Modeling and Predicting Emotion in Music

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 observed music signal. It is because of this dynamic nature that systems that 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 also how those relationships evolve *over time*. In this work, we discuss modeling such relationships using a conditional random field (CRF), a powerful graphical model that is trained to predict the conditional probability $p(\mathbf{y}|\mathbf{x})$ for a sequence of labels $\mathbf{y}$ given a sequence of features $\mathbf{x}$. 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. In selecting acoustic features for music emotion recognition, we discuss the application of regression-based deep belief networks (DBNs) to learn features directly from magnitude spectra. These features are specifically optimized for the prediction of emotion, and the trained models can potentially provide new insight into the relationships between music and emotion.**

I. Introduction

The medium of music has evolved specifically for the expression of emotions, and it is natural for us to organize music in terms of its emotional associations. But while such organization is a natural process for humans, quantifying it empirically proves to be a very difficult task. Myriad features, such as harmony, timbre, interpretation, and lyrics affect emotion, and the mood of a piece may also change over its duration. But in developing automated systems to organize music in terms of emotional content, we are faced with a problem that oftentimes lacks a well-defined answer; there may be considerable disagreement regarding the perception and interpretation of the emotions of a song or ambiguity within the piece itself. When compared to other music information retrieval tasks (e.g., genre identification), the identification of musical mood is still in its early stages, though it has received increasing attention among the music information retrieval (Music-IR) research community in recent years [1].

Collecting human judgements is necessary for deriving emotion labels and associations, which presents difficulty given the variability of perceptions between listeners that evolve over time in synchrony with the music. In collecting this data, we have investigated approaches using both serious games and Amazon's Mechanical Turk (MTurk) [2], [3]. In both approaches we designed activities to collect second-by-second labels of music using the two-dimensional, arousal-valence (A-V) model of human emotion, where valence indicates positive vs. negative emotions and arousal reflects emotional intensity [4]. This representation provides quantitative labels that are well-suited to computational methods for parameter estimation. The activities were designed specifically to capture A-V labels dynamically (over time) to reflect emotion changes in synchrony with music and also to collect a distribution of labels across multiple players for a given song or even a moment within a song. In our serious games approach, we created MoodSwings, a two-player online collaborative activity, wherein each subject provided a check against the other, reducing the probability of nonsense labels [2]. To investigate any biases in the data due to collaborative labeling, we also investigated annotating the same corpus with a more traditional single paid annotator approach using MTurk [3]. In this approach, we collected labels highly correlated with those collected in the game, essentially finding no biases due to its collaborative nature. Furthermore, we also found the ability to collect data at a much higher rate than the game, but at the cost of dealing with a large increase in noise due to the monetary incentives.

No dominant feature representation for music emotion recognition has yet emerged. Current methods typically focus on combining several feature domains (e.g. loudness, timbre, harmony, rhythm), oftentimes as many as possible, followed by feature selection and dimensionality reduction techniques. While these methods can lead to enhanced classification performance, they leave much to be desired in terms of understanding the complex relationship between acoustic content and emotional associations. In this talk, we will discuss using deep belief networks (DBNs) to learn representations of music audio that are specifically optimized for the prediction of emotion [5], [6].

In previous work, we have investigated modeling emotional responses to music as a time-varying *stochastic distribution* [7] within the A-V model of human emotions. Such models are inflexible to A-V distribution variation across multiple songs, and we have therefore also investigated *heatmap* representations [8]. To obtain A-V heatmaps, we model the relationships between acoustic parameters and emotion space classes 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 $\mathbf{11 \times 11}$ 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.

II. Data Collection Methods

In our serious games approach, we designed MoodSwings, a collaborative online game that leverages crowdsourcing to collect mood ratings [2]. The game board is based on the A-V space, where the valence dimension represents positive versus negative emotions and arousal represents high versus low energy [4]. Anonymously-partnered players label song clips together during each round, scoring points based on the overlap between their cursors, which encourages consensus. Bonus points are awarded to a player whose partner moves towards him/her, encouraging competition and discouraging players from blindly following their partners to score points. We recently initiated a redesign effort, investigating gameplay improvements suggested by an analysis of collected labels [9]. However, we have not addressed concerns about the game structure biasing annotations.

In order to investigate biases and potentially faster data collection, we designed a simplified labeling task for MTurk, shown in Figure 1. Single workers provide A-V labels for clips from our dataset, consisting of 240 15-second clips, which are extended to 30 seconds to give workers additional annotation practice [10]. As in MoodSwings, we collect per-second labels, but no partner is present and no points are awarded. Workers are given detailed instructions describing the A-V space. They navigate to a website that hosts the task and label 11 randomly-chosen clips. The first clip is a practice round, omitted from our analysis. The third and ninth are identical, randomly chosen from a set of 10 "verification clips," which are evaluated to identify unsatisfactory work. Workers are given a 6-digit verification code to enter on the MTurk website as proof of completion, which, if successful, earns workers $0.25 per HIT. This new dataset has been made available to the research community,[1] and is well annotated, containing 16.93 ± 2.690 ratings per song and $4,064$ label sequences.

III. Feature Learning with Deep Belief Networks

The ambiguous nature of musical emotion makes it an especially interesting problem for the application of feature learning. Using deep belief networks (DBNs) [11], [12], [13], we develop methods for the learning of emotion-based acoustic representations directly from magnitude spectra. The topology of a trained DBN is identical to that of a multi-layer perceptron (MLP) or neural network, but DBNs employ a far superior training procedure involving a secondary topology, which is later removed. DBN training begins with an unsupervised pre-training approach using greedily-trained restricted Boltzman

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

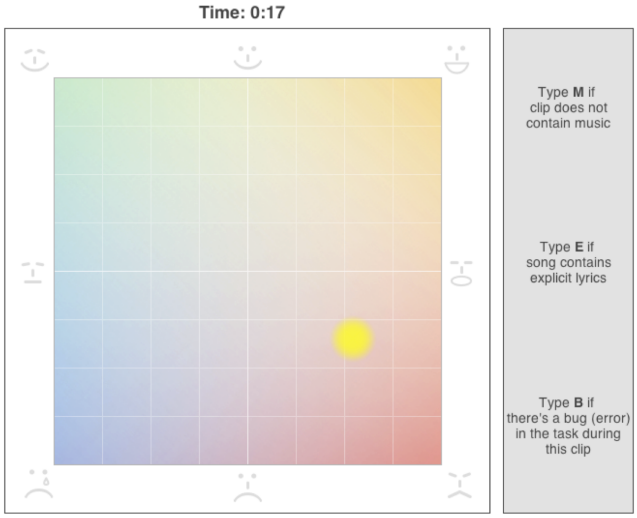

Fig. 1. Screenshot of labeling task deployed on MTurk, depicting the A-V space and a yellow orb as the annotator's cursor. A sidebar provides additional instructions, e.g. workers may type "B" if they encounter bugs in the task.

machines (RBMs) [11], [12], [13]. The general approach is to attach a logistic regression layer for classification after pre-training and to then use gradient descent to perform the fine-tuning. In this approach, we implement the DBN to learn feature detectors for a regression problem and instead attach a linear regression layer. Our approach uses conjugate gradient fine-tuning, which we found to provide more accurate feature detectors for our regression problem.

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 that reduces the dimensionality of the data and the other that reconstructs it.

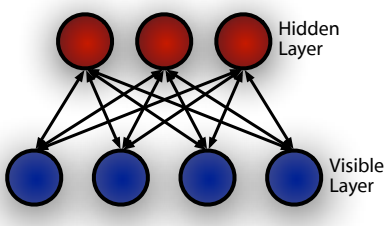

Fig. 2. Restricted Boltzman machine topology.

During RBM learning the individual node connections are treated as a Markov chain, and the all of the units in one layer are updated in parallel given the states of the other layer; this is repeated until the system reaches equilibrium. We train our RBMs using contrastive divergence, which runs the Markov chain for n full steps before computing the correlation between hidden and visible layer, and can be seen as minimizing the difference of two KL-divergences,

$$\mathrm{KL}(P^0||P_\theta^\infty) - \mathrm{KL}(P_\theta^n||P_\theta^\infty) \quad (1)$$

where P^0 is the distribution of the data, and P_θ^∞ is the equilibrium distribution of the model [12]. 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. We wish the output of our DBN to be continuous A-V coordinates, and we therefore 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,[2] 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.

IV. 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 [14], [15] for further details.

A. 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 3.

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

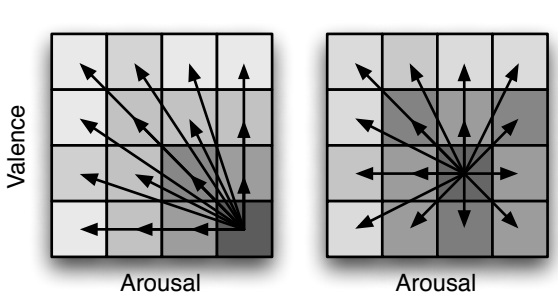

Fig. 3. Heatmap visualization of CRF transition probabilities. Actual discretization is 11×11.

B. 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++,[3] 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.

V. Discussion and Future Work

The deep belief network is a powerful topology for music emotion recognition for both learning informative feature domains, as well as providing insight into the direct relationship between emotion and acoustic content. The overall performance could potentially be increased by more advanced regression algorithms that are more robust to the high dimensionality of the data. Furthermore, other optimization methods for fine-tuning the deep structure could be investigated as well as alternate error metrics. The approach, which models the output of the regression layer as a single point in the A-V space, could be expanded to a metric that provides better knowledge of the emotion space distribution. Such an approach could model the A-V space as a heatmap, using the same approach as with the CRF.

Many problems in music information retrieval lack a singular dominant feature, and thus much attention has been given to feature selection and dimensionality reduction methods. These methods can be helpful in increasing classification performance, but leave much to be desired in terms of mathematical understanding of the predicted process. Research in deep learning is still in the early stages but offers the potential to inform many of these tasks both conceptually, as well as in raw performance.

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

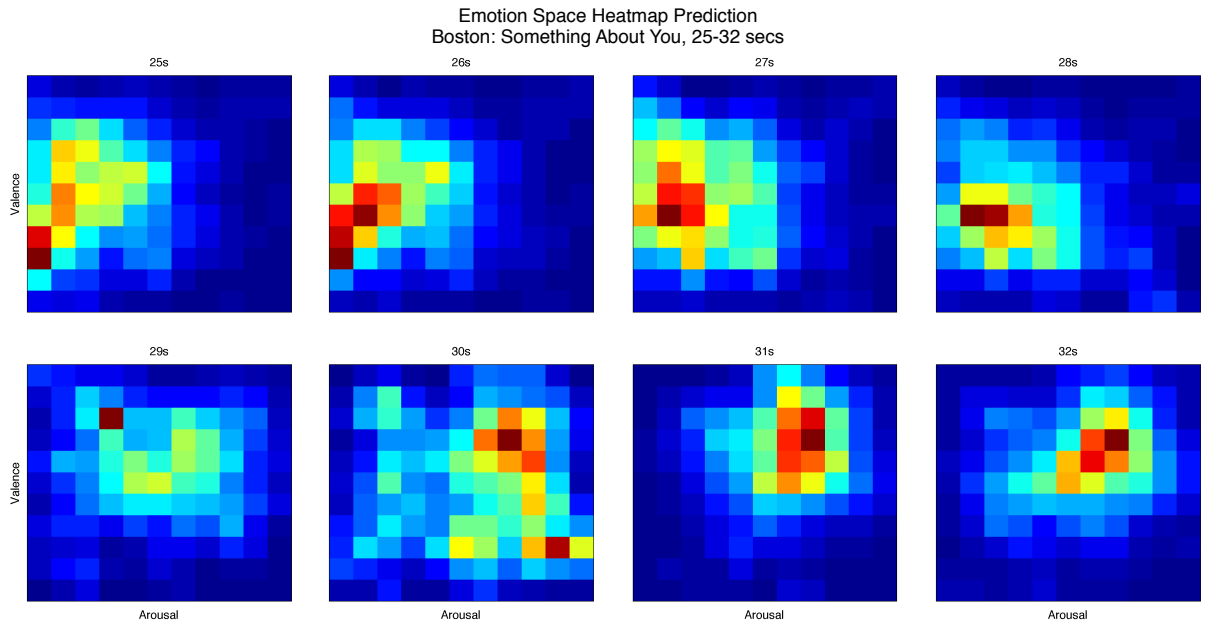

Fig. 4. Emotion space heatmap prediction using conditional random fields. Shown is the predicted emotion from the beginning of the song "Something About You," by Boston. These figures demonstrate the system tracking the emotion through the low-energy, negative-emotion introduction, and through the transition at second 29 into a high-energy, positive emotion rock verse. In these figures, red indicates the highest density and blue is the lowest [8].

In a future approach, the CRF performance could be improved by developing a model that 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.

In the talk, we will show results and discuss the effectiveness of CRFs for modeling the relationships between acoustic features and emotion space parameters. Furthermore, we will discuss feature learning in musical emotion recognition and demonstrate its use in providing us with computational models to potentially learn more about the relationships between the acoustic and affective domains. In looking to improve emotion-prediction performance, we will also discuss several potential directions in the development of models that incorporate multiple spectral time-scales to derive musical emotion. We will provide additional results for these approaches and discuss the tradeoffs associated with the increased input dimensionality as a result of the additional data.

Acknowledgment

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

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

[3] 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.

[4] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford, U.K.: Oxford Univ. Press, 1989.

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

[6] ——, "Modeling the acoustic structure of musical emotion with deep belief networks," in *NIPS Workshop on Music and Machine Learning*, 2011.

[7] ——, "Prediction of time-varying musical mood distributions from audio," in *ISMIR*, Utrecht, Netherlands, 2010.

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

[9] B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim, "Improving music emotion labeling using human computation," in *ACM SIGKDD HCOMP Workshop*, Washington, D.C., 2010.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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