International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

# A Review on Music Emotion Recognition

**Devangi Doshi**

Mukesh Patel School of Technology and Management Engineering, Mumbai, India

**Abstract:** ***Music is a form of art and cultural activity that has an ability to make us feel in a certain way. Music Emotion Recognition is an upcoming area of research in Music Information Retrieval community. The paper describes a short review on how emotion recognition can implemented. The collection of emotion datasets along with music datasets are used to determine an emotion for a particular music piece using certain Artificial Intelligence algorithms. Representation methods for music files and emotions have been described for algorithm implementation. A case study gives an insight of suitable algorithms.***

**Keywords:** music, emotions, artificial intelligence

## 1. Introduction

Music Information Retrieval (MIR) has increased research in automated systems for searching and organizing music related data due to vast and easily accessible digital music libraries. Although genre and artist based retrieval has a greater attention for MIR, Music Emotion Recognition (MER) is an upcoming field for the same. Computationally determining the music emotions requires knowledge of machine learning, neural networks, signal processing, and psychology and musictheory. In this paper Section II talks about Emotion Representation, Section III describes Audio Representation, Section IV tells us about the music features, Section V deals with the algorithms that can be used and Section VI provides a pathwayfor Music Emotion Recognition.

## 2. Emotion Representation

Music is often referred as language of emotions [3]. It is capable of conveying as well as inducing basic emotions particularly like happiness and sadness, also complex aesthetic emotions like nostalgia and wonder. Strong and positive emotional responses are evoked through music. Music can also elicit mixed emotions like simultaneous perceptions and feelings of both happiness and sadness, and positive evaluations of sad sounding emotions. Thus difficulties have been faced to model emotions.

The emotions datasets can be formed by annotating music clips [1]. Annotation games, surveys, lyrics are few techniques for collecting emotional content in music. Surveys can be done where subject is asked to annotate music clip from a set of adjectives/tags. Another approach is by annotation games like TagATune, MoodSwings, ListenGame, MajorMiner and HerdIt. Lyricator system provides an emotional score for a song based on its lyrical content. These types of datasets are helpful in modeling of emotions. Other ways of annotations include Web-Documents and Social Tags (which requires knowledge of text mining andnatural language processing).

Today, most MIR systems use either categorical psychometrics or dimensional psychometrics for emotion representation.

### 2.1 Categorical Psychometrics

Finding and organizing data into a set of emotional descriptors based on their relevance is defined as a categorical approach. The annual Music Information Research Evaluation exchange (MIREX), a community based framework for evaluating MIR systems and algorithms based on mood classification has categorized songs into five clusters, shown inTable1[1].

**Table 1**

| Clusters | MoodAdjectives |
|---|---|
| Cluster1 | passionate, rousing, confident, boisterous, rowdy |
| Cluster2 | rollicking, cheerful, fun, sweet, amiable/goodnatured |
| Cluster3 | literate, poignant, wistful, bittersweet, autumnal, brooding |
| Cluster4 | humorous, silly, crampy, quirky, whimsical, witty, wry |
| Cluster5 | aggressive, fiery, tense intense, volatile, visceral |

The Geneva Emotional Music Scale (GEMS) developed by *Zentner et.al* [6] is domain specific categorical emotional model. The GEMS scale consists of 45 labels consistently chosen for describing emotive states. They are organized in three-level hierarchy with middle level of 9 general categories like wonder, tenderness, transcendence, nostalgia, calmness, power, joy, tension and sadness. These are based on surveys where participants are choose terms that describe the induce emotions. The categoricalapproach is widely in MER.

### 2.2 Scalar/Dimensional Psychometrics

Mood can be represented as a point on a plane to discriminate emotions. Most widely used is a two-dimensional Valence-Arousal (V-A) space where arousal ranges from high to low and valence ranges from positive to negative.

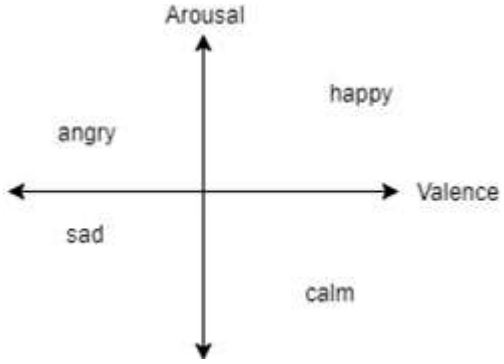


**Figure 1**

We can observe this from the Figure 1 given above. E.g. positive valence and negative arousal describes "calmness" whereas negative valence and positive arousal describes "angry".

## 3. Audio Representation

A key component is audio content representation for a powerful machine learning tools for classification and retrieval. A process of three-stages for compact song representation is as follows:

1) Short Time Frames: each song is processed to time series of low feature vectors.
2) Encoding: Each feature vector is encoded using a pre-calculated dictionary, a codebook of k "basis vectors".
3) Pooling: The coded frame vectors are pooled together to represent a music clip.

### 3.1 Short Time Frames

Spectral features are suggested to capture timbral qualities for low level feature vectors. Following steps are performed based on Mel Frequency Scale (MFS).

1) After converting into small frames, Discrete Fourier Transform is obtained for each frame
2) The loudness of signal is found to be logarithmic, hence logarithm of amplitude is calculated.
3) Mel Frequency bins is achieved to emphasize meaningful frequencies.
4) Decorrelation of components and reduction of feature dimension is achieved by Discrete Cosine Transform or Principle Component Analysis.

### 3.2 Encoding

It is observed that combination of short time frames along with pooling is not suitable for song representation,[2] hence, encoding is implemented. The Least Absolute Shrinkage and Selection Operator (LASSO) is an optimization criterion for linear regression that selects few regression coefficients to have effective magnitude while the rest are nullified or shrunk. This technique is also referred as "sparse coding".

The second encoding technique used is Vector Quantization (VQ): A continuous multi dimensional vector space is quantized to discrete finite set of bins. Each frame vector is quantized to the nearest codeword in codebook (Euclidean Distance is calculated) [3].

Another technique for encoding is Cosine Similarity (CS). In this technique, instead of calculating the Euclidean distance, dot product i.e. similarity is calculated between codeword and the feature vector.

Comparison of these techniques have concluded that VQ is more robust, havingsmooth and controlled change in parameter compared to LASSO and CS.

### 3.3 Pooling

Pooling relevant features over smaller segments and then aggregate them by averaging overall segments in a song is recommended than summarizing all short time features by computing statistical summaries as it dilutes their local discriminative characteristics.

These representation methods applied to low level features is recommended to represent various aspects of musical audio for MER.

## 4. Audio Features

A rich set of audio features based on temporal and spectral representation is extracted, in order to analyze music from audio content. MIR Toolbox is widely used for this extraction. In [5], two different types of features (mean and standard deviation) with total of 55 features are extracted and categorized into four perceptual dimension (classes) viz. Dynamics (loudness and quietness of music), Rhythm (beats), Harmony (consonance) and Spectral (timbral quality). An attempt is made to select the best individual class that provides best features however; a combination of classes has better results. E.g. a combination of rhythm, harmony and spectral have provided better results. These feature combinations is an input to an algorithm along with the emotion for emotion recognition.

## 5. Algorithms

### 5.1Neural Networks

Neural Networks has gained popularity in prediction and analysis domain of various applications like Natural Language Processing, Speech Recognition, Stock Market prediction and Image processing. Neural Networks uses data from various sources as an input which is passed into a processing unit where database is stored for analysis or prediction to produce an output for deducing certain probable outcome. Generally NN is of 3-layers where every neuron of each layer is connected to all the neurons of the next layer. If it were for emotion prediction, the input could be audio feature vectors and emotion vectors and the output would be the probability of a given audio to represent an emotion. The hidden layers will compute the probability for an emotion to correspond to an audio clip.

If $X=(X_1, X_2, X_3.....X_n)$ is a training example and $A=(A_1,A_2, A_3.....A_n)$ is an annotation vector, both X and A are an input to Neural Network, the hidden layers does the processing where similarity between $X_i$ and $A_i$ is determined. The output shows the highest probability of an emotion to represent a music clip.

### 5.2 Support Vector Machines

Support Vector Machine is a supervised learning algorithm used for classification and regression analysis. In this technique training samples are marked as belonging to one or the other category. SVM builds a model that assigns new example to one or the other category. It is a non-probabilistic binary classifier. SVM has a wide range of applications. Some of the examples are Spam Classifier, Image Classification and Bioinformatics.

If SVM is to be implemented then the tags/annotation and the audio content is an input. The SVM models similar audio clips within a tag i.e. examples of same categories form a cluster and a clear gap between different categories makes it easy for discrimination of emotion. When a new audio clip is an input to the classifier the model learns to fit in this audio clip into appropriate tag category by predicting which category they belong to.

## 6. Case Study

### 6.1 Using Semantic Embedding Recurrent Neural Networks

#### 6.1.1 Datasets

In this paper[3] ,the datasets selected by the author are tested upon two methods:

1) Regression: Training samples are annotated by vector of real numbers.
2) K-class classification: Each training sample is annotated by single label.

The first dataset uses GEMS model of emotions. Various music styles and artist were supplied by Magnatune. Emotify game was used to gather the annotations which were tagged by a respondent with at least 3 out of 9 mid level GEMS emotions. The annotation matrix is a 9-element vector in range [0,1] indicating percentage of tagged music with emotion. The music piece is 1 minute long, with 400 files in dataset, split evenly between four music genres: rock, pop, electronic and classical.

The second dataset, LastFM, contains music files split into anger, happy, sad and relaxed. The annotations are based on 30 most popular LastFM tags. 7digital.com is the origination of the music files. This dataset is jotted down into Valence-Arousal plane: (1,1) for happy, (1,-1) for relaxed, (-1,1)for angry and (-1,-1)for sad. The music files are 30 seconds long with 638 files in “angry” class, 752 in “happy”, 749 in “relaxed” and 763 in “sad”.

### 6.2 Terminologies

Feature Learning concept is used to create a data representation to pass it to a classifier or regressor for emotion detection. Feature learning is closely associated with neural networks where initial layer learns representation of data and passes it to final layer or classifier. Multi label embedding which has good results in multi label annotations inspires feature learning technique to transform feature vector into feature space where data points resemble their relations in annotation space. For this semi supervised embedding approach is used where loss function is calculated to incorporate annotated data. This learning process is to create feature space where training samples with same class are close to each other and with different classes are far away from each other.

Recurrent Neural Network along with Gating mechanism causes a more complex neural network where a recurrent layer is replaced by a “unit”, consisting of interconnected layers. The gating mechanism decides whether the output should pass through or not. Gated Recurrent Unit (GRU) is generally considered to be used.

### 6.3 Working

A set of n training samples $X=(X_1, X_2, X_3 \ldots..X_n)$ where every sample is $X=(x_{i1}, x_{i2}, x_{i3}\ldots...x_{ili})$ a series containing vectors of the same size. The annotation vector represents either real number for regression problem or k-1 zeros and a single one for k-class classification.

Let $H = (h_{i1}, h_{i2}, h_{i3} \ldots...h_{ili})$ was output of GRU resulting from input sequence X. A feature vector $f_i$ is defined by

$$f_i = \sum_{j=1}^{l_i} \frac{h_{ij}}{l_i}$$

Let F and A be the music representation data and annotation matrix where i-th row of matrixFcorrespond to i-th row of matrixA. Calculationof loss function was used to determine featurespace in which similarity between two musicvectors and corresponding annotations.This wasachieved by using cosine similarity.

$$cos(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

$$L(F, A) = \|F'F'^{\top} - A'A'^{\top}\|$$ Loss function can be defined as:

Where F’ and A’ are normalized vectors of F and A. This loss function helps in optimizing weights using gradient descent method. Learned features are an input to a separate classifier.

### 6.4 Experimental conditions

A 10-fold cross-validation was used for proposed approaches. Theano python library was used as it is efficient for GPU usage and easy gradient computation. The representation of music files is a spectrogram with Mel frequency scale (40 bins) and log-scale for power. MIR Toolbox was used to extract spectrogram using default parameters: frames 50ms long with 25ms overlap. 600 vectors of 160 size for 30-sec music file is result of input vector to RNN that contains two frames of spectrogram and deltas for these frames in each vector.

For classification and regression, SVM and SVR are used for which Radial Basis Function Kernel is used. The size of neural network is two GRU layers where one layer is 100 neurons and other is 50 neurons. For increased learning

speed and preventing overfitting in each epoch only a fragment of song is randomly selected.

### 6.5 Experiment Results

The results of Emotify game dataset shows that GRU+Semantic Embedding+SVR has better results. However it is observed that the harmonic features for emotion detections failswiththismethod[3].

The results of LastFM dataset is given below [3]:

**Table 2**

| | Accuracy | VectorSize |
|---|---|---|
| Spec+Dyn | 0.523 | 39 |
| Spec+Rhy | 0.523 | 37 |
| Spec+Dyn+Har | 0.54 | 47 |
| Semantic Embedding GRU+SVM | 0.542 | 50 |

As can be seen from Table 2 Semantic Embedding GRU+ SVMhascomparableresults.

## 7. Results

The method described in this paper has a better approach as compared to already existing methods by replacing standard music features with feature learning approach and using SVM classifier on learned features on two different datasets. The limitation of this paper is low depth of architecture and low time scales.

### 7.1 Using Hierarchical SVM Classifiers

#### 7.1.1Datasets

In this paper, a total of 219 classical music samples from two datasets are used where first dataset is of 270 music samples of 30 secs from 211 pieces of Western classical music were selected which are denoted by six graduate students to label emotions including happiness, tension, sadness and peace based on emotions perceived. The author selected a total of 175 music clips were picked with 49 for happy, 38 for tensional, 477 for peaceful and 41for sad.

In the second dataset two music therapist selected and annotated 60 music clips of 180 seconds out of which 45 samples were selected(13 for happy , 11 for tensional, 9 for peaceful and 12 for sad).

#### 7.1.2 Working

The following steps has been carried in [4] for MER:

##### 7.1.2.1 Data acquisition

Each music clip was converted to a standard recording format: mono channel PCM with sampling rate of 22,050 Hz and 16-bit resolution.

##### 7.1.2.2Feature Generation

In [4], 35 features were obtained from rhythm, dynamic. Pitch and timbre from converted music recording.

**1) Dynamics**

Following five dynamic features were generated:

x1: Mean_Loudness, the average Loudness foreach32-ms frame.
x2: Var_Loudness, the variance of Loudness foreach32-ms frame.

x3: Range_Loudness, the difference between the maximal and minimal Loudness for each 32-msframe.
x4: RMS_Loudness, the root mean square valueofLoudness foreach32-msframe.
x5: Low-energy_Rate, the percentage of

32-ms frames with less-than-average RMS_Loudness energy for each audio recording.

**2) RhythmFeatures**

Followingfeatureswereg enerated:
x6: Tempo, the average peaks of the Loudness perminutefor eachaudiorecording.
x7: Var_Rhythm, the variance of a note length foreachmusic audiorecording.
x8: Articulation, the average of all notes' attack time ratio for each music audio recording. The attack time ratio is defined as the ratio of attack time and note length
x9: Median_Slope, the median of all attack slopes for each music audio recording. The attack slope is defined as the slope from valley to peak at teach note.
x10: Max_Slope, the maximum attack slope foreachmusic audiorecording.

**3) Pitch Features**

Following features were obtained:
x11: Mean_Pitchsemitone, the average Pitch semitone for each music audio recording.
x12:Median_Pitch semitone,the median Pitch semitone for each music audio recording.
x13:Var_Pitch semitone,the variance of Pitch semitone for each music audio recording.
x14: Max_Pitchstep, the maximum Pitch step for each music audio recording.
x15: Min_Pitchstep, the minimum Pitch step for each music audio recording.
x16: Mean_Pitchstep, the average Pitch step for each music audio recording.
x17: Best_Modality, the most possible mode of each music recording in which a total of 12 possible modes including C, C♯, D, E ♭, E, F,F♯,G,G♯,A,B ♭,andB.
x18: In harmonicity, the degree to which the frequencies of partial tones depart from whole multiples of the fundamental frequency
x19: Roughness, the amount of partials that departs from multiples of the fundamental frequency.

**4) Timbre Features**

A total of 16 timbre relatedfeatureswereg enerated.
x20: Brightness_1500Hz, the percentage of energy above 1500Hz.
x21: Brightness_3000Hz, the percentage of energy above 3000Hz.
x22: Spectral_Rolloff, the frequency of the percentage of energy which is less than 15%. Another 13 features are generated based on Mel-frequency cepstrum (MFC) analysis which is a representation of the short term power spectrum of a sound signal on the Mel scale.

### 7.2 Feature Selection

The author used kernel based class separability for feature selection (KBCS).

### 7.3 Feature Extraction

The author employed the nonparametric weighted feature extraction (NWFE) for feature extraction. The idea of this method is to assign every sample with different weights and to define new nonparametric between-class and within-class scatter matrices. The goal of the NWFE method is to find a linear transformation which maximizes the between-class scatter and minimizes the within-classscatter.

### 7.4 Classifier Construction

The first classifier, node I, separates music samples into a group of low arousal and a group of high arousal. The second and third classifiers: node II and node III, are used to discriminate positive and negative valence from each level of arousal, respectively. With the hierarchical SVM classifiers, four music emotions viz. happy, tensional, peaceful and sad can be assigned to the corresponding quadrants of the 2D music emotion space. KBCS and NWFE methods were utilized at every node to find the best feature set for different separation targets at each node.

## 8. Experiment and Results

In this paper, the author employed a 5-fold cross validation to test the method and the accuracies of each SVM classifier. Node I of the trained hierarchical SVMs classifier was used to discriminate music samples with high and low arousal. The KBCS+NWFE methods were used to rank and reduce the normalized music features. At node I of the SVM classifier, the best average accuracy of dataset A was obtained when 15 features were used, and the best average accuracy of dataset B was reached when 5 features were used. Node II and node III are both responsible for classifying positive and negative valence. The same feature selection and extraction procedures were performed as those used in node I. The best average accuracy was achieved at 10 features for both node II and III in dataset A, and was achieved by 5 features for both node II and III in dataset B. The overall accuracy found is 86.94% in dataset A and 92.33% in dataset B for discriminating four music emotions based on the corresponding quadrants of the 2D emotion space.

## 9. Conclusion

MER has gained interest in past decades. It is still an ongoing research. From the case study I, a suggestion of increasing the number of neurons and additional layers along with the time scale of music file might provide better results. Based on the description of case study II, extracting more number features and increasing the number of music files shall be helpful for model analysis. A greater understanding of music moods and emotions can be achieved by collaborations of MIR researchers, psychologist and neuroscientists.

Future of MER might get extended into various fields of technology like Robotics, where robots might express emotions listening to music. Exploration of various approaches brings a hope ofprogress in this field.

## References

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, D. Turnbul, "Music emotion recognition: a state of the art review", Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), pp. 255-266, 2010.

[2] Y. Vaizman, B. McFee, and G. Lanckriet, ”Codebook based audio feature representation for music information retrieval,” IEEE/ACM Transactions on Acoustics, Speech and Signal Processing, vol.22,no.10,pp.1483-1493,2014.

[3] Jan Jakubik, Halina Kwasnicka,”Music EMotion Analysis using Semantic Embedding Recurrent Neural Networks” INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference.

[4] Wei-Chun Chang,Jeen Shing Wang and Yu-Liang Hsu, “A music Emotion Recognition with Hierarchical SVM based Classifier”,Computer, Consumer and Control (IS3C), 2014 International Symposium,published by IEEE

[5] Y. Song, S. Dixon, M. Pearce”Evaluation of Musical Features for Music Emotion Classification”, Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR), pp.523-528,2012.

[6] A. Aljanaki, F. Wiering, and R. Veltkamp, ”Computational modeling of induced emotion using GEMS,” Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR),pp.373-378,2014.

[7] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, ”A Regression Approach to Music Emotion Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, pp. 448-457,2008.