MUSIC EMOTION RECOGNITION WITH ADAPTIVE AGGREGATION OF GAUSSIAN PROCESS REGRESSORS

Satoru Fukayama and Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper describes a novel method for estimating the emotions elicited by a piece of music from its acoustic signals. Previous research in this field has centered on finding effective acoustic features and regression methods to relate features to emotions. The state-of-the-art method is based on a multi-stage regression, which aggregates the results from different regressors trained with training data. However, after training, the aggregation happens in a fixed way and cannot be adapted to acoustic signals with different musical properties. We propose a method that adapts the aggregation by taking into account new acoustic signal inputs. Since we cannot know the emotions elicited by new inputs beforehand, we need a way of adapting the aggregation weights. We do so by exploiting the deviation observed in the training data using Gaussian process regressions. We confirmed with an experiment comparing different aggregation approaches that our adaptive aggregation is effective in improving recognition accuracy.

Index Terms— Music emotion recognition, Gaussian process regression, Adaptive aggregation, Multi-level regression

1. INTRODUCTION

Music emotion recognition is a task to estimate how a piece of music affects the emotions of a listener. It works by extracting features from music audio and applying a regression model or classification model to map those features into an emotion representation. Psychological studies have proposed two-dimensional values in the Arousal-Valence (AV) plane for representing emotion [1]. The task we focus on in this research is to estimate the AV value from the musical audio input. Specifically, the AV values are estimated for segments of music lasting 30 seconds. This problem setting is the same as that of the Emotion in Music Task at MediaEval Workshops, which is the leading evaluation campaign on music emotion recognition [2, 3].

Previous methods have used carefully selected features to estimate emotions. Linear regression methods such as multivariate regression analysis have been used to map those features to emotions [4, 5]. Feature selection algorithms are also applied to find important features to improve the estimation [6]. Recently, emotion recognition has been improved by automatically tuning the importance of each feature instead of carefully choosing useful ones. One approach is multi-level regression which creates individual regressors using each feature and then aggregates the results from those regressors with another regressor [7, 8]. Another approach is to apply non-linear regression methods and make use of the non-linear dimensionality reduction. Powerful regressors such as neural networks [9, 10, 11], support vector regressions [12], and Gaussian process regressions [13] have been used.

Although state-of-the-art methods have significantly im-

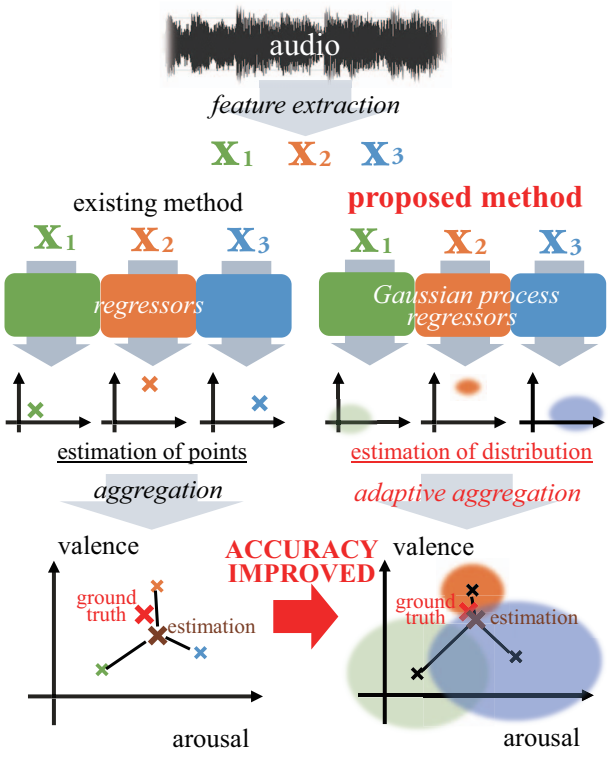

Fig. 1. Fixed aggregation vs. proposed adaptive aggregation of Gaussian process regressors for music emotion recognition.

proved performance, they cannot change the importance of each feature once it has been trained using training data. It is not clear whether all kinds of emotions on the AV plane can be estimated from a fixed importance for each feature. As we show in our research, the importance of features differs when estimating low-arousal and high-arousal areas. However, there is as yet no method that can adaptively change the importance of each feature depending on what emotion we want to estimate.

We propose a novel method of emotion estimation that can adjust the importance of each feature depending on the audio input. The technical difficulty here is the estimation of the importance before knowing the emotion of the audio input. If we know the emotion of the audio beforehand, one strategy is to adjust the importance by optimizing it to match the result to that emotion. However, in our case, the emotion can be known only once the estimation has been performed; hence, this strategy cannot be taken.

We overcome this difficulty by using Gaussian process (GP) regression. The overview and novelty of our method are shown in Figure 1. GP regression can predict the mean and variance of an estimated AV value from a new audio input. By preparing multiple GP regressors wherein each is trained with a different feature, we obtain the means and variances of the estimated AV values considering each feature. The feature is not important when the feature values in the training data take various values for similar AV values, and this leads to a large variance when an AV value is estimated with GP regression from a new piece of audio. Thus, we can obtain the importance of each feature by calculating the variance of AV values with individual GP regressors. Since the importance of each feature varies depending on the audio input, we aggregate the results from the individual regressors adaptively to estimate the emotion. The mathematical formulation for this procedure is described in Section 3, and in Section 4 we show that it can improve the performance of music emotion recognition.

2. RELATED WORK

Our research can be related to previous music emotion recognition methods from two viewpoints: selecting features and regression methods. In this section, we summarize some important insights from the review papers [14, 15, 16].

2.1. Features for emotion recognition

It is usually argued that AV values are related to the musical content, such as tempo, pitch, loudness, timbre for arousal, and mode and harmony for valence [17]. In particular, audio features, such as mel-frequency cepstral coefficients (MFCCs), chroma (i.e., pitch class profile), spectral descriptors (e.g., spectral contrast, centroid, flux, rolloff, and flatness) are used [6]. Text information is often combined with audio features [14]. For example, lyric and metadata such as genre tags and social media tags can be used [16, 18].

2.2. Regression methods

Emotion recognition is usually seen as a problem of regression between features and AV values. The regression can be linear or non-linear. Multivariate regression is a powerful linear method used in emotion recognition [4]. Non-linear regressors include support vector regression [12], deep recurrent neural networks [19, 10], continuous conditional random fields [11], conditional random field [20], k-nearest neighbors [21] and Gaussian process [13].

Multi-level regression is used to gather information from individual regressors which are trained with individual feature. Schmidt et al. proposed fusing results from individual regressors to improve performance [7]. They attempted two different fusion topologies: in one case, the secondary regressors receive only AV estimates respectively; in the other, the secondary regressors receive both arousal and valence values from the first-level regressors.

3. ADAPTIVE AGGREGATION OF GAUSSIAN PROCESS REGRESSORS

Given an excerpt of music audio with a fixed length, we want to estimate the AV values by aggregating the predictions of N different features $\mathbf{x_1}, \ldots, \mathbf{x_N}$. Since the same discussion applies to estimating the arousal and the valence, we will only discuss the case of estimating one of them, and will represent its value as y. The prediction from the n^{th} feature is y_n, and our goal is to find an appropriate weight w_n for this feature.

3.1. Quantifying the importance of features

Let us think of a transform which maps the n^{th} feature $\mathbf{x_n}$ into an emotion y_n. If we know that the ground truth of emotion for the given audio is y^*, the estimation error ϵ_n for this feature can be defined as $\epsilon_n = y^* - y_n$.

Suppose we have two different estimates for emotion, y_n and y_m, from the n^{th} and m^{th} features, respectively. If ϵ_n is more likely to take a value closer to 0 compared with ϵ_m, we can say the n^{th} feature is more important than the m^{th} feature in estimating the true value y^*. To quantify the importance of features, we assume that every ϵ_n follows a Gaussian distribution: $\epsilon_n \sim \mathcal{N}\left(0, \sigma_n^2\right)$. The n^{th} feature is more important in estimation than the m^{th} feature when $\sigma_n^2 < \sigma_m^2$.

3.2. Adaptive aggregation

Now we will show how to aggregate the estimation results obtained from N different features. From the assumption that each estimation error $\epsilon_n, (n = 1, \ldots, N)$ follows a Gaussian

distribution, we can derive N probability distributions:

$$P_n(y) = \mathcal{N}\left(y_n, \sigma_n^2\right), n = 1, \ldots, N \quad (1)$$

The maximum likelihood estimation of y can be understood to be a value that maximizes the joint probability P_{J}, which is calculated by multiplying $P_n(y)$ for all $n = 1, \ldots, N$ [22]. We get

$$P_{\mathrm{J}}(y) = \prod_{n=1}^{N} P_n(y) \propto \frac{1}{\sigma_1 \cdots \sigma_N} \exp\left(-\frac{1}{2}\xi^2\right) \quad (2)$$

where

$$\xi^2 = \sum_{n=1}^{N} \frac{(y - y_n)^2}{\sigma_n^2}. \quad (3)$$

To maximize $P_{\mathrm{J}}(y)$ with respect to y, we need to minimize ξ^2. By solving $\frac{\mathrm{d}\xi^2}{\mathrm{d}y} = 0$, we obtain the maximum likelihood estimation of y:

$$y = \frac{1}{W} \sum_{n=1}^{N} w_n y_n, \quad w_n = \frac{1}{\sigma_n^2}, \quad (4)$$

where $W = \sum_{n=1}^{N} w_n$. We have thus estimated the emotion y as a weighted average wherein the weights are the normalized inversed squares of the variances.

3.3. Gaussian process regressors

The discussion so far indicates that if we can get Gaussian distributions $P_n(y), n = 1, \ldots, N$ as equation (1), we can calculate the optimal estimation of y. This final subsection describes how we can obtain those distributions from training data by using Gaussian process regression.

Let $\{\mathbf{x}_n^{(1)} \ldots \mathbf{x}_n^{(K)}, y_n^{(1)} \ldots y_n^{(K)}\}$ be the training data having K data points for the n^{th} feature. Moreover, let $\mathbf{y}_n$ be a vector containing the training data, $\mathbf{y}_n = \left(y_n^{(1)} \ldots y_n^{(K)}\right)^{\mathrm{T}}$, where T denotes the transpose of a vector. According to the derivation of GP regression [13], we get a Gaussian distribution $P_n(y) = \mathcal{N}\left(y_n, \sigma_n^2\right)$ where the mean y_n and the variance σ_n^2 are

$$y_n = \mathbf{k}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_{\mathrm{obs}}^2 \mathbf{I}\right)^{-1} \mathbf{y}_n, \quad (5)$$

$$\sigma_n^2 = k(\mathbf{x}_n, \mathbf{x}_n) + \sigma_{\mathrm{obs}}^2 - \mathbf{k}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_{\mathrm{obs}}^2 \mathbf{I}\right)^{-1} \mathbf{k}_*. \quad (6)$$

Here, $k(\cdot, \cdot)$ is a squared-exponential covariance function:

$$k\left(\mathbf{x}_n^{(i)}, \mathbf{x}_n^{(j)}\right) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}\left(\mathbf{x}_n^{(i)} - \mathbf{x}_n^{(j)}\right)^2\right), \quad (7)$$

where $\mathbf{I}$ denotes the identity matrix,
and $\mathbf{k}_* = \left(k\left(\mathbf{x}_n, \mathbf{x}_n^{(1)}\right) \ldots k\left(\mathbf{x}_n, \mathbf{x}_n^{(K)}\right)\right)$, $\mathbf{K} = K_{ij} = k\left(\mathbf{x}_n^{(i)}, \mathbf{x}_n^{(j)}\right)$. The parameters σ_{obs}, σ_{f} and l are optimized through training to maximize the log-likelihood of the training data.

4. EXPERIMENT ON MUSIC EMOTION RECOGNITION

We conducted an experiment to verify the effectiveness of our method. We estimated the AV values from fixed lengths (30 seconds) of music audio. We compared the emotion recognition accuracies obtained with different methods.

4.1. Compared methods

There were two points of comparison. First, we compared the performance of our proposed adaptive aggregation approach to that of an existing multi-level regression approach [7, 8]. Second, we compared both of these to a baseline which does not use aggregation. All three methods tested were based on GP regression. They are:

(1) Adaptive aggregation of results from GP regressors (the proposed method)

(2) Fixed aggregation of results from GP regressors using multivariate regression, and

(3) GP regression using all features as a single feature (no aggregation).

4.2. Experiment procedure

We followed three steps in our experiment. First, low-level audio descriptors related to emotion were extracted from the audio. Specifically, we used a set of descriptors from the state-of-the-art method that came in first place at the MediaEval Emotion in Music task in 2014 [10]. This is the official set of descriptors used in the 2013 INTERSPEECH Computational Paralinguistics Challenge (ComPareE) [23]. It contains 65 types of descriptors, their first-order derivatives, and statistics (mean, standard deviation, skewness, max., min.) calculated from an audio clip. As a result, we obtained a 6373-dimensional vector describing the content of the audio.

Second, we generated features and constructed GP regressors. We created 3 different feature sets by dividing the original 6373-dimensional vector into three, collecting those descriptors related to the spectral descriptors, MFCCs, and others, respectively. To construct the GP regressors, we used the Emotion in Music Database as training data, which consists of 744 audio clips and annotations of AV values on a scale from 1 to 9 [24], and we normalized them to a scale from -1.0 to 1.0. We used 619 randomly chosen clips as training data and the rest (125 clips) as evaluation data. The Constrained Optimization BY Linear Approximation (COBYLA) method [25] was used to optimize the GP parameters.

Finally, the results from the regressors were aggregated to obtain the AV value estimate. We calculated the AV value as a weighted average of the means of the estimated Gaussian distributions from the GP regressors, wherein the weights were set to the normalized inversed square of the variances. When

method	arousal		valence	
	R^2	RMSE	R^2	RMSE
(1) adaptive aggregation of GP regressors	**0.636 ± 0.040**	**0.206 ± 0.014**	**0.413 ± 0.043**	**0.230 ± 0.011**
(2) fixed aggregation of GP regressors	0.619 ± 0.042	0.211 ± 0.014	0.397 ± 0.064	0.234 ± 0.013
(3) GP regression	0.488 ± 0.027	0.238 ± 0.053	0.399 ± 0.073	0.233 ± 0.012

Table 1. Comparison of methods of estimating emotion from audio in the MediaEval 2013 Emotion in Music development set and 10-fold cross-validation. The proposed adaptive method, evaluated with the R-squared value (official score in MediaEval task), outperformed the fixed aggregation method by 2.7% for estimating arousal and 4.0% for valence.

conducting the fixed aggregation, we split the training data (619 clips) into 309 clips and 310 clips and used the former clips to construct the GP regressors and the latter ones for the multivariate regression between the results from the GP regressors and ground-truth.

4.3. Evaluation metrics

We calculated the R-squared value (R^2) and the root mean squared error (RMSE) by following the official evaluation scheme used in the MediaEval Emotion in Music task. Since R^2 has several definitions, we chose the one which was used in the previous music emotion recognition research [13]: $R^2 = 1 - \frac{\sum_k \left(y^{*(k)} - y^{(k)}\right)^2}{\sum_k \left(y^{*(k)} - \bar{y^*}\right)^2}$, where $y^{*(k)}$ is the ground truth for the k^{th} audio clip, and $\bar{y^*}$ is the mean of the ground truth values. The estimated values are close to the ground truth data when R^2 is close to 1.0, and RMSE is close to 0.0. Both these measures were calculated with 10 different combinations of training data and evaluation data, and the means and standard deviations of the values were used to measure the performance (10-fold cross-validation).

4.4. Results

The results shown in Table 1 indicate that adaptive aggregation of regressors improved accuracy. For both arousal and valence estimations, adaptive aggregation of GP regressors performed the best out of the three methods. Comparing adaptive aggregation with fixed aggregation, we can see the R-squared value for adaptive aggregation had a 2.7% improvement for arousal and a 4.0% improvement for valence. For arousal, both aggregation approaches performed better than the GP regression using all features as a single feature, but for valence, only the proposed adaptive aggregation method gave an improvement.

5. DISCUSSION

We confirmed that adaptive aggregation of regressors can improve the performance of music emotion recognition. GP regression was suitable for our aggregation methodology since it can predict not only the estimation values but also the variances.

The results imply that the importance of each feature set can differ depending on the emotion that we are estimating. We actually found that the feature set most important for estimating arousal value around 0.425 was the MFCC-related feature set (weight 0.52), whereas for estimating arousal value around -0.5, the spectral-related feature set had the biggest importance (weight 0.60). Although we could not find a simple explanation for which is the most important feature set in which part of the AV plane, we confirmed that using fixed weights for aggregating the regressors is too coarse an approximation to reflect the relative importance of each feature set at all points in the plane.

Our method matched the performance of one state-of-the-art approach, but did not outperform it. The best performance using the same dataset reported so far is $R^2 = 0.704$ for estimating arousal by using deep recurrent neural networks [19]. However, considering that the highest performance using a single GP regressor nearly matches that one ($R^2 = 0.695 \pm 0.046$) with a novel selection of features and kernel functions [13], trying various features and kernel functions may eventually yield even higher estimation performance.

The way of determining the optimal number of feature sets for aggregation remains to be investigated. Although adding new feature sets can increase the adaptiveness of our method, adding too many feature sets by dividing a feature set into too many sets can decrease the benefit of the non-linearity of GP regression.

6. CONCLUSION

We proposed a music emotion recognition method based on a novel aggregation of Gaussian process regressors. The method exploits the variance of the training data and adapts the weights for aggregating the results depending on the input audio. Experimental results showed that our method could actually improve the estimation performance. Our future work includes investigating the effect of varying the number of feature sets.

Acknowledgements

The authors thank Jordan B. L. Smith for helping to polish this paper. This work was supported in part by OngaCREST, JST.

7. REFERENCES

[1] J. A. Russel, "A circumplex model of affect," *Journal of Personal Social Psychology*, vol. 39, pp. 1161–1178, June 1980.

[2] M. Soleymani, M. N. Caro, E. M. Schmidt, and Y.-H. Yang, "The MediaEval 2013 Brave new Task: Emotion in Music," in *Proceedings of MediaEval 2013 Workshop*, 2013.

[3] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in Music task at MediaEval 2014," in *Proceedings of MediaEval 2014 Workshop*, 2014.

[4] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of ISMIR 2009*, 2009, pp. 621–626.

[5] A. Aljanaki, F. Wiering, and R. C. Veltkamp, "MIRUtecht participation in MediaEval 2013: emotion in music task," in *Proceedings of MediaEval 2013 Workshop*, 2013.

[6] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of MIR 2010*, 2010, pp. 267–273.

[7] E. M. Schmidt, F. Eyben, and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proceedings of ISMIR 2010*, 2010, pp. 465–470.

[8] Y. Fan and M. Xu, "MediaEval 2014: THU-HCSIL approach to emotion in music task using multi-level regression," in *Proceedings of MediaEval 2014 Workshop*, 2014.

[9] F. Weninger, F. Eyben, and B. Schuller, "The TUM approach to the MediaEval music emotion task using generic affective audio features," in *Proceedings of MediaEval 2013 Workshop*, 2013.

[10] E. Coutinho, F. Weninger, B. Schuller, and K. R. Scherer, "The Munich LSTM-RNN approach to the MediaEval 2014 ´Emotion in Music´ task," in *Proceedings of MediaEval 2014 Workshop*, 2014.

[11] V. Imbrasaite and P. Robinson, "Music emotion tracking with continuous conditional neural fields and relative representation," in *Proceedings of MediaEval 2014 Workshop*, 2014.

[12] B. Han, S. Rho, R. B. Dannenberg, and E. Hwang, "SMERS: music emotion recognition using support vector regression," in *Proceedings of ISMIR 2009*, 2009, pp. 651–656.

[13] K. Markov and T. Matsui, "Music genre and emotion recognition using Gaussian process," *IEEE Access*, vol. 2, pp. 688–697, June 2014.

[14] 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 *Proceedings of ISMIR 2010*, 2010, pp. 255–266.

[15] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: a review," *ACM Transaction on intelligent systems and technology*, vol. 3, pp. 40:1–40:30, 2012.

[16] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: implications for content and context-based models," in *Proceedings of CMMR 2012 (Lecture Notes in Computer Science 7900)*, 2012, pp. 228–252.

[17] A. Gabrielsson, "The influence of musical structure on emotional expression," *Music and emotion: Theory and research*, pp. 223–248, 2001.

[18] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Proceedings of ICMLA 2008*, 2008, pp. 688–693.

[19] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *Proceedings of ICASSP 2014*. IEEE, 2014, pp. 5449–5453.

[20] W. Yang, K. Cai, B. Wu, Y. Wang, X. Chen, D. Yang, and A. Horner, "Beatsens' solution for MediaEval 2014 emotion in music task," in *Proceedings of MediaEval 2014 Workshop*, 2014.

[21] A. Wieczorkowska, "Towards extracting emotions from music," in *Proceedings of the International Workshop on Intelligent Media Technology for Communicative Intelligence*, 2004, pp. 228–238.

[22] J. Hartung, G. Knapp, and B. K. Sinha, "Statistical meta-analysis with applications," *Wiley Series in Probability and Statistics*, 2008.

[23] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, pp. 1–12, May 2013.

[24] M. Soleymani, M. N. Caro, and E. M. Schmidt, "1000 songs for emotional analysis of music," in *Proceedings of CroudMM 2013*. ACM, 2013.

[25] A. R. Conn, K. Scheinberg, and P. L. Toint, "On the convergence of derivative-free methods for unconstrained optimization," *Approximation theory and optimization*, pp. 83–108, 1997.