

# EXPLOITING GENRE FOR MUSIC EMOTION CLASSIFICATION

Yu-Ching Lin<sup>1</sup>, Yi-Hsuan Yang<sup>1</sup>, Homer H. Chen<sup>1</sup>, I-Bin Liao<sup>2</sup>, Yeh-Chin Ho<sup>2</sup>

<sup>1</sup>National Taiwan University

<sup>2</sup>Telecommunication Laboratories, Chunghwa Telecom

## ABSTRACT

Genre and emotion have been applied to content-based music retrieval and organization; however, the intrinsic correlation between them has not been explored. In this paper we present a statistical association analysis to examine such intrinsic correlation and propose a two-layer scheme that exploits the correlation for emotion classification. Significant improvement of classification accuracy over the traditional single-layer scheme is obtained.

**Index Terms**—Music genre classification, association analysis, music emotion classification.

## 1. INTRODUCTION

Due to the explosive growth of music recordings in recent years, content-based music organization has emerged as an attractive means for accessing large-scale music collection. Genre and emotion classifications are key elements of content-based music organization.

Genre, by which a song is classified into classical, jazz, hip-hop, etc., has been used to describe the intrinsic form of music for a long time. A recent user study [1] shows that, besides keywords such as artist, song title, and lyric, genre is the most popular cues for searching music, as is evident by the fact that most music websites provide the genre metadata of music recordings. Great progress has been made in automatic genre classification, see [2] for a comprehensive review.

Emotion, as one of the preeminent functions of music [3], is also an important means for music classification. An emotion-based music retrieval system [4] provides users the functionality for retrieving music according to emotion. Comparing to genre classification, emotion classification is more challenging because of the subtlety of emotion and the difficulty of collecting subjective annotation of emotion.

Genre and emotion provide complementary descriptions of music content and often correlate with each other. For example, a rock song is often aggressive, and a rhythm and blues (R&B) song is more likely to be sentimental. Despite the existence of salient correlation between them, genre classification and emotion classification have been studied separately without considering the inter-relation.

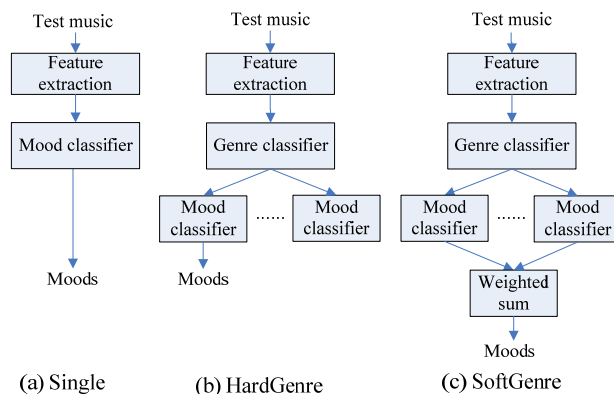


Figure 1. (a) Traditional single-layer mood classification scheme, (b) the proposed two-layer scheme, and (c) the soft version of the two-layer scheme.

In this paper, based on the statistical evidence that genre and emotion are correlated, we propose a two-layer scheme to exploit the correlation between them. As genre metadata are more stable and easier to collect, we use genre metadata to aid emotion classification. The genre of a song is predicted in the first layer, and then we can apply a genre-specific emotion classification model in the second layer (Figures 1b and 1c). We show by experiments that the two-layer scheme significantly outperforms the traditional single-layer scheme.

The remainder of the paper is organized as follows. Section 2 dwells on the adopted dataset. Section 3 presents the statistical association analysis of genre and emotion. The proposed two-layer emotion classification scheme is described in Section 4 and the evaluation results in Section 5. Section 6 concludes the paper.

## 2. DATA COLLECTION

In order to analyze the correlation between genre and emotion, we build a dataset composed of genre and emotion metadata from the famous music review website named All Music Guide (AMG, <http://www.allmusic.com>). This dataset, called whole set, contains the annotation of 6490 albums. Each of them is annotated with a single genre and more than one emotion. See Table I for a summary.

The audio files of a subset of the whole set are acquired from personal collections for the evaluation of music

Table I  
Dataset description

Dataset	# albums	# genres	# emotions
whole set	6490	12	184
cluster set	300	6	12

Table II  
The resulting 12 emotion clusters and some of the associated emotion labels in each cluster

Cluster	Associated emotions			# songs
1	Rustic	Self-Conscious	Sparse	161
2	Bright	Reverent'	Sparkling	269
3	Acerbic	Bitter	Ironic	330
4	Aggressive	Fiery	Manic	701
5	Carefree	Cheerful	Happy	729
6	Bleak	Brooding	Ominous	369
7	Delicate	Gentle	Intimate	704
8	Atmospheric	Ethereal	Hypnotic	326
9	Angry	Harsh	Hostile	463
10	Humorous	Quirky	Silly	440
11	Ambitious	Dramatic	Enigmatic	407
12	Hedonistic	Outrageous	Reckless	301

emotion classification. The dataset, called cluster set, is composed of 1535 songs collected from 300 albums, with the six equally distributed genres: blues, country, jazz, R&B, rap, and rock. We propagate the album annotation to the song annotation, so each song of the album has the same annotation of the album.

Since the 184 emotion labels contain several synonyms, a set of more representative emotion clusters can be acquired by grouping them. Because the emotion labels are obtained from album annotation, we conduct the grouping process in the annotation of album. We measure the similarity  $S(e_1, e_2)$  of two emotion labels,  $e_1$  and  $e_2$ , by the frequency they co-occur as follows

$$S(e_1, e_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}, \quad (1)$$

where  $E_1$  is the set of albums annotated with  $e_1$ ,  $E_2$  is the set of albums annotated with  $e_2$ , and  $|E|$  denotes the cardinality of a set. Based on the similarity defined above, we construct an affinity matrix of emotion labels and employ spectral clustering [5] to group them. The number of clusters is determined by the Eigengap technique proposed in [5]. After removing those emotion clusters with that half or more of the albums are annotated in our dataset, we obtain the 12 emotion clusters shown in Table II. It can be found that the emotions in the same cluster are often synonyms of the same feeling, while emotions in different clusters share little similarity.

### 3. STATISTICAL ASSOCIATION ANALYSIS

To examine the overall correlation between emotion and genre, we employ chi-square test — the widely used

Table III  
Association test result

	$\chi^2$	$\chi_0^2(0.05)$	Cramer's V
+ rock	9393.6	233.9	0.227
- rock	6057.6	177.4	0.353
cluster Set	1550.8	73.3	0.225

significance test of association [7]. Based on a null hypothesis, the chi-square statistic  $\chi^2$  is computed to indicate how far the observed distribution departs from the expected distribution of null hypothesis. If  $\chi^2$  is larger than the critical value  $\chi_0^2$  of a significance level, the null hypothesis is rejected. We test the null hypothesis that genre and emotion are independent at significance level 0.05. Because about half of the whole set are rock, we also conduct the test without rock albums to eliminate the bias. Table III shows the result. We can see that all chi-square tests reject the null hypothesis, suggesting that genre and emotion are correlated.

To estimate the associative strength, we also compute Cramer's V [7]. The value of it ranges from zero to one, and a larger value indicates a stronger association. The result is shown in Table III. Considering that not all genre-emotion pairs are correlated, for example the passion of love is too general to be associated with a specific genre, we may conclude that the reported Cramer's V are moderately high enough, and a certain degree of correlation does exist.

Note that this result contradicts the conclusion in [6], that genre and emotion are independent. The authors of [6] examine the correlation of genre-emotion pairs and find the significantly correlated emotions are shared by all genres. Ignoring that the most co-occurred emotions of each genre are quite different, they conclude that one can model genre and emotion separately. Our experiment result shows the integral correlation between genre and emotion does exist.

### 4. SYSTEM DESCRIPTION

Since genre and emotion are correlated, it is interesting to see whether the incorporation of genre metadata improves the accuracy of emotion classification. We thus propose a two-layer scheme that the emotion classifier specifically for a single genre is used. Designing such genre-specific classifiers is based on two reasons. First, since emotion and genre are correlated, we may set up different emotion priors for each genre-specific classifier. For example, a rap song is less likely to be relaxing than a jazz song. Second, as a happy song of rock music and a happy song of jazz music may sound substantially different, predicting emotion may become easier if each genre-specific classifier only needs to focus on a single genre of music.

As shown in Figure 1(a), a typical single-layer emotion classification system is composed of two parts: feature extraction and model training. An emotion classifier learns the relationships between audio features and emotion labels during the training process. According to AMG, a music recording can be associated with more than one emotion.

We thereby treat emotion classification as a multi-label classification problem [10]. We denote this single-layer scheme as “Single” hereafter.

Figures 1(b) shows the proposed two-layer emotion classification scheme, denoted as “HardGenre.” In the training phase, a genre classifier is trained in the first layer. The training dataset is then divided according to the genres, and the genre-specific multi-label emotion classifiers are trained separately using the songs of the corresponding genre in the second layer. In the testing phase, the system classifies genre in the first layer and classifies emotion with the emotion classifier of the predicted genre in the second layer. For example, if a song is predicted as rock in the first layer, the emotion classifier trained with rock songs will be employed in the second layer.

The intrinsic form of some songs in our dataset is mixed with the musical elements of two genres. For example, an R&B singing may be put in a rap song as one of the verse<sup>1</sup>. Based on this observation, we further propose a soft version of HardGenre in Figures 1(c). We denote this scheme as “SoftGenre,” which utilizes soft-decision genre classification in the first layer. Rather than predicting genre in a deterministic way, we compute the likelihood  $P(g|x)$  of a song  $x$  being one of the genre types  $g$ , and then aggregate the predictions of each genre-specific emotion classifier by weighted combination of  $P(g|x)$ . If the weighted sum of an emotion label is larger than 0.5, we predict that the song has the emotion label. Clearly, HardGenre is the special case of SoftGenre with total weight on a single genre type.

In our implementation, each song is converted to a uniform format (22050Hz, mono channel PCM WAV) and normalized to the same volume label to ensure fair comparison. Feature extraction is done by a well-known music processing system, Marsyas [8], which generates a total of 436 audio features: 68 timbral textual features, 48 pitch content features, 8 rhythmic content features, 120 linear-prediction based features, and 192 MPEG-7 features. With these spectral attributes, temporal traits and musical characteristics, the features can thoroughly represent the audio content. LIBSVM [9] is employed for genre classification, and a multi-label variant of LIBSVM, which is extended from the one-against-all multi-class method, is employed for emotion classification.

## 5. EXPERIMENT

We adopt ten-fold cross validation for evaluation. The whole dataset<sup>2</sup> is randomly divided into 10 parts, 9 of them for training and the rest for testing, and the partition is iteratively performed until each fold is held out once. The above process is repeated 20 times and the result is obtained by averaging the evaluation measures.

<sup>1</sup> Verse is one of the sectional forms of music

<sup>2</sup> Data is available: <http://mpac.ee.ntu.edu.tw/~vagante/genreEmo>

### 5.1. Evaluation Measures

The standard evaluation measurements for multi-label classification problems are the macro-average F-measure (Macro-avg) and micro-average F-measure (Micro-avg) [10] respectively as,

$$\frac{1}{d} \sum_{j=1}^d \frac{2 \sum_{i=1}^{k_j} \widehat{y}_{ij} y_{ij}}{\sum_{i=1}^{k_j} \widehat{y}_{ij} + \sum_{i=1}^{k_j} y_{ij}}, \quad (2)$$

$$\frac{2 \sum_{j=1}^d \sum_{i=1}^{k_j} \widehat{y}_{ij} y_{ij}}{\sum_{j=1}^d \sum_{i=1}^{k_j} \widehat{y}_{ij} + \sum_{j=1}^d \sum_{i=1}^{k_j} y_{ij}}, \quad (3)$$

where  $d$  denotes the number of class labels,  $k_j$  is the number of songs of label  $j$ ,  $\widehat{y}_{ij} \in \{0,1\}$  denotes the prediction of song  $i$  in label  $j$ , and  $y_{ij}$  is the ground truth value. Macro-avg is the equally-weighted mean of the F-measure of each label, while Micro-avg accounts for the prediction of all samples and calculates F-measure across all labels. Therefore, Macro-avg respects each label equally, while Micro-avg respects the overall prediction result of all labels.

### 5.2. Two-layer emotion classification

In the experiments, the following four emotion classification schemes are compared. Single is the traditional single-layer scheme, while HardGenre and SoftGenre are the proposed two-layer schemes. To estimate the maximal improvement the two-layer scheme can make, we put an oracle genre classifier that predicts the genre of a song according to the ground truth in the first layer of the HardGenre scheme. We refer to this scheme as “TrueGenre.” Though the oracle genre classifier is not perfect in that we have approximated the song-level ground truth using album-level annotations, we regard the performance of the TrueGenre scheme as an upper bound which defines both the viability of the two-layer scheme and the F-measures that is reasonable for our system to achieve.

Figure 2 illustrates the F-measures of the 12 emotion clusters in the four emotion classification schemes. As TrueGenre brings about improvement for all emotion clusters, we can see that the emotion classification can be much improved if genre classification has been done perfectly. When the genre of a song is predicted by the system and some incorrect predictions are made (the genre classification accuracy is about 58.98%), HardGenre still offers improvement for most emotion clusters, especially for emotion clusters whose F-measures are low under the Single scheme. On the other hand, we can also see that the errors made in genre classification have negative impacts on the prediction of emotion clusters, such as clusters 4, 5, and 7. In our experiment result, we find that the mis-classified songs often have two comparable likelihoods of the different genres, which is consistent to the observation mentioned in Section 4 that the intrinsic form of music may be a mixed

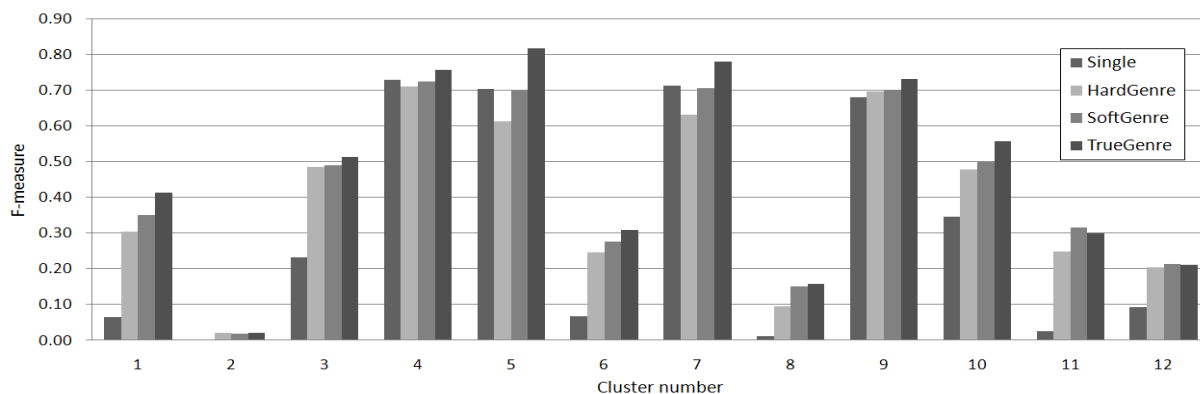


Figure 2. F-measures of the four emotion classification schemes based on the 12 emotion clusters listed in Table II.

Table IV  
Experiment results of the four emotion classification schemes

System	Macro-avg	Micro-avg
Single	0.31	0.52
HardGenre	0.40*	0.51
SoftGenre	0.43*	0.56*
TrueGenre	0.46*	0.60*

\*the scheme has significant improvement over the Single scheme at significant level 0.05

one of two genres. Using deterministic genre classification is not suitable in this case. This problem can be mitigated by the use of soft decision in genre classification. From Figure 2, we can see that SoftGenre compensates for the errors made by the deterministic genre classification and has improvement in all emotion clusters.

Table IV shows the Macro-avg and Micro-avg of the four emotion classification schemes. We can see that the proposed two-layer schemes have significant improvement over the traditional scheme, Single. With the use of SoftGenre, the Macro-avg and Micro-avg can be improved to 0.43 and 0.56, which are very close to that of TrueGenre. A closer look at Table IV reveals the HardGenre scheme has improvement in the Macro-avg but not in the Micro-avg. In Table II, we can see that clusters 4, 5, and 7 with the smaller F-measures contain the largest number of songs among the emotion clusters. Even if the classification of the rest of the clusters is improved, small decrease of the F-measures of the three clusters still greatly affects the Micro-avg. The SoftGenre compensates for this drawback, and we can see that both Macro-avg and Micro-avg are improved to a similar degree to the ones of TrueGenre.

## 6. CONCLUSION

In this paper, the correlation between music genre and emotion is carefully examined and exploited to improve the performance of an emotion classification system. With the proposed two-layer scheme, significant improvement of classification accuracy over the conventional single-layer

scheme is achieved. Despite of errors made in the first layer of genre classification, we have demonstrated that the utilization of soft genre decisions makes the two layer scheme achieve a similar accuracy as if the true genres are known in prior.

## 7. ACKNOWLEDGMENTS

This work is supported by a grant from the National Science Council of Taiwan under NSC 97-2221-E-002-111-MY3.

## 8. REFERENCES

- [1] H. Lee and J. S. Downie, "Survey of music information needs, uses, and seeking behaviours: preliminary findings," Proc. ISMIR 2004, pp. 441–446.
- [2] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," IEEE Signal Processing Magazine, vol. 23, no. 2, 2006, pp. 133–141.
- [3] D. Huron, "Perceptual and cognitive applications in music information retrieval," Proc. ISMIR, 2000.
- [4] Y.-H. Yang et al, "A regression approach to music emotion recognition," IEEE Trans. Audio, Speech and Language Processing, vol. 16, no. 2, 2008, pp. 448–457.
- [5] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," Proc. NIPS, 2001, pp. 849–856.
- [6] X. Hu and J. S. Downie, "Exploring mood metadata: relationships with genre, artist and usage metadata," Proc. ISMIR, 2007, pp. 67–72.
- [7] A. Agresti, Categorical data analysis, John Wiley & Sons Publications, 2002.
- [8] G. Tzanetakis and P. Cook, "Marsyas: a framework for audio analysis," Organized Sound, vol. 4, no. 3, 2000, pp. 169–175.
- [9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," 2001.
- [10] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," International Journal of Data Warehousing and Mining, vol. 3, no. 3, 2007, pp. 1–13.