

# Dynamic Music Emotion Recognition Using State-Space Models

Konstantin Markov  
Human Interface Laboratory  
The University of Aizu  
Fukushima, Japan  
markov@u-aizu.ac.jp

Tomoko Matsui  
Department of Statistical Modeling  
Institute of Statistical Mathematics  
Tokyo, Japan  
tmatsui@ism.ac.jp

## ABSTRACT

This paper describes the temporal music emotion recognition system developed at the University of Aizu for the Emotion in Music task of the MediaEval 2014 benchmark evaluation campaign. The arousal-valence trajectory prediction is cast as a time series filtering task and is modeled by a state-space models. These models include standard linear model (Kalman filter) as well as novel non-linear, non-parametric Gaussian Processes based dynamic system. The music signal was parametrized using standard features extracted with the Marsyas toolkit. Based on the preliminary results obtained from small random validation set, clear advantage of any feature or model could not be observed.

## 1. INTRODUCTION

Gaussian Processes (GPs) [4] are becoming more and more popular in the Machine Learning community for their ability to learn highly non-linear mappings between two continuous data spaces. Previously, we have successfully applied GPs for static music emotion recognition [3]. Dynamic or continuous emotion estimation is more difficult task and there are several approaches to solve it. The simplest one is to assume that for a relatively short period of time emotion is constant and apply static emotion recognition methods. A better approach is to consider emotion trajectory as a time varying process and try to track it or use time series modeling techniques. In [5], authors use Kalman filters to model emotion evolution in time for each of four data partitions. For evaluation, KL divergence between the predicted and reference A-V points distributions is measured assuming "perfect" test samples partitioning. Our approach is similar since we also use data partitioning, however, we apply model selection method. In addition, we present novel dynamic music emotion model based on GPs. The task and the database used in this evaluation are described in detail in the Emotion in Music overview paper [1].

## 2. STATE-SPACE MODELS

State-space models (SSM) are widely used in time series analysis, prediction, and modeling. They consist of latent state variable  $\mathbf{x}_t \in \mathbb{R}^e$  and observable measurement variable  $\mathbf{y}_t \in \mathbb{R}^d$  which are related as follows:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \mathbf{v}_{t-1} \quad (1)$$

$$\mathbf{y}_t = g(\mathbf{x}_t) + \mathbf{w}_t \quad (2)$$

where  $f()$  and  $g()$  are unknown functions governing temporal state dynamics and state-to-measurement mapping respectively. System and observation noises  $\mathbf{v}_t$  and  $\mathbf{w}_t$  are assumed to be independent. Probabilistically, a SSM can also be defined using two distributions:  $p(\mathbf{x}_t|\mathbf{x}_{t-1})$  and  $p(\mathbf{y}_t|\mathbf{x}_t)$ . For a sequence of  $T$  measurements, the filtering task is to approximate  $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ , while approximating  $p(\mathbf{x}_t|\mathbf{y}_{1:T})$  is the goal of the Rauch-Tung-Striebel (RTS) smoothing.

For continuous music emotion recognition,  $\mathbf{x}_t$  would represent the unknown A-V vector and  $\mathbf{y}_t$  would correspond to feature vector(s). SSM learning in our case is simplified since the state A-V labels are given for the training and  $f()$  and  $g()$  can be learned independently.

### 2.1 Kalman filter

The Kalman filter is a linear SSM where  $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$  and  $g(\mathbf{x}) = \mathbf{B}\mathbf{x}$  with  $\mathbf{A}$  and  $\mathbf{B}$  being unknown parameters, and  $\mathbf{v}$  and  $\mathbf{w}$  are zero mean Gaussian noises. Thus, both  $p(\mathbf{x}_t|\mathbf{x}_{t-1})$  and  $p(\mathbf{y}_t|\mathbf{x}_t)$  become Gaussians and simple analytic solution for the filtering and smoothing tasks can be obtained.

### 2.2 Gaussian Process dynamic system

When  $f()$  and  $g()$  are modeled by GPs, we get a Gaussian Process dynamic system. Such SSMs have been recently proposed, but lack efficient and commonly adopted algorithms for learning and inference. Availability of A-V values for training, however, makes the learning task easy since each target dimension of  $f()$  and  $g()$  can be learned independently using GP regression training algorithm. For the inference, however, there is no straightforward solution. One can always opt for Monte Carlo sampling algorithms, but they are notoriously slow. We used the solution proposed in [2]. It is based on analytic moment matching to derive Gaussian approximations to the filtering and smoothing distributions.

## 3. EXPERIMENTS

The development dataset was randomly split into training and validation sets having 600 and 144 clips each. Full cross-validation scenario was not adopted due to time constraints.

### 3.1 Feature extraction

Features were extracted from the audio signal which was first downsampled to 22050 kHz. Using the Marsyas toolkit we obtained features such as *mfcc*, *spfe* including zero-crossing rate, spectral flux, centroid, and rolloff, and spectral crest factor *scf*. All feature vectors were calculated from 512 samples frames with no overlap. First order statistics were calculated for windows of 1 sec. with 0.5 sec. overlap. Thus,

**Table 1: Kalman filter and linear RTS smoother AROUSAL results. 144 clips validation set.**

Features	KF		RTS	
	Corr.Coef.	RMSE	Corr.Coef.	RMSE
Single model				
mfcc	0.2062	0.2894	0.1070	0.3008
spfe	0.1976	0.2860	0.0998	0.3109
mfcc+spfe	0.2326	0.2378	0.0894	0.2291
mfcc+scf	0.1171	0.2288	0.1611	0.2188
baseline	0.2791	0.3631	0.1898	0.4027
Multiple models				
mfcc	0.1698	0.1384	0.0991	0.1284
spfe	0.0957	0.1874	0.0292	0.1772
mfcc+spfe	0.2022	0.1290	0.1246	0.1277
mfcc+scf	0.0059	0.1613	0.0253	0.1615
baseline	0.0212	0.2276	0.0236	0.2259

for the last 30 seconds of each clip there were 61 feature vectors. In addition to these features, we also used the features from the MediaEval2014 *baseline* system [1].

### 3.2 Data clustering

In a way similar to [5], we clustered all training clips into four clusters based on their static A-V values. Separate SSMs were trained from each cluster’s data. During the test, the trajectory obtained from the model which showed the best match, i.e. the highest likelihood, was taken as the prediction result.

## 4. RESULTS

In order to see the effect of data clustering, we also evaluated linear system trained on all 600 clips. Tables 1 and 2 show the average correlation coefficient as well as the average RMS error with respect to different features for Arousal and Valence respectively. As can be seen, clustered multiple models show lower correlation, but smaller RMSE. It is possible that the clustering has reduced the amount of training for each model resulting in less accurate parameter estimation. Table 3 shows results of the GP based system evaluation with multiple models. Single model was not used due to prohibitive memory requirements. Compared to the corresponding multiple model results of the linear system, only Valence shows some improvement.

Using the official test set consisting of 1000 clips we were able to evaluate only the Kalman filter base system due to time limitations. Results using the baseline features as well as couple of Marsyas feature sets are presented in Table 4.

## 5. CONCLUSIONS

We presented two state-space model based dynamic music emotion recognition systems - one linear and one based on Gaussian Processes. The preliminary results did not show clear advantage of any system or feature set. This is probably due to the small size of the validation set. More detailed experiments involving more data are planned for the future.

## 6. REFERENCES

- [1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at MediaEval 2014. In *MediaEval 2014 Workshop*, Barcelona, Spain, Oct 2014.

**Table 2: Kalman filter and linear RTS smoother VALENCE results. 144 clips validation set.**

Features	KF		RTS	
	Corr.Coef.	RMSE	Corr.Coef.	RMSE
Single model				
mfcc	0.0411	0.6262	0.0598	0.7082
spfe	0.0332	0.3945	0.0464	0.4710
mfcc+spfe	0.0304	0.6208	0.0725	0.6978
mfcc+scf	0.1545	0.6692	0.1401	0.7231
baseline	0.0753	0.2681	0.0779	0.2996
Multiple models				
mfcc	-0.082	0.1847	-0.042	0.1915
spfe	-0.055	0.2353	-0.060	0.2497
mfcc+spfe	-0.054	0.1866	-0.068	0.1914
mfcc+scf	0.0149	0.1688	-0.008	0.1703
baseline	-0.080	0.2425	-0.058	0.2497

**Table 3: GP filter and GP-RTS smoother results. Multiple models. 144 clips validation set.**

Features	GP-F		GP-RTS	
	Corr.Coef.	RMSE	Corr.Coef.	RMSE
AROUSAL				
mfcc	0.0436	0.3088	0.0743	0.3207
spfe	0.0582	0.3048	0.0714	0.3486
baseline	-0.0073	0.3025	0.0393	0.3444
VALENCE				
mfcc	0.0217	0.2766	0.0313	0.3083
spfe	0.0283	0.3297	-0.003	0.3515
baseline	-0.011	0.3891	-0.020	0.4431

- [2] M. Deisenroth, R. Turner, M. Huber, U. Hanebeck, and C. Rasmussen. Robust filtering and smoothing with gaussian processes. *Automatic Control, IEEE Transactions on*, 57(7):1865–1871, 2012.
- [3] K. Markov, M. Iwata, and T. Matsui. Music emotion recognition using gaussian processes. In *Proceedings of the ACM multimedia 2013 workshop on Crowdsourcing for Multimedia*, CrowdMM. ACM, 2013.
- [4] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, 2006.
- [5] E. Schmidt and Y. Kim. Prediction of time-varying musical mood distributions using kalman filtering. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 655–660, Dec 2010.

**Table 4: Kalman filter results using the 1000 clips test set.**

Features	Corr.Coef.	RMSE
AROUSAL		
mfcc+spfe	0.2735±0.4522	0.3733±0.1027
mfcc+scf	0.1622±0.5754	0.3541±0.0990
baseline	0.2063±0.5720	0.0804±0.0505
VALENCE		
mfcc+spfe	0.0469±0.4326	0.2002±0.0971
mfcc+scf	0.0265±0.4378	0.1338±0.0806
baseline	0.1665±0.5166	0.1385±0.0723