A Feature Survey for Emotion Classification of Western Popular Music

Scott Beveridge[1] and Don Knox[2]

[1] Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany
bevest@idmt.fraunhofer.de
[2] Glasgow Caledonian University, Cowcaddens Road, Glasgow, Scotland
d.knox@gcu.ac.uk

Abstract. In this paper we propose a feature set for emotion classification of Western popular music. We show that by surveying a range of common feature extraction methods, a set of five features can model emotion with good accuracy. To evaluate the system we implement an independent feature evaluation paradigm aimed at testing the property of generalizability; the ability of a machine learning algorithm to maintain good performance over different data sets.

Keywords: Music emotion classification, popular music, support vector machine

1 Introduction

Developing computational models of musical emotions is a multidisciplinary task including the fields of music psychology, musicology and computer science. Early research in this area focussed primarily on the classical music repertoire (1; 2; 3). From a musicological perspective this bias is easy to understand. Classical content provides well structured and well defined emotional ideas by means of motif, movement and form. In comparison, popular music tends to be more sonically and emotionally homogeneous owing perhaps to the commercial nature of the genre. Nevertheless, it is important that in 'real world' applications of emotion classification this type of content is taken into account.

The aim of this paper is to identify a subset of musical features that characterizes emotion in Western popular music. It achieves this by examining six of the most commonly occurring feature extraction toolboxes in the Music Emotion Classification (MEC) literature. To test the robustness of this approach we adopt a number of feature selection and classification algorithms in the context of an independent feature evaluation paradigm. The objective is to examine model generalizability, the property of a machine learning model that ensures good performance over multiple unrelated data sets. Evidence of generalizability supports the universality of the selected features.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London
All rights remain with the authors.

2 Feature Space

The feature spaces considered in this research are shown in Table 1. These algorithms extract low to mid-level acoustical and psychoacoustical features from the spectral representation of music clips. In all cases default parameters are used that include factors such as sample rate, bit depth, frame size and hop factor.

Toolbox	Number of features
MIRtoolbox	376
PsySound3	24
Marsyas 0.4	124
Marsyas 0.1	32
Sound Description Toolbox	187
Lu Implementation	71
All features	814

Table 1. Feature extraction toolboxes

Due to its demonstrated success in Music Emotion Recognition (MER) applications (4; 5) the MIRtoolbox for Matlab is used as an experimental baseline. The current version (1.3.4) provides a base set of 376 features derived from the statistics of frame-level features. PsySound3 creates features based on psychoacoustic models and is represented by 24 core features including those used in research by Yang *et al* (6). Marsyas, which was perhaps the first extraction framework to be developed for MIR, is implemented in two forms. The first version of the toolbox (0.1) contains a subset extractor which enjoyed success in early genre recognition tasks (7). The newest iteration of Marsyas (0.4) is also evaluated as it includes an extended feature extractor that is widely used in MIR. The Sound Description Toolbox has been included as it contains a number of MPEG-7 standard descriptors as well as perceptual and spectral features. Finally, the framework implemented in research by Lu *et al* (3) has been included. Although not publicly available this framework was recreated by the authors (8) due to its excellent performance in classical MER.

3 Emotion Space

Emotion concepts are defined on the basis of the circumplex model proposed by Russell (9). The circumplex model represents the emotion space with two orthogonal bipolar dimensions of *arousal* and *valence* (Figure 1). For the classification task the quadrants created by these dimensions are used as the target emotion concepts. These are anxious, exuberant, depressed and content.

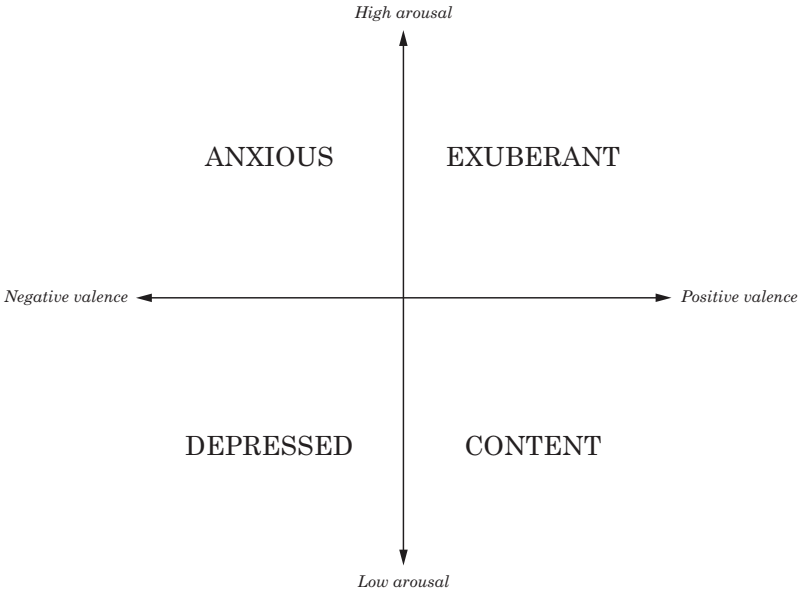

Fig. 1. The circumplex model

4 Musical Corpora

4.1 LastFM100

Two independent corpora are implemented in the classification framework. The first is derived from the LastFM database and consists of 25 tracks representing each quadrant of the circumplex model. These tracks were obtained by querying the LastFM database using the publicly available Application Programming Interface (API)[3]. This data acquisition technique has been used successfully in previous studies (10; 11) and is considered reliable due to its large number of active users.

4.2 Yang40

The second corpus was sourced from published research conduced by Yang *et al* (12). It contains 60 popular music tracks evaluated by 40 participants based on the dimensions arousal and valence. As this analysis requires discrete classes, each track was generalized into quadrants of the circumplex model determined by its arousal and valence values. Carrying out this process led to an uneven distribution across the emotion classes. To address this issue, random instances were removed from each class. This resulted in 10 instances or tracks per class.

[3] `www.last.fm/api`

5 Methodology

5.1 General Approach

A two-stage classification methodology was performed to determine the most representative feature set for Western popular music. First, initial models were constructed and evaluated in a traditional supervised train/test procedure. Using 10 x 10 fold stratified cross-validation these models were compared with respect to classification accuracy. In the second stage, the highest performing models were chosen for validation with the Yang40 corpus. As this data set is completely independent, this stage tests generalizability. High and consistent accuracy across data sets indicates high discriminative value of selected features. Each feature extraction toolbox was tested in isolation and then concatenated into a combined feature space named *Combined*.

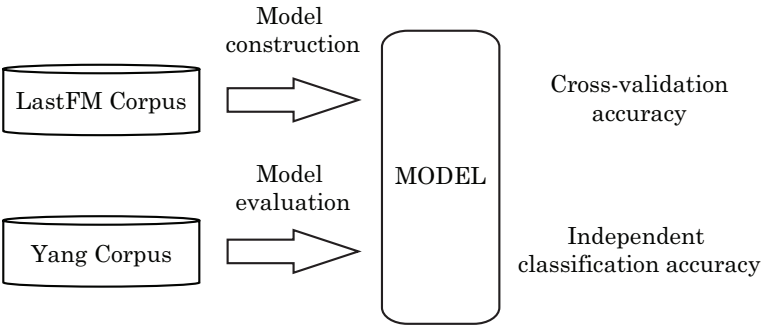

Fig. 2. Model training evaluation

5.2 Feature Selection and Classification

With such a high dimensional feature space it was necessary to apply feature reduction techniques. These included attribute subset and single attribute methods including InfoGainAttributeEval (InfoGain), CfsSubsetEval (Cfs), and ReliefFAttributeEval (ReliefF) (13). Based on the number of instances in the cross-validation data set and the need for parsimony, the 10 highest ranking features were chosen to represent the final feature space. Three classification models were also chosen for the analysis, K-Nearest Neighbours (K-NN), Naive Bayes and Support Vector Machines (SVM). A detailed explanation on the operation of these models is beyond the scope of this paper, however their inclusion was made based on good performance in previous emotion classification tasks (14; 15; 16; 17).

Performance of the feature selection/classification models are reported in terms of accuracy as defined by the Music Information Retrieval Evaluation eXchange (MIREX) in the 2007 Audio Music Mood Classification (AMC) task[4] (1).

$$\text{accuracy} = \frac{\text{number of correctly classified songs}}{\text{total number of songs}} \tag{1}$$

Modelling and feature selection were implemented in the WEKA environment, a freeware machine learning suite developed by the university of Waikato, New Zealand[5].

6 Results

6.1 Cross-validation

Table 2 shows classification accuracies for 10 runs of 10 fold cross-validation. For each feature extraction toolbox all feature selection/classification permutations are tested. The figures in bold show the highest performing models for each extraction toolbox including the concatenated *Combined* version.

These results show a clear boundary in performance between the Lu Implementation and *Combined* feature spaces and the rest of the feature toolboxes. The highest performing model for the *Combined* feature space has an accuracy of 0.64 with InfoGain feature selection and Naive Bayes classifier. The Lu Implementation is marginally higher with an accuracy of 0.65 using a combination of ReliefF feature selection and SVM classifier. The remaining toolboxes have accuracies ranging from 0.41 (Marsyas 0.1) to 0.48 (Sound Description Toolbox). This difference in performance is evident across all feature selection/classification approaches. As a result, the Lu Implementation modelled with SVM/ReliefF and *Combined* feature space modelled with Naive Bayes/InfoGain were chosen for independent validation with the Yang40 corpus.

6.2 Independent evaluation

The results of the independent evaluation step are shown in Table 3. With the 10 highest ranking features the Lu Implementation shows an accuracy of 0.65 and the *Combined* feature space 0.68. When expressed as percentages, this shows that 65 and 68% of the instances in the Yang40 corpus were correctly classified. As an additional measure the number of features used for classification were reduced to 5 and then 3. The aim was to determine how the steep drop in features might affect overall classification performance. Using only the top 5 ranked features, accuracies of 0.60 and 0.65 were achieved with the Lu Implementation and *Combined* feature spaces respectively. Using 3 features, classification accuracy dropped to 0.58 and 0.62. This small drop of between 6 and 7% shows the strong predictive power of these features.

[4] `http://www.music-ir.org/mirex/wiki/2007:Audio_Music_Mood_Classification`
[5] `http://www.cs.waikato.ac.nz/ml/weka/`

	InfoGain	ReliefF	Cfs
KNN			
MIR Toolbox	0.36	0.38	0.37
Marsyas 0.4	**0.45**	0.41	0.43
Marsyas 0.1	0.39	0.40	0.41
PsySound3	0.40	0.36	0.35
Sound Description Toolbox	0.39	0.41	0.44
Lu Implementation	0.60	0.62	0.63
Combined	0.63	0.64	0.57
Naive Bayes			
MIR Toolbox	0.41	**0.44**	0.40
Marsyas 0.4	0.40	0.40	0.43
Marsyas 0.1	**0.41**	0.41	0.40
PsySound3	0.41	0.42	0.40
Sound Description Toolbox	**0.48**	0.45	0.43
Lu Implementation	0.64	0.65	0.65
Combined	**0.64**	0.61	0.61
SVM			
MIR Toolbox	0.40	0.42	0.39
Marsyas 0.4	0.38	0.38	0.41
Marsyas 0.1	0.40	0.39	0.37
PsySound3	0.42	**0.43**	0.39
Sound Description Toolbox	0.44	0.45	0.43
Lu Implementation	0.62	**0.65**	0.62
Combined	0.62	0.59	0.58

Table 2. Model accuracies for 10 x 10 fold cross-validation

	Number of features		
Toolbox	10	5	3
Lu Implementation	0.65	0.60	0.58
Combined	0.68	0.65	0.62

Table 3. Classification Accuracy with 10, 5 and 3 features

7 Discussion

An important insight into the sonic properties of Western popular music is given in the reduced ranked feature set in Table 4. The two highest ranking features are statistics of frame-level values of spectral centroid. Indicating the 'centre of mass' of the spectrum, these features are a measure of high frequency content or brightness. Spectral flux, the third feature has been shown to be a useful perceptual indicator of music instrument timbre (18). The following two features relate to measures of intensity or loudness. These are Intensity ratio in sub band three as defined by Lu in (3), and sharpness, a perceptual measure of loudness relating to critical bandwidth. The sixth feature is a measure of tonal centre and is calculated as the mean of frame-wise centroid values of the chromagram. Spectral entropy is ranked seventh and gives an indication of the presence of predominant peaks in the signal. This is based on the Shannon entropy used in information theory (19). The next feature is Spectral Rolloff, an estimation of the amount of high frequency energy in a signal. The ninth ranked feature is Spectral Dissonance from the PsySound3 toolbox. Spectral Dissonance is a measure of the interference or *roughness* of spectral components. The final feature is the mean of the zerocross rate across frames. Zerocross is considered as a general measure of noisiness.

Overall, the ranking in Table 4 shows the importance of timbral characteristics in the recognition of emotion in popular music. The slight bias towards these features also suggests that modern production techniques, in particular over-compression, leads to homogeneity in terms of intensity or perceived loudness.

Rank	Feature name	Toolbox
1	Spectral Centroid Std	Lu Implementation
2	Spectral Centroid Variance	Lu Implementation
3	Spectral Flux Mean	Marsyas 0.1
4	Intensity Ratio Sub band 3 Mean	Lu Implementation
5	Sharpness Mean	PsySound3
6	Tonal Chromagram Centroid Mean	MIR Toolbox
7	Spectral Entropy Mean	MIR Toolbox
8	Spectral Rolloff Std	MIR Toolbox
9	Spectral Dissonance Std	PsySound3
10	Zerocross Mean	MIR Toolbox

Table 4. Ranked features from Combined feature space

8 Conclusions

By surveying six commonly used feature extraction toolboxes we present a compact feature subset for characterizing emotion in Western popular music. The efficacy of this feature space is tested using a combination of feature selection and classification algorithms in a unique feature evaluation paradigm. By examining model generalizability we have shown the potential universality of these features in an emotion classification task. The composition of the final feature set shows a bias towards spectrally derived acoustic features, reinforcing the idea that modern production techniques remove some important information carried by intensity or loudness features.

9 Acknowledgements

The research presented in this paper is part of the SyncGlobal project. SyncGlobal is a 2-year collaborative research project between Piranha Womex AG, Bach Technology GmbH, 4FreindsOnly AG and the Fraunhofer IDMT in Ilmenau, Germany. The project is co-financed by the Germany Ministry of Education and Research in the framework of the SME innovation program (FKZ 01/S11007).

Bibliography

[1] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*. Oxford University Press, 2001.

[2] D. Liu, N. Zhang, and H. Zhu, "Form and mood recognition of johann strauss's waltz centos," *The Chinese Journal of Electronics*, vol. 12, no. 4, pp. 587–593, 2003.

[3] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[4] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of the 10th International Society for Music Information Retrieval (ISMIR) Conference*, pp. 621–626, 2009.

[5] J. C. Wang, H. Y. Lo, S. K. Jeng, and H. M. Wang, "MIREX 2010: Audio classification using semantic transformations and classifier ensemble," tech. rep., Institute of Information Science, Academia Sinica, Taipei, Taiwan, 2010. Available from URL `http://www.music-ir.org/mirex/abstracts/2010/WLJW2.pdf`, accessed January 2011.

[6] 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.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[8] D. Knox, S. Beveridge, L. Mitchell, and R. A. R. MacDonald, "Acoustic analysis and mood classification of pain-relieving music," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1673–1682, 2011.

[9] J. Russell, "A circumplex model of emotions," *The Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[10] X. Hu, M. Bay, and J. S. Downie, "Creating a simplified music mood classification ground-truth set," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR07)*, pp. 309–310, 2007.

[11] Y. C. Lin, Y. H. Yang, H. H. Chen, I. B. Liao, and Y. C. Ho, "Exploiting genre for music emotion classification," in *IEEE International Conference on Multimedia and Expo, (ICME 2009)*, pp. 618–621, 2009.

[12] Y. H. Yang, Y. F. Su, Y. C. Lin, and H. H. Chen, "Music emotion recognition: the role of individuality," in *Proceedings of the international workshop on Human-centered multimedia*, pp. 13–22, 2007.

[13] E. Frank and I. H. Witten, *Data Mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 2005.

[14] G. Tzanetakis and P. Cook, *Manipulation, analysis and retrieval systems for audio signals*. PhD thesis, 2002.

[15] D. Turnbull, L. Barrington, and G. Lanckriet, "Modelling music and words using a multi-class naive bayes approach," in *Proceedings of the 7th International Society for Music Information Retrieval (ISMIR) Conference*, 2006.

[16] K. Bischoff, C. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach," in *Proceedings of the International Society for Music Information Retrieval Conference, Kobe, Japan*, 2009.

[17] R. P. Panda, Renato; Paiva, "Using support vector machines for automatic mood tracking in audio music," in *Audio Engineering Society Convention 130*, 2011.

[18] S. Le Groux and P. Verschure, "Emotional responses to the perceptual dimensions of timbre: A pilot study using physically informed sound synthesis," in *Proceedings of the 7th International Symposium on Computer Music Modeling*, 2010.

[19] C. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 1948.