Source Separation Improves Music Emotion Recognition

Jieping Xu[1], Xirong Li* [1,2], Yun Hao[3], Gang Yang[1]
[1]Multimedia Computing Lab, School of Information, Renmin University of China, 100872 China
[2]Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, 100872 China
[3]Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, USA
{xjieping,xirong,yanggang}@ruc.edu.cn, yunhao2@illinois.edu

ABSTRACT

Despite the impressive progress in music emotion recognition, it remains unclear what aspect of a song, i.e., singing voice and accompanied music, carries more emotional information. As an initial attempt to answer the question, we introduce *source separation* into a standard music emotion recognition system. This allows us to compare systems with and without source separation, and consequently reveal the influence of singing voice and accompanied music on emotion recognition. Classification experiments on a set of 267 songs with last.fm annotations verify the new finding that source separation improves song music emotion recognition.

Categories and Subject Descriptors

H.5.5 [**INFORMATION INTERFACES AND PRESENTATION**]: Sound and Music Computing—*Systems*

Keywords

Music emotion recognition, source separation

1. INTRODUCTION

Music is often referred to as a language of emotion. Performers and listeners communicate in this language with the utilization of varied acoustic cues such as tempo, sound level, spectrum, and articulation [3]. Automatically classifying music into a set of predefined emotion classes such as 'happy', 'sad', and 'anger' is useful for music organization and personalized music recommendation. Hence, music emotion recognition is an important topic in music information retrieval [2, 4, 6].

To predict emotions for a novel piece of music, the state of the art first segments the music into a number of short clips [9], as illustrated in Figure 1. Each clip is represented by multiple audio features describing timbre, melody, rhythm, etc. Then, each of the clips is classified using pretrained Support Vector Machines (SVMs) classifiers, and the classification results are aggregated as the final prediction.

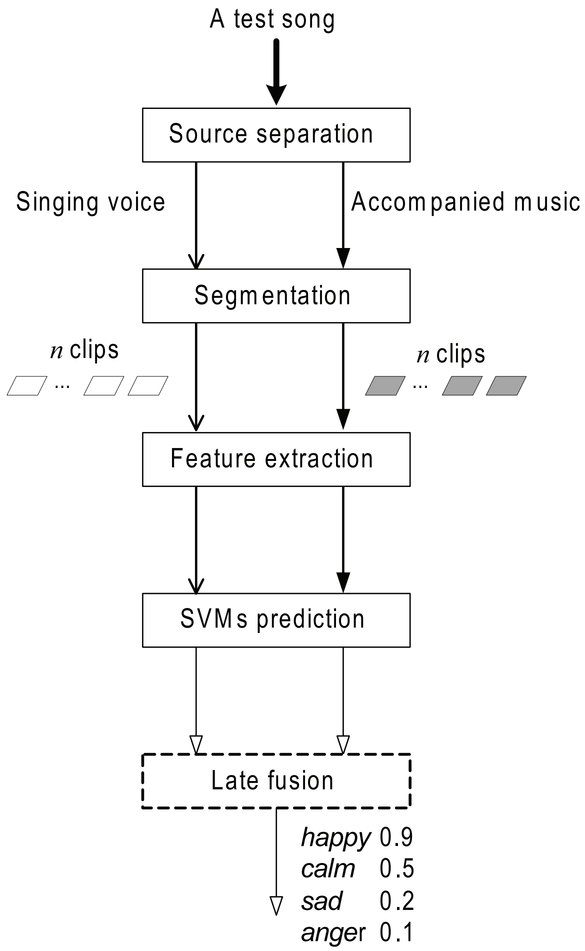

Figure 1: A conceptual diagram of the proposed music emotion recognition system. Different from previous works, we introduce source separation into the system, allowing us to study which aspect of song music is more important for emotion recognition.

Notice that music is often a mixture of multiple sound tracks. A song is singing voice blended with accompanied music, e.g., instrumental sounds. A psychological experiment by Siegwart and Scherer in 1995 [11] studied the vocal

*Corresponding author.

ICMR'14, April 01 - 04, Glasgow, United Kingdom
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2782-4/14/04$15.00.
http://dx.doi.org/10.1145/2578726.2578784.

expression of emotion in opera music. They reported that when it comes to voice samples, the utilization of acoustic cues differs substantially. Represented by timbre features, singing voice alone might be effective in distinguishing ‘calm’ from ‘sad’, but less effective when mixed with accompanied music. However, existing works on music emotion recognition do not consider separating singing voice and accompanied music. As a consequence, it remains unknown which of the singing voice, the accompanied music, or their combination has a larger influence on the expression of music emotion. A more fundamental question is *what aspect of a song is carrying more emotional information?*

In order to reveal the impact of the different aspects of a song on expressing its emotion, we propose to leverage *source separation* for music emotion recognition. In signal processing, source separation is to recover each signal from the combined signal in which several signals have been mixed together [8]. While source separation has been considered for music information retrieval [13], its use for music emotion recognition is, to the best of our knowledge, non-existing in the literature.

2. MUSIC EMOTION RECOGNITION BASED ON SEPARATED SOURCES

For a given song, we aim to build a system that can automatically predict the main emotion that a common user would perceive from the song. The main novelty is that we introduce a source separation component into the system, as illustrated in Figure 1. By doing so, the system can predict music emotions based on the individual sources, let it be singing voice or accompanied music.

Next, we describe source separation in Section 2.1, and the other part of the classification system in Section 2.2.

2.1 Source Separation

For a given song, we need source separation to separate singing voices from accompanied music. Our focus is to study the influence of the (imperfectly) separated sources on emotion recognition. So instead of developing new source separation methods, we opt to use established ones. In that regard, we employ the popular Flexible Audio Source Separation Toolbox[1] (FASST).

FASST implements a generic audio source separation framework [8], based on a library of structured source models that enable the incorporation of prior knowledge about each source via user-specifiable constraints. Time-mixed signals are taken as the sum of magnitude or power spectrogram of each sound signal. In particular, FASST models the music sources in each time-frequency bin by random variables which follow zero-mean Gaussian distributions. Each source is modeled by a spectral component overall all time-frequency bins, while each spectral component is given by a model based on nonnegative matrix factorization with the Itakura-Saito divergence [8]. Figure 2 shows the spectrograms of two songs and their separated singing voice and accompanied music. In the spectrogram of the singing voice, the energy concentrates on the low-frequency domain, showing obvious singing characteristics. The qualitative result shows that the singing voice and the accompanied music are reasonably separated.

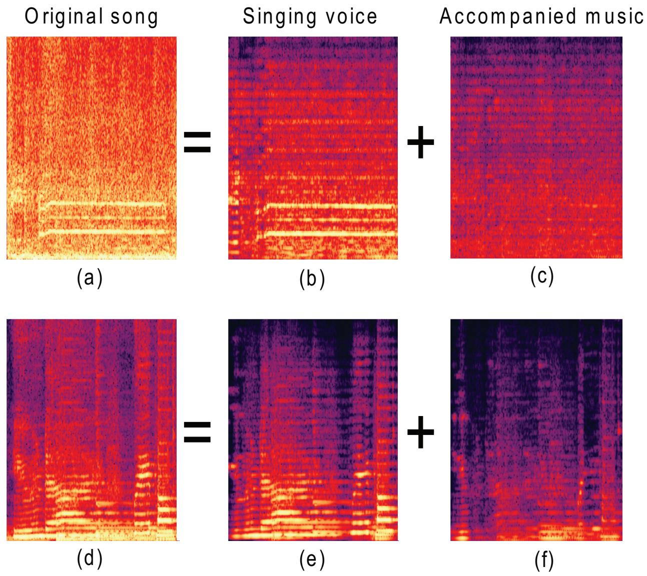

Figure 2: Source separation via spectrogram decomposition. The top row is a song of ‘anger’, while the bottom row is a song of ‘calm’.

2.2 The Classification System

As music emotion is often expressed in a short duration, a common practice is to segment each song into a number of short fixed-length clips and build a clip-based classification system [4]. So feature extraction, training, and prediction are all performed at the clip level.

For the ease of consistent description, let x be a song in consideration, and $\{c_i | i = 1, \ldots, m\}$ be a set of m predefined emotion classes. By source separation, x is now decomposed into two parts, namely the singing voice and the accompanied music, denoted as y and z, respectively. For the two parts, we segment them into n fixed-length clips, notated as $\{y_1, \ldots, y_n\}$ and $\{z_1, \ldots, z_n\}$. In contrast, $\{x_1, \ldots, x_n\}$ represent the clips of the original song without source separation. For each class c_i, we use $p_i(y_j)$ to denote its classifier, which outputs the probability that the j-th clip from the singing voice is a positive instance of the class. In a similar fashion, we define $p_i(x_j)$ and $p_i(z_j)$.

Classifier Training. We choose SVMs for its well recognized performance, as justified by MIREX, a leading benchmark for music emotion recognition [12]. To propagate annotations from the song level to the clip level, we simply treat all clips of a song as positive training examples of a class if the song is labeled as a positive instance of the class. For each class and for each source, we train a binary SVMs classifier in a one-versus-all manner using the LibSVM software [1]. The RBF kernel is used, with SVMs parameters optimized by multi-fold cross validation on the training data.

Prediction. The predicted class of a test song is obtained by aggregating the predictions of its clips. In particular, for the singing voice based system, we write the decision rule as

$$\operatorname*{argmax}_i \sum_{j=1}^{n} p_i(y_j). \qquad (1)$$

[1] http://bass-db.gforge.inria.fr/fasst

Table 1: Audio features used in our system. Each clip is represented by a combined 84-dimensional feature vector.

Type	Feature	Dim
Timbre	MFCC	26
	LPC	20
	Spectral Centroid	2
	Spectral Rolloff Point	2
	Spectral Flux	2
	Spectral Variability	2
	Zero Crossings	2
	Compactness	2
Intensity	Root Mean Square	2
	Fraction of Low Energy Windows	2
Other	Method of Moments	10
Melody	Chroma Feature	12

In a similar vein, the decision rule for the accompanied music based system is

$$\underset{i}{\operatorname{argmax}} \sum_{j=1}^{n} p_i(z_j). \tag{2}$$

Putting Eq. (1) and Eq. (2) together, we have late fusion of the two sources, that is,

$$\underset{i}{\operatorname{argmax}} \sum_{j=1}^{n} \left(p_i(y_j) + p_i(z_j)\right). \tag{3}$$

Feature Extraction. We adopt audio features that are found to be useful for music emotion recognition [2,5,6]. In particular, for each clip, we extract a combined 84-dimensional feature vector using jAudio [7]. The combined feature consists of a 72-dimensional acoustic feature and a 12-dimensional chroma feature, see Table 1. The acoustic feature includes timbre, intensity and Method of Moments. By identifying spectral components that differ by a musical octave, the chroma feature is relatively robust with respect to changes in timbre and accompanied music [5].

In sum, by comparing $p_i(x_j)$, $p_i(y_j)$ and $p_i(z_j)$, we verify whether source separation improves music emotion recognition, and which source is more important for expressing the emotion.

3. EVALUATION

3.1 Experimental Setup

Data. When the paper was written, no music data with emotion ground truth is available in the public literature including the MIREX forum which does not release data but asks participants to submit their softwares. Moreover, the emotional perception of a song can be subjective sometimes. So we turn to `last.fm`, a music tagging website where a song is tagged by many users independently. The accumulated frequency of a tag thus reflects the consensus of user feelings about the song. We consider the following four basic emotions, i.e., 'anger', 'calm', 'happy', and 'sad', each of which corresponds to one of the four quadrants in the Russell V-A space [10]. We select from `last.fm` western popular songs that are labeled with one of the four tags, and build a set of 267 songs. We empirically set the length of a clip to be 15 seconds, resulting in 4,542 clips in total, see Table 2 for the data statistics. The dataset is randomly split into two parts, 70% for training and 30% for testing.

Table 2: A set of 267 songs and 4,542 clips used in our experiments, 70% of the set for training and the other 30% for testing.

Class	No. of songs	No. of clips
anger	57	877
calm	76	1,301
happy	64	1,087
sad	70	1,277

Implemented Systems. To justify necessity of source separation, we implement and compare the following four systems:

1) Baseline: The baseline system that makes prediction based on original songs with no source separation,

2) Singing voice: Prediction based on the separated singing voice alone, see Eq. (1),

3) Background music: Prediction based on the separated accompanied music alone, see Eq. (2),

4) Late fusion: Combining the prediction of *2)* and *3)*, see Eq. (3).

Evaluation criterion. For each class, its accuracy is computed as the number of correctly predicted songs divided by the number of songs that are predicted as this class.

3.2 Experiments

Experiment 1. Does source separation help? As shown in Figure 3, the systems built upon the separated sources clearly outperform the baseline system. Compared to the baseline which obtains an accuracy of 0.371, the proposed system, either using the singing voice or using the accompanied music, scores a higher accuracy of 0.476 and 0.491, respectively. For a better understanding of the results, we report confusion matrices in Table 3. Without source separation, a song of 'calm' tends to be misclassified as 'sad' with an error rate of 0.478. After source separation, this error rate is reduced to 0.261 and 0.391 when using the two sources separately. Hence, we conclude that source separation is helpful for recognizing emotions of song music.

Experiment 2. Which source carries more emotion? As shown in Figure 3, comparing the two sources, the background sound is consistently better than the baseline for all the four classes, while the singing voice is worse than the baseline for the category 'anger'. By examining the confusion matrices, we observe that for the singing voice based system, 35.3% of the songs of 'anger' are misclassified as 'happy', 23.5% misclassified as 'calm', and 17.7% as 'sad'. The result suggests the audio features extracted from the singing voice are less robust than their counterparts from the accompanied music for recognizing the emotion of 'anger' in songs. In terms of the overall performance, the accompanied music is slightly better than the singing voice.

Experiment 3. Does late fusion help? As shown in Figure 3 and Table 3, late fusion lifts the performance, with an accuracy of 0.532. Note that the gain is obtained by using fully automated source separation, without the need of extra manual annotation. This result shows that combining the two separated sources in a late fusion manner is helpful.

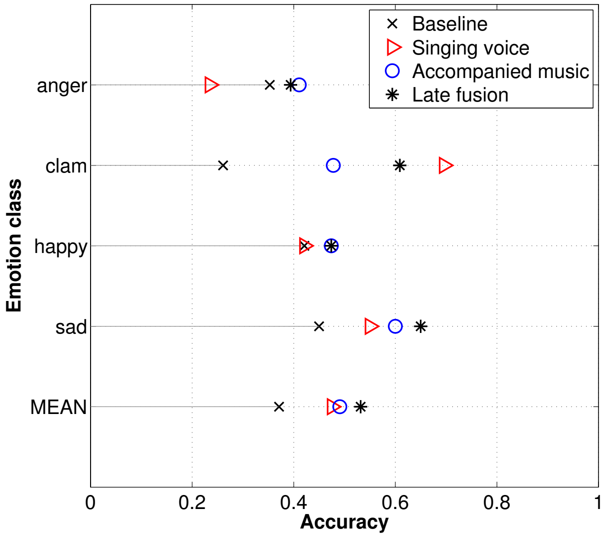

Figure 3: Comparing music emotion recognition with and without source separation. The systems using the separated sources outperform the baseline system. Combining the systems by late fusion further improves the classification accuracy.

Table 3: Confusion matrices.

	Predicted class			
	anger	calm	happy	sad
1) Baseline				
anger	**0.353**	0.177	0.293	0.177
calm	0	0.261	0.261	**0.478**
happy	0.316	0.158	**0.421**	0.105
sad	0.250	0.150	0.150	**0.450**
2) Singing voice				
anger	0.235	0.235	**0.353**	0.177
calm	0	**0.696**	0.043	0.261
happy	0.263	0.158	**0.421**	0.158
sad	0.100	0.100	0.250	**0.550**
3) Accompanied music				
anger	**0.411**	0.177	0.177	0.235
calm	0	**0.478**	0.131	0.391
happy	0.316	0.158	**0.474**	0.052
sad	0.100	0.050	0.250	**0.600**
4) Late fusion				
anger	**0.394**	0.176	0.215	0.215
calm	0	**0.609**	0.087	0.304
happy	0.316	0.158	**0.474**	0.052
sad	0.100	0.050	0.200	**0.650**

The result also suggests that singing voice and accompanied music are complementary to each other to some extent.

4. CONCLUSIONS

This paper presents a pilot study towards answering the question of what aspect of song music carries more information for expressing emotions. Our contribution is by introducing source separation into a standard music emotion classification system, we reveal the influence of the two separated sources, i.e., singing voice and accompanied music, on the classification performance. Experiments on a set of 267 songs with `last.fm` annotations support the following conclusions. Source separation is helpful. Compared to the baseline without source separation, the systems built on accompanied music improves the accuracy from 0.371 to 0.491. Comparing the two sources, accompanied music is more robust than singing voice. Combining the two sources by late fusion brings further improvements, reaching an accuracy of 0.532. These results suggest the joint use of source separation and late fusion for song music emotion recognition.

Acknowledgements. This work was supported by NSFC (61303184, 61003205), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (14XNLQ01, 13XNLF05), SRFDP (20130004120006), and SRF for ROCS, SEM.

5. REFERENCES

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[2] Y. Feng, Y. Zhuang, and Y. Pan. Popular music retrieval by detecting mood. In *ISMIR*, 2003.

[3] P. Juslin. Cue utilization in communication of music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(6):1797–1813, 2000.

[4] Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *ISMIR*, 2000.

[5] A. Klapuri, A. Eronen, and J. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):342–355, 2006.

[6] L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transaction on Audio, Speech, and Language Processing*, 14(1):5–18, 2006.

[7] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle. jAudio: A feature extraction library. In *ISMIR*, 2005.

[8] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, 2012.

[9] R. Panda and R. Paiva. Music emotion classification: Dataset acquisition and comparative analysis. *DAFx*, 2012.

[10] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[11] H. Siegwart and K. Scherer. Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi. *Journal of Voice*, 9(3):249–260, 1995.

[12] D. Stephen. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[13] E. Vincent and N. Ono. Music source separation and its applications to MIR. In *ISMIR*, 2010.