Modeling Expressed Emotions in Music using Pairwise Comparisons

Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen, and Jan Larsen *

Technical University of Denmark,
Department of Informatics and Mathematical Modeling,
Richard Petersens Plads B321, 2800 Lyngby, Denmark
{jenma; jenb; bjje; jl}@imm.dtu.dk

Abstract. We introduce a two-alternative forced-choice experimental paradigm to quantify expressed emotions in music using the two well-known arousal and valence (AV) dimensions. In order to produce AV scores from the pairwise comparisons and to visualize the locations of excerpts in the AV space, we introduce a flexible Gaussian process (GP) framework which learns from the pairwise comparisons directly. A novel dataset is used to evaluate the proposed framework and learning curves show that the proposed framework needs relative few comparisons in order to achieve satisfactory performance. This is further supported by visualizing the learned locations of excerpts in the AV space. Finally, by examining the predictive performance of the user-specific models we show the importance of modeling subjects individually due to significant subjective differences.

Keywords: expressed emotion, pairwise comparison, Gaussian process

1 Introduction

In recent years Music Emotion Recognition has gathered increasing attention within the Music Information Retrieval (MIR) community and is motivated by the possibility to recommend music that expresses a certain mood or emotion.

The design approach to automatically predict the expressed emotion in music has been to describe music by structural information such as audio features and/or lyrical features. Different models of emotion, e.g., categorical [1] or dimensional [2], have been chosen and depending on these, various approaches have been taken to gather emotional ground truth data [3]. When using dimensional models such as the well established *arousal* and *valence* (AV) model [2] the majority of approaches has been to use different variations of self-report direct scaling listening experiments [4].

* This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886, and in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

Direct-scaling methods are fast ways of obtaining a large amount of data. However, the inherent subjective nature of both induced and expressed emotion, often makes anchors difficult to define and the use of them inappropriate due to risks of unexpected communication biases. These biases occur because users become uncertain about the meaning of scales, anchors or labels [5]. On the other hand, lack of anchors and reference points makes direct-scaling experiments susceptible to drift and inconsistent ratings. These effects are almost impossible to get rid of, but are rarely modeled directly. Instead, the issue is typically addressed through outlier removal or simply by averaging across users [6], thus neglecting individual user interpretation and user behavior in the assessment of expressed emotion in music.

Pairwise experiments eliminates the need for an absolute reference anchor, due to the embedded relative nature of pairwise comparisons which persists the relation to previous comparisons. However, pairwise experiments scale badly with the number of musical excerpts which they accommodate in [7] by a tournament based approach that limits the number of comparisons and transforms the pairwise judgments into possible rankings. Subsequently, they use the transformed rankings to model emotions.

In this paper, we present a novel dataset obtained by conducting a controlled pairwise experiment measuring expressed emotion in music on the dimensions of valence and arousal. In contrast to previous work, we learn from pairwise comparisons, directly, in a principled probabilistic manner using a flexible Gaussian process model which implies a latent but interpretable valence and arousal function. Using this latent function we visualize excerpts in a 2D valance and arousal space which is directly available from the principled modeling framework. Furthermore the framework accounts for inconsistent pairwise judgments by participants and their individual differences when quantifying the expressed emotion in music. We show that the framework needs relatively few comparisons in order to predict comparisons satisfactory, which is shown using computed learning curves. The learning curves show the misclassification error as a function of the number of (randomly chosen) pairwise comparisons.

2 Experiment

A listening experiment was conducted to obtain pairwise comparisons of expressed emotion in music using a two-alternative forced-choice paradigm. 20 different 15 second excerpts were chosen from the USPOP2002[1] dataset. The 20 excerpts were chosen such that a linear regression model developed in previous work [8] maps exactly 5 excerpts into each quadrant of the two dimensional AV space. A subjective evaluation was performed to verify that the emotional expression throughout each excerpt was considered constant.

A sound booth provided neutral surroundings for the experiment and the excerpts were played back using headphones to the 8 participants (2 female,

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

6 male). Written and verbal instructions were given prior to each session to ensure that subjects understood the purpose of the experiment and were familiar with the two emotional dimensions of valence and arousal. Each participant compared all 190 possible unique combinations. For the arousal dimension, participants were asked the question *Which sound clip was the most excited, active, awake?*. For the valence dimension the question was *Which sound clip was the most positive, glad, happy?*. The two dimensions were evaluated individually in random order. The details of the experiment are available in [9].

3 Pairwise-Observation based Regression

We aim to construct a model for the dataset given the audio excerpts in the set $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ with $n = 20$ distinct excerpts, each described by an input vector $\mathbf{x}_i$ of audio features extracted from the excerpt. For each test subject the dataset comprises of all $m = 190$ combinations of pairwise comparisons between any two distinct excerpts, u and v, where $\mathbf{x}_u \in \mathcal{X}$ and $\mathbf{x}_v \in \mathcal{X}$. Formally, we denote the output set (for each subject) as $\mathcal{Y} = \{(d_k; u_k, v_k) | k = 1, ..., m\}$, where $d_k \in \{-1, 1\}$ indicates which of the two excerpts that had the highest valence or arousal. $d_k = -1$ means that the u_k'th excerpt is picked over the v_k'th and visa versa when $d_k = 1$.

We model the pairwise choice, d_k, between two distinct excerpts, u and v, as a function of the difference between two functional values, $f(\mathbf{x}_u)$ and $f(\mathbf{x}_v)$. The function $f : \mathcal{X} \to \mathbb{R}$ thereby defines an internal, but latent absolute reference of either valence or arousal as a function of the excerpt represented by the audio features.

Given a function, $f(\cdot)$, we can define the likelihood of observing the choice d_k directly as the conditional distribution.

$$p\left(d_k | \mathbf{f}_k\right) = \Phi\left(d_k \frac{f\left(\mathbf{x}_{v_k}\right) - f\left(\mathbf{x}_{u_k}\right)}{\sqrt{2}}\right), \tag{1}$$

where $\Phi(x)$ is the cumulative Gaussian (with zero mean and unity variance) and $\mathbf{f}_k = \left[f\left(\mathbf{x}_{u_k}\right), f\left(\mathbf{x}_{v_k}\right)\right]^\top$. This classical choice model can be dated back to Thurstone and his fundamental definition of *The Law of Comparative Judgment* [10].

We consider the likelihood in a Bayesian setting such that $p\left(\mathbf{f} | \mathcal{Y}, \mathcal{X}\right) = p\left(\mathcal{Y} | \mathbf{f}\right) p(\mathbf{f} | \mathcal{X}) / p\left(\mathcal{Y} | \mathcal{X}\right)$ where we assume that the likelihood factorizes, i.e., $p\left(\mathcal{Y} | \mathbf{f}\right) = \prod_{k=1}^{m} p\left(d_k | \mathbf{f}_k\right)$.

In this work we consider a specific prior, namely a Gaussian Process (GP), first considered with the pairwise likelihood in [11]. A GP is typically defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [12]. By $f\left(\mathbf{x}\right) \sim \mathcal{GP}\left(\mathbf{0}, k(\mathbf{x}, \mathbf{x}')\right)$ we denote that the function $f(\mathbf{x})$ is modeled by a zero-mean GP with covariance function $k(\mathbf{x}, \mathbf{x}')$. The fundamental consequence of this formulation is that the GP can be considered a distribution over functions, defined as $p\left(\mathbf{f} | \mathcal{X}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$ for any finite set of of function values $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$, where $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Bayes relation leads directly to the posterior distribution over $\mathbf{f}$, which is not analytical tractable. Instead, we use the *Laplace Approximation* to approximate the posterior with a multivariate Gaussian distribution[1].

To predict the pairwise choice d_t on an unseen comparison between excerpts r and s, where $\mathbf{x}_r, \mathbf{x}_s \in \mathcal{X}$, we first consider the predictive distribution of $f(\mathbf{x}_r)$ and $f(\mathbf{x}_s)$. Given the GP, we can write the joint distribution between $\mathbf{f} \sim p(\mathbf{f}|\mathcal{Y}, \mathcal{X})$ and the test variables $\mathbf{f}_t = [f(\mathbf{x}_r), f(\mathbf{x}_s)]^T$ as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_t \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_t \\ \mathbf{k}_t^T & \mathbf{K}_t \end{bmatrix} \right), \tag{2}$$

where $\mathbf{k}_t$ is a matrix with elements $[\mathbf{k}_t]_{i,2} = k(\mathbf{x}_i, \mathbf{x}_s)$ and $[\mathbf{k}_t]_{i,1} = k(\mathbf{x}_i, \mathbf{x}_r)$ with $\mathbf{x}_i$ being a training input.

The conditional $p(\mathbf{f}_t|\mathbf{f})$ is directly available from Eq. (2) as a Gaussian too. The predictive distribution is given as $p(\mathbf{f}_t|\mathcal{Y}, \mathcal{X}) = \int p(\mathbf{f}_t|\mathbf{f})\, p(\mathbf{f}|\mathcal{Y}, \mathcal{X})\, d\mathbf{f}$, and with the posterior approximated with the Gaussian from the Laplace approximation then $p(\mathbf{f}_t|\mathcal{Y}, \mathcal{X})$ will also be Gaussian given by $\mathcal{N}(\mathbf{f}_t|\boldsymbol{\mu}^*, \mathbf{K}^*)$ with $\boldsymbol{\mu}^* = \mathbf{k}_t^T \mathbf{K}^{-1} \hat{\mathbf{f}}$ and $\mathbf{K}^* = \mathbf{K}_t - \mathbf{k}_t^T (\mathbf{I} + \mathbf{W}\mathbf{K})\, \mathbf{k}_t$, where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from the Laplace approximation (see [13]). In this paper we are only interested in the binary choice d_t, which is determined by which of $f(\mathbf{x}_r)$ or $f(\mathbf{x}_s)$ that dominates[2].

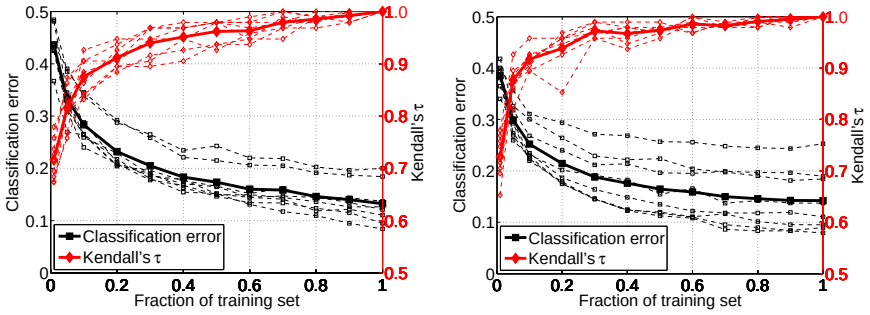
The zero-mean GP is fully defined by the covariance function, $k(\mathbf{x}, \mathbf{x}')$. In the emotion dataset each input instance is an excerpt described by the vector $\mathbf{x}$ containing the audio features for each time frame which is naturally modeled with a probability density, $p(\mathbf{x})$. We apply the probability product (PP) kernel [14] in order to support these types of distributional inputs. The PP kernel is defined directly as an inner product as $k(\mathbf{x}, \mathbf{x}') = \int [p(\mathbf{x})\, p(\mathbf{x}')]^q d\mathbf{x}$. We fix $q = 1/2$, leading to the Hellinger divergence [14]. In order to model the audio feature distribution for each excerpt, we resort to a (finite) Gaussian Mixture Model (GMM). Hence, $p(\mathbf{x})$ is given by $p(\mathbf{x}) = \sum_{z=1}^{N_z} p(z)\, p(\mathbf{x}|z)$, where $p(\mathbf{x}|z) = \mathcal{N}(\mathbf{x}|\mu_z, \sigma_z)$ is a standard Gaussian distribution. The kernel is expressed in closed form [14] as $k(p(\mathbf{x}), p(\mathbf{x}')) = \sum_z \sum_{z'} (p(z)\, p(z'))^q \tilde{k}(p(\mathbf{x}|\theta_z), p(\mathbf{x}'|\theta_{z'}))$ where $\tilde{k}(p(\mathbf{x}|\theta_z), p(\mathbf{x}'|\theta_{z'}))$ is the probability product kernel between two single components - also available in closed form [14].

4 Modeling Expressed Emotion

In this section we evaluate the ability of the proposed framework to capture the underlying structure of expressed emotions based on pairwise comparisons, directly. We apply the GP model using the probability product (PP) kernel described in Section 3 with the inputs based on a set of audio features extracted

[1] More details can be found in e.g. [13].

[2] With the pairwise GP model the predictive distribution of d_t can also be computed analytically (see [13]) and used to express the uncertainty in the prediction relevant for e.g. sequential designs, reject regions etc.



(a) Valence (b) Arousal

Fig. 1. Classification error learning curves and Kendall's τ for 10-fold CV on comparisons. Bold lines are mean curves across subjects and dash lines are curves for individual subjects. Notice, that for the classification error learning curves, the baseline performance corresponds to an error of 0.5, obtained by simply randomly guessing the pairwise outcome.

from the 20 excerpts. By investigating various combinations of features we obtained the best performance using two sets of commonly used audio features. The first set is the Mel-frequency cepstral coefficients (MFCC), which describe the short-term power spectrum of the signal. Secondly, we included spectral contrast features and features describing the spectrum of the Hanning windowed audio. Based on an initial evaluation, we fix the number of components in the GMM used in the PP Kernel to $N_z = 3$ components and train the individual GMMs by a standard EM algorithm with K-means initialization. Alternatively, measures such as the Bayesian Information Criterion (BIC) could be used to objectively set the model complexity for each excerpt.

4.1 Results: Learning Curves

Learning curves for the individual subjects are computed using 10-fold cross validation (CV) in which a fraction (90%) of the total number of pairwise comparisons constitutes the complete training set. Each point on the learning curve is an average over 10 randomly chosen and equally-sized subsets from the complete training set. The Kendall's τ rank correlation coefficient is computed in order to relate our results to that of e.g. [7] and other typical ranking based applications. The Kendall's τ is a measure of correlation between rankings and is defined as $\tau = (N_s - N_d)/N_t$ where N_s is the number of correctly ranked pairs, N_d is the number of incorrectly ranked pairs and N_t is the total number of pairs. The reported Kendall's τ is in all cases calculated with respect to the predicted ranks using all the excerpts.

Figure 1 displays the computed learning curves. With the entire training set included the mean classification errors across subjects for valence and arousal are 0.13 and 0.14, respectively. On average this corresponds to a misclassified comparison in every 7.5 and 7'th comparison for valence and arousal, respectively.

For valence, the mean classification error across users is below 0.2 with 40% of the training data included, whereas only 30% of the training data is needed to obtain similar performance for arousal. This indicates that the model for arousal can be learned slightly faster than valence. Using 30% of the training data the Kendall's τ is 0.94 and 0.97, respectively, indicating a good ranking performance using only a fraction of the training data.

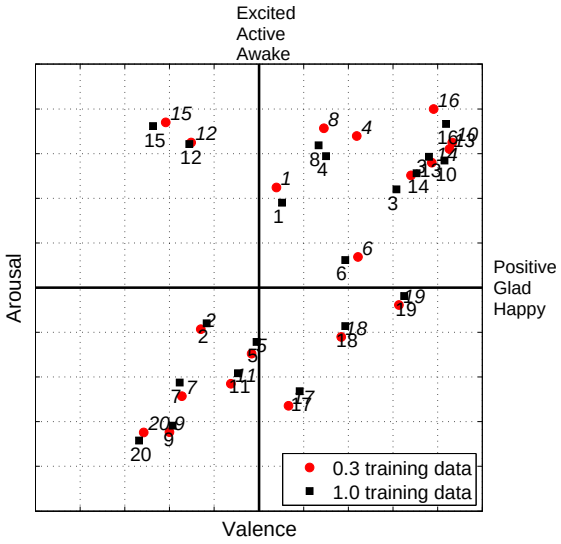
When considering the learning curves for individual users we notice significant individual differences between users—especially for arousal. Using the entire training set in the arousal experiment, the user for which the model performs best results in an error of 0.08 whereas the worst results in an error of 0.25. In the valence experiment the best and worst performances result in classification errors of 0.08 and 0.2, respectively.

4.2 Results: AV space

The learning curves show the pure predictive power of the model on unseen comparisons, but may be difficult to interpret in terms of the typical AV space. To address this we show that the latent regression function $f(\cdot)$ provides an internal but unit free representation of the AV scores. The only step required is a normalization which ensures that the latent values are comparable across folds and subjects. In Figure 2 the predicted AV scores are shown when the entire training set is included and when only 30% is included. The latter corresponds to 51 comparisons in total or an average of 2.5 comparisons per excerpt. The results are summarized by averaging across the predicted values for each user. 15 of the 20 excerpts are positioned in the typical high-valence high-arousal and low-valence low-arousal quadrants, 2 excerpts are clearly in the low-valance high-arousal quadrant and 3 excerpts are in the high-valance low-arousal quadrant of the AV space. The minor difference in predictive performance between 30% and the entire training dataset does not lead to any significant change in AV scores, which is in line with the reported Kendall's τ measure.

4.3 Discussion

The results clearly indicate that it is possible to model expressed emotions in music by directly modeling pairwise comparisons in the proposed Gaussian process framework using subject specific models. An interesting point is the large difference in predictive performance between subjects given the specific models. These differences can be attributed to the specific model choice (including kernel) or simply to subject inconsistency in the pairwise decisions. The less impressive predictive performance for certain subjects is presumably a combination of the two effects, although given the very flexible nature of the Gaussian process model, we mainly attribute the effect to subjects being inconsistent due to for example mental drift. Hence, individual user behavior, consistency and discriminative ability are important aspects of modeling expressed emotion in music and other cognitive experiments, and thus also a critical part when aggregating subjects in large datasets.



No.	Song name
1	311 - T and p combo
2	A-Ha - Living a boys adventure
3	Abba - Thats me
4	Acdc - What do you do for money honey
5	Aaliyah - The one i gave my heart to
6	Aerosmith - Mother popcorn
7	Alanis Morissette - These r the thoughts
8	Alice Cooper - Im your gun
9	Alice in Chains - Killer is me
10	Aretha Franklin - A change
11	Moby - Everloving
12	Rammstein - Feuer frei
13	Santana - Maria caracoles
14	Stevie Wonder - Another star
15	Tool - Hooker with a pen..
16	Toto - We made it
17	Tricky - Your name
18	U2 - Babyface
19	Ub40 - Version girl
20	Zz top - Hot blue and righteous

Fig. 2. AV values computed by averaging the latent function across folds and repetitions and normalizing for each individual model for each participant. Red circles: 30% of training set is used. Black squares: entire training set is used.

The flexibility and interpolation abilities of Gaussian Processes allow the number of comparisons to be significantly lower than the otherwise quadratic scaling of unique comparisons. This aspect and the overall performance should of course be examined further by considering a large scale dataset and the use of several model variations. In addition, the learning rates can be improved by combining the pairwise approach with active learning or sequential design methods, which in turn select only pairwise comparisons that maximize some information criterion.

We plan to investigate how to apply multi-task (MT) or transfer learning to the special case of pairwise comparisons, such that we learn one unifying model taking subjects differences into account instead of multiple independent subject-specific models. A very appealing method is to include MT learning in the kernel of the GP [15], but this might not be directly applicable in the pairwise case.

5 Conclusion

We introduced a two-alternative forced-choice experimental paradigm for quantifying expressed emotions in music in the typical arousal and valance (AV) dimensions. We proposed a flexible probabilistic Gaussian process framework to model the latent AV scales directly from the pairwise comparisons. The framework was evaluated on a novel dataset and resulted in promising error rates for both arousal and valence using as little as 30% of the training set corresponding to 2.5 comparisons per excerpt. We visualized AV scores in the well-known two dimensional AV space by exploiting the latent function in the Gaussian process

model, showing the application of the model in a standard scenario. Finally we especially draw attention to the importance of maintaining individual models for subjects due to the apparent inconsistency of certain subjects and general subject differences.

References

1. K. Hevner, "Experimental studies of the elements of expression in music," *American journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.
2. J.A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.
3. Y.E. Kim, E.M. Schmidt, Raymond Migneco, B.G. Morton, Patrick Richardson, Jeffrey Scott, J.A. Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proc. of the 11th Intl. Society for Music Information Retrieval (ISMIR) Conf*, 2010, pp. 255–266.
4. E. Schubert, *Measurement and time series analysis of emotion in music*, Ph.D. thesis, University of New South Wales, 1999.
5. M. Zentner and T. Eerola, *Handbook of Music and Emotion - Theory, Research, Application*, chapter 8 - Self-report measures and models, Oxford University Press, 2010.
6. A. Huq, J. P. Bello, and R. Rowe, "Automated Music Emotion Recognition: A Systematic Evaluation," *Journal of New Music Research*, vol. 39, no. 3, pp. 227–244, Sept. 2010.
7. Y.-H. Yang and H.H. Chen, "Ranking-Based Emotion Recognition for Music Organization and Retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, May 2011.
8. J. Madsen, *Modeling of Emotions expressed in Music using Audio features*, DTU Informatics, Master Thesis, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6036, 2011.
9. J. Madsen, *Experimental Protocol for Modelling Expressed Emotion in Music*, DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6246, 2012.
10. L. L. Thurstone, "A law of comparative judgement.," *Psychological Review*, vol. 34, 1927.
11. W. Chu and Z. Ghahramani, "Preference learning with Gaussian Processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.
12. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
13. B.S. Jensen and J.B. Nielsen, *Pairwise Judgements and Absolute Ratings with Gaussian Process Priors*, Technical Report, DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6151, September 2011.
14. T. Jebara and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.
15. E.V. Bonilla, F.V. Agakov, and C.K.I. Williams, "Kernel multi-task learning using task-specific features," *Proceedings of the 11th AISTATS*, 2007.