Automatically Estimating Emotion in Music with Deep Long-Short Term Memory Recurrent Neural Networks

Eduardo Coutinho[1,2], George Trigeorgis[1], Stefanos Zafeiriou[1], Björn Schuller[1]

[1]Department of Computing, Imperial College London, London, United Kingdom
[2]School of Music, University of Liverpool, Liverpool, United Kingdom
{e.coutinho, g.trigeorgis, s.zafeiriou, bjoern.schuller}@imperial.ac.uk

ABSTRACT

In this paper we describe our approach for the MediaEval's "Emotion in Music" task. Our method consists of deep Long-Short Term Memory Recurrent Neural Networks (LSTM-RNN) for dynamic Arousal and Valence regression, using acoustic and psychoacoustic features extracted from the songs that have been previously proven as effective for emotion prediction in music. Results on the challenge test demonstrate an excellent performance for Arousal estimation ($r = 0.613 \pm 0.278$), but not for Valence ($r = 0.026 \pm 0.500$). Issues regarding the quality of the test set annotations' reliability and distributions are indicated as plausible justifications for these results. By using a subset of the development set that was left out for performance estimation, we could determine that the performance of our approach may be underestimated for Valence (Arousal: $r = 0.596 \pm 0.386$; Valence: $r = 0.458 \pm 0.551$).

1. INTRODUCTION

The MediaEval 2015 "Emotion in Music" task comprises three subtasks with the goal of finding the best combination of methods and features for the time-continuous estimation of Arousal and Valence: **Subtask 1** - Evaluating the best feature sets for the time-continuous prediction of emotion in music; **Subtask 2** - Evaluating the best regression approaches using a fixed feature set provided by the organisers; **Subtask 3** - Evaluating the best overall approaches (the choice of features and regressor is free). The development set consists of a subset of 431 songs used in last year's competition (a total of 1 263) [1]. These pieces were selected for being annotated by at least 5 raters, and yielding good agreement levels (Cronbach's alpha ≥ 0.6). The test set comprises 58 new songs taken from freely available databases. Unlike the development set which includes only 45 seconds excerpts of the original songs, the songs in the test set are complete. For full details on the challenge tasks and database, please refer to [2].

2. METHODOLOGY

Feature sets. We used two features sets in our experiments, both of which were used in the first and last authors' submissions to last year's challenge [5]. The first feature set (*FS1*) is used this year by the organisers as the baseline set. It consists of the official set of 65 low-level acoustic descriptors (LLDs) from the 2013 INTERSPEECH Computational Paralinguistics Challenge (ComPareE; [15]), plus their first order derivates (130 LLDs, in total). The mean and standard deviation functionals of each LLD over 1 s time windows with 50 % overlap (step size of 0.5 s) are also calculated in order to adapt the LLDs to the challenge requirements. This results in 260 features exported at a rate of 2 Hz. All features were extracted using openSMILE ([8]). The second feature set (*FS2*) consists of the same acoustic features included in *FS1*, plus four features – Sensory Dissonance (*SDiss*), Roughness (R), Tempo (T), and Event Density (*ED*). These features correspond to two psychoacoustic dimensions strongly associated with the communication of emotion in music and speech (e. g., [4]). *SDiss* and R are instances of *Roughness*, whereas T and *ED* are indicators of the pace of music (*Duration* measures). The four features were extracted with the MIR Toolbox [10]. For estimating SDiss we used Sethares ([12]) formula, and for R Vassilakis algorithm ([13]). In relation to the *Duration*-related features, we used the *mirtempo* and *mireventdensity* functions to estimate, respectively, T and *ED*. T is measured in beats-per-minute (BPM) and *ED* as the number of note onsets per second. *FS2* was submitted as the mandatory run for *Subtask 1*.

Regressor: LSTM-RNN Similarly to the first and last author's approach to last year's edition of this challenge [5], and given the importance of the temporal context in emotional responses to music (e. g., [4]), we considered the use of deep LSTM-RNN [9] as regressors. An LSTM-RNN network is similar to an RNN except that the nonlinear hidden units are replaced by a special kind of memory blocks which overcome the vanishing gradient problem of RNNs. Each memory block comprises one or more self-connected memory cells and three multiplicative units – input, output, and forget gates – which provide the cells with analogues of write, read, and reset operations. The multiplicative gates allow LSTM memory cells to store and access information over long sequences (and corresponding periods of time) and to learn a weighting profile of the contribution of other moments in time for a decision at a specific moment in time. LSTM-RNN have been previously used on the context of time-continuous predictions of emotion in music (e.g., [5, 3, 16]).

Models training We used a multi-task learning framework for the joint learning of Arousal and Valence time-continuous values. The development set was divided into 11 folds using a modulus based scheme. A 10-fold cross-

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany

validation procedure was used in the development phase for parameter optimisation, and the extra fold was used to estimate the performance of our optimised model on the official test set. Our basic architecture consisted of deep LSTM-RNN with 3 hidden layers. Given that unsupervised pre-training of models has been demonstrated empirically to help converge speed, and to guide the learning process towards basins of attraction of minima that lead to better generalisation [7], we pre-trained the first hidden layer of the model. We used an unsupervised pre-training strategy consisting of de-noising LSTM-RNN auto-encoders (DAE, [14]). We first created a LSTM-RNN with a single hidden layer trained to predict the input features ($y(t) = x(t)$). In order to avoid over-fitting, in each training epoch and timestep t, we added a noise vector n to $x(t)$, sampled from a Gaussian distribution with zero mean and variance n. The development and test sets from last year's challenge was used to train the DAE. After determining the auto-encoder weights, the second and third hidden layers were added (and the output layer replaced by the regression variables). The number of memory blocks in each hidden layer (including the pre-trained layer), the learning rate (LR), and the standard deviation of the Gaussian noise applied to the input activations (σ; used to alleviate the effects of over-fitting when pre-training the first layer) were sequentially optimised (a momentum of 0.9 was used for all tests). An early stopping strategy was also used to further avoid overfitting the training data – training was stopped after 20 iterations without improvement of the performance (sum of squared errors) on the validation set. The instances in the training set of each fold were presented in random order to the model. Both the input and output data were standardised to the mean and standard deviation of the training sets in each fold. We computed 5 trials of the same model each with randomised initial weights in the range [-0.1,0.1].

Runs We submitted four runs for the whole challenge. The specifics of each run are as follows: *Run 1* consisted of the predictions of our model using the baseline features (*FS1*); The submitted predictions consisted of the average over a number of LSTM-RNN outputs selected from all folds and trials. The selected folds and trials were determined by minimising the root mean squared error ($RMSE$) on the small test set created to estimate the predictive power of our models before submission. *Run 2* was similar to *Run 1* but using *FS2*; *Run 3* was similar to *Run 1*, except that the selected folds and trials were selected by minimising the Concordance Correlation Coefficient (CCC) [11], which is a combined measure of precision (like $RMSE$) and similarity (like Pearson's linear correlation coefficient r).

3. RESULTS AND CONCLUSIONS

In Table 1, we report the official challenge metrics (r and $RMSE$) calculated individually for each music piece and averaged across all pieces (standard deviations are also given) of the challenge's official test set (a) and the team's test set (b). The analysis of the results obtained this year indicate that all our runs performed better than the baseline for Arousal. On the official test set, runs 3 and 4 led to lowest $RMSE$ (0.234 and 0.236, respectively) and runs 2 and 4 to the highest r (0.613). Thus, *Run 3* led to the best compromise between both measures. This run consists of the average outputs of two LSTM-RNNs with three layers (200+150+25) and *FS2* as input. The model's hyper-parameters and the number of networks used to estimate Arousal and Valence were optimised using the CCC ($LR = 5 * 10^{-6}$, noise $\sigma = 0.3$). In relation to Valence, our models perform below the baseline on the official test set (see Table 1 b)). One possible reason for this may be the low quality of the Valence annotations obtained for the test set this year (the average Cronbach's α [6] across all test pieces for Valence is 0.29). In Table 1 b) we show the performance estimated on another test set consisting of a subset of the development set that was left out exclusively to estimate the performance of the runs submitted to the challenge. As it can be seen, Valence predictions yield much better results, while the Arousal performance on the team test set is comparable to the one reached with the official test set. Furthermore, in terms of r ($RMSE$ cannot be compared), these results are noticeably better than the best results in last year's challenge (*me14*), which can be due to the use of more reliable targets during training or the extra hidden layer added to the model. Another possibility is that our models over-fitted the development data in aspects that are not directly visible. According to the organisers, the development set annotations yield a high correlation between Arousal and Valence, whereas the test set does not. It could thus be that, the models are picking up on this particularity of the development set, which gives unwanted effects for new music where Arousal and Valence are not correlated.

Table 1: Results on the official test set (a) and team's test set (b). *CB*: challenge baseline; *me14*: best team results in the 2014 challenge.

		Run	Arousal	Valence
a)	$RMSE$	2	0.242±0.116	0.373±0.195
		3	0.234±0.114	0.372±0.190
		4	0.236±0.114	0.375±0.191
		CB	0.270±0.110	0.366±0.180
	r	2	0.611±0.254	0.004±0.505
		3	0.599±0.287	0.017±0.492
		4	0.613±0.278	0.026±0.500
		CB	0.360±0.260	0.010±0.380
b)	$RMSE$	2	0.206±0.128	0.212±0.116
		3	0.221±0.119	0.185±0.119
		4	0.220±0.121	0.183±0.110
		me14	0.102±0.052	0.079±0.048
	r	2	0.532±0.421	0.394±0.509
		3	0.596±0.386	0.458±0.551
		4	0.591±0.386	0.456±0.543
		me14	0.354±0.455	0.198±0.492

In future studies, apart from verifying this possibility, we will further investigate optimal pre-training strategies for deep LSTM-RNNs. Further, beyond the expert-given feature sets employed here, we will consider opportunities of end-to-end deep learning strategies.

4. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the the European Union's Horizon 2020 research and innovation programme under grant agreement no. 645378 (ARIA-VALUSPA).

5. REFERENCES

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

[2] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[3] E. Coutinho, J. Deng, and B. Schuller. Transfer learning emotion manifestation across music and speech. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 3592–3598, Beijing, China, 2014.

[4] E. Coutinho and N. Dibben. Psychoacoustic cues to emotion in speech prosody and music. *Cognition & emotion*, 27(4):658–684, 2013.

[5] E. Coutinho, F. Weninger, B. Schuller, and K. R. Scherer. The munich lstm-rnn approach to the mediaeval 2014 "emotion in music" task. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, pages 5–6, Wurzen, Germany, 2014.

[6] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

[7] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660, 2010.

[8] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*, pages 835–838, Barcelona, Spain, October 2013.

[9] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[10] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[11] L. I.-K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989.

[12] W. A. Sethares. *Tuning, timbre, spectrum, scale*, volume 2. Springer, 2005.

[13] P. Vassilakis. Auditory roughness estimation of complex spectra – roughness degrees and dissonance ratings of harmonic intervals revisited. *The Journal of the Acoustical Society of America*, 110(5):2755–2755, 2001.

[14] P. Vincent, Y. B. Larochelle, and P. A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning(ICML'08)*, pages 1096–1103. ACM, 2008.

[15] 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 Psychology*, 4(Article ID 292):1–12, May 2013.

[16] F. J. Weninger, F. Eyben, and B. Schuller. On-Line Continuous-Time Music Mood Regression with Deep Recurrent Neural Networks. In *Proceedings 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5449–5453, Florence, Italy, May 2014. IEEE, IEEE.