Predicting Time-Varying Musical Emotion Distributions from Multi-Track Audio

Jeffrey Scott, Erik M. Schmidt, Matthew Prockup, Brandon Morton, and Youngmoo E. Kim

Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
{jjscott, eschmidt, mprockup, bmorton, ykim}@drexel.edu

Abstract. Music exists primarily as a medium for the expression of emotions, but quantifying such emotional content empirically proves a very difficult task. Myriad features comprise emotion, and as such music theory provides no rigorous foundation for analysis (e.g. key, mode, tempo, harmony, timbre, and loudness all play some roll), and the weight of individual musical features may vary due to the expressiveness of different performers. In previous work, we have shown that the ambiguities of emotions make the determination of a single, unequivocal response label for the mood of a piece of music unrealistic, and we have instead chosen to model human response labels to music in the arousal-valence (A-V) representation of affect as a *stochastic distribution*. Using multi-track sources, we seek to better understand these distributions by analyzing our content at the performer level for different instruments, thus allowing the use of instrument-level features and the ability to isolate affect as a result of different performers. Following from the time-varying nature of music, we analyze 30-second clips on one-second intervals, investigating several regression techniques for the automatic parameterization of emotion-space distributions from acoustic data. We compare the results of the individual instruments to the predictions from the entire instrument mixture as well as ensemble methods used to combine the individual regressors from the separate instruments.

Keywords: emotion, mood, machine learning, regression, music, multi-track

1 Introduction

There has been a growing interest in the music information retrieval (Music-IR) research community gravitating towards methods to model and predict musical emotion using both content based and semantic methods [1]. It is natural for humans to organize music in terms of emotional associations, and the recent explosion of vast and easily accessible music libraries has created high demand for automated tools for cataloging, classifying and exploring large volumes of music content. Crowdsourcing methods provide very promising results, but do not perform well outside of music that is highly-popular, and therefore leave much

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

to be desired given the long-tailed distribution of music popularity. The recent surge of investigations applying content-based methods to model and predict emotional affect have generally focused on combining several feature domains (e.g. loudness, timbre, harmony, rhythm), in some cases as many as possible, and performing dimensionality reduction techniques such as principal component analysis (PCA). While using these methods may in many cases provide enhanced classification performance, they provide little help in understanding the contribution of these features to musical emotion.

In this paper, we employ multi-track sources for music emotion recognition, allowing us to extract instrument-level acoustic features while avoiding corruption that would usually occur as a result of noise induced by the other instruments. The perceptual nature of musical emotion necessarily requires supervised machine learning, and we therefore collect time-varying ground truth data for all of our multi-track files. As in previous work, we collect data via a Mechanical Turk human intelligence task (HIT) where participants are paid to provide time-varying annotations in arousal-valence (A-V) model of affect, where valence indicates positive vs negative emotion, and arousal indicates emotional intensity [2]. In this initial investigation we obtain these annotations on our full multi-track audio files, thus framing the task as predicting the mixed emotion from the individual instrument sources. Furthermore, we model our collected A-V data for each moment in a song as a *stochastic distribution*, and find that the labels can be well represented as a two-dimensional A-V Gaussian distribution.

In isolating specific instruments we gain the ability to extract specific acoustic features targeted at each instrument, allowing us to find the most informative domain for each. In addition, we also isolate specific performers, potentially allowing us to take into account performer-level affect as a result of musical expression. We build upon our previous work modeling time-varying emotion-space distributions, and seek to develop new models to best combine this multi-track data [3–5]. We investigate multiple methods for automatically parameterizing an A-V Gaussian distribution, effectively creating functional mappings from acoustic features directly to emotion space distribution parameters.

2 Background

Prior work in modeling musical emotion has explored content based and semantic methods as well as combinations of both models [1]. Much of the work in content based methods focuses on training supervised machine learning models to predict classes of emotion, such as happy, joyful, sad or depressed. Several works also attempt to classify songs into discretized regions of the arousal-valence mood space [6–8].

In addition to classification, several authors have successfully applied regression methods to project from high dimensional acoustic feature vectors directly into the two dimensional A-V space [9, 8]. To our knowledge, no one has attempted to leverage the separate audio streams available in multi-track recordings to enhance emotion prediction using content based methods.

3 Dataset

We selected 50 songs spanning 50 unique artists from the RockBand® game and created five monaural stem files for each song. This is the same dataset (plus 2 additional songs) that we used in a previous paper for performing analyses on multi-track data[10, 11]. A stem may contain one or more instruments from a single instrument class. For example, the vocal track may have one lead voice or a lead and harmony or even several harmonies as well as doubles of those harmonies. Each stem only contains one instrument class (i.e. bass, drums, vocals) excepting the backup track which can contain audio from more than one instrument class. For each song there are a total of six audio files - backup, bass, drums, guitar, vocals and the full mix, which is a linear combination of the individual instruments.

To label the data, we employed an annotation process based on the MoodSwings game outlined in [2]. We used Amazon's Mechanical Turk and rejected the data of users who did not pass the verification criteria of consistent labeling on the same song and similarity to expert annotations. For the 50 songs in our corpus there is an average of 18.48 ± 3.05 labels for each second with a maximum of 25 and a minimum of 12. A 40 second clip was selected for each song and the data of the first 10 seconds was discarded due to the time it takes a user to decide on the emotional content of the song [12]. As a result, we are using 30 second clips for our time varying prediction of musical emotion distributions.

4 Experiments

The experiments we perform are similar in scope to those presented in a previous paper which utilized a different dataset [4]. This allows us to verify that we attain comparable results using instrument mixtures and provides a baseline to compare the results from the audio content of individual instruments.

4.1 Overview

Acoustic features are extracted from each of the five individual instrument files as well as the final mix and are described in more detail in Section 4.2. We use linear regression to calculate the projection from the feature domain of each track to the parameters of the Gaussian distribution that models the labels at a given time.

$$[f_1^{(t)} \cdots f_m^{(t)}]\, \mathbf{W}_t = [\mu_v^{(t)}\ \mu_a^{(t)}\ \mathbf{\Sigma}_{11}^{(t)}\ \mathbf{\Sigma}_{12}^{(t)}\ \mathbf{\Sigma}_{22}^{(t)}] \tag{1}$$

Here $[f_1^{(t)} \cdots f_t^{(t)}]$ are the acoustic features, $\mathbf{W}_t$ is the projection matrix, μ_a and μ_v are the means of the arousal and valence dimensions, respectively, and $\mathbf{\Sigma}$ is the 2×2 covariance matrix. For an unknown song, $\mathbf{W}_t$ is used to predict the distribution parameters in the A-V space from the features for track t. The regressor for each track can be used on its own to predict A-V means and covariances.

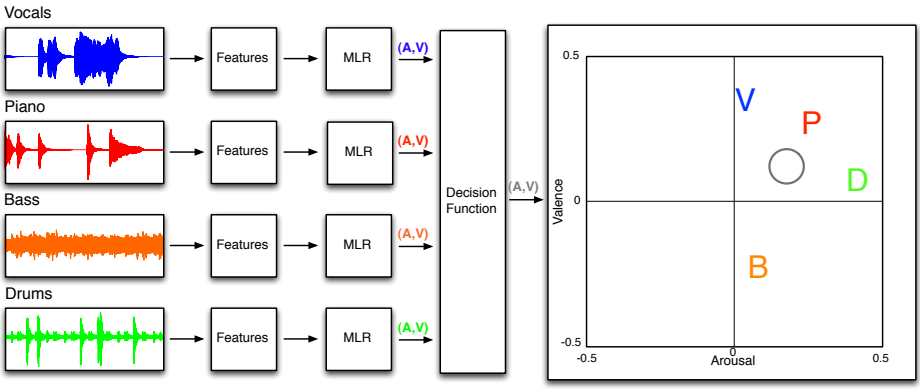


Fig. 1: Acoustic features are computed on each individual instrument file and a regression matrix is computed to project from features to a distribution in the A-V space. A different distribution is computed for each instrument (B/D/P/V) and the mean of the distribution parameters (gray circle) is used as the final A-V distribution.

We also investigate combinations of the individual regressors to reduce the error produced by a single instrument model. In these cases, the final prediction is a weighted combination of the predictions from each individual regressor

$$\boldsymbol{\theta} = \sum_{k=1}^{K} \pi_K \theta_K \tag{2}$$

where $\theta = [\mu_v \ \mu_a \ \boldsymbol{\Sigma}_{11} \ \boldsymbol{\Sigma}_{12} \ \boldsymbol{\Sigma}_{22}]$ and π_k is the mixture coefficient for each regressor. In this paper, we try the simplest case which averages the predicted distribution parameters to produce the final distribution parameter vector. Figure 1 depicts the test process for an unknown song.

Having a small dataset of only 50 songs, we perform leave-one-out cross validation (LOOCV), training on 49 songs and testing on the remaining song. This process is repeated until every song has been used as a test song.

4.2 Acoustic Features

We investigate the performance of a variety of acoustic features that are typically used throughout the music information retrieval (Music-IR) community including MFCCs, chroma, spectrum statistics and spectral contrast features. The audio files are down-sampled to 22050 Hz and the features are aggregated over one second windows to align with the second by second labels attained from the annotation task. Table 1 lists the features used in our experiments [13–16].

Feature	Description
MFCC	Mel-Frequency Cepstral Coefficients (20 dimensions)
Chroma Autocorrelation	The autocorrelation of the 12 dimensional chroma vector
Spectral Contrast	Energy in spectral peaks and valleys
Statistical Spectrum Descriptors	Statistics of the spectrum (spectral shape)

Table 1: Acoustic features used in the experiments.

5 Results

We perform experiments using the audio of individual instruments, the full instrument mixture and combinations of the individual instruments. We also compare the results of using different features for each track.

Table 2 shows the results for the regressors trained on individual instruments. The mean average error is the average euclidean distance of the predicted mean of the distribution from the true mean of the distribution across all cross validation folds. Since we are modeling distributions and not just singular A-V coordinates, we also compute the one-way Kullback-Liebler (KL) Divergence from the projected distribution to the true distribution of the collected A-V labels. The table shows the average KL divergence for each regressor averaged across all cross validation folds. We observe that the best regressor for bass, drums and vocals is attained using spectral contrast features and the best regressor for the backup and drum tracks is computed using spectral shape features. It is notable that chroma features perform particularly poor in terms of KL divergence but are only slightly worse than the other features at predicting the means of the distribution.

We also consider combinations of regressors which are detailed in Table 3. The 'Best Single' row shows the best performing single regressor in terms of A-V mean prediction using each feature. The second row in the table includes the results of averaging the predicted distribution parameters for all five individual instrument models for the given feature. Lastly, 'Final Mix' lists the average distance between the predicted and true A-V mean when projecting from features computed on the final mixed track. We note that averaging the models improves performance for all of the best single models excepting the spectral contrast feature. Comparing the averaged models to the prediction from the final mix, the averaged single instrument regressors perform better for MFCCs and spectral shape features but do not perform as well as the final mixes when using chroma or spectral contrast features.

In Figure 2 we see examples of both the predicted and actual distributions for a 30 second clip from the song *Hysteria* by Muse. Both the true and estimated distributions get darker over time as do the data points of the individual users. The predictions for the individual instruments (a-e) are shown along with the average of the predictions for all the instruments (f).

Feature	Instrument	Average Mean Distance	Average KL Divergence
MFCC	Backup	0.152 ± 0.083	1.89 ± 2.34
	Bass	0.141 ± 0.070	1.26 ± 1.29
	Drums	0.140 ± 0.075	1.17 ± 1.52
	Guitar	0.133 ± 0.066	1.22 ± 1.40
	Vocals	0.134 ± 0.071	1.41 ± 1.81
Spectral Contrast	Backup	0.145 ± 0.125	1.86 ± 5.93
	Bass	$\mathbf{0.140 \pm 0.076}$	$\mathbf{1.21 \pm 1.38}$
	Drums	0.139 ± 0.071	1.20 ± 1.88
	Guitar	$\mathbf{0.125 \pm 0.063}$	$\mathbf{1.06 \pm 1.42}$
	Vocals	$\mathbf{0.129 \pm 0.065}$	$\mathbf{1.00 \pm 1.32}$
Spectral Shape	Backup	$\mathbf{0.132 \pm 0.068}$	$\mathbf{1.25 \pm 1.91}$
	Bass	0.142 ± 0.071	1.31 ± 1.63
	Drums	$\mathbf{0.131 \pm 0.072}$	$\mathbf{1.03 \pm 1.38}$
	Guitar	0.134 ± 0.063	1.12 ± 1.42
	Vocals	0.133 ± 0.067	1.12 ± 1.47
Chroma	Backup	0.153 ± 0.084	10.85 ± 15.6
	Bass	0.159 ± 0.084	5.35 ± 6.13
	Drums	0.162 ± 0.089	2.87 ± 3.01
	Guitar	0.147 ± 0.074	2.66 ± 4.33
	Vocals	0.154 ± 0.078	5.99 ± 10.4

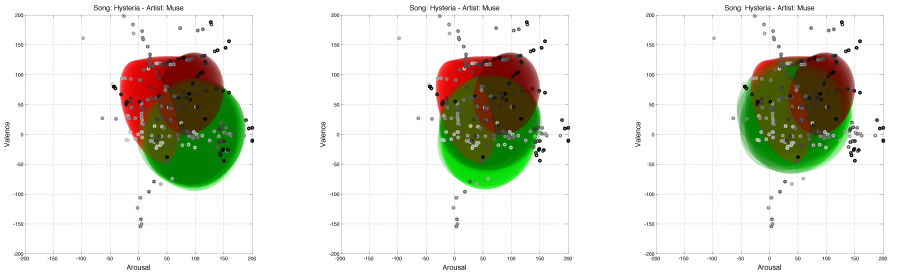
Table 2: Mean average error between actual and predicted means in the A-V coordinate space as well as Kullback-Leibler (KL) divergence between actual and predicted distributions. The value of the best performing feature for each instrument is in bold.

6 Discussion

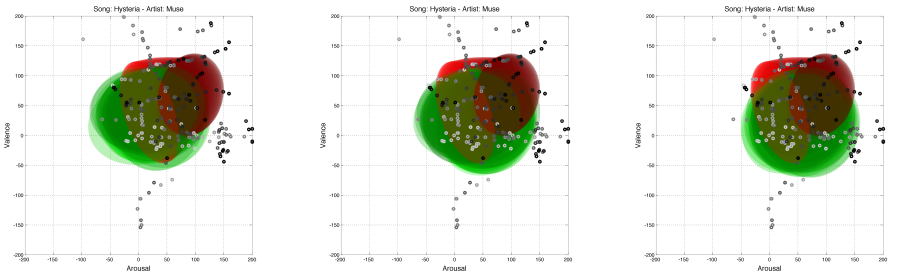
In this initial work we demonstrate the potential of utilizing multi-track representations of songs for modeling and predicting time varying musical emotion distributions. We achieved performance on par with what we have shown previously with a different corpus using similar techniques and a simple averaging of a set of regressors trained on individual instruments. Using more advanced techniques to determine the optimal combinations and weights of instruments and features could provide significant performance gains compared to averaging the output of all the models. There are a variety of ensemble methods for regres-

	Features			
	Chroma	Contrast	MFCC	Shape
Best Single	0.147 ± 0.074	0.125 ± 0.063	0.133 ± 0.066	0.131 ± 0.072
Avg Models	0.142 ± 0.075	0.126 ± 0.066	$\mathbf{0.124 \pm 0.061}$	$\mathbf{0.129 \pm 0.064}$
Final Mix	$\mathbf{0.141 \pm 0.073}$	$\mathbf{0.124 \pm 0.066}$	0.129 ± 0.069	0.132 ± 0.066

Table 3: Results from different combinations of single instrument regressors



(a) Bass (b) Backup (c) Guitar



(d) Drums (e) Vocal (f) Averaged Prediction

Fig. 2: Actual (red) and predicted (green) distributions for *Hysteria* by Muse. The color of the distribution gets darker over time as does the color of the individual data points.

sion that would be applicable to learning better feature and model combinations for regression in the A-V space. We hope to infer, from the results of such experiments, whether certain instruments contribute more to invoking emotional responses from humans.

The results shown in these experiments are encouraging, especially in the performance gains in the case of the MFCC features. An interesting result is that each individual instrument spectral contrast prediction performs better than that of MFCCs, but the MFCC multi-track combination is the top performer equal with spectral contrast on the full mix. This result highlights that the highest performing feature on a single track might not be the same one that offers the most new information to the aggregate track prediction. As a result, in future work we plan to investigate feature selection for this application, performing a number of experiments with different acoustic feature combinations to determine the best acoustic feature for each instrument in the multi-track prediction system.

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. J. Speck, E. Schmidt, and B. Morton, "A comparitive study of collaborative vs. traditional musical mood annotation," in *ISMIR*, Miami, FL, 2011.
3. 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.
4. E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *ISMIR*, Utrecht, Netherlands, 2010.
5. ——, "Prediction of time-varying musical mood distributions using Kalman filtering," in *IEEE ICMLA*, Washinton, D.C., 2010.
6. 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.
7. K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach," in *Proceedings of the 10th International Society for Music Information Conference*, Kobe, Japan, 2009.
8. B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "Smers: Music emotion recognition using support vector regression," in *ISMIR*, Kobe, Japan, 2009.
9. H. Chen and Y. Yang, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE TASLP*, no. 99, 2011.
10. J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," in *SMPC*, Padova, Italy, 2011.
11. J. Scott and Y. E. Kim, "Analysis of acoustice features for automated multi-track mixing," in *ISMIR*, Miami, Florida, 2011.
12. 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.
13. 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.
14. 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.
15. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE TASSP*, vol. 28, no. 4, 1980.
16. T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music." in *Proc. of the Intl. Computer Music Conf.*, 1999.