View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Online Research Database In Technology

Technical University of Denmark

Modeling Temporal Structure in Music for Emotion Prediction using Pairwise Comparisons

Madsen, Jens; Jensen, Bjørn Sand; Larsen, Jan

Published in:
Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)

Publication date:
2014

Link back to DTU Orbit

Citation (APA):
Madsen, J., Jensen, B. S., & Larsen, J. (2014). Modeling Temporal Structure in Music for Emotion Prediction using Pairwise Comparisons. In Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014) (pp. 319-324). International Society for Music Information Retrieval.

DTU Library
Technical Information Center of Denmark

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

MODELING TEMPORAL STRUCTURE IN MUSIC FOR EMOTION PREDICTION USING PAIRWISE COMPARISONS

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

ABSTRACT

The temporal structure of music is essential for the cognitive processes related to the emotions expressed in music. However, such temporal information is often disregarded in typical Music Information Retrieval modeling tasks of predicting higher-level cognitive or semantic aspects of music such as emotions, genre, and similarity. This paper addresses the specific hypothesis whether temporal information is essential for predicting expressed emotions in music, as a prototypical example of a cognitive aspect of music. We propose to test this hypothesis using a novel processing pipeline: 1) Extracting audio features for each track resulting in a multivariate "feature time series". 2) Using generative models to represent these time series (acquiring a complete track representation). Specifically, we explore the Gaussian Mixture model, Vector Quantization, Autoregressive model, Markov and Hidden Markov models. 3) Utilizing the generative models in a discriminative setting by selecting the Probability Product Kernel as the natural kernel for all considered track representations. We evaluate the representations using a kernel based model specifically extended to support the robust two-alternative forced choice self-report paradigm, used for eliciting expressed emotions in music. The methods are evaluated using two data sets and show increased predictive performance using temporal information, thus supporting the overall hypothesis.

1. INTRODUCTION

The ability of music to represent and evoke emotions is an attractive and yet a very complex quality. This is partly a result of the dynamic temporal structures in music, which are a key aspect in understanding and creating predictive models of more complex cognitive aspects of music such as the emotions expressed in music. So far the approach of creating predictive models of emotions expressed in music has relied on three major aspects. First, self-reported annotations (rankings, ratings, comparisons, tags, etc.) for quantifying the emotions expressed in music. Secondly, finding a suitable audio representation (using audio or lyrical features), and finally associating the two aspects using machine learning methods with the aim to create predictive models of the annotations describing the emotions expressed in music. However the audio representation has typically relied on classic audio-feature extraction, often neglecting how this audio representation is later used in the predictive models.

We propose to extend how the audio is represented by including *feature representation* as an additional aspect, which is illustrated on Figure 1. Specifically, we focus on including the temporal aspect of music using the added feature representation [10], which is often disregarded in the classic audio-representation approaches. In Music Information Retrieval (MIR), audio streams are often represented with frame-based features, where the signal is divided into frames of samples with various lengths depending on the musical aspect which is to be analyzed. Feature extraction based on the enframed signal results in multivariate time series of feature values (often vectors). In order to use these features in a discriminative setting (i.e. predicting tags, emotion, genre, etc.), they are often represented using the mean, a single or mixtures of Gaussians (GMM). This can reduce the time series to a single vector and make the features easy to use in traditional linear models or kernel machines such as the Support Vector Machine (SVM). The major problem here is that this approach disregards all temporal information in the extracted features. The frames could be randomized and would still have the same representation, however this randomization makes no sense musically.

In modeling the emotions expressed in music, the temporal aspect of emotion has been centered on how the labels are acquired and treated, not on how the musical content is treated. E.g. in [5] they used a Conditional Random Field (CRF) model to essentially smooth the predicted labels of an SVM, thus still not providing temporal information re-

© Jens Madsen, Bjørn Sand Jensen, Jan Larsen.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Jens Madsen, Bjørn Sand Jensen, Jan Larsen. "Modeling Temporal Structure in Music for Emotion Prediction using Pairwise Comparisons", 15th International Society for Music Information Retrieval Conference, 2014.

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.

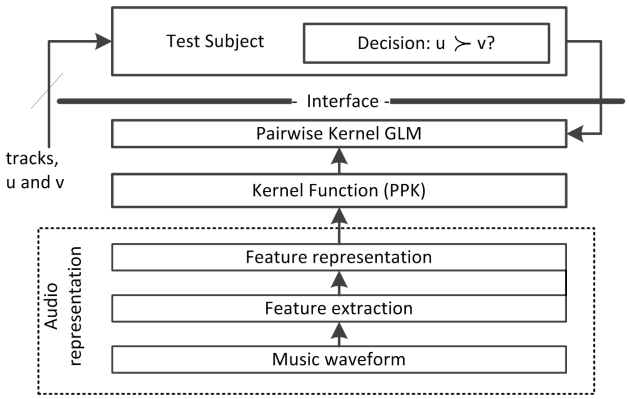

Figure 1. Modeling pipeline.

garding the features. In [12] a step to include some temporal information regarding the audio features was made, by including some first and second order Markov properties for their CRF model, however still averaging the features for one second windows. Other approaches have ranged from simple feature stacking in [13] to actually using a generative temporal model to represent features in [17]. The latter showed that using a Dynamical Texture Mixture model to represent the feature time series of MFCCs, taking temporal dynamics into account, carried a substantial amount of information about the emotional content. In the present work, in contrast to prior work, we focus on creating a common framework by using generative models to represent the multivariate feature time series for the application of modeling aspects related to the emotions expressed in music. Since very little work has been done within this field, we make a broad comparison of a multitude of generative models of time series data. We consider how the time series are modeled on two aspects: whether the observations are continuous or discrete, and whether temporal information should be taken into account or not. This results in four different combinations, which we investigate: 1) a continuous, temporal, independent representation which includes the mean, single Gaussian and GMM models; 2) a temporal, dependent, continuous representation using Autoregressive models; 3) a discretized features representation using vector quantization in a temporally independent Vector Quantization (VQ) model; and finally 4) a representation including the temporal aspect fitting Markov and Hidden Markov Models (HMM) on the discretized data. A multitude of these models have never been used in MIR as a track-based representation in this specific setting. To use these generative models in a discriminative setting, the Product Probability Kernel (PPK) is selected as the natural kernel for all the feature representations considered. We extend a kernel-generalized linear model (kGLM) model specifically for pairwise observations for use in predicting emotions expressed in music. We specifically focus on the feature representation and the modeling pipeline and therefore use simple, well-known, frequently used MFCC features. In total, eighteen different models are investigated on two datasets of pairwise comparisons evaluated on the valence and arousal dimensions.

2. FEATURE REPRESENTATION

In order to model higher order cognitive aspects of music, we first consider standard audio feature extraction which results in a frame-based, vector space representation of the music track. Given T frames, we obtain a collection of T vectors with each vector at time t denoted by $\mathbf{x}_t \in \mathbb{R}^D$, where D is the dimension of the feature space.The main concern here is how to obtain a track-level representation of the sequence of feature vectors for use in subsequent modelling steps. In the following, we will outline a number of different possibilities — and all these can be considered as probabilistic densities over either a single feature vector or a sequence of such (see also Table. 1).

Continuous: When considering the original feature space, i.e. the sequence of multivariate random variables, a vast number of representations have been proposed depending on whether the temporal aspects are ignored (i.e. considering each frame independently of all others) or modeling the temporal dynamics by temporal models.

In the time-independent case, we consider the feature as a bag-of-frames, and compute moments of the independent samples; namely the mean. Including higher order moments will naturally lead to the popular choice of representing the time-collapsed time series by a multivariate Gaussian distribution (or other continuous distributions). Generalizing this leads to mixtures of distributions such as the GMM (or another universal mixture of other distributions) used in an abundance of papers on music modeling and similarity (e.g. [1, 7]).

Instead of ignoring the temporal aspects, we can model the sequence of multivariate feature frames using well-known temporal models. The simplest models include AR models [10].

Discrete: In the discrete case, where features are naturally discrete or the original continuous feature space can be quantized using VQ with a finite set of codewords resulting in a dictionary(found e.g. using K-means). Given this dictionary each feature frame is subsequently assigned a specific codeword in a 1-of-P encoding such that a frame at time t is defined as vector $\tilde{\mathbf{x}}_t$ with one non-zero element.

At the track level and time-independent case, each frame is encoded as a Multinomial distribution with a single draw, $\tilde{\mathbf{x}} \sim \mathrm{Multinomial}(\boldsymbol{\lambda}, 1)$, where $\boldsymbol{\lambda}$ denotes the probability of occurrence for each codeword and is computed on the basis of the histogram of codewords for the entire track. In the time-dependent case, the sequence of codewords, $\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_T$, can be modeled by a relatively simple (first order) Markov model, and by introducing hidden states this may be extended to the (homogeneous) Hidden Markov model with Multinomial observations ($\mathrm{HMM}_{\mathrm{disc}}$).

2.1 Estimating the Representation

The probabilistic representations are all defined in terms of parametric densities which in all cases are estimated using standard maximum likelihood estimation (see e.g. [2]). Model selection, i.e. the number of mixture components, AR order, and number of hidden states, is performed using

Obs.	Time	Representation	Density Model	$\boldsymbol{\theta}$	Base
Continuous	Indp.	Mean	$p(\mathbf{x}\|\boldsymbol{\theta}) \equiv \delta(\boldsymbol{\mu})$	μ, σ	Gaussian
		Gaussian	$p(\mathbf{x}\|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}\|\mu, \boldsymbol{\Sigma})$	$\mu, \boldsymbol{\Sigma}$	Gaussian
		GMM	$p(\mathbf{x}\|\boldsymbol{\theta}) = \sum_{i=1}^{L} \lambda_i \mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$	$\{\lambda_i, \mu_i, \boldsymbol{\Sigma}_i\}^{i=1:L}$	Gaussian
	Temp.	AR	$p(\mathbf{x}_0, \mathbf{x}_1, .., \mathbf{x}_P\|\boldsymbol{\theta}) = \mathcal{N}\left([\mathbf{x}_0, \mathbf{x}_1, .., \mathbf{x}_P]^\top \|\mathbf{m}, \boldsymbol{\Sigma}_{\|\mathbf{A},\mathbf{C}}\right)$	$\mathbf{m}, \boldsymbol{\Sigma}_{\|\mathbf{A},\mathbf{C}}$	Gaussian
Discrete	Indp.	VQ	$p(\tilde{\mathbf{x}}\|\boldsymbol{\theta}) = \boldsymbol{\lambda}$	$\boldsymbol{\lambda}$	Multinomial
	Temp.	Markov	$p(\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1, .., \tilde{\mathbf{x}}_T\|\boldsymbol{\theta}) = \boldsymbol{\lambda}_{\tilde{\mathbf{x}}_0} \prod_{t=1}^{T} \boldsymbol{\Lambda}_{\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}}$	$\boldsymbol{\lambda}, \boldsymbol{\Lambda}$	Multinomial
		HMM_{disc}	$p(\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1, .., \tilde{\mathbf{x}}_T\|\boldsymbol{\theta}) = \sum_{\mathbf{z}_{0:T}} \boldsymbol{\lambda}_{\mathbf{z}_0} \prod_{t=1}^{T} \boldsymbol{\Lambda}_{\mathbf{z}_t, \mathbf{z}_{t-1}} \boldsymbol{\Phi}_t$	$\boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}$	Multinomial

Table 1. Continuous, features, $\mathbf{x} \in \mathbb{R}^D$, L is the number of components in the GMM, P indicates the order of the AR model, $\mathbf{A}$ and $\mathbf{C}$ are the coefficients and noise covariance in the AR model respectively and T indicates the length of the sequence. Discrete, VQ: $\tilde{\mathbf{x}} \sim \text{Multinomial}(\boldsymbol{\lambda})$, $\boldsymbol{\Lambda}_{\mathbf{z}_t, \mathbf{z}_{t-1}} = p(\mathbf{z}_t|\mathbf{z}_{t-1})$, $\boldsymbol{\Lambda}_{\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_{t-1}} = p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-1})$, $\boldsymbol{\Phi}_t = p(\tilde{\mathbf{x}}_t|\mathbf{z}_t)$. The basic Mean representation is often used in the MIR field in combination with a so-called squared exponential kernel [2], which is equivalent to formulating a PPK with a Gaussian with the given mean and a common, diagonal covariance matrix corresponding to the length scale which can be found by cross-validation and specifically using $q = 1$ in the PPK.

Bayesian Information Criterion (BIC, for GMM and HMM), or in the case of the AR model, CV was used.

2.2 Kernel Function

The various track-level representations outlined above are all described in terms of a probability density as outlined in Table 1, for which a natural kernel function is the Probability Product Kernel [6]. The PPK forms a common ground for comparison and is defined as,

$$k\left(p(\mathbf{x}|\boldsymbol{\theta}), p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right) = \int \left(p(\mathbf{x}|\boldsymbol{\theta})\, p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right)^q d\mathbf{x}, \quad (1)$$

where $q > 0$ is a free model parameter. The parameters of the density model, $\boldsymbol{\theta}$, obviously depend on the particular representation and are outlined in Tab.1. All the densities discussed previously result in (recursive) analytical computations. [6, 11]. [1]

3. PAIRWISE KERNEL GLM

The pairwise paradigm is a robust elicitation method to the more traditional direct scaling approach and is reviewed extensively in [8]. This paradigm requires a non-traditional modeling approach for which we derive a relatively simple kernel version of the Bradley-Terry-Luce model [3] for pairwise comparisons. The non-kernel version was used for this particular task in [9].

In order to formulate the model, we will for now assume a standard vector representation for each of N audio excerpts collected in the set $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., N\}$, where $\mathbf{x}_i \in \mathbb{R}^D$, denotes a standard, D dimensional audio feature vector for excerpt i. In the pairwise paradigm, any two distinct excerpts with index u and v, where $\mathbf{x}_u \in \mathcal{X}$ and $\mathbf{x}_v \in \mathcal{X}$, can be compared in terms of a given aspect (such as arousal/valance). With M such comparisons we denote the output set as $\mathcal{Y} = \{(y_m; u_m, v_m) | m = 1, ..., M\}$, where $y_m \in \{-1, +1\}$ indicates which of the two excerpts had the highest valence (or arousal). $y_m = -1$ means that the u_m'th excerpt is picked over the v_m'th and visa versa when $y_m = 1$.

The basic assumption is that the choice, y_m, between the two distinct excerpts, u and v, can be modeled as the difference between two function values, $f(\mathbf{x}_u)$ and $f(\mathbf{x}_v)$. The function $f : \mathcal{X} \to \mathbb{R}$ hereby defines an internal, but latent, absolute reference of valence (or arousal) as a function of the excerpt (represented by the audio features, $\mathbf{x}$).

Modeling such comparisons can be accomplished by the Bradley-Terry-Luce model [3, 16], here referred to more generally as the (logistic) pairwise GLM model. The choice model assumes logistically distributed noise [16] on the individual function value, and the likelihood of observing a particular choice, y_m, for a given comparison m therefore becomes

$$p(y_m|\mathbf{f}_m) \equiv \frac{1}{1 + \mathrm{e}^{-y_m \cdot z_m}}, \quad (2)$$

with $z_m = f(\mathbf{x}_{u_m}) - f(\mathbf{x}_{v_m})$ and $\mathbf{f}_m = [f(\mathbf{x}_{u_m}), f(\mathbf{x}_{v_m})]^T$. The main question is how the function, $f(\cdot)$, is modeled. In the following, we derive a kernel version of this model in the framework of kernel Generalized Linear Models (kGLM). We start by assuming a linear and parametric model of the form $\mathbf{f}_i = \mathbf{x}_i \mathbf{w}^\top$ and consider the likelihood defined in Eq. (2). The argument, z_m, is now redefined such that $z_m = \left(\mathbf{x}_{u_m}\mathbf{w}^\top - \mathbf{x}_{v_m}\mathbf{w}^\top\right)$. We assume that the model parameterized by $\mathbf{w}$ is the same for the first and second input, i.e. $\mathbf{x}_{u_m}$ and $\mathbf{x}_{v_m}$. This results in a projection from the audio features $\mathbf{x}$ into the dimensions of valence (or arousal) given by $\mathbf{w}$, which is the same for all excerpts. Plugging this into the likelihood function we obtain:

$$p(y_m|\mathbf{x}_{u_m}, \mathbf{x}_{v_m}, \mathbf{w}) = \frac{1}{1 + \mathrm{e}^{-y_m\left(\left(\mathbf{x}_{u_m} - \mathbf{x}_{u_m}\right)\mathbf{w}^\top\right)}}. \quad (3)$$

[1] It should be noted that using the PPK does not require the same length T of the sequences (the musical excerpts). For latent variable models, such as the HMM, the number of latent states in the models can also be different. The observation space, including the dimensionality D, is the only thing that has to be the same.

Following a maximum likelihood approach, the effective cost function, $\psi(\cdot)$, defined as the negative log likelihood is:

$$\psi_{GLM}(\mathbf{w}) = -\sum_{m=1}^{M} \log p\left(y_m | \mathbf{x}_{u_m}, \mathbf{x}_{v_m}, \mathbf{w}\right). \quad (4)$$

Here we assume that the likelihood factorizes over the observations, i.e. $p(\mathcal{Y}|\mathbf{f}) = \prod_{m=1}^{M} p(y_m|\mathbf{f}_m)$. Furthermore, a regularized version of the model is easily formulated as

$$\psi_{GLM-L2}(\mathbf{w}) = \psi_{GLM} + \gamma \|\mathbf{w}\|_2^2, \quad (5)$$

where the regularization parameter γ is to be found using cross-validation, for example, as adopted here. This cost is still continuous and is solved with a L-BFGS method.

This basic pairwise GLM model has previously been used to model emotion in music [9]. In this work, the pairwise GLM model is extended to a general regularized kernel formulation allowing for both linear and non-linear models. First, consider an unknown non-linear map of an element $\mathbf{x} \in \mathcal{X}$ into a Hilbert space, $\mathcal{H}$, i.e., $\varphi(\mathbf{x}) : \mathcal{X} \mapsto \mathcal{H}$. Thus, the argument z_m is now given as

$$z_m = \left(\varphi(\mathbf{x}_{u_m}) - \varphi(\mathbf{x}_{v_m})\right) \mathbf{w}^T \quad (6)$$

The *representer theorem* [14] states that the weights, $\mathbf{w}$ — despite the difference between mapped instances — can be written as a linear combination of the inputs such that

$$\mathbf{w} = \sum_{l=1}^{M} \alpha_l \left(\varphi(\mathbf{x}_{u_l}) - \varphi(\mathbf{x}_{v_l})\right). \quad (7)$$

Inserting this into Eq. (6) and applying the "kernel trick" [2], i.e. exploiting that $\langle \varphi(\mathbf{x}) \varphi(\mathbf{x}') \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}')$, we obtain

$$\begin{aligned} z_m &= \left(\varphi(\mathbf{x}_{u_m}) - \varphi(\mathbf{x}_{v_m})\right) \sum_{l=1}^{M} \alpha_l \left(\varphi(\mathbf{x}_{u_l}) - \varphi(\mathbf{x}_{v_l})\right) \\ &= \sum_{l=1}^{M} \alpha_l (\varphi(\mathbf{x}_{u_m}) \varphi(\mathbf{x}_{u_l}) - \varphi(\mathbf{x}_{u_m}) \varphi(\mathbf{x}_{v_l}) \\ &- \varphi(\mathbf{x}_{v_m}) \varphi(\mathbf{x}_{u_l}) + \varphi(\mathbf{x}_{v_m}) \varphi(\mathbf{x}_{v_l})) \\ &= \sum_{l=1}^{M} \alpha_l (k(\mathbf{x}_{u_m}, \mathbf{x}_{u_l}) - k(\mathbf{x}_{u_m}, \mathbf{x}_{v_l}) \\ &- k(\mathbf{x}_{v_m}, \mathbf{x}_{u_l}) + k(\mathbf{x}_{v_m}, \mathbf{x}_{v_l})) \\ &= \sum_{l=1}^{M} \alpha_l k(\{\mathbf{x}_{u_m}, \mathbf{x}_{v_m}\}, \{\mathbf{x}_{u_l}, \mathbf{x}_{v_l}\}). \end{aligned} \quad (8)$$

Thus, the pairwise kernel GLM formulation leads exactly to standard kernel GLM like [19], where the only difference is the kernel function which is now a (valid) kernel between two sets of pairwise comparisons [2]. If the kernel function is the linear kernel, we obtain the basic pairwise logistic regression presented in Eq. (3), but the the kernel formulation easily allows for non-vectorial inputs as provided by the PPK. The general cost function for the kGLM model is defined as,

$$\psi_{kGLM-L2}(\boldsymbol{\alpha}) = -\sum_{m=1}^{M} \log p\left(y_m | \boldsymbol{\alpha}, \mathbf{K}\right) + \gamma \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha},$$

i.e., dependent on the kernel matrix, $\mathbf{K}$, and parameters $\boldsymbol{\alpha}$. It is of the same form as for the basic model and we can apply standard optimization techniques. Predictions for unseen input pairs $\{\mathbf{x}_r, \mathbf{x}_s\}$ are easily calculated as

$$\begin{aligned} \Delta f_{rs} &= f(\mathbf{x}_r) - f(\mathbf{x}_s) \quad (9) \\ &= \sum_{m=1}^{M} \alpha_m \, k(\{\mathbf{x}_{u_m}, \mathbf{x}_{v_m}\}, \{\mathbf{x}_r, \mathbf{x}_s\}). \quad (10) \end{aligned}$$

Thus, predictions exist only as *delta* predictions. However it is easy to obtain a "true" latent (arbitrary scale) function for a single output by aggregating all the delta predictions.

4. DATASET & EVALUATION APPROACH

To evaluate the different feature representations, two datasets are used. The first dataset (*IMM*) consists of $N_{\text{IMM}} = 20$ excerpts and is described in [8]. It comprises all $M_{\text{IMM}} = 190$ unique pairwise comparisons of 20 different 15-second excerpts, chosen from the USPOP2002 [3] dataset. 13 participants (3 female, 10 male) were compared on both the dimensions of valence and arousal. The second dataset (*YANG*) [18] consists of $M_{\text{YANG}} = 7752$ pairwise comparisons made by multiple annotators on different parts of the $N_{\text{YANG}} = 1240$ different Chinese 30-second excerpts on the dimension of valence. 20 MFCC features have been extracted for all excerpts by the MA toolbox [4].

4.1 Performance Evaluation

In order to evaluate the performance of the proposed representation of the multivariate feature time series we compute learning curves. We use the so-called Leave-One-Excerpt-Out cross validation, which ensures that all comparisons with a given excerpt are left out in each fold, differing from previous work [9]. Each point on the learning curve is the result of models trained on a fraction of all available comparisons in the training set. To obtain robust learning curves, an average of 10-20 repetitions is used. Furthermore a 'win'-based baseline ($Base_{low}$) as suggested in [8] is used. This baseline represents a model with no information from features. We use the McNemar paired test with the *Null* hypothesis that two models are the same between each model and the baseline, if $p < 0.05$ then the models can be rejected as equal on a 5% significance level.

5. RESULTS

We consider the pairwise classification error on the two outlined datasets with the kGLM-L2 model, using the outlined pairwise kernel function combined with the PPK kernel (q=1/2). For the *YANG* dataset a single regularization parameter γ was estimated using 20-fold cross validation used

[2] In the Gaussian Process setting this kernel is also known as the Pairwise Judgment kernel [4], and can easily be applied for pairwise leaning using other kernel machines such as support vector machines

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

[4] http://www.pampalk.at/ma/

Obs.	Time	Models	Training set size						
			1%	5%	10%	20%	40%	80 %	100 %
Continuous	Indp.	Mean	0.468	0.386	0.347	0.310	0.277	0.260	0.252
		$\mathcal{N}(\mathbf{x}\|\mu,\sigma)$	0.464	0.394	0.358	0.328	0.297	0.279	0.274
		$\mathcal{N}(\mathbf{x}\|\mu,\mathbf{\Sigma})$	**0.440**	0.366	0.328	0.295	0.259	0.253	0.246
		GMM_{diag}	0.458	0.378	0.341	0.304	0.274	0.258	0.254
		GMM_{full}	0.441	0.362	0.329	0.297	0.269	0.255	0.252
	Temp.	DAR_{CV}	0.447	0.360	**0.316**	**0.283**	**0.251**	*0.235*	**0.228**
		VAR_{CV}	0.457	**0.354**	0.316	0.286	0.265	0.251	0.248
Discrete	Indp.	$\text{VQ}_{p=256}$	0.459	0.392	0.353	0.327	0.297	0.280	0.279*
		$\text{VQ}_{p=512}$	0.459	0.394	0.353	0.322	0.290	0.272	0.269
		$\text{VQ}_{p=1024}$	0.463	0.396	0.355	0.320	0.289	0.273	0.271
	Temp.	$\text{Markov}_{p=8}$	0.454	0.372	0.333	0.297	0.269	0.254	0.244
		$\text{Markov}_{p=16}$	0.450	0.369	0.332	0.299	0.271	0.257	0.251
		$\text{Markov}_{p=24}$	0.455	0.371	0.330	0.297	0.270	0.254	0.248
		$\text{Markov}_{p=32}$	0.458	0.378	0.338	0.306	0.278	0.263	0.256
		$\text{HMM}_{p=8}$	0.461	0.375	0.335	0.297	0.267	0.250	0.246
		$\text{HMM}_{p=16}$	0.451	0.370	0.328	*0.291*	*0.256*	**0.235**	*0.228*
		$\text{HMM}_{p=24}$	*0.441*	*0.366*	*0.328*	0.293	0.263	0.245	0.240
		$\text{HMM}_{p=32}$	0.460	0.373	0.337	0.299	0.268	0.251	0.247
		Baseline	0.485	0.413	0.396	0.354	0.319	0.290	0.285

Table 2. Classification error on the *IMM* dataset applying the pairwise kGLM-L2 model on the **valence** dimension. Results are averages of 20 folds, 13 subjects and 20 repetitions. McNemar paired tests between each model and baseline all result in $p \ll 0.001$ except for results marked with * which has $p > 0.05$ with sample size of 4940.

across all folds in the CV. The quantization of the multivariate time series, is performed using a standard online K-means algorithm [15]. Due to the inherent difficulty of estimating the number of codewords, we choose a selection specifically (8, 16, 24 and 32) for the Markov and HMM models and (256, 512 and 1024) for the VQ models. We compare results between two major categories, namely with continuous or discretized observation space and whether temporal information is included or not.

The results for the IMM dataset for valence are presented in Table 2. For continuous observations we see a clear increase in performance between the Diagonal AR (DAR) model of up to 0.018 and 0.024, compared to traditional Multivariate Gaussian and mean models respectively. With discretized observations, an improvement of performance when including temporal information is again observed of 0.025 comparing the Markov and VQ models. Increasing the complexity of the temporal representation with latent states in the HMM model, an increase of performance is again obtained of 0.016. Predicting the dimension of arousal shown on Table 3, the DAR is again the best performing model using all training data, outperforming the traditional temporal-independent models with 0.015. For discretized data the HMM is the best performing model where we again see that increasing the complexity of the temporal representation increases the predictive performance. Considering the *YANG* dataset, the results are shown in Table 4. Applying the Vector AR models (VAR), a performance gain is again observed compared to the standard representations like e.g. Gaussian or GMM. For discretized data, the temporal aspects again improve the performance, although we do not see a clear picture that increasing the complexity of the temporal representation increases the performance; the selection of the number of hidden states could be an issue here.

Obs.	Time	Models	Training set size						
			1%	5%	10%	20%	40%	80 %	100 %
Continuous	Indp.	Mean	0.368	*0.258*	*0.230*	0.215	0.202	0.190	0.190
		$\mathcal{N}(\mathbf{x}\|\mu,\sigma)$	0.378	0.267	0.241	0.221	0.205	0.190	0.185
		$\mathcal{N}(\mathbf{x}\|\mu,\mathbf{\Sigma})$	0.377	0.301	0.268	0.239	0.216	0.208	0.201
		GMM_{diag}	0.390	0.328	0.301	0.277	0.257	0.243	0.236
		GMM_{full}	*0.367*	0.303	0.279	0.249	0.226	0.216	0.215
	Temp.	DAR_{CV}	0.411	0.288	0.243	0.216	**0.197**	**0.181**	**0.170**
		VAR_{CV}	0.393	0.278	0.238	*0.213*	0.197	0.183	0.176
Discrete	Indp.	$\text{VQ}_{p=256}$	**0.351**	**0.241**	**0.221**	**0.208**	0.197	0.186	0.183
		$\text{VQ}_{p=512}$	0.356	0.253	0.226	0.211	0.199	0.190	0.189
		$\text{VQ}_{p=1024}$	0.360	0.268	0.240	0.219	0.200	0.191	0.190
	Temp.	$\text{Markov}_{p=8}$	0.375	0.265	0.238	0.220	0.205	0.194	0.188
		$\text{Markov}_{p=16}$	0.371	0.259	0.230	0.210	*0.197*	0.185	0.182
		$\text{Markov}_{p=24}$	0.373	0.275	0.249	0.230	0.213	0.202	0.200
		$\text{Markov}_{p=32}$	0.374	0.278	0.249	0.229	0.212	0.198	0.192
		$\text{HMM}_{p=8}$	0.410	0.310	0.265	0.235	0.211	0.194	0.191
		$\text{HMM}_{p=16}$	0.407	0.313	0.271	0.235	0.203	0.185	0.181
		$\text{HMM}_{p=24}$	0.369	0.258	0.233	0.215	0.197	*0.183*	*0.181*
		$\text{HMM}_{p=32}$	0.414	0.322	0.282	0.245	0.216	0.200	0.194
		Baseline	0.483	0.417	0.401	0.355	0.303	0.278	0.269

Table 3. Classification error on the *IMM* dataset applying the pairwise kGLM-L2 model on the **arousal** dimension. Results are averages of 20 folds, 13 participants and 20 repetitions. McNemar paired tests between each model and baseline all result in $p \ll 0.001$ with a sample size of 4940.

Obs.	Time	Models	Training set size						
			1%	5%	10%	20%	40%	80 %	100 %
Continuous	Indp.	Mean	0.331	0.300	0.283	0.266	0.248	0.235	0.233
		$\mathcal{N}(\mathbf{x}\|\mu,\sigma)$	0.312	0.291	0.282	0.272	0.262	0.251	0.249
		$\mathcal{N}(\mathbf{x}\|\mu,\mathbf{\Sigma})$	0.293	0.277	0.266	0.255	0.241	0.226	0.220
		GMM_{diag}	0.302	0.281	0.268	0.255	0.239	0.224	0.219
		GMM_{full}	0.293	0.276	0.263	0.249	0.233	0.218	0.214
	Temp.	$\text{DAR}_{p=10}$	0.302	0.272	0.262	0.251	0.241	0.231	0.230
		$\text{VAR}_{p=4}$	**0.281**	**0.260**	**0.249**	**0.236**	**0.223**	**0.210**	**0.206**
Discrete	Indp.	$\text{VQ}_{p=256}$	0.304	0.289	0.280	0.274	0.268	0.264	0.224
		$\text{VQ}_{p=512}$	0.303	0.286	0.276	0.269	0.261	0.254	0.253
		$\text{VQ}_{p=1024}$	*0.300*	*0.281*	0.271	0.261	0.253	0.245	0.243
	Temp.	$\text{Markov}_{p=8}$	0.322	0.297	0.285	0.273	0.258	0.243	0.238
		$\text{Markov}_{p=16}$	0.317	0.287	0.272	0.257	0.239	0.224	0.219
		$\text{Markov}_{p=24}$	0.314	0.287	*0.270*	*0.252*	*0.235*	*0.221*	*0.217*
		$\text{Markov}_{p=32}$	0.317	0.292	0.275	0.255	0.238	0.223	0.217
		$\text{HMM}_{p=8}$	0.359	0.320	0.306	0.295	0.282	0.267	0.255
		$\text{HMM}_{p=16}$	0.354	0.324	0.316	0.307	0.297	0.289	0.233
		$\text{HMM}_{p=24}$	0.344	0.308	0.290	0.273	0.254	0.236	0.234
		$\text{HMM}_{p=32}$	0.344	0.307	0.290	0.272	0.254	0.235	0.231
		Baseline	0.500	0.502	0.502	0.502	0.503	0.502	0.499

Table 4. Classification error on the *YANG* dataset applying the pairwise kGLM-L2 model on the **valence** dimension. Results are averages of 1240 folds and 10 repetitions. McNemar paired test between each model and baseline results in $p \ll 0.001$. Sample size of test was 7752.

6. DISCUSSION

In essence we are looking for a way of representing an entire track based on the simple features extracted. That is, we are trying to find generative models that can capture meaningful information coded in the features specifically for coding aspects related to the emotions expressed in music.

Results showed that simplifying the observation space using VQ is useful when predicting the arousal data. Introducing temporal coding of VQ features by simple Markov models already provides a significant performance gain, and adding latent dimensions (i.e. complexity) a further gain is obtained. This performance gain can be attributed to the temporal changes in features and potentially hidden structures in the features not coded in each frame of the features but, by their longer term temporal structures, captured by the models.

We see the same trend with the continuous observations, i.e. including temporal information significantly increases

predictive performance. These results are specific for the features used, the complexity, and potentially the model choice might differ if other features were utilized. Future work will reveal if other structures can be found in features that describe different aspects of music; structures that are relevant for describing and predicting aspects regarding emotions expressed in music.

Another consideration when using the generative models is that the entire feature time series is replaced as such by the model, since the distances between tracks are now between the models trained on each of the tracks and not directly on the features [5]. These models still have to be estimated, which takes time, but this can be done offline and provide a substantial compression of the features used.

7. CONCLUSION

In this work we presented a general approach for evaluating various track-level representations for music emotion prediction, focusing on the benefit of modeling temporal aspects of music. Specifically, we considered datasets based on robust, pairwise paradigms for which we extended a particular kernel-based model forming a common ground for comparing different track-level representations of music using the probability product kernel. A wide range of generative models for track-level representations was considered on two datasets, focusing on evaluating both using continuous and discretized observations. Modeling both the valence and arousal dimensions of expressed emotion showed a clear gain in applying temporal modeling on both the datasets included in this work. In conclusion, we have found evidence for the hypothesis that a statistically significant gain is obtained in predictive performance by representing the temporal aspect of music for emotion prediction using MFCC's.

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] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] R. D. Bock and J. V. Jones. *The measurement and prediction of judgment and choice*. Holden-day, 1968.

[4] F. Huszar. A GP classification approach to preference learning. In *NIPS Workshop on Choice Models and Preference Learning*, pages 1–4, 2011.

[5] V. Imbrasaite, T. Baltrusaitis, and P. Robinson. Emotion tracking in music using continuous conditional random fields and relative feature representation. In *ICME AAM Workshop*, 2013.

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

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

[8] 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.

[9] J. Madsen, B. S. Jensen, J. Larsen, and J. B. Nielsen. Towards predicting expressed emotion in music from pairwise comparisons. In *9th Sound and Music Computing Conference (SMC) Illusions*, July 2012.

[10] 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.

[11] 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.

[12] E. M. Schmidt and Y. E. Kim. Modeling musical emotion dynamics with conditional random fields. In *12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[13] E. M. Schmidt, J. Scott, and Y. E. Kim. Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *13th International Conference on Music Information Retrieval (ISMIR)*, 2012.

[14] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Computational Learning Theory*, 2111:416–426, 2001.

[15] D. Sculley. Web-scale k-means clustering. *International World Wide Web Conference*, pages 1177–1178, 2010.

[16] K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[17] Y. Vaizman, R. Y. Granot, and G. Lanckriet. Modeling dynamic patterns for emotional content in music. In *12th International Conference on Music Information Retrieval (ISMIR)*, pages 747–752, 2011.

[18] Y-H. Yang and H.H. Chen. Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774, May 2011.

[19] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. In *Journal of Computational and Graphical Statistics*, pages 1081–1088. MIT Press, 2001.

[5] We do note that using a single model across an entire musical track could potentially be over simplifying the representation, in our case only small 15-30-second excerpts were used and for entire tracks some segmentation would be appropriate.