ON-LINE CONTINUOUS-TIME MUSIC MOOD REGRESSION WITH DEEP RECURRENT NEURAL NETWORKS

Felix Weninger[1], *Florian Eyben*[1], *Björn Schuller*[1,2]

[1] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2] Department of Computing, Imperial College London, U.K.
{weninger,eyben,schuller}@tum.de

ABSTRACT

This paper proposes a novel machine learning approach for the task of on-line continuous-time music mood regression, i.e., low-latency prediction of the time-varying arousal and valence in musical pieces. On the front-end, a large set of segmental acoustic features is extracted to model short-term variations. Then, multi-variate regression is performed by deep recurrent neural networks to model longer-range context and capture the time-varying emotional profile of musical pieces appropriately. Evaluation is done on the 2013 MediaEval Challenge corpus consisting of 1 000 pieces annotated in continous time and continuous arousal and valence by crowdsourcing. In the result, recurrent neural networks outperform SVR and feedforward neural networks both in continuous-time and static music mood regression, and achieve an R^2 of up to .70 and .50 with arousal and valence annotations.

Index Terms— music information retrieval; emotion recognition; recurrent neural networks

1. INTRODUCTION

Music mood recognition, i. e., automatic determination of the perceived emotion, is a highly promising topic in music information retrieval, with applications in the organization of music collections and recommendation. Early work on music mood recognition started as a special case of music tagging, by using categorical labels such as *happy* or *sad* [1]. However, such categorical taxonomies are often ambiguous [2]; hence, many recent studies – e.g., [3–6] – use a dimensional model of affect proposed by Russell [7], describing emotional tags as points in the plane spanned by the arousal and valence axes. This turns the problem of emotion prediction into a two-dimensional regression problem [2]. Besides, in most music pieces, the emotion is not static, but rather varies over time; for example, composers often contrast passages of different emotional content, e.g., lively vs. calm, with each other – this holds both for classical and popular music. This observation, together with the above, calls for an emotion model that is continuous both in time and in value.

From the machine learning perspective, time-continuous recognition comes with the need for context-sensitive models, since the emotion in a certain time interval depends on past and future input to some (variable) extent. In the music mood recognition domain, a few classes of context-sensitive models have been proposed; among them are graphical models such as conditional random fields [8], and recurrent neural networks (RNNs) [6], which can also perform regression. RNNs can access previous hidden layer activations to deliver the emotion prediction for the current time frame. In this study, we adopt the Long Short-Term Memory (LSTM) type of RNN since it has shown superior performance in music information retrieval tasks such as onset detection [9] and transcription [10].

Since emotion in speech and music evolves much slower than typical acoustic features derived from short-time spectra, there is a need for aggregating these features to match the time scales. In this study, we apply functionals such as moments, percentiles and regression coefficients to the 'low-level' feature contours, resulting in generic affective features which are highly effective for describing emotion in speech, music, and sound, and across these domains [11]. Next, we propose to use a deep RNN structure to reduce these feature vectors over multiple time steps to hidden representations.

Our methods are evaluated on the 2013 MediaEval Music Emotion task, the first major comparative evaluation campaign for music mood recognition. While the focus is on low-latency prediction of emotion from short observation intervals, we also address song level estimates, which are useful, e.g., for sorting a music archive by mood. Our results are given in Section 5, indicating the effectiveness of the proposed method for prediction of emotional profiles and overall mood of musical pieces, compared to a support vector regression (SVR) and feedforward neural network baseline.

2. RELATED WORK

One of the earliest studies on music mood regression by SVR was presented by [2]. [12] describes on-line continuous-time continuous-valued *speech* emotion recognition with multi-task LSTM-RNN networks, but does not consider feature reduction. [5] models time-varying emotion in an aggregated fashion by performing regression on the distribution, rather than context modeling across periods with different emotion. Small-scale evaluations of (shallow) recurrent neural networks for music mood recognition have been performed in [6, 13]. [14] uses a feedforward deep belief network for unsupervised feature generation from short-term features, which can be seen as a replacement for the rule-based functional extraction step which is performed in our work; however, the actual regression is done by linear regression, without context modeling.

3. METHODOLOGY

3.1. Acoustic Feature Brute-Forcing

The front-end of our approach consists of supra-segmental features calculated by applying statistical functionals, such as mean and mo-

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 338164 (ERC Starting Grant iHEARu).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Table 1: 70 provided low-level descriptors (LLD) (top) and applied functionals (bottom). [1]: only applied to F0; [2]: not applied to voicing related LLD; [3]: only applied to voicing related LLD

4 energy related LLD
Sum of auditory spectrum (loudness).
Sum of RASTA-style filtered auditory spectrum.
RMS Energy, Zero-Crossing Rate.
64 spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0–8 kHz).
MFCC 1–14.
Spectral energy (logarithmic) 250–650 Hz, 1 k–4 kHz.
Log. spectral slope 0–1 kHz 1–5 kHz, and 0–5 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90.
Log. spectral entropy, harmonicity
Psychoacoustic Sharpness
CHROMA (Pitch Class Profiles) 1–12
2 voicing related LLD
F0 by SHS with Viterbi smoothing, Probability of voicing

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
5 % percentile ($\approx$ min), 95 % percentile ($\approx$ max)
percentile range 5 %–95 %
standard deviation
gain of linear prediction (LP), LP Coefficients 1–5
amplitude mean of peaks[2], of minima [2]
amplitude range of peaks[2]
mean value of peaks – arithmetic mean[2]
mean / std.dev. of inter peak distances[2]
mean / std.dev. of rising / falling slopes[2]
Functionals applied to LLD only
arithmetic mean, root quadratic mean, flatness
rel. duration LLD is above 25 / 50 / 75 / 90% range
rel. duration LLD is rising
rel. duration LLD has positive curvature
mean, max, min, std. dev. of segment length[1]
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames[1]
skewness[3], kurtosis[3]

ments, to the contours of frame-wise low-level descriptors (LLDs), such as chroma features, MFCCs or energy, over fixed length segments. In this study, non-overlapping segments of one second length are used, which corresponds to the annotation interval. We use a set of purely acoustic affective features based on the baseline feature set of the 2013 Computational Paralinguistics Evaluation (ComParE) campaign [15]. It has been shown in [11] 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 [11]. While our MediaEval submission was based on the ComParE set, for the present study, we modified and reduced the ComParE feature set in order to adapt to the specific task of music mood recognition. For example, chroma features are added, and purely human voice related features such as jitter and shimmer were removed. Furthermore, a few rather redundant functionals (such as various types of means) or functionals not connected to the task (position of max/min) were eliminated. In the result, our feature set contains 4 777 features. The LLDs and functionals are shown in detail in Tables 1 and 1.

3.2. Deep Recurrent Neural Networks

The neural network architecture we adopt in this study is based on Long Short-Term Memory (LSTM) deep recurrent neural networks (RNNs) [16]. A deep LSTM-RNN can be described as an automaton-like structure mapping from a sequence of observations to a sequence of output features. These mappings are defined by activation weights and a non-linear activation function as in a standard multi-layer perceptron. However, recurrent connections allow to access activations from past time frames. To solve the problem of exponential weight decay (or blowup) in the recurrent connections, the LSTM concept introduces an internal state variable ('memory cell') whose content is modified in each timestep by so-called input and forget gates [17], instead of simply having a recurrent connection with constant weight. In other words, memory is modeled explicitly instead of implicitly (by recursion), as in traditional RNNs. The output of each layer of LSTM cells is determined by a non-linear function of the cell states, scaled by the output gate. Mathematically, the following iterative procedure is executed in a N-layer deep RNN:

$$\mathbf{h}_t^{(0)} := \mathbf{x}_t, \tag{1}$$

$$\mathbf{c}_t^{(n)} := \mathbf{f}_t^{(n)} \otimes \mathbf{c}_{t-1}^{(n)} + \mathbf{i}_t^{(n)} \otimes \tanh\left(\mathbf{W}^{(n-1),(n)}\mathbf{h}_t^{(n-1)} + \mathbf{W}^{(n),(n)}\mathbf{h}_{t-1}^{(n)} + \mathbf{b}^{(n)}\right), \tag{2}$$

$$\mathbf{h}_t^{(n)} := \mathbf{o}_t^{(n)} \otimes \tanh(\mathbf{c}_t^{(n)}),$$

$$\hat{\mathbf{y}}_t := \mathbf{W}^{(N),(N+1)}\mathbf{h}_t^{(N)} + \mathbf{b}^{(N+1)}. \tag{3}$$

In the above, $\mathbf{h}_t^{(n)}$ denotes the hidden feature representation of time frame t in the level n units, $n = 1, \ldots, N$. The 0-th layer is the input layer and the $N + 1$-th layer the output layer. Analogously, $\mathbf{c}_t^{(n)}$, $\mathbf{f}_t^{(n)}$, $\mathbf{i}_t^{(n)}$, and $\mathbf{o}_t^{(n)}$ denote the dynamic cell state, forget gate, input gate, and output gate activations. $\mathbf{W}^{(n-1),(n)}$ and $\mathbf{W}^{(n),(n)}$ denote weight matrices for feedforward and recurrent connections and $\mathbf{b}^{(n)}$ stands for bias vectors (with superscripts denoting layer indices). The input gate activations $\mathbf{i}_t^{(n)}$ regulate the 'influx' from the feedforward and recurrent connections. $\mathbf{f}_t^{(n)}$, $\mathbf{i}_t^{(n)}$, and $\mathbf{o}_t^{(n)}$ are calculated in a similar fashion as $\mathbf{c}_t^{(n)}$ (2) – see [16] for details. The weight matrices and bias vectors are all learnt from training sequences, minimizing the average sum of squared errors per sequence.

In our application, the weight matrix $\mathbf{W}^{(0),(1)}$ maps high-dimensional input features $\mathbf{x}_t$ to lower-dimensional hidden layer features $\mathbf{h}_t^{(1)}$, and the recurrent weight matrix $\mathbf{W}^{(1),(1)}$ serves to aggregate information from multiple time steps in $\mathbf{h}_t^{(1)}$. Thus, information from the low-level descriptors is aggregated hierarchically: once by application of functionals, then by the RNN. The remaining weights then define a mapping from $\mathbf{h}_t^{(1)}$ to the emotion labels.

3.3. Multi-Task Learning

Multi-task learning can be used as a regularization to improve generalization of neural networks. We consider joint learning of arousal and valence, as well as joint learning of the instantaneous emotion (arousal/valence) with their dynamics. As a simple measurement for the dynamics, we use delta regression coefficients as typically done in speech recognition, i.e., adding 2×2 future and past targets in a regression formula. Since the future timesteps are only used in training, this does not conflict with the goal of an on-line system.

3.4. De-Noising Auto-Encoder Pre-Training

To add more structure to deep networks, pre-training of the first hidden layer(s) is often used, especially with low amounts of training data [14]. In this study, we use pre-training based on the de-noising auto-encoder principle [18]. We first create a LSTM network with a single hidden layer trained to predict the input features, i.e., $\mathbf{y}_t = \mathbf{x}_t$. To avoid over-fitting, in each training epoch and timestep t, we add a noise vector $\mathbf{n}$ to $\mathbf{x}_t$ (Eqn. 1) which is sampled from a Gaussian distribution with zero mean and variance σ_n. After determining the auto-encoder weights a second hidden layer is added and the rest of the weights is trained, this time using the regression targets and keeping the first layer weights constant.

4. EXPERIMENTAL SETUP

4.1. MediaEval Emotion in Music Database

Our evaluation database is the official corpus of the MediaEval 2013 Emotion in Music challenge. Of the 1 000 song database presented in [19], 700 songs were assigned to the development set and 300 songs to the test set. In this study, we only use the development set as the test set labels have not been released yet. Genres are balanced and cover classical music (including contemporary pieces), blues, jazz, rock and pop. Songs are annotated via crowd-sourcing on Amazon Mechanical Turk (AMT) in the dimensions arousal and valence. Time continuous observer annotations are averaged and 're-sampled' at 1 Hz such that the 'ground truth' for every segment of 1 s length corresponds to the average of all available annotations within that segment. After pre-selection, 100 qualified annotators participated in the AMT experiment. Annotation was done by mouse movements, separately for arousal and valence.

4.2. Classifier Training and Evaluation

We evaluate in 10-fold cross validation on the development 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, i.e. segments of one song do not occur in more than one fold.

We compare the proposed LSTM-RNN approach to feedforward neural networks (FNN) and Support Vector Regression (SVR). 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 neural networks. SVR is chosen for its capability to handle large feature spaces by L2 regularization and its popularity in music mood regression [2, 3, 11]. The regression targets are standardized independently to zero mean and unit variance, which is important for multi-task neural networks minimizing the sum of squares error function, which is sensitive to the scaling of the target variables with respect to each other. The complexity constant for SVR training was chosen as 10^{-4} based on our experiments on the MediaEval development set described in [20]. Neural networks have one or two hidden layers with 192, 256, or 384 units (cf. Section 5 for details on the topologies). Gradient descent with 25 sequences per weight update is used for training. An early stopping strategy is used, using a held out part of each fold's training set. 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. The same amount of noise is used in de-noising auto-encoder pre-training ($\sigma_n = 0.6$). Sequences are presented in random order during training. SVR models are trained with Weka [21] using Sequential Minimal Optimization (SMO). BLSTM-RNNs are trained with our open-source CUDA RecuRrent Neural Network Toolkit (CURRENNT)[1] for reproducibility. All 'hyper'-parameters not mentioned in the above are left at the toolkits' defaults.

	Arousal			Valence		
Model	R^2	MLE	$\overline{\tau}$	R^2	MLE	$\overline{\tau}$
LSTM-RNN	**.635**	**.071**	**.222**	**.421**	.078	**.174**
FNN	.557	.076	.096	.298	.083	.038
FNN(stack2)	.576	.073	.109	.320	**.075**	.041
SVR	.566	.074	.132	.312	.075	.059

Table 2: Single-Task Regression on MediaEval 2013 development set in 10-fold cross-validation: LSTM-RNN vs. FNN and FNN with two input frames concatenated (stack2).

	Arousal			Valence		
Tasks	R^2	MLE	$\overline{\tau}$	R^2	MLE	$\overline{\tau}$
A+ΔA / V+ΔV	**.638**	**.071**	**.223**	.425	.078	.181
A+V	.626	.072	.208	.433	.079	**.192**
A+V+ΔA+ΔV	.634	.071	.207	.433	**.078**	.183

Table 3: Multi-Task Nets: Effects of different training targets.

4.3. Evaluation Metrics

We report the official challenge metrics [19], determination coefficient (R^2) and average Kendall's τ per song ($\overline{\tau}$). The latter is a measure of how well the emotional profile of each song is captured by the regressor, as opposed to overall correlation. For example, a system that predicts the correct average emotion of the song in each segment would have high R^2 but zero $\overline{\tau}$. Mean linear error (MLE) corresponding to the range $[-0.5, +0.5]$ is provided for reference.

5. EXPERIMENTAL RESULTS

5.1. Single-Task Regression

In a first set of experiments, we compare the performance of different single-task regressors (LSTM-RNN, FNN, and SVR). That is, we train separate models for arousal and valence. We also investigate 'stacking' input feature vectors for the FNN in order to provide context. We restrict the context to one additional (past) frame to keep the number of parameters in the input connection reasonable (9 554 instead of 4 777 input features). Table 2 shows the results for single-task regression. It can be seen that LSTM-RNN deliver best performance in terms of all of the considered evaluation metrics. The most interesting result is perhaps that $\overline{\tau}$ increases drastically when using LSTM-RNNs. There is a slight performance gain by adding one frame of context at the expense of doubling the parameter size of the first layer, but from the results it can hardly be expected that adding more frames would reach the performance of LSTM.

5.2. Multi-Task Regression

In a second set of experiments, we investigate the benefit of adding the deltas of the emotion profile as targets, as well as learning arousal and valence models separately. Thus, multi-task models have two

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

LSize	#L	Pre	Arousal R^2	Arousal MLE	Arousal $\overline{\tau}$	Valence R^2	Valence MLE	Valence $\overline{\tau}$
192	1	–	.610	.072	.147	.402	.071	.093
256	1	–	.607	.073	.141	.385	.072	.079
384	1	–	.606	.074	.150	.378	.072	.070
192	2	–	.643	.070	.211	.423	.070	**.203**
256	2	–	.634	.071	.207	**.433**	.078	.183
384	2	–	**.644**	.070	**.229**	.413	.070	.186
192	2	✓	.631	**.068**	.222	.421	.070	.194
256	2	✓	.624	.069	.216	.416	**.069**	.179
384	2	✓	.635	.068	.214	.403	.070	.171

Table 4: 1- and 2-layer nets: Effects of network topologies (layer size: LSize, number of layers: #L) and of pre-training the first layer (pre). LSTM-RNNs.

Model	Arousal R^2	Arousal MLE	Valence R^2	Valence MLE
SVR/Song	0.541	0.088	0.320	0.100
Avg SVR/1 s	0.689	**0.059**	0.458	**0.062**
Avg LSTM	**0.704**	0.085	**0.500**	0.087

Table 5: Song-level results by averaging predictions (SVR or LSTM, 1 s functionals) or taking functionals over whole songs (SVR/Song).

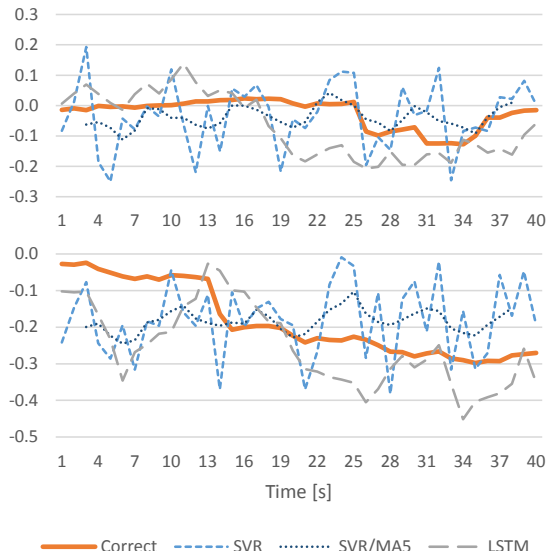
or four regression targets (arousal and valence, arousal plus delta, valence plus delta, arousal and valence plus deltas). Table 3 shows the results for these models, when LSTM-RNNs are used. It can be seen that adding delta regression coefficients improves performance; the improvement in $\overline{\tau}$ for valence is statistically significant ($p < .05$ according to a one-tailed t-test). Learning arousal and valence prediction in a single net further improves R^2 and $\overline{\tau}$ for valence, but the performance in arousal prediction decreases. Thus, there seems to be a trade-off between the benefit of regularization to learning valence recognition (which is harder, according to the single-task results), and precise learning of the 'easier' arousal task. Still, the four-task network delivers the best average R^2 for arousal and valence (.528). For the sake of clarity, we always use four training targets in the ongoing.

5.3. Deep Networks and Pre-Training

Third, we show the effectiveness of deep neural networks for the task of predicting emotion from large feature sets. In Table 4, we display the results obtained with a single hidden layer and the LSTM-RNN architecture. It can be seen that these are largely inferior (statistically significant) to the results reported above. When increasing the layer size, performance further decreases. This shows that the simple single hidden layer model is inadequate for the given task; more than layer is needed so that the first layer can perform the feature reduction task. For networks with two hidden layers, there is no clear trend as to which layer size performs best; by tuning the layer size, we can only achieve a slight gain to .229 / .203 average τ. This clearly shows that depth is more important than breadth in our task. Finally, if we pre-train the first layer, there is no significant gain in performance. This could indicate that the gain by including the regression targets in training the first layer (and thus having $\mathbf{h}_t^{(1)}$ targeted to the regression task) outweighs the gain of pre-training – note that we also use noise on the input features when not pre-training the first layer.

In Figure 1, we show an example from the MediaEval development set (song 130). It can be seen that both SVR and LSTM

Fig. 1: Arousal (top) and valence (bottom) of song #130 from the MediaEval development set: Ground truth (continuous line), SVR, and LSTM predictions; SVR predictions smoothed by moving average (order 5, MA5).



deliver reasonable predictions of the average mood in the shown 40 second clip. However, the contour of the SVR prediction, even when smoothed by a moving average filter, hardly represents the actual contour of the emotion ($\tau = -.016$ for arousal, $\tau = .103$ for valence). The LSTM prediction is much smoother and more in line with the correct emotional profile ($\tau = .338$ for arousal and $\tau = .538$ for valence), although there is some overshooting.

In Table 5 we show the performance of predicting the average song emotion, once by extending the functionals over the whole song and using SVR, once by using the SVR predictions as above and taking the average per song, and finally by averaging LSTM predictions (two layers, 384 units, 4-task). From the results it seems that taking functionals over whole songs is rather unstable, probably because such long feature contours can hardly be captured by simple statistics. Comparing the averaging of regression outputs on 1 second functionals, LSTMs deliver better correlations with the song-level annotation than SVR at the expense of increased MLE. Similar to the shapes in Figure 1 we often observe that the LSTM overshoots on changes of the emotion contour, while SVR predictions are closer to the average ground truth without capturing much of the time-varying emotional profile.

6. CONCLUSIONS AND OUTLOOK

We have introduced an effective approach for music mood regression, combining acoustic feature brute-forcing and RNN-based context-sensitive feature reduction and regression. On the MediaEval challenge task we achieved significant gains with respect to SVR modeling. The proposed method can be used in low-latency settings and is already real-time capable on a standard PC. Yet, future work will concentrate on effective combination with feature selection to further decrease complexity. We will also investigate unsupervised pre-training using large amounts of unlabeled music data.

7. REFERENCES

[1] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proc. of SIGIR*, 2003, pp. 375–376.

[2] Y.H. Yang, Y.C. Lin, Y.F. Su, and 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.

[3] S. Rho, B.J. Han, and E. Hwang, "SVR-based music mood classification and context-based music recommendation," in *Proc. ACM Multimedia*, Beijing, China, 2009, pp. 713–716.

[4] B. Schuller, F. Weninger, and J. Dorfner, "Multi-Modal Non-Prototypical Music Mood Analysis in Continuous Space: Reliability and Performances," in *Proceedings 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, Miami, FL, October 2011, ISMIR, pp. 759–764, ISMIR, (acceptance rate: 59 %).

[5] E.M. Schmidt and Y.E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. of ISMIR*, 2010, pp. 465–470.

[6] E. Coutinho and A. Cangelosi, "A neural network model for the prediction of musical emotions," in *Advances in Cognitive Systems*, S. Nefti-Meziani and J. Grey, Eds., pp. 331–368. IET Publisher, London, UK, 2010, ISBN: 978-1849190756.

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

[8] E.M. Schmidt and Y.E. Kim, "Modeling musical emotion dynamics with conditional random fields," in *Proc. of ISMIR*, Miami, FL, USA, 2011, pp. 777–782.

[9] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal Onset Detection with Bidirectional Long-Short Term Memory Neural Networks," in *Proceedings 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, Utrecht, The Netherlands, 2010, pp. 589–594, ISMIR.

[10] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 121–124.

[11] 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.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "A Multi-Task Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech," *ACM Transactions on Interactive Intelligent Systems, Special Issue on Affective Interaction in Natural Environments*, vol. 2, no. 1, March 2012, 29 pages.

[13] N.N. Vempala and F.A. Russo, "Predicting emotion from music audio features using neural networks," in *Proc. of 9th International Symposium on Computer Music Modelling and Retrieval (CMMR)*, London, UK, 2012, pp. 336–343.

[14] E.M. Schmidt, J. Scott, and Y.E. Kim, "Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion," in *Proc. of ISMIR*, 2012, pp. 325–330.

[15] 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, 2013, pp. 148–152, ISCA.

[16] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, Technische Universität München, 2008.

[17] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[18] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of ICML*, 2008, pp. 1096–1103.

[19] 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, 2013, ACM, to appear.

[20] F. Weninger, F. Eyben, and B. Schuller, "The TUM approach to the MediaEval music emotion task using generic affective audio features," in *Proc. of MediaEval 2013 held in conjunction with ACM MM*, Barcelona, Spain, October 2013.

[21] 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.