IEEE Access
practical innovations : open solutions

Received May 28, 2014, accepted June 16, 2014, date of publication June 25, 2014, date of current version July 9, 2014.

Digital Object Identifier 10.1109/ACCESS.2014.2333095

Music Genre and Emotion Recognition Using Gaussian Processes

KONSTANTIN MARKOV[1], (Member, IEEE), AND TOMOKO MATSUI[2], (Member, IEEE)
[1]Division of Information Systems, University of Aizu, Aizuwakamatsu 965-8580, Japan
[2]Department of Statistical Modeling, Institute of Statistical Mathematics, Tokyo 106-8569, Japan

Corresponding author: K. Markov (markov@u-aizu.ac.jp)

ABSTRACT Gaussian Processes (GPs) are Bayesian nonparametric models that are becoming more and more popular for their superior capabilities to capture highly nonlinear data relationships in various tasks, such as dimensionality reduction, time series analysis, novelty detection, as well as classical regression and classification tasks. In this paper, we investigate the feasibility and applicability of GP models for music genre classification and music emotion estimation. These are two of the main tasks in the music information retrieval (MIR) field. So far, the support vector machine (SVM) has been the dominant model used in MIR systems. Like SVM, GP models are based on kernel functions and Gram matrices; but, in contrast, they produce truly probabilistic outputs with an explicit degree of prediction uncertainty. In addition, there exist algorithms for GP hyperparameter learning—something the SVM framework lacks. In this paper, we built two systems, one for music genre classification and another for music emotion estimation using both SVM and GP models, and compared their performances on two databases of similar size. In all cases, the music audio signal was processed in the same way, and the effects of different feature extraction methods and their various combinations were also investigated. The evaluation experiments clearly showed that in both music genre classification and music emotion estimation tasks the GP performed consistently better than the SVM. The GP achieved a 13.6% relative genre classification error reduction and up to an 11% absolute increase of the coefficient of determination in the emotion estimation task.

INDEX TERMS Music genre classification, music emotion estimation, Gaussian processes.

I. INTRODUCTION

Alot of music data have become available recently either locally or over the Internet but in order for users to benefit from them, an efficient music information retrieval technology is necessary. Research in this area has focused on tasks such as genre classification, artist identification, music mood estimation, cover song identification, music annotation, melody extraction, etc. which facilitate efficient music search and recommendation services, intelligent play-list generation and other attractive applications. Information sources for MIR can be: 1) text based - music related Internet sites, social networks, lyrics, etc; 2) audio based - the music signal itself; or 3) mixed text and audio. In this study, we concern ourselves with audio based music genre classification and music emotion[1] estimation tasks.

Genre classification has been one of the most widely researched tasks since the work of Tzanetakis and Cook [1] sparked interest in this area. It is a classical supervised classification task where given labeled data, i.e. songs with their true genre type coming from a finite set of categories (genres), the goal is to predict the genre of an unlabeled music piece. Human categorization of music appears natural, yet it can be inconsistent, changing and, in some cases, may even seem arbitrary. This is probably because human judgements are influenced not only by the audio signal, but also by other factors, such as artist fashion, dance styles, lyrics, social and political attachments, religious believes, etc. [2]. In addition, new genres constantly appear while others become forgotten or irrelevant. Thus, it is impossible to come up with a commonly agreed set of music genres. In MIR studies, researchers usually limit the number of genres to about ten of the most popular and easily distinguished types. Each genre classification system consists of minimum two blocks: feature extractor and classifier. Studies in music processing have investigated various feature types and their extraction algorithms [1], [3], [4]. Carefully crafted music features such as chroma vectors are mostly used for some specific tasks, for example, music transcription or music scene analysis [5]. On the other hand, spectrum and its derivatives are also widely

[1] We assume that the terms "mood" and "emotion" have the same meaning and use them interchangeably.

adopted for music pattern classification. Various methods for building music genre classifiers have been studied, ranging from Support Vector Machines (SVM) to compressive sampling models [6]. However, in most of the studies, parametric models have been utilized. Learning approaches include instances of supervised, semi-supervised [7], and unsupervised [8] methods.

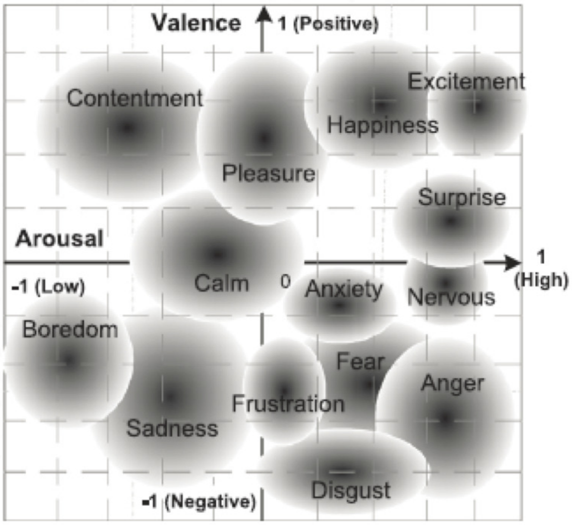


FIGURE 1. Two dimensional (Valence-Arousal) affective space of emotions [32]. Different regions correspond to different categorical emotions.

Although users are more likely to use genres or artists names when searching or categorizing music, the main power of music is in its ability to communicate and trigger emotions in listeners. Thus, determining computationally the emotional content of music is an important task. Existing automatic systems for music mood recognition are based on emotion representation which can be either categorical or dimensional [9], [10]. Categorical approaches involve finding emotional descriptors, usually adjectives, which can be arranged into groups. Given the perceptual nature of human emotion, it is difficult to come up with an intuitive and coherent set of adjectives and their specific grouping. To alleviate the challenge of ensuring consistent interpretation of mood categories, some studies propose to describe emotion using continuous multidimensional metrics defined on low-dimensional spaces. Most widely accepted is the Russell's two-dimensional *Valence-Arousal* (VA) space [11] where emotions are represented by points in the VA plane. Figure 1 shows the space where some regions are associated with distinct mood categories. In this paper, we assume that the task of music emotion recognition is to automatically find the point in the VA plane which corresponds to the emotion induced by a given music piece. Since the Valence and Arousal are by definition continuous and independent parameters, we can estimate them separately using the same music feature representation and two different regression models. Prior studies focused on searching for mood specific acoustic features have not found any dominant single one [12], so the most commonly used are those employed in the other MIR tasks as well. Regression models, such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), or Adaboost.RT, as well as Multi-Level Least-Squares or regression trees have [10] been successfully applied to music emotion estimation. Model learning is again supervised and requires labeled training data. Finding consistent mood labels in terms of VA values is even more challenging than obtaining genre labels since emotion interpretation can be very subjective and varies among listeners. It requires music annotation by multiple experts which, is expensive, time consuming, and labor intensive [13].

Gaussian Processes have been known as non-parametric Bayesian models for quite some time, but just recently have attracted attention of researchers from other fields than statistics and applied mathematics. After the work of Rasmussen and Williams [14] which introduced GPs for the machine learning tasks of classification and regression, many researchers have utilized GPs in various practical applications. As SVMs, they are also based on kernel functions and Gram matrices, and can be used as their plug-in replacement. The advantage of GPs with respect to SVMs is that their predictions are truly probabilistic and that they provide a measure of the output uncertainty. Another big plus is the availability of algorithms for their hyper parameter learning. The downside is that the GP training complexity is $\mathcal{O}(n^3)$, which makes them difficult to use in large scale tasks. Several sparse approximation methods have been proposed [15], [16], but this problem has not yet been fully solved and is a topic of an ongoing research.

The goal of this work is to investigate the applicability of Gaussian Process models to music genre and emotion recognition tasks and to compare their performance with the current state-of-the-art Support Vector Machines. Some of our preliminary studies [17], [18] had shown that GPs can be a feasible alternative to SVMs, but more careful and thorough investigation was necessary in order to confirm those findings. Here, using two databases of similar size for each task, the same set of features and the same experimental settings, we evaluated both the GP and SVM models and compared their performances. Results clearly show that GPs outperform SVMs in both tasks. Genre classification accuracy of the GPs was higher in all cases and the gain in the Valence estimation, which is considered much more difficult than Arousal estimation, was up to 11% absolute in terms of R^2 metric. We have to note that, since each music piece in our experiments was represented by a single feature vector, this may not be classified as a large scale evaluation whereby SVMs could have practical advantage because of their lower computational complexity. In this regard, further research involving sparse GP learning and inference methods is necessary.

II. RELATED STUDIES

As we mentioned in the previous section, music genre classification is one of the most popular MIR tasks. What also contributes to its popularity is public availability of music data, such as GTZAN [1] and ISMIR2004 [19] databases. A good review of the feature extraction and modeling methods for genre classification is given in [20], where results obtained by various research teams using the same GTZAN data are also compared. According to this review, the most widely applied classifier is the SVM, which achieves accuracy between 70.4% and 79.8% depending on the features used. Finding good music signal representation is an important task and some studies have proposed carefully designed features, such as perceptually based acoustic features [21], or modulation spectrum analysis features [22]. Both of them achieve very good results, but require long duration signals to work. Conventional features used for genre classification can be divided into "low-level" features including timbre (zero crossing rate, spectral centroid, flux, rolloff, MFCC, and others) and temporal (amplitude modulation or auto-regressive coefficients) features, as well as ""mid-level" features, such as rhythm, pitch and harmony [20]. On the other hand, it is also possible to apply unsupervised learning methods to find some "high level" representations of the "low-level" features, and then use them as a new type of features. This can be accomplished using Non-Negative Matrix Factorization (NMF), sparse coding [7], or Deep Neural Networks (DNN) [23]. In both cases, genre classification is done by standard classifiers, SVM, and Neural Network (additional DNN layer), respectively.

Audio based emotion prediction research has been focused on the challenge of finding the best combination of features, learning methods, and mood representation schemes [9]. Categorical emotion recognition is similar to the genre classification and the approaches are similar as well. In one of the earliest studies, features representing timbre, rhythm, and pitch have been used in SVM based system to classify music into 13 mood categories [24]. With 499 hand-labeled 30-sec. clips, an accuracy of 45% was achieved. In 2007, music emotion classification was included in the MIR evaluation exchange (MIREX) benchmarks and the best performance of 61.5% was again achieved using SVM classifier [25]. However, recent studies have suggested that regression approaches using continuous mood representation can perform better than categorical classifiers [26]. Support Vector Regression (SVR) was applied in [12] to map music clips, each represented by a single feature vector, into two dimensional VA space. After principal component analysis (PCA) based feature dimensionality reduction, this system achieved R^2 scores of 0.58 and 0.28 for arousal and valence, respectively. Later, this approach was extended by representing perceived emotion of a clip as a probability distribution in the emotion plane [27]. It also is possible to combine categorical and continuous emotion representations by quantizing the VA space and apply emotion cluster classification using SVM [28], or another regression model, trained for each cluster [29]. It can be argued that emotions are not necessarily constant, but can vary during the course of a song. In this case, time-varying emotion estimation or emotion tracking methods are required. One approach is to divide a piece of music into segments short enough to assume that mood does not change within each segment, and then use standard emotion recognition techniques [30]. Another study [31] considers arousal and valence as latent states of a linear dynamical system and applies Kalman filter to recover the mood dynamics over time.

Although Gaussian Processes have become popular in machine learning community and have been used in such tasks as object categorization in computer vision [33] or economics and environmental studies [34], there are still few GP applications in the field of signal processing. In one such application, GP regression model is applied to time domain voice activity detection and speech enhancement [35]. In [36], using GP, researchers estimate speakers likability given recordings of their voices. Another recent study employs GPs for head-related transfer function (HRTF) estimation in acoustic scene analysis [37]. Lately, several extensions and new models based on GPs have been developed. For example, Gaussian Process latent variable model (GP-LVM) was introduced for non-linear dimensionality reduction [38], but have also been applied to image reconstruction [39] and human motion modeling [40]. Another promising extension is the Gaussian Process Dynamic Model (GPDM) [41]. It is a non-linear dynamical system which can learn the mapping between two continuous variables spaces. One of the first applications of GPDM in audio signal processing was for speech phoneme classification [42]. Although the absolute classification accuracy of the GPDM was not high, in certain conditions they outperformed the conventional hidden Markov model (HMM). In [43], GPDM is used as a model for non-parametric speech representation and speech synthesis. Similar to GPDM is the GP based state-space model [44], [45]. It is essentially a non-linear Kalman filter and is very useful for time series processing. Compared to some approximate Gaussian filters, such as the Extended Kalman filter (EKF) and the Unscented Kalman filter (UKL), it gives exact expected values in the prediction and filter steps.

III. GAUSSIAN PROCESSES

Gaussian processes are used to describe distributions over functions. Formally, the GP is defined as a collection of random variables any finite number of which has a joint Gaussian distribution [14]. It is completely specified by its mean and covariance functions. For a real process $f(\boldsymbol{x})$, the mean function $m(\boldsymbol{x})$ and the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ are defined as

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})] \\ k(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))]. \end{aligned} \tag{1}$$

Thus, the GP can be written as

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')). \tag{2}$$

A GP prior over function $f(\boldsymbol{x})$ implies that for any finite number of inputs $\boldsymbol{X} = \{\boldsymbol{x}_i\} \in \mathbb{R}^d$, $i = 1, \ldots, n$, the vector of function values $\boldsymbol{f} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]^T = [f_1, \dots, f_n]^T$ has a multivariate Gaussian distribution

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}) \tag{3}$$

where the mean $\boldsymbol{m}$ is often assumed to be zero. The covariance matrix $\boldsymbol{K}$ has the following form

$$\boldsymbol{K} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \tag{4}$$

and characterizes the correlation between different points in the process. For $k(\boldsymbol{x}, \boldsymbol{x}')$, any kernel function which produces symmetric and semi-definite covariance matrix can be used.

IV. GAUSSIAN PROCESS REGRESSION

Given input data vectors $\boldsymbol{X} = \{\boldsymbol{x}_i\}, i = 1, \dots, n$ and their corresponding target values $\boldsymbol{y} = \{y_i\}$, in the simplest regression task, y and $\boldsymbol{x}$ are related as

$$y = f(\boldsymbol{x}) + \varepsilon \tag{5}$$

where the latent function $f(\boldsymbol{x})$ is unknown and ε is often assumed to be a zero mean Gaussian noise, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. Putting a GP prior over $f(\boldsymbol{x})$ allows us to marginalize it out, which means that we do not need to specify its form and parameters. This makes our model very flexible and powerful since $f(\boldsymbol{x})$ can be any non-linear function of unlimited complexity.

In practice, targets y_i are assumed to be conditionally independent given f_i, so that the likelihood can be factorized as

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_1^n p(y_i|f_i) \tag{6}$$

where $p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma_n^2)$, according to our observation model Eq.(5). Since $\boldsymbol{f}$ has normal distribution, i.e. $\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$, it follows that $\boldsymbol{y}$ is also a Gaussian random vector

$$p(\boldsymbol{y}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{K} + \sigma_n^2 \boldsymbol{I}). \tag{7}$$

Given some new (test) input $\boldsymbol{x}_*$, we can now estimate the unknown target y_* and, more importantly, its distribution. Graphically, the relationship between all involved variables can be represented as shown in Fig.(2). To find y_*, we first obtain the joint probability of training targets $\boldsymbol{y}$ and $f_* = f(\boldsymbol{x}_*)$, which is Gaussian

$$p(\boldsymbol{y}, f_*|\boldsymbol{x}_*, \boldsymbol{X}) = \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{K} + \sigma_n^2 \boldsymbol{I} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & k(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix}\right) \tag{8}$$

where $\boldsymbol{k}_*^T = [k(\boldsymbol{x}_1, \boldsymbol{x}_*), \dots, k(\boldsymbol{x}_n, \boldsymbol{x}_*)]$. Then, from this distribution, it is easy to obtain the conditional $p(f_*|\boldsymbol{y}, \boldsymbol{x}_*, \boldsymbol{X})$, which is also Gaussian

$$p(f_*|\boldsymbol{y}, \boldsymbol{x}_*, \boldsymbol{X}) = \mathcal{N}(f_*|\mu_{f_*}, \sigma_{f_*}^2) \tag{9}$$

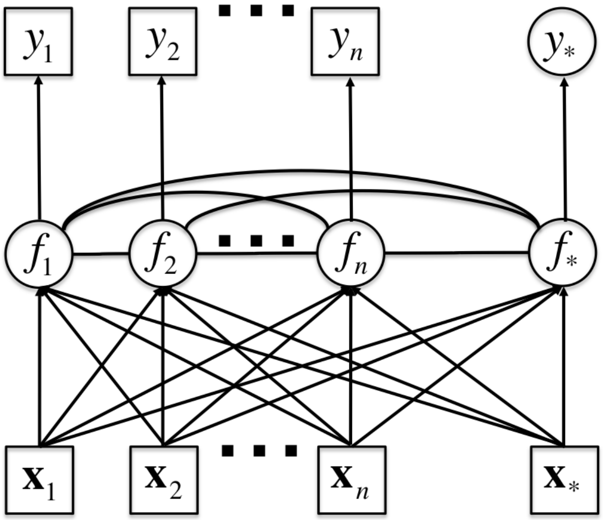


FIGURE 2. Graphical representation of observable $\boldsymbol{x}$, $\boldsymbol{y}$, (enclosed in squares), latent $\boldsymbol{f}$, and unobservable $\boldsymbol{y}_*$ (enclosed in circles) variable relationships in Gaussian Process based regression task.

with mean and variance

$$\mu_{f_*} = \boldsymbol{k}_*^T (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{y}, \tag{10}$$

$$\sigma_{f_*}^2 = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{k}_* \tag{11}$$

It is worth noting that the mean μ_{f_*} is a linear combination of the observed targets $\boldsymbol{y}$. It can also be viewed as a linear combination of the kernel functions $k(\boldsymbol{x}_*, \boldsymbol{x}_i)$. On the other hand, the variance $\sigma_{f_*}^2$ depends only on inputs $\boldsymbol{X}$.

To find out the predictive distribution of y_*, we marginalize out f_*

$$\begin{aligned} p(y_*|\boldsymbol{y}, \boldsymbol{x}_*, \boldsymbol{X}) &= \int p(y_*|f_*) p(f_*|\boldsymbol{y}, \boldsymbol{x}_*, \boldsymbol{X}) df_* \\ &= \mathcal{N}(y_*|\mu_{y_*}, \sigma_{y_*}^2) \end{aligned} \tag{12}$$

where it is easy to show that for homoscedastic likelihood, as in our case, the predictive mean and variance are [46]

$$\mu_{y_*} = \mu_{f_*}, \text{ and} \tag{13}$$

$$\sigma_{y_*}^2 = \sigma_{f_*}^2 + \sigma_n^2. \tag{14}$$

Making this mean our predicted target, $y_{pred} = \mu_{y_*}$ will minimize the risk for a squared loss function $(y_{true} - y_{pred})^2$. The variance $\sigma_{y_*}^2$, on the other hand, shows the model uncertainty about y_{pred}.

A. PARAMETER LEARNING

Until now, we have considered fixed covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$, but in general, it is parameterized by some parameter vector $\boldsymbol{\theta}$. This introduces *hyper-parameters* to GP, which are unknown and, in practice, very little information about them is available. A Bayesian approach to their estimation would require a *hyper-prior* $p(\boldsymbol{\theta})$ and evaluation of the following posterior

$$p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X})} = \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{15}$$

where the likelihood $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})$ is actually the GP marginal likelihood over function values $\boldsymbol{f}$

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta})d\boldsymbol{f}. \tag{16}$$

However, the evaluation of the integral in Eq.(15) can be difficult and as an approximation we may directly maximize Eq.(16) w.r.t. the hyper-parameters $\boldsymbol{\theta}$. This is known as maximum likelihood II (ML-II) type hyper-parameter estimation. Since both the GP prior $\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{K})$ and the likelihood $\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f},\sigma_n^2\boldsymbol{I})$ are Gaussians, the logarithm of Eq.(16) can be obtained analytically

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}_y| - \frac{n}{2}\log 2\pi \tag{17}$$

where $\boldsymbol{K}_y = \boldsymbol{K} + \sigma_n^2\boldsymbol{I}$ is the covariance matrix of the noisy targets $\boldsymbol{y}$. Hyper parameters $\boldsymbol{\theta} = \{\sigma_n^2, \boldsymbol{\theta}_k\}$ include the noise variance and parameters of the kernel function. Those which maximize Eq.(17) can be found using gradient based optimization method. Partial derivatives for each θ_i are found from

$$\begin{aligned}\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i} &= -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\frac{\partial\boldsymbol{K}_y}{\partial\theta_i}\boldsymbol{K}_y^{-1}\boldsymbol{y}\\ &\quad -\frac{1}{2}\mathtt{tr}(\boldsymbol{K}_y^{-1}\frac{\partial\boldsymbol{K}_y}{\partial\theta_i})\end{aligned} \tag{18}$$

where for $\theta_i = \sigma_n^2$ we have

$$\frac{\partial\boldsymbol{K}_y}{\partial\sigma_n^2} = \sigma_n^2\boldsymbol{I}. \tag{19}$$

Usually, kernel function parameters are all positive, which would require constrained optimization. In practice, this problem is easily solved by optimizing with respect to the logarithm of the parameters, so simple unconstrained optimization algorithms can be used.

V. GAUSSIAN PROCESS CLASSIFICATION

For binary classification, given training data vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ with corresponding labels $y_i \in \{-1,+1\}$, we would like to predict the class membership probability of a test point $\boldsymbol{x}_*$. This is done using an unconstrained latent function $f(\boldsymbol{x})$ with GP prior and mapping its value into the unit interval [0, 1] by means of a sigmoid shaped function [47]. Common choice for such function is the logistic function or the cumulative density function Φ of the standard Gaussian distribution. When the sigmoid is point symmetric, the likelihood $p(y|\boldsymbol{x})$ can be written as $\mathrm{sig}(y \cdot f(\boldsymbol{x}))$.

Let $\boldsymbol{X} = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]$ be the training data matrix, $\boldsymbol{y} = [y_1,\ldots,y_n]^T$ be the vector of target values, and $\boldsymbol{f} = [f_1,\ldots,f_n]^T$ with $f_i = f(\boldsymbol{x}_i)$ be the vector of latent function values. Given the latent function, the class labels are assumed independent Bernoulli variables and therefore the likelihood can be factorized as

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{n} p(y_i|f_i) = \prod_{i=1}^{n}\mathrm{sig}(y_if_i) \tag{20}$$

Using the Bayes' rule and since by definition $p(\boldsymbol{f}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K})$, we can express the posterior distribution over the latent values as

$$p(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{\int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f}} \tag{21}$$

$$= \frac{\mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K})}{p(\boldsymbol{y}|\boldsymbol{X})}\prod_{i=1}^{n}\mathrm{sig}(y_if_i). \tag{22}$$

Unfortunately, both the likelihood $p(\boldsymbol{y}|\boldsymbol{f})$ and the marginal $p(\boldsymbol{y}|\boldsymbol{X})$ are non-Gaussian, so an analytic solution is impossible. Approximations in this case are either based on a Gaussian approximation to the posterior or Markov Chain Monte Carlo (MCMC) sampling [47].

For a test vector $\boldsymbol{x}_*$, we first find predictive distribution for the corresponding latent variable f_* by marginalizing over the training set latent variables

$$p(f_*|\boldsymbol{x}_*,\boldsymbol{y},\boldsymbol{X}) = \int p(f_*|\boldsymbol{f},\boldsymbol{x}_*,\boldsymbol{X})p(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})d\boldsymbol{f} \tag{23}$$

where the conditional prior

$$p(f_*|\boldsymbol{f},\boldsymbol{x}_*,\boldsymbol{X}) = \mathcal{N}(f_*|\boldsymbol{k}_*^T\boldsymbol{K}^{-1}\boldsymbol{f}, k(\boldsymbol{x}_*,\boldsymbol{x}_*) - \boldsymbol{k}_*^T\boldsymbol{K}^{-1}\boldsymbol{k}_*) \tag{24}$$

is Gaussian.

Finally, the predictive class membership probability is obtained by averaging out the test latent variable

$$\begin{aligned}p(y_*|\boldsymbol{x}_*,\boldsymbol{y},\boldsymbol{X}) &= \int p(y_*|f_*)p(f_*|\boldsymbol{x}_*,\boldsymbol{y},\boldsymbol{X})df_*\\ &= \int \mathrm{sig}(y_*f_*)p(f_*|\boldsymbol{x}_*,\boldsymbol{y},\boldsymbol{X})df_*\end{aligned} \tag{25}$$

A Gaussian approximation to the posterior of Eq.(21), $q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}|\bar{\boldsymbol{f}},\boldsymbol{A})$ gives rise to an approximate predictive distribution for the test data, i.e. $q(f_*|\boldsymbol{x}_*,\boldsymbol{y},\boldsymbol{X}) = \mathcal{N}(f_*|\mu_*,\sigma_*^2)$, with mean and variance

$$\begin{aligned}\mu_* &= \boldsymbol{k}_*^T\boldsymbol{K}^{-1}\bar{\boldsymbol{f}}\\ \sigma_*^2 &= k(\boldsymbol{x}_*,\boldsymbol{x}_*) - \boldsymbol{k}_*^T(\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{K}^{-1})\boldsymbol{k}_*\end{aligned} \tag{26}$$

When the cumulative Gaussian density function Φ is used as a likelihood function, the approximate probability of $\boldsymbol{x}_*$ having label $y_* = +1$ can be calculated analytically

$$\begin{aligned}q(y_* = +1|\boldsymbol{x}_*,\boldsymbol{y},\boldsymbol{X}) &= \int \Phi(f_*)\mathcal{N}(f_*|\mu_*,\sigma_*^2)df_*\\ &= \Phi(\frac{\mu_*}{\sqrt{1+\sigma_*^2}})\end{aligned} \tag{27}$$

The parameters $\bar{\boldsymbol{f}}$ and $\boldsymbol{A}$ of the posterior approximation can be found using either the Laplace's method or the Expectation Propagation (EP) algorithm [48].

A. PARAMETER LEARNING

As in the case of Gaussian Process regression, kernel function parameters can be learned by marginal likelihood $p(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\theta})$ maximization. However, in this case, the likelihood $p(\mathbf{y}|\boldsymbol{f})$ is no longer Gaussian and analytic solution does not exist. Again, Laplace or EP approximation can be used. For the maximization, good candidates are gradient based methods, such as the conjugate gradient optimization or the BFGS algorithm [49].

B. RELATION TO SVM

For the soft margin support vector machine, the optimization problem is defined as

$$\min_{\boldsymbol{w}, w_0} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(1 - y_i f_i) \tag{28}$$
$$\text{s.t. } 1 - y_i f_i \geq 0, \; i = 1, \ldots, n$$

where $f_i = f(\boldsymbol{x}_i) = \boldsymbol{w}\boldsymbol{x}_i + w_0$ and the solution has the form $\boldsymbol{w} = \sum_i \lambda_i y_i \boldsymbol{x}_i = \sum_i \alpha_i \boldsymbol{x}_i$. Thus, the square norm of $\boldsymbol{w}$ becomes

$$\|\boldsymbol{w}\|^2 = \sum_{i,j} \alpha_i \alpha_j \boldsymbol{x}_i \boldsymbol{x}_j \tag{29}$$

which in matrix form and using kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ instead of $\boldsymbol{x}_i\boldsymbol{x}_j$ is

$$\|\boldsymbol{w}\|^2 = \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} = \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f} \tag{30}$$

where $\boldsymbol{f} = \boldsymbol{K\alpha}$. Then, substituting $\|\boldsymbol{w}\|^2$ in Eq.(28) we obtain the following objective function

$$\frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1}\boldsymbol{f} + C\sum_{i=1}^{n}(1 - y_i f_i) \tag{31}$$
$$\text{s.t. } 1 - y_i f_i \geq 0, \; i = 1, \ldots, n$$

which requires constrained optimization. On the other hand, in the GP classification, during the posterior approximation we need to find the maximum a posteriori value $\bar{\boldsymbol{f}}$ of $p(\boldsymbol{f}|\mathbf{y}, \boldsymbol{X})$ by maximizing the $\log p(\mathbf{y}|\boldsymbol{f}) + \log p(\boldsymbol{f}|\boldsymbol{X})$ which becomes

$$\log p(\mathbf{y}|\boldsymbol{f}) - \frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1}\boldsymbol{f} - \frac{1}{2}\log|\boldsymbol{K}| - \frac{n}{2}\log 2\pi \tag{32}$$

when using zero mean GP prior $\mathcal{N}(\boldsymbol{f}|0, \boldsymbol{K})$. Since the last two terms are constant when the kernel is fixed, it is equivalent to minimizing the following quantity

$$\frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1}\boldsymbol{f} - \sum_{i=1}^{n}\log p(y_i|f_i) \tag{33}$$

Apparently, there is a strong similarity between the SVM optimization problem and the MAP maximization of the GP classifier. Thus, there is a close correspondence between their solutions [14]. Note that $-\sum_{i=1}^{n}\log p(y_i|f_i)$ is always positive and, therefore, no constrained optimization is required.

One big advantage of the GP classifier is that the output it produces - the prediction for $p(y = +1|\boldsymbol{x})$ - is clearly probabilistic. Furthermore, it provides a measure of uncertainty for this prediction, i.e. the predictive variance of $f(\boldsymbol{x})$. Although it is possible to give probabilistic interpretation to the SVM outputs by wrapping them with sigmoid function, this is a rather *ad hoc* procedure which also requires tuning of the sigmoid parameters [50].

VI. EXPERIMENTS WITH MUSIC EMOTION RECOGNITION

In this study, we assume that music emotion recognition is to estimate the Valence-Arousal (VA) values for a song, or a clip as in our case, given its feature representation. Separate Gaussian Process regression (GPR) and Support Vector regression (SVR) models are independently trained using the same training data and corresponding reference VA values.

The models' performance is measured in terms of R^2 measure. It is widely used to describe the goodness of fit of a statistical model and is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{34}$$

where y_i are the reference values, $\bar{y}$ is their mean, and f_i are the corresponding estimates. R^2 takes values in the range $[0, 1]$[2] with $R^2 = 1$ meaning a perfect data fit.

A. DATABASE AND FEATURE EXTRACTION

For the music emotion recognition experiments we used the ''MediaEval'2013'' database [13]. It consists of 1000 clips (each 45 seconds long) taken at random locations from 1000 different songs. They belong to the following 8 music genres: Blues, Electronic, Rock, Classical, Folk, Jazz, Country, and Pop, and were distributed uniformly, i.e. 125 songs per genre. There were 53-100 unique artists per genre, which provide a good distribution across artists. Each clip has been annotated with Arousal and Valence score on a 9 point scale by a number of annotators. In total there have been 100 annotators each of whom annotating 107.9 clips on average. The mean of annotator scores has been taken as final Arousal or Valence label for each clip.

All music audio data were monaural with sampling frequency of 44.1kHz, which was reduced to 22.05kHz before the feature extraction. In our experiments, we adopted those feature extraction methods which are widely used in music signal processing studies and can be referred to as a ''standard'' set for such tasks. They include the following:

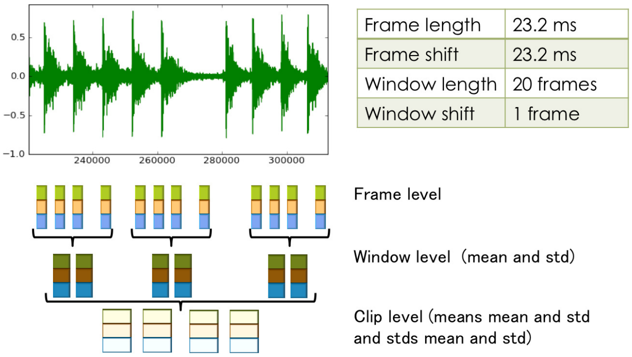
- MFCC (mel frequency cepstral coefficients) - first proposed for speech recognition - they are also one of the main features for music processing;
- LSP (line spectral pairs) - another speech related feature used for speech coding representing the LPC coefficients;
- TMBR (timbre features) - a set of four scalar features consisting of spectral centroid, spectral flux, spectral rolloff, and zero crossings;
- SCF and SFM (spectral crest factor and spectral flatness measure) - these features are subband based

[2] In practice, it can take values outside this range, which would indicate estimation failure.

measures indicating spectral shape and used to discriminate between tone-line and noise-like sounds;

- CHR (chromagram) - this feature represents the spectrum distribution of the distinct semitones and provides information about the key and mode.

We used the Marsyas tool [51], which allows all of the above features to be extracted as well as any combination of them. The analysis frame size was set to 512 points, which corresponded to 23.2msec at 22.05kHz sampling rate. There was no overlap between neighboring frames. We used default dimensionality for each feature type which for MFCC was 13, for LSP was 18, for each timbre feature was 1, and for each SFM, SCF, and CHR was 24. When multiple features are calculated, they are stacked into a single vector per frame. Further, the vector sequence is divided into windows of 20 vectors and the mean and standard deviation are calculated for each window. Finally, same statistics (mean and std) are calculated for the window means and standard deviations, and stacked into a single vector which represents the whole clip. Schematically, this process is illustrated in Fig.3.



Frame length	23.2 ms
Frame shift	23.2 ms
Window length	20 frames
Window shift	1 frame

FIGURE 3. **Feature extraction process. Mean and standard deviation are calculated for a window of several frame level feature vectors. At clip level, same statistics are obtained for means and standard deviations separately and then stacked together to form a single feature vector per clip.**

B. SVM BASELINE RESULTS

Our SVR emotion estimation system was built using the LIBSVM toolkit [52]. We chose the Radial Bases function (RBF) kernel because during our previous experiments [18] it had shown better performance than the linear kernel.

We experimented with several different feature sets, starting from only MFCC and gradually adding new features. The cost parameter C was manually optimized using grid search for each feature set. The RBF kernel scale parameter was set to its default value. Table 1 shows the SVR performance in terms of R^2 for both Arousal and Valence estimators. Results are given as mean and standard deviation of a 10-fold cross-validation experiments. Each raw of the table gives R^2 values for a different feature set, where "(1)+TMBR" indicates the feature set of case 1, i.e. MFCC, plus TMBR features. Clip level feature vector dimensionality is shown in column "Dims."

TABLE 1. **SVM performance in terms of R^2 measure. The kernel function is RBF. Results are given as mean and std values of 10-fold cross-validation.**

Case	Features	Dims	Arousal	Valence
(1)	MFCC	52	0.630 ± 0.064	0.356 ± 0.085
(2)	(1)+TMBR	68	0.630 ± 0.065	0.354 ± 0.084
(3)	(2)+SCF+SFM	260	0.686 ± 0.045	0.398 ± 0.075
(4)	(3)+CHR+LSP	388	0.587 ± 0.066	0.341 ± 0.097

TABLE 2. **GPR performance in terms of R^2 measure using squared exponential covariance and constant mean functions. Results are given as mean and std values of 10-fold cross-validation.**

Case	Features	Dims	Arousal	Valence
(1)	MFCC	52	0.623 ± 0.057	0.323 ± 0.073
(2)	(1)+TMBR	68	0.634 ± 0.063	0.340 ± 0.082
(3)	(2)+SCF+SFM	260	0.661 ± 0.061	0.391 ± 0.056
(4)	(3)+CHR+LSP	388	0.665 ± 0.048	0.442 ± 0.049

As can be seen, adding timbral features to the MFCC did not have any effect. In contrast, the spectral crest factor (SCF) and spectral flatness measure (SFM) provided significant improvement. On the other hand, the chromagram and line spectral pairs negatively affected the performance.

C. GP REGRESSION RESULTS

For the first experiments with GP regression, we used the same feature sets as with the SVR baseline, and choose similar kernel: the Squared Exponential (SE) function which is defined as

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 \exp(-\frac{1}{2l^2}(\boldsymbol{x} - \boldsymbol{x}')^T(\boldsymbol{x} - \boldsymbol{x}')) \qquad (35)$$

where the scale σ_k^2 and length l are the kernel parameters. In contrast to the SVR case, however, GRP kernel and noise (σ_n^2 from Eq.7) parameters are learned from the training data and manual tuning is not necessary. The GP mean function $m(x)$ is usually set to zero, but the GP regression and classification toolbox (GPML [46]) we used to implement our GPR system allows the mean to be set to a constant other than zero, which is also estimated during training.

Table 2 presents the GPR performance for different feature sets using squared exponential kernel and constant mean function. It is directly comparable with Table 1, which shows the SVR results. In the GPR case, we can see that the bigger the feature set, the better the R^2 score for both Arousal and Valence. The best Arousal result is worse than the SVR one, but the Valence value of 0.442 is 11% better than the 0.398 baseline.

For the following experiments we used only the full feature set (condition (4) in Table 2) because it yielded the best results. Since the GPR learning allows kernel function parameters to be estimated automatically, wide variety of differentiable kernel functions can be utilized. Beside the squared exponential (SE), we used the following functions:

- Linear (LIN) with parameter l,

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + 1)/l^2 \tag{36}$$

- Rational Quadratic (RQ) with parameters σ_k, α and l,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 (1 + \frac{1}{2\alpha l^2} (\boldsymbol{x} - \boldsymbol{x}')^T (\boldsymbol{x} - \boldsymbol{x}'))^{-\alpha} \tag{37}$$

- Matérn of 3rd degree (MAT3) with parameters σ_k and l,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_k^2 (1 + r) \exp(-r), \tag{38}$$

$$r = \sqrt{\frac{3}{l^2} (\boldsymbol{x} - \boldsymbol{x}')^T (\boldsymbol{x} - \boldsymbol{x}')}$$

It is possible to combine several kernel functions into a more complex one using sum or product operations. We experimented with various kernel combinations and the results of those which showed the best performance are summarized in the lower part of Table 3. The upper part of the table shows the results of each individual kernel function. The mean function in these experiments was constant.

TABLE 3. GPR performance in terms of R^2 measure using different kernel functions, as well as their best combinations. Results are given as mean and std values of 10-fold cross-validation.

Covariance	Arousal	Valence
LIN	0.605 ± 0.066	0.356 ± 0.063
SE	0.665 ± 0.048	0.442 ± 0.049
RQ	0.693 ± 0.041	0.472 ± 0.055
MAT3	0.685 ± 0.042	0.463 ± 0.052
LIN + RQ	0.695 ± 0.046	0.470 ± 0.056
SE + MAT3	0.694 ± 0.039	0.471 ± 0.057
SE * RQ	0.692 ± 0.040	0.473 ± 0.055

As evident from Table 3, the GPR performance greatly depends on the selected kernel function. The Rational Quadratic is the best, followed by the Matérn of 3rd degree. Small additional improvement can be achieved using composite kernels such as LIN+RQ or SE*RQ. The best GPR score is much better than the SVR score in both Arousal and Valence estimation tasks. We have to note that kernels such as RQ or MAT3 can also be used in the SVR case. The practical problem here is that SVM framework does not allow for kernel function parameter estimation as GP does. This greatly reduces the range of useful SVR kernels and makes finding their parameters a tedious task.

Finally, Fig.4 compares the best results of the GPR and SVR based systems in terms of R^2 measure. In both cases the GPR is better, achieving bigger improvement for the Valence estimation, which is traditionally the more difficult task.

VII. EXPERIMENTS WITH MUSIC GENRE CLASSIFICATION

In these experiments, we again compared the Gaussian Processes and Support Vector Machines, but in the classification task. We kept the same amount of data, feature extraction methods and cross-validation type of evaluation as in the previous regression task.

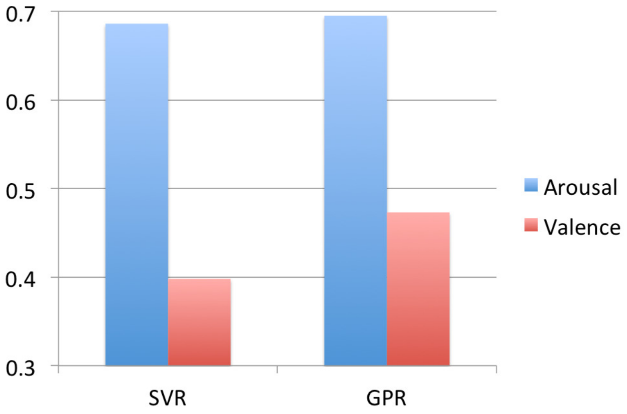


FIGURE 4. Gaussian Process (GPR) and Support Vector machine regression (SVR) best performance comparison in terms of R^2 for both the Arousal and Valence prediction tasks.

A. DATABASE AND FEATURE EXTRACTION

We used the popular GTZAN song collection [1] which consisted of 30 second long music clips belonging to one of the following 10 genres: Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, and Rock. There were 100 clips per genre and 1000 clips in total.

All 1000 clips were processed in the same way as the MediaEval'2013 data for music emotion estimation and exactly the same features were extracted as well. Again, each music clip was represented by a single feature vector consisting of two level statistics of the frame level features, as depicted in Fig.3.

B. SVM AND GP CLASSIFICATION EVALUATION

Since the SVM and GP are binary classifiers, in both cases, multiclass classification is simulated by one-versus-others setting. As in the music emotion experiments, SVM cost parameter C was manually optimized and the RBF scale was set to its default value.

In the GP classification, the likelihood function should have the shape of sigmoid. We tried two such functions:

- *Logistic* defined as

$$p(y|f) = \frac{1}{1 + \exp(-yf)}, \tag{39}$$

- and *Error function* (ERF) defined as

$$p(y|f) = \int_{-\infty}^{yf} \mathcal{N}(t) dt. \tag{40}$$

Both functions are in their simplest form, i.e. with no parameters, so they do not have to be estimated during training.

Table 4 compares SVM and GP based classification systems' performance for various feature sets. The GP model was trained using SE covariance, zero mean and ERF likelihood functions. These results clearly show that GP consistently outperforms the SVM classifier in all cases. Again the best performance is achieved with the full feature set: MFCC+TMBR+SCF+SFM+CHR+LSP.

TABLE 4. Music genre classification accuracy(%). The SVM kernel function is RBF. The GP classifier uses SE covariance, ERF likelihood and zero mean. Results are given as mean and std values of 10-fold cross-validation.

Case	Features	Dims	SVM	GP
(1)	MFCC	52	65.8 ± 6.3	68.7 ± 6.3
(2)	(1)+TMBR	68	70.2 ± 5.5	72.2 ± 6.0
(3)	(2)+SCF+SFM	260	74.9 ± 3.1	76.4 ± 3.1
(4)	(3)+CHR+LSP	388	76.5 ± 3.5	78.3 ± 3.7

TABLE 5. GP music genre classification accuracy (%). results are given as mean and std values of 10-fold cross-validation.

Covariance	Likelihood	
	ERF	Logistic
LIN	75.8 ± 3.0	75.9 ± 3.1
SE	78.3 ± 3.7	78.4 ± 3.1
RQ	78.7 ± 3.4	78.7 ± 3.3
MAT3	78.6 ± 3.2	78.6 ± 3.1
LIN + RQ	78.9 ± 3.2	79.0 ± 3.4
SE + RQ	78.9 ± 3.2	79.3 ± 3.0
LIN * RQ	78.5 ± 3.4	79.3 ± 3.3

A comparison between the two GP likelihood functions with respect to various covariance kernels and their combinations is given in Table 5. It seems that the *Logistic* function is slightly better, especially in the composite kernels case. The absolute difference between GP and SVM best results of 79.3% and 76.5% is 2.8%, which corresponds to 13.6% relative error reduction.

VIII. CONCLUSION

In this paper, we described and evaluated two systems based on Gaussian Process models for music genre and emotion recognition, respectively. In each of these tasks, Support Vector Machine is currently considered as the state-of-the-art model and therefore we used it for comparison.

The GP and SVM have many common characteristics. They are both non-parametric, kernel based models, and their implementation and usage as regressors or binary classifiers are the same. However, GP are probabilistic Bayesian predictors which in contrast to SVM produce Gaussian distributions as their output. Another advantage is the possibility of parameter learning from the training data. On the other hand, SVM provide sparse solution, i.e. only ‘‘support'' vectors are used for the inference, which can be a plus when working with large amount of data.

The evaluation experiments carried out using the MediaEval'2013 music database for emotion estimation and GTZAN corpus for genre classification have shown that GP models consistently outperform the SVM, especially in the classification task.

We have extended the GP application field into the area of music information retrieval, but there are many other unexplored research directions where GP can become viable alternative to the current state-of-the-art methods. One such direction is speech processing and recognition where high performance temporal sequences discrimination and non-linear dynamical system modeling are demanded.

REFERENCES

[1] G. Tzanetakis and P. Cook, ‘‘Musical genre classification of audio signals,'' *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[2] J. C. Lena and R. A. Peterson, ‘‘Classification as culture: Types and trajectories of music genres,'' *Amer. Soc. Rev.*, vol. 73, no. 5, pp. 697–718, Oct. 2008.

[3] R. Typke, F. Wiering, and R. C. Veltkamp, ‘‘A survey of music information retrieval systems,'' in *Proc. Int. Conf. Music Inform. Retr.*, 2005, pp. 153–160.

[4] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, ‘‘Content-based music information retrieval: Current directions and future challenges,'' *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, Apr. 2008.

[5] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York, NY, USA: Springer-Verlag, 2006.

[6] K. Chang, J.-S. Jang, and C. Iliopoulos, ‘‘Music genre classification via compressive sampling,'' in *Proc. Int. Soc. Music Inform. Retr. (ISMIR)*, 2010, pp. 387–392.

[7] K. Markov and T. Matsui, ‘‘High level feature extraction for the self-taught learning algorithm,'' *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 6, Apr. 2013.

[8] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun, ‘‘Unsupervised learning of sparse features for scalable audio classification,'' in *Proc. Int. Soc. Music Inform. Retr. (ISMIR)*, 2011, pp. 681–686.

[9] E. Kim *et al.*, ‘‘Music emotion recognition: A state of the art review,'' in *Proc. 11th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, 2010, pp. 255–266.

[10] M. Barthed, G. Fazekas, and M. Sandler, ‘‘Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models,'' in *Proc. 9th Symp. Comput. Music Model. Retr. (CMMR)*, Jun. 2012, pp. 492–507.

[11] J. A. Russell, ‘‘A circumplex model of affect,'' *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[12] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, ‘‘A regression approach to music emotion recognition,'' *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[13] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, ‘‘1000 songs for emotional analysis of music,'' in *Proc. 2nd ACM Multimedia Workshop Crowdsourcing Multimedia*, Barcelona, Spain, 2013.

[14] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2006.

[15] L. Csató and M. Opper, ‘‘Sparse on-line Gaussian processes,'' *Neural Comput.*, vol. 14, no. 3, pp. 641–668, Dec. 2002.

[16] E. Snelson and Z. Ghahramani, ‘‘Sparse Gaussian processes using pseudo-inputs,'' in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT press, 2006, pp. 1257–1264.

[17] K. Markov and T. Matsui, ‘‘Music genre classification using Gaussian process models,'' in *Proc. IEEE Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2013, pp. 1–6.

[18] K. Markov, M. Iwata, and T. Matsui, ‘‘Music emotion recognition using Gaussian processes,'' in *Proc. 2nd ACM Multimedia Workshop Crowdsourcing Multimedia*, Barcelona, Spain, Oct. 2013.

[19] P. Cano *et al.*, ‘‘ISMIR 2004 audio description contest,'' Music Technol. Group, Universitat Pompeu Fabra, Barcelona, Spain, Tech. Rep. MTG-TR-2006-02, 2006.

[20] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, ‘‘A survey of audio-based music classification and annotation,'' *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[21] I. Panagakis, C. Kotropoulos, and G. Arce, ‘‘Music genre classification using locality preserving non-negative tensor factorization and sparse representations,'' in *Proc. Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, 2009, pp. 249–254.

[22] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, ‘‘Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features,'' *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670–682, Jun. 2009.

[23] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., 2009, pp. 1096–1104.

[24] T. Li and M. Ogihara, "Detecting emotion in music," in *Proc. Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, vol. 3. Oct. 2003, pp. 239–240.

[25] G. Tzanetakis, "Marsyas submissions to MIREX 2007," in *Proc. Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, 2007.

[26] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 40:1–40:30, May 2012.

[27] Y.-H. Yang and H. H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2184–2196, Sep. 2011.

[28] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proc. Int. Conf. Multimedia Inform. Retr.*, 2010, pp. 267–274.

[29] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. 10th Int. Soc. Music Inform. Retr. Conf. (ISMIR)*, Oct. 2009, pp. 621–626.

[30] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[31] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using Kalman filtering," in *Proc. 9th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2010, pp. 655–660.

[32] H. Cai and Y. Lin, "Modeling of operators' emotion and task performance in a virtual driving environment," *Int. J. Human-Comput. Stud.*, vol. 69, no. 9, pp. 571–586, Aug. 2011.

[33] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Gaussian processes for object categorization," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 169–188, Jun. 2010.

[34] Y. Saatçi, R. Turner, and C. Rasmussen, "Gaussian process change point models," in *Proc. 27th Annu. Int. Conf. Mach. Learn.*, 2010, pp. 927–934.

[35] S. Park and S. Choi, "Gaussian process regression for voice activity detection and speech enhancement," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2008, pp. 2879–2882.

[36] D. Lu and F. Sha, "Predicting likability of speakers with Gaussian processes," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2012.

[37] T. Komatsu, T. Nishino, G. W. Peters, T. Matsui, and K. Takeda, "Modeling head-related transfer functions via spatial-temporal Gaussian process," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 301–305.

[38] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *J. Mach. Learn. Res.*, vol. 6, pp. 1783–1816, Nov. 2005.

[39] M. Titsias and N. Lawrence, "Bayesian Gaussian process latent variable model," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, May 2010.

[40] N. D. Lawrence and A. Moore, "Hierarchical Gaussian process latent variable models," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 481–488.

[41] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

[42] H. Park, S. Yun, S. Park, J. Kim, and C. D. D. Yoo, "Phoneme classification using constrained variational Gaussian process dynamical system," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 2015–2023.

[43] G. E. Henter, M. R. Frean, and W. B. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4505–4508.

[44] M. P. Deisenroth, M. F. Huber, and U. D. Hanebeck, "Analytic moment-based Gaussian process filtering," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 225–232.

[45] R. D. Turner, M. P. Deisenroth, and C. E. Rasmussen, "State-space inference and learning with Gaussian processes," in *Proc. 13th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2010, pp. 868–875.

[46] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *J. Mach. Learn. Res.s*, vol. 11, pp. 3011–3015, Jan. 2010.

[47] H. Nickisch and C. Rasmussen, "Approximations for binary Gaussian Process classification," *J. Mach. Learn. Res.*, vol. 9, pp. 2035–2078, Mar. 2008.

[48] M. Kuss and C. Rasmussen, "Assessing appropriate inference for binary Gaussian process classification," *J. Mach. Learn. Res.*, vol. 6, pp. 1679–1704, Jan. 2005.

[49] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer-Verlag, 2006.

[50] J. Platt, "Probabilities for SV machines," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 61–74.

[51] G. Tzanetakis and P. Cook, "MARSYAS: A framework for audio analysis," *Org. Sound*, vol. 4, no. 3, pp. 169–175, Dec. 1999.

[52] C.-C. Chang and C.-J. Lin. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* [Online]. *2(3)*, pp. 27:1–27:27. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

KONSTANTIN MARKOV (M'05) was born in Sofia, Bulgaria. He received the (Hons.) degree from Saint Petersburg State Polytechnic University, Saint Petersburg, Russia, and then he was a Research Engineer with the Communication Industry Research Institute, Sofia, for several years. He received the M.Sc. and Ph.D. degrees in electrical engineering from the Toyohashi University of Technology, Toyohashi, Japan, in 1996 and 1999, respectively. He was a recipient of the Best Student Paper Award from the IEICE Society in 1998. In 1999, he joined the Department of Research Development, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan, and in 2000, he became an Invited Researcher with the ATR Spoken Language Translation (SLT) Research Laboratories. He became a Senior Research Scientist at the Department of Acoustics and Speech Processing, ATR SLT. In 2009, he joined the Human Interface Laboratory of the Information Systems Division at the University of Aizu, Aizuwakamatsu, Japan. He is a member of the International Speech Communication Association. His research interests include audio signal processing, Bayesian statistical modeling, machine learning, and pattern recognition.

TOMOKO MATSUI (M'91) received the Ph.D. degree from the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in 1997. From 1988 to 2002, she was with Nippon Telegraph and Telephone, Tokyo, where she was involved in speaker and speech recognition. From 1998 to 2002, she was with the Spoken Language Translation Research Laboratory, Advanced Telecommunications Research Institute International, Kyoto, Japan, as a Senior Researcher, and was involved in speech recognition. In 2001, she was an Invited Researcher with the Department of Acoustic and Speech Research, Bell Laboratories, Murray Hill, NJ, USA, where she was involved in finding effective confidence measures for verifying speech recognition results. She is currently a Professor with the Institute of Statistical Mathematics, Tokyo, where she is involved in statistical modeling for speech and speaker recognition applications. She was a recipient of the Paper Award of the Institute of Electronics, Information, and Communication Engineers of Japan in 1993.