On Perceived Emotion in Expressive Piano Performance:

Further Experimental Evidence for the Relevance of Mid-level Perceptual Features

Shreyan Chowdhury, Gerhard Widmer

ISMIR 2021, 8 – 12 Nov, 2021

JOHANNES KEPLER
UNIVERSITY LINZ

Institute of
Computational
Perception

erc
European Research Council
Established by the European Commission

Dissonance

Registers

Articulation

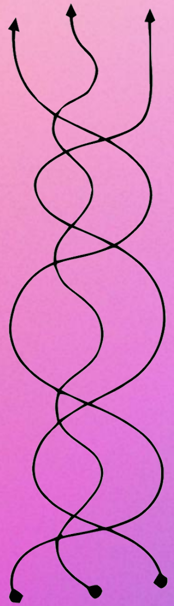
Major/minor mode

Instrumentation

Tempo

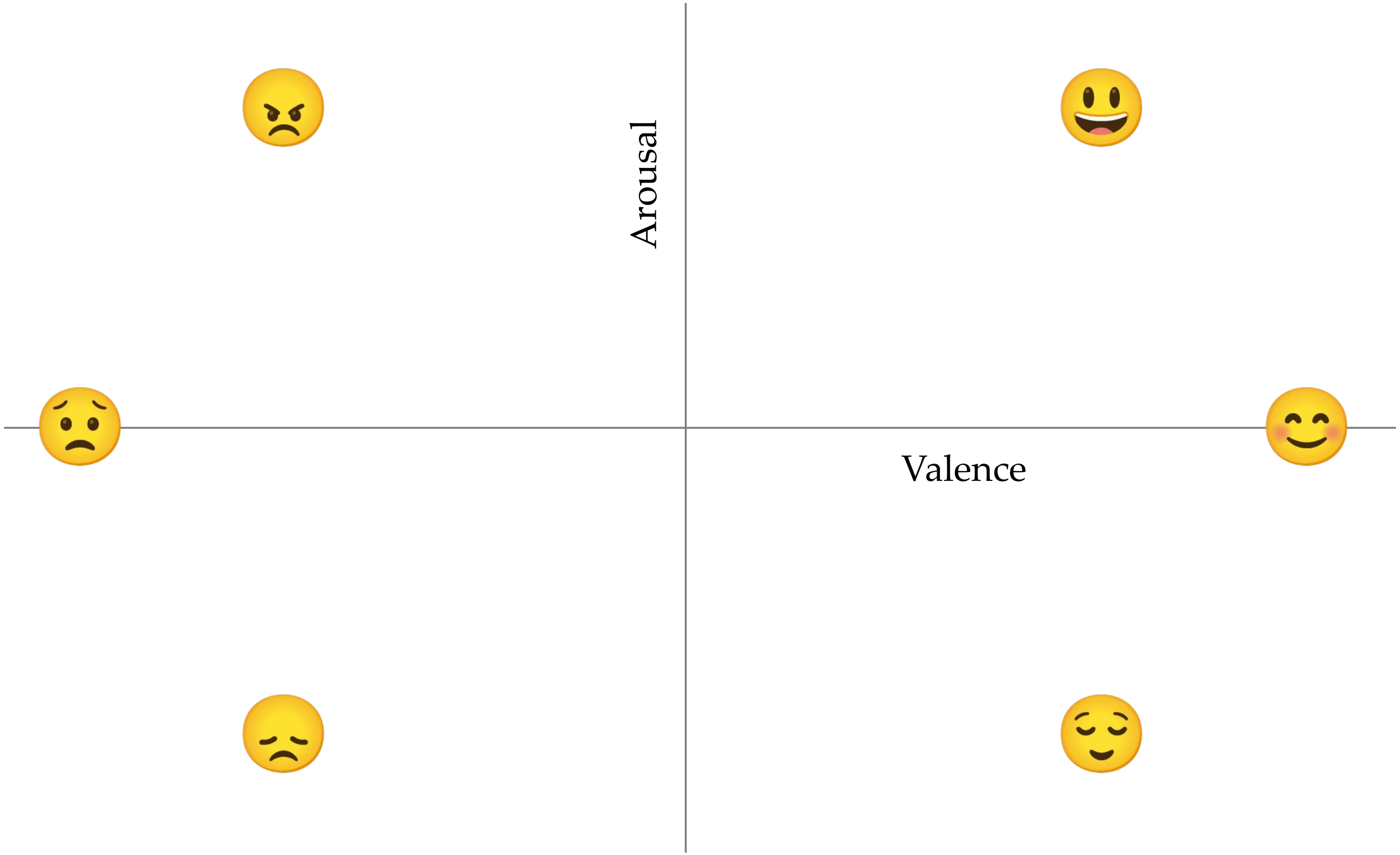
Dynamics

Intervals



Music

Emotion



Arousal

Valence

Bach Prelude No.2 in C minor

Glenn Gould

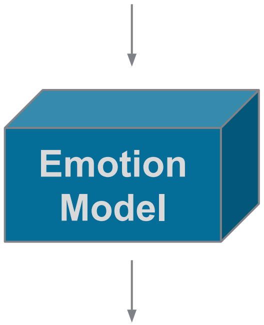
Arousal?

Valence?

Friedrich Gulda

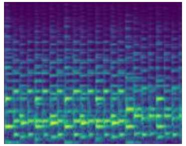
Music Emotion Recognition

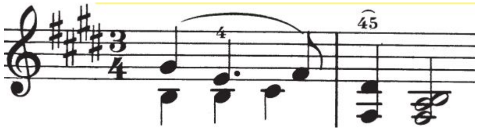
Features



Emotion Model

Emotion Predictions

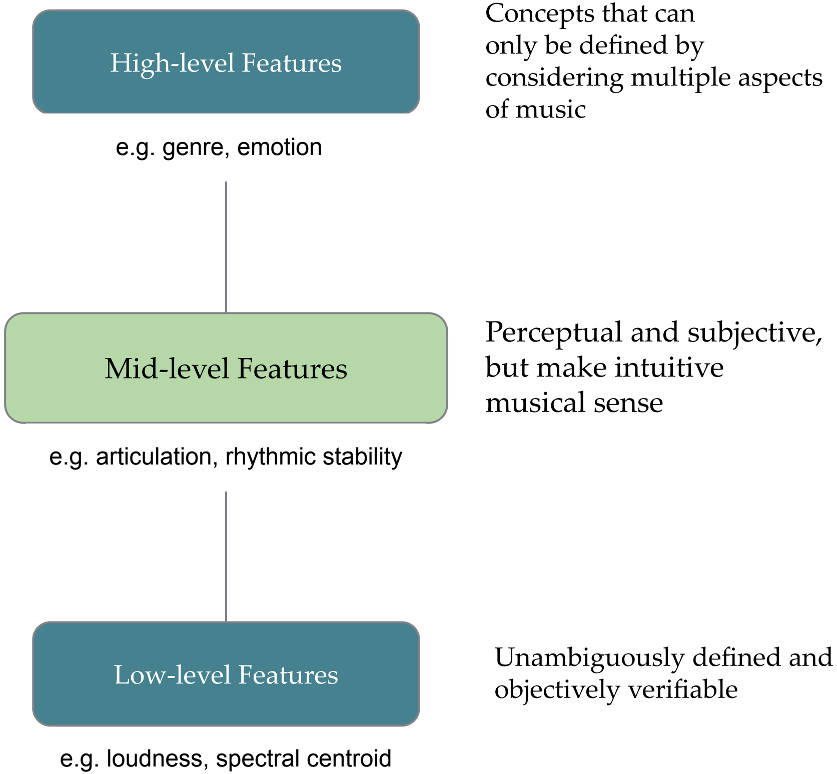




Low-level Features
- Loudness
- Onset Rate
- Spectral Centroid
- Spectral Rolloff
- ⋮

Score-based Features
- Pitch Density
- Inter Onset Interval
- Mode
- Key Strength
- ⋮

Mid-level Features



High-level Features

e.g. genre, emotion

Concepts that can only be defined by considering multiple aspects of music

Mid-level Features

e.g. articulation, rhythmic stability

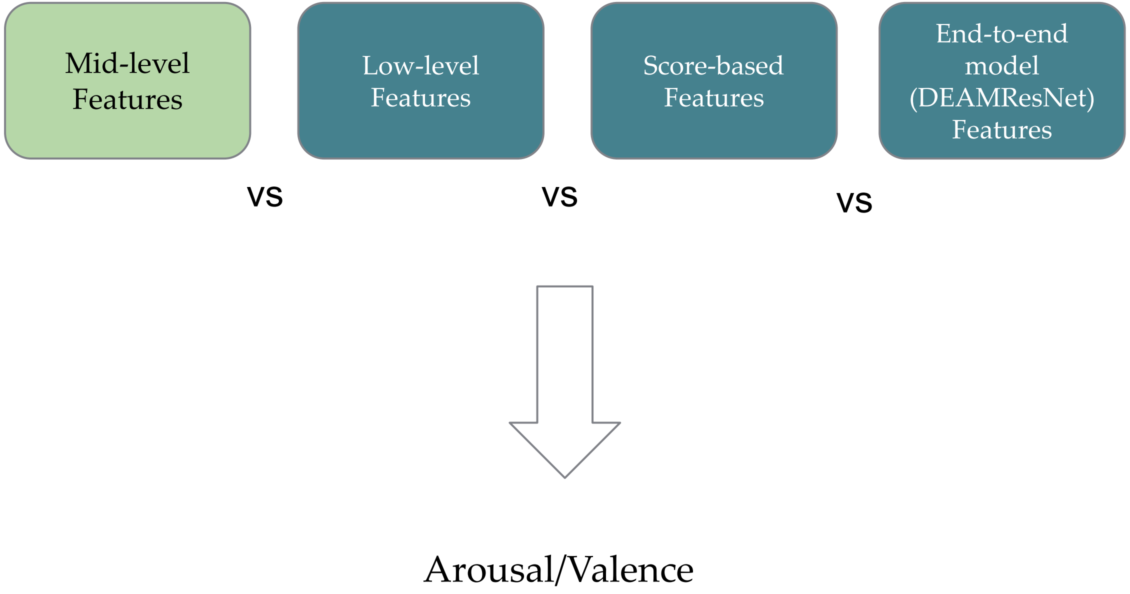
Perceptual and subjective, but make intuitive musical sense

Low-level Features

e.g. loudness, spectral centroid

Unambiguously defined and objectively verifiable

What’s in the Paper?



Mid-level Features

vs

Low-level Features

vs

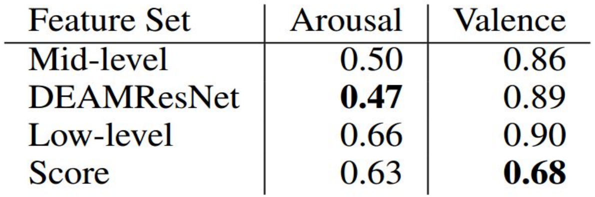
Score-based Features

vs

End-to-end model (DEAMResNet) Features

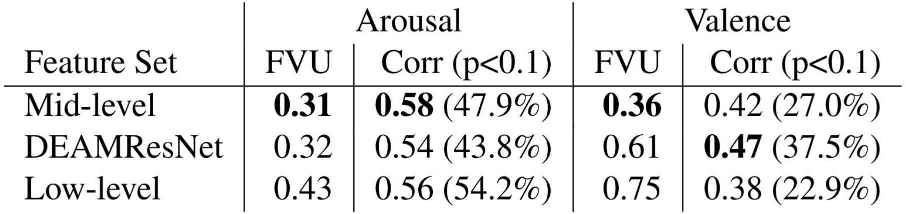
Arousal/Valence

What's in the Paper?



Feature Set	Arousal	Valence
Mid-level	0.50	0.86
DEAMResNet	**0.47**	0.89
Low-level	0.66	0.90
Score	0.63	**0.68**

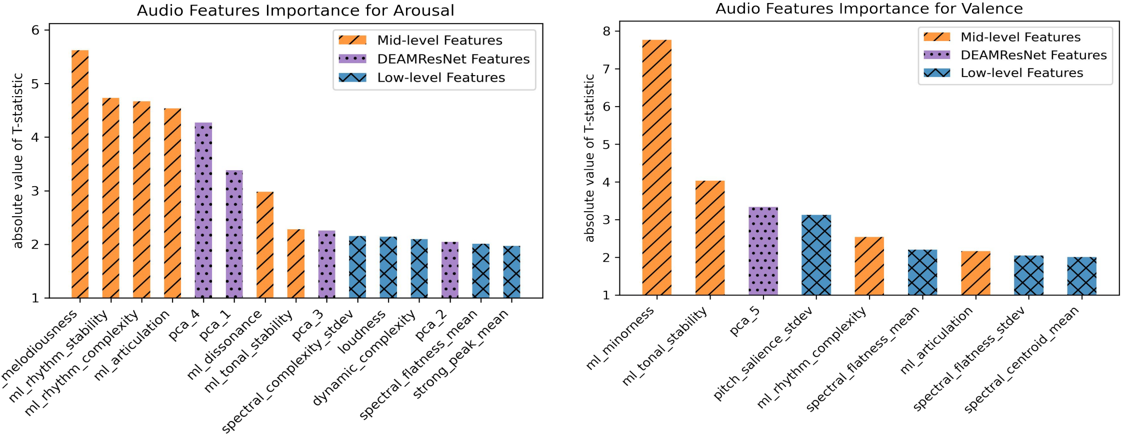
Piece-wise variation metrics (lower=better)



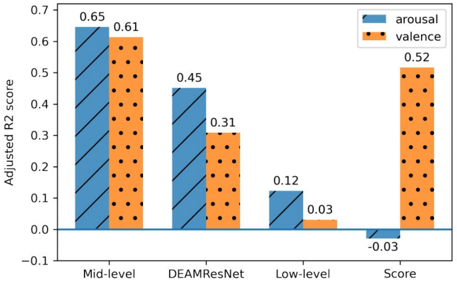
	Arousal		Valence	
Feature Set	FVU	Corr (p<0.1)	FVU	Corr (p<0.1)
Mid-level	**0.31**	**0.58** (47.9%)	**0.36**	0.42 (27.0%)
DEAMResNet	0.32	0.54 (43.8%)	0.61	**0.47** (37.5%)
Low-level	0.43	0.56 (54.2%)	0.75	0.38 (22.9%)

Performance-wise variation metrics (lower=better). Score features are the same across performances of the same piece, and are thus excluded.

What’s in the Paper?



Feature importance for modeling arousal and valence



Predictive power on outlier performances

Thank you!

https://archives.ismir.net/ismir2021/paper/000015.pdf

ON PERCEIVED EMOTION IN EXPRESSIVE PIANO PERFORMANCE: FURTHER EXPERIMENTAL EVIDENCE FOR THE RELEVANCE OF MID-LEVEL PERCEPTUAL FEATURES

Shreyan Chowdhury[1] **Gerhard Widmer**[1,2]
[1]Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2]LIT AI Lab, Linz Institute of Technology, Austria
firstname.lastname@jku.at

ABSTRACT

Despite recent advances in audio content-based music emotion recognition, a question that remains to be explored is whether an algorithm can reliably discern emotional or expressive qualities between different performances of the same piece. In the present work, we analyze several sets of features on their effectiveness in predicting arousal and valence of six different performances (by six famous pianists) of Bach's Well-Tempered Clavier Book 1. These features include low-level acoustic features, score-based features, features extracted using a pre-trained emotion model, and Mid-level perceptual features. We compare their predictive power by evaluating them on several experiments designed to test performance-wise or piece-wise variations of emotion. We find that Mid-level features show significant contribution in performance-wise variation of both arousal and valence – even better than the pre-trained emotion model. Our findings add to the evidence of Mid-level perceptual features being an important representation of musical attributes for several tasks – specifically, in this case, for capturing the expressive aspects of music that manifest as perceived emotion of a musical performance.

1. INTRODUCTION

A musical performance, particularly in the Western music tradition, is not merely a literal acoustic rendering of a notated piece or composition. Rather, the piece is transformed by the performer's own expressive performance choices, relating to such dimensions as the choice of tempo, expressive tempo and timing variations, dynamics, articulation, and so on. The emotional effect of a performance on a listener can be a consequence both of the composition itself, with its musical properties and structures, and of the performance, the way the piece was played. In fact, it has been convincingly demonstrated [1, 2] that performers are capable of communicating, with high accuracy, intended emotional qualities by their playing.

The analysis of emotion in music recordings has a long history in Music Information Retrieval, with many works addressing content-based emotion regression and classification typically using low-level or hand-crafted audio and musical features [3–6] or using deep learning based methods [7–9]. However, there has been little research on the more subtle problem of identifying emotional aspects that are due to the actual *performance*, and even less on models that can automatically recognize this from *audio* recordings. On the latter problem – the one to be addressed in this paper – the most directly relevant prior work we are aware of is [10], where 324 6-second audio snippets of different genres (classical, jazz, blues, metal, etc.) were annotated in terms of perceived emotion (valence and arousal), and various regressors were trained to predict these two dimensions from a set of standard audio features. The regression models were then used to predict valence-arousal trajectories over 5 different recordings of 4 Chopin pieces, but no ground truth in terms of human emotion annotations was collected. The relevance of the model predictions was evaluated only indirectly, by comparing similarity scores between predicted profiles with overall performance similarity ratings by three human listeners, which showed some non-negligible correlations.

In a recent focused study [11], Battcock & Schutz (referred to as "B&S" henceforth) investigate how three specific score-based cues (Mode, Pitch Height, and Attack Rate [1]) work together to convey emotion in J.S. Bach's preludes and fugues collected in his *Well-tempered Clavier (WTC)*. They used recordings of the complete WTC Book 1 (48 pieces) of one famous pianist (Friedrich Gulda) as stimuli for human listeners to rate each performance on perceived arousal and valence. Their findings suggest that within this set of performances, arousal is significantly correlated with attack rate and valence is affected by both the attack rate and the mode. However, that study was based on only one set of performances, making it impossible to decide whether the human emotion ratings used as ground truth really reflect aspects of the compositions themselves, or whether they were also (or even predominantly) affected by the specific (and, in some cases, rather unconventional)

Attribution: S. Chowdhury, and G. Widmer, "On Perceived Emotion in Expressive Piano Performance: Further Experimental Evidence for the Relevance of Mid-level Perceptual Features", in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, Online, 2021.

[1] Actually, attack rate as computed by B&S is also informed by the average tempo of the performance; thus, it is not strictly a score-only feature.