Affective Feature Design and Predicting Continuous Affective Dimensions from Music

Naveen Kumar, Rahul Gupta, Tanaya Guha, Colin Vaz
Maarten Van Segbroeck, Jangwon Kim, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL)
University of Southern California, Los Angeles
http://sail.usc.edu

ABSTRACT

This paper presents affective features designed for music and develops a method to predict dynamic emotion ratings along the arousal and valence dimensions. We learn a model to predict continuous time emotion ratings based on combination of global and local features. This allows us to exploit information from both the scales to make a more robust prediction.

1. INTRODUCTION

The *MediaEval 2014 Emotion in Music* challenge consists of two tasks: desiging affective features, and predicting continuous emotion dimensions (arousal and valence) from music [1]. These tasks are important to our understanding of audio-based emotion prediction [2, 3] which finds applications in interaction modeling, automatic assessment of socio-emotional state of people.

2. AFFECTIVE FEATURE DESIGN

Designing features that can correlate well with human affective dimensions is critical. Below, we describe three new features designed to capture affect from music.

Compressibility features (comp): We hypothesize that stronger emotion like high arousal or high valence is evoked by complex interplay of various musical components, and therefore, the complexity of a music signal may be correlated with affect. We measure the complexity of a music signal by its compressibility i.e. how much the signal can be compressed. Intuitively, the more a given signal can be compressed, the lower is its complexity. This quantity is related to a theoretical measure of data complexity (Kolmogorov complexity) which is in general a non-computable quantity. In practice, Kolmogorov complexity is often approximated by the length of the compressed data.

We first convert each mp3 music file to raw audio format, and compress it using a lossless audio codec (FLAC). The compressed file length of the music file form the global compressibility feature (*global comp*). We also use this idea to create a dynamic feature (*dynamic comp*) where each 0.5 sec segment of the music file is compressed in the same manner, and a dynamic feature of compressibility is created.

Median Spectral Band Energy (MSBE) This feature is motivated by the observation that high arousal and valence songs often involve numerous instruments playing in tandem to crate the perception of a rich sound. This gives rise to a large spectral bandwidth. We propse to use the median spectral energy across bands as a robust metric to capture this effect. This feature is extracted at a global level, one value for each song.

Spectral Centre of Mass (SCOM) We also found the spectral center of mass to be typically low for lower arousal songs. This is again correlated with the fact that high arousal songs are usually more wideband. SCOM is also a static feature that computes a single value per song.

Table 1 below presents correlations between the proposed features and their corresponding static or dynamic emotion ratings.

3. PREDICTING CONTINUOUS EMOTION RATINGS

To predict continous arousal/ valence emotion ratings, we try to incorporate information from both the local and global scales. This is performed by extracting features at the global and local scales (every 0.5s). Each of these systems is discussed in detail below.

Frame Level Prediction We use dynamic features extracted at an interval of 0.5s for directly predicting continuous arousal and valence ratings for each song . The dynamic features sets openSMILE and dynamic complexity are extracted for this purpose.

We train separate linear regression models for arousal and valence over all frames in the training set which is used to directly make independent predictions for each frame in the test set. Finally, the resulting predictions are smoothed over time using a moving average filter to incorporate the smoothness expected of human annotations.

Predicting dynamic ratings using global features
In addition to frame level predictions, we also hypothesize that the dynamic ratings of each song are also affected by global factors of a song. Hence, we also try to predict the dynamic ratings, using features extracted over the entire 45 second clip of songs. To predict the dynamic ratings using static features, we first parametrize the ratings using a Haar transform. The haar coefficients for each song's ratings are then used as alternate labels for our models. This particular choice of label space was motivated by the smooth and sometimes piecewise constant nature of annotated ratings. In addition, it can also be seen from Fig. 3 that only a few

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Correlation results for the proposed features on the training and test sets

Static Features	Train		Test		Run#
	Arousal	Valence	Arousal	Valence	
global comp	0.656	0.394	0.529	0.278	1
MSBE	0.518	0.262	-	-	-
SCOM	0.579	0.353	-	-	-
Dynamic Features	Arousal	Valence	Arousal	Valence	Run#
dynamic comp	0.190	0.063	0.192	0.135	2

Table 2: Correlation and RMSE results for submitted predictions on the test set

Features	System	Arousal		Valence		Run#
		ρ	rmse	ρ	rmse	
openSMILE	baseline	0.18	0.15	0.11	0.12	-
openSMILE	Lin.Reg.	0.28 ($\tau = 6.5s$)	0.88	0.14 ($\tau = 20s$)	0.32	3
openSMILE	Lin.Reg. + Normalization	**0.28**	**0.13**	**0.14**	**0.10**	4
comp, SCOM, MSBE	PLSR on Haar coeff	0.22	0.12	0.11	0.09	5

of the coefficients are strongly correlated with the emotion ratings allowing for a sparse and robust representation.

We compute 64 Haar coefficients to encode the emotion dynamics over the length of each song. We learn a Partial-Least Squares Regression (PLSR) model to predict each of the haar coefficients (for both arousal and valence) as a label using global features such as static compressibility, sCOM and MSBE. The predicted haar coefficients are finally used to reconstruct back the dynamic emotion ratings via an inverse Haar transform. This method incorporates the temporal smoothness constraint within the algorithm by performing prediction in a label space where the ratings have an inherent sparse representation. More importantly, this system captures those aspects of emotion dynamics that are governed by global characteristics of the song.

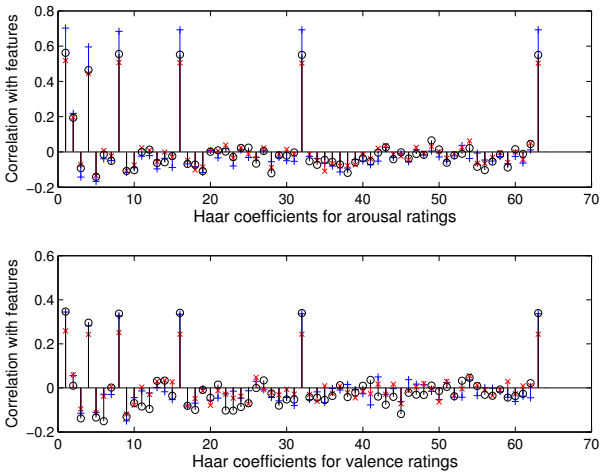

Figure 1: Each arousal valence rating is converted to 64 dimensional label space of haar-coefficients. Figure shows correlation of each haar-coefficient with the three static features (indicated by +,X,O).

4. RESULTS AND EVALUATION

We submitted 3 runs for each of our system to the challenge. Systems 1 and 2 used frame level prediction using linear regression models. We use an opensmile feature set comprising approximately 6000 features to predict arousal and valence ratings at an interval of 0.5 seconds. Given that the affective dimensions evolve smoothly over time, we incorporate context from neighboring frames. For each frame, we compute the unweighted average of affective dimension predictions over a window centered at that frame. The context window lengths τ for arousal and valence (system 1 submission) are 6.5 and 20 seconds long respectively. The window lengths reflect the fact that valence evolves slower as compared to arousal. For system 2, we additionally normalize the system 1 predictions for songs which have outlier predictions lying outside the range $[-1, 1]$. This helps reduce the RMSE. For system 3, we observe that smoothing doesn't help as much because the model inherently already ensures smoothness by choice of an inherently sparse label space.

Correlation and RMSE between predicted and annotated emotion ratings in reported for each tasks, averaged over songs is reported as an evaluation metric for all systems.

5. CONCLUSION

From our experiments, we observe that dynamic emotion ratings in a song depend not only on local characteristics of the music, but also on overall global features of a song. We also note that it helps to take into account context from adjacent frames. This is evident from the improved prediction results obtained by smoothing predictions using a moving average filter.

6. REFERENCES

[1] A. Aljanaki, Y. Yang, and M. Soleymani. Emotion in music task at mediaeval 2014. In *Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17*, 2014.

[2] D. Bone, C. Lee, and S. Narayanan. Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features.

[3] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *Affective Computing, IEEE Transactions on*, 3(2):184–198, 2012.