A Music Labeling Model Based on Traditional Chinese Music Characteristics for Emotional Regulation

Zhenghao He*
Tongji University
2954383503@qq.com

Ruifan Chen*
Tongji University
rfchenmusic@gmail.com

Yayue Hou
Tongji University
hanzhechun123@gmail.com

Fei Xie
Tongji University
xief@tongji.edu.cn

Xiaoliang Gong
Tongji University
gxllshsh@tongji.edu.cn

Anthony G Cohn
University of Leeds, Tongji University
a.g.cohn@leeds.ac.uk

ABSTRACT

The effectiveness of emotion regulation based on traditional Chinese music has been verified in clinical trials over thousands of years, but the reasons are unclear. This paper aims to use feature engineering to find effective music features which are effective for classifying different types of music and thus to try to provide an automatic recognition framework for building music libraries that can be used for mood regulation and music therapy. In this work, five modes (equivalent to the scales of Western music) of traditional Chinese music repertoire which can be used to regulate loneliness, anxiety, anger, joy, and fear are used. Features including Chroma, Mel-spectrogram, Tonnetz, and full feature vector features, are extracted for different length fragments of a piece of music which are then used to build a classification model for the five modes using a convolutional neural network (CNN). The results show that the highest 5-classes classification accuracy, 71.09%, is achieved from a Mel map of 5s music clips. A music mode labeling model is then constructed using a weighted combination of the different individual feature models. This model was then qualitatively evaluated on 13 pieces of music in different musical styles, and the results were reasonable from a music theory perspective. In future work, this music labeling model will be tested on more types of tracks to better assess its reliability.

CCS CONCEPTS

• **Computing methodologies** → Machine learning.

KEYWORDS

Mood regulation, music feature engineering, convolutional neural network, music labeling model

*These authors contributed to the work equally and should be regarded as co-first authors.

ICSCA 2023, February 23–25, 2023, Kuantan, Malaysia

ACM ISBN 978-1-4503-9858-9/23/02…$15.00
https://doi.org/10.1145/3587828.3587837

ACM Reference Format:
Zhenghao He, Ruifan Chen, Yayue Hou, Fei Xie, Xiaoliang Gong, and Anthony G Cohn. 2023. A Music Labeling Model Based on Traditional Chinese Music Characteristics for Emotional Regulation. In *2023 12th International Conference on Software and Computer Applications (ICSCA 2023), February 23–25, 2023, Kuantan, Malaysia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3587828.3587837

1 INTRODUCTION

In psychology, emotion is related to a particular physiological pattern, such as a positive or negative experience [1]. It is a physiological response transmitted via neural or hormonal or a combination of both [2]. If people are in a state of extreme emotion and this cannot be adjusted in time, whether the extreme emotions are positive or negative emotions, it may affect their mental health. These effects include but are not limited to the reduction of concentration and memory, the negative effect of decision-making, communication ability [3]. Previous research has shown music can play a positive role in improving the level of patients' physical state, relieving pressure and anxiety, and reducing pain [4]. In psychology, the process of using music to regulate patients' symptoms is called music therapy, whose essence is to introduce music-related experiences in the process of treating patients to meet the patient's physiological, emotional, and cognitive needs [5]. Many articles have proved the efficacy of music treatment, e.g. [6][7][8].

In recent years, the role of traditional Chinese music for emotional regulation role has been investigated. For example, Chen et al. [9] categorized some pieces of traditional Chinese music according to the data collected in clinical trials, and classified more than 60 traditional pieces of Chinese music according to their tone[1], including some pieces whose classification is controversial, and assigned the correct type label to each piece of music. Xu et al. [10] combined qualitative analysis and quantitative analysis with oral evaluation and physiological recordings, and divided 120 pieces of traditional Chinese music into five types: 'calm, fresh, happy, express and talk'.

However, most current research on emotion regulation in traditional Chinese music is based on clinical trials, rather than exploring the possible causes of its internal human body regulation from the music itself. In addition, these studies only examined the names of the tracks when classifying them, and did not compare different versions of the same music existing on current streaming platforms.

[1]There are five *tones* in traditional Chinese music (see column 1 of Table 1) which are equivalent to the do-re-mi-sol-la in western music, each of which is associated with its own scale. Different *modes* are formed by starting a scale at a different tone.

Table 1: The number, average length and proportion of the five tone tracks

Tone	Number	Average Duration(s)	Quantitative Proportion by number of tracks (%)	Quantitative proportion by total duration %)
Gong	17	418.14	27.87	30.00
Shang	7	487.31	11.48	14.39
Jue	12	393.04	19.67	19.90
Zhi	16	242.36	26.23	16.49
Yu	9	505.95	14.75	19.21

Zhang et al. [11] represented traditional Chinese musical pieces as textual patterns according to a certain pattern and utilized K-means clustering algorithm to classify the data. However, their data only consist of short fragments and they do not attempt to classify longer fragments. A team from Konkuk University of South Korea built a dual-format database for Korean traditional music and then used SVMs, decision trees, and random forests to classify the music pieces into emotional types and reached an accuracy of 0.883, 0.959, and 0.969 respectively [12]. Byte Dance's team extracted a piano scroll diagram from GiantMIDI-Piano, a transcription-based dataset and then used a convolutional neural network (CNN) and a recurrent neural network (RNN) to predict labels such as genre and composer achieving an accuracy of 0.739 for a coarse-grained 10 categorisation and 0.489 for a finer-grained 100 label set [13]. However, it is aimed at popular music and classical music: there is no study of traditional Chinese music.

In this paper, we use machine learning algorithms to categorize a corpus of processed traditional Chinese music clips based on feature vectors capturing key aspects of each clip. This enables us to explore which features are most effective in classification for human mood regulation. This music labelling classification model can be used to add further labelled pieces of music to the initial library thus giving a larger repertoire for emotion regulation.
In section 2 we describe the datasets and the main methods. Then in section 3 we give the results from our classifier and prediction algorithms before concluding in section 4.

2 METHODS

In sections 2.1 and 2.2 we overview the two datasets used in this paper. The first is used to train the classifier and only contains traditional Chinese music. The second which is used for testing predications made by the trained classifier contains Chinese music, but also Indian and Western music. Then in section 2.3 we describe the feature extraction methods used to represent the musical pieces in the databases. In section 2.4 describe the CNN used to build a model to classify music into one of the five tones. Section 2.5 gives the steps used to construct the system which can tag a new piece of music from the models built in section 2.4.

2.1 The Dataset for Training the Classifier

This work uses 61 traditional Chinese music tracks with five tones 'Gong, Shang, Jue, Zhi, and Yu' as shown in Table 1. These are the ones mentioned above in the work of Chen et al. [9], and the associated tones were obtained using clinical data. The above mentioned 10 controversial songs [9] were not included in the dataset.

Data sets with fragment lengths of 5s, 10s, 20s, and 30s of all tracks were constructed based on the original audio files to expand the size of the dataset. The numbers of the fragments in the different tone types are shown in Table 2. The fragments were created by calling the "librosa" [14] library to read the original audio, segment according to the target fragment length, and then store the new audio fragments. The 'AudioSegment' module in the 'pydub' library was then used to convert the fragments to mp3 format to reduce the storage space. Fragment cropping started with the head of the audio file and was performed continuously in chronological order, and segments at the tail of the audio file shorter than the target time were omitted.

2.2 The Dataset for the Prediction task

To evaluate the trained classifier, we first selected five traditional Chinese music pieces, namely 'Yang Whip Urging Horses to Transport Grain', 'Spring to Jingjiang', 'Wisdom Fighting', 'Day of Celebration', and 'Yellow Ying yin', all of which were played by the Huaxia Chinese Orchestra. In addition, in order to verify the generalization ability of the labeler, this article also selected four pieces of Western classical music and four pieces of Indian pop music as tests. The four Western classical music pieces are 'The Third Movement of the Sonata of Sorrow' composed by Beethoven, the 'First Movement of the Moonlight Sonata' composed by Beethoven, the 'Waltz No. 1' composed by Chopin and 'Clair de Lune' composed by Debussy. The four Indian pop songs came from the NMED-H dataset [15], 'Ainvayi Ainvayi', 'Daaru Desi', 'Haule Haule', and 'Malang'.

2.3 Feature Extraction

In this study, the Librosa library [14] was used to extract the frequency domain features and the time domain features of each track clip. Frequency domain features are mainly related to music features associated with tone and harmony pitch, while time domain features are mainly associated with rhythm features such as speed and rhythm. We use the Chroma, Mel-spectrogram, and Tonnetz features in the Librosa library as frequency domain features. To capture the music's rhythm, we select features such as music speed and note intensity distribution chart from the Librosa library.

The full set of features used in this work are: 'chroma_stft', 'chroma_cqt', 'chroma_cens', 'melspectrogram', 'mfcc', 'rms', 'spectral_centroid', 'spectral_bandwidth', 'spectral_contrast', 'spectral_flatness', 'spectral_rolloff', 'poly_features', 'tonnetz', 'zero_crossing_rate', and 'tempo'. The full feature vector is divided into three parts, which characterize the global characteristics of

Table 2: The numbers of the fragments in the different tone types [15]

Tone	Number of 5s fragments	Number of 10s fragments	Number of 20s fragments	Number of 30s fragments
Gong	1414	703	346	229
Shang	680	338	167	111
Jue	939	466	231	150
Zhi	776	363	187	123
Yu	905	451	223	148

the music, the melody, and the rhythm. Each part contains 45 eigenvalues. These eigenvalues are the mean, standard deviation, and variance values of two-dimensional signal features similar to Chroma plots.

2.4 Convolutional Neural Networks (CNN) Classifiers

The CNN classifier used in this work is the most commonly used deep learning model. For music segments, the temporal correlation of the features is strong which makes a CNN an appropriate technique. The data used for training are all the feature matrices of all the audio segments, whose abscissa is time. The constructer of the CNN has two convolutional layers composed of 64 and 128 cores (7 * 7) that extract the local features of the data, followed by two convolutional layers composed of 256 and 512 cores (3 * 3) to extract global features of the data. Since the training data is feature rich, all pooling layers in the model use average pooling to retain as many original features as possible. While in the experiments reported in section 3 below, only traditional Chinese music was used, in order to try to create a more general model able to handle other types of music in the future, several dropout layers with different discard rates were added at the end of the model to limit any overfitting to the training data. The data set is divided into a training set, verification set and test set in a ratio of 6:2:2. Separate CNN models were built for each of the feature types, and these are combined into a single prediction model as described below.

3 EXPERIMENTS AND DISCUSSION

3.1 Classification of the five music modes

We built four different convolutional neural networks to classify the dataset from §2.1 into the five tones, one for each of the feature types in the first column of Table 4. The parameters used for training the CNN are shown in table 3. In the five-classification experiment, the model corresponding to the full feature vector has a relatively stable performance, and the results are shown in Table 4. It is worth noting that in the CNN-based model training, the Mel map increased the training time by an order of magnitude due to its large data volume compared with the other feature types. Among the 16 feature-fragment size combinations, the Mel map on the 5s dataset reached an average accuracy of 71.09 ± 1.14% on the test set, which is higher than the correlation model of the full feature vector in this five-class classification experiment.

The full feature vector contains the most complete set of music theory features. Although it is only composed of the mean, variance, or standard deviation of each feature, it still performs the best across all the fragment lengths. The Chroma map and Mel map are both frequency domain features and performed next best. The overall performance of Tonnetz map was poor. This suggests that the melodic characteristics of music are more important than harmony information (such as the chords used). This may be the role of traditional Chinese music in emotion regulation since it focusses more on melody than harmonies.

We now consider how to combine these different models into a single model to predict a label for a new piece of music. This will be achieved by using a weighted combination of the best predictors from Table 4. When selecting these models, models with low accuracy on the test set during training were screened out. In the remaining models, the results determined as correct and the results determined as controversial are considered as correct results, and then the five models with the highest accuracy rate are selected to form the prediction module for music tagging. According to Table 4[2], the Chroma map has the lowest correct rate, and the other two features have the same accuracy, so the Chroma map has the lowest weight. Comparing the test set accuracy of 10s, 20s full feature vector and 10s Mel map, the two full feature vector test sets have the same accuracy, about 10% higher than the 10s Mel map, so the full feature vector has the highest weight, followed by the 10s Mel map. The final distribution is shown in Table 5. The exact weights were determined empirically but may not be a global optimum.

3.2 Music labeling prediction

Our music labeling model is composed of three steps.

Step 1. The preprocessing operations described in section 2.3 and used during the classification task are applied to the track whose label is to be predicted.

Step 2. The five models corresponding to each of the feature types shown in the first column of Table 5 are applied to yield individual predictions for the label of the piece.

Step 3. The weights from Table 5 are used to combine the scores step 2. The label with the highest score will be the output music mode label.

This model was tested on the 13 music tracks overviewed in section 2.2 and the prediction of the music mode labeler is shown in Table 6. For the five traditional Chinese music songs, the music labeler gives labels that meet the characteristics of the music. For example, the music 'Wisdom' is a solo, shasrp and bright tone, high melody[3], in line with the systematic prediction of 'Shang' music's

[2]In computing these figures, if a prediction partly accords with the ground truth, the partial credit is assigned pro-rata. E.g. If the result of classification is gong, and the ground truth is gong or shang, then a 0.5 credit is given.

[3]Note that the descriptions of each piece and the interpretation of the fit of these to the predictions was made by the authors.

Table 3: The parameters used for training the CNN

Feature type	batch_size	epochs	callbacks	optimizer	loss
Chroma	64	100	EarlyStopping(patience=20)	Adam (lr=0.0005)	categorical_crossentropy
Mel	64	100	EarlyStopping(patience=20)	Adam (lr=0.0005)	categorical_crossentropy
Tonnetz	64	100	EarlyStopping(patience=20)	Adam (lr=0.0005)	categorical_crossentropy
Full feature vector	64	1000	EarlyStopping(patience=200)	Adam (lr=0.0005)	categorical_crossentropy

Table 4: Results of the experiments classifying the dataset in §2.1 into the five tones using a convolutional neural network

Feature type	Fragment length(s)	Training set average accuracy(%)	Test set average accuracy(%)	Average training duration (s)
Chroma	5	98.43±2.83	60.95±1.91	1283.85
	10	99.00±1.26	59.36±2.15	973.92
	20	98.67±1.74	51.47±4.34	802.27
	30	95.96±2.46	46.54±3.71	586.71
Mel	5	98.59±0.79	71.09±1.14	20665.88
	10	91.23±8.84	53.20±4.30	13322.82
	20	89.31±12.61	40.26±3.98	11193.11
	30	77.02±13.96	36.73±3.69	17457.01
Tonnetz	5	78.40±14.52	48.97±2.88	1155.38
	10	84.93±18.20	47.31±4.73	1005.28
	20	73.82±11.68	39.09±4.17	687.65
	30	61.14±13.02	32.88±4.84	667.88
Full feature vector	5	95.93±2.28	68.21±1.28	1382.79
	10	90.23±6.73	65.18±2.03	615.16
	20	93.29±3.72	62.90±2.97	510.55
	30	80.00±8.91	58.89±4.73	377.37

Table 5: The selected weights and features for the music prediction model

Model feature type	**Weight(%)**
10s Full feature vector	30
20s Full feature vector	30
10s Mel	20
5s Chroma	10
10s Chroma	10

high characteristics. The piece 'celebrating day' is a festival music, bright rhythm, that conforms to the prediction of 'Gong' music's cheerful characteristics. For more complex music, our music tagging system can give labels for separate sections of the musical piece according to their characteristics. The tone of the initial and end sections of 'The whip urged the horse to carry grain and work hard' are cheerful, which corresponds to the label of 'Zhi' given by our music labeling model. However, the middle of the music is slow; though the flute tone is bright, the melody uses more techniques, and the melody is solemn, which fits with the solemn 'Gong' label given by the music labeling model.

For the four Western classical music tracks, the music tagging system also gave consistent labels. Although 'Pathos Sonata third movement' and 'Waltz No. 1' are lively, the melody goes more up the scale, and the music has a rich classical sense, so this music labeling model represents the high 'Gong' tone; 'Moonlight Sonata first movement' and 'Clair de lune' are both piano music depicting the moonlight, and the music is hazy, so the prediction model represents the soft 'Yu' tone.

For the four Indian pop songs, the music tagging system gave a 'Zhi' result. The music style is cheerful, while the four Indian pop songs are all biased to Indian dance music, which is more cheerful than the traditional Chinese music used for training, so the results given by the music label are reasonable and reliable. However, it should also consider the different effects of different dance styles on emotional regulation, that is, other Indian pop music may have a cheerful tone but different listening sense, which may be more suitable to regulate other types of emotions, but the label given by

Table 6: The prediction of the music mode labeler on the test data from §2.2

Music	Prediction Result
The whip urged the horse to carry grain and work hard	Gong/Zhi
Spring to Jing River	Gong/Jue
Wisdom	Shang
Pathos Sonata third movement	Zhi
Moonlight Sonata first movement	Yu
Waltz No. 1	Shang
Clair de lune	Yu
Ainvayi Ainvayi	Zhi
Daaru Desi	Zhi
Haulw Haule	Zhi
Malang	Zhi

our current system may not be appropriate. To sum up, in order to obtain a more robust and reliable labeler, the model within the labeler needs to be trained and strengthened with more data input; but for now, the results of the classifier seem relatively credible.

4 CONCLUSION

This paper exploits a music dataset whose tone labels were obtained clinically and demonstrates a method for predicting these via a machine learning model. Melodic features such as scales and modes are the most important musical features in the data studied in this paper, while harmonic and rhythmic features of the music are also important for the mood-regulating utility of the music.

We conducted experiments with a convolutional neural network in order to build a good classifier for mode classification by selecting the most useful features in the frequency and time domains. Using both machine learning metrics and musical theory, this paper evaluated and compared the models built using machine learning, and summarizes the degree of adaptation of the models corresponding to different feature types in different algorithms and different tasks. Among all the features, the full feature vector is the most adaptable, and its corresponding models have excellent and stable performance in different algorithms and different task contexts.

Finally, this paper explores the features that are suitable for characterizing the emotion regulation effects of traditional Chinese music in the context of music theory, and observes and compares these features to find the most suitable features for building a label generator for a music repertoire that has not yet been analyzed clinically for its type, and finally shows how this can be used to add new pieces not already having a label to a music library that can be used for emotion regulation.

Based on a traditional Chinese music library verified by clinical data, this work builds a relatively reliable music tagging system. However, due to the limitations of training data there is still room for improvement.

The amount of training data used in this paper is small, and the training data are all traditional Chinese music, and the music type is relatively simple. In subsequent studies, other existing data sets related to the emotion regulation of music could be introduced into the training data set to enrich the diversity of training data genres and styles.

The features used in this work are directly present in the music or can be computed after simple calculations, which can be extracted from the audio signal of low-level features. To improve the stability and generalization of the music labeling model, other mathematical feature extraction methods could be introduced to calculate features.

In such subsequent work, with the introduction of more algorithms, more feature types, and more data, the model type and weighting algorithm which integrates the individual model predictions could be optimized accordingly to obtain more applicable and accurate music tags.

ACKNOWLEDGMENTS

The 16th "Experimental Teaching Reform Fund" Project of Tongji University. Grant no: 0800104311.

Tongji University National Computer and Information Technology Experimental Teaching Demonstration Center Funding. Grant no: 0800120010.

Tongji University 2021 Curriculum Civic Science Project. Grant no: 4250104078/086.

REFERENCES

[1] Daniel L. Psychology Second Edition. M. New York: Worth Publishers, 2011: 310.
[2] Damasio A R. . Emotion in the perspective of an integrated nervous system. J. Brain research reviews, 1998, 26(2-3): 84.
[3] PR Goldin, K Mcrae , W Ramel, J J Gross. The neural bases of emotion regulation: reappraisal and suppression of negative emotion. J. Biological psychiatry, 2008, 63(6): 577.
[4] Ye Z. The Study of Music Therapy for Depression under the Background of Brain Science. J . Academic Journal of Humanities & Social Sciences, 2021, 4(2): 2.
[5] Stegemann, Thomas, Geretsegger, Monika, P Quoc, Eva, Riedl, Hannah, Smetana. Music therapy and other music-based interventions in pediatric health care: An overview. Medicines, 2019, 6(1), 25.
[6] M Geretsegger, C Elefant, KA Mössler, C Gold. Music therapy for people with autism spectrum disorder[EB/OL]. (2022-05-09)[2022-05-16]. https://www.cochranelibrary.com/cdsr/doi/10.1002/ 14651858.CD004381.pub4/full.
[7] Geipel, Josephine, Koenig, Julian, Hillecke, Thomas, K., Resch, Franz, Kaess, Michael. Music-based interventions to reduce internalizing symptoms in children and adolescents: A meta-analysis. Journal of affective disorders, 2018, 225: 647-656.
[8] Bieleninik Ł, Ghetti C, Gold C. Music therapy for preterm infants and their parents: A meta-analysis. J. Pediatrics, 2016, 138(3).
[9] Chen Yukun, Geng Shaohui, Li Jiangbo, Bao Yu, Zhang Qing, Yang Liping, Liu Shuning. Study of Music and Clinical Application of Depression . J. Chinese Journal of Traditional Chinese Medicine, 2019,34 (9): 4.
[10] Xu Rui, Yang Qiuli, Luo Yuejia "Research on the Emotional Regulation Database of Chinese National Music." Abstract of the 18th National Psychological Academic

Conference - Psychology and Social Development 2015
[11] Zhang Liumei, Jiang Fanzhi, Li Jiao, Ma Gang, Liu Tianshi. "K-means clustering analysis of Chinese traditional folk music based on midi music textualization." 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP) 2021.
[12] Lee J S, Lee M C, Jang D. Korean Traditional Music Genre Classification Using Sample and MIDI Phrases. J. KSII Transactions on Internet and Information Systems (TIIS), 2018, 12(4): 1869-1886.
[13] Kong Q, Choi K, Wang Y. Large-Scale MIDI-based Composer Classification. J. arXiv preprint arXiv:2010.14805, 2020.
[14] Brian McFee, Colin Raffel, Dawen Liang. librosa: Audio and Music Signal Analysis in Python.14th Python in Science Conference (SciPy 2015),2015
[15] Blair Kaneshiro, Duc T. Nguyen, Jacek P. Dmochowski, Anthony M. Norcia, and Jonathan Berger (2016). Naturalistic Music EEG Dataset - Hindi (NMED-H). Stanford Digital Repository. Available at: http://purl.stanford.edu/sd922db3535