# EXPLORATION OF MUSIC EMOTION RECOGNITION BASED ON MIDI

**Yi Lin, Xiaoou Chen and Deshun Yang**
Institute of Computer Science & Technology, Peking University
lin.yi.tz@gmail.com
{chenxiaoou, yangdeshun}@pku.edu.cn

## ABSTRACT

Audio and lyric features are commonly considered in the research of music emotion recognition, whereas MIDI features are rarely used. Some research revealed that among the features employed in music emotion recognition, lyric has the best performance on valence, MIDI takes the second place, and audio is the worst. However, lyric cannot be found in some music types, such as instrumental music. In this case, MIDI features can be considered as a choice for music emotion recognition on valence dimension.

In this presented work, we systematically explored the effect and value of using MIDI features for music emotion recognition. Emotion recognition was treated as a regression problem in this paper. We also discussed the emotion regression performance of three aspects of music in terms of edited MIDI: chorus, melody, and accompaniment. We found that the MIDI features performed better than audio features on valence. And under the realistic conditions, converted MIDI performed better than edited MIDI on valence. We found that melody was more important to valence regression than accompaniment, which was in contrary to arousal. We also found that the chorus part of an edited MIDI might contain as sufficient information as the entire edited MIDI for valence regression.

## 1. INTRODUCTION AND RELATED WORKS

Music is a natural carrier to express and convey emotion. Some emotion models have been developed to describe emotion state. Russell's two-dimensional valence-arousal (V-A) model [4] consisted of two independent dimension of valence and arousal. Valence stands for appraisal of polarity and arousal stands for the intensity of emotion. Mehrabian [10] extended this approach and developed a three-dimensional pleasure-arousal-dominance (PAD) model, where dimension P distinguishes the positive-negative quality of emotion, dimension A refers to the intensity of physical activity and mental alertness, and dimension D refers to the degree of control. In this paper, we focused on Russell's V-A model, especially the valence (V) dimension, which corresponds to pleasure (P) dimension in PAD model. Several types of feature have been developed to represent a piece of music. Audio and lyric features are commonly considered in the research of music emotion recognition, whereas MIDI features are rarely used [1]. Emotion recognition can be viewed as a multiclass-multilabel classification or regression problem [1]. In this paper emotion recognition was treated as a regression problem.

This paper focused on music emotion recognition with MIDI features, which can be extracted directly from MIDI files. Unlike audio data, MIDI is a kind of the electronic score. Symbolic representations of music (such as key, pitch, tempo, etc.), which are high-level musicological symbolic features reflecting music concepts, can be easily extracted and calculated from MIDI by toolkits such as jSymbolic [7]. Therefore, MIDI features may be more effective on music emotion regression.

Oliveira and Cardoso [3] constructed a dataset of 96 western tonal music (film music) pieces, which lasted between 20 seconds to 1 minute. These pieces were in MIDI format and each piece might express only one type of affective content. 80 listeners were asked to label these musical pieces with affective labels on valence and arousal, respectively. Both musicological symbolic features (e.g., tempo, note duration, note density, etc) and acoustical features (such as MFCCs) were extracted. After feature selection, they performed SVM regression and received correlation coefficient of 81.21% on valence and 84.14% on arousal from 8-fold cross validation experiment. Then Oliveira and Cardoso [3] selected the most important features that were identified separately during feature selection results for valence and arousal and performed the 8-fold cross validation of SVM regression again. The most important features selected were all symbolic features. The performance evaluated in terms of correlation coefficient reached 71.5% on valence and 79.14% on arousal. Oliveira and Cardoso's work demonstrated that MIDI is effective on music emotion recognition.

Oliveira and Cardoso's work [3] only focused on MIDI, whilst Guan et al. [2] compared music emotion recognition performance among audio, lyric, and MIDI features. They presented an AdaBoost approach on 1687 Chinese songs. A wave file and a lyric file were collected for each song and a MIDI files was converted from the

wave file. Audio, lyric, and MIDI features were extracted separately. Guan et al. [2] applied their AdaBoost.RM approach on a multi-modal feature set in which audio, lyric and MIDI features were combined together. The performance of the experiment was 74.2% on valence in terms of correlation coefficient. They also applied regression to audio, lyric, and MIDI features, respectively. The results showed that the regressor built using lyric features yielded the best performance on valence; the performance was 62.3% in terms of correlation coefficient, whereas the regressor built using audio features had the poorest performance of 47.3% and the regressor built using MIDI features was in between, 54.1%.

Based on findings from the related works, there were two aspects of views towards MIDI. Oliveira and Cardoso [3] regarded MIDI as an object on which the emotion value is labeled. Although they received a good result on emotion regression, it was very difficult to find a satisfactorily-composed MIDI file for a song. A satisfactorily composed MIDI file should be a MIDI that sounds exactly the same to the original song when listened to. Instead Guan et al. [2] regarded MIDI as an intermediate representation of music audio data. They firstly converted audio files into MIDI files and then extracted MIDI features from the MIDI files. This was a much easier way for obtaining example MIDI files, as there were large amount of audio files available on the Internet.

In our work, we took Guan et al.'s [2] view and regarded MIDI as an intermediate representation of music audio data. We provided a systematic exploration of the effect and value of using MIDI features for music emotion recognition. Two types of MIDI files were. One type of MIDI files was converted from audio files, e.g., mp3, wav, etc. MIDI files obtained in this way were referred as converted MIDI. The other type of MIDI files was composed by musicians and composers via a score editing tool such as Guitar pro [5]. MIDI files obtained in this way were referred as edited MIDI. Oliveira and Cardoso [3] only considered edited MIDI and Guan et al. [2] only considered converted MIDI, whereas we considered both in our work.

Our work consisted of two parts. In part one we compared the music emotion regression performance of audio, lyric, and MIDI features. We carried out emotion regression with audio, lyric, and MIDI features separately rather than combined them together. We also compared the music emotion regression performance between converted MIDI and edited MIDI; this work was not included in the paper of Guan et al [2]. Moreover, we played the 653 edited MIDI files back acoustically, recorded that, and converted that back to MIDI automatically to investigate what helped to predict emotion. In part two we investigated the edited MIDI from three aspects of music: chorus, melody, and accompaniment.

The paper is organized as follows. Section 2 presents the datasets and features. Section 3 reports the experiments and results. Section 4 provides the analysis of the experiment results. Section 5 concludes the key findings.

## 2. DATASETS AND FEATURES

### 2.1 Datasets

We used Guan et al.'s 2500-Chinese-song list, where each song was annotated with a PAD label [2]. Datasets were constructed according to the 2500-Chinese-song list. All these Chinese songs were composed based on Western 12-tone equal temperament.

In our work, 7 datasets were constructed as described in the following section (including four sets of edited MIDI files, one set of audio files, one set of lyric files and one set of converted MIDI files). Among the four types of data (edited MIDI, converted MIDI, audio, and lyric), the edited MIDI data was the most difficult to be collected among the four kinds of data: edited MIDI, converted MIDI, audio, and lyric. Therefore we firstly constructed our edited MIDI datasets to determine how many songs could be included in our work. Then we constructed audio dataset and lyric dataset. Finally we constructed converted MIDI dataset in which converted MIDIs came from audios.

#### *2.1.1 Edited MIDI*

We applied two ways for collecting edited MIDIs. One way was to download edited MIDI files from Internet that were composed by musicians and composers. In our work, the notes of Edited MIDI files that collected in this way are entered directly into a sequencer program or notation software rather than being recorded directly from playing on a MIDI instrument. The other way was to download scores from Internet that were written by musicians and composers and then translated the scores into MIDI files via a score editing tool called Guitar pro [5].

The final edited MIDI dataset contained 653 MIDI pieces, which lasted between 15 seconds to 7 minutes. This dataset was a subset of the 2500 Chinese songs. Each MIDI in the dataset contains melody, accompaniment, and at least two different timbres, so that the MIDI could sound as similar as possible to the original songs when listened to. If multiple MIDI versions of a song were available, the one that was the closest to the original song when listened to would be retained. This dataset was referred as edit-MIDI.

We split each of the 653 MIDI pieces into two parts: melody and accompaniment. This work was carried out by manually extracting the tracks corresponding to melody and accompaniment, respectively, from one MIDI to create two new MIDIs. One represented melody and the other represented accompaniment. Thus two more datasets were constructed. The dataset, which contained 653 MIDI pieces corresponding to the melody part, was referred as edit-MIDI-melody; the other dataset, which

contained 653 MIDI pieces corresponding to the accompaniment part, was referred as edit-MIDI-accom.

Short versions of the 653 MIDIs in edit-MIDI are also collected. Each short MIDI was the chorus part of the corresponding music. Most of these pieces were used as cell phone ringtones, lasting between 15 seconds to 40 seconds. This dataset is referred as edit-MIDI-chorus.

To conclude, we built four datasets for edited MIDI: edit-MIDI, edit-MIDI-melody, edit-MIDI-accom, and edit-MIDI-chorus. It worth to mention that it has been very difficult to find for a piece of music a satisfactorily-composed edited MIDI that sounds exactly the same as the original music. The reasons are two-fold: on one hand, not all music had a MIDI; on the other hand, the composer might not be willing to share the MIDI file.

#### *2.1.2 Audio and Lyric*

For the 653 songs corresponding to the edited MIDI, we downloaded their wav audio files and lyric files, which constituted the audio dataset and lyric dataset, respectively.

#### *2.1.3 Converted MIDI*

Converted MIDI dataset contained 653 MIDI files converted from the 653 wav files. WIDI Recognition System Pro 4.0 [6] was used to help with the conversion. This dataset was referred as conv-MIDI.

There were several differences between edited MIDI and converted MIDI. Firstly, each converted MIDI pieces was of the same length as the original song, whereas the length of edited MIDI pieces varied. Secondly, there was only one timbre existed in converted MIDI; the timbre was set to be Instrumental Grand in WIDI. Edited MIDI, however, contained at least two different timbres. Finally, the converted MIDI was very different from the edited MIDI of the same song when listened to. The converted MIDI sounded like its pitches were in a mess and it was hard or even unable to distinguish how the melody went. Edited MIDI, however, sounded similar to or even the same as the original song.

### 2.2 Features

MIDI features were extracted from the MIDI files by jSymbolic [7]. For each MIDI file 112 types of MIDI features were extracted to compose a feature vector of 1022 dimensions. Audio features were extracted from the wave files by jAudio [8]. For each audio file 27 types of audio features were extracted to compose a feature vector of 112 dimensions for each song. After lyric files were pre-processed with traditional NLP tools including stop-words filtering and word segmentation, unigram features were extracted from the lyrics file to compose a feature vector of 13251 dimensions.

## 3. EXPERIMENTS AND RESULTS

Supervised feature selection was carried out on each of the feature sets to reduce the number of features and to improve the regression results. Correlation-based Feature Subset Selection with BestFirst was applied as its search method in our work [9].

Following results of regression experiments were obtained using 5-fold cross validation of SMO regression with RBFKernel [9]. SMO regression [11] implements support vector machine for regression and in our work we used Radial Basis Function (RBF) as its kernel function. The performance of regression was measured in terms of correlation coefficient (CF).

Russell's two-dimensional valence-arousal (V-A) model [4] was employed to measure music emotion.

### 3.1 Comparison among Lyric, Audio, Edited MIDI, and Converted MIDI

Firstly, regression was applied on edit-MIDI, conv-MIDI, audio, and lyric datasets separately to compare the MIDI features with commonly used audio and lyric features. The results are showed in Table 1.

Table 1 shows the regression performance of lyric, audio, conv-MIDI, and edit-MIDI on valence and arousal. It worth noting that conv-MIDI was found to perform better than audio and worse than lyric on valence. We also found that the performance of edit-MIDI was much worse than conv-MIDI.

| Dataset | V | A |
|---|---|---|
| Lyric | 78.81% | 66.52% |
| Audio | 54.45% | 76.5% |
| Conv-MIDI | 57.09% | 74.96% |
| Edit-MIDI | 46.42% | 53.37% |

**Table 1.** Performance of lyric, audio, conv-MIDI, and edit-MIDI.

To analyze the difference between edited MIDI and converted MIDI, we examined the remaining features on valence after feature selection. For conv-MIDI, there were 10 types of features remaining that consisted of 60 features. For edit-MIDI, there were 37 types of features remaining that consisted of 315 features. Among these types of remaining features, there were 7 types of features that were common to both converted MIDI and edited MIDI. The remaining feature types on valence are shown in Table 2.

In Table 2, row 1 listed the remaining feature types that belonged to converted MIDI only; row 3 listed the remaining feature types that belonged to edited MIDI only; row 2 listed the remaining feature types that belonged to both converted MIDI and edited MIDI. In row 1, it can be seen that there were only 3 types of features (e.g.

in row 1: Duration, Combined Strength of Two Strongest Rhythmic Pulses, and Rhythmic Variability) that were included in features of converted MIDI, but not in features of edited MIDI. In order to investigate the importance of these three types of features, we carried out the regression again without these 3 types of features on conv-MIDI. The performance of the experiment dropped 3.01% (from 57.09% to 54.07%) in terms of CF on valence.

| | |
|---|---|
| conv-MIDI only | Duration<br>Combined Strength of Two Strongest Rhythmic Pulses<br>Rhythmic Variability |
| conv-MIDI & edit-MIDI | Chromatic Motion<br>Strength of Strongest Rhythmic Pulse<br>Variability of Note Duration<br>Basic Pitch Histogram<br>Beat Histogram<br>Melodic Interval Histogram<br>Time Prevalence of Pitched Instruments |
| edit-MIDI only | Amount of Apreggiation<br>Average Melodic Interval<br>Brass Fraction<br>Changes of Meter<br>Dominant Spread<br>Glissando Prevalence<br>Most Common Melodic Interval Prevalence<br>Most Common Pitch Class Prevalence<br>Note Density<br>Number of Common Pitches<br>Quality<br>Fifths Pitch Histogram<br>Melodic Interval Histogram<br>Note Prevalence of Pitched Instruments<br>…… |

**Table 2.** The remaining feature groups on valence after feature selection.

### 3.2 Conversions of the Edited MIDI

In order to investigate whether it was the process of conversion from the original audio to MIDI that helped predicting emotion, we went through a two-step experiment. Firstly we played the 653 edited MIDI files back acoustically and recorded the sounds (MIDI-WAV dataset); and secondly, the files were converted back to MIDI automatically (MIDI-WAV-MIDI dataset). We then examined the changes of the regression performance on valence. The results are showed in Table 3.

| Dataset | V |
|---|---|
| Edit-MIDI | 46.42% |
| MIDI-WAV | 26.04% |
| MIDI-WAV-MIDI | 43.83% |

**Table 3.** The performance of tow conversions of the edited MIDI.

Table 3 shows how the valence regression performed on the three datasets obtained from the two conversions. The performance was measured in terms of CF. Results in Table 3 revealed that the performance of MIDI-WAV-MIDI is lower than that of Edit-MIDI by 2.59% (from 46.42% to 43.83%).

### 3.3 Melody and Accompaniment of Edited MIDIs

The edited MIDI sounded similar to, or even the same as the original song. In our work each edited MIDI file was split into two parts: melody and accompaniment. This allowed us to measure the performance of these two parts on music emotion regression separately. However, we were not able to experiment them on converted MIDI, because melody or accompaniment could not be extracted from converted MIDI.

The experiment results are showed in Table 4.

| Dataset | V | A |
|---|---|---|
| Edit-MIDI-melody | 46.26% | 44.8% |
| Edit-MIDI-accom | 39.51% | 48.94% |

**Table 4.** Performance of melody and accompaniment of edited MIDI.

Table 4 shows the performance of melody and accompaniment of edited MIDI in terms of CF. Melody was found to perform better than accompaniment on valence regression, which was in contrary to arousal.

### 3.4 The Chorus of Edited MIDI

In most cases the chorus of a song is the emphatic part that reflects music concept. The chorus may express only one type of affective content.

We collected edited MIDI files containing only the chorus part of the corresponding songs and applied emotion regression on them. The results are showed in Figure 1.

Figure 1 shows the performance of the chorus dataset with different dataset size. Meanwhile we compared the performance of the chorus dataset with the performance of the entire MIDIs dataset (i.e. Edit-MIDI dataset). The number on the vertical axis refers to the number of MIDI files and the percentage on the horizontal axis refers to the performance of regression on valence measured in CF. The black bars refer to the performance of the chorus dataset and the gray bars refer to the performance of the corresponding entire MIDIs dataset. For each dataset size, we carried out the experiments 3 times by randomly choosing data examples.

Figure 1 revealed that the performance on the chorus dataset was very close to that on the entire MIDIs dataset.

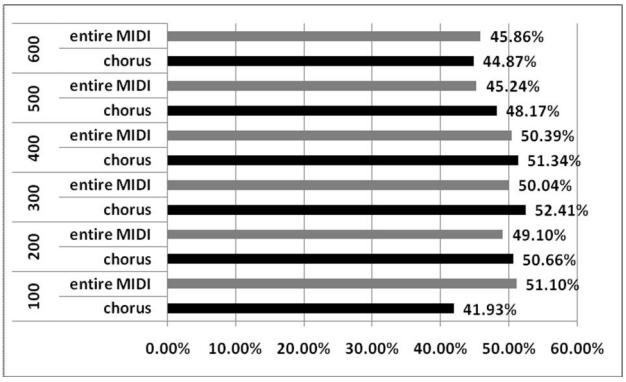


**Figure 1.** The performance of the chorus.

## 4. RESULTS ANALYSIS

The result showed in row 1 to 4 of Table 1 aligns with the work of Guan et al [2]. Lyric performed the best, converted MIDI was the second and audio performed the worst for valence regression. By converting audio to MIDI and then using MIDI features extracted from converted MIDI files, converted MIDI performed better than the audio by 2.36% on valence.

The result showed in row 3 to 4 of Table 1 indicated that the edited MIDI performs worse than the converted MIDI by 10.67% on valence.

To analyze the difference between edited MIDI and converted MIDI, we investigated the remaining features on valence after feature selection, which was shown in Table 2. The performance of the experiment dropped 3.01% (from 57.09% to 54.07%) in terms of CF on valence. This result indicated that, three types of features (Duration, Combined Strength of Two Strongest Rhythmic Pulses, and Rhythmic Variability) have stronger ability to express and distinguish emotion on valence.

Table 3 shows how the performance varied with the two conversions processes. The results indicated that the MIDI files converted from audio might not perform as well as the original edited MIDIs. From Table 1 we found that the conv-MIDI performed better than the edit-MIDI; the reason for this result might be that the source audio of conv-MIDI was much better than the source audio of MIDI-WAV-MIDI rather than that the process of conversion from the audio to MIDI that helped predicting emotion. The source audio of conv-MIDI was the original audio that people listened to, while the source audio of MIDI-WAV-MIDI was the audio files synthesized from the edited MIDI files, which were not exactly the same as the originally performed audio when listened to. If a large amount of satisfactorily-composed edited MIDIs were available, the regression performance of edited MIDIs might well be better than that of converted MIDIs. However, it was not always practical and feasible to obtain such perfect dataset. Considering the difficulty in collecting satisfactorily-composed edited MIDIs, converted MIDIs can be considered as a good choice for music emotion regression.

Table 4 shows that for edited MIDI the melody MIDI performed better than the accompaniment MIDI by 6.75% on valence and accompaniment performed better than the melody by 4.14% on arousal; this result indicates that melody is more effective and important to distinguish the positive-negative quality of affective content, and accompaniment is more effective to distinguish intensity of physical activity and mental alertness.

Figure 1 shows the regression performance on the chorus dataset and entire MIDIs dataset with different dataset size. The results showed that the performance on chorus dataset was very close to that on entire MIDIs dataset. Most of the entire MIDIs were sufficiently long to contain more than chorus part of music. On one hand, the result revealed that the use of chorus instead of the entire song did not improve the valence regression in terms of edited MIDI. On the other hand, the result showed that the chorus part of an edited MIDI might have contained as sufficient information as the whole edited MIDI for valence regression.

## 5. CONCLUSION

In this presented work, much valuable findings were obtained. Firstly, we found that the MIDI features extracted from converted MIDI files performed better than audio features that were extracted from audio files Secondly, we found the edited MIDI performed worse than converted MIDI under the realistic conditions. Therefore, the converted MIDI could be considered as a good choice for music emotion regression rather than the edited MIDI. We also compared and illustrated the differences between them based on features. The results indicated that three types of features (Duration, Combined Strength of Two Strongest Rhythmic Pulses, and Rhythmic Variability) have stronger ability to express and distinguish emotion on valence. Finally, we decomposed the edited MIDI and explored three aspects that were believed to be important to music emotion recognition: melody, accompaniment, and chorus. Two conclusions were drawn from the experimental results. One was that melody was more effective to valence regression and accompaniment to arousal; the other one was that the chorus of an edited MIDI may have contained as sufficient information as the whole edited MIDI for valence regression.

## 6. ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No.61170167) and Beijing Natural Science Foundation (4112028).

## 7. REFERENCES

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull: "Music emotion recognition: A state of the art review, " *11th International Society for Music Information Retrieval Conference*, 2010

[2] Di Guan, Xiaoou Chen, and Deshun Yang: "Music Emotion Regression based on Multi-modal Features," *9th International Symposium on Computer Music Modeling and Recognition*, 2012.

[3] A. P. Oliveira, and Amılcar Cardoso: "Modeling affective content of music: a knowledge base approach," *Sound and Music Computing Conference*, 2008.

[4] J. A. Russell, "A circumspect model of affect," *Journal of Psychology and Social Psychology*, vol. 39, no. 6, p. 1161, 1980

[5] http://www.guitar-pro.com

[6] http://www.widisoft.com

[7] C. McKay, and I. Fujinaga: "jSymbolic: A feature extractor for MIDI files," *Proceedings of the International Computer Music Conference*, 2005

[8] D. McEnnis, C. McKay, and I. Fujinaga: "jAudio: A Feature Extraction Library," *Proceedings of the International Conference on Music Information Recognition*, 2005

[9] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka

[10] Mehrabian, A.: "Framework for A Comprehensive Description and Measurement of Motional States," *Genetic, Social, and General Psychology Monographs*, vol. 121, pp. 33—361, 1995

[11] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, and K.R.K. Murthy: "Improvements to the SMO Algorithm for SVM Regression," *IEEE Transactions on Neural Networks*, 1999