PREDICTION OF TIME-VARYING MUSICAL MOOD DISTRIBUTIONS FROM AUDIO

Erik M. Schmidt and Youngmoo E. Kim
Electrical and Computer Engineering, Drexel University
{eschmidt,ykim}@drexel.edu

ABSTRACT

The appeal of music lies in its ability to express emotions, and it is natural for us to organize music in terms of emotional associations. But the ambiguities of emotions make the determination of a single, unequivocal response label for the mood of a piece of music unrealistic. We address this lack of specificity by modeling human response labels to music in the arousal-valence (A-V) representation of affect as a *stochastic distribution*. Based upon our collected data, we present and evaluate methods using multiple sets of acoustic features to estimate these mood distributions parametrically using multivariate regression. Furthermore, since the emotional content of music often varies within a song, we explore the estimation of these A-V distributions in a *time-varying* context, demonstrating the ability of our system to track changes on a short-time basis.

1. INTRODUCTION

The problem of automated recognition of emotional content (mood) within music is the subject of increasing attention among music information retrieval (MIR) researchers [1–3]. 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 emotions conveyed by a song.

A variety of methods are used for collecting mood-specific labels for music corpora, for example, annotations curated by experts (e.g., Allmusic.com) and the analysis of unstructured user-generated tags (e.g., Last.fm). While these approaches efficiently provide data for large collections, they are not well-suited for reflecting variations in the emotional content as the music changes. In prior work we created *MoodSwings* [4], 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 [5]. 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.

In previous work we have investigated short-time regression approaches for emotional modeling, developing a functional mapping from a large number of acoustic features directly to A-V space coordinates [1]. Since the application of a single, unvarying mood label across an entire song belies the time-varying nature of music, we focused on using short-time segments to track emotional changes over time. In our current work we demonstrate that not only does the emotional content change over time, but also that a distribution of (as opposed to singular) ratings is appropriate for even short time slices (down to one second). In observing the collected data, we have found that most examples can be well represented by a single two-dimensional Gaussian distribution.

To perform the mapping from acoustic features to the A-V mood space, we explore parameter prediction using multiple linear regression (MLR), partial least-squares (PLS) regression, and support vector regression (SVR). In modeling the data as a two dimensional Gaussian, our goal is to be able to predict the A-V distribution parameters $\mathcal{N}(\mu, \Sigma)$ from the acoustic content. We first evaluate the effectiveness of this system in predicting emotion distributions for 15 second clips and subsequently shorten the analysis window length to demonstrate its ability to follow changes in A-V label distributions over time.

No dominant acoustic feature has yet emerged for music emotion recognition, and previous work has focused on combining multiple feature sets [1–3, 6]. We evaluate multiple sets of acoustic features for each task, including psychoacoustic (mel-cepstrum and statistical frequency spectrum descriptors) and music-theoretic (estimated pitch chroma) representations of the labeled audio. Although the large number of potential features can present problems, rather than employing dimensionality reduction methods (e.g., principal components analysis) we explore an alternative method for combining different feature sets, using ensemble methods to determine the relative contribution of single-feature systems for improved overall performance.

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. Recent work has also indicated that regression approaches often outperform classification when using similar features [1, 2].

Yang *et al.* introduced the use of regression for mapping of high-dimensional acoustic features into the two-dimensional space [6]. Support vector regression (SVR), as well as a variety of boosting algorithms including AdaBoost.RT, 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 [2]. Their classification task consisted of a quantized version of the A-V space into 11 blocks. Using a wide variety of audio features, they initially investigated the use of classification, obtaining only $\sim$33%. Still mapping to the same 11 quantized categories, applying regression they obtained up to $\sim$95% accuracy.

Eerola *et al.* introduced the use of a three-dimensional parametric emotion model for labeling music [3]. 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, using PLS and also report peak R^2 prediction rates for 5 basic emotion classes (angry, scary, happy, sad, and tender) as ranging from 0.58 to 0.74.

3. GROUND TRUTH DATA COLLECTION

Traditional methods for collecting perceived mood labels, such as the soliciting and hiring of human subjects, can be flawed. 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.

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.

4. ACOUSTIC FEATURE COLLECTION

As previously stated, there is no single dominant feature, but rather many that play a role (e.g., loudness, timbre, harmony) in determining the emotional content of music. Since our experiments focus on the tracking of emotion over time, we chose to focus on solely on time-varying features. Our collection (Table 1) contains many features that are popular in Music-IR and speech processing, encompassing both psychoacoustic and music-theoretic representations. Instead of raw chroma we utilize the autocorrelation of each short-time chroma vector, providing a shift-invariant feature. In preliminary experiments we found this feature to perform better than raw chroma, since it promotes similarity in terms of the modes of harmony (e.g. major, minor, augmented, and diminished chords) as opposed to particular chords (e.g., A major vs. D major).

Feature	Description
Mel-frequency cepstral coefficients (MFCCs) [7]	Low-dimensional representation of the spectrum warped according to the mel-scale. 20-dimensions used.
Chroma (i.e., Pitch Class Profile) [8]	Autocorrelation of chroma is used, providing an indication of modality.
Spectral Spectrum Descriptors (SSDs) [9]	Includes spectral centroid, flux, rolloff, and flatness. Often related to timbral texture.
Spectral Contrast [10]	Rough representation of the harmonic content in the frequency domain.

Table 1. Acoustic feature collection for music emotion regression.

5. EXPERIMENTS AND RESULTS

Given the continuous nature of our problem, the prediction of a 2-d Gaussian within the A-V space, we explored several methods for multi-variate parameter regression. In these experiments we employ multiple linear regression (MLR), partial least-squares (PLS), and support vector regression (SVR) to create optimal projections from each of the acoustic feature sets described above. For our initial distribution regression experiments, we averaged feature

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

Feature/ Topology	Regression Method	Average Mean Distance	Average KL Divergence	Average Randomized KL Divergence	T-test
MFCC	MLR	0.161 ± 0.008	4.098 ± 0.513	8.516 ± 1.566	5.306
Chroma	MLR	0.185 ± 0.010	5.617 ± 0.707	7.765 ± 2.135	5.659
S. Shape	MLR	0.167 ± 0.009	4.183 ± 0.656	7.691 ± 1.573	5.582
S. Contrast	MLR	**0.151 ± 0.008**	**3.696 ± 0.657**	**8.601 ± 1.467**	**5.192**
MFCC	PLS	0.155 ± 0.008	3.863 ± 0.56712	8.306 ± 1.389	5.540
Chroma	PLS	0.183 ± 0.010	5.286 ± 0.96019	7.146 ± 1.665	5.565
S. Shape	PLS	0.151 ± 0.008	3.770 ± 0.84026	8.278 ± 1.527	4.951
S. Contrast	PLS	**0.151 ± 0.008**	**3.684 ± 0.644**	**8.700 ± 1.831**	**5.171**
MFCC	SVR	**0.140 ± 0.008**	**3.186 ± 0.597**	**7.744 ± 1.252**	**5.176**
Chroma	SVR	0.186 ± 0.008	4.831 ± 0.737	6.466 ± 0.935	5.655
S. Shape	SVR	0.176 ± 0.008	4.611 ± 0.841	7.348 ± 1.025	5.251
S. Contrast	SVR	0.150 ± 0.008	3.357 ± 0.500	7.356 ± 1.341	5.301
Stacked Features	MLR	0.152 ± 0.007	3.917 ± 0.496	9.355 ± 1.879	5.737
Fusion Unweighted	MLR	0.149 ± 0.007	3.333 ± 0.433	6.785 ± 0.996	5.879
Fusion Weighted	MLR	0.147 ± 0.007	3.280 ± 0.423	6.803 ± 1.309	5.980
M.L. Seperate	MLR	0.147 ± 0.007	3.399 ± 0.478	8.235 ± 1.598	5.598
M.L. Combined	MLR	**0.145 ± 0.007**	**3.198 ± 0.454**	**7.637 ± 1.389**	**5.551**
Stacked Features	PLS	0.145 ± 0.006	3.403 ± 0.467	8.407 ± 1.635	5.543
Fusion Unweighted	PLS	0.145 ± 0.007	3.332 ± 0.508	7.123 ± 1.461	5.681
Fusion Weighted	PLS	0.145 ± 0.006	3.309 ± 0.501	7.160 ± 1.373	5.619
M.L. Seperate	PLS	0.145 ± 0.008	3.465 ± 0.577	8.426 ± 1.705	5.433
M.L. Combined	PLS	**0.144 ± 0.007**	**3.206 ± 0.515**	**7.889 ± 1.656**	**5.485**

Table 2. Distribution regression results for fifteen second clips.

dimensions across all frames of a given 15-second music clip, thus representing each clip with a single vector of features. Preliminary experiments were performed using second- and higher-order statistics with the 15-second clips, but in all cases the inclusions of such data failed to show any significant performance gains.

In all experiments, 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. Additionally, each experiment was subject to over 50 cross-validations, varying the distribution of training and testing data sets.

5.1 Single Feature Emotion Distribution Prediction

There are many possible methods for evaluating the performance of our system. Kullback-Liebler (KL) divergence (relative entropy) is commonly used to compare probability distributions. Since the regression problem targets known distributions, our primary performance metric is the non-symmetrized (one-way) KL divergence (from the projected distribution to that of the collected A-V labels). To provide an additional qualitative metric, we also provide results as the Euclidean distance between the projected means as a normalized percentage of the A-V space. However, to provide 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 Student's T-test for paired samples to verify the statistical significance of our results.

From Table 2 it can be seen that the best performing single feature system is SVR with MFCC features at an average KL of 3.186. However, in both the MLR and PLS system the highest performing single feature is spectral contrast with 3.696 and 3.684, respectively. As the main advantage of PLS over MLR is that it observes any correlation between dimensions in the multivariate regression, it is surprising that the performance difference between the two is nearly negligible. Given our degrees of freedom (72 test samples), even our lowest T-test value (5.171) produces confidence of statistical significance greater than 99.999%.

Shown in Figure 1 is the projection of six 15-second clips into the (A-V) space resulting from multiple regression methods and acoustic features. The standard deviation of the ground truth as well as each projection is shown as an ellipse. The performance of the regression can be evaluated in terms of the total amount of overlap between a projection and its ground truth.

5.2 Feature Fusion

While most individual features perform reasonably in mapping to A-V coordinates, a method for combining information from these domains (more informed than simply concatenating the features) could potentially lead to higher performance. In this section we investigate multiple schemes for feature fusion. Given the very small performance gains and high computational overhead of SVR, we chose to narrow our focus to MLR and PLS for these

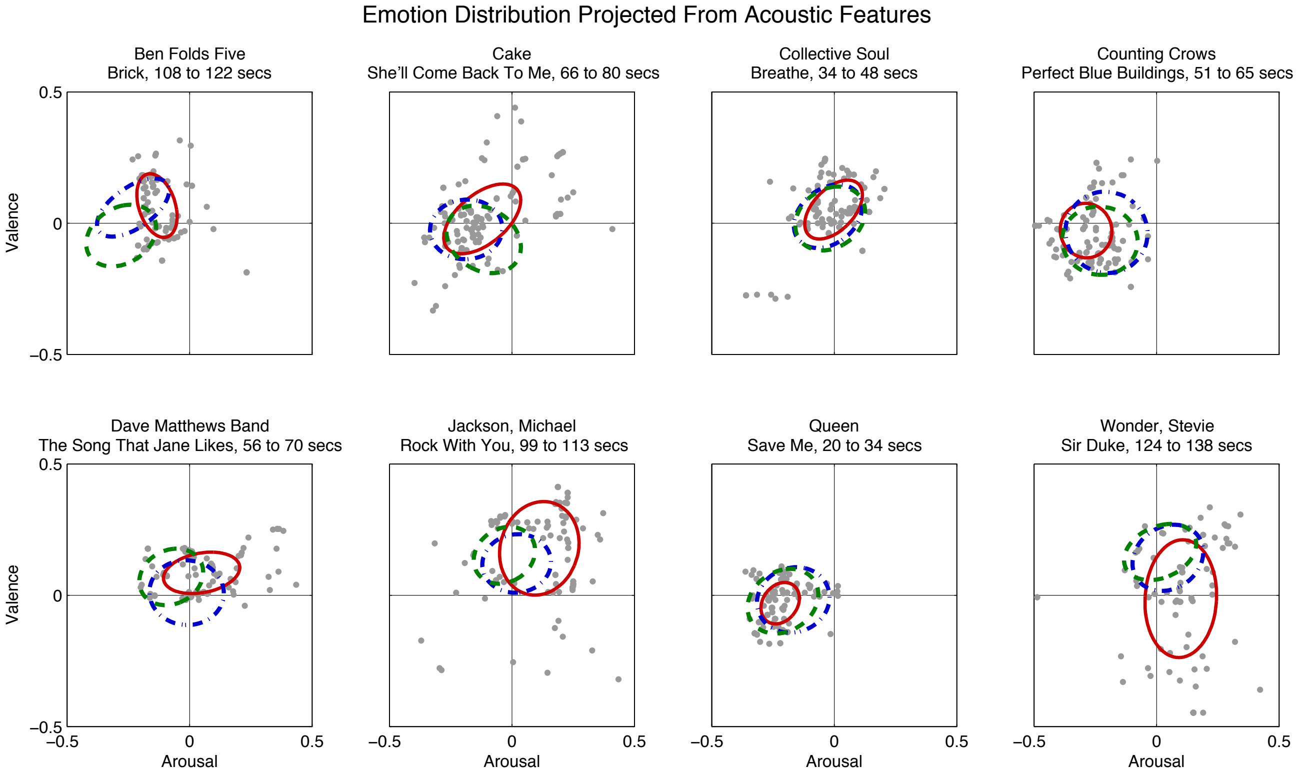

Figure 1. Collected A-V labels and distribution projections resulting from regression analysis. A-V labels: second-by-second labels per song (gray ●), Σ of collected labels (solid red ellipse), Σ of MLR projection from spectral contrast features (dash-dot blue ellipse), Σ of MLR Multi-Level combined projection (dashed green ellipse).

experiments. As our ultimate system will require many predictions over time in order to reflect emotional changes, the costs of SVR outweigh the benefits.

In our fusion results the performance for simply stacking features into one large feature vector is provided to give context to the other fusion methods. Our more simple approach consists of a fusion system that is a combination of the outputs from the individual feature regression systems. In the unweighted approach we simply average the parameter outputs from each individual feature regressor, and in the weighted approach we weight each individual feature regressor by its ability to predict a particular parameter, which is determined by leave-one-out cross-validation.

In addition, we develop a two-level regression scheme by feeding the outputs of individual regressors, each trained using distinct features, into a second-stage regressor determining the final prediction. We investigated two topologies (Figure 2): in one case the secondary arousal and valence regressors receive only arousal and valence estimates, respectively; in the second case the secondary arousal and valence regressors receive both arousal and valence estimates from the first-stage. We refer to these two topologies as multi-layer separate and multi-layer combined. In all cases the secondary regressors are trained using a leave-one-out method (on each iteration we train the first-stage regressors leaving one example out and use the estimates of that example from the first stage to train the second stage). The results for both cases are shown in Table 2.

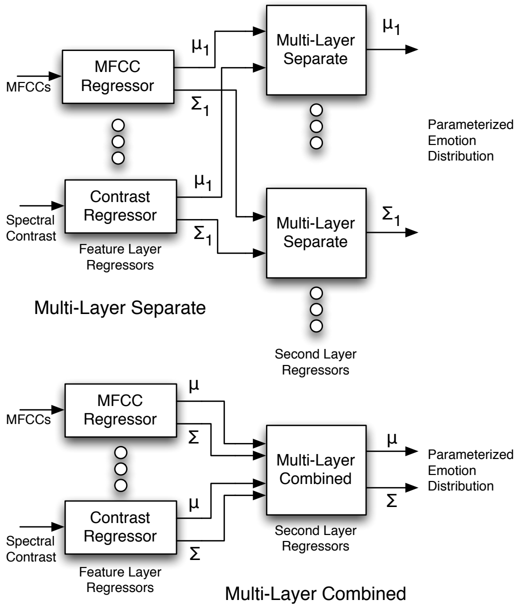

Figure 2. Multi-layer regression topologies.

5.3 Time-varying emotion distribution prediction

In attempting to predict the emotion distribution over time, we next shorten label observation rate to once per-second and attempt to regress multiple feature windows from each song. For the ground truth data collection this means that for each 15-second clip we now have 15 examples, increasing our total corpus to 3600 examples. Of course for any

Feature/ Topology	Average Mean Distance	Average KL Divergence	Average Randomized KL Divergence	T-test
MFCC	0.169 ± 0.007	14.61 ± 3.751	27.00 ± 10.33	10.77
Chroma	0.190 ± 0.007	18.71 ± 6.819	22.53 ± 6.984	9.403
S. Shape	0.173 ± 0.007	15.46 ± 6.402	24.61 ± 9.220	11.06
S. Contrast	$\mathbf{0.160 \pm 0.006}$	$\mathbf{13.61 \pm 5.007}$	$\mathbf{27.29 \pm 9.861}$	**10.23**
M.L. Combined	$\mathbf{0.154 \pm 0.006}$	$\mathbf{13.10 \pm 5.359}$	$\mathbf{28.39 \pm 10.35}$	**10.08**

Table 3. Distribution regression results for short-time (one-second) A-V labels.

experiment, multiple examples from the same song must be either all in the training or testing set. In addition, as it is clear that some past data may be necessary to accurately determine the current emotional content, we include past features and investigate the optimal feature window length.

Given the similar performance of MLR and PLS in fusion methods, for our short time analysis we will restrict ourselves to only the MLR methods. The similarity in performance is likely due to the fact that in the multilayer combined system, both MLR and PLS are able to account for the correlation between label dimensions. In moving forward with time-varying regression, we wish to be able to apply all methods in real-time as a "virtual annotator" for MoodSwings. This directly addresses the bootstrapping problem inherent to the system in cases where multiple annotators are not available at the same time. A preliminary single-user version of MoodSwings called MoodSwings Single Player,[2] which demonstrates our real-time regression system, is available online.

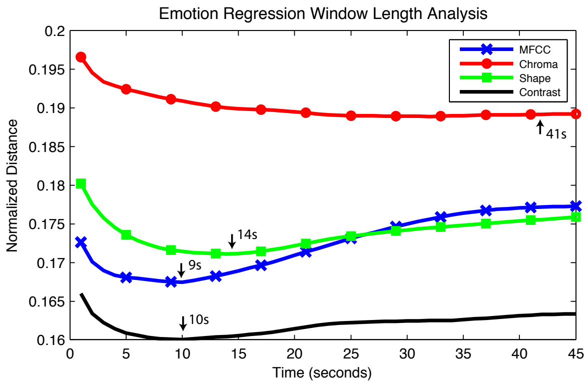

Figure 3. Window length analysis for different acoustic features.

For our time-varying approach, we seek to develop regressors that can predict the emotion for a single second using only current and past audio data. In terms of our data collection this implies that we have 15 distributions for each 15-second music clip (for 240 clips this yields a total of 3600 distributions). But we should also consider the optimal analysis window length for regression from each acoustic feature set. In Figure 3 we perform a regression analysis for each window length from 1 to 45 seconds (in increments of one second) and plot the average KL divergence from the projections to the collected distributions. As in previous experiments, the training/testing data is split 70%/30% and cross-validated 50 times. From the window length analysis in Figure 3, it can be seen that the optimal window length is not the same for all feature domains. For MFCCs we obtain the most accurate prediction using 13 seconds of past feature data, also 13 seconds for SSDs, 15 for spectral contrast, and 41 seconds for chroma. We use these feature window lengths in the regression analysis to follow.

In moving to short-time labels, it can be seen from Table 3 that our overall KL has increased, but our average distance ratings have mostly remained the same. This is most likely attributed to the fact that the underlying label covariance is less consistent due to the smaller quantity of collected A-V labels. Out T-test values have increased as well, which can be attributed to the overall increase in examples (from 240 to 3600). Considering our short-time degrees of freedom (1080 testing examples), our lowest T value (9.403) produces confidence of statistical significance (vs. randomly selected projections) higher than 99.999%. To visualize emotion regression over time, we have chosen three clips which display a clear shift in emotion distribution, plotting both the collected and projected distributions at one second intervals (Figure 4).

6. DISCUSSION AND FUTURE WORK

In working with highly subjective emotional labels, where there is not necessarily a singular rating for even the smallest time slice, it is clear that we can develop a more accurate system (in terms of predicting actual human labels) by representing the ground truth as a distribution. While accounting for potential dependence between the distribution parameters in the A-V emotion space seemed to be of high importance, some of the best performing techniques assumed total independence of parameters. In particular, combining MLR in multiple stages produces results comparable to more computationally complex methods.

One of our targeted applications, a "virtual annotator" to be used in MoodSwings, requires real-time calculation of projections, which also favors the simpler regression implementations. For the activity, the required degree of accuracy is questionable to begin with [11]. In our observations, we have found that it is more important for a virtual annotator to behave "realistically" (appropriate movement when the emotion changes) in order to keep a human participant engaged in the activity. But as we implement the

[2] MoodSwings Single Player: http://music.ece.drexel.edu/mssp

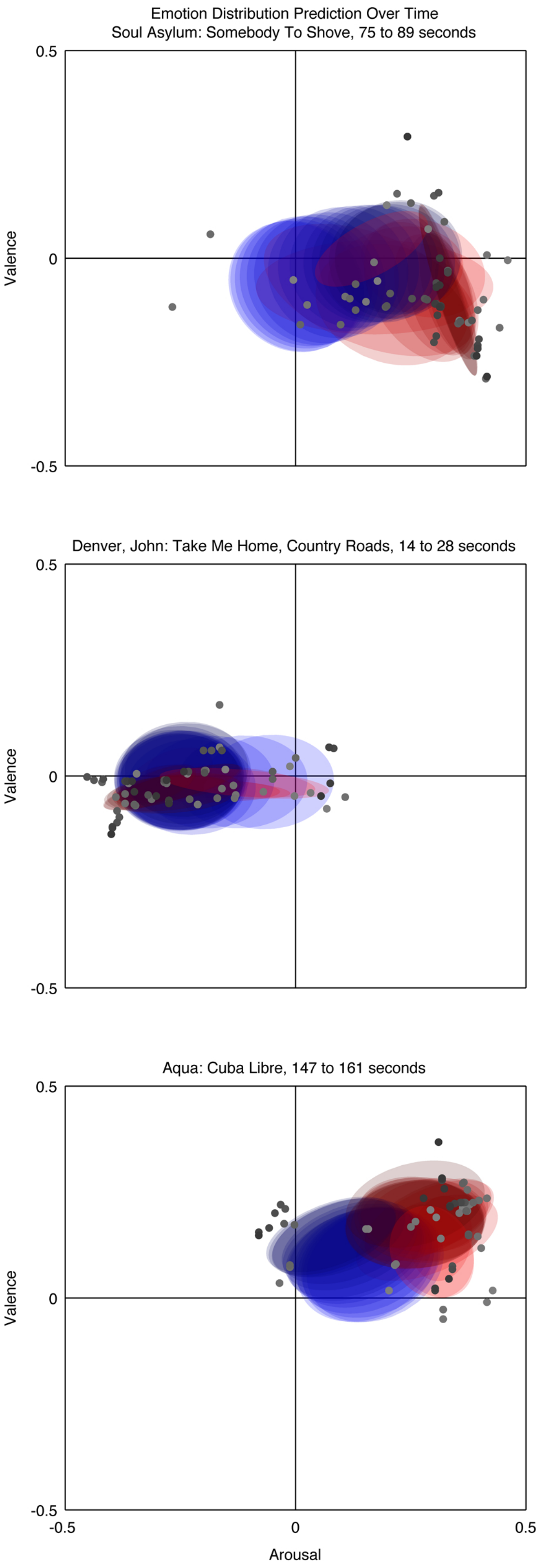

Figure 4. Time-varying emotion distribution regression results for three example 15-second music clips (markers become darker as time advances): second-by-second labels per song (gray ●), Σ of the collected labels over 1-second intervals (red ellipse), and Σ of the distribution projected from acoustic features in 1-second intervals (blue ellipse).

virtual annotator to facilitate the collection of more human data, we hope to continue increasing the accuracy of our regression system.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] 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.

[2] 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 Conf.*, Kobe, Japan, 2009.

[3] 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 Conf.*, Kobe, Japan, 2009.

[4] 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.

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

[6] 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 Transactions on*, vol. 16, no. 2, pp. 448–457, 2008.

[7] 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.

[8] T. Fujishima, “Realtime chord recognition of musical sound: a system using common lisp music.” in *Proc. of the Intl. Computer Music Conf.*, 1999.

[9] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[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*, vol. 1, 2002, pp. 113–116.

[11] B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim, “Improving music emotion labeling using human computation,” in *HCOMP ’10: Proc. of the ACM SIGKDD Workshop on Human Computation*, Washinton, D.C., 2010.