Toward Multi-Modal Music Emotion Classification

Yi-Hsuan Yang[1], Yu-Ching Lin[1], Heng-Tze Cheng[1], I-Bin Liao[2], Yeh-Chin Ho[2], and Homer H. Chen[1]

[1] National Taiwan University
[2] Telecommunication Laboratories, Chunghwa Telecom
{affige, vagante, mikejdionline}@gmail.com, {snet, ycho}@cht.com.tw, homer@cc.ee.ntu.edu.tw

Abstract. The performance of categorical music emotion classification that divides emotion into classes and uses audio features alone for emotion classification has reached a limit due to the presence of a semantic gap between the object feature level and the human cognitive level of emotion perception. Motivated by the fact that lyrics carry rich semantic information of a song, we propose a multi-modal approach to help improve categorical music emotion classification. By exploiting both the audio features and the lyrics of a song, the proposed approach improves the 4-class emotion classification accuracy from 46.6% to 57.1%. The results also show that the incorporation of lyrics significantly enhances the classification accuracy of valence.

Key words: Music emotion recognition, multi-modal fusion, lyrics, natural language processing, probabilistic latent semantic analysis

1 Introduction

Due to the explosive growth of music recordings, effective means for music retrieval and management is needed in the digital content era [1]. Classification and retrieval of music by emotion [2]-[6] has recently received increasing attention, because it is content-centric and functionally powerful.

A popular approach called music emotion classification (MEC) divides the emotions into classes and applies machine learning on audio features, such as Mel frequency cepstral coefficient (MFCC), to recognize the emotion embedded in the music signal. However, due to the semantic gap, the progress of such mono-modal approach has been stagnant. While mid-level audio features such as chord [4] or rhythmic patterns [7] have more semantic information, they cannot be reliably extracted with the state-of-the-art technology yet.

Complementary to music signal, lyrics are semantically rich and expressive and have profound impact on human perception of music [8]. It is often easy for us to tell from the lyrics whether a song expresses love, sadness, happiness, or something else. Incorporating lyrics in the analysis of music emotion is feasible because most popular songs sold in the market come with lyrics and because

most lyrics are composed in accordance with music signal [9]. One can also analyze lyrics to generate textual feature descriptions of music. Although how to use lyrics and melodies to convey emotion has been studied (see, for example, [8]), little has been reported in the literature that uses lyrics for automatic music emotion classification.

In this paper, a multi-modal approach that uses features extracted from both music signal and lyrics is proposed for music emotion classification. We adopt statistical natural language processing techniques such as bag-of-words [14] and probabilistic latent semantic analysis (PLSA) [16] to extract textual features from lyrics of any languages. We also develop a number of multi-modal methods for fusing the extracted textual features with audio features. The proposed approach is evaluated on a moderately large-scale database. The results show that the incorporation of lyrics for music emotion classification greatly improves the classification accuracy. In particular, the *late fusion by subtask merging* approach significantly outperforms the purely audio-based approach and contributes to 21% relative improvement in classification accuracy.

The remainder of the paper is organized as follows. Section 2 describes the details of the proposed multi-modal approach. Section 3 provides the result of a performance study. Section 4 reviews related work on lyrics analysis, and Section 5 concludes the paper.

2 Proposed Approach

The system diagram of the proposed multi-modal MEC approach in the training phase is shown in Fig. 1, where audio features extracted from the waveform and textual features extracted from the lyrics are used to represent a song. Two emotion classification models are trained using different modalities of the feature set and integrated by multi-modal fusion methods. The classification models are then utilized to classify the emotion of any (test) songs. Below we describe each system component in detail.

2.1 Audio Feature Extraction

To ensure fair comparison, the music samples are converted to a uniform format (22,050 Hz, 16 bits, and mono channel PCM WAV) and normalized to the same volume level. Besides, since the emotion within a music selection can vary over time [3], we apply feature extraction to the middle 30-second segment of each song and consider the classification result of the segment as the emotion of the entire song.

We use two free computer programs Marsyas [11] and PsySound [12] with default parameter values to extract a number of low-level audio features. The extracted features, which are listed in Table 1 and described in detail below, have been commonly used for MEC in pervious works [2]-[4].

Marsyas is a free software framework for rapid development and evaluation of computer audition applications. We use it to extract the well-known Mel-frequency cepstral coefficient (MFCC), a set of perceptually motivated pitch

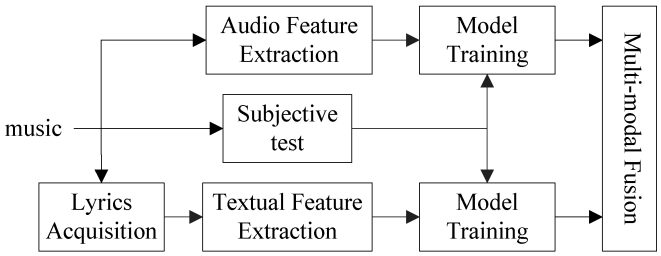

Fig. 1. System diagram of the training phase of the multi-modal music emotion recognition approach.

Table 1. Adopted feature extraction algorithms.

Modality	Method	# of features	Features
Audio	Marsyas [11]	52	Mel-frequency cepstral coefficient
	PsySound [12]	54	spectral centroid, spectral moment, spectral roughness
Textual	uni-gram [14]	4000	bag-of-words
	PLSA [16]	100	latent vectors
	bi-gram [15]	4000	bag-of-words

scale commonly used in audio signal processing [11]. The MFCCs are computed in three stages to take the temporal information of music into account. First, 13-dimension MFCCs are extracted for each short frame of 23 ms. Second, the mean and standard deviation of MFCCs are computed over a sliding texture window of 1 second. Finally, the feature vectors are collapsed into a single vector by taking again the mean and standard deviation of MFCCs over the entire 30-second segment. This gives rise to 52 MFCCs for each song.

As the name indicates, PsySound aims to model parameters of auditory sensation based on some psychoacoustic models [12]. We use it to generate 50 timbral texture features including spectral centroid and spectral moment to describe the shape properties of the FFT spectrum and cepstrum. 4 spectral roughness features are also extracted to measure dissonance, the perception of short irregularities in a sound. Any note in music that does not fall within the prevailing harmony is considered dissonant. Because of its psychoacoustical foundation, the PsySound features have been found fairly related to emotion perception [2].

2.2 Textual Feature Extraction

Lyrics are normally available on the web and downloadable with a simple crawler [13], [10]. The acquired lyrics are preprocessed with traditional information retrieval operations such as stopword removal, stemming, and tokenization [14]. As shown in Table 1, three algorithms are adopted to generate textual features.

Uni-gram A standard textual feature representation is to count the occurrence of uni-gram terms (words) in each document, and construct the bag-of-words model [14], which represents a document as a vector of terms weighted by a tfidf[3] function defined as:

$$tfidf(t_i, d_j) = \#(t_i, d_j) \log \frac{|D|}{\#D(t_i)}, \tag{1}$$

where $\#(t_i, d_j)$ denotes the frequency of term t_i occurs in document d_j, $\#D(t_i)$ the number of documents in which t_i occurs, and $|D|$ the size of the corpus. The intuition is that the importance of a term increases proportionally to its occurrence in a document, but is offset by its occurrence in the entire corpus to filter out common terms. In this way, a good combination between popularity (idf) and specificity (tf) is obtained [14]. Despite its simplicity, the unigram based bag-of-words model has shown superior performance in many information retrieval problems. We compute the tfidf for each term and select the M most frequent terms as our features (M is empirically set to 4000 in this work by a validation set).

Lyrics, however, are distinct from regular documents (e.g., news articles). First, lyrics are usually brief, and are often built from a very small vocabulary. With the *short text problem*, often there are words in a test set that do not appear in the training set [15]. Second, lyrics are often composed in a poem-like fashion. The rich *metaphors* can make word sense disambiguation [14] even more difficult. Third, lyrics are in nature recurrent because of the stanzas (group of lines arranged together in metrical length). This *recurrent structure* is not modeled by bag-of-words since word orders have been disregarded. Finally, unlike normal articles whose topics (e.g., politics, sports, and weather) are rather diverse, lyrics are *almost about love and sentiment*. This makes common stopword lists not applicable. In addition, *negation terms* such as "no" and "not" can play a more important role in lyric analysis. For example, whether there is a "not" precedent to "regret" clearly makes a difference in semantic meaning.

To address these issues, we also explore the utilization of the following two statistical natural language processing techniques to extract textual features.

PLSA PLSA has been used [15] to resolve the short text problem because it is able to discover polysems (i.e., a word that has multiple senses and multiple types of usage in different contexts) and synonymys (i.e., different words that share a similar meaning) [16]. It has been shown that PLSA increases the overlapping of semantic terms, which in turn improves the classification accuracy of short documents [15].

In PLSA [16], the joint probability between document d and term t is modeled through a latent variable z, which can be loosely thought of as a hidden class or topic. A PLSA model is parameterized by $P(t|z)$ and $P(z|d)$, which is estimated using the iterative Expectation Maximization (EM) algorithm to fit the training

[3] "tfidf" stands for term-frequency inverse-document-frequency.

corpus. Under the conditional independence assumption, the joint probability of t and d can be defined as

$$P(d,t) = P(d)P(t|d) = P(d)\sum_{z\in Z} P(t|z)P(z|d), \tag{2}$$

where Z denotes the number of latent topics. After training, $P(t|z)$ are used to estimate $P(z|q)$ for new (test) document q through a folding-in process [16]. Each component of $P(z|q)$ represents the likelihood that the document q is related to a pre-learnt latent topic z. Similarity in this *latent vector space* can be regarded as the semantic similarity between two documents. Therefore, PLSA can be viewed as a dimension reduction method ($Z \ll M$) that converts the bag-of-words model into a semantically compact form in a generative process. Z is set to 100 in this work.

Bi-gram N-gram are sequences of N consecutive words [14]. An N-gram of size 1 is a *uni-gram* (single word), size 2 is a *bi-gram* (word pairs). N-gram models are widely used to model the dependency of words. Since negation terms often reverse the meaning of the words next to them, it seems reasonable to incorporate word pairs to the bag-of-words model to take the effect of negation terms into account. To this end, we select the M most frequent uni-gram and bi-gram in the bag-of-words model and obtain a new feature representation. To avoid the situation that the single words of a word pair is doubly counted in uni-gram and bi-gram, we select frequent bi-gram first and uni-gram next.

2.3 Model Training

We adopt Thayer's arousal-valence emotion plane [17] as our taxonomy and define four emotion classes happy, angry, sad, and relaxing, according to the four quadrants of the emotion plane[4], as shown in Fig. 2. As arousal (how exciting/calming) and valence (how positive/negative) are the two basic emotion dimensions found to be most important and universal [18], we can also view the four-class emotion classification problem as the classification of high/low arousal and positive/negative valence. This view will be used in mutli-modal fusion and system evaluation.

Support vector machine (SVM) [19] is adopted to train classifiers for its superb performance shown in previous MEC works [2], [4]. SVM nonlinearly maps input feature vectors to a higher dimensional feature space by the kernel trick [19], and yields prediction functions that are expanded on a subset of support vectors. Our implementation of SVM is based on the library LIBSVM [20] with default parameter settings.

[4] This is a common taxonomy adopted in previous MEC works [3]-[5]. Though we have proposed to view the emotion plane from a continuous perspective [2], we adopt this categorical taxonomy here for quick assessing the impact of lyrics.

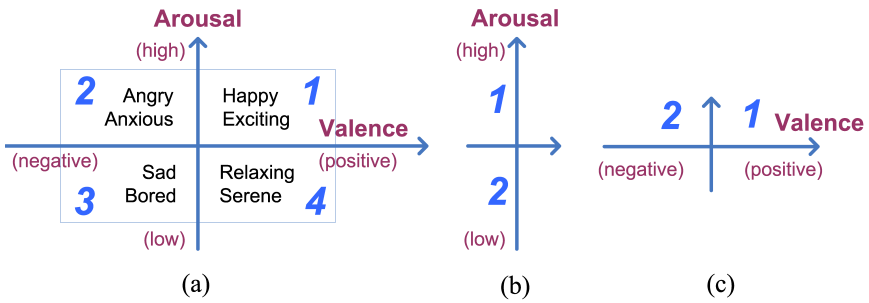

Fig. 2. (a) Thayer's arousal-valence emotion plane. We define four emotion classes according to the four quadrants of the emotion plane. We can also subdivide the four-class emotion classification to binary (b) arousal classification and (c) valence classification.

2.4 Multi-Modal Fusion

We develop and evaluate the following methods for fusing audio and text cues. To enhance readability, we denote the classification model trained by audio and textual features as M_A and M_T, respectively.

- Audio-Only (**AO**): Use audio features only and apply M_A to classify emotion. This serves as a baseline because most existing MEC work adopts it.
- Text-Only (**TO**): Use textual features only and apply M_T to classify emotion. TO is used to assess the importance of the text modality.
- Early Fusion by Feature Concatenation (**EFFC**): Concatenate the audio and textual features to a single feature vector before learning and train a single classification model. Early fusion yields a truly multi-modal feature space, but it can suffer from the difficulty to combine modalities into a common representation [21].
- Late Fusion by Linear Combination (**LFLC**$_\alpha$): Train M_A and M_T separately and combine their predictions afterwards in a linear fashion. We use SVM to produce probability estimation [20] of the class membership in each class, linearly combine the probability estimates of two SVM models, and make final decision by taking the class with highest fused value. The parameter $\alpha \in [0, 1]$ denotes the weight of the two modalities ($\alpha > 0.5$ gives more weights to text). For example, if the probability estimates of emotion for a song by M_A and M_T are $\{0, 0.1, 0.5, 0.4\}^T$ and $\{0, 0.1, 0.7, 0.2\}^T$, then the linear combination with $\alpha = 0.5$ would be $\{0, 0.1, 0.6, 0.3\}^T$, and the final decision would be class 3. Late fusion focuses on the individual strength of modalities, yet it introduces additional training efforts and the potential loss of correlation between modalities [21].
- Late Fusion by Subtask Merging (**LFSM**): Use M_A and M_T to classify arousal and valence separately and then merge the result. For example, a negative arousal (predicted by M_A) and negative valence (predicted by M_T) would be merged to class 3. We make the two modalities focus on different emotion classification subtasks because empirical test reveals audio and

text clues are complementary and useful for different subtasks. In addition, training models for arousal and valence separately has been shown adequate in [2].

3 Experimental Result

The music database is made up of 1240 Chinese pop songs, whose emotions are labeled through a subjective test. The corresponding lyrics are downloaded from the Internet by a web crawler. We build our own Chinese stopword list for stopword removal and adopt the free library LingPipe [22] for Chinese word segmentation (tokenization). Classification accuracy is evaluated by randomly selecting 760 songs as training data and 160 songs as test data, with the number of songs of each emotion class uniform. Because of this randomization, 1000 iterations are run to compute the average classification accuracy. Note the genre of our database is pop music rather than the western classical music as adopted in [3] since MEC is to facilitate music retrieval and management and since it is the pop music that dominates the everyday music listening.

3.1 Comparison of Multi-Modal Fusion Methods

Because of the different database, it is difficult to quantitatively compare the proposed approach with existing ones. Alternatively, we treat AO and TO as the two baselines, and compare the classification accuracy of different fusion methods. We first use the features extracted by Marsyas and PsySound for audio feature representation, and the uni-gram based bag-of-words model for textual feature representation. The evaluation of the other two textual feature representations is reported in later subsections.

The results are shown in Table 2. It can be observed from the first and second rows that audio features and textual features are fairly complementary. While AO yields higher accuracy for arousal classification (78%), TO performs better for valence (73%). This result implies it is promising to fuse the two modalities since they encode different parts of semantics. Note the result that audio modality yields good accuracy for arousal classification but worse accuracy for valence has been found in previous works [2], [3]. Our experiment further shows lyrics are relevant to valence, but relatively irrelevant to arousal (this is reasonable since lyrics contain sparse melodic or rhythmic information).

Table 2 also indicates that the four-class emotion classification accuracy can be significantly improved by multi-modal fusion. Among the fusion methods (rows 3-5), LFSM achieves the best classification accuracy (57.06%) and contributes a 21% relative improvement over the audio-only baseline. It can also be observed that late fusion yields better result than early fusion. This seems to imply the individual strength of the two modalities should be emphasized separately. Besides, although $\text{LFLC}_{0.5}$ is slightly worse than LFSM, its classification accuracy for valence (74.83%) is the highest among the five fusion methods. This indicates that valence can be better modeled by considering both modalities (in

Table 2. Performance comparison of variant multi-modal fusion methods for 4-class emotion classification, arousal classification, and valence classification.

#	Methods	# of features	accuracy (4-class)	accuracy (valence)	accuracy (arousal)
1	AO	106	46.63%	61.15%	78.03%
2	TO	4000	40.01%	73.32%	61.95%
3	EFFC	4106	52.48%	70.54%	77.06%
4	$LFLC_{0.5}$	106/4000	55.34%	74.83%	77.88%
5	LFSM	106/4000	57.06%	73.32%	78.03%

Table 3. Performance comparison of uni-gram and PLSA feature representations for valence classification (# of test data is fixed to 160).

Methods	# of features	# of training data		
		760	400	200
Uni-gram	4000	73.21%	67.78%	58.70%
PLSA	100	72.85%	70.59%	66.53%

a late-fusion manner), while arousal can be modeled well by audio alone. We also vary α from 0 to 1 at a step of 0.1 and find the accuracy can reach 75.18% by setting α to 0.6, which indicates again that lyrics is more related to valence than the audio part.

3.2 Evaluation for PLSA Model

To assess the short text problem, we train a PLSA model with 21661 unlabeled lyrics to convert the bag-of-words feature space to the latent vector space of dimension 100 (Z=100). We conduct performance comparison of bag-of-words and PLSA feature representations for valence classification with different numbers of training data (the number of test data is fixed to 160) to simulate different levels of the shot text problem, which is more severe with smaller number of training data since more words in the test set would not have occurred in training.

Result shown in Table 3 indicates the classification accuracy of bag-of-words degrades significantly as the number of training data decreases. In contrast, because of the incorporation of unlabeled data and the more compact feature representation, PLSA exhibits robust performance. This result shows PLSA can be applied to mitigate the short text problem effectively. However, as the number of training data is sufficient and the short text problem may no longer exists, the classification accuracy of bag-of-words becomes similar to that of PLSA.

3.3 Evaluation for Bi-Gram Model

To assess the negation-term problem, first we deliberately add common negation words such as “no” and “not” to the stoplist and remove them from the bag-of-words model. The resulting similar classification accuracy implies the effect of

negation terms is hardly modeled by uni-gram. To address this issue, another text-only classifier is trained by using both uni-gram and bi-gram. However, the incorporation of bi-gram only slightly improves the classification accuracy of valence from 73.32% to 73.79%. To better model the effect of negation terms, more advanced methods are needed.

4 Related Work

The application of text analysis to song lyrics has been explored for artist indexing [23], structure extraction, and similarity search [24]. However, there have been rare attempts to leverage the information of lyrics to MEC. Some exceptions are [5], [6] and [25], all of which use either manually or automatically generated affect lexicons to analyze lyrics. We consider these lexicon-based approaches not principled since they are not applicable to all languages. In contrast, our approach is based on statistical natural language processing and thus more general and well-grounded. Another related work for analyzing the affect of text can be found in the field of blog analysis [26], [27]. Authors in [26] also adopt bag-of-words as feature representation and SVM for model learning. Interestingly, their classification accuracy for valence classification also reaches 74%, which is very close to our result (cf. Table 2).

5 Conclusion

In this paper we have described a preliminary multi-modal approach to music emotion classification that exploits features extracted from the audio and the lyrics of a song. We apply statistical natural language processing techniques to analyze lyrics. A number of multi-modal fusion methods are developed and evaluated. Experiments on a moderately large-scale database show that lyrics indeed carry semantic information complementary to that of the music signal. By the proposed late fusion by subtask merging, we can improve the classification accuracy from 46.6% to 57.1%. Using textual features also significantly improves the accuracy of valence classification from 61.2% to 73.3%. An exploration of more natural language processing algorithms and more effective features for modeling the characteristics of lyrics is underway.

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

References

1. Casey, M. et al: Content-based music information retrieval: current directions and future challenges. Proc. IEEE, Vol. 96, No. 4 (2008) 668–696
2. Yang, Y.-H. et al: A regression approach to music emotion recognition. IEEE Trans. Audio, Speech and Language Processing, Vol. 16, No. 2 (2008) 448–457.

3. Lu, L. et al: Automatic mood detection and tracking of music audio signals. IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 1 (2006) 5–18.
4. Cheng, H.-T. et al: Automatic chord recognition for music classification and retrieval. Proc. ICME (2008) 1505–1508.
5. Yang, D. et al: Disambiguating music emotion using software agents. Proc. ISMIR (2004) 52–58.
6. Chuang, Z.-J. et al: Emotion recognition using audio features and textual contents. Proc. ICME (2004) 53–56.
7. Chua, B.-Y. et al: Perceptual rhythm determination of music signal for emotion-based classification. Proc. MMM (2006) 4–11.
8. Omar Ali, S. et al: Songs and emotions: are lyrics and melodies equal partners. Psychology of Music, Vol. 34, No. 4 (2006) 511–534.
9. Fornäs, J.: The words of music. Popular Music and Society, Vol. 26, No. 1 (2003).
10. Cai, R. et al: MusicSense: Contextual music recommendation using emotion allocation modeling. Proc. ACM Multimedia (2007) 553–556.
11. Tzanetakis, G. et al: Musical genre classification of audio signals. IEEE Trans. Speech and Audio Processing, Vol. 10, No. 5 (2002) 293–302. http://marsyas.sness.net/.
12. Cabrera, D. et al: PSYSOUND: A computer program for psychoacoustical analysis. Proc. Australian Acoustic Society Conf. (1999) 47–54. http://psysound.wikidot.com/.
13. Geleijnse, G. et al: Efficient lyrics extraction from the web. Proc. ISMIR (2006).
14. Sebastiani, F.: Machine learning in automated text categorization. ACM CSUR, Vol. 34, No. 1 (2002) 1–47.
15. Want, J. et al: Short-text classification based on ICA and LSA. Proc. ISNN (2006) 265–270.
16. Hofmann, T. et al: Probabilistic latent semantic indexing. Proc. ACM SIGIR (1999) 50–57.
17. Thayer, R. E. et al: The Biopsychology of Mood and Arousal. Oxford University Press, New York (1989).
18. Russel, A.: A circumplex model of affect. Journal of Personality & Social Science, Vol. 39, No. 6 (1980) 1161–1178.
19. Smola, A. J. et al: A tutorial on support vector regression. Statistics and Computing (2004).
20. Chang, C.-C. et al: LIBSVM: a library for support vector machines. (2001) http://www.csie.ntu.edu.tw/ cjlin/libsvm/.
21. Snoek, C. et al: Early versus late fusion in semantic video analysis. Proc. ACM Multimedia (2005) 399–402.
22. LingPipe. http://alias-i.com/lingpipe.
23. Logan, B. et al: Semantic analysis of song lyrics. Proc. ICME (2004) 827–830.
24. Mahedero, J. et al: Natural language processing of lyrics. Proc. ACM Multimedia (2005) 475–478.
25. Cho, Y.-H. and Lee, K.-J.: Automatic affect recognition using natural language processing techniques and manually built affect lexicon. IEICE Trans. Information Systems, Vol. E89, No. 12 (2006) 2964–2971.
26. Leshed, G. et al: Understanding how bloggers feel: Recognizing affect in blog posts. Proc. ACM CHI (2006).
27. Abbasi, A. et al: Affect analysis of web forums and blogs using correlation ensembles. IEEE Trans. Knowledge and Data Engineering, Vol. 20, No. 9 (2008) 1168–1180.