MODELING MUSICAL EMOTION DYNAMICS WITH CONDITIONAL RANDOM FIELDS

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

ABSTRACT

Human emotion responses to music are dynamic processes that evolve naturally over time in synchrony with the music. It is because of this dynamic nature that systems which seek to predict emotion in music must necessarily analyze such processes on short-time intervals, modeling not just the relationships between acoustic data and emotion parameters, but how those relationships evolve *over time*. In this work we seek to model such relationships using a conditional random field (CRF), a powerful graphical model which is trained to predict the conditional probability $p(\mathbf{y}|\mathbf{x})$ for a sequence of labels $\mathbf{y}$ given a sequence of features $\mathbf{x}$. Treating our features as deterministic, we retain the rich local subtleties present in the data, which is especially applicable to content-based audio analysis, given the abundance of data in these problems. We train our graphical model on the emotional responses of individual annotators in an 11×11 quantized representation of the arousal-valence (A-V) space. Our model is fully connected, and can produce estimates of the conditional probability for each A-V bin, allowing us to easily model complex emotion-space *distributions* (e.g. multimodal) as an A-V heatmap.

1. INTRODUCTION

The development of content-based systems for the prediction of emotion (mood) in music continues to be a topic of increasing attention in the Music-IR community, but thus far most approaches apply only a singular rating to a song or clip [1]. Such generalizations belie the time-varying nature of music and make emotion-based recommendation difficult, as it is very common for emotion to vary temporally throughout a song. In this work, we investigate the application of conditional random fields (CRFs) to the modeling of time-varying musical emotion. CRFs are powerful graphical models which are trained to predict the conditional probability $p(\mathbf{y}|\mathbf{x})$ for a sequence of labels $\mathbf{y}$ given a sequence of features $\mathbf{x}$. Treating our features as deterministic, we retain the rich local subtleties present in the data, which is especially promising in content-based audio analysis where there is no shortage of rich data. Furthermore, the system provides a model of both the relationships between acoustic data and emotion space parameters and also how those relationships evolve over time.

Human judgements are necessary for deriving emotion labels and associations, but perceptions of the emotional content of a given song or musical excerpt are bound to vary and reflect some degree of disagreement between listeners. Following from our previous work, we model human emotion responses to music in the arousal-valence (A-V) representation of emotion [2–4], where valence indicates positive vs. negative emotions and arousal reflects emotional intensity [5]. In our prior approaches, we modeled our emotion space distribution as a single two-dimensional Gaussian distribution, and trained multivariate regression systems to predict the parameters of the distribution directly from acoustic features [3, 4]. Using that representation, we found modeling the dynamics of the continuous parameter space to be a very challenging problem. We considered a Kalman filtering approach, but while this technique provided smooth estimates over time, the limited model complexity was unable to cover a wide variance in emotion space dynamics [4].

In applying CRFs to the problem of predicting emotion in music, instead of modeling the ambiguity of emotion *a-priori* and representing the distribution of our emotion space parameters as the ground truth, we present the training algorithm with the individual user label sequences, thus allowing the model to learn the range of emotion responses to a given piece. In our application of the CRF we must also assign emotion space meanings to the states of the model, and in doing so we discretize each label in our sequences to an 11×11 grid. While this is a significant simplification, our findings indicate that it provides sufficient granularity. Furthermore, our trained models are fully connected, and can

be used to model complex distributions in emotion as an A-V heatmap. These heatmaps can model arbitrary modes and distributions, in contrast to our previous approach, which constructed uni-modal Gaussian A-V predictions.

2. BACKGROUND

The general approach to implementing automatic mood detection from audio has been to use supervised machine learning to train statistical models based on acoustic features [1]. Chan *et al.* recently investigated modeling emotion as a distribution [6]. Their approach investigated modeling the ground truth as a Gaussian distribution as well as a heatmap and used support vector regression for the distribution prediction. However, their corpus was limited to only 60 songs, and the work only focused on applying a singular rating to an entire clip.

Conditional random fields have only just begun to gain attention as a tool for content-based audio prediction. Recently, Joder *et al.* successfully applied them to the task of audio-to-score matching, detecting more than 95% of the note onset locations to within 100 ms [7].

3. GROUND TRUTH DATA COLLECTION

In prior work, we developed an online collaborative annotation activity based on the two-dimensional A-V model [8]. In the activity, participants used a graphical interface to indicate a dynamic position within the A-V space to annotate 30-second music clips. Each subject provided a check against the other, reducing the probability of nonsense labels. The song clips used were drawn from the "uspop2002" database. [1] 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 [9]. The purpose of the MTurk activity was to provide a dataset collected through more traditional means to assess the effectiveness of the game to determine any biases induced 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 has been made available to the research community, [2] and is well annotated, containing 16.93 ± 2.690 ratings per song and $4,064$ label sequences. In this work we demonstrate the application of this densely annotated corpus for training our conditional random fields.

[1] http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

[2] http://music.ece.drexel.edu/research/emotion/moodswingsturk

3.1 Statistical Analysis

In applying relational learning methods to data, we gain the ability to model statistical dependencies from one observation to the next. To verify that our data collection exhibits such dependencies, we compute the correlation coefficients of our label sequences from one frame to the next and from the first frame of each sequence to the last. In these cases, we treat the individual discretized user labels as variables, and each second as observations of those variables. Statistics of the squared correlation coefficients (r^2) are provided for the full dataset in Table 1.

Dimension	r^2 Frame-Frame	r^2 First-Last Frame
Arousal	0.944 ± 0.093	0.507 ± 0.242
Valence	0.951 ± 0.097	0.524 ± 0.235

Table 1. Statistics of ground truth squared correlation coefficient (r^2) from one second to the next and from the first second to the last.

Overall, the dataset shows high correlation from one frame to the next, and lower correlation between the first frame and last frame. In other words, the current emotion is highly dependent upon the emotion of the prior second, and on average each sequence exhibits a significant change in emotion from beginning to end. As a result, the dataset is a good match for graphical modeling techniques.

4. ACOUSTIC FEATURE COLLECTION

In previous work we have found there to be no single dominant feature, but rather many that play a role (e.g., loudness, timbre, harmony) in determining the emotional content of music [2, 3]. Since our experiments focus on the tracking of emotion over time, we chose to focus solely on time-varying features. Our collection (Table 2) consists of the two highest performing features in prior work, Spectral Contrast and MFCCs [2, 3], as well as the Echo Nest Timbre (ENT) features.

Feature	Description
Spectral Contrast [10]	Rough representation of the harmonic content in the frequency domain.
Mel-frequency cepstral coefficients (MFCCs) [11]	Low-dimensional representation of the spectrum warped according to the mel-scale. 20 dimensions used.
Echo Nest Timbre features (ENTs) [3]	Proprietary 12-dimensional beat-synchronous timbre feature

Table 2. Acoustic feature collection for music emotion prediction.

[3] http://developer.echonest.com

ENTs have been receiving significant attention lately due to the release of the million song dataset, [4] and we therefore investigate their utility in musical emotion prediction.

5. CONDITIONAL RANDOM FIELDS

In this section we give a brief overview of conditional random fields (CRFs), mainly focused on practical considerations in implementation. The interested reader is directed to [12, 13] for further details.

5.1 Overview

Traditional approaches for graphical modeling (e.g. hidden Markov models) seek to represent the joint probability $p(\mathbf{x}, \mathbf{y})$ between sets of features $\mathbf{x}$ and labels $\mathbf{y}$. But in forcing our features into a generative model $p(\mathbf{x})$ we discard the rich local subtleties present in the data. Furthermore, in developing models for audio classification tasks, our acoustic features are naturally deterministic. With CRFs, as with logistic regression, we seek to model the conditional probability $p(\mathbf{y}|\mathbf{x})$.

CRFs are trained on sequences, and in the process of learning them we present the classification system with the individual user ratings (as opposed to statistics of all users) recorded in the MTurk task. Using a fully connected model, we are able to learn a set of transition probabilities from each class to all others. This means that at each stage in a testing sequence we can display the transition probabilities in the form of a heatmap as shown in Figure 1.

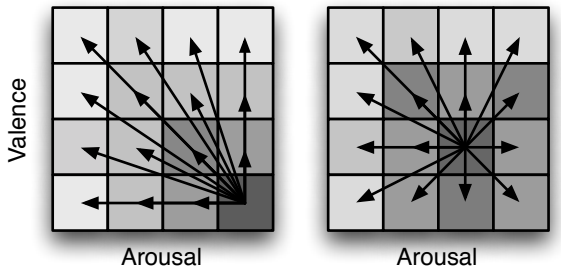

Figure 1. Heatmap visualization of CRF transition probabilities. Actual discretization is 11×11.

5.2 Feature Functions

CRFs require the specification of feature functions, which are used to specify the degree of compatibility between the features $\mathbf{x}$ and labels $\mathbf{y}$. These functions are defined over all examples, and for a single example are non-zero only for the labeled class. We train our CRFs using CRF++, [5] a highly efficient general purpose CRF toolkit written in C++. CRF++ allows the definition of both unigram and bigram features, where unigram features are related to the prediction of a single observation in a sequence (first order Markov) and bigram features are related to the prediction of pairs of observations (second order Markov). Unigram features generate a total of L×N distinct features, where L is the number of output classes and N is the number of unique features. Bigram features generate L×L×N distinct features.

[4] http://labrosa.ee.columbia.edu/millionsong/

[5] http://crfpp.sourceforge.net/

6. EXPERIMENTS AND RESULTS

In the following experiments, we investigate the use of conditional random fields for the prediction of musical emotion. As a baseline for comparing performance of the CRF in modeling the time-dependencies of our data, we additionally provide the performance for the CRF when trained on independent observations as opposed to sequences. Furthermore, to provide a baseline for comparison to our prior work [3, 4], we provide the prediction accuracy of multiple linear regression (MLR). To compute the heatmap representations for MLR, we first predict the mean and covariance of an emotion-space Gaussian density using multivariate regression, and then integrate the probability density function under each square of our heatmap.

In all experiments, to avoid the well-known "album-effect," we ensure that any songs which were recorded on the same album are either placed entirely in the training or testing set. Additionally, each experiment is subject to 5 cross-validations, varying the distribution of training and testing data sets which are split 70%/30%, respectively.

6.1 Acoustic Feature Representation

All features are initially computed using short-time analysis windows at a much higher rate than our 1-second emotion label windows. In order to reduce their frame rate to that of the labels, spectral contrast and MFCCs are simply re-windowed via averaging from their original analysis rate ($\sim$ 23 msec). The ENTs are re-windowed following their non-linear analysis frame start times to take into account their beat-synchronous nature.

Additionally, conditional random fields are highly optimized to operate on binary features, and given the high dimensionality of our data, we found it necessary to convert our features to such a representation. In doing so, each feature dimension is quantized using 10 equal energy bins, which for the 14-dimensional case of spectral contrast yields 140 binary features. In early experiments, we investigated the use of higher discretization levels as well as combining representations from multiple discretization levels (e.g. 5, 10, 20), but overall found 10 levels to offer the best performance.

6.2 Training Sequence Label Jittering

In discretizing our original label sequences to the 11×11 grid representation, our CRF models are trained on vectorized version of that space by assigning 121 classes. As a result, the neighbor-relationship of the heatmap grid-cells is lost in the vector-wise representation, and we therefore investigate how to improve the models ability to learn such relationships.

In order to ensure that the CRF learns the spatial relationships of each class, we train it on additional "jittered" versions of each label sequence. This has two benefits: it increases the overall size of our dataset, and it helps the model to learn the spatial relationships between the different classes. In applying our jitter we increase the size of our dataset by a factor of 10, creating 9 additional sequences for each sequence in our dataset. Each jittered sequence is created by adding a small amount of zero mean Gaussian noise, biasing the whole sequence by a single point. In initial experiments we modified the number of jittered sequences at multiple levels between 0 and 50, but found 10 to offer the best performance.

6.3 CRF Parameterization

As previously discussed, the training of CRFs requires the selection of feature functions. In our experiments, we elect to use three different types of features: a simple unigram node feature for each acoustic feature dimension, a unigram edge feature that models the change in each feature dimension between nodes, and a simple bigram (second order) feature that models the joint probability of the next two states for arbitrary input. The total number of binary CRF features for a selected training set is described in Table 3.

Additionally, in the case of the CRF trained on independent observations, we remove all but node features, so as to avoid an artificial decrease in performance. When presenting the training algorithm with independent examples instead of sequences, feature functions that encode time dependencies that cannot be modeled lead to large decreases in performance.

The training of graphical models such as CRFs tends to have a very high computational cost. We ran our experiments on Amazon's Elastic Compute Cloud (EC2) [6] using High-CPU Extra Large Instances (c1.xlarge) which provide access to a 64-bit platform with 8 virtual cores. Shown in Table 3 is the computation time for each feature domain the CRF was trained on as well as the number of binary features created using the specified feature functions.

[6] http://aws.amazon.com/ec2/

Feature	# CRF Features	Compute Time (hrs)
Contrast	$210,782$	11.49 ± 1.245
MFCC	$300,927$	11.81 ± 1.515
ENT	$185,009$	12.04 ± 0.461

Table 3. Computing time analysis for CRF training on each cross-validation set.

6.4 Evaluating CRF Performance

We begin our analysis by attempting to predict a singular A-V point at each second in our sequences. These predictions are taken as the means of the CRF heatmaps, which we compare to the means of the MLR Gaussian distributions. In the second stage of analysis we investigate the accuracy of the CRF heatmaps, which we compare to MLR Gaussian heatmaps.

6.4.1 A-V Mean Prediction

We compute the heatmap mean as the sum of the weighted A-V coordinate values of each bin center. For each two-dimensional heatmap we compute,

$$\mu_a = \sum_{y_a, y_v} P(y_a, y_v \,|\, x)\; y_a,$$
$$\mu_v = \sum_{y_a, y_v} P(y_a, y_v \,|\, x)\; y_v. \qquad (1)$$

where y_a and y_v are the arousal and valence coordinates of each bin center. The mean values for the ground truth distribution are computed directly in the continuous A-V space. These results are available in the third column of Table 4. Overall we see the best performance (minimum mean ℓ^2 error) of 0.122 using the CRF with MFCCs, which is significantly improved over the best result with MLR, which is spectral contrast at 0.140.

6.4.2 Heatmap Prediction Evaluation

As previously stated, because the CRF is a fully connected model, we can use the transition probabilities to construct an A-V heatmap. But the ground truth heatmap must be estimated empirically as a two dimensional histogram, which is a difficult task. In traditional generative estimation the goal is to fit a probabilistic model to data, and derive a smooth function, even with a small dataset. But with histograms, a small amount of data can lead to sparse, blocky estimates, and a massive amount of data is needed to achieve the true smooth distribution.

As a result of this we have chosen the earth mover's distance (EMD) [14] to be our primary metric for comparing these histograms, which can be thought of as the minimum cost of transforming one heatmap into the other. Using this metric we can take into account the weight of adjacent bins, which overall provides a more accurate comparison of the

Acoustic Feature	Prediction Method	A-V Mean ℓ^2 Error	Heatmap Earth Mover's Distance	Heatmap Error Unsmoothed ($\times 10^{-2}$)	Heatmap Error Smoothed G.T. ($\times 10^{-2}$)	Heatmap Error Smoothed ($\times 10^{-2}$)
Contrast	CRF	0.130 ± 0.007	0.180 ± 0.007	1.300 ± 0.007	0.539 ± 0.002	0.342 ± 0.0142
MFCC	CRF	$\mathbf{0.122 \pm 0.004}$	$\mathbf{0.173 \pm 0.004}$	$\mathbf{1.300 \pm 0.000}$	$\mathbf{0.541 \pm 0.010}$	$\mathbf{0.326 \pm 0.008}$
ENT	CRF	0.130 ± 0.004	0.179 ± 0.003	1.300 ± 0.009	0.510 ± 0.010	0.337 ± 0.009
Contrast	CRF-I	0.138 ± 0.006	0.188 ± 0.005	1.323 ± 0.007	0.452 ± 0.012	0.355 ± 0.011
MFCC	CRF-I	$\mathbf{0.135 \pm 0.004}$	$\mathbf{0.186 \pm 0.003}$	$\mathbf{1.319 \pm 0.006}$	$\mathbf{0.459 \pm 0.007}$	$\mathbf{0.350 \pm 0.008}$
ENT	CRF-I	0.144 ± 0.005	0.194 ± 0.004	1.331 ± 0.005	0.446 ± 0.007	0.367 ± 0.009
Contrast	MLR	$\mathbf{0.140 \pm 0.005}$	$\mathbf{0.213 \pm 0.009}$	$\mathbf{1.082 \pm 0.010}$	$\mathbf{0.580 \pm 0.018}$	$\mathbf{0.460 \pm 0.018}$
MFCC	MLR	0.141 ± 0.005	0.208 ± 0.008	1.076 ± 0.009	0.570 ± 0.021	0.448 ± 0.021
ENT	MLR	0.153 ± 0.005	0.204 ± 0.007	1.068 ± 0.009	0.560 ± 0.018	0.440 ± 0.018

Table 4. Emotion prediction results for conditional random fields (CRF) trained on sequence examples as well as independent examples (CRF-I). Multiple linear regression (MLR) provided as baseline.

two heatmaps. These results are in the fourth column of Table 4, where we find the CRF to be the best performer with an EMD of 0.173, which is significantly better than the CRF trained on independent samples at 0.186 and MLR at 0.213.

But we also investigate the absolute pixel error between the predicted and ground truth heatmaps. These results are shown in the fifth column of Table 4, and we find that MLR appears to be performing slightly better than the CRF. This result is not surprising given the sparsity that our ground truth heatmaps exhibit, which is to be expected with 121 histogram bins computed from an average of 16.93 ratings. The MLR method which predicts Gaussian distributions guarantees a smooth distribution, which will produce a lower pixel error if the ground truth is sparse or blocky than the CRF which takes arbitrary shapes. But it can be easily demonstrated that the CRF is more accurate by applying a simple smoothing function to the ground truth.

To smooth out the blocking artifacts from sparsity we apply a simple 2-d Gaussian filter. This process applies a light smoothing without altering the mean of the data. These results are shown in the sixth column of Table 4. Here the CRF performs slightly better, and the performance similarity is most likely because the CRF is producing rough edges compared to the smooth MLR predictions that are computed from the Gaussian PDF. An interesting result is that the independently learned CRFs perform the best here. This is most likely because they produce more uniform transition probabilities due to their training method.

To compensate for blocking artifacts in the CRF predictions, we apply a smoothing filter to them as well. Initial experiments showed applying the same filter to the MLR heatmaps improved performance there too, so to keep our analysis consistent we apply the filter them as well. We examine the differences in heatmaps using mean absolute error, and these results are shown in the seventh column of Table 4. In these results we see again that the CRF is performing significantly better than MLR.

6.4.3 Visualizing the Results

Shown in Figure 2 are the CRF heatmap predictions for eight seconds of the song "Something About You," by Boston. The colormap of these heatmaps assigns red to areas of high density, blue to low, and uses the color spectrum to assign colors in between. This clip was selected because of the large change in emotion that occurs at second 29, where the song transitions from a low-energy, negative-emotion introduction into a high-energy, positive-emotion hard-rock verse. The system tracks the transition very accurately, showing a brief amount of uncertainty at second 30 in terms of positive or negative emotion, and finally settles on positive emotion at second 31. Prediction videos using the system are also available online. [7]

7. DISCUSSION AND FUTURE WORK

We have demonstrated conditional random fields to be a powerful tool for modeling time-varying musical emotion. The CRF approach is shown to be superior to MLR both at predicting single A-V mean values as well as full emotion space heatmaps. Overall, the best performing feature for CRF prediction is MFCCs, which differs from our MLR method, where spectral contrast performs best. This perhaps indicates that there is more information to be gained out of MFCCs when modeling the temporal evolution of emotion.

Using the earth mover's distance we are able to better analyze the similarity between heatmaps by also taking into account adjacent bin densities. While the MLR method appears to perform slightly higher when the ground truth distributions are not smoothed, this is a result of blocking artifacts in the ground truth. The the Gaussian density is a smooth function, which is much more likely to be similar to a sparse ground truth distribution than the CRF predictions, which take on arbitrary shapes and are not necessarily

[7] http://music.ece.drexel.edu/research/emotion

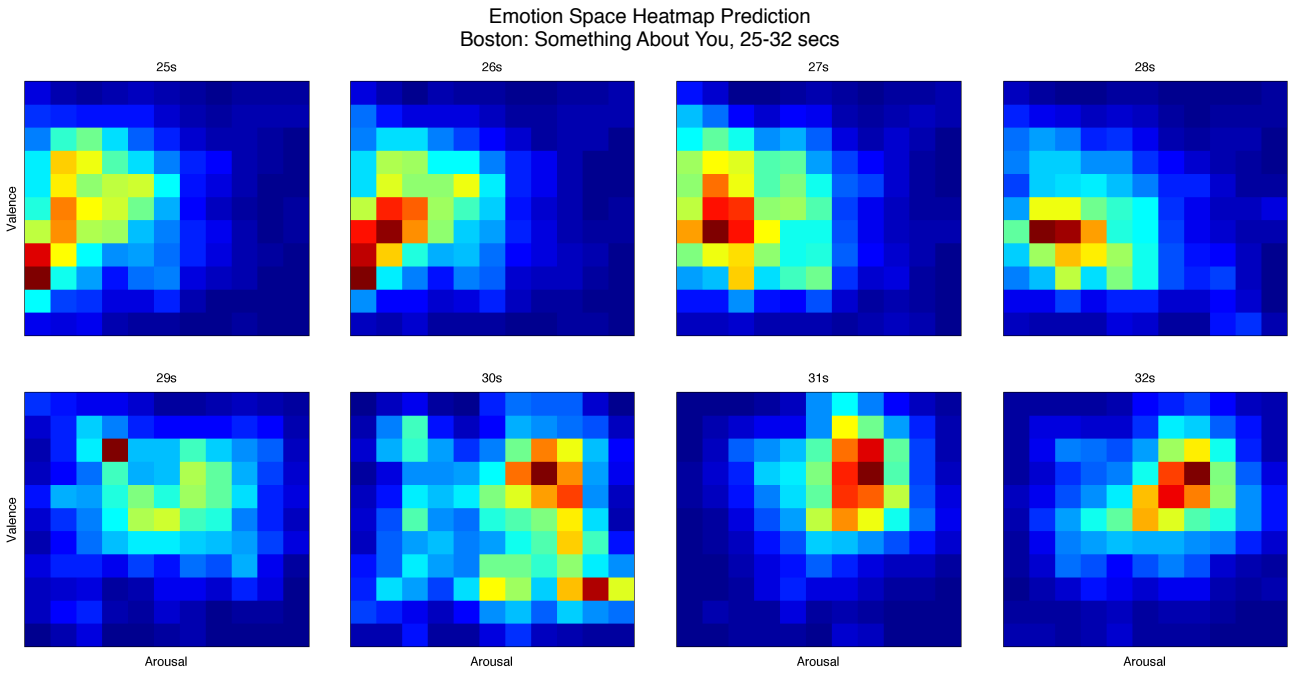

Figure 2. Emotion space heatmap prediction using conditional random fields. Shown is the predicted emotion from the beginning of the song "Something About You," by Boston. These figures demonstrate the system tracking the emotion through the low-energy, negative-emotion introduction, and through the transition at second 29 into a high-energy, positive emotion rock verse. In these figures, red indicates the highest density and blue is the lowest.

as smooth. Overall, the ground truth representation could significantly benefit from more data.

In a future approach, the CRF performance could be improved by developing a model which can encapsulate the A-V spatial relationships between CRF nodes, which could potentially produce smoother estimates without any need for label jittering. In such a model, we could also limit the connections between local heatmap pixels, thus allowing us the ability to tradeoff model complexity for the flexibility of our emotion space distribution flexibility.

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] 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.

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

[4] ——, "Prediction of time-varying musical mood distributions using Kalman filtering," in *IEEE ICMLA*, Washinton, D.C., 2010.

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

[6] H. Chen and Y. Yang, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE TASLP*, no. 99, 2011.

[7] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE TASLP*, no. 99, 2011.

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

[9] 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.

[10] 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*, 2002.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE TASSP*, vol. 28, no. 4, 1980.

[12] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2007, ch. 4, pp. 93–127.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML*, 2001.

[14] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *ICCV*, 2009.