Music Emotion Classification: Fuzzy K-nearest Neighbor Classifier

Myo Thin Zar Aung, Ei Mon Mon Swe
University of Computer Studies, Yangon
myo25myo@gmail.com , eimonmonswe@gmail.com

Abstract

Music expresses emotion. A number of audio extracted features have influence on the perceived emotional expression of music and 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. According to different approaches, we can provide the music emotion classification. In this paper, we consider a fuzzy k-nearest neighbor classifier to classify music emotion. For each music segment, this approach determines how likely the song segment belongs to an emotion class. This fuzzy classifier is 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.

1. Introduction

Music is important to our daily life. Humans, by nature, are emotionally affected by music. 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 important tasks for various applications, such as song selection in mobile devices, music recommendation systems, TV and radio programs and music therapy.

Music classification by perceived emotion is one of the most important research topics, for it is content-based and functionally more powerful. 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 fuzzy k-nearest neighbor classifier 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. Section 2 gives an overview of the classification system. Feature extraction is described in Section 3 and Section 4 expresses the theory of this system. Experimental results are given in Section 5 and conclusions and references are described in Section 6 and 7.

2. Music emotion classification

In this 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 1 and 2, while details are described in the following sub-sections.

2.1. Pre-processing

In preprocessing procedure, music clips are down sampled to 22050 Hz, 16 bit, and mono channel signals from files with CD-quality format. Preprocessing is performed in both the MG and the EC stages.

2.2. Model generator

First, we collect a large number of popular songs and choose a 20-second segment from strong emotion segment in each song as the initial music dataset. And then, volunteers are asked to classify the songs subjectively. If less than half of the subjects have the same emotional response (group 1, 2, 3 or 4) to a song segment, this segment is

discarded. The remaining segments are preprocessed and saved in WAV format. Then, we extract features using spectral flux, spectral centroid, intensity and MFCC. Figure 1 is the first part of this system. This describes the generalization mode for training data.

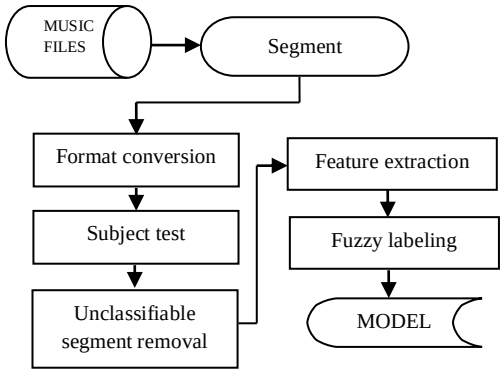

Figure 1. Block diagram of model generator

2.3. Emotion Classifier

We start by choosing a music segment with an unknown emotion and preprocessing it according to the procedure described in emotion classifier. Then, features are extracted by using spectral flux, spectral centroid, intensity and MFCC. And then, we classify this segment by using fuzzy k-nearest neighbor classifier. In Figure 2, the unknown data is classified by using classification method.

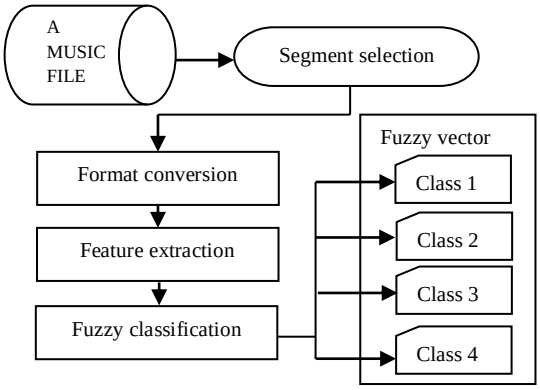

Figure 2. Block diagram of emotion classifier

3. Feature Extraction

The first step in any classification problem is to identify the features that are to be use for classification. Several different features have been suggested in music emotion classification.

Extracting features should be compressive (representing the very well), compact (requiring a small amount of storage), and effective (not requiring much computation for extracting).

In the case of audio signals, features may be related to the main dimension of music including melody, harmony, timbre, and spatial location. Before extracting, we segment the entire song every 20 second and divide frames with 32ms. These frames overlap according to1/3 overlapping between frames to increase and classify the segments sequentially. To extract the feature of these segments, we use spectral flux, spectral centriod, intensity and mel-frequency cepstral coefficient **(MFCC)**.

3.1. Mel-Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficients (MFCCs) are used for speech recognition and music modeling [6]. To derive MFCCs features, the signal was divided into frames and the amplitude spectrum was calculated for each frame. Next, its logarithm was taken and converted to Mel scale. Finally, the discrete cosine transform was implemented. We selected the first 13MFCCs.

3.2. Spectral Flux

Spectral flux is a measure of how quickly the power spectrum of a signal is changing and calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. The spectral flux can be used to determine the timbre of an audio signal or in onset detection among other things. Spectral Flux is defined as the squared difference normalize between the normalize magnitudes of successive spectral distribution.

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \qquad (1)$$

Where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t, and previous frame $t-1$. The spectral flux is a measure of the amount of local spectral change.

3.3. Spectral Centroid

The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the “center of mass” of the spectrum is. It has a robust connection with the impression of “brightness” of a sound. It is calculated as the weighted mean of the frequencies present in the signal, with their magnitude as the weights:

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \quad (2)$$

where $x(n)$ represents the magnitude of bin number n, and $f(n)$ represents the center frequency of that bin.

3.4. Intensity

Since human's hearing organs are simulated by the loudness of signals, human auditory system is greatly affected by the intensity of music. The intensity features is an essential feature in mood detection. In this system, the intensity feature is composed of the spectrum sum of the signal and the spectrum distribution in each sub-band.

$$I(n) = \sum_{k=0}^{\omega/2} A(n,k) \quad (3)$$

$$D_i(n) = \frac{1}{I(n)} \sum_{k=L_i}^{H_i} A(n,k) \quad (4)$$

$I(n)$ is the intensity of i^{th}-frame and $D_i(n)$ is the intensity ration of the i^{th} sub-band. L_i and H_i are the lower and upper bound of the i^{th} sub-band, respectively, and A (n,k) is the absolute value of the k^{th} FFT coefficients of the nth frame.

Another set of 3 features that relate to timbre texture were extracted from the Short-Term Fourier Transform (FFT): spectral flux, spectral centroid, intensity. For each of the 16 aforementioned features (13 MFCCs, 3 FFT) we calculated the mean and standard deviation of overall frames.

4. 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 fuzzy k-nearest neighbor classifiers adopted in our work are described next.

4.1 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 [5] is a combination of fuzzy logic and k-NN classifier that 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:

$$\mu_{uc} = \frac{\sum_{i=1}^{k} w_i \mu_{ic}}{\sum_{i=1}^{k} w_i} \quad (5)$$

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:

$$w_i = d_{iu}^{-2} \quad (6)$$

With Eq. (6) we get the C×1 fuzzy vector μ_u indicating music emotion strength (C = 4 in our system) of the input sample:

$$\mu_u = \{\mu_{u1}, ..., \mu_{uc}, ..., \mu_{uC}\}^t \quad (7)$$

$$\sum_{c=1}^{C} \mu_{uc} = 1 \quad (8)$$

In fuzzy labeling we compute μ_i, the fuzzy vector of the training sample. Several methods have been developed ([1], [4], [5]) and can be generalized as:

$$\mu_{uc} = \begin{cases} \beta + (n_c/K) * (1-\beta) \\ (n_c/K) * (1-\beta) \end{cases} \quad (9)$$

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).

5. Experimental and Results

We collected 200 popular songs from various albums and choose a 20-second segment from each song. Subjects were asked to label the emotion of the segments. After removing the unclassifiable ones, as explained in Section (EC), we obtained segments with their group of emotion annotated by the subjects. In Table 1, the accuracy of the training data is described by subjectively. Table 2 describes the accuracy of the unknown data by using the Fuzzy k-nearest neighbor classifier.

Table 1. Training results with FKNN

	Happy	Fear	Sad	Angry
Happy	99.98%	0%	0%	0.2%
Fear	0%	98.99%	0%	1.01%
Sad	0%	0%	99.97%	0.3%
Angry	0%	0.1%	0%	99.99%

Table 2. No Training results with FKNN

	Happy	Fear	Sad	Angry
Happy	89.23%	3.47%	2.00%	5.30%
Fear	9.24%	84.27%	2.35%	4.14%
Sad	4.31%	3.04%	85.34%	7.31%
Angry	9.77%	1.92%	3.09%	83.99%

6. Conclusions

In this paper, we have proposed an emotion classification system for music. The system is described by a fuzzy emotion classification system that can measure the relative strength of music emotion. The 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 emotion in a song.

7. References

[1] Yi-Hsuan Yang, Chia-Chu Liu, and Homer H. Chen ,"Music Emotion Classification: A Fuzzy Approach," Graduate Institute of Communication Engineering, National Taiwan University. 1 Roosevelt Rd. Sec4. Taipel, 10617, Taiwan R.O.C. 886-2-3366-3549.

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

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

[4] 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.

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

[6] B. Logan. Mel frequency cepstral coefficients for music modeling. In Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR 2000), Plymouth, Massachusetts, 2000.

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