The TUM Approach to the MediaEval Music Emotion Task Using Generic Affective Audio Features

Felix Weninger, Florian Eyben
Machine Intelligence & Signal Processing Group,
MMK, Technische Universität München
80290 Munich, Germany
{weninger,eyben}@tum.de

Björn Schuller
Department of Computing
Imperial College London
London SW7 2AZ, UK
schuller@IEEE.org

ABSTRACT

This paper describes the TUM approach for the MediaEval Emotion in Music task which consists of non-prototypical music retrieved from the web, annotated by crowdsourcing. We use Support Vector Machines and BLSTM recurrent neural networks for static and dynamic arousal and valence regression. A generic set of acoustic features is used that has been proven effective for affect prediction across multiple domains. In the result, the best models explain 64 and 48 % of the annotations' variance for arousal and valence in the static case, and an average Kendall's tau with the songs' emotion contour of .18 and .12 is achieved in the dynamic case.

1. INTRODUCTION

The 2013 MediaEval 'Emotion in Music' task is to provide continuous valued arousal and valence estimates both for whole songs (static) and sequences of one second long segments (dynamic). For details on the task, we refer to the paper describing the task [1]. In the following we describe our approach.

2. METHOD

Our approach is based on supra-segmental features calculated by applying statistical functionals, such as mean and moments, to the contours of frame-wise low-level descriptors (LLDs), such as MFCCs or energy, over either fixed length segments (one second, corresponding to the annotated intervals in the corpus) or whole songs. In particular, we use the set of affective features developed as baseline for the 2013 Computational Paralinguistics Evaluation (ComParE) campaign [2]. It has been shown in [3] that this set provides robust cross-domain assessment of emotion (continuous arousal and valence) in speech, music, and acoustic events. Despite its rather 'brute-force' nature, it has been shown to outperform a more hand-crafted set of musically motivated features for the task of music mood regression.

The ComParE feature set contains 6 373 features. LLDs include auditory weighted frequency bands, their sum (corresponding to loudness), spectral measures such as centroid, roll-of point, skewness, sharpness, and spectral flux. Furthermore, voicing related LLDs such as fundamental frequency (corresponding to 'main melody') and harmonics-to-noise ratio (corresponding to 'percussiveness') are added. Delta regression coefficients (weighted discrete derivatives) are added to capture time dynamics. Statistical functionals include mean, moments, quartiles, 1- and 99-percentiles, as well as contour related measurements such as (relative) rise and fall times, amplitudes and standard deviations of local maxima ('peaks'), and linear and quadratic regression coefficients. An exhaustive list of the LLDs and functionals along with a detailed analysis of feature relevance for music mood regression is found in [3]. Extraction of acoustic features is done with our open-source toolkit openSMILE [4] which can be used 'out-of-the-box' to extract the ComParE set, so that our features can be reproduced by the interested reader. Prior to feature extraction, songs are normalized to -3 dB maximum amplitude using 'sox'. This is done to remove noise in energy-related features and improve generalization.

As regressors, we use Support Vector Regression (SVR) for song-level regression and bidirectional Long Short-Term Memory recurrent neural networks (BLSTM-RNNs) for dynamic regression. Both use the same input features, normalized to the range $[-1, +1]$ for SVR and standardized to zero mean and unit variance (on the training data) for BLSTM-RNNs. Separate SVR models are trained for arousal and valence regression while BLSTM-RNNs learn both arousal and valence prediction in a multi-task learning fashion. For BLSTM-RNNs, the regression targets are standardized as well. In addition, we investigate adding delta regression coefficients of the arousal and valence targets as additional regression tasks, in order to improve modeling of the dynamic emotion profile. The complexity constant for SVR training was varied from 10^{-4} to 10^{-1}. BLSTM-RNNs with two hidden layers (128 LSTM units per layer and direction) are used; thus, the first layer performs information reduction to a 128-dimensional feature set. The segments of each song are processed in order, forming sequences. Gradient descent with 25 sequences per weight update is used for training. An early stopping strategy is used, using a held out validation set in each fold. Training is stopped after a maximum of 100 iterations or after 20 iterations without improving the validation set error (sum of squared errors). To alleviate over-fitting to the high dimensional input feature set, Gaussian noise with zero mean and standard deviation 0.6 is added to the input activations, and sequences are presented in random order during training. SVR models are trained with Weka [5] using Sequential Minimal Optimization (SMO). BLSTM-RNNs are trained with our open-source CUDA RecuRrent Neural Network Toolkit (CURRENNT)[1] for further reproducibility. All hyper-parameters not mentioned in the above are left at the toolkits' defaults.

3. RESULTS

Table 1 shows the results on the development set (700 songs, 28 k segments). We use 10-fold cross validation on the development

[1] https://sourceforge.net/p/currennt

(a) Song level, SVR

	Arousal		Valence	
C	R^2	MLE	R^2	MLE
10^{-4}	.593	.090	.346	.099
10^{-3}	.656	**.078**	.419	.091
10^{-2}	.611	.087	.343	.104
10^{-1}	.580	.092	.323	.107

(b) Segment level, BLSTM

	Arousal			Valence		
Tasks	R^2	MLE	$\overline{\tau}$	R^2	MLE	$\overline{\tau}$
A+V	**.627**	.073	.140	**.427**	**.078**	**.152**
A+V+Δ	.626	**.072**	**.161**	.415	.079	.146

(c) Song level, BLSTM (average segment level predictions)

	Arousal		Valence	
Tasks	R^2	MLE	R^2	MLE
A+V	.682	.081	.495	**.087**
A+V+Δ	**.684**	.080	**.499**	.088

Table 1: Development set results (10-fold cross-validation). Best results per task (song / segment level) are printed in bold face.

set. Evaluation measures are computed on the entire development set (not by averaging across folds). The fold subdivision follows a simple modulo based scheme (song ID modulo 10), and is thus easily reproducible and song independent (in the case of regression on segments). We report the official challenge metrics, determination coefficient (R^2) for whole song regression and average Kendall's τ per song ($\overline{\tau}$) for segment regression, along with mean linear error (MLE). MLE is calculated after scaling the annotations to the range $[-0.5, +0.5]$. On segment level, we also report R^2 (across all segments) to assess the overall regression performance without taking into account the modeling of the emotional profile of a song.

In short, we observe that (a) SVR performance is very sensitive to the complexity parameter; (b) R^2 on segment level is very high compared to $\overline{\tau}$, indicating the difficulty of estimating the dynamics of the annotation contour within a song instead of the overall emotion; (c) adding deltas to the regression targets improves $\overline{\tau}$ for arousal, but not valence prediction; (d) best song level results in terms of R^2 are obtained by averaging BLSTM predictions, outperforming SVR by a large margin for valence (.499 vs. .419). In the following the configurations for our test set runs are summarized.

- Static task (song level):
  1. SVR: SVR with $C = 10^{-3}$, trained on the entire development set
  2. BLSTM-PA-Song: BLSTM-RNNs trained on the 10 training folds of the development set; segment level predictions averaged within songs and across networks
  3. BLSTM-WA-Song: BLSTM-RNN trained on the 10 training folds of the development set by weight averaging; segment level predictions averaged within songs
- Dynamic task (segment level):
  1. BLSTM-PA-Seg: BLSTM-RNNs trained on the 10 training folds of the development set; predictions averaged across networks
  2. BLSTM-WA-Seg: BLSTM-RNNs trained on the 10 training folds of the development set by weight averaging

(a) Song level ('Static task')

	Arousal		Valence	
Run name	R^2	MLE	R^2	MLE
SVR	.646	.083	.421	.095
BLSTM-PA-Song	.642	.085	.477	.090
BLSTM-WA-Song	.643	.085	.473	.091

(b) Segment level ('Dynamic task')

	Arousal		Valence	
Run name	$\overline{\tau}$	MLE	$\overline{\tau}$	MLE
BLSTM-PA-Seg	.180	.072	.124	.075
BLSTM-WA-Seg	.174	.073	.111	.076

Table 2: Test set results.

To deliver BLSTM predictions on the test set, we either average the predictions of the 10 networks trained on the development set (PA), or average their weights and run additional training iterations on the entire development set (WA).

Table 2 shows that BLSTM-RNNs outperform SVR on the song level for valence while being on par for arousal. This is consistent with the development set results. On the segment level, the WA strategy delivers slightly worse results in terms of $\overline{\tau}$ than PA while using a 10 times smaller model.

4. CONCLUSION

We have presented the TUM approach to the 2013 MediaEval Emotion in Music task. Best results on the static (song level) task were obtained by averaging time-varying predictions of a BLSTM-RNN. BLSTM-RNNs also delivered consistent improvements over the baseline in the dynamic task.

5. REFERENCES

[1] M. Soleymani, M. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proc. of CrowdMM (held in conjunction with ACM MM)*. Barcelona, Spain: ACM, 2013, to appear.

[2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[3] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, no. Article ID 292, pp. 1–12, May 2013.

[4] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM*. Barcelona, Spain: ACM, October 2013, 4 pages, to appear.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.