Time-continuous estimation of emotion in music with recurrent neural networks

Thomas Pellegrini, Valentin Barrière
Université de Toulouse, IRIT, Toulouse, France
thomas.pellegrini@irit.fr

ABSTRACT

In this paper, we describe the IRIT's approach used for the MediaEval 2015 "Emotion in Music" task. The goal was to predict two real-valued emotion dimensions, namely valence and arousal, in a time-continuous fashion. We chose to use recurrent neural networks (RNN) for their sequence modeling capabilities. Hyperparameter tuning was performed through a 10-fold cross-validation setup on the 431 songs of the development subset. With the baseline set of 260 acoustic features, our best system achieved averaged root mean squared errors of 0.250 and 0.238, and Pearson's correlation coefficients of 0.703 and 0.692, for valence and arousal, respectively. These results were obtained by first making predictions with an RNN comprised of only 10 hidden units, smoothed by a moving average filter, and used as input to a second RNN to generate the final predictions. This system gave our best results on the official test data subset for arousal (RMSE=0.247, r=0.588), but not for Valence. Valence predictions were much worse (RMSE=0.365, r=0.029). This may be explained by the fact that in the development subset, valence and arousal values were very correlated (r=0.626), and this was not the case with the test data. Finally, slight improvements over these figures were obtained by adding spectral flatness and spectral valley features to the baseline set.

1. INTRODUCTION

Music Emotion Recognition still is a hot topic in Music Information Retrieval. In [15], the authors list four main issues that explain why MER is a challenging and very interesting scientific task: 1) ambiguity and granularity of emotion description, 2) heavy cognitive load of emotion annotation, 3) subjectivity of emotional perception, 4) semantic gap between low-level acoustic features and high-level human perception. It consists of either labeling songs and music pieces as a whole, thus involving a classification task, or estimating emotion dimensions in continuous time and space domains, being then a regression task applied to time series. This last case is the objective of the current challenge. For a complete description of the task and corpus involved in challenge, the reader may refer to [4].

For continuous-space MER, many machine learning (ML) techniques were reported in the literature [11]. In the MediaEval 2014 challenge edition, a variety of techniques were used: simple and multi-level linear regression models [12], Support Vector machines for regression (SVR) [9], conditional random fields [14], Long Short-Term Memory and Recurrent Neural Networks (LSTM-RNN) [7]. This last approach was the one that achieved the best results. Following these results, we chose to use RNNs. All the ML models were developed using the Theano toolbox [5].

2. METHODOLOGY

In order to tune and test prediction models, we ran 10-fold cross-validation (CV) experiments on the development data subset. Once the best model was selected and tuned within this setup, a single model was trained on the whole development subset, and used to generate predictions on the official evaluation data subset.

The input data were zero-mean and unit-variance normalized. Standard PCA, PCA with a Gaussian kernel and denoising autoencoders with Gaussian noise were tested to further process the data, but no improvement was achieved with any of these techniques.

We chose to use recurrent neural networks (RNN) for their time sequence modeling capabilities. We used the Elman [8] model type, in which recurrent connections feed the hidden layer. The activations of the hidden layer at time $t-1$ are stored and fed back to the same layer at time t together with the data input. A tanh activation function and a softmax function were used for the hidden layer and the final layer with two outputs (for arousal and valence), respectively. The layer weights were trained with the standard mean-root-squared cost function. Weights were updated after each forward pass on a single song via the momentum update rule.

The hyperparameters were tuned with the 10-fold CV setup. The best model was comprised of 10 hidden units, trained with a 1.0×10^{-3} learning rate and a 1.0×10^{-2} regularization coefficient with both $L1$ and $L2$ norms. To further limit overfitting, an early stopping strategy was used: the models were all trained with 50 iterations only. This number of iterations was set empirically.

A moving average filter was used to smooth the predictions. Its size was tuned in the 10-fold CV setup, and the best one was a window of 13 points. To avoid unwanted border effects, the first and last 6 points, corresponding to the filter delay, were equaled to the unfiltered predictions.

Another post-processing step was tested. It consisted of feeding another RNN with the predictions of the first RNN. By looking at the output, one can see that the second RNN further smoothed the predictions.

Copyright is held by the author/owner(s).
MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

Table 1: 10-fold cross-validation (CV) and official evaluation test (Eval) results. *lr*: linear regression model, *BSL*: baseline results provided by the organizers, *rnn*: RNN, *rnn2*: RNN fed with the predictions of the first RNN.

System	CV				Eval			
	Valence		Arousal		Valence		Arousal	
	RMSE	r	RMSE	r	RMSE	r	RMSE	r
lr, 260 feat.	.275	.637	.254	.646	N/A	N/A	N/A	N/A
BSL, 260 feat.	N/A	N/A	N/A	N/A	.366 ± .18	.01 ± .38	.27 ± .11	.36 ± .26
rnn, 260 feat.	.261	.675	.246	.670	.377 ± .181	.017 ± .420	.259 ± .112	.518 ± .238
+ smoothing	.254	.694	.239	.689	.365 ± .188	.029 ± .476	.247 ± .116	.588 ± .235
+ rnn2	.250	.703	.238	.692	N/A	N/A	N/A	N/A
rnn, 268 feat.	.259	.678	.245	.673	.373 ± .180	.023 ± .422	.254 ± .106	.532 ± .224
+ smoothing	.252	.697	.238	.692	.361 ± .187	.044 ± .487	.243 ± .111	.612 ± .216
+ rnn2	.249	.706	.238	.694	.371 ± .194	.044 ± .525	.244 ± .115	.635 ± .222

The challenge rules also allowed to use our own acoustic features. To complete the 260 baseline features, a set of 29 acoustic feature types were extracted with the ESSENTIA toolbox [6], which is a toolbox specifically designed for Music Information Retrieval. The 29 feature types such as Bark and Erb bands that use several frequency bands resulted in a total of 196 real values per audio frame. The same frame rate as the baseline feature one was used (0.5s window duration and hop size). We chose the feature types among a large list, from the spectral domain mainly, such as the so-called spectral "contrast", "valley", "complexity", but also a few features from the time domain, such as "danceability". For a complete list and description of the available feature extraction algorithms, the reader may refer to the ESSENTIA API documentation Web page [2].

In order to select useful features, we tested each feature type by adding them one at a time to the baseline feature set. Only three feature types were found to improve the baseline CV performance: two variants of spectral flatness and a feature called "spectral valley". The two spectral flatness features use two different frequency scales: the Bark and the Equivalent Rectangular Bandwidth (ERB) scales. 25 Bark bands were used, as computed in ESSENTIA [1, 3]. The ERB scale consists of applying a frequency domain filterbank using gammatone filters [13]. Spectral flatness provides a way to quantify how noise-like a sound is, as opposed to being tone-like. Spectral valley is a feature derived from the so-called *spectral contrast* feature, which represents the relative spectral distribution [10]. This feature was shown to perform better than Mel frequency cepstral coefficients in the task of music type classification [10].

3. RESULTS

Results are shown in Table 1, for both the cross-validation experiments and the runs on the official evaluation test data subset, referred to as 'CV' and 'Eval', respectively. The results are reported in terms of root-mean-squared error (RMSE) and Pearson's correlation coefficient (r).

Generally speaking, valence predictions were less accurate than the arousal ones, unlike the performance results reported in the 2014's edition, as reported in [7], for example. Concerning the CV results, the simple linear regression model (*lr*) was outperformed by the RNN model with the baseline 260 features, with 0.275 and 0.261 RMSE values for valence, 0.254 and 0.246 for arousal, respectively. Since the number of runs was limited, we did not submit predictions with *lr* on Eval. As expected, this shows that the sequential modeling capabilities of the RNN are useful for this task. Adding the extra 8 features brought slight improvement (*rnn, 268feat.*). Smoothing the network predictions brought further improvement, using either 260 or 268 features as input. Finally, using the predictions as input to a second RNN brought slight improvement too. The best system achieved averaged RMSE of 0.250 and 0.238, and Pearson's correlation coefficients of 0.703 and 0.692, for valence and arousal, respectively.

Concerning the Eval results, this system also gave the best results on the official test data subset but for arousal (RMSE=0.247, r=0.588) only. Valence predictions were much worse (RMSE=0.365, r=0.029). This may be explained by the fact that in the development subset, valence and arousal values were very correlated (r=.626), and this was not the case with the test data, as hypothesized by the challenge organizers. This performance discrepancy was also observed by the organizers that provided baseline results ('BSL') using a linear regression model [4]. Only our best three arousal predictions outperformed the BSL results significantly.

4. CONCLUSIONS

In this paper, we described our experiments using RNNs for the 2015 MediaEval Emotion in Music task. As expected, the sequence modeling capabilities revealed useful for this task since basic linear regression models were outperformed in our cross-validation experiments. Prediction smoothing also revealed useful. The best results were obtained when using smoothed predictions fed into a second RNN for both valence and arousal in our CV experiments, and only for arousal on the official test set. The observed performance discrepancy between the valence and arousal variables may be due to the differences between the development and test data: valence and arousal values were very much correlated in the development dataset, and much less in the test data set. Concerning the acoustic feature set, slight improvements were obtained by adding spectral flatness and spectral valley features to the baseline feature set. As future work, we plan to further explore denoising encoders, LSTM-RNNs, since our first experiments with these models did not show improvement compared to the use of basic RNNs.

5. REFERENCES

[1] The bark frequency scale. http://ccrma.stanford.edu/~jos/bbt/Bark_Frequency_Scale.html. Accessed: 2015-08-24.

[2] Essentia algorithm documentation web page. http://essentia.upf.edu/documentation/algorithms_reference.html. Accessed: 2015-08-24.

[3] The essentia bark documentation page. http://essentia.upf.edu/documentation/reference/std_BarkBands.html. Accessed: 2015-08-24.

[4] 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.

[5] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a cpu and gpu math expression compiler. In *Proc. of the Python for scientific computing conference (SciPy)*, volume 4, page 3, Austin, 2010.

[6] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, and et al. ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *Proc. International Society for Music Information Retrieval Conference (ISMIR'13)*, pages 493–498, Curitiba, 2013.

[7] E. Coutinho, F. Weninger, B. Schuller, and K. Scherer. The Munich LSTM-RNN Approach to the MediaEval 2014 âĂIJEmotion in MusicâĂİ Task. In *Working Notes Proceedings of the MediaEval 2014 Workshop*, Barcelona, 2014.

[8] J. Elman. Finding structure in time. *Cognitive Science*, 14(2), 1990.

[9] V. Imbrasaite and P. Robinson. Music emotion tracking with continuous conditional neural fields and relative representation. 2014.

[10] D. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. Music type classification by spectral contrast feature. In *Proc. ICME*, volume 1, pages 113–116, Lausanne, 2002.

[11] Y. Kim, E. Schmidt, R. igneco, O. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. Emotion recognition: a state of the art review. In *11th International Society for Music Information and Retrieval Conference*, Utrecht, 2010.

[12] N. Kumar, R. Gupta, T. Guha, C. Vaz, M. Van Segbroeck, J. Kim, and S. Narayanan. Affective feature design and predicting continuous affective dimensions from music. In *MediaEval Workshop*, Barcelona, 2014.

[13] B. C. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:3:750–753, 1983.

[14] W. Yang, K. Cai, B. Wu, Y. Wang, X. Chen, D. Yang, and A. Horner. BeatsensâĂŹ Solution for MediaEval 2014 Emotion in Music Task. 2014.

[15] Y.-H. Yang and H. H. Chen. *Music emotion recognition.* CRC Press, 2011.