Feature Selection for Content-Based, Time-Varying Musical Emotion Regression

Erik M. Schmidt[1], Douglas Turnbull[2], and Youngmoo E. Kim[1]
Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104[1]
Department of Computer Science, Swarthmore College, Swarthmore PA 19081[2]
{eschmidt, ykim}@drexel.edu[1]
turnbull@cs.swarthmore.edu[2]

ABSTRACT

In developing automated systems to recognize the emotional content of music, we are faced with a problem spanning two disparate domains: the space of human emotions and the acoustic signal of music. To address this problem, we must develop models for both data collected from humans describing their perceptions of musical mood and quantitative features derived from the audio signal. In previous work, we have presented a collaborative game, *MoodSwings*, which records dynamic (per-second) mood ratings from multiple players within the two-dimensional *Arousal-Valence* representation of emotion. Using this data, we present a system linking models of acoustic features and human data to provide estimates of the emotional content of music according to the arousal-valence space. Furthermore, in keeping with the dynamic nature of musical mood we demonstrate the potential of this approach to track the emotional changes in a song *over time*. We investigate the utility of a range of acoustic features based on psychoacoustic and music-theoretic representations of the audio for this application. Finally, a simplified version of our system is re-incorporated into MoodSwings as a simulated partner for single-players, providing a potential platform for furthering perceptual studies and modeling of musical mood.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.5.5 [**Sound and Music Computing**]: Systems

General Terms

Algorithms, Experimentation

Keywords

Emotion recognition, audio features, machine learning, regression

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

1. INTRODUCTION

The training of supervised machine learning systems for determining the emotional content in music necessarily requires human-labeled "ground truth" observations, but the task often lacks a singular well-defined answer. A variety of factors contribute to a person's perception of musical mood, and there is bound to be some variation and disagreement between user ratings of the same content. Inspired by other "games with a purpose", we created MoodSwings [1] to collect second-by-second labels of music clips using the well-known *Arousal-Valence* (A-V) space representation of emotions, where valence reflects positive vs. negative emotions and arousal indicates emotional intensity [2]. The game was designed specifically to capture data reflecting the time-varying nature of musical mood and also to collect a distribution of labels across multiple players for a given song or even a moment within a song.

Because of the time-varying nature of music, developing systems to automatically organize an entire song or clip using a single mood label, as in prior approaches [3, 4, 5, 6, 7], undoubtedly leads to imprecise classifications. Using initial data collected by MoodSwings, it is instead our ultimate goal to track the emotional content of music over time. In order to take full advantage of the A-V space, we formulate our problem as a regression; developing a functional mapping from high-dimensional acoustic features to emotion space coordinates. This mapping is first implemented as a least-squares regression and later improved using support vector regression (SVR). We first demonstrate preliminary results of our system in tracking the emotional content of music over short time windows, and later implement a simplified system to be used as a simulated player "AI" for single-player MoodSwings games.

In searching for the most informative features for mood detection, no single dominant feature (e.g., loudness, timbre, and harmony all play some role) has yet emerged [8]. In our experiments, we also investigated 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.

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. Li and Ogihara [3] used acoustic features related to timbre, rhythm, and pitch to train support vector ma-

chines (SVMs) to classify music into 13 mood categories. Using a hand-labeled library of 499 music clips (30-seconds each), they achieved an accuracy of ∼45%, with 50% of the database used for training and testing, respectively.

Lu, Liu, and Zhang [4] pursued mood detection and tracking (following dynamic mood changes during a song) using a variety of acoustic features related to intensity, timbre, and rhythm. Their classifier used Gaussian Mixture Models (GMMs) for Thayer's four principal mood quadrants in the valence-arousal representation. The system was trained using a set of 800 classical music clips (from a data set of 250 pieces), each 20 seconds in duration, hand labeled to one of the 4 quadrants. Their system achieved an accuracy of ∼85% when trained on 75% of the clips and tested on the remaining 25%.

Xiao, Dellandrea, Dou, and Chen [5] investigated the optimal segment length for mood classification, using the same classification scheme as [4]. Using 60 unique classical pieces, each piece is broken down into segments of length 4s, 8s, 16s, and 32s, and labeled by two reviewers. They found that their system reached its peak classification performance when 16s clips were used in both the training and testing, achieving a classification performance of 88.46%.

In 2007, the Music Information Research Evaluation eXchange (MIREX) first included a "beta" task on audio music mood classification with 8 systems submitted. The audio clips used for this task were assigned to one of 5 mood clusters, aggregated from mood labels (adjectives) taken from the All Music Guide. Using 600 30-second hand-labeled clips, the clips were selected to be distributed equally among the 5 mood clusters that were used in the evaluations. All participants performed reasonably well (far higher than chance) with the highest performing system achieving correct classifications slightly over 60% of the time [9]. It should be noted that several of the systems were primarily designed for the genre classification task, but were also appropriated to the mood classification task [6].

Most similar to our work, Yang, Lin, Su, and Chen [7] introduced the use of regression for mapping of high-dimensional acoustic features to the A-V space. Support vector regression [10], as well as variety of boosting algorithms including AdaBoost.RT [11], are 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 in their study. The work is primarily focused on the regression methods themselves as opposed to acoustic feature selection and analysis. The feature set used consists of 114 dimensions computed using publicly available extraction tools, which were then reduced to a tractable number of dimensions using principal component analysis (PCA).

3. GROUND TRUTH DATA COLLECTION

As discussed in [1], traditional methods for collecting perceived mood labels, such as the soliciting and hiring of human subjects, can be flawed. MoodSwings is a game for online collaborative annotation based on the two-dimensional arousal-valence model. In the game, players position their cursor within the A-V space while competing (and collaborating) with a partner player to correctly annotate five 30-second music clips. Glimpses of the partner's position are provided every three seconds and scoring is based on the amount of overlap between the players' cursors, which encourages consensus and discourages nonsensical labeling. As an additional incentive for proactive and independent labeling, bonus points are awarded to the player who first reaches a particular location, making it impossible to out-score an opponent by simply following their cursor.

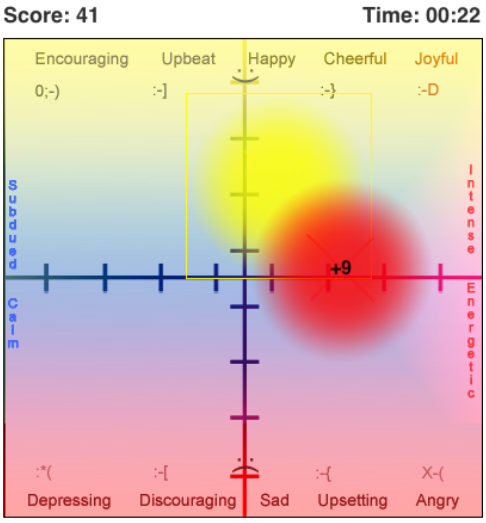

Figure 1: The MoodSwings gameboard.

3.1 Summary of Data Collection

The song clips used in MoodSwings are drawn from the uspop2002 database [12], and overall we have collected over 100,000 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, via labels collected through the game, 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.

Although we used the MoodSwings Lite corpus as the basis for classification, the original (uniform) distribution across the quadrants shifted slightly as more labels were collected for each individual clip. The final distribution of "ground truth" class labels is given in Table 1.

Class No.	Arousal	Valence	No. of Examples
1	high	high	72
2	low	high	51
3	low	low	56
4	high	low	61

Table 1: Quadrant-based class assignments of all MoodSwings Lite music clips.

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 contains many features that are popular in music information retrieval and speech processing encompassing both psychoacoustic as well as music-theoretic representations.

4.1 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCCs) are among the most widely used acoustic features in speech and audio processing. MFCCs are essentially a low-dimensional representation of the spectrum warped according to the mel-scale, which reflects the nonlinear frequency sensitivity of the human auditory system [13].

4.2 Chroma

The chromagram is a well-established method for estimating the western pitch class components within a short time-interval [14]. It is essentially a circular version of the logarithmically warped spectrogram, where the frequencies corresponding to chroma in different octaves are grouped together and summed to estimate the energy at each of the 12 pitch classes. Using this feature, it is sometimes possible to obtain an indication of the overall musical key and modality.

4.3 Statistical Spectrum Descriptors (SSD)

In music and audio processing, statistical spectrum descriptors are often related to timbral texture [15]. For each spectral shape function, we begin by dividing the data into short-overlapping segments, applying a Hanning window, and computing the magnitude DFT. A short explanation of each of the SSD features can be seen in Table 2.

Feature	Description
Centroid	The weighted-average (center of mass) of the spectrum
Flux	The Euclidean distance between successive spectral frames
Rolloff	The frequency beneath which a given proportion of the total spectral energy lies, typically 85%
Flatness	Quantifies how close the spectral distribution is to uniform (white)

Table 2: Description of spectral shape features.

4.4 Octave-Based Spectral Contrast

Many spectral features perform some averaging of the spectral distribution, which results in a loss of spectral information (note that two different spectra can yield very similar results for many spectral shape features). Spectral contrast features provide 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 [16].

5. EXPERIMENTS AND RESULTS

In order to properly make the case for regression methods, we first investigated the use of classification techniques involving discrete emotion classes based upon the four quadrants of the A-V space. We next investigated the emotion regression of the same clips using traditional least-squares regression as well as support vector regression (SVR). For both classification and regression, we averaged feature dimensions across all frames of a given 15-second music clip, thus representing each clip with a single vector of features. Although this is a significant reduction, it provides a consistent representation that facilitates direct comparisons between the various classification and regression methods.

Since it is our ultimate goal to track the emotional content over time, higher-order statistics of the features become less and less meaningful as we shorten the window length. In emotion tracking we investigated regression performance at shorter window lengths to develop a system that we ultimately implemented back into MoodSwings as the second player "AI" for single player games.

In all experiments, classification and regression, we divided the MoodSwings Lite corpus 70%/30% between training and testing samples. To avoid the "album-effect" [17, 18], 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 Classification

Support vector machines (SVMs) were chosen as our primary classification method, based upon their past successful application to similar music classification tasks (e.g., artist and genre classification) [19]. Using kernel methods, SVMs can be used to construct non-linear decision boundaries and have proven to be robust to some types of noise [20, 21, 22]. Given the binary nature of the SVM, to solve our four-class problem we implemented four one-versus-all SVMs, choosing the binary classifier providing the highest confidence as the class estimate.

Shown in Table 3 are the 4-way classification results using the individual acoustic feature sets, as well as stacking all features and stacking only MFCCs and spectral contrast. As there is no single dominant feature for emotion classification, it is expected that a method which fuses multiple features is necessary to obtain higher performance.

Feature Type	Accuracy
MFCC	$47.74 \pm 5.31\%$
Chromagram	$38.97 \pm 5.60\%$
Spectral Shape	$36.99 \pm 4.79\%$
Spectral Contrast	$48.67 \pm 6.10\%$
All Features Stacked	$38.24 \pm 4.60\%$
MFCC & Spec. Contrast Stacked	$\mathbf{50.18 \pm 4.18\%}$

Table 3: Results for four-way mood classification.

In dividing the data into discrete classes, clips for which A-V labels are in fact quite similar may be categorized into completely different classes. Such severe quantization of essentially continuous label data is likely the primary factor resulting in generally poor 4-way classification performance.

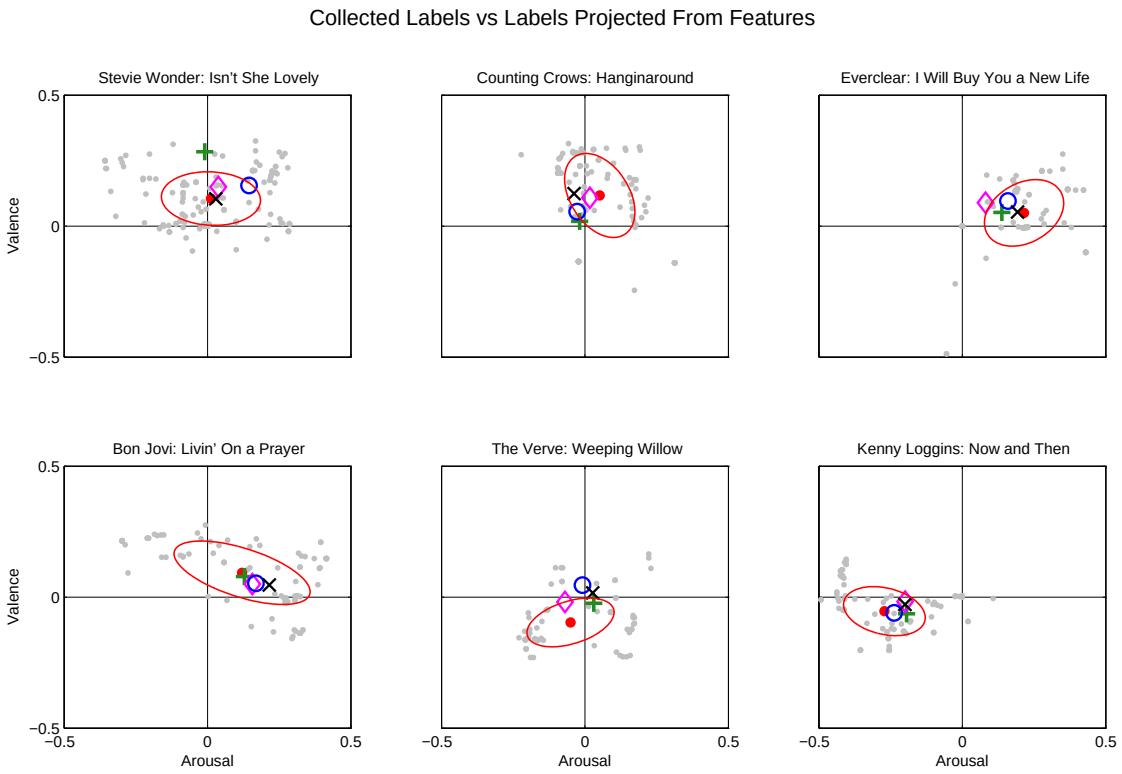

Figure 2: Collected A-V labels and projections resulting from regression analysis. A-V labels: second-by-second labels per song (gray •), μ of collected labels (red •), σ of collected labels (red ellipse). Projections: least-squares spectral contrast projection (green +), least-squares MFCC projection (magenta ⋄), SVR spectral contrast projection (blue O), and SVR MFCC projection (black X).

5.2 Regression

Given the continuous nature of the collected A-V labels and the myriad problems produced by discrete emotional classes, we investigated multiple regression methods for mapping acoustic features into the A-V space. We start by performing regression on the same 15-second clips, but then move to regression over time, investigating multiple window lengths for optimal performance, and finally integrate a regression system back into MoodSwings as the second player "AI".

As with classification, we implement supervised methods for our regression, using least-squares and support vector regression (SVR) [10] methods to create optimal projections from mean acoustic features to mean A-V values. There are many possible methods for evaluating regression performance. Our primary performance metric is the average Euclidean distance between the projected coordinates and the collected A-V labels, as a normalized percentage of the A-V space. 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 prediction for each sample. The resulting distance to the corresponding A-V label was compared to that of another randomly selected A-V label 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.

5.2.1 Feature Selection

Without the need for forcing the data into discrete classes, we achieve qualitatively higher performance. For single feature sets, regression using least-squares regression and spectral contrast features results in the smallest average deviation (14.90%) from the mean labels of the test samples. The corresponding T-test value for this case (5.28), given the degrees of freedom (72 test data samples), provides >99.9% confidence of statistical significance.

Support vector regression using single acoustic features results in a smaller average deviation in almost all cases, with the highest performance (lowest average distance) using MFCCs. Again, the T-test indicates very high confidence of statistical significance.

The primary advantage of regression over classification is for A-V labels close to an axis, where a very accurate regression projection can still lead to a misclassification, according to the discrete class labels. Shown in Figure 2 is the projection of six 15-second clips into the arousal-valence space resulting from multiple regression methods and acoustic features. The performance of the regression can be evaluated both in terms of the distance from the mean of the collected labels and also whether or not the regression label

Least-Squares Regression			
Feature	Avg. Distance	Avg. Rand. Dist.	T-test
MFCC	0.158 ± 0.008	0.232 ± 0.015	4.271
Chroma	0.197 ± 0.009	0.207 ± 0.010	0.591
S. Shape	0.163 ± 0.009	0.222 ± 0.011	3.557
S. Contrast	$\mathbf{0.149 \pm 0.010}$	$\mathbf{0.238 \pm 0.014}$	**5.280**
All Feat. Stacked	0.154 ± 0.008	0.256 ± 0.015	5.796
MFCC & S.C.	0.159 ± 0.008	0.248 ± 0.015	5.116
Support Vector Regression			
MFCC	$\mathbf{0.138 \pm 0.007}$	$\mathbf{0.235 \pm 0.014}$	**5.538**
Chroma	0.178 ± 0.009	0.213 ± 0.012	2.149
S. Shape	0.170 ± 0.010	0.231 ± 0.014	3.422
S. Contrast	0.146 ± 0.009	0.231 ± 0.016	5.028
All Feat. Stacked	0.169 ± 0.008	0.237 ± 0.015	3.723
MFCC & S.C.	0.141 ± 0.007	0.241 ± 0.013	5.629

Table 4: Regression results for individual sets of acoustic features.

falls within the first standard deviation of the labels (shown as an ellipse).

5.2.2 *Multi-level Regression*

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. We implemented 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: 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 will refer to these two topologies as multi-level separate and multi-level combined.

In all cases the secondary regressors employ linear least-squares and 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 5.

Least-Squares Regression			
Topology	Avg. Distance	Avg. Rand. Dist.	T-test
Seperate	0.144 ± 0.006	0.233 ± 0.016	5.246
Combined	$\mathbf{0.144 \pm 0.006}$	$\mathbf{0.232 \pm 0.013}$	**5.317**
Support Vector Regression			
Seperate	0.138 ± 0.006	0.235 ± 0.014	5.451
Combined	$\mathbf{0.137 \pm 0.005}$	$\mathbf{0.213 \pm 0.012}$	**5.896**

Table 5: Multi-layer regression results.

5.2.3 *Emotion Tracking Over Time*

Although the projection from acoustic features to A-V coordinates is quite promising, the ultimate goal of our research is to be able to track the changes of emotion within music *over time*. Shown in Figure 3 is the performance of spectral contrast regression for varying window lengths. It can be seen that it is possible to obtain the short-time information with only minimal loss in individual frame accuracy.

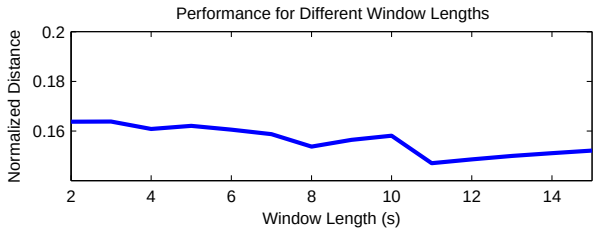

Figure 3: Performance data for varying window lengths.

Using these shorter window lengths we sought to implement our regression system to track the emotional changes within our clips over time. Shown in Figure 4 are the projections of three clips of audio obtained using only spectral contrast features and our most most simple (least-squares) regression technique. In this example, each clip has been broken down into three five second examples. Each individual label, which is shown as a dot, gets darker over time to show the overall emotional shift of the song. Of each five-second segment, we compute the mean as our ground truth and display that as a red X. Using our trained emotion regressor, we use the acoustic features from the five second chunks and project the predicted A-V position over time as well (blue O). As we continue to expand the label collection and pursue more intelligent feature fusion methods for regression, we are working towards the goal of accurately reflecting changing emotions within any song on a short-time basis.

5.2.4 *MoodSwings "AI"*

To further test our system, and to obtain additional human feedback, we have incorporated a simplified version of our regression into MoodSwings to compensate for one of our most major issues with the game–that oftentimes no human partners are available online. Single-player games can be played using "partner" labels recorded from a prior game, but this eliminates songs for which annotations have not been previously recorded. Our previous solution was to intersperse use of an "AI" opponent that generates random labels centered around the player's position, which is relatively easy to detect and can be highly frustrating for a player. In our new solution, we implement our regression system limiting ourselves to only simple least-squares regression and spectral contrast features. During gameplay, we provide the second player's annotations at one-second intervals, which are projected from only the average spectral contrast over a two-second window. The "AI" is available for demo on the web anytime using MoodSwings Single Player[1] (SP), a limited version of the game which only uses the "AI" as the secondary player.

6. DISCUSSION AND FUTURE WORK

In working with a continuously labeled space such as A-V it is clear that regression provides a more informative

[1]MoodSwings SP: `http://music.ece.drexel.edu/mssp`

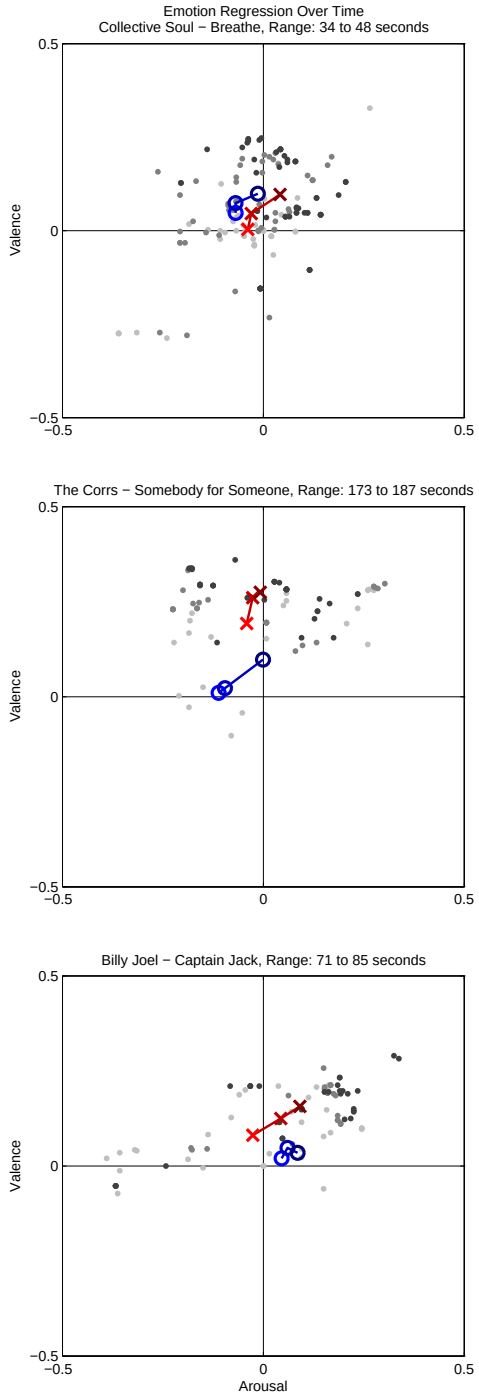

Figure 4: A-V labels and projections over time for three example 15-second music clips (markers become darker as time advances): second-by-second labels per song (gray •), mean of the collected labels over 5-second intervals (red X), and projection from acoustic features in 5-second intervals (blue O).

result over classification that is less sensitive to small variations (e.g., near the quadrant boundaries). In examining acoustic features for classification and regression, spectral contrast and MFCCs consistently provided the best performance across the classification and regression tasks, while spectral shape and chroma did not perform to expectations. MFCCs and spectral contrast capture different aspects of frequency-domain variation, so it is somewhat surprising that the combination of the two features improved classification but not regression performance. This is likely due to the "curse of dimensionality" (exponentially more data is required to fill a volume in feature space as more dimensions are added). Thus, adding feature dimensions, rather than leading to more informed decisions, could have hindered overall performance. The relative scaling of different features also presents problems, since variations in magnitudes may lead to artificially inflated or reduced contributions from particular features.

The regression results are quite promising, even using the most elementary techniques (least-squares). In order to improve our regression system, we plan to continue pursuing techniques to appropriately combine information from multiple acoustic features for audio mood estimation. That is, instead of choosing a subset of features or performing dimensionality reduction (e.g., principal components analysis) on a combined feature set, we train a separate system for each feature set and use ensemble methods to determine the relative contribution of each single-feature system to improve overall performance. In addition to (or in place of) multilevel regression, feature fusion methods could be used to combine information spaces. For example, the residual error of the trained regressors could be used as a measure of the confidence of each projection, which could weight higher and lower performing feature projections accordingly.

The collection of accurate labels is clearly a crucial component for developing systems to organize music by emotion. We believe that these experiments demonstrate the MoodSwings game to be a powerful tool for such data collection, turning the labeling process into a fun activity, and capturing information reflecting the time-varying nature of musical mood. We believe that the performance of the current regression system could be believable as an anonymous "human" partner, and greatly improves our ability to collect additional data when a secondary partner is not available.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] Y. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proc. International Conference on Music Information Retrieval*, Philadelphia, PA, September 2008.

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

[3] T. Li and M. Ogihara, "Detecting emotion in music," in *Proceedings of the 4th International Conference on Music Information Retreival*, Baltimore, MD, October 2003.

[4] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[5] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "What is the best segment duration for music mood analysis?"

International Workshop on Content-Based Multimedia Indexing (CBMI 2008), pp. 17–24, May 2008.

[6] J. S. Downie, "The 2007 MIREX results overview," MIREX 2007. [Online]. Available: http://www.music-ir.org/mirex2007/

[7] 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.

[8] X. Hu, J. Downie, C. Laurier, M. Bay, and A. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," *Proceedings of the 9th International Conference on Music Information Retrieval*.

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

[10] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[11] D. Shrestha and D. Solomatine, "Experiments with adaboost. rt, an improved boosting scheme for regression," *Neural computation*, vol. 18, no. 7, pp. 1678–1710, 2006.

[12] D. P. W. Ellis, A. Berenzweig, and B. Whitman, "The "uspop2002" pop music data set." [Online]. Available: http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

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

[14] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2001.

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

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

[17] Y. Kim, D. Williamson, and S. Pilli, "Towards quantifying the "album effect" in artist identification," in *Proceedings of ISMIR 2006 Seventh International Conference on Music Information Retrieval*, September 2006.

[18] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Johannes Kepler University, Linz, March 2006.

[19] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis, "Support vector machine active learning for music retrieval," *Multimedia Systems*, vol. 12, no. 1, pp. 3–13, Aug 2006.

[20] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[21] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–295, 1995.