Enhance Popular Music Emotion Regression by Importing Structure Information

Xing Wang, Yuqian Wu, Xiaoou Chen, Deshun Yang*
*Institute of Computer Science and Technology, Peking University, Beijing, P.R. China 100871
E-mail:{icstwangxing,wuyuqian,chenxiaoou,yangdeshun}@pku.edu.cn

***Abstract*—Emotion is a useful mean to organize music library, and automatic music emotion recognition is drawing more and more attention. Music structure information is imported to improve the result for music emotion regression. Music dataset with emotion and structure annotations is built, and features concerning lyrics, audio and midi are extracted. For each emotion dimension, regressors are built using different features on different type of segments in order to find the best segment for music emotion regression. Results show that structure information can help improve emotion regression. Verse is good for pleasure recognition, while chorus is good for arousal and dominance. The difference between verse and chorus can also help improve regressors.**

I. Introduction

There have been a number of studies on music emotion recognition (MER) recent years. Most work analyze music emotion based on the whole song or 30 seconds segments. However, popular music has regular structure and each type of structure has its own functions. The verse sections describe the background, while the chorus sections are summary and of greater emotional intensity. But it has not been explored which schema to choose segments for MER is the best. In addition, the verses are usually arranged to make the choruses more impressive. The difference between structures might also be helpful for MER.

In this work, we explore how structure affect the emotion regressors. Continuous PAD (Pleasure - Arousal - Dominance) emotion model is adopted to represent music emotion [1]. P distinguishes the positive-negative quality of emotional states, A refers to the intensity of physical activity and mental alertness, and D is defined in terms of control versus lack of control. A music data set with emotion annotation and structure annotation is built, and features of different types and different music structures are extracted to build emotion regressors. By comparing different structures and different features, we come to the following conclusion: Using verse or chorus segment of songs is better than simply choosing segments based on time or using the whole song. For pleasure dimension, Verses are more useful; For arousal and dominance dimension, Choruses are more useful. The Verse-Chorus difference can contribute to the improvement of pleasure recognition.

II. Related Work

There are three main granularities for the music excerpts used in MER. The most commonly used music excerpt is a 20-30 second segment because they think the emotion in short segments is stable. The segments are chosen based on time position, structure position or chosen manually [2-5]. But each song has main emotion which might be different from the emotion of segments. The second type of music excerpt is the whole song. This is adopted by work such as Hu [7] and Guan [8]. They extract feature from the whole song, which neglect the different function of different structures. The third type of music excerpt is at the frame level [9][10]. The emotions are collected continuously using interface such as MoodSwing. These work face the same one-main-emotion problem with the first type. In this work, we annotate emotion based on the whole song, try to improve the regressors by importing structure information and analysis how different structures affect the emotion recognition.

Though music structure have been imported to help other music information retrieval task, they are usually not taken into consideration in MER. Namunu proposed a novel beat space segmentation music structure analysis method to help music semantics understanding such as music transcription, summarization, retrieval and streaming [11]. For MER task, Carlos [12] extracted three 30-second music segments from the beginning, middle and end of each song. Ensemble learning is adopted to merge the results of different classifiers built using different segments. These type of segment are not related with the structure such as verse and chorus. Few study, if any, has been conducted to investigate the influence of music segmentation on emotion recognition [13].

III. Music Structure

A. Typical music structures

A typical structure arrangement might be as followings: Intro - Verse1 - Verse2 - Chorus - Chorus - Instrument - Verse2 - Chorus - Chorus - Bridge - Chorus - Outro. Typical structures and their functions are listed below [14].

Introduction is a unique section that comes at the beginning of the piece. Verse is the main part of a song. The story is usually presented in this type of sections. Chorus is the element of the song that repeats at least once both musically and lyrically. It is almost always of greater musical and emotional intensity than the verse. Bridge is usually used for avoid the monotonous continuous chorus sections. Instrument solo is a transitional section designed to showcase an instrumentalist. Outro is used for ending a song.

B. Statistics of music structure

The average length and ratio of different structures is shown in TABLE I. Choruses and verses appear more than other structures, they occupy about 70% of the whole song together.

TABLE I
PROPERTIES OF EACH TYPE OF STRUCTURE

	Intro	Verse	Chorus	Inst.	Bridge	Outro
# per song	1.00	3.37	4.12	1.21	1.14	1.00
Length	25.74	75.61	93.16	24.87	21.25	28.39
Ratio	0.10	0.31	0.38	0.10	0.09	0.11

C. Segments Statistics

Besides segments based on music structure, we also extract three commonly used segments for music emotion recognition. SSE is segment of 30 seconds with strongest energy. AFTER30 is the segment from 30s to 60s. MID30 is the 30 seconds in the middle of the song. TABLE II shows the time length of each type of structure for above types of segments. The value is the average length of subsection in the segments by removing zero. We can find that the most common structure type is chorus and chorus occupies a large portion of the SSE . Verse and Instrument also occur very commonly because they are usually ahead of chorus section, but their average length is very short because they are the very short beginning of the SSE.

TABLE II
SSE PROPERTIES

Seg.	Intro	Verse	Chorus	Inst.	Bridge	Outro
SSE	14.9	7.6	24.5	7.1	10.5	8.1
MID30	-	16.8	12.3	13.9	11.2	-
AFTER30	8.3	24.1	9.5	9.5	13.5	-

IV. DATASET AND FEATURES

A. Dataset

The dataset consists of 507 songs with PAD value annotations. 14 volunteers whose ages range from 22 to 40 use Self Assessment Manikins (SAM) to annotate the songs with integer PAD values ranging from -4 to 4. There are seven annotations in average and at least five annotations for each song. The final emotion values are the average value of all annotations. The structure of song is annotated by one expert manually.

B. Audio features

Audio Features are extracted with MIRtoolbox [15]. The Features extracted can be grouped into several groups as shown in TABLE III. The features are grouped into four perceptual dimensions of music listening according to Song's work [16]. Feature values are calculated for each frame and then summarized for each excerpt. Then summarized feature values are calculated based on the values of all frames. For Low energy, Fluctuation Peak Pos., Fluctuation Mag. and Fluctuation Centroid, only the mean value is calculated For all other features, six value are calculated: mean, std, slope, PeriodFreq, PeriodAmp, PeriodEntropy.

TABLE III
AUDIO FEATURES GROUPS

Group	#	Features	#	Features
Dynamics	6	RMS	1	Low energy
Rhythm	1	Fluctuation Peak Pos.	6	tempo
	1	Fluctuation Peak Mag.	6	Attack Time
	1	Fluctuation Centroid	6	Attack Slope
Spec.	6	Centroid	6	Brightness
	6	Spread	6	Skewness
	6	Kurtosis	6	Rolloff95
	6	Rolloff85	6	Spectral Entropy
	6	Flatness	6	Roughness
	6	Irregularity	6	Spectral flux
	6	Zero crossing rate	78	MFCC
	78	DMFCC	78	DDMFCC
Harmony	6	Chromagram Peak Pos.	6	Key clarity
	6	Chromagram Peak Mag,	6	Key mode
	6	HCDF		

C. Midi features

We transcript audio wav files into mid files using WIDI[1]. JSymbolic [17] are then adopted to extracted high level musical feature from mid file. Features are divided into 7 groups: Instrument, Texture, Dynamics, Pitch statistics, Melody, Chord, and Rhythm.

There are flaws with the WIDI and jSymbolic toolkit which will affect the regressors built using MIDI features. Chord features, which affect pleasure dimension a lot, are not implemented in current jSymbolic toolkit. Instrument features have little meaning as the WIDI cannot distinguish between different instruments, and all the notes are piano notes. Texture features are calculated from multiple tracks. It will not work because WIDI transcripts the music into only one track. Besides, the notes which are separated originally overlap with each other in the only track. This will affect features concerning melody.

D. Lyrics features

Vector Space Model (VSM) is adopted to extract features from lyrics. As we use a Chinese dataset, the pre-processing is different from that of English lyrics. Chinese sentences do not have space between words and special tools are needed for word segmentation. Word segmentation is done on lyrics using MMseg package[2] to get bag of words B_m for a song m. For each emotion dimension e, the feature vector $\boldsymbol{\mu}_m$ is built based on B_m. Binary feature vector $\boldsymbol{\mu}_m$ is used for regression [8].

$$\boldsymbol{\mu}_m = [u_1, ..., u_n]^T \tag{1}$$

where u_i is 1 if w_i occurs in B_m else 0, $w_i \in L_e$ and L_e is the lexicon used for VSM, n is the size of L_e, $e \in \{P, A, D\}$.

Lexicon of a large size will bring noise and consume huge amount of computational resources. Chi-square feature

[1]http://www.widisoft.com/english/products.html
[2]https://pypi.python.org/pypi/mmseg/1.3.0

selection is adopted to reduce the length of feature vector, which is quite effective for text classification. The χ^2 statistic measures the lack of independent between a feature word w and a emotion class c and is zero if w and c are independent. All words are ordered by decreasing χ^2 statistics value, the first 2000 words are selected for the lyrics lexicon L_e.

V. Experiment

In this section, we will firstly introduce the measures and machine learning package used in the experiments. Then we will show results for regressors built on different types of features from different music structures. Further more, we explore how different kinds of audio and midi features affect the regressors. At last, we will show the effect of difference between sections on the MER.

A. Measures and tools

Correlation coefficient (CF) statistic developed by Karl Pearson is adopted to measure the performances of regressors. We use 5-fold cross validation to evaluate the performance and the mean CF values are reported. The feature selection for lyrics features is only done with the training dataset to avoid over-fitting.

Weka, a machine learning toolkit is used in our experiment. SMOReg is used to build regressors with RBF Kernel and default parameters.

B. Comparison of different structures

Firstly, results of regressor built with different features based on different kinds of segment will be presented to show how structure of music affects the emotion recognition. Then, we will dive deep into audio features and midi features in detail to find the best features and segment combination. At last, we will show results concerning the effect of differences between verse and chorus on emotion.

TABLE IV
Comparison of regression results using different features

	Fea.	Verse	Chorus	SSE	AFTER30	MID30	All
P	Lyrics	0.449	0.388	-	-	-	**0.583**
	Audio	**0.608**	0.562	0.477	0.541	0.527	0.582
	MIDI	**0.497**	0.468	0.332	0.476	0.404	0.417
A	Lyrics	0.139	**0.237**	-	-	-	0.222
	Audio	0.713	0.819	0.755	0.646	0.694	**0.822**
	MIDI	0.641	**0.761**	0.674	0.594	0.677	0.692
D	Lyrics	0.211	0.230	-	-	-	**0.380**
	Audio	0.545	**0.653**	0.566	0.549	0.537	0.648
	MIDI	0.481	**0.592**	0.533	0.469	0.537	0.568

1) Regression using different features based on different excerpts: Firstly, we can find from TABLE IV that the best performance is achieved using only verse or chorus except for lyrics features on pleasure and dominance dimension, they are even better than the whole song as shown in column 'All'. Verses are better than choruses on pleasure dimension, while choruses are better than verses on arousal and dominance dimension. Besides, we can find that commonly used segments - SSE, MID30 and AFTER30 usually cannot get the best performance. The performance of these segments is related with the verse and chorus ratio which is shown in TABLE II. AFTER30 segments have larger ratio on verses and they are good for pleasure recognition. SSE segments have lager ratio on chorus and they are good for the arousal and dominance recognition. We do not conduct experiment using SSE, AFTER30, and MID30 based on lyrics features because text segments based on these structures are meaningless.

TABLE V
Comparison of results for different audio features

	Fea.	Verse	Chorus	SSE	AFTER30	MID30	All
P	Dynamics	0.356	0.192	0.001	0.308	0.093	0.196
	Rhythm	0.511	0.398	0.400	0.472	0.431	0.429
	Spec.	**0.582**	0.517	0.466	0.529	0.493	0.545
	Harmony	0.457	0.456	0.334	0.424	0.436	0.490
	All	**0.608**	0.562	0.477	0.541	0.527	0.582
A	Dynamics	0.323	0.291	0.171	0.293	0.217	0.332
	Rhythm	0.509	0.493	*0.524*	0.439	0.377	0.509
	Spec.	0.697	**0.798**	0.725	0.621	0.679	0.799
	Harmony	0.666	0.770	0.711	0.643	0.692	0.762
	Audio-all	0.713	**0.819**	0.755	0.646	0.694	0.822
D	Dynamics	0.292	0.193	0.076	0.263	0.102	0.249
	Rhythm	0.382	0.310	0.376	0.339	0.251	0.358
	Spec.	0.537	**0.644**	0.556	0.517	0.536	0.630
	Harmony	0.530	0.634	0.565	0.535	0.552	0.637
	Audio-all	0.545	**0.653**	0.566	0.549	0.537	0.648

2) Regression using different audio features based on different excerpts : Results for regressors built on different kinds of audio features is shown in TABLE V. Firstly, structure information is helpful for the regressors. For pleasure and dominance, best performance is achieved when using only verse and chorus sections. For arousal, it is better to use the whole song. SSE, AFTER30 and MID30 do not perform as verse, chorus or the whole song. Secondly, there are dominant audio features. Although regressors built on the concatenated audio features perform best, they are not significantly better than the spectral features alone. Music is essentially a sequence of note with different frequencies, which is best summarized by the spectral features. Dynamics features contain too few information to build effective regressors. Finally, we get an unexpected result. We can get the best result using SSE segment when building regressor for arousal based on rhythm features. It could be because that the segments with strongest energy have regular loud beat, which makes it easy to extract rhythm features accurately to build better regressor.

3) Regression using different MIDI features based on different excerpts : Results for regressors built on different kinds of midi features is shown in TABLE VI. The single char in the features column is the capital letter of the feature type. For example, 'I' means that the regressor is built with instrument-related features. And string means concatenated features. For example, 'DPM' means that features concatenated from dynamics, pitches and melody are used.

Firstly, for single type of features and merged features, best performance can be achieved using only verse or chorus

TABLE VI
COMPARISON OF RESULTS FOR DIFFERENT MIDI FEATURES

	Fea.	Verse	Chorus	SSE	AFTER30	MID30	All
P	I	0.093	-0.060	-0.011	0.050	-0.042	-0.030
	T	0.000	0.000	0.000	0.000	0.007	0.000
	D	0.191	0.212	0.210	0.273	0.170	0.164
	P	0.249	0.189	0.064	0.416	0.238	0.051
	M	**0.453**	0.399	0.291	0.463	0.373	0.410
	R	0.057	0.115	0.046	0.236	0.024	-0.028
	DPM	0.494	0.447	0.329	0.478	0.408	0.426
	DPMR	**0.498**	0.469	0.333	0.472	0.406	0.408
	All	0.497	0.468	0.332	0.476	0.404	0.417
A	I	-0.061	-0.070	-0.056	-0.033	-0.049	-0.052
	T	0.026	0.026	0.026	0.026	0.032	0.026
	D	0.318	0.439	0.369	0.362	0.339	0.341
	P	0.548	0.578	0.548	0.574	0.581	0.428
	M	0.611	**0.713**	0.637	0.564	0.629	0.677
	R	0.179	0.229	0.340	0.401	0.369	0.105
	DPM	0.629	0.744	0.664	0.590	0.664	0.685
	DPMR	0.638	0.759	0.674	0.594	0.676	0.688
	All	0.641	**0.761**	0.674	0.594	0.677	0.692
D	I	-0.089	-0.034	-0.048	-0.075	-0.037	0.016
	T	-0.001	-0.001	-0.001	-0.001	-0.028	-0.001
	D	0.238	0.331	0.323	0.334	0.268	0.266
	P	0.396	0.411	0.361	0.462	0.443	0.358
	M	0.479	**0.573**	0.514	0.448	0.506	0.528
	R	0.041	0.153	0.279	0.312	0.223	0.065
	DPM	0.490	0.587	0.533	0.467	0.528	0.563
	DPMR	0.482	**0.592**	0.538	0.469	0.536	0.568
	All	0.481	0.592	0.533	0.469	0.537	0.568

part. When we compare different features, we can come to the following conclusion: Merged feature are better than single feature alone. Pitch features are similar with audio spectral features, but they are not the best single features. Melody features are the best single features on three emotion dimensions. Instrument and texture features are meaningless because the instrument and track information are lost when transcript music into midi using WIDI.

C. Difference between verse and chorus

Song would be boring if all sections are similar. The chorus is the core of a song, and it should be different from verse in order to be emphasized. We get the feature difference between verse and chorus (VCD) by subtracting the corresponding element in chorus features vector and the element in verse's. Regression results for VCD features are shown in TABLE VII. We can find that the VCD features are helpful for building pleasure regressor using MIDI features. It is not helpful for the other situations.

TABLE VII
VERSE CHORUS DIFFERENCE

	Emo.	VCD	All	VCD+All
Audio	P	0.478	0.5822	0.5875
	A	0.4517	0.8224	0.8181
	D	0.4068	0.6484	0.6344
MIDI	P	0.3081	0.4169	**0.4673**
	A	0.2638	0.6918	0.6886
	D	0.2269	0.568	0.563

VI. CONCLUSIONS

In this work, we import structure information to help improve music emotion regression. By building regressor using different kinds of features for different emotion dimensions, we can find that verse and chorus are significantly helpful for emotion regressors. Besides, by importing the difference between verse and chorus, result for pleasure regressor built with midi features is improved.

ACKNOWLEDGMENT

Project supported by the Natural Science Foundation of China (Multi- modal Music Emotion Recognition technology research No.61170167) & Beijing Natural Science Foundation (Multimodal Chinese song emotion recognition No.4112028)

REFERENCES

[1] Mehrabian, A.: Framework for A Comprehensive Description and Measurement of Motional States. Genetic, Social, and General Psychology Monographs, vol. 121, pp. 33–361(1995)

[2] LEMAN, M., VERMEULEN, V., VOOGDT, L. D., MOELANTS, D., AND LESAFFRE, M. 2005. Prediction of musical affect using a combination of acoustic structural cues. J. New Music Res. 34, 1, 39–67.

[3] CHENG, H.-T., YANG, Y.-H., LIN, Y.-C., AND CHEN, H.-H. 2009. Multimodal structure segmentation and analysis of music using audio and textual information. In Proceedings of the IEEE International Symposium on Circuits and Systems. 1677–1680.

[4] HU, X., DOWNIE, J. S., LAURIER, C., BAY, M., AND EHMANN, A. F. 2008. The 2007 MIREX audio mood clas- sification task: Lessons learned. In Proceedings of the International Conference on Music Information Retrieval. 462–467.

[5] YANG, Y.-H., SU, Y.-F., LIN, Y.-C., AND CHEN, H. H. 2007. Music emotion recognition: The role of indi- viduality. In Proceedings of the ACM International Workshop on Human-Centered Multimedia. 13–21. http://mpac.ee.ntu.edu.tw/? yihsuan/MER/hcm07/.

[6] Yi-Hsuan Yang, Xiao Hu: Cross-cultural Music Mood Classification: A Comparison on English and Chinese Songs. ISMIR 2012: 19-24

[7] Yajie Hu, Xiaoou Chen, Deshun Yang: Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method. ISMIR 2009: 123-128

[8] Di Guan, Xiaoou Chen and Deshun Yang: Music Emotion Regression Based on Multi-modal Features. CMMR 2012: 70-77.

[9] KORHONEN, M. D., CLAUSI, D. A., AND JERNIGAN, M. E. 2006. Modeling emotional content of music using system identification. IEEE Trans. Syst. Man Cyber. 36, 3, 588–599.

[10] SCHMIDT, E. M., TURNBULL, D., AND KIM, Y. E. 2010. Feature selection for content-based, time-varying musical emotion regression. In Proceedings of the ACM International Conference on Multimedia Information Retrieval. 267–274.

[11] Namunu C. Maddage, Changsheng Xu, Mohan S. Kankanhalli, and Xi Shao. 2004. Content-based music structure analysis with applications to music semantics understanding. In Proceedings of the 12th annual ACM international conference on Multimedia (MULTIMEDIA '04). ACM, New York, NY, USA, 112-119.

[12] Carlos Nascimento Silla Jr., Alessandro L. Koerich, Celso A. A. Kaestner: Improving automatic music genre classification with hybrid content-based feature vectors. SAC 2010: 1702-1707.

[13] Yi-Hsuan Yang and Homer H. Chen. 2012. Machine Recognition of Music Emotion: A Review. ACM Trans. Intell. Syst. Technol. 3, 3, Article 40 (May 2012), 30 pages.

[14] http://en.wikipedia.org/wiki/Song_structure_(popular_music)

[15] O. Lartillot and P. Toiviainen. MIR in matlab (II): A toolbox for musical feature extraction from audio. In Proceedings of 5th International Conference on Music Information Retrieval, 2007.

[16] Yading Song, Simon Dixon, Marcus Pearce: EVALUATION OF MUSICAL FEATURES FOR EMOTION CLASSIFICATION. ISMIR 2012: 523-528.

[17] Mckay, Cory. Automatic genre classification of MIDI recordings. M.A. Thesis. McGill University, 2004.