ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 164

International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)

Music Emotion Recognition Using a Variant of Recurrent Neural Network

Huaping Liu,Yong Fang* and Qinghua Huang

Shanghai Institute of Advanced Communication and Data Science,
Key laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai University, Shanghai, 200444, P. R. China
*Corresponding author

***Abstract*—Searching music by emotion has always been strongly needed by users. Since music streaming applications usually have tens millions of music pieces in database, it is impossible to label emotion tags for each music piece manually. It is desired that an intelligent algorithm can recognize emotion expressed by music automatically. Valence-Arousal model is a widely used for representing emotion, but the angle of vectors on V-A plane labeled by different raters usually varies greatly, which makes it difficult to train classification models. We are trying to introduce a label space defined by pairs of antonyms, which makes emotion label relatively objective. We also used a variant model of recurrent neural network in the paper, in this model, RNN is a mean to extract features from melody, and with other features calculated by normal machine learning algorithms, this model can make a good prediction of emotions.**

Keywords—music emotion; harmonics and percussive; chroma recurrent neural network

I. INTRODUCTION

Music emotion recognition (MER) is a young, but fast expanding field, stimulated by the interest music industry to improve automatic music categorization methods for large-scale online music collection [1]. In [2], an analysis of written music queries from creative professionals showed that 80% of the queries for production music contain emotional terms, making them one of the most salient and important components of exploratory music search. Many researchers suggest that emotion can be scaled and measured by a continuum of descriptors or simple multi-dimensional metrics. According to seminal work by Russell and Thayer [3], emotion labels may be organized into low-dimensional models. Most noted is the two dimensional Valence-Arousal (V-A) space [4-7], where emotions exist on a plane along independent axes of arousal (intensity), and valence (an appraisal of polarity). But the truth is, value labeled on respective varies greatly by different raters.

In year 2013 and 2015, LSTM-RNN based music emotion recognition have got the best both for arousal and valence performance. Such In[8], presents a method consists of deep Long-Short Term Memory Recurrent Neural Networks (LSTM-RNN) for dynamic Arousal and Valence regression, using acoustic and psychoacoustic features extracted from the songs that have been previously proven as effective for emotion prediction in music. In[9], considering the high context correlation among the music feature sequence, and study several multi-scale approaches at different levels, including acoustic feature with Deep Brief Networks (DBNs) followed a modified Auto Encoder (AE), bi-directional Long-Short Term Memory Recurrent Neural Networks (BLSTM-RNNS) based multi-scale regression fusion with Extreme Learning Machine (ELM), and hierarchical prediction with Support Vector Regression (SVR). In recent years, many studies are based on deep learning and V-A model.

This paper is organized as follows: Section II introduces the valence (positive or negation emotions expressed in music) and arousal (energy of the music) two dimension emotion model and it's disadvantage in the practical application, Section III describes the optimization method and constructs the new emotion model, Section IV evaluates the optimization performance. Section V concludes this paper.

II. VALENCE-AROUSAL EMOTION MODEL

Take DEAM Dataset as an example, the average standard deviation on arousal for each song by different raters reaches 1.4667 out of 10 and the average standard deviation on valence each song by different raters reaches 1.5219 out of 10. This results in the great variation in the angle represented in V-A space, the standard deviation of angle can be as large as 21.38 degree for a single song, which is about half the distance from 'miserable' to 'angry'. It is reasonable to have such variance in angle empirically, because there is a strong tendency that music can invoke more than one emotions at the same time. According to the data set, it is very likely that one feels delighted and excited at the same time when listening to a certain piece of music. While some music can be miserable to some people and angry to some others. So emotions is fuzzy by itself, using an accurate angle on a V-A Space might not be a good choice for supervised leaning.

In this paper, we used label space defined by pairs of antonyms. The assumption that each emotion has its antonyms on the V-A space has been well examined by Russell and Thayer. So for each emotion we concern, we can always find a pair of antonymous emotions on the V-A space as label space. Emotion vector rated by raters can always be categorized into the label emotion ($\vec{Label}_e$) or its antonymous ($\vec{Label}_a$) by the value it dot multiplies with the label emotion vector.

$$Label = Mean(Sgn(\vec{Label}_e \cdot \vec{Label}_a)) \quad (1)$$

And the ratio of labels on each side ($\vec{Label}_e$ and $\vec{Label}_a$) makes the fusion of emotion a certain set of label space. For instance, if 7 out of 10 raters feel more happy than miserable and 3 out of 10 raters feel more miserable than happy, such music piece can be labeled 70% happy and 30% miserable.

III. NEW MODEL CONSTRUCTION

It is revealed by correlation that signature of arousal depends highly on four features, energy of harmonic sound, energy of percussive sound, tempo of harmonic sound, tempo of percussive sound. Music with high tempo and loud percussive sound is surely to have high arousal and music with low tempo and low percussive sound is more likely to be less in arousal. Valence has a relatively low correlation with these four features, but depends more on features extracted by RNN.

The new model construction is depicted Fig1.

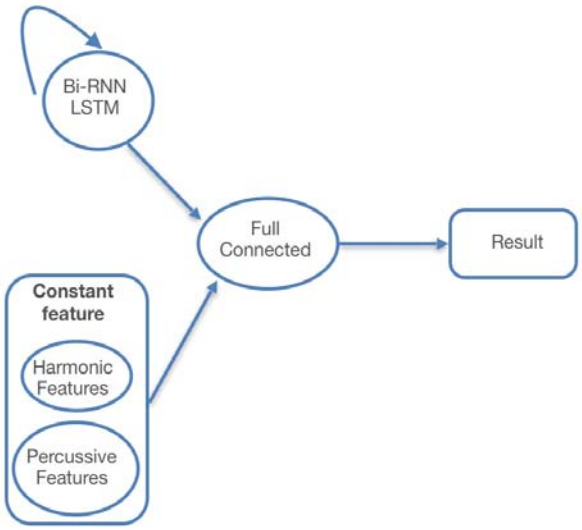

FIGURE I. NEW MODEL CONSTRUCTION

Bi-RNN (LSTM) is a two-way recurrent neural network with long short term memory capability. The input of the training is the chroma spectrum of the music for 10 seconds. Constant feature divides a piece of music into two parts, harmonic and percussive, calculates beats and energy of the two parts, and connects them with the last layer of Bi-RNN; Full connected is a two layer full link layer, and then output is classified information. In the output of classification, we regard the opposite two emotions as two separate categories, the opposite two categories as a dimension, and then we determine the tendency on this dimension by the percentage of the two opposite emotions in fusion output.

A. Harmonic Percussive Source Separation(HPSS) by Median Filters

Median filter is applied in this paper to separate harmonic sound and percussive sound from a music sound source. This method is first posted by Derry Fitzgerald [10]. First of all, short-time Fourier transform is applied to generate spectrogram of the music piece. Harmonic sound can be proximately regarded as horizontal lines on spectrogram while percussive sound can be regarded as vertical lines. So the HPSS problem is transferred to using a Non-negative Matrix Factorization (NMF) to split the spectrogram into horizontal and vertical lines. More specifically, the NMF is a binary mask to differentiate harmonic and percussive. For each element in the $N \times M$ matrix of spectrogram χ, mv is the median value of the K elements next to it along the vertical axis and mh is the median value of the K elements next to it along the horizontal axis. So the binary mask for harmonic on the $\chi(m,n)$ is 1 only if mh of $\chi(m,n)$ is greater than mv of $\chi(m,n)$. And binary mask for percussive is simply the negation of the mask for harmonic. With these median filters we can separate a spectrogram of a sound into spectrogram of harmonic sound and spectrogram of percussive sound.

In the paper invention calculates music mood by separating harmonic and percussive tones, and can accurately identify the arousal degree of music mood.

B. Music Tempo Estimation[11]

After HPSS, we should calculate tempo of harmonic sound and percussive sound respectively. Usually the two tempos are the same, but they are different in essence so we still treat them as different. In this paper a tempo-gram is used to estimate tempo of a piece of music. After short-time Fourier transform with windows, we first extract onset strength from it according to the energy change between frames. A tempo-gram is a time-pulse representation of an audio signal laid out such that it indicates the variation of onset strength over time, when given a specific time lag l or a BPM value τ. Since tempo is considered as constant in a given piece of music, we only need to calculate variation in long term temporal structure, auto-correlation-based tempo-gram is then used to pick the best BPM to illustrate music pieces tempo, and the BPM peak where has the closest mean local auto-correlation and global auto-correlation.

Get the harmonic and percussive of the fragment, and calculate tempo and energy of the two parts as static characteristics, and fed into constant feature.

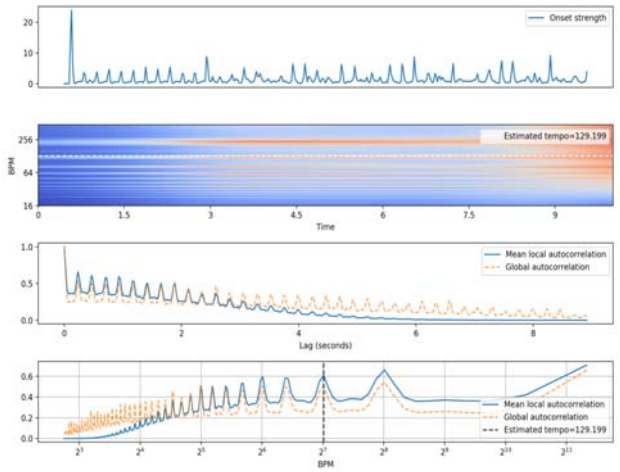

FIGURE II. TEMPO ESTIMATION

Fig2 shows the calculation process of the tempo parameter.

C. *Chroma Features*

For our RNN input representation, we investigate using Chroma features extracted from spectrogram. The main idea of chroma features is to aggregate all spectral information that relates to a given pitch class into a single coefficient [12]. Given a pitch-based log-frequency spectro-gram $y_{LF}: Z \times [0{:}127] \to R_{\geq 0}$, as defined in

$$y_{LF}(m,p) := \sum_{k \in P(p)} \left| \chi(m,k) \right|^2 \tag{2}$$

a chroma representation or chroma-gram Z × [0 : 11] → R≥0 can be derived by summing up

$$C(m,c) := \sum_{\{p \in [0:127] \mid p \bmod 12 = c\}} y_{LF}(m,p) \tag{3}$$

all pitch coefficients that belong to the same chroma for $c \in [0,11]$.

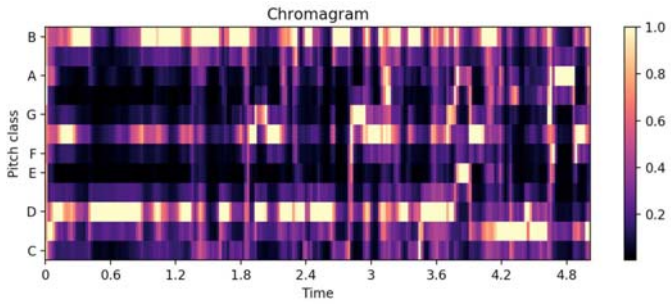

FIGURE III. CHROMA FEATURES OF A 10 SECOND MUSIC PIECE

The function of chroma is to get the melody of audio from the energy distribution in the spectrum. Each frame in the spectrum corresponds to a twelve-dimensional vector, corresponding to twelve tones of an octave degree.

Each frame of harmonic sound or percussive sound has a twelve-dimensional feature. Such as:

{0.10943639, 0.10766678,0.10823173, 0.14889599, 0.14798909, 0.0811433, 0.13909055, 0.44898109 1. , 0.64003491, 0.23333309, 0.14314128}

As the basic of tonality, chroma features are fed into Bi-RNN(LSTM) of the graph. This part is a recurrent neural network unit with long-term and short-term memory, and a verctor can be obtained.

IV. NEURAL NETWORK COMBINING RNN AND CONSTANT FEATURES

When a music piece is given, two types of features are extracted from the music piece. First is constant features, which are unrelated to time. Constant features are energy of harmonic sound, energy of percussive sound, tempo of harmonic sound, tempo of percussive sound. As mentioned above, the music piece is separated by HPSS algorithm into harmonic and percussive sounds, and using tempo estimation algorithms we can get tempo of respective sound. The energy over each STFT window can be calculated with the equation below. So there are four features in the constant features.

$$y(m,k) := \left| \chi(m,k) \right|^2 \tag{4}$$

And there are also dynamic features, which are chroma features, changing over time. RNN is applied to extract encoded features from the chroma sequence [14-16]. The RNN model is a Bi-RNN cell with a LSTM cell (forget bias = 1.0), each Bi-RNN cell has an input of 12-dimension vector and 512 hidden nodes. Each frame of chroma features are fed into the Bi-RNN cells and be encoded into a new feature vector.

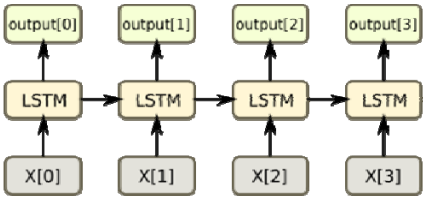

FIGURE IV. RNN WITH LSTM

This vector is then fed into a full connected neural network this full connected NN has two layers, each has 256 hidden nodes. And later connected to a soft max activation gate and use gradient descent optimizer to reduce cross entropy between prediction and ground truth.

V. EVALUATION OF EXPERIMENTAL RESULTS

The experiment is trained and tested on the DEAM dataset [1]. This dataset consists of 1802 excerpts and full songs annotated with valence and arousal values both continuously (per-second) and over the whole song. The whole dataset is

trained and tested on two label spaces: one is calm-excited space, and another is joy-sad space. It can be 81% accurate on predicting music excitement and 78% accurate predicting joyless. Although it might not be as intelligent as human brain, but we can modify this model by training model on different label spaces to better serve different user cases.

TABLE I. CONFUSION MATRIX ON EXCITED-CALM SPACE

	Excited	Calm
Excited	0.83	0.17
Calm	0.24	0.76

TABLE II. CONFUSION MATRIX ON JOY-SAD SPACE

	Joy	Sad
Joy	0.64	0.36
Sad	0.08	0.92

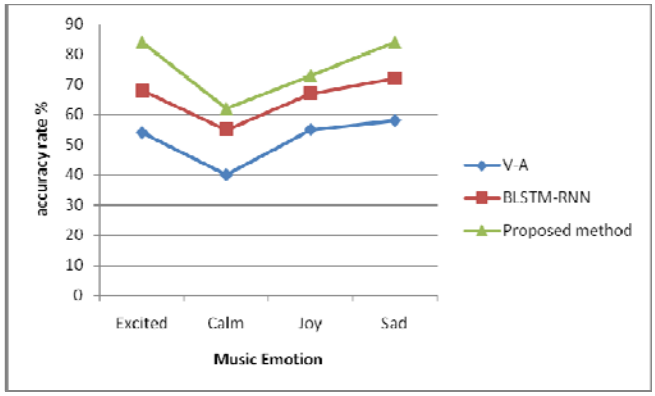

FIGURE V. SUBJECTIVE ASSESSMENT ACCURACY CURVE

Fig.5 is the subjective test result, have 253 volunteers, each person listens to 100 clips and marks the value of music in A-V space, only include four emotions: excited, calm, Joy and sad. The three algorithms are compared and tested, V-A[2], BLSTM-RNN[8] and the proposed method.

VI. CONCLUSION

Using fusion of antonymous can better describe emotions in the context of emotion recognition and it is easier to train models in supervised learning. Tempo and energy of harmonic and percussive sound in music pieces can be a very useful feature in emotion recognition, and with RNN encoding chroma, algorithm can be very intelligent telling emotion of music. This algorithm is still not able to understand implicit feelings in music, but it can be useful selecting music with strong emotions and recommend it to users.

ACKNOWLEDGMENT

The work was supported by the Key Support Projects of Shanghai Science and Technology Committee (16010500100).

REFERENCES

[1] Alijanaki A, Yang YH, Soleymani M. Developing a benchmark for emotional analysis of music. PLos ONE 12(3)| DOI:10.1371/ journal.pone.0173392, March 10, 2017

[2] Inskip C, Macfarlane A, Rafferty P. Towards the disintermediation of creative music search: analysing queries to determine import facets. International Journal on Digital Libraries. 2012; 12(2): 137-147.

[3] James A Russell, A Circumplex Model of Affect, University of British Columbia [J]. Journal of Personality and Social Psychology 39(6):1161-1178, Dec 1980.

[4] Yang YH, Lin YC, Su YF, Chen HH. A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech, and Language Processing. 2008; 16(2):448-457.

[5] Eerola T. Modeling emotion in music: Advances in conceptual, contextual and validity issues. In: Procddings of AES International Conference; 2014

[6] Barthet M, Fazekas G, Sandler M. Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based modles. In:Proceedings of International Symposium on Computer Music Modelling & Retrieval; 2012. p. 492-507.

[7] Hu X, Yang YH. Cross-dataset and cross-cultural music mood prediction: A case on Western and Chinese Pop Songs. IEEE Transaction on Affective Computing. 2016; PP(99)

[8] Mingxing Xu, Xinxing Li, Haishu Xianyu, Jiashen Tian, Fanhang Meng and Wenxiao Chen. Multi-scale Approaches to the MediaEval 2015 "Emotion in Music" Task. MediaEval 2015 Workshop, Sept.14-15

[9] Coutinho E, Trigeorgis G, Zafeirious S, Schuller B. Automatically estimation emotion in music with deep long-short term memory recurrent neural networks. In:Working Notes Proceeding of the MediaEval 2015 Workshop; 2015

[10] Derry FitzGrald, Harmonic/Percussive Separation Using Median Filtering. Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10), Graz, Austria, September 6-10, 2010

[11] Aggelos Gkiokas, Vassilis Katsouros, George Carayannis and Themos Stajylakis. Music Tempo Estimation and Beat Tracking by Appling Source Separation and Metrical Relation[C]. IEEE. Kyoto, Japan, 2012.

[12] Ringeval F, Schuller B, Valstar M, Cowie R, Pantic M. AVEC 2015: The 5th International Audio/Visual Emotion Challenge and Workshop. In: Proceedings of ACM International Conference on Multimedia; 2015. p. 1335–1336.

[13] Aljanaki A, Yang YH, Soleymani M. Emotion in Music task at MediaEval 2015. In: Working Notes Proceedings of the MediaEval 2015 Workshop; 2015.

[14] Aljanaki A, Wiering F, Veltkamp RC. Studying emotion induced by music through a crowdsourcing game. Information Processing and Management. 2016; 52(1):115–128.

[15] Wang JC, Yang YH, Wang HM, Jeng SK. Modeling the affective content of music with a Gaussian mixture model. IEEE Transactions on Affective Computing. 2015; 6(1):56–68.

[16] Chen YA, Wang JC, Yang YH, Chen HH. Linear regression-based adaptation of music emotion recognition models for personalization. In: Proceedings of IEEE International Conference Acoustics, Speech & Signal Processing; 2014. p. 2149–2153.