Learning Combinations of Multiple Feature Representations for Music Emotion Prediction

Jens Madsen, Bjørn Sand Jensen and Jan Larsen
Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, Building 321
2800 Kongens Lyngby, Denmark
jenma@dtu.dk, bjje@dtu.dk, janla@dtu.dk

ABSTRACT

Music consists of several structures and patterns evolving through time which greatly influences the human decoding of higher-level cognitive aspects of music like the emotions expressed in music. For tasks, such as genre, tag and emotion recognition, these structures have often been identified and used as individual and non-temporal features and representations. In this work, we address the hypothesis whether using multiple temporal and non-temporal representations of different features is beneficial for modeling music structure with the aim to predict the emotions expressed in music. We test this hypothesis by representing temporal and non-temporal structures using generative models of multiple audio features. The representations are used in a discriminative setting via the Product Probability Kernel and the Gaussian Process model enabling Multiple Kernel Learning, finding optimized combinations of both features and temporal/ non-temporal representations. We show the increased predictive performance using the combination of different features and representations along with the great interpretive prospects of this approach.

Keywords

Music emotion prediction; expressed emotions; pairwise comparisons; multiple kernel learning; Gaussian process

1. INTRODUCTION

Music is ubiquitous and the use of music by individuals varies from introvert reflection to extrovert expression. As pointed out by studies in social [25] and music psychology [13], one of the main reasons why people listen to music is to regulate their emotional state. Whether it concerns the change, intensification or release of the emotions people experience. This insight and opportunity has led the Music Information Retrieval (MIR) community to focus on emotion, to both navigate in large music archives and as the core aspect in recommendation.

In order to integrate emotion as a viable, robust and scalable element in modern services, three research areas are of interest within MIR, namely 1) the elicitation of the emotions expressed in music, 2) the representation of the audio, and 3) the creation of a model that uses the representation and predicts the human annotations; hence, allowing us to model the emotions expressed in large scale music archives.

In this work we focus on the representation of the audio. Previously the focus on audio representation has been on 1) handcrafted audio features [28, 15], often derived from a mixture between signal processing and musicology using an agnostic approach, and 2) using (sparse) representation directly from tempo-spectral representations [26] such as spectrograms. In previous work, we showed how a times-series of frame-based features can be represented probabilistically in mathematical models and is beneficial for modeling emotions expressed in music [16]. We introduced the use of generative models as feature representation, both coding temporal and non-temporal aspects, showing an improved predictive performance. However, explaining higher-order cognitive aspects using single features and feature representations is a simplified view of the complex structures found in music. The temporal patterns found in tonal features like chroma are likely to be different from the temporal loudness patterns, as shown in [17]. Likewise, several temporal patterns could potentially be found in a single feature. Hence, there is a need to investigate the combination of both features and representations, e.g. capturing different temporal and non-temporal patterns for each feature, which is the focus of this work.

Previous work in combining multiple features for modeling high-level cognitive aspects has often been performed by stacking [3, 14] standard features in a vector space representation. Some work has been done in kernel representations for genre classification, where [5] compared different ways of learning representations of features using simple Gaussian kernels computed on each feature separately, using an Support Vector Machine (SVM) based Multiple Kernel Learning (MKL) approach to find the optimal combination of each feature. In [2] different modalities were combined, consisting of acoustic content and social context features, using MKL for semantic tag prediction. They showed that combining kernels improved the predictive performance using a kernel for social context features as well as a Probability Product Kernel (PPK) between Gaussian Mixture models (GMMs) which was used to summarize MFCC and Chroma features.

In this work, we test the hypothesis whether combining features and feature representations through a discriminative model is beneficial for capturing emotions expressed in music. We use generative models as feature representation in the form of probability densities capturing multiple temporal as well as non-temporal structures in music. Through the PPK [10] we define correlations between excerpts, but contrary to previous work in the music field, we deploy a hierarchical, non-parametric Bayesian model with Gaussian process prior in a MKL setting. We further use a weakly informative prior

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.

ASM'15, October 30, 2015, Brisbane, Australia.

ACM 978-1-4503-3750-2/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2813524.2813534.

on each kernel weight allowing us to find sparse combination of kernels each representing features and feature representations.

We test our hypothesis on a dataset with pairwise comparisons on the dimensions of valence and arousal using three different audio features (MFCC, Chroma and Loudness) and a multitude of temporal/non-temporal feature representations of each, resulting in 83 different feature/feature representation combinations.

2. FEATURE REPRESENTATION

In order to automatically extract and represent meaningful aspects of the music signal, relevant for a subsequent modeling step, we follow standard approaches in modeling of music. This implies that we first extract a number of standard audio features, $j \in \{1, \dots, J\}$, which result in a frame-based, vector space representation. With T_n such frames, we obtain a collection of T_n vectors, $\mathbf{X}_n^{(j)} = [\mathbf{x}_{n,1}^{(j)}, \dots, \mathbf{x}_{n,T_n}^{(j)}]$, where $\mathbf{x}_{n,t}^{(j)} \in \mathbb{R}^{D^{(j)}}$ is the $D^{(j)}$ dimensional feature vector of feature j at time t for track n.

The next general modeling choice concerns the representation of the T vectors on a track level — and how to capture the temporal and non-temporal aspects, which we hypothesize is important in predicting higher order cognitive aspects such as emotion. In order to provide a common and comparable representation we choose to represent all the tracks and different features as a probability density, $p(\mathbf{X}^{(j)}|\boldsymbol{\theta}_n^{(j)})$, where $\mathbf{X}^{(j)} = [\mathbf{x}_{n,1}^{(j)}, \dots, \mathbf{x}_{n,T}^{(j)}]$ and T is the length of an arbitrary feature time-series j. $\boldsymbol{\theta}_n^{(j)}$ is the parameters of a specific density which characterizes track n for feature j. This allows for a principled statistical inclusion of different features supporting multiple assumptions regarding the temporal structure.

The density based representation supports a wide variety of statistical models for multivariate data and sequences. In particular, we consider a broad selection of density models grouped by the temporal assumptions and further distinguished by whether they are based on a discrete encoding (vector quantization).

Non-Temporal: In this case the frame-based vectors are considered independent in time and we use well-known Gaussian Mixture Models [1, 12] in the continuous feature space and a basic Vector Quantization (VQ) through a multinomial density (found e.g. using K-means) in the discrete case.

Temporal: In this case the features are considered temporally dependent and we use Auto Regressive (AR) [19] models in the continuous feature space and Markov and Hidden Markov Models (HMM) in the discrete, vector quantized space.

A detailed overview of the models is given in [16, 17]. The parameters, $\boldsymbol{\theta}$, in the track representations / densities are fitted using maximum likelihood methods.

3. MODEL

With the aim to incorporate all the aforementioned representations and automatically select the relevant ones, we formulate a hierarchical statistical model, which can handle the density based representation and the particular pairwise observations.

First, we collect the inputs, i.e., the representation of features, in a set,

$$\mathcal{X} = \left\{ p\left(\mathbf{X}^{(j)}|\boldsymbol{\theta}_n^{(j,q)}\right) | n{=}1,...,N \wedge j{=}1,...,J \wedge q{=}1,...,Q_j \right\},$$

which contains the available combinations of N excerpts, J features, and Q different types of probability densities (e.g. HMM, AR, GMM).

The dataset under consideration contains pairwise comparison between two audio excerpts, $u \in \{1, \dots, N\}$ and $v \in \{1, \dots, N\}$.

$$\mathcal{Y} = \{(y_m; u_m, v_m) | m{=}1,...,M\},$$

where $y_m \in \{-1, 1\}$ indicates which of the two excerpts that had the highest valence or arousal and M denotes the number of comparisons. $y_m = -1$ means that the u_m'th excerpt is picked over the v_m'th and vice versa when $y_m = 1$.

The goal of the the model is to predict the ranking of excerpts on the dimensions of valence and arousal from the pairwise comparisons directly. In order to map from the feature representation to the observations, we select a relatively simple non-parametric hierarchical model [24]. The main assumption is that the pairwise observation, y_m, between two distinct excerpts, u and v, with individual representations defined by $\boldsymbol{\theta}_{u_m}$ and $\boldsymbol{\theta}_{v_m}$ can be modeled as the difference between two function values one for each excerpt, $f_{u_m} = f\left(p\left(\mathbf{X}|\boldsymbol{\theta}_{u_m}\right)\right)$ and $f_{v_m} = f\left(p\left(\mathbf{X}|\boldsymbol{\theta}_{v_m}\right)\right)$. The likelihood is then given as

$$p\left(y_m|\mathbf{f}_m\right) \equiv \Phi\left(y_m \times \left(f_{u_m} - f_{v_m}\right)\right) \tag{1}$$

where $\mathbf{f}_m = [f_{u_m}, f_{v_m}]^T$ and Φ is the cumulative Gaussian function this likelihood is also known as the Probit likelihood [27, 4]. The function, f, hereby defines an internal, but latent absolute reference of e.g. valence or arousal as a function of the excerpt represented by the audio features, which maps from a feature representation to a real number, i.e., $f : p\left(\mathbf{X}|\boldsymbol{\theta}\right) \rightarrow \mathbb{R}$. Given the uncertainty about the problem complexity, we choose to model this function in a non-parametric manner by considering the function values, $\mathbf{f} = [f_1, \dots, f_N]$, directly as parameters and placing a (zero-mean) Gaussian process prior on $\mathbf{f}$ defined via a covariance function [24, 4], implying that $\mathbf{f}$ are random variables.

We compactly describe the model through the following process

$$\alpha^{(j,q)}|\nu, \eta \sim \text{half student - t}\left(\nu, \eta\right) \tag{2}$$

$$\mathrm{k}\left(p\left(\mathbf{X}|\boldsymbol{\theta}\right), \cdot\right) \equiv \sum_{j=1}^{J} \sum_{q=1}^{Q_j} \alpha^{(j,q)} \mathrm{k}\left(p\left(\mathbf{X}^{(j)}|\boldsymbol{\theta}^{(j,q)}\right), \cdot\right) \tag{3}$$

$$f_{u_m}, f_{v_m}|\mathcal{X}, \mathrm{k}(\cdot,\cdot), \boldsymbol{\alpha} \sim \mathcal{GP}\left(\mathbf{0}, \mathrm{k}\left(p\left(\mathbf{X}|\boldsymbol{\theta}\right), \cdot\right)\right) \tag{4}$$

$$\pi_m|f_{u_m}, f_{v_m} \equiv \Phi\left(f_{u_m} - f_{v_m}\right) \tag{5}$$

$$y_m|\pi_m \sim Bernoulli_{\pm 1}(\pi_m) \tag{6}$$

where $\boldsymbol{\alpha} = \{\forall j, q : \alpha^{(j,q)}\}$. The feature representations are included via the sum in the covariance function, like in standard MKL [8]. $Bernoulli_{\pm 1}$ simply denotes the Bernoulli distribution returning ± 1 instead of $0/1$.

In order to be statistically consistent, the GP-view on MKL only requires that $\alpha^{(j,q)} > 0$ hence not having the restriction of sum-to-one or other dependencies between the α's as in a risk minimization setting [23]. While a Dirichlet prior on $\boldsymbol{\alpha}$ would provide some interpretive power [7] we here opt for individual, non-informative priors based on the half student-t[6], not restricting the total variance on $\mathbf{f}$ yet still promoting small values of $\boldsymbol{\alpha}$ if not found to be relevant. A particular down side of the model is that the joint inference about $\mathbf{f}$ and $\boldsymbol{\alpha}$ is not convex.

The core component in the model is the Gaussian process prior, which is fully defined by the covariance function or a weighted sum of valid covariance functions, $\mathrm{k}\left(p\left(\mathbf{x}|\boldsymbol{\theta}\right), \cdot\right)$, and the choice is very critical to the performance of the model. Since the excerpts are conveniently represented as densities over their features, the natural covariance function is the PPK [9]. The PPK forms a common

ground for comparison and integration of the different representations and is defined as,

$$\mathrm{k}\left(p\left(\mathbf{X}|\boldsymbol{\theta}\right), p\left(\mathbf{X}|\boldsymbol{\theta}'\right)\right) = \int \left(p\left(\mathbf{X}|\boldsymbol{\theta}\right) p\left(\mathbf{X}|\boldsymbol{\theta}'\right)\right)^{\rho} d\mathbf{X}, \quad (7)$$

where $\rho > 0$ is a free model parameter we fix to $\rho = 0.5$. The parameters of the density model, $\boldsymbol{\theta}$, obviously depend on the particular representation. All the densities discussed previously result in (recursive) analytical computations [9, 20]. We further normalize the covariance such that each element in the diagonal is one.

3.1 Inference & Predictions

The variables of interest in this paper are the actual parameters which enter the likelihood, $\mathbf{f}$, and the weights, α on the combinations which indicates the relative importance of the feature representation. In the present work we limit the investigation such that Bayesian inference is only conducted over the $\mathbf{f}$ parameters and rely on MAP (type-II) estimation of the parameters α. We further fix the parameters in the hyperprior (ν and η). The required posterior is now given as

$$p\left(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \boldsymbol{\alpha}, \nu, \eta\right) = \frac{p\left(\boldsymbol{\alpha}|\nu, \eta\right) p\left(\mathbf{f}|\boldsymbol{\alpha}, \mathcal{X}\right) \prod_{m=1}^{M} p\left(y_m|f_{u_m}, f_{v_m}\right)}{p\left(\mathcal{Y}|\mathcal{X}, \boldsymbol{\alpha}, \nu, \eta\right)}$$

where $p\left(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \boldsymbol{\alpha}, \nu, \eta\right)$ is analytical intractable and for fast and robust inference we use the *Laplace Approximation* as previously suggested for a similar model by [4].

Given the analytical approximation to $p\left(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \boldsymbol{\alpha}, \nu, \eta\right)$, the marginal likelihood (or evidence), $p\left(\mathcal{Y}|\mathcal{X}, \boldsymbol{\alpha}, \nu, \eta\right)$, is also available and is used [24] to find point estimates of the hyper parameters. Here, we use it specifically to find $\boldsymbol{\alpha}$ in a fast manner using standard gradient based optimization.

Prediction of the pairwise choice, y_t, between test excerpts r and s with feature representations $p\left(\mathbf{X}^{(j)}|\boldsymbol{\theta}_r^{(j,q)}\right), p\left(\mathbf{X}^{(j)}|\boldsymbol{\theta}_s^{(j,q)}\right) \in \mathcal{X}$, is done by first considering the joint predictive distribution of $\mathbf{f}_t = [f_r, f_s]$ which is given as

$$p\left(\mathbf{f}_t|\mathcal{Y}, \mathcal{X}, \boldsymbol{\alpha}, \nu, \eta\right) = \int p\left(\mathbf{f}_t|\mathbf{f}\right) p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\alpha}, \nu, \eta\right) d\mathbf{f}, \quad (8)$$

which can easily be shown to be a two-dimensional Gaussian due to the conditioning properties of a GP and the Laplace approximation of the posterior. It is given as $\mathcal{N}\left(\mathbf{f}_t|\boldsymbol{\mu}^*, \mathbf{K}^*\right)$ where $\boldsymbol{\mu}^* = \mathbf{k}_t^T \mathbf{K}^{-1} \hat{\mathbf{f}}$, with $\mathbf{K}$ being the covariance matrix of the training inputs. $\mathbf{K}^* = \mathbf{K}_t - \mathbf{k}_t^T \left(\mathbf{I} + \mathbf{W}\mathbf{K}\right) \mathbf{k}_t$, where $\mathbf{K}_t$ is the covariance matrix of the test inputs with $\hat{\mathbf{f}}$ and $\mathbf{W}$ defined by the Laplace approximation (see [11]). $\mathbf{k}_t$ is a matrix with elements $[\mathbf{k}_t]_{i,2} = \mathrm{k}\left(p\left(\mathbf{X}|\boldsymbol{\theta}_n\right), p\left(\boldsymbol{X}|\boldsymbol{\theta}_s\right)\right)$ and $[\mathbf{k}_t]_{i,1} = \mathrm{k}\left(p\left(\mathbf{X}|\boldsymbol{\theta}_n\right), p\left(\mathbf{X}|\boldsymbol{\theta}_r\right)\right)$, where n indexes a training excerpt.

An advantage of the Gaussian process model is the availability of predictive uncertainties available through the predictive distribution defined as $p\left(y_t|r, s, \mathcal{X}, \mathcal{Y}\right) = \int p\left(y_t|\mathbf{f_t}\right) p\left(\mathbf{f}_t|\mathcal{X}, \mathcal{Y}, \boldsymbol{\alpha}, \nu, \eta\right) d\mathbf{f}_t$, where $p\left(y_t|\mathbf{f_t}\right)$ is the likelihood. The binary choice can simply be determined by which of f_r or f_s that dominates, without the need to calculate the full predictive distribution.

4. DATASET & EVALUATION APPROACH

The dataset used was described in [15] which comprises of all 190 unique pairwise comparisons of 20 different 15 second excerpts, chosen from the USPOP2002[1] dataset. 13 participants (3 female, 10 male) compared on both the dimensions of valence and arousal.

4.1 Features

We represent the harmonic content of a short time window of audio by computing the spectral energy at frequencies that correspond to each of the 12 notes in a standard chromatic scale. These so-called chroma features are extracted every 250 ms using [22] resulting in a 12-dimensional time-series. We represent the loudness as a function of time by coefficients of energy in each 24 Bark band [21]. We compute the energy for each 23 ms using [18] resulting in a 24-dimensional time-series. 20 Mel-Frequency Cepstral Coefficient (MFCC)[1] are computed to capture timbre like qualities in the musical signal. For the discrete feature representation, codewords have been found using standard k-means trained on the excerpts in each cross validation fold, to reduce overfitting.

4.2 Performance Evaluation

In order to compare the performance of the proposed model, we evaluate the following three conditions

- **Single Representation:** Only one $\alpha_l \neq 0$ i.e. a single feature representation is included, which serves as a baseline (best performing single representation). l runs over all features and feature representations J and Q and a noise term.
- **Tied Representation:** The weights are tied for all feature representations, i.e. $\alpha_l = \hat{\alpha}\ \forall l = [1, L-1]$. The $\alpha_{l=L}$ on the noise term is learned independently from the tied values.
- **Multiple Representation:** The weights are learned individually with no restrictions for all features and feature representations J and Q and a noise term.

The performance of the three conditions is measured in terms of two metrics: the predictive classification error on both arousal and valence dimension and the predictive log-likelihood which accounts for the uncertainty in the predictions (and is thus considered a better realistic measure of generalization). We use the McNemar paired test to test the significance of the results of the proposed method. The $Null$ hypothesis is that two models are the same. If $p < 0.05$ then the models can be rejected as equal on a 5% significance level.

5. RESULTS

In this section we present the results of using multiple features and feature representations for modeling emotions expressed in music. Moreover we interpret the resulting weighting of each feature/feature representation in modeling both the valence and arousal dimension.

The initialization of the kernel weights $\boldsymbol{\alpha}$ is performed by an initial greedy approach using forward selection based on the marginal likelihood. This is done due to the rather high number of kernels used in the present study and using this simple procedure for initialization results in a much better performance of the GP model. The weights for the kernels not chosen in the forward selection is initialized by drawing from the prior.

The error rate is significantly lower across all participants for the valence and arousal data, when comparing the models based on Multiple and Single representations (*Multiple model* vs. *Single model*). The *Multiple models* outperforms the models based on Tied

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

[1] `http://www.pampalk.at/ma/`

Measure	Representation	Participant													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Average
Error rate	Single	0.182	0.132	0.132	0.129	0.153	0.224	0.245	0.129	0.139	0.168	0.134	0.132	0.118	0.155
	Tied	0.205	0.150	0.161	0.179	0.176	0.253	0.284	0.139	0.213	0.211	0.171	0.161	0.163	0.190
	Multiple	0.166	0.113	0.113	0.087	0.124	0.182	0.229	0.105	0.105	0.147	0.129	0.118	0.087	0.131
Negative log-likelihood	Single	174.4	171.3	164.8	158.5	170.2	201.2	207.1	169.2	174.3	170.6	168.4	146.9	165.1	172.5
	Tied	208.9	215.4	206.6	224.0	196.1	217.7	227.6	215.6	211.8	202.6	213.9	198.6	215.7	211.9
	Multiple	162.7	156.6	149.7	150.3	156.7	182.5	199.8	150.0	152.7	155.4	153.1	139.5	157.6	159.0

Table 1: Comparison of the classification error rate for the arousal dimension (average of 20-folds) and negative predictive log-likelihood (sum of 20-folds) between the Single, Tied and Multiple settings (lower is better). The McNemar test between the Multiple and both Single and Tied representations results in p$<<$0.001

Measure	Representation	Participant													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Average
Error rate	Single	0.261	0.189	0.205	0.197	0.247	0.211	0.189	0.192	0.184	0.221	0.218	0.195	0.203	0.209
	Tied	0.368	0.232	0.253	0.226	0.387	0.295	0.397	0.242	0.232	0.353	0.300	0.237	0.268	0.292
	Multiple	0.229	0.155	0.161	0.153	0.182	0.155	0.147	0.168	0.147	0.158	0.189	0.147	0.137	0.164
Negative log-likelihood	Single	225.4	200.9	218.6	208.8	221.5	230.8	214.4	180.6	193.5	208.6	228.8	209.9	221.1	212.5
	Tied	245.9	225.8	233.6	225.8	254.9	237.6	256.6	216.1	219.9	261.9	238.0	232.2	236.8	237.3
	Multiple	214.8	182.9	206.3	184.0	209.4	220.8	202.9	177.0	171.8	194.3	224.3	182.6	205.4	198.2

Table 2: Comparison of the classification error rate for the valence dimension (average of 20-folds) and negative predictive log-likelihood (sum of 20-folds) between the Single, Tied and Multiple settings (lower is better). The McNemar test between the Multiple and both Single and Tied representations results in p$<<$0.001

representations (*Tied models*) in all cases. It is evident that learning the individual kernel weights is clearly an advantage compared to the *Tied model* especially for the Valence data.

Looking at the predictive likelihood in Table 1 and 2 we see a similar pattern, both the valence and arousal data is explained better by the *Multiple model* compared to the best *Single model* across all participants.

The error rate for the *Single models* results in 0.1551 as an average across all subjects for the arousal dimension and 0.2087 for the valence dimension, which is significantly lower than previous work [16] (p-value $<<$ 0.001). The best performing single feature/feature representation combination is the Markov model with 16 codewords of Loudness features producing an error rate of 0.1704 for arousal data and for valence the HMM model of order $p = 24$ on Chroma features producing an error rate of 0.2289.

In Figure 1 and 2 the kernel weights $\boldsymbol{\alpha}$ are presented as results from the Gaussian process model trained on arousal and valence data respectively. Interpreting the α values we can clearly see the different aspects being coded for the two dimensions. For the arousal data there is a clear concentration of α around the low order Diagonal AR models (DAR) (p=1-3) for the MFCC feature. The simple Markov models and to some degree the HMM are favored in encoding Loudness both using 8, 16 and 24 codewords. The VQ representation of MFCC and Mean of Loudness are the primary non-temporal feature representation learned, these trends are seen across multiple subjects.

For the valence dimension shown in Figure 2 we see a similar pattern of DAR models being selected, here the MFCC features are picked and with slightly higher orders, directly translated to longer temporal dynamics captured. For the discrete representations on the valence data the HMMs are favored for both coding the MFCC, Loudness and Chroma features, indicating the need for more complex temporal structures.

6. DISCUSSION

In this work we proposed a probabilistic framework for modeling the emotions expressed in music not only including multiple features, but also multiple probabilistic feature representations capturing both temporal and non-temporal aspects.

We first note that the obtained *Multiple models* are significantly better than the best *Single models*. Across both emotion dimensions we see improved performance when learning each kernel weight as compared to only including a single representation. We also note that both *Single* and *Multiple models* perform better than previous reported performance. The *Tied models* proved to be less successful judged by both the predictive likelihood and error rate. This suggest that the idea of simply applying a naive summation of all kernels into standard methods such as SVMs, GPs or kGLM is not a viable approach. It really calls for actual tuning of individual weights, for learning multiple feature and feature representation combinations. This is however not viable via an exhaustive search and — as noted by us and others — this leaves MKL the only viable solution.

We explored the potential of the outlined approach for interpretation of which different both temporal and non-temporal aspects is important in representing higher-order cognitive aspects in music, showing that the method relatively robustly identifies a certain subset of the representations across test subjects. We foresee that the method through the learning of explicit temporal dynamics will become an important tool in understanding and analyzing the temporal aspects of music, which was explored in [17]. We see the expressed emotions in music as a prototypical example of higher order cognitive aspects and categorization of music and audio in general and this approach can easily be extended to other areas. Furthermore this method can be applied to any task where multiple both temporal and non-temporal aspects can be coded in time-series data.

A particular limitation of the current model is the relative simple type-II based inference of the weights, which shows great sensitivity to initialization. We foresee that future work in developing more efficient and sophisticated inference schemes will further improve the performance of the proposed probabilistic model.

7. CONCLUSION

This work presents a novel approach of combining multiple features and feature representations by using generative models as

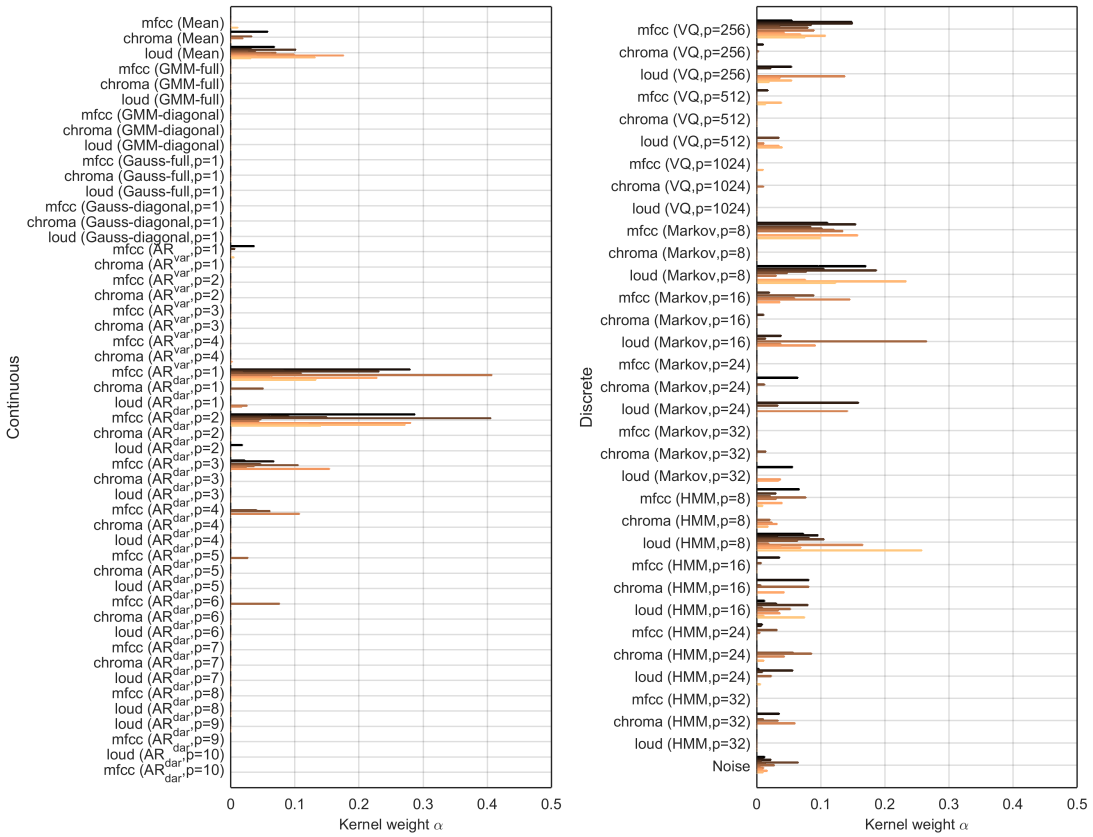

Figure 1: Arousal:The normalized kernel weights α. Each color correspond to the weights for each 13 subjects individually. p indicates the order of the model or the number of codewords in the VQ, Markov and HMM model case.

feature representation coding both temporal and non-temporal aspects of the music. Using the Product Probability Kernel to compute covariance between each feature representation, we present the Gaussian process model utilized in a Multiple Kernel Learning setting enabling us to find optimized combinations of both features and feature representations. We show greatly improved performance of the error rate and predictive likelihood, on both valence and arousal dimensions.

8. REFERENCES

[1] J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *3rd International Conference on Music Information Retrieval (ISMIR)*, pages 157–163, 2002.

[2] L. Barrington, M. Yazdani, D. Turnbull, and G. R. Lanckriet. Combining feature kernels for semantic music retrieval. In *ISMIR*, pages 614–619, 2008.

[3] M. Barthet, G. Fazekas, and M. Sandler. 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*, pages 19–22, June 2012.

[4] W. Chu and Z. Ghahramani. Preference learning with Gaussian Processes. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pages 137–144, 2005.

[5] Z. Fu, G. Lu, K.-M. Ting, and D. Zhang. On feature combination for music classification. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 453–462. Springer, 2010.

[6] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[7] M. Girolami and S. Rogers. Hierarchic bayesian models for kernel learning. *Icml 2005 - Proceedings of the 22nd International Conference on Machine Learning, Icml Proc. Int. Conf. Mach. Learn*, 119:241–248, 2005.

[8] M. Gonen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[9] T. Jebara and A. Howard. Probability Product Kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[10] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[11] B. Jensen and J. Nielsen. *Pairwise Judgements and Absolute Ratings with Gaussian Process Priors*. Technical Report, DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6151, September 2011.

[12] J. H. Jensen, D. P. W. Ellis, M. G. Christensen, and S. H. Jensen. Evaluation of distance measures between gaussian mixture models of mfccs. In *8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[13] P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

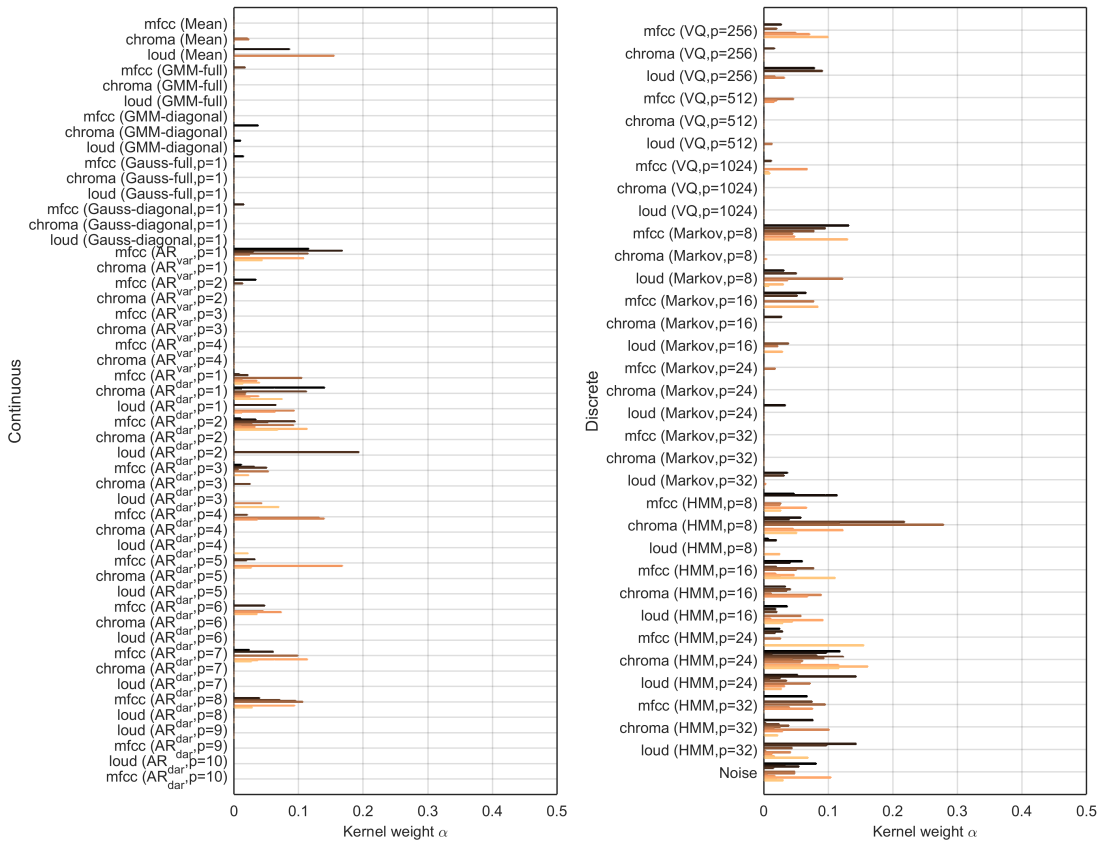

Figure 2: Valence: The normalized kernel weights α. Each color correspond to the average weights across the 20 folds for each 13 subjects individually.

[14] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[15] J. Madsen, B. S. Jensen, and J. Larsen. Predictive modeling of expressed emotions in music using pairwise comparisons. *From Sounds to Music and Emotions, Springer Berlin Heidelberg*, pages 253–277, Jan 2013.

[16] J. Madsen, B. S. Jensen, and J. Larsen. Modeling temporal structure in music for emotion prediction using pairwise comparisons. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[17] J. Madsen, B. S. Jensen, and J. Larsen. Affective modeling of music using probabilistic features representations. *In submission*, July 2015.

[18] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. an easy to use and efficient audio feature extraction software. In *11th International Conference on Music Information Retrieval (ISMIR)*, 2010.

[19] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen. Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1654–1664, 2007.

[20] A. Meng and J. Shawe-Taylor. An investigation of feature models for music genre classification using the support vector classifier. In *International Conference on Music Information Retrieval*, pages 604–609, 2005.

[21] B. C. Moore, B. R. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

[22] M. Müller and S. Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

[23] H. Nickisch and M. Seeger. Multiple kernel learning: A unifying probabilistic viewpoint. *Icml 2005 - Proceedings of the 22nd International Conference on Machine Learning, Icml Proc. Int. Conf. Mach. Learn*, 2012.

[24] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[25] P. J. Rentfrow. The role of music in everyday life: Current directions in the social psychology of music. *Social and Personality Psychology Compass*, 6(5):402–416, 2012.

[26] L. Su, C.-C. Yeh, J.-Y. Liu, J.-C. Wang, and Y.-H. Yang. A systematic evaluation of the bag-of-frames representation for music information retrieval. *Multimedia, IEEE Transactions on*, 16(5):1188–1200, Aug 2014.

[27] L. L. Thurstone. A law of comparative judgement. *Psychological Review*, 34, 1927.

[28] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):40, 2012.