On Perceived Emotion in Expressive Piano Performance: Further Experimental Evidence for the Relevance of Mid-level Perceptual Features

Shreyan Chowdhury, Gerhard Widmer

Institute of Computational Perception
Johannes Kepler University Linz

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

INTRODUCTION

Perceived emotion of in music can be affected by several musical qualities. Music Emotion Recognition (MER) models have typically used features derived from audio content or midi/score to learn and predict emotion. Recent methods have used end-to-end audio to emotion neural networks.

Can we do better by using mid-level perceptual features [1]?

Research Questions:

Evaluate and compare feature sets on their predictive capacity for Arousal and Valence (A/V).

1. A/V fitting using each feature set
2. Importance of individual features in each set
3. Modeling variation of A/V between pieces
4. Modeling variation of A/V between different performances of the same piece

Feature Sets:

1. Low-level features
2. Score-level features
3. DEAMResNet features (end-to-end audio to emotion model)
4. Mid-level Features

Data:

- Recordings of Bach's Well-Tempered Clavier Book 1 (48 pieces) by six famous pianists: Gould, Gulda, Hewitt, Richter, Schiff, Tureck.
- A/V annotations gathered from participants in a listening experiment and averaged for each recording.
- Each recording annotated by 29 participants.
- Audio stimuli limited to first 8 measures of each piece.

References

[1] Chowdhury et al. "Towards Explainable Music Emotion Recognition: The Route via Mid-level Features" (ISMIR 2019)

[2] Koutini et al. "Emotion and theme recognition in music with Frequency-Aware RF-Regularized CNNs" (arXiv:2007.13503)

[3] Chowdhury & Widmer "Towards Explaining Expressive Qualities in Piano Recordings: Transfer of Explanatory Features via Acoustic Domain Adaptation" (ICASSP 2021)

[4] Aljanaki & Soleymani "A data-driven approach to mid-level perceptual musical feature modeling" (ISMIR 2018)

[5] Aljanaki et al. "Developing a benchmark for emotional analysis of music" (PloS one 12 (3), e0173392)

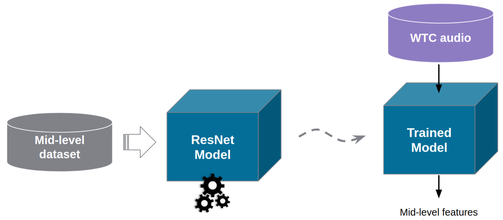
FEATURE EXTRACTION USING NEURAL NETWORK MODELS

Mid-level Features

Model: receptive-field regularised ResNet [2].

Trained using audio and annotations from Midlevel Dataset [4].

Adapted for piano music using unsupervised domain adaptation [3]



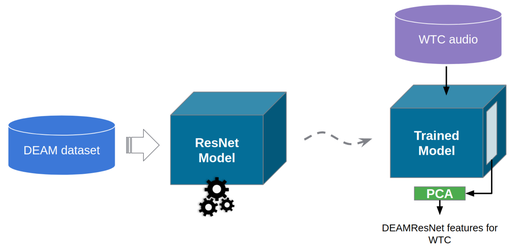
DEAMResNet Features

Features from end-to-end audio to emotion model

Model: receptive-field regularised ResNet [2].

Trained on DEAM Dataset [5].

Adapted for piano music using unsupervised domain adaptation [3]



A/V FITTING USING EACH FEATURE SET

	Piece-wise		Pianist-wise		LOO	
Feature Set	A	V	A	V	A	V
Mid-level	**0.68**	0.63	**0.68**	0.64	**0.69**	0.65
DEAMResNet	0.67	0.37	0.61	0.41	0.68	0.43
Low-level	0.54	0.20	−0.11	−0.05	0.57	0.30
Score	0.08	**0.67**	0.39	**0.75**	0.37	**0.74**

Adjusted R2 score
for different cross-validation splits. A: Arousal, V: Valence, LOO: Leave-One-Out

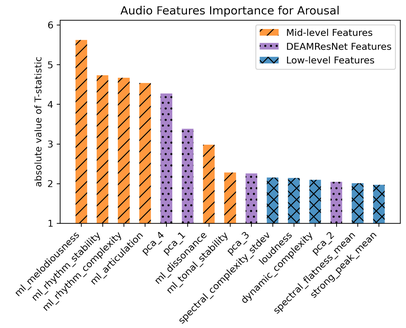
PERFORMANCE-WISE VARIATION*

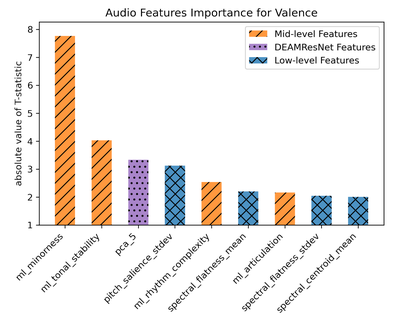
	Arousal		Valence	
Feature Set	FVU	Corr (p<0.1)	FVU	Corr (p<0.1)
Mid-level	**0.31**	**0.58** (47.9%)	**0.36**	0.42 (27.0%)
DEAMResNet	0.32	0.54 (43.8%)	0.61	**0.47** (37.5%)
Low-level	0.43	0.56 (54.2%)	0.75	0.38 (22.9%)

Evaluation metrics for performance-wise variation.
FVU: Fraction of Variance Unexplained. Corr: Pearson's correlation coefficient.

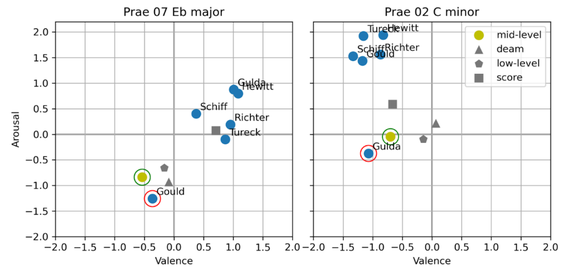
* See paper for piece-wise variation results (omitted here for brevity)

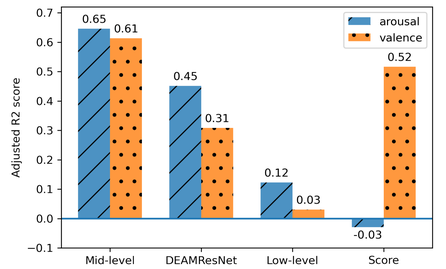
FEATURE IMPORTANCE





GENERALIZING TO OUTLIER PERFORMANCES





(Top) Emotions for two pieces with the outlier performance marked in red; Predictions using different feature sets are also plotted.
(Bottom) Adjusted R2 score for predicting emotion on a test set with 48 outlier performances

TAKEAWAYS

- Mid-level features are useful for modeling emotion in general, and differences in emotion between performances in particular.
- Mid-level features have good generalizing capacity.
- Valence is hard to model using audio content derived features. Score features generally perform better more consistently for valence.

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")

erc European Research Council Established by the European Commission