2010 Ninth International Conference on Machine Learning and Applications

Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering

Erik M. Schmidt and Youngmoo E. Kim
Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
Philadelphia, PA 19104 USA
{*eschmidt, ykim*}*@drexel.edu*

***Abstract*—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. In previous work, we have modeled human response labels to music in the arousal-valence (A-V) representation of affect as a time-varying, *stochastic distribution* reflecting the ambiguous nature of the perception of mood. These distributions are used to predict A-V responses from acoustic features of the music alone via multi-variate regression. In this paper, we extend our framework to account for multiple regression mappings contingent upon a general location in A-V space. Furthermore, we model A-V state as the latent variable of a linear dynamical system, more explicitly capturing the dynamics of musical mood. We validate this extension using a "genie-bounded" approach, in which we assume that a piece of music is correctly clustered in A-V space *a priori*, demonstrating significantly higher theoretical performance than the previous single-regressor approach.**

***Keywords*-Emotion recognition, audio features, regression, linear dynamical systems, Kalman filtering**

I. 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]. However, relatively little attention has been given to systems that model how musical emotion changes over time, with most work aimed at applying a singular rating to an entire song or clip. 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. As online retailers continue to amass music libraries reaching into the millions of songs, emotion-based music search has the potential to revolutionize how we buy and listen to our music. For instance, such a system would not only allow the user to search for "happy" songs, but perhaps songs that start off as "angry," "sad," or "frustrated" and become "happy" as the song progresses.

In prior work, we created *MoodSwings* [2], an online collaborative activity designed 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 [3]. The game was 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. This method potentially provides quantitative labels that are well-suited to computational methods for parameter estimation.

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. In developing computational systems for recognizing musical affect, this lack of specificity presents significant challenges for the traditional approach of using supervised machine learning systems for classification. Instead of viewing musical mood as a singular label or value, the modeling of emotional "ground-truth" as a *probability distribution* potentially provides a more realistic and accurate reflection of the perceived emotion conveyed by a song. In previous work, we demonstrated that at reasonably short time slices (less than 15 seconds), A-V emotion can be well represented with a single two-dimensional Gaussian distribution [4]. In that approach, distribution parameters were estimated independently at each emotion space time analysis window, which often led to noisy estimates, especially in the covariances. In this work, we employ a Kalman preprocessing step in estimating our ground truth distribution, providing a more accurate estimation at each time instance, as well as a more accurate estimation of how the distribution evolves.

In previous work, we investigated using short time segments to track emotional changes throughout short clips of music [4], [5]. While these methods provided reasonable accuracy, they were all memoryless, and therefore provided no probabilistic model of the temporal dependence in musical emotion. In this new work, we seek to model musical emotion using a systems approach, employing a linear dynamical system (LDS), a model that has been subject to increasing attention in the Music-IR community [6]. Using LDS, we are able to model not just the mapping from features to emotion but also how these emotion labels evolve over time. As we find that our emotion space time-series distribution parameters clearly do not evolve in the same way across our entire music corpus, we also develop a mixture approach, partitioning the data into different A-V

978-0-7695-4300-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICMLA.2010.101

clusters and using the "genie bound" to prove that this will enhance overall performance.

While in previous approaches we have focused on combining multiple feature domains, we have also found that octave-based spectral contrast consistently provided more accurate results alone than any other feature by itself [4], [5]. Since the focus of this work is on developing specific time-varying machine learning techniques, we will restrict ourselves to only this feature.

II. 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]. Recent work has also indicated that regression approaches often outperform classification when using similar features [5], [7].

Yang *et al.* introduced the use of regression for mapping of high-dimensional acoustic features into the two-dimensional space [8]. Support vector regression (SVR), as well as a variety of boosting algorithms were applied to solve the regression problem. The ground-truth A-V labels were collected by recruiting 253 college students to annotate the data, and only one label was collected per clip. Compiling a wide corpus of features totaling 114 feature dimensions, they applied principal component analysis (PCA) before regression.

Further confirming the robustness of regression for A-V emotion prediction, Han *et al.* demonstrated that regression approaches can outperform classification when applied to the same problem [7]. Their classification task consisted of a version of the A-V space quantized into 11 blocks. Switching from classification to regression, they saw an increase from ∼33% to ∼95% accuracy.

Eerola *et al.* introduced the use of a three-dimensional parametric emotion model for labeling music [9]. In their work, they investigated multiple regression approaches including partial least-squares (PLS) regression, an approach that considers correlation between label dimensions. They achieve R^2 performance of 0.72, 0.85, and 0.79 for valence, activity, and tension, respectively.

III. Ground Truth Data Collection

Traditional methods for collecting perceived mood labels, such as the soliciting and hiring of human subjects, can be flawed [2]. In MoodSwings, participants use a graphical interface to indicate a dynamic position within the A-V space to annotate five 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,[1] and overall we have collected over 150,000 individual A-V labels spanning more than 1,000 songs.

A. MoodSwings Lite Corpus

Since the database consists entirely of popular music, the labels collected thus far display an expected bias towards high-valence and high-arousal values. Although inclusion of this bias could be useful for optimizing classification performance, it is not as helpful for learning a mapping from acoustic features that provides coverage of the entire emotion space. Because of this trend, we developed a reduced dataset consisting of 15-second music clips from 240 songs, selected using the original label set, to approximate an even distribution across the four primary quadrants of the A-V space. These clips were subjected to intense focus within the game in order to form a corpus, referred to here as MoodSwings Lite, with significantly more labels per song clip, which is used in this analysis.

B. Data Preprocessing

To construct our time-varying corpus for supervised machine learning, the collected A-V labels from each 15-second clip are analyzed at 1-second intervals. In prior work, we found A-V emotion can be well represented with a two-dimensional Gaussian distribution at such intervals, and we estimated the parameters at each time instance independently, computing the sample mean and minimum variance unbiased estimator of the covariance [4]. While the assumption of temporal independence allows estimation methods that are computationally efficient, it ignores the time-varying nature of music and undoubtedly introduces noise. Furthermore, the temporal starting point for a MoodSwings match is randomly selected for each game, so a particular labeler does not necessarily have the opportunity to weigh in on all 15 time intervals in our clips. As a result, we often have different sample sizes at each interval. For example see the left-hand plot of Figure 1: shown in gray are the individual second-by-second labels collected from the MoodSwings game and in red ellipses are the estimates of the distribution; both become darker as time progresses. While the general drift of the distribution is captured, there is a heavy amount of noise in the covariance ellipses.

To address this issue and form a better approximation of our ground truth distribution, we apply preprocessing using a Kalman/Rauch-Tung-Striebel (RTS) smoother [10], [11]. Using this approach we model our labels $\mathbf{y}$ as noisy observations of the true distribution of $\mathbf{x}$,

$$\mathbf{x}_t = \mathbf{x}_{t-1} + w_t, \qquad (1)$$

$$\mathbf{y}_t = \mathbf{x}_t + v_t. \qquad (2)$$

Gaussian noise sources w and v parametrized experimentally, representing the interfering noise corrupting our labels

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

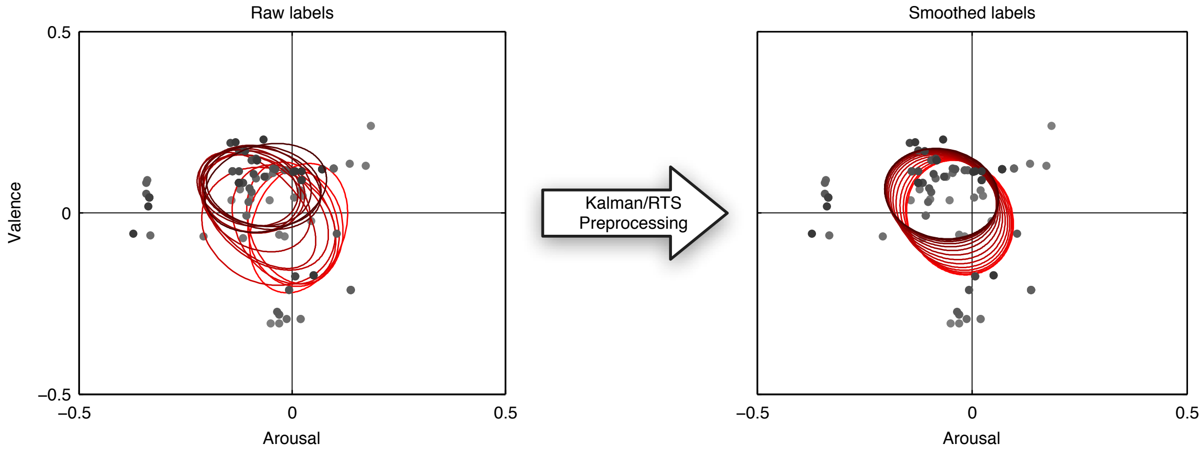

Figure 1. Emotion label preprocessing: gray dots indicate individual second-by-second labels collected from the MoodSwings game. Red ellipses are the estimates of the distribution; both become darker as time progresses.

which we assume to be zero mean,

$$w \sim \mathcal{N}(0, Q) \tag{3}$$

$$v \sim \mathcal{N}(0, R) \tag{4}$$

Thus, the values of Q and R are chosen experimentally to provide the desired amount of smoothing. Defining our initial conditions as the mean and covariance of our noisy labels, we next perform Kalman smoothing to estimate the clean distribution $\mathbf{x}$ of our noisy process $\mathbf{y}$. The Kalman smoothing equations are discussed in detail in Section V-B in Equations 12-19. It is important to note that in this special case of the Kalman filter, because the observation matrix C and dynamics matrix A have been omitted, they simply reduce to identity matrices in the standard Kalman/RTS equations.

Seen in the right half of Figure 1 is the smoothed version of the estimate, which provides a much more reasonable representation of the true distribution. It can be seen that the distribution moves consistently forward, and the unrealistic movement in the covariance ellipses has been removed.

IV. Octave-Based Spectral Contrast

Since the focus of this work is time-varying machine learning methods, we restrict ourselves in terms of features to only octave-based spectral contrast. As previously stated, we have found in previous work that this feature consistently provided more accurate results alone than any other feature by itself [4], [5].

The spectral contrast feature provides a rough representation of the harmonic content in the frequency domain based upon the identification of peaks and valleys in the spectrum separated into different frequency sub-bands [12]. Whereas the other spectrum-based features (Mel-frequency cepstral coefficients, statistical spectrum descriptors) provide information about the wideband characteristics of the signal, spectral contrast focuses on the narrowband (harmonic) content, which is often the dominant component in musical signals.

V. Methods

We investigate multiple methods for our time varying multi-variate prediction. The first system investigated is multiple linear regression (MLR), which makes predictions at each time slice independently. Next, we incorporate feedback using a linear dynamical system (LDS) and attempt to model how the emotion space distribution evolves over time with Kalman/RTS smoothing. Finally, we investigate mixtures of systems both for MLR and LDS and establish a "genie bound" for the performance upper limit on both systems.

A. Multiple Linear Regression

Our first approach simply models the relation from acoustic features to A-V coordinates using least-squares regression. That is, if we have some features $\mathbf{y}$ and emotion labels $\mathbf{x}$ (adopting the Kalman notation we will use in the latter approaches), we can define a projection matrix P that is valid for all labels,

$$P\mathbf{y}_t = \mathbf{x}_t. \tag{5}$$

In this case, each column vector in $\mathbf{y}$ is the spectral contrast at one second intervals and in $\mathbf{x}$ is the arousal and valence means and covariance (a total of five dimensions as two of the covariance values will be redundant), also at one second intervals.

B. Kalman Filtering

As we know each frame in our time-varying prediction is not independent from the previous one, we wish to develop a probabilistic model of how our emotion space distribution evolves over time. To do this, we start with the simplest model: we simply learn the time dependence of our emotion labels $\mathbf{x}$, and use that to restrict the way our emotion predictions evolve. That is, we wish to learn a simple dynamics model A such that,

$$\mathbf{x}_t = A\mathbf{x}_{t-1}. \tag{6}$$

Given this model, our emotion space mapping could be represented as a linear dynamical system as follows,

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + w_t, \tag{7}$$

$$\mathbf{y}_t = C\mathbf{x}_t + v_t. \tag{8}$$

Because the LDS models a zero mean Gaussian process, we remove the means of our labels and features prior to training, notated as $\bar{x}$ and $\bar{y}$, respectively. These values are stored such that $\bar{y}$ can be removed from unlabeled features $\mathbf{y}$ in the testing phase, and so $\bar{x}$ can be used to apply the proper bias to the estimate of the labels $\mathbf{x}$.

To use this new structure, we can learn C through least-squares just as we did with P in MLR, but it must function such that,

$$C\mathbf{x}_t = \mathbf{y}_t. \tag{9}$$

In learning the dynamics matrix A, we performed initial experiments using least-squares but opted for a constraint generation approach. Using this approach, we achieved slightly better results, which is likely attributed to the fact that it guarantees a stable solution for A. Using constraint generation, the problem is formed as a convex optimization one and efficiently solved through quadratic programming [13].

As in the data preprocessing, we model our driving noise w and observation noise v as zero mean Gaussian,

$$w \sim \mathcal{N}(0, Q) \tag{10}$$

$$v \sim \mathcal{N}(0, R) \tag{11}$$

In this case, we can compute the values for Q and R directly from the residuals of A and C.

Finally, given an unknown testing example $\mathbf{y}$, we remove the known bias $\bar{y}$ and form an estimate of its emotion space distribution by first performing the following forward Kalman filter recursions [10], [11]:

$$\mathbf{x}_t^{t-1} = A\mathbf{x}_{t-1}^{t-1} \tag{12}$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A' + Q \tag{13}$$

$$K_t = V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1} \tag{14}$$

$$\mathbf{x}_t^t = \mathbf{x}_t^{t-1} + K_t(\mathbf{y}_t - C\mathbf{x}_t^{t-1}) \tag{15}$$

$$V_t^t = V_t^{t-1} - K_tCV_t^{t-1} \tag{16}$$

Next, we perform the following backward (RTS) recursions:

$$J_{t-1} = V_{t-1}^{t-1}A'(V_t^{t-1})^{-1} \tag{17}$$

$$\mathbf{x}_{t-1}^T = x_{t-1}^{t-1} + J_{t-1}(\mathbf{x}_t^T - A\mathbf{x}_{t-1}^{t-1}) \tag{18}$$

$$V_{t-1}^T = V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1})J_{t-1}' \tag{19}$$

Adding the known bias, we form our final estimate as,

$$\mathbf{x}_t^{\text{FINAL}} = \mathbf{x}_t^T + \bar{x}. \tag{20}$$

C. Kalman Filter Mixtures

While a single LDS makes for an interesting approach to model the temporal dependence in music emotion prediction, we do not expect that we can represent the dynamics in all 240 clips with a single A matrix, and we certainly should be able to achieve better results with more than one C. In this approach, we attempt to train multiple systems and establish a "genie bound" for the prediction of musical emotion using mixtures, and to compare that to using MLR on the same clusters. We now add an assignment variable z,

$$\mathbf{x}_t = A_z\mathbf{x}_{t-1} + w_t, \tag{21}$$

$$\mathbf{y}_t = C_z\mathbf{x}_t + v_t. \tag{22}$$

To form our mixtures, we first perform k-means clustering on a three dimensional vector for each sequence: the average arousal, valence, and flux. The flux is computed between each time instance vector in our label sequences, which as previously discussed is made up of the arousal and valence means and covariance at one second intervals. This allows clustering to occur both spatially and dynamically.

As in the previous experiments, the feature means $\bar{y}$ are computed across all clusters and saved for removal before both training and testing. In the case of the labels, we now compute a specific mean for each cluster $\bar{x}_z$, which is removed before training and saved to bias testing estimates for the corresponding cluster.

VI. Experiments and Results

To quantify the accuracy of the predictions, we first examine the distance from the predicted to the true distributions. To do this, we apply the non-symmetrized (one-way) Kullback-Leibler (KL) divergence, a measure of relative entropy between two distributions. Since we have applied the assumption that emotion space distributions are Gaussian, the KL divergence has closed-form,

$$\text{KL}(p||q) \equiv \int p(x)\log\frac{p(x)}{q(x)}dx = E_p\left\{\log\frac{p(X)}{q(X)}\right\} \tag{23}$$

$$\text{KL}(p||q) = \log\frac{|\Sigma_q|}{|\Sigma_p|} + tr(\Sigma_q^{-1}\Sigma_p^{-1}) + (\mu_q - \mu_p)^T\Sigma_q^{-1}(\mu_q - \mu_p) - d. \tag{24}$$

As an additional qualitative metric, we also provide the Euclidean distance between the projected means as a normalized percentage of the A-V space. However, to provide

Feature/ Topology	Average Mean Distance	Average KL Divergence	Average Randomized KL Divergence	T-test
MLR Noisy [4]	0.169 ± 0.007	14.61 ± 3.751	27.00 ± 10.33	10.77
MLR	0.160 ± 0.007	4.576 ± 0.642	9.531 ± 1.856	18.35
Kalman	0.160 ± 0.007	4.650 ± 0.652	8.870 ± 1.633	16.85
MLR Mixture	0.109 ± 0.007	3.179 ± 0.539	12.00 ± 1.899	19.68
Kalman Mixture	0.109 ± 0.007	2.881 ± 0.568	12.34 ± 1.973	19.63

Table I
RESULTS FOR EMOTION DISTRIBUTION PREDICTION OVER TIME.

context to KL values and to benchmark the significance of the regression results, we compared the projections to those of an essentially random baseline. Given a trained regressor and a set of labeled testing examples, we first determined an A-V distribution for each sample. The resulting KL divergence to the corresponding A-V distribution was compared to that of another randomly selected A-V distribution from the test set. Comparing these cases over 50 cross-validations, we computed a Student's T-test for paired samples to verify the statistical significance of our results.

As should be expected, these T-test values produce extremely high confidence values. For our paired t-test on 865 testing examples, even the system which returned the smallest T value (10.77) produces over 99.999% confidence.

A. Multiple Linear Regression

While MLR predicts each time instance independently from the previous one, it provides a reasonable amount of accuracy, a result consistent with our prior work [5], [4]. Shown in Table I are the results where MLR has an average error in the mean prediction of 0.160 in the normalized space. This is a significant improvement over our previous work where it was off by 0.169 on average [4], which can be directly attributed to the fact that there was no label preprocessing. However, the largest improvement can be seen in the KL-divergence values which have reduced from 14.61 to 4.576, on average.

B. Kalman Filtering

The Kalman filtering approach provides quite reasonable results as well. The total error is similar to MLR, but this result is not surprising due to the fact that we have restricted our evolution model to a single A matrix, which is a significant simplification. While the average distance remains identical, the increase from 4.576 to 4.650 in KL is a nearly negligible change. As this demonstrates that all of our clips do not fit easily into exactly the same dynamics model, it makes a strong case for mixture systems.

C. Kalman Filter Mixtures

The Kalman mixture provides the best result of any system, using only four clusters we achieve an average KL of 2.881, which is a significant improvement over both the MLR system at 4.576 and the MLR mixture at 3.179. In terms of mean error, however, Kalman and MLR mixtures produce nearly identical results, with normalized distances of about 0.109.

Shown in Figure 2 are the the emotion space predictions of six fifteen second clips using this method. Shown in gray are the individual A-V ratings collected from the MoodSwings game, in red are the distribution estimates, and in blue are the predictions from acoustic features; all three become darker as time progresses.

VII. Discussion and Future Work

In working with highly subjective emotion labels where there is not necessarily a singular rating for even the smallest time slice, it is clear that we can develop the most accurate representation of our ground truth using a distribution. Using a Kalman filtering approach, we have been able to form robust estimates of our distribution and how it evolves over time.

In terms of predicting these time-varying distributions from acoustic content, we have demonstrated our system's general approach to be a powerful representation, yielding the ability to model and track emotion space distribution changes over time. Through the "genie bound," we have shown that using mixtures we can more accurately model both the different spacial clusters as well as dynamics in the A-V space. In the future, we wish to develop systems to accurately predict the cluster of an unknown testing example. This problem could be solved through traditional classification approaches, though added difficulty is presented in terms of clusters that are often of unequal sizes (especially in terms of dynamics clustering), and making decisions based on the prior negates the benefit of those added clusters.

A future system could potentially pair our emotion space estimation with a song segmentation approach. That is, if we could first segment our songs into multiple sections (e.g. verse, chorus, bridge), it would be a reasonable assumption that the emotion space distribution would be much more stationary in such a range, and our Kalman filtering approach could provide highly robust estimates. With such a system, we could also potentially represent songs as Gaussian mixtures in A-V space, enabling more robust emotion-based music recommendation.

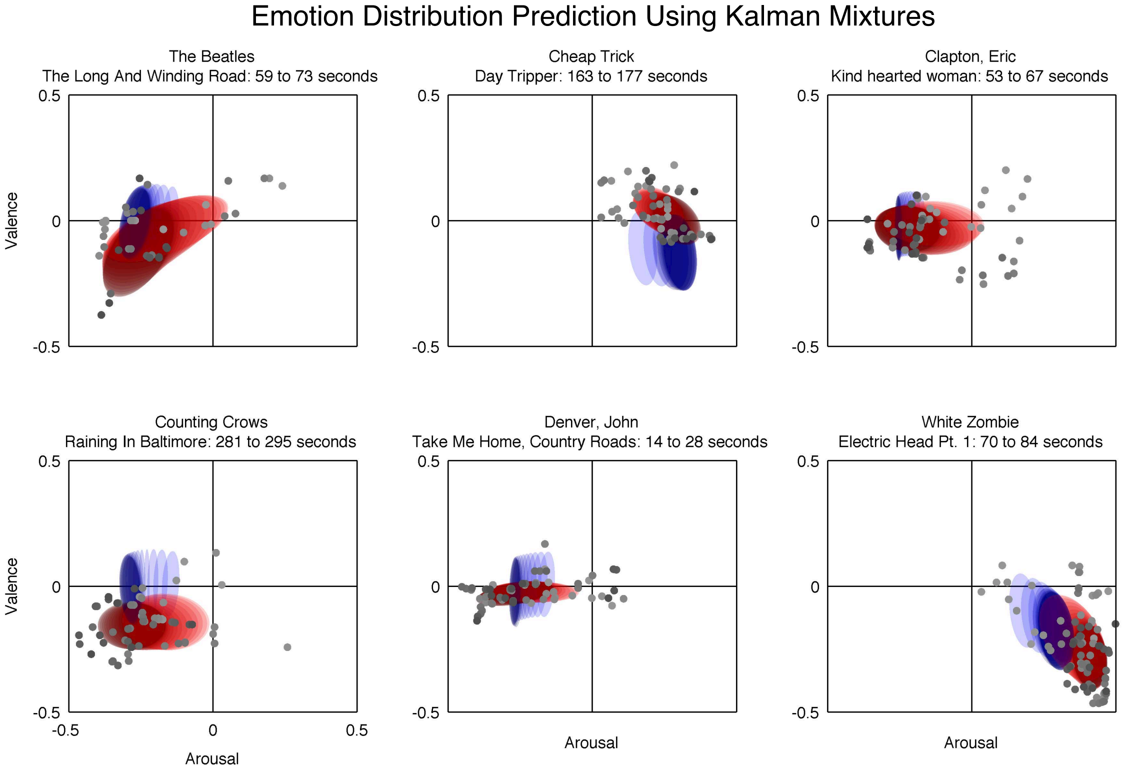

Figure 2. Emotion label preprocessing: gray dots indicate individual second-by-second labels collected from the MoodSwings game, red ellipses indicate the estimates of the distribution, and blue ellipses indicate the predictions using Kalman filter mixtures; all three become darker as time progresses.

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 *Proc. of the 11th Intl. Society for Music Information Retrieval (ISMIR) Conf.*, Utrecht, Netherlands, 2010.

[2] Y. E. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. of the 9th Intl. Conf. on Music Information Retrieval*, Philadelphia, PA, September 2008.

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

[4] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. of the 11th Intl. Society for Music Information Retrieval (ISMIR) Conf.*, Utrecht, Netherlands, 2010.

[5] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *MIR '10: Proc. of the Intl. Conf. on Multimedia Information Retrieval*, Philadelphia, PA, 2010, pp. 267–274.

[6] L. Barrington, A. Chan, and G. Lanckriet, "Modeling music as a dynamic texture," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 18, no. 3, pp. 602–612, 2010.

[7] B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression," in *Proc. of the 10th Intl. Society for Music Information Retrieval (ISMIR) Conf.*, Kobe, Japan, 2009.

[8] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, "A regression approach to music emotion recognition," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 16, no. 2, pp. 448–457, 2008.

[9] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. of the 10th Intl. Society for Music Information Retrieval (ISMIR) Conf.*, Kobe, Japan, 2009.

[10] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[11] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," University of Toronto, Tech. Rep. CRG-TR-96-2, 1996.

[12] 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.

[13] S. Siddiqi, B. Boots, and G. Gordon, "A constraint generation approach to learning stable linear dynamical systems," in *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press, 2008, pp. 1329–1336.