Automatic Music Mood Recognition using Support Vector Regression

Manisha Sarode
Department of Electronics and Telecommunication
JSPM's Rajarshi Shahu
College of Engineering,
Savitribai Phule Pune University,
Pune, Maharashtra, India

D. G. Bhalke, PhD
Department of Electronics and Telecommunication,
JSPM's Rajarshi Shahu
College of Engineering,
Savitribai Phule Pune University,
Pune, Maharashtra, India

ABSTRACT

Music is a dialect of feelings, and henceforth music feeling could be helpful in music understanding, proposal, recovery and some other music-related applications. Numerous issues for music feeling acknowledgment have been tended to by various teaches, for example, physiology, brain science, intellectual science and musicology. Music emotion regression is considered more appropriate than classification for music emotion retrieval, since it resolves some of the ambiguities of emotion classes. We present a music emotion recognition system based on support vector regression (SVR) method. The process of recognition consists of three steps: (i) Several music features have been extracted from music signal; (ii) those features have been mapped into various emotion categories on Thayer's two-dimensional emotion model; (iii) two regression functions have been trained using SVR and then arousal and valence values are predicted.

General Terms

Support vector regression (SVR), Thayer's two-dimensional emotion model, Regression theory, Arousal and Valence Modeling, Emotion Visualization, Timbral Features.

Keywords

Arousal, Valence, music emotion recognition (MER), support vector regression.

1. INTRODUCTION

Music assumes an imperative part in mankind's history, significantly all the more so in the computerized age. At no other time has such a substantial gathering of music been made and got to day by day by individuals. As the measure of substance keeps on blasting, the way music data is sorted out needs to develop with a specific end goal to take care of the always expanding demand for simple and viable data access. Music grouping and recovery by feeling is a conceivable methodology, for it is content-driven and practically effective. Feeling acknowledgment from music sign is a testing errand because of the accompanying reasons. To start with, feeling discernment is inherently subjective, and individuals can see diverse feelings for the same tune. This subjectivity issue makes the execution assessment of a MER framework on a very basic level troublesome in light of the fact that a typical concession to the characterization result is difficult to acquire. Second, it is difficult to portray feeling all around in light of the fact that the descriptive words used to depict feelings might be equivocal, and the utilization of descriptors for the same feeling can change from individual to individual. Third, it is still puzzling how music inspires feeling. What inborn component of music, assuming any, makes a particular passionate reaction in the audience is still a long way from surely knew. Numerous PC researchers have concentrated on music recovery by utilizing musical meta-information, (for example, title, kind or inclination) as well as low-level feature analysis (such as pitch, tempo or rhythm), while music psychologists have been interested in studying how music communicates emotion.

Currently, there is no standard technique to gauge and investigate feeling in music. Be that as it may, a mental model of feeling has discovered expanding use in computational studies. Thayer's two-dimensional emotion model in Fig.1 offers a basic yet entirely compelling model for putting feeling in a two-dimensional space. In the model, the amount of arousal and valence is measured along the vertical and horizontal axis, respectively. [11]

Conventional disposition and feeling research in music has concentrated on finding mental and physiological elements that impact feeling acknowledgment and grouping. Amid the 1980s, a few feeling models were proposed, which were to a great extent in view of the dimensional methodology for feeling rating. The dimensional methodology concentrates on distinguishing feelings in light of their area on a little number of measurements, for example, valence and activity. We attempt to build up a music emotion recognition system for predicting the arousal and valence of a song based on audio content.

In any case, even with the emotion plane, the downright scientific classification of emotion classes is still characteristically vague. Every feeling class speaks to a territory in the feeling plane, and the feeling states inside every region may differ a great deal.

For instance, the main quadrant of the emotion plane contains emotions, for example, energized, upbeat, and satisfied, which are distinctive in nature. This vagueness confounds the subjects in the subjective test and befuddles the clients while recovering a music piece as indicated by their emotion states. An option is to see the emotion plane as a ceaseless space and perceive every purpose of the plane as an emotion state. Along these lines, the uncertainty connected with emotion classes or descriptive words can be effectively dodged following no straight out classes are required.

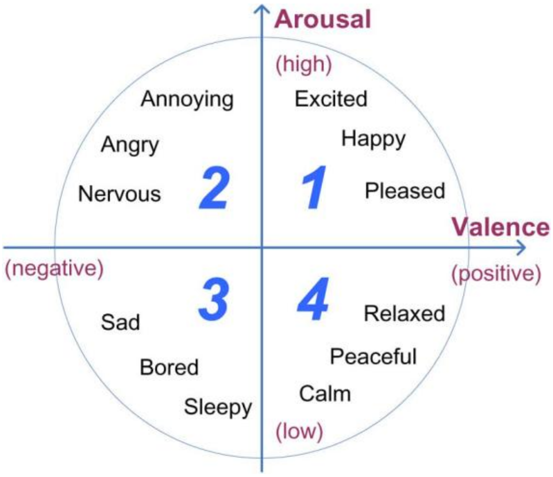


Fig.1 Thayer's arousal-valence emotion plane

2. RELATED WORK

In spite of a lot of exertion that has been made for MER as of late, little consideration has been paid to see the feeling plane from a continuous viewpoint. A few special cases can be found in the music emotion variation detection (MEVD) field [7], where the emotion substance of music is measured as a time-varying continuous variable, and some factual techniques are created to anticipate the emotion variety. Nonetheless, distinguishing the emotion variety is not the same as representing each song separately as a point in the emotion plane.

2.1 Arousal and Valence Modeling (AV Modeling)

To identify the emotion variety in video arrangements, AV demonstrating is proposed in [8] to figure the AV values. The arousal and valence models are weighted blends of some part capacities that are figured along the course of events.

The subsequent arousal and valence bends are joined to frame a full of affective curve, making it simple to follow the emotion variety of video substance and to recognize the portions with high emotional content. The segment capacities utilized for arousal are the motion vectors between back to back video outlines, the adjustments in shot lengths, and the energy of sound. Valence is demonstrated by the pitch of sound.

2.2 Regression Approach

Regression theory [16] is an all around considered hypothesis going for anticipating a genuine worth from observed variables (or features). No temporal data or geometric operation is required. Consequently, defining MER as a regression issue is by all accounts a promising methodology.

Since the AV qualities are seen upon as genuine qualities from the continuous perspective, the regression theory can be all around connected to straightforwardly anticipate arousal and valence.

To detail MER as regressor, the accompanying contemplations are considered. Equation(1) gives generic formula.

$$\mathcal{E} = 1/N \sum_{i=0}^{N} (y_i - \mathrm{R}(x_i))\text{\^{}}2 \qquad (1)$$

1) Domain of R: The Thayer's emotion plane is seen as a direction space crossed by arousal and valence, where every value is confined.

2) Ground truth: The ground truth is set by means of a subjective test by averaging the suppositions about the AV estimations of every music test.

3) Feature extraction: The extracted features should be applicable to emotion recognition for the regressor to be exact.

4) Regression calculation: Although regression theory has been all around concentrated on and numerous great regressors are promptly accessible, the execution of a regressor is case dependent.

5) Training fashion: There is a sure level of dependency amongst arousal and valence. In this way, we train A-V together.

3. SYSTEM DESCRIPTION

Proposed MER system represents each music selection as a point in the emotion plane and provides a friendly user interface for music retrieval and management. The system diagram is shown in Fig. 2.

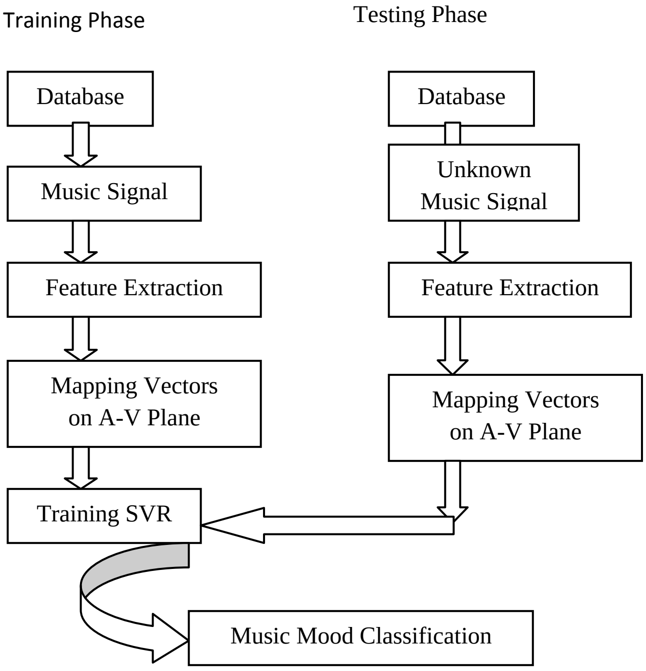


Fig.2 General block diagram of proposed work

The Block Diagram is divided into two phases, viz. testing phase and the training phase. The testing phase uses known music signal, from which the musical features are extracted and mapped on the A-V plane. The training is done using support vector regression approach and music emotion is classified.

In the test phase the system will classify the music emotion for unknown music signals. Based on the classification done for the testing dataset, the system will identify the closest vales and classify the emotion for each music sample.

3.1 Data Collection and Preprocessing

The music database is comprised of various prevalent melodies chosen from various Western and Indian collections. Two criteria are utilized as a part of the choice: 1) These tunes ought to be appropriated consistently in every quadrant of the feeling plane. 2) Each music test ought to express a specific predominant emotion.

To compare the fragments decently, the music tests are changed over to a uniform configuration (44100 Hz, 16 bits, and mono channel PCM WAV) and standardized to the same volume level. Also, since the feeling inside a music determination can change after some time, for every tune we physically select a 25-s fragment (generally the tune part) that is illustrative of the melody and communicates a specific overwhelming feeling [7]. Note we trim music physically since the execution of existing music thumb nailing calculations are considered not sufficiently strong [17].

3.2 Feature Extraction

After preprocessing, we extract musical features and construct a feature space. Extracted are 19 timbral features [Spectral centroid, spectral crest factor, spectral decrease, flatness, spectral flux, spectral kurtosis, spectral mfccs, spectral pitch chroma, spectral roll off, spectral skewness, spectral slope, spectral spread, spectral tonal power ratio, time acf coeff, time peak envelope, spectral predictivity ratio, time rms, time std, time zero crossing rate] Spectral centroid, spectral rolloff, and spectral flux describe spectral shape properties, zero-crossing measures the noisiness of the signal, and MFCC is a nonmusical pitch scale normally used in audio and speech signal processing.

Spectral contrast features capture the relative spectral information in each sub-band and utilize the spectral peak, spectral valley, and their dynamics as features. The spectral contrast features also roughly reflect the relative distribution of the harmonic and non-harmonic components in the spectrum.

3.3 Subjective Test

The ground truth of the AV values are set through the subjective test.

The following rules are followed for the subjective test.

1) Label the evoking emotion rather than the perceived one.
2) Express the general feelings in response to melody, lyrics, and singing (vocal) of the song.. We do not attempt to ignore the influences of the lyrics and singing even though the related features are not considered so far.
3) Music emotion perception is in nature subjective. Annotated are personal feelings.

3.4 Regressor Training

The inputs from feature extraction and subjective test are then used to train the following regression algorithm: support vector regression (SVR). SVR nonlinearly maps input feature vectors to a higher dimensional feature space by the kernel trick and yields prediction functions that are expanded on a subset of support vectors. As its name indicates, SVR is an extension of support vector classification, which has been found superior to existing machine learning methods in many cases.

3.5 Emotion Visualization

Connected with the AV values, every music sample is envisioned as a point in the emotion plane, and the comparability between music samples can be evaluated by computing the Euclidean separation in the emotion plane. A user interface that supports music retrieval/recommendation by determining a point in the emotion plane can be acknowledged without further naming the unseen music. Such an user interface can be of awesome use in managing vast scale music databases.

4. PERFORMANCE STUDY

We run a series of experiments to evaluate the performance of the regression approach. Different training and testing databases are created. We have taken 25 music samples for each emotion of happy, sad, angry, peaceful, bored, calm, sleepy, excited, relaxed, pleasant, nervous and annoying. A total training database consists of 300 music samples.

We have taken a 60-40 ratio for training and testing databases. A total of 120 music samples are taken for testing purpose.

4.1 Estimated Accuracy for different categories of emotions

Table.1 Accuracy results

Emotion	Accuracy
Angry	90%
Annoying	60%
Nervous	50%
Excited	90%
Happy	80%
Pleasant	80%
Peaceful	70%
Relaxed	60%
Calm	80%
Sad	80%
Bored	50%
Sleepy	60%

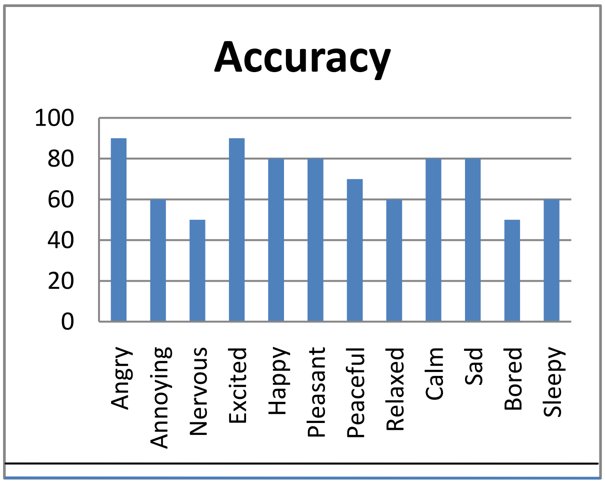


Fig3. Accuracy results in graphical representation

4.2 Estimated sample output of emotion classification

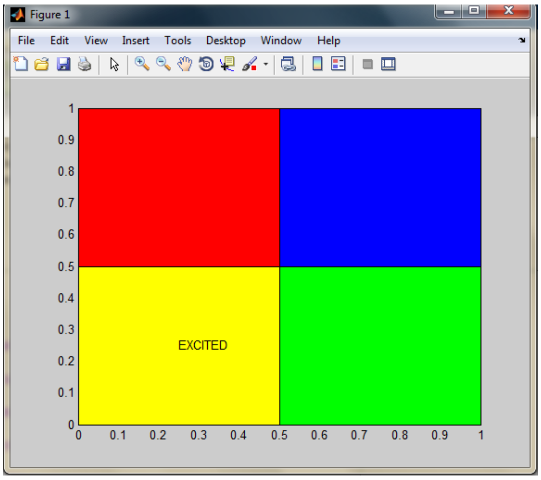
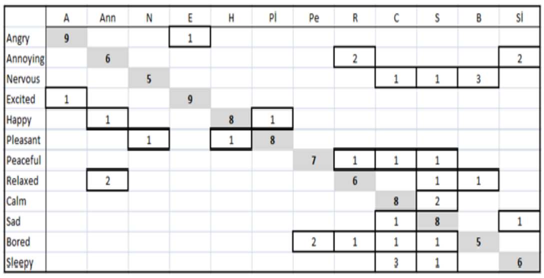


Fig.4 AV plane emotion recognition result

4.3 Results with Confusion Matrix



	A	Ann	N	E	H	Pl	Pe	R	C	S	B	Sl
Angry	9			1								
Annoying		6						2				2
Nervous			5						1	1	3	
Excited	1			9								
Happy		1			8	1						
Pleasant			1		1	8						
Peaceful							7	1	1	1		
Relaxed		2						6		1	1	
Calm									8	2		
Sad									1	8		1
Bored							2	1	1	1	5	
Sleepy									3	1		6

Fig5 Confusion Matrix

5. CONCLUSION

In proposed work, music selection is quantified as a point in the arousal-valence emotion plane. In addition, since there is more opportunity in depicting a melody, the subjectivity issue is eased to some degree. The exactness of the AV calculation decides the reasonability of the MER framework. We adopt the support vector regression for direct estimation of the AV values. Through an extensive performance study on the selection of data space, feature space and regressor we have demonstrated the effectiveness of the regression approach. The accuracy obtained for different emotions is- Happy- 80%, sad- 80%, angry- 90%, peaceful- 70%, Annoying- 60%, Nervous- 50%, Excited- 90%, Pleasant- 80%, Relaxed- 60%, Calm- 80%, Bored- 50% and Sleepy- 60%.

The overall accuracy of the proposed system is 75%, which is 25% more than the previously designed systems used for emotion recognition. We found the dominating features to be the spectral Centroid, spectral decrease, spectral flatness, spectral mfcc and time zero crossing rate. A user interface that supports music retrieval/recommendation by determining a point in the emotion plane is acknowledged without further naming the unseen music. Such an user interface can be of awesome use in managing vast scale music databases.

6. ACKNOWLEDGMENTS

We are also grateful to Dr. D.G.Bhalke for assistance with the techniques and to moderate this paper which in line improved the manuscript significantly.

7. REFERENCES

[1] Y. Feng, Y. Zhuang, and Y. Pan, “Popular music retrieval by detecting mood,” *Proc. ACM SIGIR*, pp. 375–376, 2003.

[2] T. Li and M. Ogihara, “Content-based music similarity search and emotion detection,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. 17–21.

[3] L. Lu, D. Liu, and H.-J. Zhang, “Automatic mood detection and tracking of music audio signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[4] Yi-Hsuan Yang; Yu-Ching Lin; Ya-Fan Su; Chen, H.H., "A Regression Approach to Music Emotion Recognition," in *Audio, Speech, and Language Processing, IEEE Transactions on*, vol.16, no.2, pp.448-457, Feb. 2008

[5] T.-L. Wu and S.-K. Jeng, “Extraction of segments of significant emotional expressions in music,” in *Proc. Int. Workshop Comput. Music Audio Technol.*, 2006, pp. 76–80.

[6] M.-Y. Wang, N.-Y. Zhang, and H.-C. Zhu, “User-adaptive music emotion recognition,” in *Proc. Int. Conf. Sig. Process.*, 2004, pp.1352–1355.

[7] Y.-H. Yang, C.-C. Liu, and H. H. Chen, “Music emotion classification: A fuzzy approach,” in *Proc. ACMMultimedia*, Santa Barbara, CA, 2006, pp. 81–84.

[8] A. Hanjalic and L.-Q. Xu, “Affective video content representation and modeling,” *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb.2005.

[9] M. D. Korhonen, D. A. Clausi, and M. E. Jernigan, “Modeling emotional content of music using system identification,” *IEEE Trans. Syst. Man., Cybern.*, vol. 36, no. 3, pp. 588–599, Jun. 2006.

[10] E. Schubert, “Measurement and time series analysis of emotion in music,” Ph.D. dissertation, School of Music Music Education, Univ. New South Wales, Sydney, NSW, Australia, 1999.

[11] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York: Oxford Univ. Press, 1989.

[12] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statist. Comput.*, pp. 199–222, 2004.

[13] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302,Jul. 2002.

[14] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, “Temporal feature integration for music genre classification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1654–1663, Jul. 2007.

[15] Hsu, Chih-Wei., and Lin, Chih-Jen: “A comparison of methods for multiclass Support vector machines,” *IEEE Trans. on Neural Networks*, Vol.13(2), pp.415-425, 2002.

[16] A. Sen and M. Srivastava, *Regression Analysis: Theory, Methods, And Applications*. New York: Springer, 1990.

[17] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, “Content-based music structure analysis with applications to music semantics understanding,” in *Proc. ACM Multimedia*, 2004, pp. 112–119