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

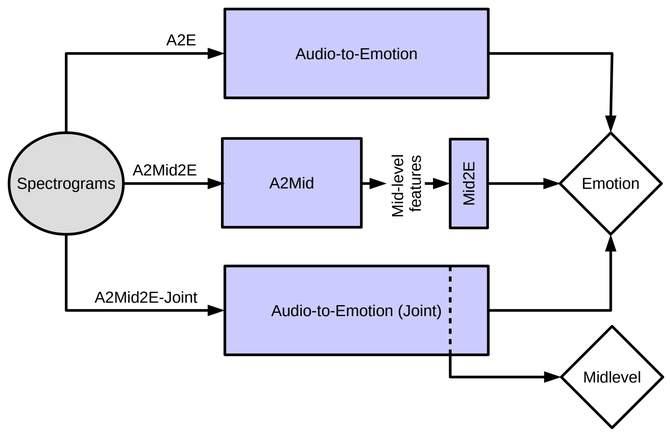
INTRODUCTION

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

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

Approach:
Train a model with an intermediate linear layer on data annotated with perceptual feature ratings and emotion ratings. The weights of the linear layer can then be used to study the effects of each perceptual feature on each final emotion prediction.

ARCHITECTURE



VGG-like network

DATASETS

Mid-level Perceptual Features Dataset: Aljanaki et al., 2018
Mid-level perceptual features are musical qualities that are supposed to be meaningful and intuitively recognizable by most listeners, without requiring music theoretic knowledge.

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): Eerola et al., 2011

Valence	**Energy**	**Tension**	**Anger**
Fear	**Happy**	**Sad**	**Tender**

EXPERIMENTS

Training Schemes:
- *A2E* (Audio-to-Emotion) Predict emotion values directly from spectrogram (baseline).
- *A2Mid2E* (Audio-to-Mid-level-to-Emotion) Learn a spectrogram to mid-level feature extractor, and a mid-level to emotion predictor separately.
- *A2Mid2E-Joint* (Audio-to-Mid-level-to-Emotion-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).

$$\text{effect}_j^{(i)} = w_j x_j^{(i)}$$

- 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

RESULTS

Evaluation metric: Pearson’s Correlation

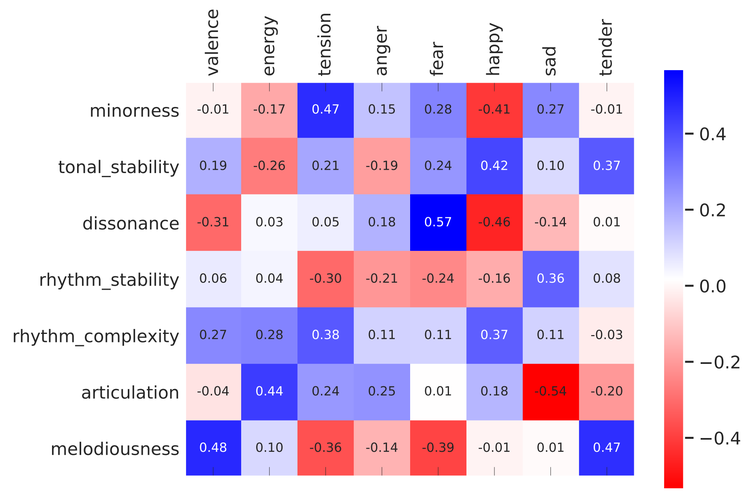
Model	Avg. Emotion Correlation
A2E	0.76
A2Mid2E	0.71
A2Mid2E-Joint	0.75

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

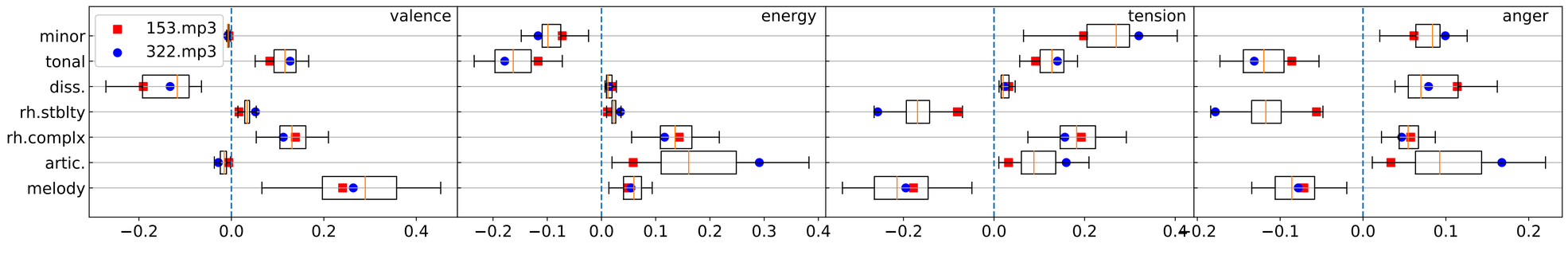
CoE = “Cost of Explainability”

Explainable predictions of emotion from music can be obtained by introducing an **intermediate representation** of mid-level perceptual features in the predictor model.

Model-level explanations visualized as linear layer weights (right), and song-level explanations visualized as effects plots (below).



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



Effects Plot for the dataset (boxes), and two songs (dots) with similar emotion profile but different mid-level features profile

Scan for paper/poster/ demo

shreyan.chowdhury@jku.at

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 670035 ("Con Espressione”)

European Research Council
Established by the European Commission