Towards Explainable Emotion Recognition in Music: The Route via Mid-level Features

Shreyan Chowdhury, Andreu Vall, Verena Haunschmid, Gerhard Widmer

Institute of Computational Perception
Johannes Kepler University Linz

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

INTRODUCTION

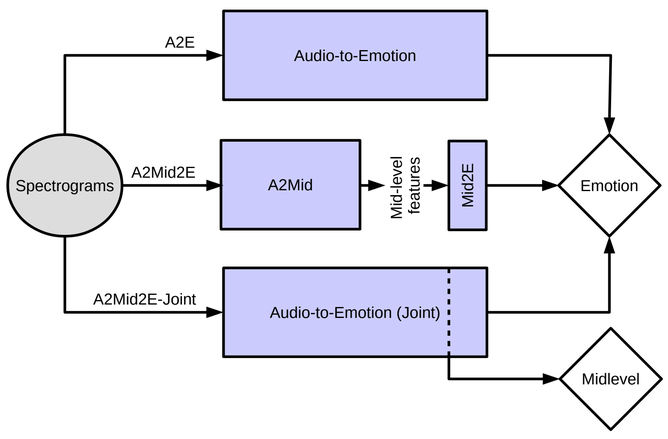
The Problem:
It is difficult to interpret emotional predictions in terms of musical content.

The Goal:
To give musically or perceptually meaningful justifications or explanations for predictions.

Data:
Audio clips rated with

- *Mid-level perceptual features*, which are musical qualities that are supposed to be meaningful and intuitively recognizable by most listeners, without requiring music theoretic knowledge
- *Emotion* ratings

ARCHITECTURE



DATASETS

Mid-level Perceptual Features Dataset

Perceptual Feature	Question asked to human raters
Melodiousness	To which excerpt do you feel like singing along?
Articulation	Which has more sounds with staccato articulation?
Rhythmic Stability	Which is easier to march along with?
Rhythmic Complexity	Difficult to repeat by tapping? Difficult to find the meter? Rhythm has many layers?
Dissonance	Noisier timbre? Has more dissonant intervals?
Tonal Stability	Easier to determine the tonic and key?
Modality (‘Minorness’)	Which song would have more minor chords?

Soundtracks Dataset (Emotion Ratings)

Valence	Energy	Tension	Anger
Fear	Happy	Sad	Tender

EXPERIMENTS

Training Schemes

- *A2E* Predict emotion values directly from spectrogram (baseline).
- *A2Mid2E* Learn a spectrogram to mid-level feature extractor, and a mid-level to emotion predictor separately.
- *A2Mid2E-Joint* Learn mid-level feature extractor and emotion predictor jointly

Song-level Explanations

- *Effects:* weights times feature values for the linear layer. Distribution over a set of examples is plotted as boxplots *(called effects plots)*.
- For a particular song, the effects of each feature contributing to a prediction can be plotted and visualized.
- Example case: two songs with similar emotion profile but different mid-level feature profile

Parameters:

Input: 313x149 Mel-spectrograms

Annotation ranges:

Mid-level: 0.1 – 1.0

Emotion: 0.1 – 0.78

Loss: Mean Squared Error

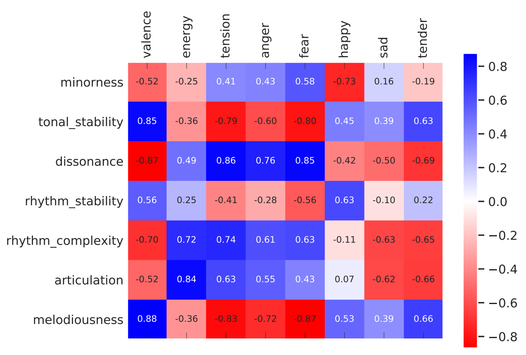
Optimizer: Adam

Evaluation metric: Pearson’s Correlation

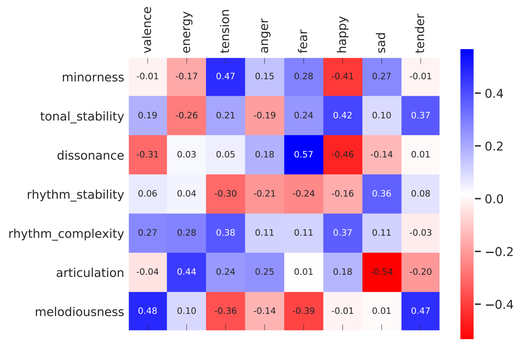
RESULTS

Model	Avg. Emotion Correlation
A2E	0.76
A2Mid2E	0.71
A2Mid2E-Joint	0.75

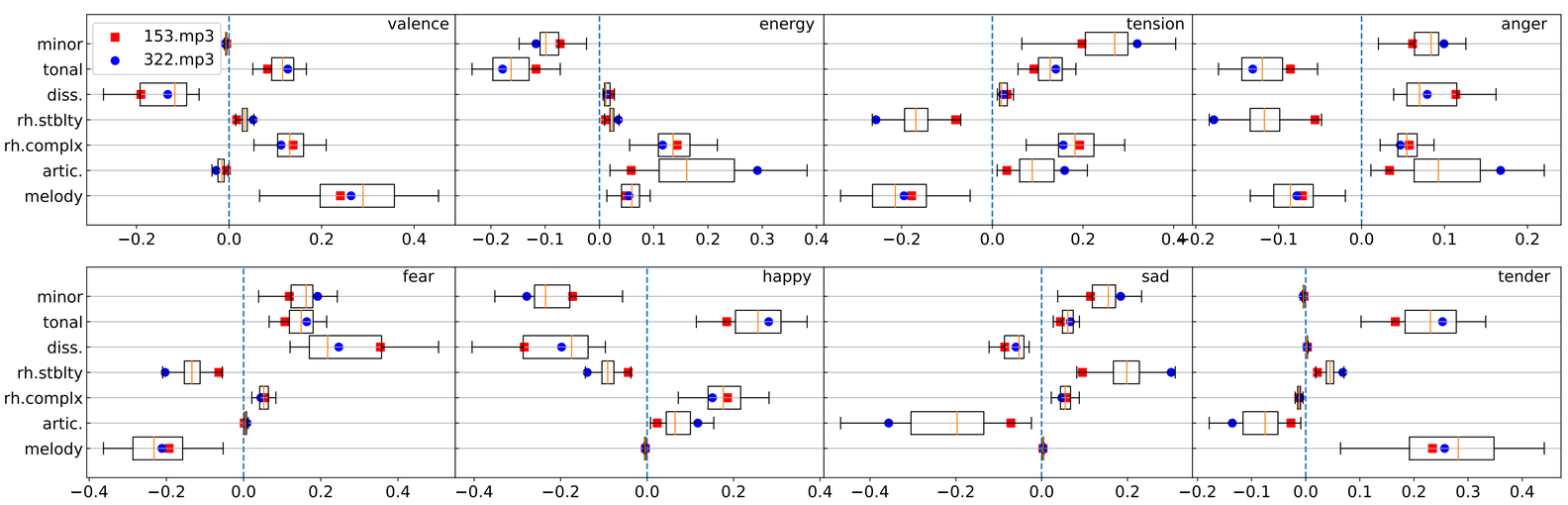
Cost Type	Cost Value
$CoE_{A2Mid2E}$	0.05
$CoE_{A2Mid2E\text{-}Joint}$	0.01



Pairwise correlation between mid-level and emotion annotations.



Weights from the linear layer of the ‘A2Mid2E-Joint’ model.



Effects Plot

European Research Council
Established by the European Commission