misclassification.

The results for FKNN and FNM after feature selection are shown in Tables 4 and 5. We can see great improvement in the overall accuracy. Moreover, we note that FNM still performs better (78.33% vs. 70.88%). Figures 6 and 7 show the distribution of emotions after translating fuzzy vectors to valence and arousal.

4.2 Plot of Music Emotion Variation

As an example, we segment Rene Liu's "Love You Very Much", and use FNM to track the variation of arousal and valence values in the song. The results are shown in Figure 8.

We consider the time points with rapid change of arousal or valence as the border between different emotions and divide the song into several sub-emotion units denoted as I, II, III, IV, V, VI, and VII, as shown in Figure 8. We observe the following relation: The sub-units II, IV, and VI with higher arousal correspond to the chorus while others represent the Intro, Verse1, Verse2, middle-eight, and Outro of the song. This relation indicates an interesting link between emotion detection and music structure analysis.

5. CONCLUSIONS

In this paper, we have described a fuzzy emotion classification system that can measure the relative strength of music emotion. This approach performs better than conventional deterministic approaches because it is able to incorporate the subjective nature of emotion perception in the classification. We have also presented a music emotion variation detection scheme to track the variation of emotions in a song.

6. ACKNOWLEDGMENTS

This work was supported in part by grants from Intel and the National Science Council of Taiwan under contracts NSC 94-2219-E-002-016 and NSC 94-2725-E-002-006-PAE.

7. REFERENCES

[1] Wang, M., Zhang, N., and Zhu, H., "User-Adaptive Music Emotion Recognition," IEEE, Int. Conf. Signal Processing, pp. 1352-1355, 2004.

[2] Liu, D., Lu, L., and Zhang, H. J., "Automatic Mood Detection from Acoustic Music Data," ISMIR, 2003.

[3] Yang, D., and Lee, W., "Disambiguating Music Emotion Using Software Agents," ISMIR, 2004.

[4] Seif El-Nasr, M., Yen, J., and Ioerger, T., "FLAME – Fuzzy logic adaptive mode of emotions," Autonomous Agents and Mult-iAgent Systems, 3, pp. 219□257, 2000.

[5] Thayer, R. E., "The Biopsychology of Mood and Arousal," Oxford University Press, 1989.

[6] PsySound, http://members.tripod.com/~densil/.

[7] Schubert, E., "Measurement and Time Series Analysis of Emotion in Music," Ph. D. Thesis, UNSW, 1999.

[8] Keller, J. M., Gray, M. R., and Givens, J. A., "A Fuzzy k-Nearest Neighbor Algorithm," IEEE Trans. Syst. Man. Cybern., vol. SMC-15(4), pp. 580-585, 1985.

[9] Han, J. H. et al, "A Fuzzy K-NN Algorithm Using Weights from the Variance of Membership Values," CVPR, 1999.

[10] Tran, D. et al, "Fuzzy Nearest Prototype Classifier Applied to Speaker Identification," ESIT, 1999.

[11] Sever, H., "Knowledge Structuring for Database Mining and Text Retrieval Using Past Optimal Queries," PhD Thesis, The University of Southwestern Louisiana, 1995.

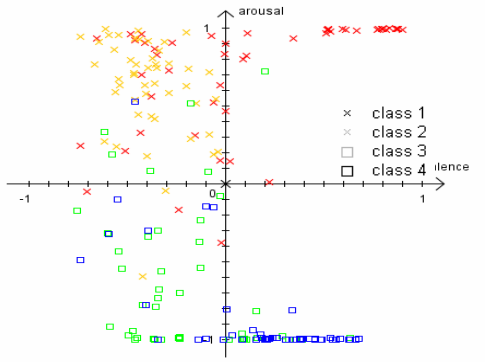

Figure 6. 2D plane for FKNN (after feature selection).

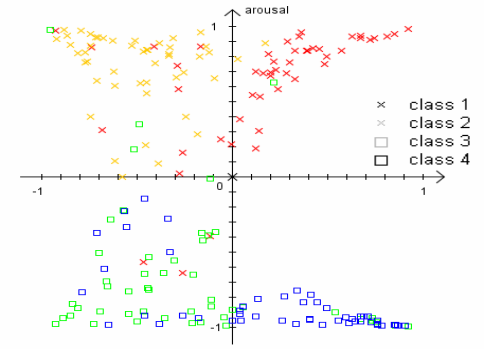

Figure 7. 2D plane for FNM (after feature selection).

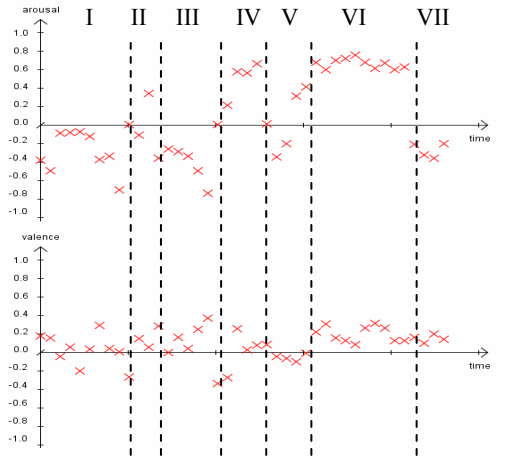

Figure 8. Arousal and valence variation of "Love You Very Much", using FNM.