FEATURE LEARNING IN DYNAMIC ENVIRONMENTS: MODELING THE ACOUSTIC STRUCTURE OF MUSICAL EMOTION

Erik M. Schmidt, Jeffrey Scott, and Youngmoo E. Kim
Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
{eschmidt,jjscott,ykim}@drexel.edu

ABSTRACT

While emotion-based music organization is a natural process for humans, quantifying it empirically proves to be a very difficult task, and as such no dominant feature representation for music emotion recognition has yet emerged. Much of the difficulty in developing emotion-based features is the ambiguity of the ground-truth. Even using the smallest time window, opinions about emotion are bound to vary and reflect some disagreement between listeners. In previous work, we have modeled human response labels to music in the arousal-valence (A-V) emotion space with time-varying stochastic distributions. Current methods for automatic detection of emotion in music seek performance increases by combining several feature domains (e.g. loudness, timbre, harmony, rhythm). Such work has focused largely in dimensionality reduction for minor classification performance gains, but has provided little insight into the relationship between audio and emotional associations. In this work, we seek to employ regression-based deep belief networks to learn features directly from magnitude spectra. Taking into account the dynamic nature of music, we investigate combining multiple timescales of aggregated magnitude spectra as a basis for feature learning.

1. INTRODUCTION

The problem of automated recognition of emotional content (mood) within music has been the subject of increasing attention among the music information retrieval (Music-IR) research community [1]. While there has been much progress in machine learning systems for estimating human emotional response to music, little progress has been made in terms of intuitive feature representations. Current methods generally focus on combining several feature domains (e.g. loudness, timbre, harmony, rhythm) and performing dimensionality reduction techniques to extract the most relevant informaiton. In many cases these methods have failed to provide enhanced classification performance, and they leave much to be desired in terms of understanding the complex relationship between emotional associations and acoustic content.

The Music Information Retrieval Evaluation eXchange (MIREX) [1] audio mood classification task provides an excellent illustration of this. Shown in Figure 1 is the performance of MIREX submissions for each year. The first year MIREX ran the task it received 9 submissions, and the best performing system achieved 61.50% performance on the 6-class problem using a feature space spanning 16-dimensions [2]. Each year the task has received a larger number of submissions, with exponentially larger feature libraries, but have failed to produce significant performance gains. Most recently, in 2010 the task received 36 submissions with the best system mining a 70-dimensional feature space, but achieved only 64.17% [3]. These results perhaps indicate that the data necessary for informing systems for this problem is not present in any current feature set.

Human judgments are necessary for deriving emotion labels and associations, but individual perceptions of the emotional content of a given song or musical excerpt are bound to vary and reflect some degree of disagreement between listeners. This lack of specificity presents significant challenges for developing informative feature representations for content-based music emotion prediction. In previous work we have investigated modeling emotional responses to music as both a singular point [4] as well as a stochastic distribution [5] over the arousal-valence (A-V) space of emotional affect. In this two dimensional representation valence indicates positive versus negative emotion and arousal reflects emotional intensity [6].

The ambiguous nature of musical emotion makes it an especially interesting problem for the application of feature learning. Using deep belief networks (DBNs) [7] we develop methods for learning emotion-based acoustic representations directly from magnitude spectra. In previous work, we have found these models to be powerful methods for generating reduced dimensionality representations of raw input spectra [8]. In that approach, we learned features directly from spectra at the 20msec rate of our windowed Short-Time Fourier Transform (STFT). In doing so, we provided a direct comparison between the DBN model for extracting features to standard acoustic representations such as MFCCs.

[1] http://www.music-ir.org/mirex

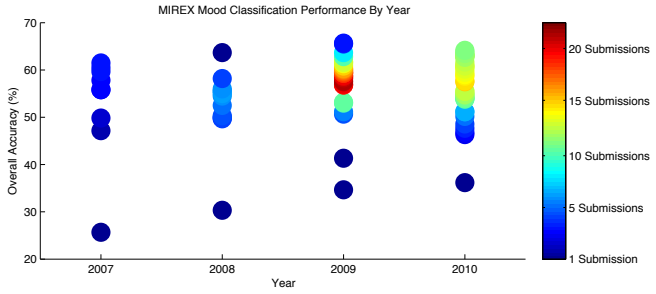

Figure 1. MIREX mood classification task performance by year.

But as humans require a larger time window than 20msec to determine emotions, and building upon that work we seek to improve our performance by informing our learned feature representations with spectra aggregated at multiple timescales. In addition, we investigate a universal background model (UBM) approach to feature learning. As DBN training follows an unsupervised pretraining, we investigate bootstrapping a much larger unlabeled dataset in developing our models. Given the challenges of collecting emotion annotated data, pretraining on a limited dataset is insufficient to form a general music model for finetuning. By bootstrapping a larger dataset we demonstrate significant improvement in using our DBN models for emotion prediction, and modest gains when using our learned features in a separate supervised machine learning approach.

We compare the learned feature representations to other state-of-the-art representations investigated in prior work [4, 5, 9]. In these experiments, we use the DBN hidden layer outputs as features to predict the training labels using a separate linear regression model. In all experiments we show that the features generated by the DBN outperform all other features, and that the topology is especially promising in providing insight into the relationship between acoustic data and emotional associations.

2. BACKGROUND

Feature learning has only recently gained attention in the machine listening community. Lee *et al.* was the first to apply deep belief networks to acoustic signals, employing an unsupervised convolutional approach [10]. Their system employed PCA to provide a dimensionality reduced representation of the magnitude spectrum as input to the DBN and showed slight improvement over MFCCs for speaker, gender, and phoneme detection.

Hamel and Eck applied deep belief networks (DBNs) to the problems of musical genre identification and autotagging [11]. Their approach used raw magnitude spectra as the input to their DBNs, which were constructed from three layers, employing fifty units at each layer. The system was trained using a greedy-wise pre-training and fine-tuned on a genre classification dataset, consisting of 1000 30-second clips. The system took 104 hours to train, and as a result was not cross-validated. Applied to a genre classification task, the learned features achieved a classification accuracy of 0.843, which was an increase over MFCCs at 0.790. The learned model was also used to inform an autotagging algorithm, which scored 0.73 in terms of mean accuracy, a slight improvement over MFCCs at 0.70.

3. GROUND TRUTH DATA COLLECTION

In prior work, we developed an online collaborative annotation activity based on the two-dimensional A-V model [12]. In this activity, participants use a graphical interface to indicate a dynamic position within the A-V space to annotate 30-second music clips. Each subject provides a check against the other, reducing the probability of nonsense labels. The song clips used are drawn from the "uspop2002" database. [2] Using initial game data, we constructed a corpus of 240 15-second music clips, which were selected to approximate an even distribution across the four primary quadrants of the A-V space.

In more recent work we have developed a Mechanical Turk (MTurk) activity to collect annotations on the same dataset [13]. The purpose of the MTurk activity was to provide a dataset collected through more traditional means to assess the effectiveness of the game, specifically to determine any biases created though collaborative labeling. Overall, the datasets were shown to be highly correlated, with arousal $r = 0.712$, and a valence $r = 0.846$. This new dataset is available to the research community, [3] and is densely annotated, containing $4,064$ label sequences in total, 16.93 ± 2.690 ratings per song. In this work we demonstrate the application of this densely annotated corpus for emotion-based feature learning.

4. ACOUSTIC FEATURE COLLECTION

Since our focus is on learning features that are specifically tuned to emotion prediction, we limit our comparisons to features that performed well in previous work. The features are also commonly used in the machine listening community and provide a reasonable baseline for testing. Our collection (Table 1) consists of the two highest performing features in prior work, Spectral Contrast and MFCCs [4,5], as well as the Echo Nest Timbre (ENT) features.

Feature	Description
Spectral Contrast [14]	Rough representation of the harmonic content in the frequency domain.
Mel-frequency cepstral coefficients (MFCCs) [15]	Low-dimensional representation of the spectrum warped according to the mel-scale. 20 dimensions used.
Echo Nest Timbre features (ENTs) [4]	Proprietary 12-dimensional beat-synchronous timbre feature

Table 1. Acoustic feature collection for music emotion prediction.

[2] http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html
[3] http://music.ece.drexel.edu/research/emotion/moodswingsturk
[4] http://developer.echonest.com

5. DEEP BELIEF NETWORKS

A fully trained deep belief network shares an identical topology to a neural network, though they offer a far-superior training procedure, which begins with an unsupervised pre-training that models the hidden layers as restricted Boltzman machines (RBMs) [7, 16, 17] . A graphical depiction of an RBM is shown in Figure 2. An RBM is a generative model that contains only a single hidden layer, and in simplistic terms they can be thought of two sets of basis vectors, one which reduces the dimensionality of the data and the other that reconstructs it.

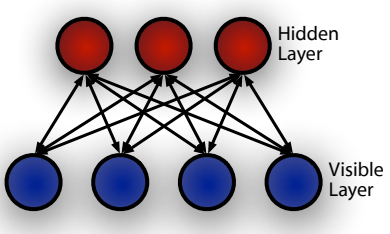

Figure 2. Restricted Boltzman machine topology.

RBMs are Markov random fields (MRFs) with hidden units, in a two layer architecture where we have visible units $\mathbf{v}$ and hidden units $\mathbf{h}$. This has an energy function of the form,

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in \text{freq}} b_i v_i - \sum_{j \in \text{features}} c_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (1)$$

where in this case the input is a spectrogram $\mathbf{v} \in \mathbb{R}^{1 \times I}$ and the hidden layer $\mathbf{h} \in \mathbb{R}^{1 \times J}$. The model has parameters $\mathbf{W} \in \mathbb{R}^{I \times J}$, with biases $\mathbf{c} \in \mathbb{R}^{1 \times J}$ and $\mathbf{v} \in \mathbb{R}^{1 \times I}$. During pre-training, we learn restricted Boltzman machines "greedily," where we learn them one at a time from the bottom up. That is, after we learn the first RBM we retain only the forward weights, and use them to create the input for training the next RBM layer.

As in the typical approach to deep learning, after pre-training we form a multi-layer perceptron using only the forward weights of the RBM layers. However, in typical approaches the final step is to attach logistic regression layer to the output of the MLP, and the full system is fine-tuned for classification using gradient descent. Since our goal is to learn feature detectors for a regression problem, we instead attach a simple linear regression layer and report the prediction error for fine-tuning as the mean squared error of the estimators. Squared error is chosen as opposed to Euclidean error for speed and numerical stability, as both functions have the same minimum. Furthermore, we elect to do our fine-tuning using conjugate gradient optimization, which we found to outperform gradient descent for our topology during initial testing.

We trained our DBNs using Theano, [5] a Python-based package for symbolic math compilation, and Scipy's optimization toolbox for the conjugate gradient optimization. Theano is an extremely powerful tool for machine learning problems because it combines the simplicity of Python with the power of compiled C, which can target the CPU or GPU.

[5] http://deeplearning.net/software/theano/

6. EXPERIMENTS AND RESULTS

In the following experiments we investigate employing deep belief networks for emotion-based acoustic feature learning. In all experiments, the model training is cross-validated 5 times, dividing the dataset into 50% training, 20% verification, and 30% testing. To avoid the well-known album-effect, we ensured that any songs that were recorded on the same album were either placed entirely in the training or testing set.

All learned features are then evaluated in the context of multiple linear regression (MLR), as we have investigated in prior work [4,5,18]. MLR provides extremely high computational efficiency, making it ideal for discriminating between relative usefulness of many feature domains.

6.1 Short-time Feature Learning

In the first set of experiments, we investigate learning features directly from short-time magnitude spectra. We have investigated this approach in prior work in the context of a different dataset [4], and we investigate it here to compare performance with the Turk dataset and to provide a baseline for our further work. As with our previous work, we use 3 hidden layers in all experiments, each containing 50 nodes. Furthermore, we run pre-training for 50 epochs with a learning rate of 0.001. During the conjugate gradient fine-tuning stage we attach an additional multiple linear regression (MLR) layer to the output of the DBN. As this stage is supervised, for each input example $\mathbf{x}_i$, we train the model to produce the emotion space parameter vector $\mathbf{y}_i$,

$$\mathbf{y}_i = [\mu_a,\ \mu_v]. \quad (2)$$

Shown in Table 2, are the results for employing the learned features for multiple linear regression. Features are first extracted on 20msec intervals, and then appropriately aggregated to match the one second intervals of our labels. Results for these features which are learned from single frames are shown as DBN-SF. We additionally show the KL-divergence for the Gaussian ground-truth representation used in prior work [5, 18]. Where we develop regressors to predict the parameterization vector $\mathbf{y}_i$ of a two-dimensional Gaussian in A-V space,

$$\mathbf{y}_i = [\mu_a,\ \mu_v,\ \sigma_{aa}^2,\ \sigma_{vv}^2,\ \sigma_{av}^2]. \quad (3)$$

6.2 Multi-frame Feature Learning

While future work on more sophisticated fine-tuning approaches or better stochastic models in pre-training may improve performance, the largest issue is the inherent limitation in using a single short-time window. Human emotional associations necessarily require more than a ∼20ms short-time window, and thus future approaches must take into account the variation of acoustic data over a larger period of time. In these experiments we investigate the development of models that incorporate multiple spectral

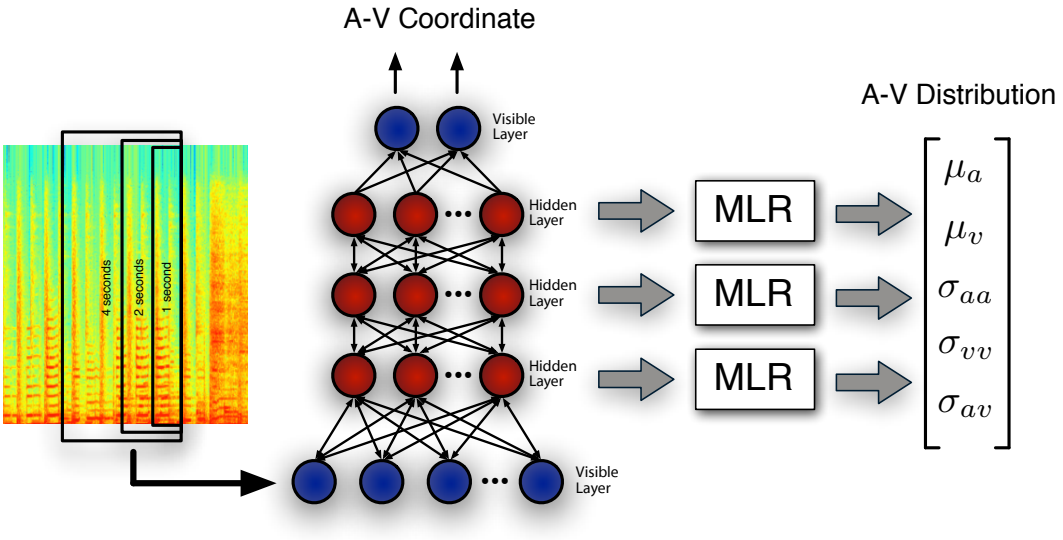

Figure 3. Feature learning system architecture showing the temporal aggregation, deep belief network and subsequent training of linear regressors to predict multi-dimensional A-V distributions.

windows to derive musical emotion. Taking spectral aggregations of the past one second, past two seconds, and past four seconds, we concatenate the resulting vectors as inputs to the system. As each spectrum frame is 257-dimensional vector, the total DBN input is now 771 dimensions. A diagram showing the multi-rate temporal integration, DBN training and linear regression to the emotion space is shown in Figure 3. Results for multi-frame (MF) feature learning can be found in Table 2 labeled as DBN-MF.

For this new approach, we provide visualizations of the learned features. Figure 4 shows the input spectrogram in log-magnitude for proper visualization, though we do not take the log for the actual model input.

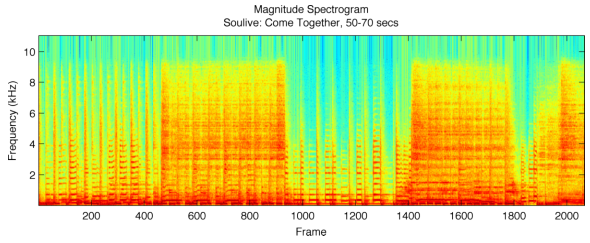

Figure 4. Log-magnitude spectrogram of input audio.

In the original spectrogram (Figure 4) we see the verse transition into the first chorus of the Soulive rendition of the Beatles song *Come Together* starting around frame 500. We see a similar pattern in the spectrogram between frames 1488-1800, which is the only other part of the clip where the percussion includes cymbals. Shown in Figure 5 are the resulting features from the intermediary layer outputs. Note that the structural information in the spectrogram is retained in the hidden layer outputs rendered in Figure 5.

We also wish to investigate the reconstruction of the original input aggregated spectrogram from the hidden layer outputs. Figure 6 depicts this reconstruction which

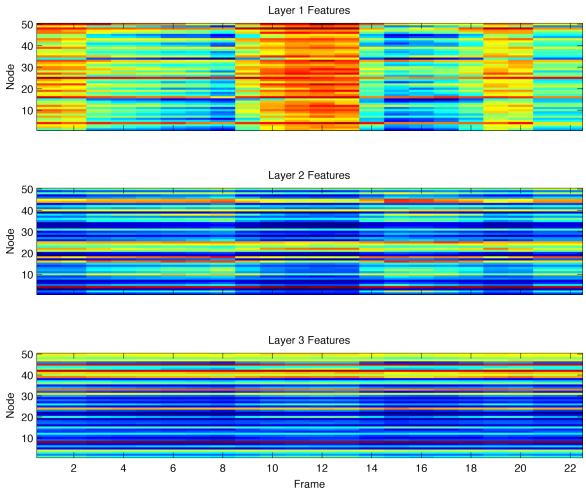

Figure 5. DBN hidden layer outputs using the aggregate spectral frames as input.

was generated using the method outlined in previous work [8]. Due to the concatenations of multiple time scale aggregations, we adjust the y-axis to display the correct frequency values for each. The top contains the last one second aggregations, below that is the aggregation from the last two seconds, and the last four seconds is at the bottom.

6.3 Universal Background Model Feature Learning

In order to improve our results with the multi-frame approach, we seek to harness the power of our much larger unlabeled music dataset. As DBN training relies on a two step training process, the first of which is unsupervised, there is no reason we should not use every piece of available data. In training our RBMs with our larger dataset, we get a much more accurate portrail of the distribution of music, and therefore create a much more accurate music model, which we can then finetune for musical emotion,

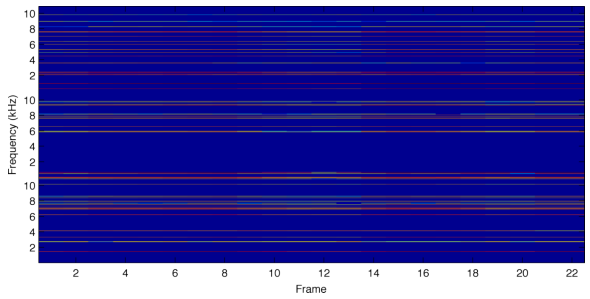

Figure 6. Reconstruction of the original aggregated spectrogram used as the DBN input. Top is last one second aggregates, middle last 2 seconds, bottom last 4 seconds.

or any other supervised machine learning problem. As this model is a general music model, we refer to it as a universal background model (UBM). For our larger dataset we use the uspop2002 dataset in its entirety, which contains nearly 8000 songs. Even after aggregating our spectra at one-second intervals this adds up to ∼26 GB of training data. Results for the universal background model approach are shown in Table 2 as DBN-UBM.

Feature Type	Average Mean Distance	Average KL Divergence
MFCC	0.140 ± 0.005	1.28 ± 0.157
Chroma	0.182 ± 0.006	3.33 ± 0.294
Spectral Shape	0.153 ± 0.006	1.51 ± 0.160
Spectral Contrast	$\mathbf{0.138 \pm 0.005}$	$\mathbf{1.29 \pm 0.160}$
ENT	0.151 ± 0.006	1.41 ± 0.175
DBN-SF Model Error	0.203 ± 0.009	-
DBN-SF Layer 1	0.138 ± 0.005	1.25 ± 0.142
DBN-SF Layer 2	$\mathbf{0.133 \pm 0.004}$	$\mathbf{1.19 \pm 0.129}$
DBN-SF Layer 3	0.133 ± 0.002	1.21 ± 0.180
DBN-MF Model Error	0.194 ± 0.032	-
DBN-MF Layer 1	0.131 ± 0.006	1.15 ± 0.106
DBN-MF Layer 2	0.131 ± 0.004	1.14 ± 0.107
DBN-MF Layer 3	$\mathbf{0.129 \pm 0.004}$	$\mathbf{1.12 \pm 0.114}$
DBN-UBM Model Error	0.140 ± 0.015	-
DBN-UBM Layer 1	0.129 ± 0.006	1.12 ± 0.091
DBN-UBM Layer 2	0.128 ± 0.004	1.13 ± 0.097
DBN-UBM Layer 3	$\mathbf{0.128 \pm 0.004}$	$\mathbf{1.11 \pm 0.090}$

Table 2. Emotion regression results for fifteen second clips. DBN-SF are features learned from single frames (SF), DBN-MF are features learned from multi-frame (MF) aggregations, and DBN-UBM are features learned with a universal background model (UBM) approach to DBN pretraining. KL-divergence is not applicable to model error.

7. DISCUSSION AND FUTURE WORK

In looking at the first set of results for learning features from single frames (DBN-SF), we see second layer features perform best for this method, outperforming spectral contrast, which is the best performing standard feature. This result is consistent with prior work [8], though here we find the DBN-SF features to be better than spectral contrast both in predicting single points and distributions. In that work we found the DBN features to be more accurate in terms of mean prediction and spectral contrast to perform slightly better in terms of KL, though we strongly emphasized an incorrect mean to be a much worse an error than an incorrectly sized or rotated covariance.

In trying to improve our features by including multiple timescales we see improvement in mean error from 0.133 to 0.129, which is encouraging. In analyzing the reconstructed spectra from first layer features, we get a very interesting result, which is similar to our prior work with [8]. The overall representation if very sparse in terms of frequency and seems to target very specific frequencies to contribute to the overall emotion features. Analyzing the features in Figure 5, we note that there is most definitely an emotion change as we progress from the slower and heavily minor sounding verse into the higher tempo rock chorus. We see changes reflected in all three layers' features in that area of the clip. We also note that it appears as if the features don't exactly line up with the spectrogram, which is a result of including past data in our feature computation. When the spectrum changes abruptly it takes several frames for our model to catch up. We do not see this as a limitation as humans have a reaction time too, which perhaps is reflected in the fact that these features are better suited for time-varying emotion prediction.

At a normalized error of 0.128, our simple MLR method with DBN features outperforms two of the three features investigated in our prior work with conditional random fields (CRFs) [19], which is a much more sophisticated method. Furthermore, while the performance increase between the third layer features of DBN-MF and DBN-UMB is small, the performance of the DBN model itself is reduced from 0.194 to 0.140, which we find to be highly encouraging. These results indicate that UBM pretraining is providing us a model that is much better suited for emotion finetuning.

In future work, we plan to investigate shrinking layer sizes in the UBM approach where we can perhaps take better advantage of the dimensionality reduction power of the RBM. Furthermore, we see that it may be interesting to investigate multiple stages of finetuning. We would first follow the approach of [7] for reducing the dimensionality of unlabeled data. It may be possible to gain a more accurate UBM by applying a finetuning stage that involved unraveling the model to reconstruct the unlabeled data. Those model parameters could then be adapted to emotion, or any other type of music prediction.

8. ACKNOWLEDGMENT

This work is supported by National Science Foundation award IIS-0644151.

9. REFERENCES

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull,

"Music emotion recognition: A state of the art review," in *ISMIR*, Utrecht, Netherlands, 2010.

[2] G. Tzanetakis, "Marsyas submissions to MIREX 2007," MIREX 2007.

[3] J.-C. Wang, H.-Y. Lo, S.-K. Jeng, and H.-M. Wang, "Mirex 2010: Audio classification using semantic transformation and classifier ensemble," in *MIREX*, 2010.

[4] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *ACM MIR*, Philadelphia, PA, 2010.

[5] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *ISMIR*, Utrecht, Netherlands, 2010.

[6] J. A. Russell, "A complex model of affect," *J. Personality Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[7] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[8] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *WASPAA*, New Paltz, NY, 2011.

[9] E. M. Schmidt, M. Prochup, J. Scott, B. Dolhansky, B. G. Morton, and Y. E. Kim, "Relating perceptual and feature space invariances in music emotion recognition," in *CMMR*, London, U.K., 2012.

[10] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *NIPS*. MIT Press, 2009.

[11] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in *ISMIR*, Utrecht, Netherlands, 2010.

[12] Y. E. Kim, E. Schmidt, and L. Emelle, "MoodSwings: A collaborative game for music mood label collection," in *ISMIR*, Philadelphia, PA, September 2008.

[13] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim, "A comparative study of collaborative vs. traditional annotation methods," in *ISMIR*, Miami, Florida, 2011.

[14] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai, "Music type classification by spectral contrast feature," in *Proc. Intl. Conf. on Multimedia and Expo*, vol. 1, 2002, pp. 113–116.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[16] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *NIPS*. MIT Press, 2007.

[18] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using Kalman filtering," in *Proc. of the 9th IEEE Intl. Conf. on Machine Learning and Applications (ICMLA)*, Washington, D.C., 2010.

[19] ——, "Modeling musical emotion dynamics with conditional random fields," in *ISMIR*, Miami, FL, 2011.