Explaining Perceived Emotion Predictions in Music: An Attentive Approach

Sanga Chaki, Pranjal Doshi, Sourangshu Bhattacharya, Priyadarshi Patnaik
Indian Institute of Technology, Kharagpur, India

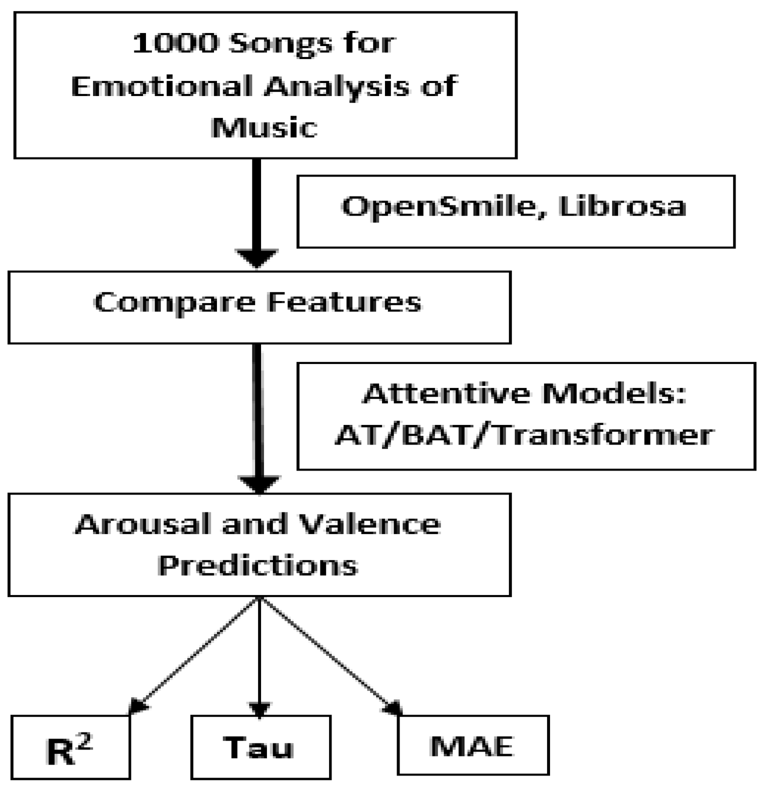
Motivation

1. Time-continuous prediction of self reported musical emotions. Few studies on design of deep learning models for this problem.
2. Challenges: a) Perceived emotion may depend on relation between music frames. b) Subjective and contextual nature of problem.

Contributions

1. Attentive LSTM based approach for emotion prediction from music clips.
2. Significant improvement of emotion prediction over vanilla LSTM.
3. Spectral features perform at par with the ComPare feature set.
4. Attention Map Analysis: Identification of music segments responsible for emotion perceived.

Attention Based Models



Process flowchart

- When listening to music, emotion at t^{th} second influenced by music context. *Attend* on those parts of input sequence, which are more relevant for t^{th} output, using *alignment* model.
- **General Mechanism**: Let model output be $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. At time t, y_t is a function of present hidden state (h_t), previous output (y_{t-1}) and unique context vector (c_t).

$$p(y_t \mid y_1, ..., y_{t-1}, \mathbf{x}) = g(h_t, y_{t-1}, c_t) \quad (1)$$

Where, $$c_t = \sum_{j=1}^{t-1} \alpha_{tj} h_j \quad (2)$$

For each output y_t, alignments between h_{t-1} and each of h_j are calculated, $1 \leq j \leq (t-2)$. $$e_{tj} = a(h_{t-1}, h_j) \quad (3)$$

Each e_{tj} used to calculate attention weights for each h_j.

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{t-1} exp(e_{tk})} \quad (4)$$

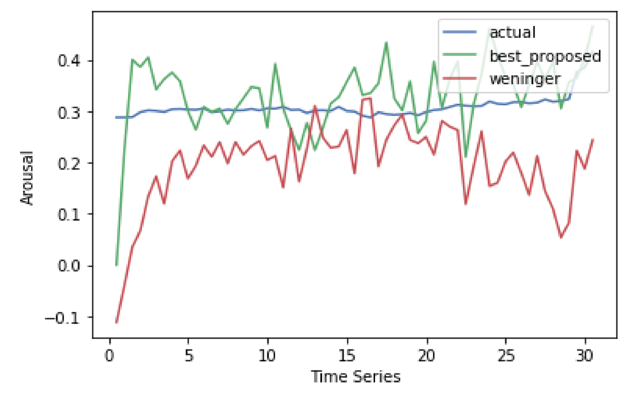
- **Types: Attention Model (AT), Backward Attention Model (AT), Transformers.**

References

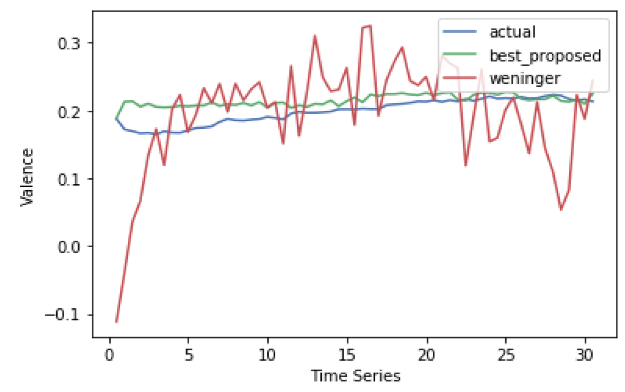
[1] F Weninger et. al. On-line continuous-time music mood regression with deep recurrent neural networks. In *ICASSP 2014*.

Exp1(a)-Model Selection

Aim:To find the best attention based model.



Arousal Comparison



Valence Comparison

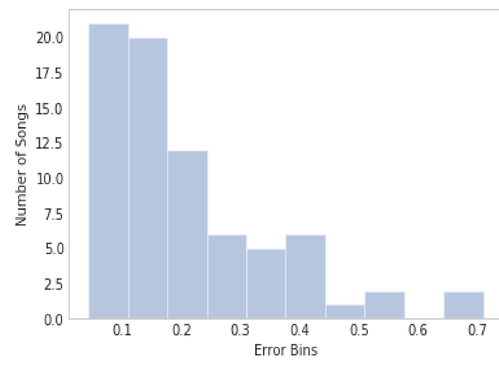
Dynamic A-V Predictions for Clip 584

Best results across different models

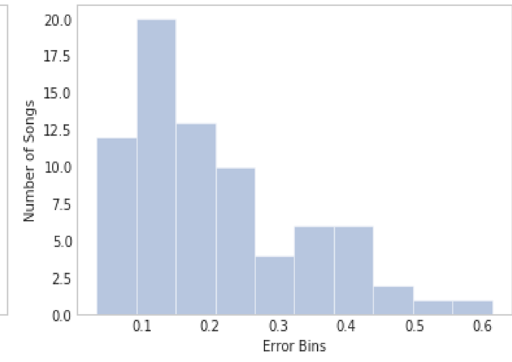
Dimension	Best Model	R^2	$\overline{\tau}$	MAE
Arousal	AT(2048_1024)	**0.78**	0.24	0.11
Valence	AT(400)	**0.53**	0.08	0.16

Exp1(b)-Error Analysis

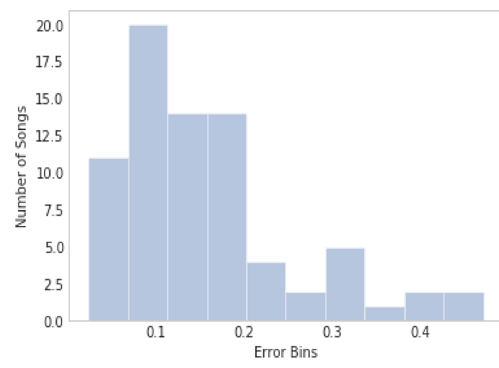
Aim: To observe patterns, biases in best models' predictions wrt to the baseline [1].



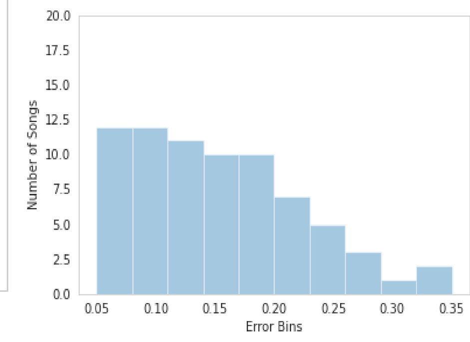
Arousal-Best



Arousal-Baseline



Valence-Best



Valence-Baseline

Error Histograms over Validation Set

Exp2 - Exploring Other Feature Sets

Aim: To explore smaller feature-sets which might produce similar/better results over same dataset.

Feature Sets for Arousal Prediction

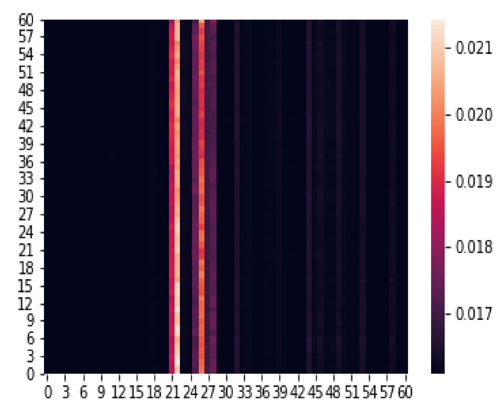
Features Used	# Features	Best Model	R^2_A	$\overline{\tau}_A$	MAE_A
Chroma(STFT+CQT)	24	AT_64	0.15	0.04	0.19
CQT on Audio clip	252	AT_64	0.45	0.06	0.17
Chroma+CQT	276	AT_64	0.57	0.07	0.14
Spectral Features	197	AT_64	**0.70**	0.03	0.12

Feature Sets for Valence Prediction

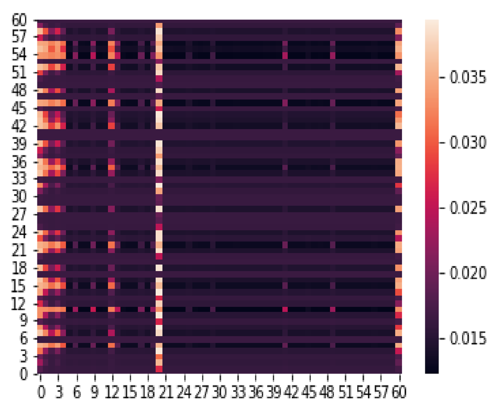
Features Used	# Features	Best Model	R^2_V	$\overline{\tau}_V$	MAE_V
Chroma(STFT+CQT)	24	AT_64	0.01	0.002	0.09
CQT on Audio clip	252	AT_64	0.07	0.01	0.17
Chroma+CQT	276	AT_64	0.17	0.06	0.14
Spectral Features	197	AT_128	**0.35**	0.07	0.16

Exp3-Attention Map Analysis

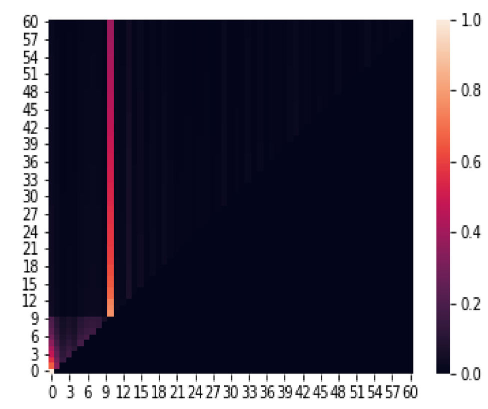
Aim: To demonstrate clip-frames which are attended to during emotion prediction using best *AT* and *BAT* models. To obtain insights into specific audio features of those frames conducive to certain emotion perception.



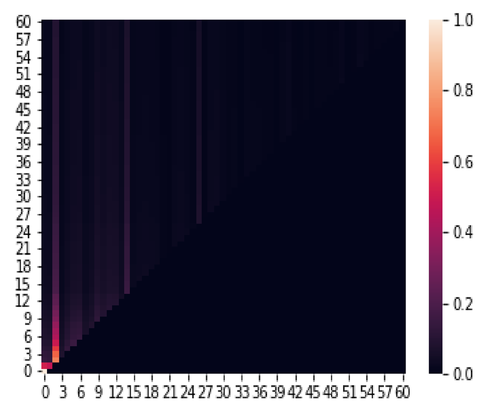
Clip206-Arousal



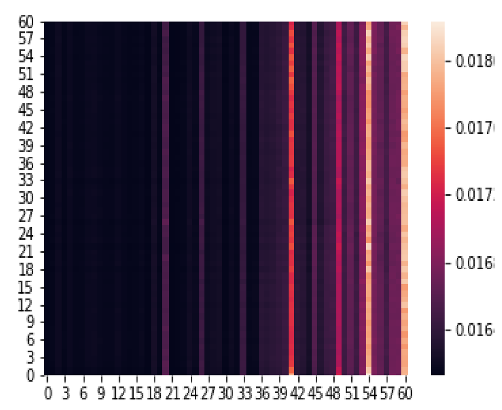
Clip206-Valence



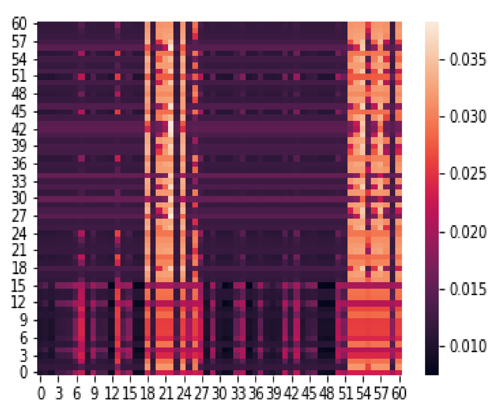
Clip60-Arousal



Clip308-Arousal

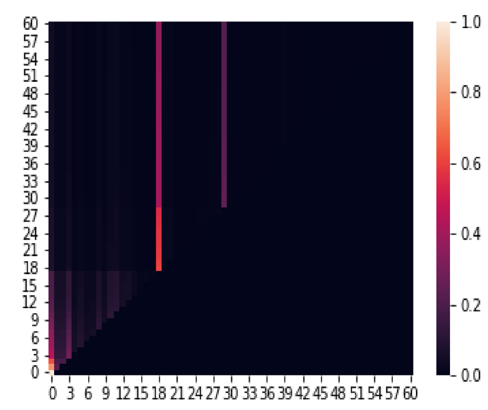


Clip978-Arousal

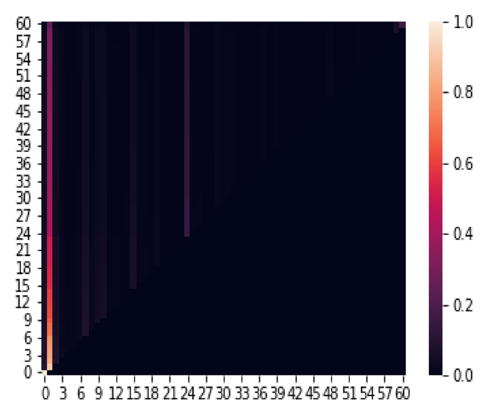


Clip978-Valence

Attention Maps using AT models.

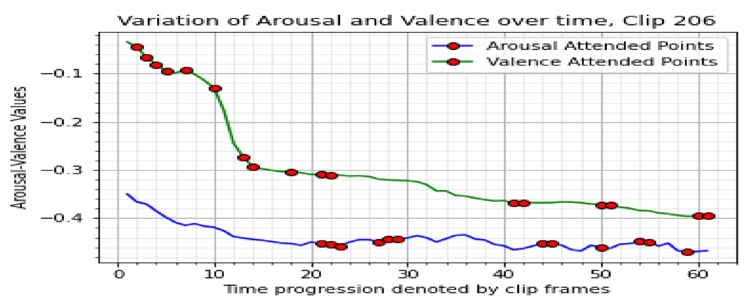


Clip60-Valence

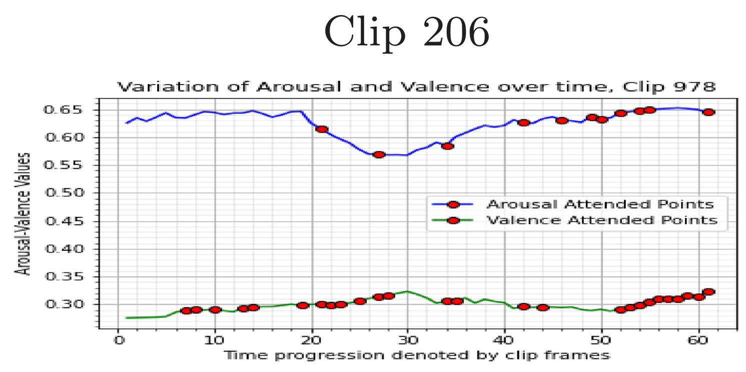


Clip308-Valence

Attention Maps using BAT models.

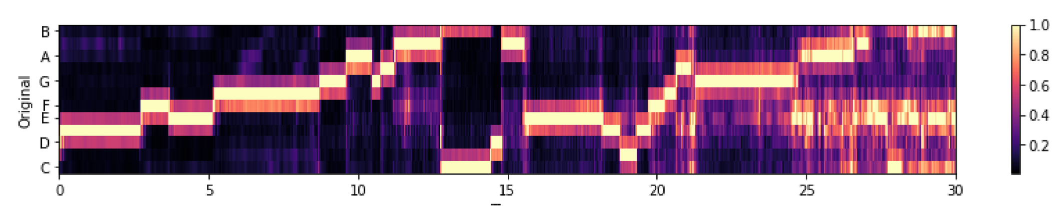


Clip 206

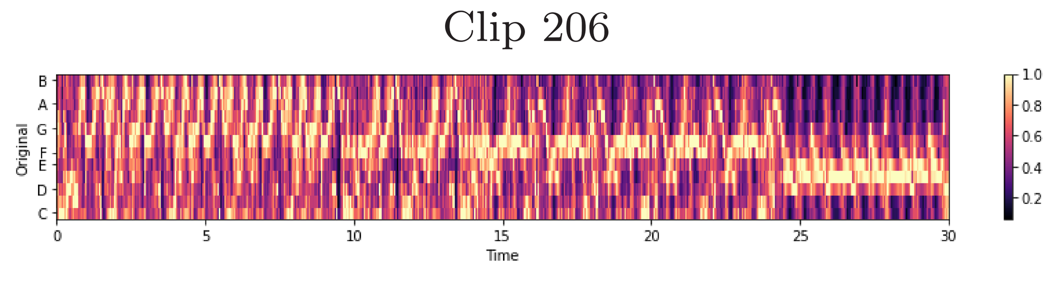


Clip 978

Attended frames vs ground truth



Clip 206



Clip 978

Chromagrams for Attention Map Analysis