MODELING TEMPORAL STRUCTURE IN MUSIC FOR EMOTION PREDICTION USING PAIRWISE COMPARISONS

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,bjje,janla}@dtu.dk

INTRODUCTION

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).

3) Utilizing the **generative models in a discriminative setting** by selecting the Probability Product Kernel (**PPK**) 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 **(2AFC)** self-report paradigm, used for eliciting expressed emotions in music.

FEATURE REPRESENTATION

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 continuous temporal dependent representation, using **Autoregressive models**.

3) a discretized temporally independent representation, using vector quantization in a **Vector Quantization (VQ)** model.

4) a discretized temporally dependent, using **Markov** and **Hidden Markov Models (HMM).**

The kernel used for all representations is the PPK $k\left(p\left(\mathbf{x}|\boldsymbol{\theta}\right),p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right)=\int\left(p\left(\mathbf{x}|\boldsymbol{\theta}\right)p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right)^{q}d\mathbf{x}$

PAIRWISE KERNEL GLM

The pairwise paradigm is a robust elicitation method to the more traditional direct scaling approach. This paradigm requires a **non-traditional modeling approach** for which we derive a relatively simple kernel version of the **Bradley-Terry-Luce model** for pairwise comparisons. We define a set of feature vectors for the N audio excerpts $\mathcal{X}=\{\mathbf{x}_i|i=1,...,N\}$ where $\mathbf{x}_i\in\mathbb{R}^D$ denotes the D dimensional audio feature vector for excerpt i. In the pairwise case we have two excerpts u and v, where $\mathbf{x}_u\in\mathcal{X}$ and $\mathbf{x}_v\in\mathcal{X}$ We write the likelihood for a single comparison as $p\left(y_m|\mathbf{f}_m\right)\equiv\frac{1}{1+\mathrm{e}^{-y_m\cdot z_m}}$ where $y_m\in\{-1,+1\}$ indicating whether excerpt u or v was the highest on either the valence and arousal scale. Furthermore $\mathbf{f}_m=[f(\mathbf{x}_{u_m}),f(\mathbf{x}_{v_m})]^T$ and $z_m=f(\mathbf{x}_{u_m})-f(\mathbf{x}_{v_m})$ hence modeling the latent function as the difference between functional values. The likelihood for all M comparisons can then be written as

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

We use the "kernel trick" and adding L2 regularization we end up with the likelihood

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

DATASET

1. Dataset. (**IMM**): $\mathbf{N_{IMM}=20}$ excerpts. $\mathbf{M_{IMM}=190}$ unique pairwise comparisons of 20 different 15-second excerpts, chosen from the USPOP2002 dataset. 13 participants (3 female, 10 male) were compared on both the dimensions of **valence** and **arousal**. (total 4940 comparisons)

2. Dataset (**YANG**) consists of $\mathbf{M_{YANG}=7752}$ pairwise comparisons made by multiple annotators on different parts of the $\mathbf{N_{YANG}=1240}$ different Chinese 30-second excerpts on the dimension of **valence**. **20 MFCC features** have been extracted for all excerpts by the MA toolbox.

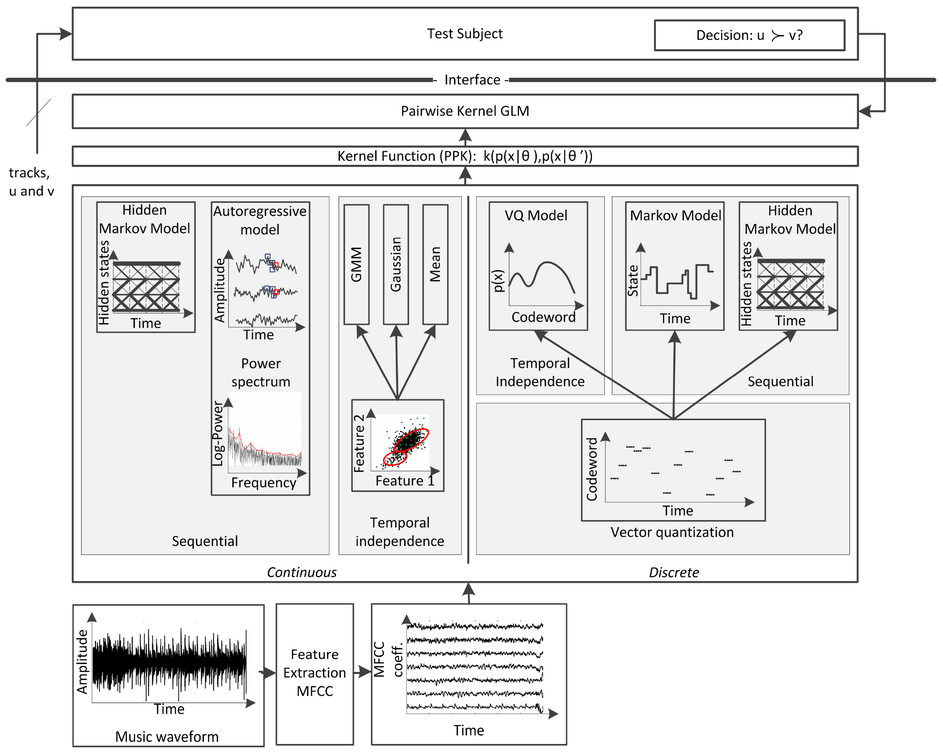

Figure 1. Overview of processing pipeline for the representation of features using generative models for the use of prediction the emotions expressed in music using pairwise comparisons.

RESULTS

For **both datasets**, both on the dimensions of **valence** and **arousal** we see **a significant increase in predictive performance when including temporal information in the feature representation**. This is both the case of **continuous** and **discretized observations**.

DISCUSSION/CONCLUSION

In essence we seek an approach to obtain a complete track representation, using generative models as feature representation we see an increase in predictive performance and provides a significant compression of features. Specifically we see that for

Discrete observations

- Simplifying the observation space using VQ is useful when predicting the arousal data.
- Introducing temporal coding of VQ features by simple Markov models provides a significant performance gain.
- Adding latent dimensions (i.e. complexity) a further gain is obtained using a HMM.

Continuous observations

- Adding temporal information to the feature representation for continuous observations provide a significant performance gain.

In conclusion **we see evidence for the hypothesis** that adding temporal information for the representation of audio features for prediction the emotions expressed in music adds statistical significant performance gain.

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.	$\mathrm{VQ}_{p=256}$	0.459	0.392	0.353	0.327	0.297	0.280	0.279*
		$\mathrm{VQ}_{p=512}$	0.459	0.394	0.353	0.322	0.290	0.272	0.269
		$\mathrm{VQ}_{p=1024}$	0.463	0.396	0.355	0.320	0.289	0.273	0.271
	Temp.	$\mathrm{Markov}_{p=8}$	0.454	0.372	0.333	0.297	0.269	0.254	0.244
		$\mathrm{Markov}_{p=16}$	0.450	0.369	0.332	0.299	0.271	0.257	0.251
		$\mathrm{Markov}_{p=24}$	0.455	0.371	0.330	0.297	0.270	0.254	0.248
		$\mathrm{Markov}_{p=32}$	0.458	0.378	0.338	0.306	0.278	0.263	0.256
		$\mathrm{HMM}_{p=8}$	0.461	0.375	0.335	0.297	0.267	0.250	0.246
		$\mathrm{HMM}_{p=16}$	0.451	0.370	0.328	*0.291*	*0.256*	**0.235**	*0.228*
		$\mathrm{HMM}_{p=24}$	*0.441*	*0.366*	*0.328*	0.293	0.263	0.245	0.240
		$\mathrm{HMM}_{p=32}$	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 1. 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 <<0.001 except for results marked with * which has p > 0:05 with sample size of 4940.

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.	$\mathrm{VQ}_{p=256}$	**0.351**	**0.241**	**0.221**	**0.208**	0.197	0.186	0.183
		$\mathrm{VQ}_{p=512}$	0.356	0.253	0.226	0.211	0.199	0.190	0.189
		$\mathrm{VQ}_{p=1024}$	0.360	0.268	0.240	0.219	0.200	0.191	0.190
	Temp.	$\mathrm{Markov}_{p=8}$	0.375	0.265	0.238	0.220	0.205	0.194	0.188
		$\mathrm{Markov}_{p=16}$	0.371	0.259	0.230	0.210	*0.197*	0.185	0.182
		$\mathrm{Markov}_{p=24}$	0.373	0.275	0.249	0.230	0.213	0.202	0.200
		$\mathrm{Markov}_{p=32}$	0.374	0.278	0.249	0.229	0.212	0.198	0.192
		$\mathrm{HMM}_{p=8}$	0.410	0.310	0.265	0.235	0.211	0.194	0.191
		$\mathrm{HMM}_{p=16}$	0.407	0.313	0.271	0.235	0.203	0.185	0.181
		$\mathrm{HMM}_{p=24}$	0.369	0.258	0.233	0.215	0.197	*0.183*	*0.181*
		$\mathrm{HMM}_{p=32}$	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 2. Classification error on the **IMM** dataset applying the pairwise kGLM-L2 model on the **arousal** dimension. Results are averages of 20 folds, 13 subjects and 20 repetitions. McNemar paired tests between each model and baseline all result in p <<0.001 with 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.	$\mathrm{DAR}_{p=10}$	0.302	0.272	0.262	0.251	0.241	0.231	0.230
		$\mathrm{VAR}_{p=4}$	**0.281**	**0.260**	**0.249**	**0.236**	**0.223**	**0.210**	**0.206**
Discrete	Indp.	$\mathrm{VQ}_{p=256}$	0.304	0.289	0.280	0.274	0.268	0.264	0.224
		$\mathrm{VQ}_{p=512}$	0.303	0.286	0.276	0.269	0.261	0.254	0.253
		$\mathrm{VQ}_{p=1024}$	*0.300*	*0.281*	0.271	0.261	0.253	0.245	0.243
	Temp.	$\mathrm{Markov}_{p=8}$	0.322	0.297	0.285	0.273	0.258	0.243	0.238
		$\mathrm{Markov}_{p=16}$	0.317	0.287	0.272	0.257	0.239	0.224	0.219
		$\mathrm{Markov}_{p=24}$	0.314	0.287	*0.270*	*0.252*	*0.235*	*0.221*	*0.217*
		$\mathrm{Markov}_{p=32}$	0.317	0.292	0.275	0.255	0.238	0.223	0.217
		$\mathrm{HMM}_{p=8}$	0.359	0.320	0.306	0.295	0.282	0.267	0.255
		$\mathrm{HMM}_{p=16}$	0.354	0.324	0.316	0.307	0.297	0.289	0.233
		$\mathrm{HMM}_{p=24}$	0.344	0.308	0.290	0.273	0.254	0.236	0.234
		$\mathrm{HMM}_{p=32}$	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 3. 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 0:001. Sample size of test was 7752.