Predictive Modeling of Expressed Emotions in Music Using Pairwise Comparisons

Jens Madsen, Bjørn Sand Jensen, and Jan Larsen*

Department of Applied Mathematics and Computer Science,
Technical University of Denmark,
Matematiktorvet Building 303B, 2800 Kongens Lyngby, Denmark
{jenma,bjje,janla}@dtu.dk

Abstract. We introduce a two-alternative forced-choice (2AFC) experimental paradigm to quantify expressed emotions in music using the arousal and valence (AV) dimensions. A wide range of well-known audio features are investigated for predicting the expressed emotions in music using learning curves and essential baselines. We furthermore investigate the scalability issues of using 2AFC in quantifying emotions expressed in music on large-scale music databases. The possibility of dividing the annotation task between multiple individuals, while pooling individuals' comparisons is investigated by looking at the subjective differences of ranking emotion in the AV space. We find this to be problematic due to the large variation in subjects' rankings of excerpts. Finally, solving scalability issues by reducing the number of pairwise comparisons is analyzed. We compare two active learning schemes to selecting comparisons at random by using learning curves. We show that a suitable predictive model of expressed valence in music can be achieved from only 15% of the total number of comparisons when using the Expected Value of Information (EVOI) active learning scheme. For the arousal dimension we require 9% of the total number of comparisons.

Keywords: expressed emotion, pairwise comparison, Gaussian process, active learning.

1 Introduction

With the ever growing availability of music through streaming services, and with access to large music collections becoming the norm, the ability to easy-to-navigate-and-explore music databases has become increasingly pertinent. This problem has created the need to use alternative methods to organize and retrieve musical tracks, one being cognitive aspects such as emotions. The reasoning behind using emotions dates back to Darwin, who argued that music was a predecessor to speech in communicating emotions or intents [6]. This alternative seems appealing and a natural way of thinking about music, since most people can relate to happy or sad music, for example. The aspects about music that

* This publication only reflects the authors' views.

M. Aramaki et al. (Eds.): CMMR 2012, LNCS 7900, pp. 253–277, 2013.

express or induce emotions have been studied extensively by music psychologists [13]. The Music Information Retrieval (MIR) community has been building on their work with the aim to create automatic systems for recognition of emotions and organization of music based on emotion. The approach by music psychologists have been to exhaustively make experiments with human subjects/users to quantify emotions and analyze this data. To annotate the massive collections of music using a fully manual approach is not feasible and has resulted in the increased attention on automatic Music Emotion Recognition (MER).

The approach to automatically predict the expressed emotion in music has typically relied on describing music by structural information such as audio features and/or lyrics features. Controlled experiments have been conducted to obtain data describing the emotions expressed or induced in music. Machine learning methods have subsequently been applied to create predictive models of emotion, from the structural information describing music, predicting the emotional descriptors [1]. The reasoning behind using the emotions expressed in music and not induced (which describes how the subject feels as a result of the musical stimuli) has mainly been due to the availability of data. The mechanisms that are involved in the induction of emotions by music [12] are daunting. To potentially model this highly subjective aspect, a great deal of additional data about the user and context should be available in order to recognize the user's general state of mind. We see that to solve the MER, three main topics should be investigated: namely how to represent the audio using feature extraction; the machine learning methods to predict annotations, evaluations, rankings, ratings, etc.; and the method of quantifying and representing the emotions expressed in music. In the present work we want to look more closely into the aspect of quantifying the emotions expressed in music using an alternative experimental paradigm to gather more accurate ground truth data.

Music psychologists have offered different models to represent emotions in music, e.g., categorical [8] or dimensional [25], and depending on these, various approaches have been taken to gather emotional ground truth data [14]. When using dimensional models such as the well established *arousal* and *valence* (AV) model [25] the majority of approaches are based on different variations of self-report listening experiments using direct scaling [26].

Direct-scaling methods are fast ways of obtaining a large amount of data. However, they are susceptible to drift, inconsistency and potential saturation of the scales. Some of these issues could potentially be remedied by introducing anchors or reference points; hence, implicitly using relative rating aspects. However, anchors are problematic due to the inherent subjective nature of the quantification of emotion expressed in music, which makes them difficult to define, and the use of them will be inappropriate due to risks of unexpected communication biases [31]. Relative experiments, such as pairwise comparisons, eliminate 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. This was accommodated in [30] by a tournament-based approach that limits the

number of comparisons. Furthermore they introduce chaining, that is, inserting additional comparisons based on subjects' judgments and disregarding potential noise on the subjects' decisions. Multiple participants' judgments are pooled to form a large data set that is transformed into rankings which are then used to model emotions expressed in music.

However, the connection between the artist expressing emotions through music and how each individual experiences it will inherently vary. This experience is to be captured using a model of emotions using an experiment. The setup of this experiment alone gives rise to subjective differences such as interpretation and understanding of the experimental instruction, understanding and use of the scales, and projection of the emotional experience into the cognitive AV representation. Besides this, a multitude of aspects and biases can effect the judgments by participants [31]. Most of these effects are almost impossible to eliminate, but are rarely modeled directly. The issue is typically addressed through outlier removal or simply by averaging ratings for each excerpt across users [11], thus neglecting individual user interpretation and user behavior in the assessment of expressed emotion in music. For pairwise comparisons this approach is also very difficult. In previous work [20] we showed the potentially great subjective difference in the ranking of emotions, both in valence and arousal, which is due to the inherently different subjective judgments by participants.

The main objective in this work is to propose and evaluate a robust and scalable predictive model of valence and arousal, despite the adverse noise and inconsistencies committed by the participants. Our solution to this challenge is based on a two-alternative forced-choice (2AFC) approach, with the responses modeled in a Thurstonian framework with a principled noise model and a flexible non-parametric Bayesian modeling approach. This provides a supervised model, which has previously been applied in [20,21] for analyzing the ranking of excerpts in the AV space. In this work, we do not focus on the ranking, but the predictive properties of the approach, i.e., whether the model can predict the pairwise relations for new unseen excerpts.

Firstly, the predictive setting requires structural information describing the audio excerpt, so-called features (or covariates) from which new unseen comparisons can be predicted based on observed audio excerpts. Audio features and the representation of audio excerpts are still an open question in many audio modeling domains and particularly in emotion recognition. In this work we investigate the effect of various common audio features in a single mean/variance representation, given the proposed predictive approach.

Secondly, to model and understand the complex aspects of emotion requires extensive and costly experimentation. In the 2AFC paradigm the number of comparisons scales quadratically with the number of excerpts. This is not a favorable property of the current methodology. Given the best set of features (selected from the feature set investigation) we investigate two solutions to this problem: we consider the common approach of dividing the rating task between multiple individuals and/or pooling individuals' ratings [30]. Based on the rankings, we show that such an approach is not recommendable in the predictive

case, due to large subject variability. This is in line with previous work [20] on ranking. We furthermore propose and evaluate an alternative approach, namely sequential experimental design (or active learning) for reducing the number of comparisons required. In the Bayesian modeling approach deployed, this is an easy extension of the methodology. We show that faster learning rates can be obtained by applying a principled Bayesian optimal sequential design approach.

The investigation of the outlined aspects requires that all possible unique comparisons are made on both valence and arousal dimensions. Furthermore, to show variation across users, it is required to test on a reasonable number of subjects. Compared to previous work [20,21], the experimental part in this work is based on an extended data set using the 2AFC experimental paradigm quantifying the expressed emotion in music on the dimensions of valence and arousal. Finally, we discuss various extensions and open issues, outlining future research directions and possibilities.

Outline. In Sect. 2 the general methodology for examining the outlined aspects is introduced. This includes a relatively technical presentation of the modeling framework. The underlying experiment and data is described in Sect. 3, and Sect. 4 contains the experimental results including a description of the most important aspects. The results are discussed in Sect. 5, and finally Sect. 6 concludes the paper.

2 Methodology

Cognitive aspects, such as emotion, can be elicited in a number of ways which can be divided into self-report, observational indirect behavioral measures [29], psychophysiological [9] and functional neuroimaging [15]. Self-reporting approaches rely on human test subjects to actually be able to express the directed aspects, albeit using some experimental paradigm. This work focuses on self-report methods, thus asking direct questions to the user in order to elicit his or her understanding and representation of the cognitive aspect under investigation. This requires careful consideration regarding the experimental paradigm and subsequent analysis/modeling aspects.

When quantifying a cognitive aspect using either unipolar or bipolar scales, assuming that one can arrange the cognitive aspect in such a manner that we can ask the question if one element is more or less than the other. In this case we can use relative quantification methods to obtain a ranking of objects in that dimension. How the objects are arranged in the internal representation of the cognitive aspect is not being asked directly but acquired indirectly, i.e., indirect scaling. The question to the subject is not to place the object for evaluation on the scale, but cognitively a much simpler question, namely to compare objects. The argument is that simple questions about cognitive aspects provide a robust approach in obtaining information. The simplest of such indirectly scaling methods is the two-alternative forced-choice model (2AFC). Participants are simply asked which of the two objects presented has the most/highest (or least/lowest) of a given cognitive aspect, which is the approach we use in this work.

In the present setting, we look into the cognitive aspect of expressed emotion in music. To quantify this we use an experimental paradigm relying on the two-dimensional valence and arousal model, which consist of two bipolar dimensions, namely valence, ranging from happy to sad, and arousal ranging from excited to sleepy [25]. This dimensional approach naturally allows us to use the robust relative paradigm.

With this in mind, the general framework for the proposed 2AFC for eliciting and modeling general cognitive aspects is outlined in Fig. 1. Here we aim to elicit and model the users' cognitive representation of emotion, thus we present the user with a general set of instructions regarding the task and intent of the experiment. There are obvious elements of bias that can be introduced here and care has to be taken to ensure that the underlying idea of the experiment is understood to reduce bias.

The Thurstonian based paradigm in essence starts with **step A** in Fig. 1, where an experimental design mechanism will select two musical excerpts, indexed u and v, out of total of N. These two excerpts constitute a paired set for comparison indexed by k and denoted ε_k, out of K possible comparisons.

In **step B**, excerpts u_k and v_k are presented to the user through a user interface (UI), which provides instructions, asking the user to compare the two excerpts either on the valence or arousal dimension. Understanding and interpretation of the UI and the instructions given can vary between subjects and bias and variance can be introduced at this stage.

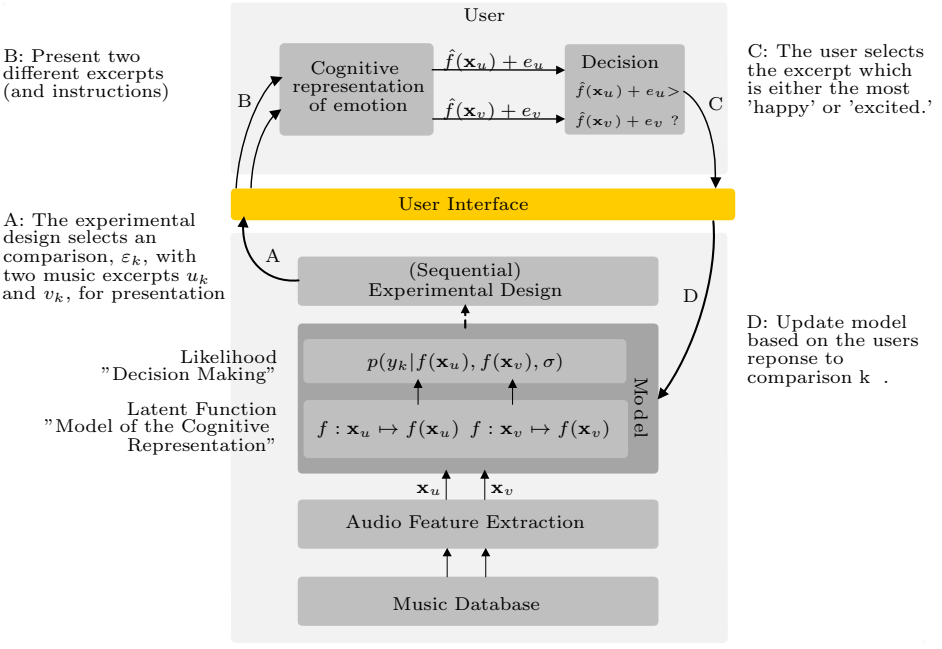


Fig. 1. Overview of the methodology from a system perspective

Table 1. Notation overview

System Element		Description	Notation
Music Database		Excerpt index	$u, v, r, s \in [1:N]$
		Number of excerpts	N
Audio Features		Audio feature representation	$\mathbf{x} \in \mathbb{R}^D$
		of excerpt (model input)	e.g. $\mathbf{x}_u$,$\mathbf{x}_v$
		A test input (to model)	$\mathbf{x}_*$
		A set of inputs (to model)	$\mathcal{X} = \{\mathbf{x}_i \vert i = 1..N\}$
User		Comparison with two inputs	$\hat{\varepsilon}_k = \{u_k, v_k\}$
		Response to a comparison	$y_k \in \{-1, +1\}$
		Number of comparisons	K
		Internal 'value' of an object in respect to a given cognitive aspect.	$\hat{f}(\mathbf{x})$
		Internal noise (independent of other inputs)	$e \sim \mathcal{N}(0, \sigma)$
		Internal basis for decision making	$\hat{f}(\mathbf{x}) + e$
Model (non-parametric)		Comparison	$\varepsilon_k = \{\mathbf{x}_{u_k}, \mathbf{x}_{v_k}\}$
		A set of K comparisons	$\mathcal{E} = \{\varepsilon_i \vert i = 1..K\}$
		A set of responses	$\mathcal{Y} = \{(y_k; \varepsilon_k) \vert k = 1..K\}$
		Hyperparameters in the model	$\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathcal{GP}}, \boldsymbol{\theta}_{\mathcal{L}}\}$
	Response	Likelihood . . .of observing a particular response given the function.	$p(y_k \vert f(\mathbf{x}_{u_k}), f(\mathbf{x}_{v_k}), \boldsymbol{\theta}_{\mathcal{L}}) = p(y_k \vert \mathbf{f}_k, \boldsymbol{\theta}_{\mathcal{L}})$
	Function	Function	$f : \mathbb{R}^D \to \mathbb{R}$ i.e. $\mathbf{x} \mapsto f(\mathbf{x})$
		Single value (a random variable)	$f(\mathbf{x})$
		Multiple values (L random variables)	$\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_L)]^\top$
		. . .for a particular comparison	$\mathbf{f}_k = [f(\mathbf{x}_{u_k}), f(\mathbf{x}_{v_k})]^\top$

In **step C** users convert their internal cognitive representation of the musical excerpts into a representation that can be used to compare the two based on the instructions given, which in our case comprise questions representing valence and arousal. Our assumption is that humans have an internal value $\hat{f}(\mathbf{x}_i) + e_i$ representing the valence or arousal value of a given excerpt $\mathbf{x}_i$ indexed by i. Given the great number of uncertainties involved in the self-report, we reasonably assume there is uncertainty on $\hat{f}(\mathbf{x})$ which is denoted $e \sim \mathcal{N}(0, \sigma)$. Prior to step C the user decides which of the two excerpts $\hat{f}(\mathbf{x}_u) + e_u$ and $\hat{f}(\mathbf{x}_v) + e_v$ is the largest given the cognitive dimension, and makes a decision which modelled by additive noise denoted $y_k \in \{-1, +1\}$, where the subject's selection is illustrated by step C in Figure 1.

In **step D** the analysis and modeling of the user's response takes place. With the aim of a predictive model, i.e., predicting the pairwise responses for unseen

music excerpts, this calls for a special modeling approach. The method applies a principled statistical modeling approach, relying on a choice model taking into account the noise, e, on the (assumed) internal representation. Secondly, the modeling approach places this choice model (likelihood function) in a Bayesian modeling framework, allowing for predictive capabilities. This results in a mathematical representation of the assumed internal representation of emotion, denoted $f(\mathbf{x})$, for a given excerpt. This representation like the internal, only makes sense when compared to the representation of other excerpts. The technical aspect of the modeling approach is described in the following sub-sections.

2.1 Likelihood

The decision process underlying 2AFC was considered in the seminal paper of Thurstone [27]. The main assumption is that the choice between two excerpts is based on the internal 'value' for each object which has a particular additive noise element. The decision is then based on the probability of the noisy internal 'value' of u or v being larger. If the additive noise is assumed to be distributed according to a Normal distribution, and independent from object to object, then the well-know probit choice model is obtained [28]. The probit choice model defines the likelihood of observing a particular response $y_k \in \{-1, +1\}$ as

$$p\left(y_k | f\left(\mathbf{x}_{u_k}\right), f\left(\mathbf{x}_{v_k}\right), \boldsymbol{\theta}_{\mathcal{L}}\right) = \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma}\right) \tag{1}$$

where $\Phi(\cdot)$ denotes the cumulative Normal distribution. The function values $f(\mathbf{x}_u)$ and $f(\mathbf{x}_v)$ are the model variables representing the assumed internal representation. However, the likelihood is seen to be dependent on the difference between the two (assumed) internal representations, in effect this means that the function itself has no absolute meaning and decisions are only based on differences. The noise variance on the (assumed) internal representation is denoted σ and provides a simple model of the internal noise process.

2.2 Latent Function

Given the response and likelihood function defined in Equ. (1), the remaining question relates to the latent function $f : \mathcal{X} \rightarrow \mathbb{R}$ defining the function values,$f(\mathbf{x})$, for each input, $\mathbf{x} \in \mathcal{X}$.

In this work we propose a non-parametric approach, in essence directly estimating values for individual $f(\mathbf{x})$'s, i.e., not through a parametric function (e.g. $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$). This is mainly motivated by the fact that the complexity of the underlying representation is virtually unknown, i.e., whether the problem is linear or non-linear is an open question which is best evaluated by allowing for very flexible function classes.

The non-parametric approach provides extreme flexibility, and we consider this in a Bayesian setting where we first assume that the likelihood factorizes, i.e., $p\left(\mathcal{Y}|\mathbf{f}\right) = \prod_{k=1}^{K} p\left(y_k|\mathbf{f}_k, \boldsymbol{\theta}_{\mathcal{L}}\right)$. This in effect means that, given the cognitive

representation, represented by $f(\cdot)$, we assume that there are no dependencies between the responses to the different comparisons. Thus, it is essential that the experimental procedure does not introduce a particular order of comparisons which may cause dependencies and systematic errors.

Given the factorized likelihood and placing a prior on the individual function values, $p(\mathbf{f}|\mathcal{X})$, the Bayesian approach directly provides the inference schema via Bayes relation. I.e. when keeping the hyperparameters, $\boldsymbol{\theta}$, constant, the posterior is directly given by

$$p(\mathbf{f}|\mathcal{X},\mathcal{Y},\boldsymbol{\theta}) = \frac{p(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_{\mathcal{GP}}) \prod_{k=1}^{K} p\left(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}}\right)}{p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta})} \tag{2}$$

The natural prior for the individual function values is a Gaussian Process (GP) [24]. This was first considered with the pairwise probit likelihood in [4]. A GP is defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [24]. The GP provides a mean for each individual $f(x)$, and correlates the functional values through a correlation function which implies some notion of smoothness; the only constraint on the function. With a zero-mean function, such a GP is denoted by $f\left(\mathbf{x}\right) \sim \mathcal{GP}\left(\mathbf{0}, \mathrm{k}\left(\mathbf{x},\mathbf{x}'\right)\right)$ with covariance function $\mathrm{k}(\mathbf{x},\mathbf{x}')$. The fundamental consequence is that the GP can be considered a distribution over functions, which is denoted as $p\left(\mathbf{f}|\mathcal{X}\right) = \mathcal{N}\left(\mathbf{0},\mathbf{K}\right)$ for any finite set of N function values $\mathbf{f} = [f(\mathbf{x}_1),...,f(\mathbf{x}_N)]^\top$, where $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$. This means that the correlation between a function value is defined by the input $\mathbf{x}$, for example audio features. The correlation function allows prediction by calculating the correlation between a new input and already observed inputs in terms of their audio features.

A common covariance function is the so-called squared exponential (SE) covariance function defined as $\mathrm{k}\left(\mathbf{x},\mathbf{x}'\right) = \sigma_f^2 \exp\left(-\left\|\mathbf{x}-\mathbf{x}'\right\|_2^2 \Big/ \sigma_l^2\right)$, where σ_f is a variance term and σ_l is the length scale, in effect, defining the scale of the correlation in the input space. This means that σ_ℓ defines how correlated two excerpts are in terms of their features. A special case arises when $\sigma_l \to 0$ which implies that the function values of two inputs are uncorrelated. In this case, knowing the functional of one input cannot be used to predict the function value of another due to the lack of correlation. On the other hand when $\sigma_l \to \infty$ the functional values are fully correlated i.e., the same.

For robustness, we provide a simple extension to the original model proposed in [4] by placing hyperpriors on the likelihood and covariance parameters, which act as simple regularization during model estimation. The posterior then yields $p(\mathbf{f}|\mathcal{X},\mathcal{Y},\boldsymbol{\theta}) \propto p\left(\boldsymbol{\theta}_{\mathcal{L}}|\cdot\right) p\left(\boldsymbol{\theta}_{\mathcal{GP}}|\cdot\right) p(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_{\mathcal{GP}}) p\left(\mathcal{Y}|\mathbf{f}\right)$, where $p(\boldsymbol{\theta}|\cdot)$ is a fixed prior distribution on the hyperparameters and a half student-t is selected in this work.

Inference. Given the particular likelihood, the posterior is not analytically tractable. We therefore resort to approximation and in particular the relatively simple Laplace approximation [24], which provides a multivariate Gaussian approximation to the posterior.

The hyperparameters in the likelihood and covariance functions are point estimates (i.e., not distributions) and are estimated by maximizing the model evidence defined as the denominator in Equ. 2. The evidence provides a principled approach to select the values of $\boldsymbol{\theta}$ which provides the model that (approximately) is better at explaining the observed data (see e.g. [2,24]). The maximization is performed using standard gradient methods.

Predictions. To predict the pairwise choice y_* 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 $\mathbf{f}_* = [f(\mathbf{x}_r), f(\mathbf{x}_s)]^T$ as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_* \\ \mathbf{k}_*^T & \mathbf{K}_* \end{bmatrix} \right), \tag{3}$$

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

The conditional $p(\mathbf{f}_*|\mathbf{f})$ is directly available from Equ. (3) as a Gaussian too. The predictive distribution is given as $p(\mathbf{f}_*|\mathcal{Y}, \mathcal{X}) = \int p(\mathbf{f}_*|\mathbf{f})\, p(\mathbf{f}|\mathcal{Y}, \mathcal{X})\, d\mathbf{f}$, and with the posterior approximated with the Gaussian from the Laplace approximation then $p(\mathbf{f}_*|\mathcal{Y}, \mathcal{X})$ will also be Gaussian given by $\mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}^*, \mathbf{K}^*)$ with $\boldsymbol{\mu}^* = \mathbf{k}_*^T \mathbf{K}^{-1} \hat{\mathbf{f}}$ and $\mathbf{K}^* = \mathbf{K}_* - \mathbf{k}_*^T (\mathbf{I} + \mathbf{W}\mathbf{K})^{-1} \mathbf{W}\ \mathbf{k}_*$, where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from the Laplace approximation (see [24]). In this paper, are often interested in the binary choice y_*, which is simply determined by which of $f(\mathbf{x}_r)$ or $f(\mathbf{x}_s)$ is the largest.

2.3 Sequential Experimental Design

The acquisition of pairwise observations can be a daunting and costly task if the database contains many excerpts due to the quadratic scaling of the number of possible comparisons. An obvious way to reduce the number of comparisons is only to conduct a fixed subset of the possible comparisons in line with classical experimental design. In this work we propose to obtain the most relevant experiments by sequential experimental design, also known as active learning in the machine learning community. In this case comparisons (each with two inputs) are selected in a sequential manner based on the information provided when conducting the particular comparison. The information considered here is based on the entropy of the predictive distribution or change in the entropy.

We consider the set of comparisons conducted so far, $\mathcal{E}_a$, which gives rise to a set of unique inputs $\mathcal{X}_a$ and a response set $\mathcal{Y}_a$ which are all denoted as active set(s). Secondly, we consider a set of candidate comparisons, $\mathcal{E}_c$, , which gives rise to a set of unique inputs $\mathcal{X}_c$ and an unknown response set $\mathcal{Y}_c$. The task is to select the next comparison $\varepsilon_* = \{\mathbf{x}_{u_*}, \mathbf{x}_{v_*}\}$ from $\mathcal{E}_c$. The following three cases is considered for solving this task:

Random: The next pairwise comparison is selected at random from the set of candidate comparisons.

VOI (Value of Information): Selection of the next comparison with the maximum entropy (i.e., uncertainty) of the predictive distribution of the model[1], $S\left(\mathbf{f}_*|\varepsilon_*, \mathcal{E}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right)$.
The next comparison is simply selected by $\arg\max_{\varepsilon_* \in \mathcal{E}_c} S\left(\mathbf{f}_*|\varepsilon_*, \mathcal{E}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right)$.
The predictive distribution is a bivariate normal distribution which has the entropy [5], $S\left(\mathbf{f}_*|\varepsilon_*, \mathcal{E}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) = \frac{1}{2}\log\left(\left(2\cdot\pi\cdot e\right)^D|\mathbf{K}^*|\right)$. Where $|\mathbf{K}^*|$ denotes the determinant of the (predictive) covariance matrix.

EVOI (Expected Value of Information): In the Bayesian framework it is possible to evaluate the expected entropy change of the posterior which was suggested in the work of Lindley [18]. Hence, the information of conducting a particular comparison is the change in entropy of the posterior i.e.,

$$\Delta S\left(\mathbf{f}\right) = S\left(\mathbf{f}|y_*, \varepsilon_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) - S\left(\mathbf{f}|\mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right)$$

The expectation in regards to y can be shown to yield [19]

$$\mathrm{EVOI}\left(\varepsilon_*\right) = \sum_{y\in\{-1,1\}} p\left(y_*|\varepsilon_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) \Delta S\left(\mathbf{f}|y_*, \varepsilon_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) \tag{4}$$

$$= \sum_{y\in\{-1,1\}} \int p\left(y_*|\mathbf{f}_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) p\left(\mathbf{f}_*|\varepsilon_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) \log p\left(y_*|\mathbf{f}_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) d\mathbf{f}_*$$

$$- \sum_{y\in\{-1,1\}} p\left(y_*|\varepsilon_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) \log p\left(y_*|\varepsilon_*, \mathcal{X}_a, \mathcal{Y}_a, \boldsymbol{\theta}\right) \tag{5}$$

Thus, the next comparison is chosen as $\arg\max_{\varepsilon_* \in \mathcal{E}_c} \mathrm{EVOI}\left(\varepsilon_*\right)$. The (inner) integral is analytical intractable and requires numerical methods. This is feasibly only due to the low dimensionality (which is effectively only one, since considering the difference distribution). An analytical approximation has been proposed for standard classification [10]; however, here we rely on numerical integration based on adaptive Gauss-Kronrod quadrature.

2.4 Evaluation

In order to evaluate the performance of the proposed modeling approach, we use a specific Cross Validation (CV) approach and baselines for verification and significance testing. When dealing with pairwise comparisons the way the cross validation is set up is a key issue.

[1] Alternatively we may consider the predictive uncertainty on the response, y_*. See e.g. [3] for a general discussion of various information criterion.

Cross Validation
In previous work [21] we evaluated the ability of the GP framework to rank excerpts on the dimensions of valence and arousal using learning curves. To obtain the learning curves, Leave-One-Out CV was used and in each fold a fraction of comparisons was left out. These comparisons are potentially connected and thus, to evaluate the ability of the model to predict an unseen excerpts rank, all comparisons with an excerpt must be left out in each fold. Thus in the present work we use a Leave-One-Excerpt-Out (LOEO) method. Learning curves are computed as a function of the fraction of all available comparisons, evaluating the question of how many pairwise comparisons are needed to obtain a competitive predictive model. Each point on the learning curves is computed as an average of 50 randomly chosen equally-sized subsets from the complete training set. The reasoning behind this is that testing all unique possible combinations of e.g. choosing 8 out of 15 excerpts is exhausting, so random repetitions are used to obtain robust learning curves.

Baselines
Three basic baselines are introduced that consider the distribution of the pairwise comparisons, namely a random baseline ($Base_{rnd}$) and two that only predict one class ($Base_{+1}$ and $Base_{-1}$), i.e., excerpt u always greater than excerpt v, or vice versa. This takes into account that the data set is not balanced between the two outcomes of $+1$ and -1. An additional baseline ($Base_{upper}$) is introduced. Given a model type, a baseline model of same type is trained on both training and test data and evaluated on the test data for that given CV fold. This provides an upper limit of how well it is possible for that given model and features can perform. Furthermore, a baseline model $Base_{low}$ is introduced that only uses information from the comparisons available in each CV fold (not the audio features). The model ranks excerpts using a tournament approach, counting the number of times a specific excerpt has been ranked greater than another. The number of wins is assigned to each excerpt's f value. All excerpts that have no f assignment are given the average f value of all available f values. To predict the test data in each CV fold, the assigned f values are used, and for f values that are equal a random choice is made with equal probability of either class. This naive baseline model serves as a lower limit, which all models have to perform better than.

Significance Testing
To ensure that each of the trained models perform better than $Base_{low}$ we use the McNemar paired test with the $Null$ hypothesis that two models are the same, if $p < 0.05$ then the models can be rejected as equal on a 5% significance level.

AV-Space Visualization
In the principled probabilistic GP framework the latent function $f(\cdot)$ is directly available to compare rankings between models. However for visualization to

compare the rankings we use a reference numerical space. The ranking of excerpts, given by $f(\cdot)$, is assigned the same functional value as the reference space, preserving the ranking of excerpts, but losing the relative distance given by $f(\cdot)$. This allows us to average rankings across users, folds and repetitions.

3 Experiment and Data

3.1 Experiment

A listening experiment was conducted to obtain pairwise comparisons of expressed emotion in music using the 2AFC experimental paradigm. A total of 20 different 15 second excerpts were chosen, in the middle of each track, from the USPOP2002[2] data set as shown in Table 2. The 20 excerpts were chosen such that a linear regression model developed in previous work [19] maps 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. This fact, and using short 15 second excerpts, should reduce any temporal change in the expressed emotion thus making post-ratings applicable. A sound booth provided neutral surroundings for the experiment to

Table 2. Excerpts used in experiment

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 - I'm 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

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

reduce any potential bias of induced emotions. The excerpts were played back using closed headphones to the 13 participants (3 female, 10 male) age 16-29, average 20.4 years old, recruited from a local high school and university. Participants had a musical training of 0-15 years, on average 2 years, and listened to 0-15 hours of music every day, on average 3.5 hours. 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. Furthermore instructions were given ensuring that participants focused on the expressed emotions of the musical excerpts. Each participant compared all 190 possible unique combinations. To reduce any systematic connection between comparisons, each comparison was chosen randomly. For the arousal dimension, participants were asked the question *Which sound clip was the most exciting, active, awake?*. For the valence dimension the question was *Which sound clip was the most positive, glad, happy?*. The reasoning behind these question lies in the communication of the dimensions of valence and arousal, pilot experiments showed a lack of understanding when fewer words were used. The two dimensions were evaluated independently and which of the two dimensions should be evaluated first was chosen randomly. The total time for the experiment was 4 hours, each session taking 1 hour in order to reduce any fatigue. After the experiments, participants rated their understanding of the experiment, the results can be seen in Table 3.

The understanding of the experiment and the scales was generally high, and it was noted that people rated the audio higher than the lyrics as a source of their judgments of the emotions expressed in music. The experiment had two atypical participants, one had low overall understanding of the experiment because he did not find the scales appropriate, and the other did understand the experiment, but did not understand the scales or found them inappropriate.

Table 3. Results of post-experiment questions to the 13 participants. All ratings were performed on a continuous scale, here normalized to 0-1. Results are presented as: minimum-maximum (average).

Question	Rating
General understanding	0.36-0.99 (0.70)
Understanding of scales	0.34-1.00 (0.84)
Appropriateness of scales	0.36-0.99 (0.78)
Lyrics, source of expressed emotion	0.00-0.74 (0.43)
Audio, source of expressed emotion	0.18-1.00 (0.69)

3.2 Audio Features

In order to represent the 15 second musical excerpts in later mathematical models, each excerpt is represented by audio features. These are extracted using four standard feature-extraction toolboxes, the MIR[17], CT[23], YAAFE[22], and

MA[3] toolboxes, and furthermore the Echonest API[4]. An overview is given in Table 4 of the features used from these toolboxes.

Due to the vast number of features used in MIR, the main standard features are grouped. In addition, the Echonest timbre and pitch features have been extracted, resulting in a total of 18 groups of features. The audio features have been extracted on different time scales, e.g., MFCCs result in 1292 samples for 15 seconds of audio data, whereas pitch produce 301 samples. Often the approach to integrate the feature time series over time is to assume that the distribution of feature samples is Gaussian and subsequently the mean and variance are used to represent the entire feature time series. In the present work, Gaussian distributions are fitted where appropriate and beta distributions are fitted where the distribution has a high skewness. The entire time series is represented by the mean and standard deviation of the fitted distributions.

4 Experimental Results

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 squared exponential (SE) kernel described in Sect. 2 with the inputs based on the groups of audio features described in Sect. 3.2 extracted from the 20 excerpts. The kernel was initialized with $\sigma_l = 1$ and $\sigma_f = 2$, furthermore the half student-t [7] hyperprior is initialized with $df = 4$ and $scale = 6$. We present three different investigations into the modeling of expressed emotions using the 2AFC paradigm. First a performance evaluation of the 18 groups of features is performed finding the best combination of features. These features are used in all subsequent results. Second, to investigate the scaling issues of 2AFC, the subjective variation in the model's predictive performance is investigated, along with a visualization of the subjective variation in rankings. Third, the question of how many pairwise comparisons are needed to obtain a predictive model of expressed emotions in music is investigated. This is evaluated using three different methods of selecting pairwise comparisons in an experimental setup, namely using the EVOI or VOI active learning methods or choosing comparisons randomly.

4.1 Performance of Features

The performance of the GP framework using the 18 different feature groupings is evaluated using LOEO learning curves. The predictive performance for the valence dimension is shown in Table 5. The single best performing feature, modeling the valence dimension is the Fluctuations feature resulting in a classification error of 0.2389 using the entire training set. For valence the Echonest pitch feature perform worse than Chroma and Pitch features from the CT toolbox although the timbre features perform slightly better than the MFCC features which are

[3] `http://www.pampalk.at/ma/`
[4] `http://the.echonest.com/`

Table 4. Acoustic features used for emotion prediction

Feature	Description	Dimension(s)
Mel-frequency cepstral coefficients (MFCCs)[1]	The discrete cosine transform of the log-transformed short-time power spectrum on the logarithmic mel-scale.	20
Envelope (En)	Statistics computed on the distribution of the extracted temporal envelope.	7
Chromagram CENS, CRP [23]	The short-time energy spectrum is computed and summed appropriately to form each pitch class. Furthermore statistical derivatives are computed to discard timbre-related information.	12 12 12
Sonogram (Sono)	Short-time spectrum filtered using an outer-ear model and scaled using the critical-band rate scale. An inner-ear model is applied to compute cochlea spectral masking.	23
Pulse clarity [16]	Ease of the perception by listeners of the underlying rhythmic or metrical pulsation in music.	7
Loudness [22]	Loudness is the energy in each critical band.	24
Spectral descriptors (sd) [22] (sd2) [17]	Short-time spectrum is described by statistical measures e.g., flux, roll-off, slope, variation, etc.	9 15
Mode, key, key strength [17]	Major vs. Minor, tonal centroid and tonal clarity.	10
Tempo [17]	The tempo is estimated by detecting periodicities on the onset detection curve.	2
Fluctuation Pattern [17]	Models the perceived fluctuation of amplitude-modulated tones.	15
Pitch [23]	Audio signal decomposed into 88 frequency bands with center frequencies corresponding to the pitches A0 to C8 using an elliptic multirate filterbank.	88
Roughness [17]	Roughness or dissonance, averaging the dissonance between all possible pairs of peaks in the spectrum.	2
Spectral Crest factor [22]	Spectral crest factor per log-spaced band of 1/4 octave.	23
Echonest *Timbre*	Proprietary features to describe timbre.	12
Echonest *Pitch* [17]	Proprietary chroma-like features.	12

Table 5. Valence: Classification error learning curves as an average of 50 repetitions and 13 individual user models, using both mean and standard deviation of the features. McNemar test between all points on the learning curve and $Base_{low}$ resulted in $p < 0.05$ for all models except results marked with *, with a sample size of 12.350.

Training size	5%	7%	10%	20%	40%	60%	80%	100%
MFCC	0.4904	0.4354	0.3726	0.3143	0.2856	0.2770	0.2719	0.2650
Envelope	**0.3733**	**0.3545**	0.3336	0.3104	0.2920	0.2842	0.2810	0.2755
Chroma	0.4114*	0.3966*	0.3740	0.3262	0.2862	0.2748	0.2695	0.2658
CENS	0.4353	0.4139	0.3881	0.3471	0.3065	0.2948	0.2901*	0.2824
CRP	0.4466	0.4310	0.4111	0.3656	0.3066	0.2925	0.2876	0.2826
Sonogram	0.4954	0.4360	0.3749	0.3163	0.2884	0.2787	0.2747	0.2704
Pulse clarity	0.4866	0.4357	0.3856	0.3336	0.3026	0.2930	0.2879	0.2810
Loudness	0.4898	0.4310	0.3684	0.3117	0.2854	0.2768	0.2712	0.2664
Spec. disc.	0.4443	0.4151	0.3753	0.3263	0.2939	0.2857	0.2827	0.2794
Spec. disc. 2	0.4516	0.4084	0.3668	0.3209	0.2916	0.2830	0.2781	0.2751
Key	0.5303	0.4752	0.4104	0.3370	0.2998	0.2918	0.2879	0.2830*
Tempo	0.4440	0.4244	0.3956	0.3559*	0.3158	0.2985	0.2933	0.2883
Fluctuations	0.4015	0.3584	**0.3141**	**0.2730**	**0.2507**	**0.2433**	**0.2386**	**0.2340**
Pitch	0.4022	0.3844	0.3602	0.3204	0.2926	0.2831	0.2786	0.2737
Roughness	0.4078	0.3974	0.3783	0.3313	0.2832	0.2695	0.2660	0.2605
Spec. crest	0.4829	0.4289	0.3764	0.3227	0.2994	0.2942	0.2933	0.2923
Echo. timbre	0.4859	0.4297	0.3692	0.3127	0.2859	0.2767	0.2732	0.2672
Echo. pitch	0.5244	0.4643	0.3991*	0.3275	0.2942	0.2841	0.2790	0.2743
$Base_{low}$	0.4096	0.3951	0.3987	0.3552	0.3184	0.2969	0.2893	0.2850

said to describe timbre. Including both mean and variance of the features showed different performance for the different features, therefore the best performing for valence and arousal was chosen resulting in both mean and variance for valence and only mean for arousal.

The learning curves showing the predictive performance on unseen comparisons on the arousal dimension are shown in Table 6. The single best performing feature, using the entire training set is Loudness resulting in an error rate of 0.1862. Here a picture of pitch and timbre related features seem to show a good level of performance.

Using a simple forward feature selection method. the best performing combination of features for valence are fluctuation pattern, spectral crest flatness per band, envelope statistics, roughness, CRP and Chroma resulting in an error of 0.1960 using the mean of the features. It should be noted that using only the 4 first produces an error of 0.1980. For arousal the best performing combination was Spectral descriptors, CRP, Chroma, Pitch, Roughness and Envelope statistics using mean and standard deviation of the features results in an error of 0.1688. All models trained for predicting valence and arousal are tested with McNemar's paired test against the $Base_{low}$, with the $Null$ hypothesis that two models are the same, all resulted in $p < 0.05$ rejecting the $Null$ hypothesis of being equal at a 5% significance level.

Table 6. Arousal: Classification error learning curves as an average of 50 repetitions and 13 individual user models, using only the mean of the features. McNemar test between all points on the learning curve and $Base_{low}$ resulted in $p < 0.05$ for all models except results marked with *, with a sample size of 12.350.

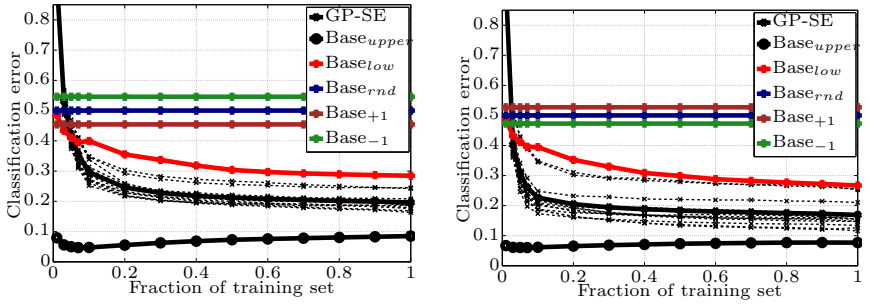
Training size	5%	7%	10%	20%	40%	60%	80%	100%
MFCC	0.3402	0.2860	0.2455	0.2243	0.2092	0.2030	0.1990	0.1949
Envelope	0.4110*	0.4032	0.3911	0.3745	0.3183	0.2847	0.2780	0.2761
Chroma	0.3598	0.3460	0.3227	0.2832	0.2510	0.2403	0.2360	0.2346
CENS	0.3942	0.3735	0.3422	0.2994	0.2760	0.2676	0.2640	0.2621
CRP	0.4475	0.4336	0.4115	0.3581	0.2997	0.2790	0.2735	0.2729
Sonogram	0.3325	0.2824	0.2476	0.2244	0.2118	0.2061	0.2033	0.2026
Pulse clarity	0.4620	0.4129	0.3698	0.3281	0.2964	0.2831	0.2767*	0.2725
Loudness	0.3261	0.2708	**0.2334**	**0.2118**	**0.1996**	**0.1944**	**0.1907**	**0.1862**
Spec. disc.	0.2909	0.2684	0.2476	0.2261	0.2033	0.1948	0.1931	0.1951
Spec. disc. 2	0.3566	0.3223	0.2928	0.2593	0.2313	0.2212	0.2172	0.2138
Key	0.5078	0.4557	0.4059	0.3450	0.3073*	0.2959	0.2926	0.2953
Tempo	0.4416	0.4286	0.4159	0.3804	0.3270	0.3043	0.2953	0.2955
Fluctuations	0.4750	0.4247	0.3688	0.3117	0.2835	0.2731	0.2672	0.2644*
Pitch	0.3173	0.2950	0.2668	0.2453	0.2301	0.2254	0.2230	0.2202
Roughness	**0.2541**	**0.2444**	0.2367	0.2304	0.2236	0.2190	0.2168	0.2170
Spectral crest	0.4645	0.4165	0.3717	0.3285	0.2979	0.2866*	0.2828	0.2838
Echo. timbre	0.3726	0.3203	0.2797	0.2524	0.2366	0.2292	0.2258	0.2219
Echo. pitch	0.3776	0.3264	0.2822	0.2492	0.2249	0.2151	0.2089	0.2059
$Base_{low}$	0.4122	0.3954	0.3956	0.3517	0.3087	0.2879	0.2768	0.2702

4.2 Subjective Variation

By letting multiple test participants rate the same musical excerpts and model these responses individually we can explore the subjective differences in greater detail.

Learning Curves

To evaluate the differences between subjects in how well the model predicts their pairwise comparisons, the LOEO learning curves for each individual are shown in Fig. 2. The $Base_{low}$ and $Base_{upper}$ described in Sect. 2.4 are shown, which indicate the window in which the proposed model is expected to perform. In Fig. 2(b) the individual learning curves are shown, computed by using the best performing combination of features as mentioned in Sect. 4.1. The difference in performance between the average of all individual models and the $Base_{upper}$ is 0.0919. Compared to the $Base_{low}$ we see a difference of 0.0982, showing a large improvement. The models trained in the data for participants 6 and 7 results in a classification error of 0.2553 and 0.2526 respectively, compared with the average of 0.1688 for the arousal dimension. Post-experiment ratings show that participant 6 rated a low rating of understanding and appropriateness of the scales of 0.3033 and 0.3172 respectively, although participant 7 rated a high understanding. In Fig. 2(a) the individual learning curves for the valence dimension are shown. Participants 1 and 5 have an error rate when using the whole training



(a) Valence (b) Arousal

Fig. 2. Individual classification error learning curves; Dashed black lines: individually trained models, Bold black crosses: average across individual models

set of 0.2421 and 0.2447 respectively compared to the average of 0.2257. Participant 5 rated in the post questionnaire a lack of understanding of the scales used in the experiment and furthermore did not find them appropriate. Participant 1 on the other hand did not rate any such lack of understanding. To investigate if there is an underlying linear connection between the models' classification error and the participants' post-questionnaire ratings, simple correlation analysis was made for all questions, a correlation of 0.13 for the appropriateness of the scales and the arousal was found and even less for the other questions, so no significant correlation was found. Comparing the average performance of the individual models and $Base_{upper}$, the difference in performance is 0.1109 using the whole training set. Furthermore comparing it to $Base_{low}$ the difference in performance is 0.0887, showing an improvement of using audio features compared to only using comparisons.

AV Space

The Gaussian Process framework can, given the features, predict the pairwise comparisons given by each participant on unseen excerpts. This on the other hand does not necessarily mean that participants' rankings of excerpts on the dimensions of valence and arousal are the same, which was investigated in previous work [20]. These variations in rankings of excerpts between subjects are visualized in the AV space on Fig. 3 using the method mentioned in Sect. 2.4. Excerpts 5, 2, 7, 9 and 20 in the low-valence low-arousal quadrant of the AV space show a relatively low variation in ranking, both in the dimension of valence and arousal, whereas the excerpts in the low-valence high-arousal quadrant, namely excerpts 12 and 15, have a high variation in both dimensions. It is evident that participants agree on the ranking of some excerpts and fundamentally disagree on some.

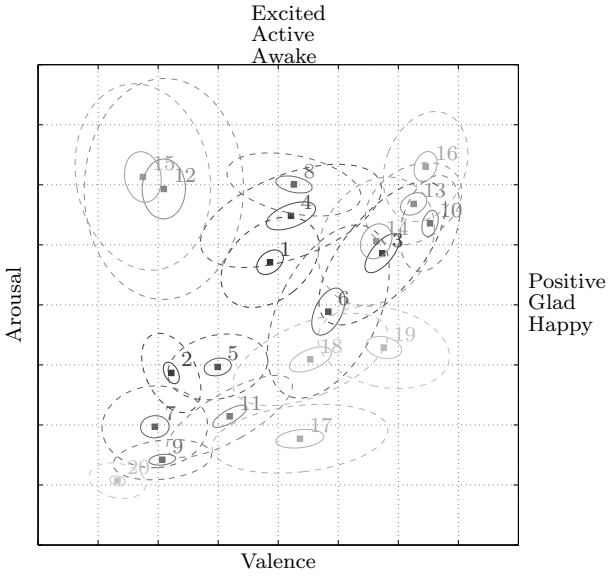


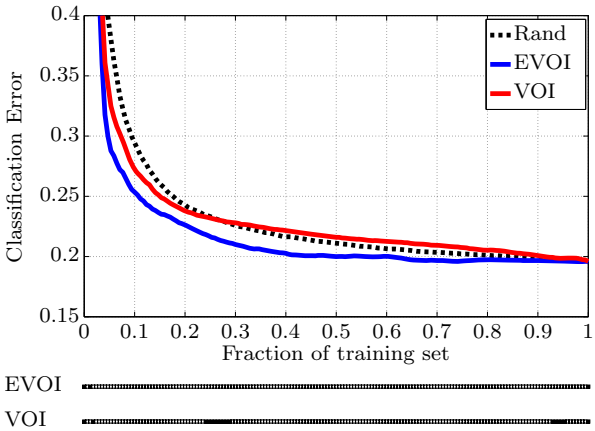
Fig. 3. Variation in ranking of excerpts in the valence and arousal space. Solid lines: 5% percentile, dashed line 50% percentile. Number refers to Table 2.

4.3 Reducing the Number of Required Comparisons

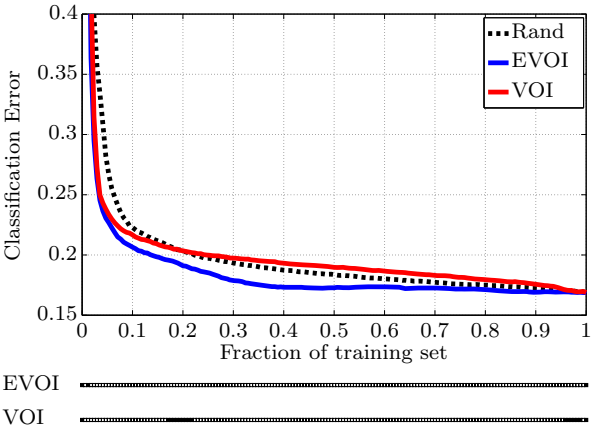
In this section we investigate how the model performs using only a fraction of available comparisons in predicting comparisons for unseen excerpts and to visualize the subsequent change in ranking of excerpts in the AV space.

Learning Curves

We investigate how many comparisons are needed to obtain a predictive model using LOEO learning curves. The traditional method of selecting a comparison in an experimental setup is simply to choose one at random from the comparisons defined by the experiment. This was the procedure in the listening experiment described in Sect. 3. But on the other hand this might not be the optimal way of choosing what comparisons should be judged by participants. Therefore we simulate if these comparisons can be chosen in alternative ways that can potentially improve the performance and decrease the number of comparisons needed to obtain a predictive model. As described in Sect. 2.3 we compare the procedure of using random selection of comparisons and the EVOI and VOI model. On Fig. 4 we see the three methods in detailed learning curves with a McNemar paired test between the model selecting comparisons at random and the EVOI and VOI models. The largest performance gains using the sequential design method EVOI are seen on the valence dimension using 4% of the training data, improving 0.105 and for arousal at 2.5% improving 0.106. Visually it is apparent that the EVOI model produces the largest improvement compared to selecting comparisons randomly. The difference after



(a) Valence



(b) Arousal

Fig. 4. Classification error learning curves comparing the EVOI, VOI and Rand models. The secondary graph below the learning curves shows filled squares when $p > 0.05$ and white when $p < 0.05$ using the McNemar's paired test. The test is performed between the the Rand model and the two EVOI and VOI.

10% of the training data is 0.041 decreasing to 0.015 at 20% with the same performance gain until 40% and gain in performance is obtained until all comparisons are judged for the valence dimension. On the arousal dimension the improvement after 4 comparisons is 0.104 and from 10% to 50% an improvement is achieved around 0.015 and 0.010. For arousal the VOI model improves the performance around 0.08 in the beginning of the learning curve at around 2-3%. Using 20% of the training set and above, selecting comparisons at random results in a better performance than selecting with the VOI model for arousal.

To evaluate the number of comparisons needed to obtain a predictive model we set a 95% performance threshold, using the entire training set. The EVOI model

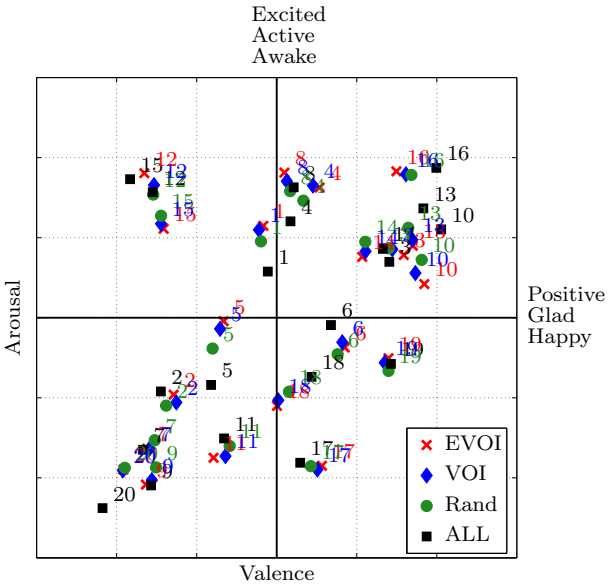


Fig. 5. AV space visualizing the change in ranking of the models trained on a fraction of available comparisons. EVOI model trained on 15.20% and 8.77% of the training set, VOI model trained on 21.64% and 14.04% and model selecting comparisons randomly (Rand) on 23.39% and 15.79% for valence and arousal respectively. Numbers refer to Table 2. Method of visualization in the AV space is described in Sect. 2.4.

achieves this performance corresponding to 0.2362 using only 15.2% of the training set, whereas the VOI model reaches this level using 21.64% and with random selection at 23.39% for the valence dimension. On the arousal dimension, the threshold performance corresponds to an error rate of 0.2104, choosing comparisons at random the model reaches this 95% performance level at 15.79% of the comparisons in the training set, the VOI model at 14.04% and the EVOI at 8.77%.

AV Space

Using a threshold we ensure that we reach a certain predictive performance, the consequence this has on the ranking of the excerpts in the AV space on the other hand could potentially be dramatic. Therefore we visualize the ranking of excerpts using the threshold discussed in the last section. The reference point to compare the change in rankings is the model trained on all comparisons for each subject individually. The rankings are visualized in the AV space on Fig. 5. Judging by the position of the excerpts in the AV space, the change in ranking is relative small, although on some excerpts the ranking does change, using the 95% performance threshold ensures that we have a good predictive performance and still reach the final ranking.

5 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. How to represent music using structural information is a key issue in MIR and the field of MER. In this work we use audio features and the optimal combination is found using learning curves and forward-feature selection. On the data set deployed, we find the gain of using audio features to predict pairwise comparisons on the dimensions of valence and arousal is 0.09 and 0.10, respectively. To make this comparison it is essential to have a proper baseline model which we introduce using the novel baseline $Base_{low}$. The baseline makes predictions solely by looking at the comparisons, and by disregarding any other information. The baseline performs similarly to a model with $\sigma_l \to 0$, resulting in no correlation between any excerpts as mentioned in Sect. 2.2. We can therefore ensure that we do capture some underlying structure represented in the music excerpts that describes aspects related to the expressed emotions in music.

Furthermore we observe a small gain in performance on the learning curves when including more comparisons for prediction. One aspect could be attributed to the pairwise comparisons, but the $Base_{upper}$ shows a very high performance, and given the flexibility of the GP model, it is plausible that this lower performance can be attributed to the audio feature representation.

The issue of scalability is addressed in the present work by investigating the possibility of using multiple participants to make judgments on subsets of a larger data set, and subsequently pooling this data to obtain one large data set. This is investigated by having 13 subjects make comparisons on the same data set and training individual models on their comparisons. The GP framework can model each individual well, although a few models show a relatively higher error rate than others. These can be attributed to lack of understanding of the experiment, scales and appropriateness of scales. Although no clear connection can be attributed solely to the post-questionnaire answers by participants as investigated by using simple correlation analysis. Either they reported incorrectly or the model and features do not capture their interpretation of the experiment. If one used comparisons from these subjects it could increase the noise in the larger data set. When visualizing the ranking in the AV space, as investigated in previous work, we furthermore see a large subjective difference in both valence and arousal for some excerpts. Even though individual models are trained, the difference in rankings would make the solution to the scalability of the 2AFC by pooling subsets of data sets problematic at best.

An alternative method in making 2AFC scalable for evaluating large music collections is to reduce the number of pairwise comparisons, which we investigate by detailed learning curves. The full Bayesian active-learning method EVOI shows the ability of potentially substantially reducing the required number of comparisons needed to obtain a predictive model down to only 15.2% of the comparisons for valence, resulting in 1.3 comparisons per excerpt, and 8.77%, resulting in 0.75 comparisons per excerpt. Although this result is obtained by sampling from the experimental data, the results are promising. Future work can

look into the performance achieved by following the active learning principle applied in the experimental design. In addition, more efficient methods of relative experimental designs should be investigated to obtain multiple pairwise comparisons and still preserving the robustness that the 2AFC provides. Furthermore, based on the findings in present work, more extensive work should be done to find features or representations of features that describe and capture the aspects that express emotions in music.

6 Conclusion

We introduced a two-alternative forced-choice experimental paradigm for quantifying expressed emotions in music along the well-accepted arousal and valance (AV) dimensions. We proposed a flexible probabilistic Gaussian process framework to model the latent AV dimensions directly from the pairwise comparisons. The framework was evaluated on a novel data set and resulted in promising predictive error rates. Comparing the performance of 18 different selections of features, the best performing combination was used to evaluate scalability issues related to the 2AFC experimental paradigm. The possibility of using multiple subjects to evaluate subsets of data, pooled to create a large data set was shown to potentially be problematic due to large individual differences in ranking excerpts on the valence and arousal dimensions. Furthermore, the scalability of the 2AFC and the possibility of using only a fraction of all potential pairwise comparisons was investigated. By applying the active learning method, Expected Value of Information, we showed that a suitable predictive model for arousal and valence can be obtained using as little as 9% and 15% of the total number of possible comparisons, respectively.

Acknowledgments. This work was supported 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.

References

1. Barthet, M., Fazekas, G., Sandler, M.: Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. In: 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) Music and Emotions, pp. 19–22 (June 2012)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
3. Chaloner, K., Verdinelli, I.: Bayesian experimental design: A review. Statistical Science 10(3), 273–304 (1995)
4. Chu, W., Ghahramani, Z.: Preference learning with Gaussian Processes. In: ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning, pp. 137–144 (2005)
5. Cover, T., Thomas, J.: Elements of information theory. Wiley (1991)
6. Cross, I.: The nature of music and its evolution. In: Oxford Handbook of Music Psychology, pp. 3–13. Oxford University Press (2009)

7. Gelman, A.: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1(3), 515–533 (2006)
8. Hevner, K.: Experimental studies of the elements of expression in music. American Journal of Psychology 48(2), 246–268 (1936)
9. Hodges, D.A.: Psychophysiology measures. In: Music and Emotion: Theory, Research, Applications, pp. 279–312. Oxford University Press, New York (2010)
10. Houlsby, N., Hernandez-Lobato, J.M., Huszar, F., Ghahramani, Z.: Collaborative Gaussian processes for preference learning. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 2105–2113 (2012)
11. Huq, A., Bello, J.P., Rowe, R.: Automated Music Emotion Recognition: A Systematic Evaluation. Journal of New Music Research 39(3), 227–244 (2010)
12. Juslin, P.N., Vastfjall, D.: Emotional response to music: The need to consider underlying mechanism. Behavioral and Brain Sciences 31, 559–621 (2008)
13. Juslin, P.N., Sloboda, J.A. (eds.): Music and Emotion: theory, research, applications. Oxford University Press, New York (2010)
14. Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D.: Music emotion recognition: A state of the art review. In: 11th International Conference on Music Information Retrieval (ISMIR), pp. 255–266 (2010)
15. Koelsch, S., Siebel, W.A., Fritz, T.: Functional neuroimaging. In: Music and Emotion: Theory, Research, Applications, pp. 313–346. Oxford University Press, New York (2010)
16. Lartillot, O., Eerola, T., Toiviainen, P., Fornari, J.: Multi-feature modeling of pulse clarity: Design, validation, and optimization. In: 9th International Conference on Music Information Retrieval (ISMIR), pp. 521–526 (2008)
17. Lartillot, O., Toiviainen, P., Eerola, T.: A matlab toolbox for music information retrieval. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization. Springer (2008)
18. Lindley, D.V.: On a measure of the information provided by an experiment. The Annals of Mathematical Statistics 27(4), 986–1005 (1956)
19. Madsen, J.: Modeling of Emotions expressed in Music using Audio features. DTU Informatics, Master Thesis (2011), `http://www2.imm.dtu.dk/pubdb/views/publicationtextunderscoredetails.php?id=6036`
20. Madsen, J., Jensen, B.S., Larsen, J., Nielsen, J.B.: Towards predicting expressed emotion in music from pairwise comparisons. In: 9th Sound and Music Computing Conference (SMC) Illusions (July 2012)
21. Madsen, J., Nielsen, J.B., Jensen, B.S., Larsen, J.: Modeling expressed emotions in music using pairwise comparisons. In: 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) Music and Emotions (June 2012)
22. Mathieu, B., Essid, S., Fillon, T., Prado, J., Richard, G.: An easy to use and efficient audio feature extraction software. In: 11th International Conference on Music Information Retrieval, ISMIR (2010)
23. Müller, M., Ewert, S.: Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In: 12th International Conference on Music Information Retrieval (ISMIR), Miami, USA (2011)
24. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)

25. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161 (1980)
26. Schubert, E.: Measurement and time series analysis of emotion in music. Ph.D. thesis, University of New South Wales (1999)
27. Thurstone, L.L.: A law of comparative judgement. Psychological Review 34 (1927)
28. Train, K.: Discrete Choice Methods with Simulation. Cambridge University Press (2009)
29. Västfjäll, D.: Indirect perceptual, cognitive, and behavioral measures. In: Music and Emotion: Theory, Research, Applications, pp. 255–278. Oxford University Press, New York (2010)
30. Yang, Y.H., Chen, H.: Ranking-Based Emotion Recognition for Music Organization and Retrieval. IEEE Transactions on Audio, Speech, and Language Processing 19(4), 762–774 (2011)
31. Zentner, M., Eerola, T.: Self-report measures and models. In: Music and Emotion: Theory, Research, Applications, pp. 187–222. Oxford University Press, New York (2010)