Towards Better Features for Music Emotion Recognition: A Machine Learning Approach

Shreyan Chowdhury
19.08.2021 | MAPLE Lab, McMaster University

JOHANNES KEPLER UNIVERSITY LINZ

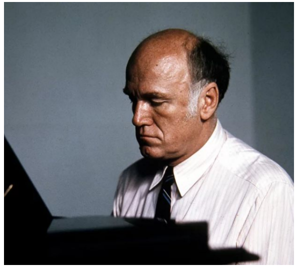
Institute of Computational Perception

Outline

- Part 1: Background
  - The Con Espressione Project
  - Machine learning refresher
  - Feature extraction for music content analysis
  - Mid-level features
- Part 2: Emotion in Bach’s Well Tempered Clavier
  - About the data
  - Feature extraction
  - Comparison of feature sets

Part 1: Background

The Con Espressione Project



The Con Espressione Project

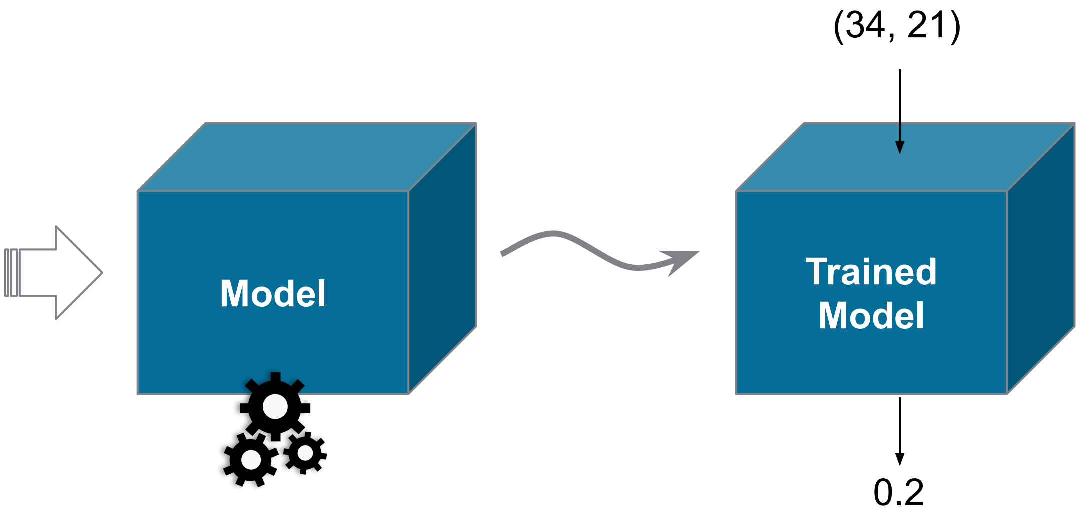
https://www.jku.at/en/institute-of-computational-perception/research/projects/con-espressione

Machine Learning Refresher

Features Labels

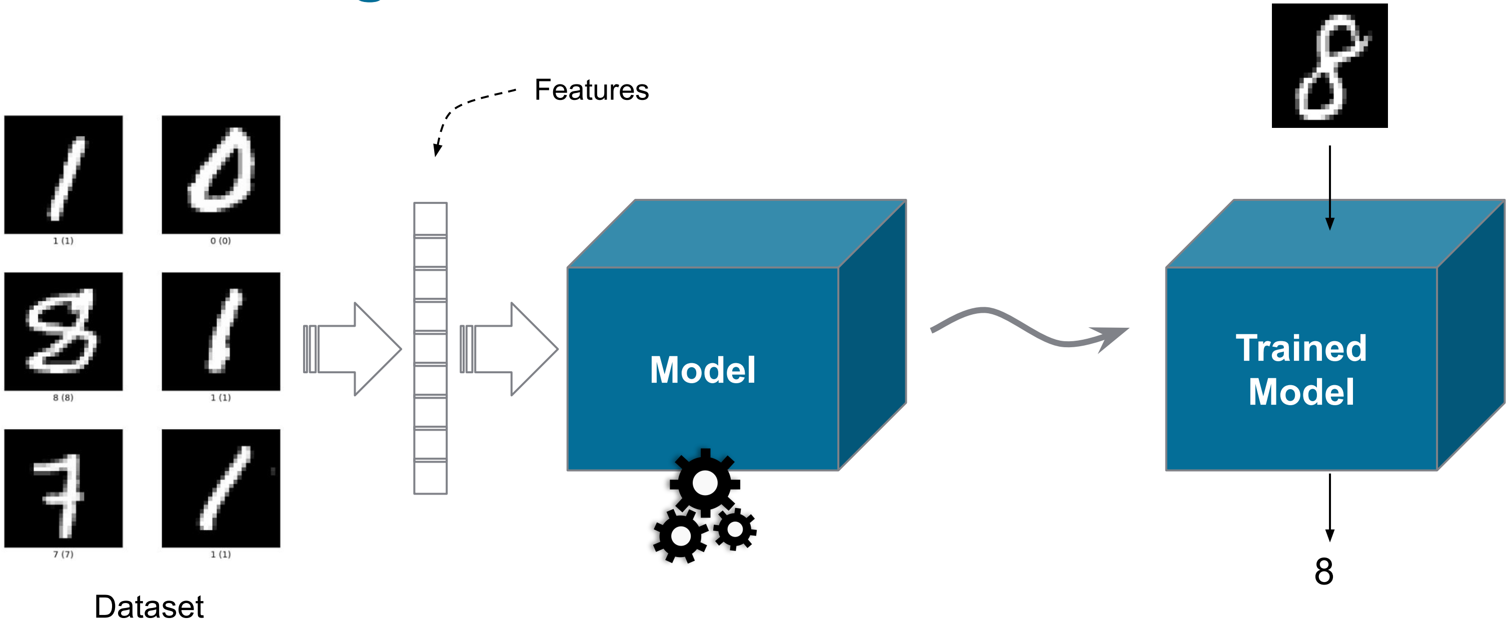
Temperature (C)	Humidity (%)	Rain
23	34	0
27	82	1
19	67	1
21	43	0

Dataset

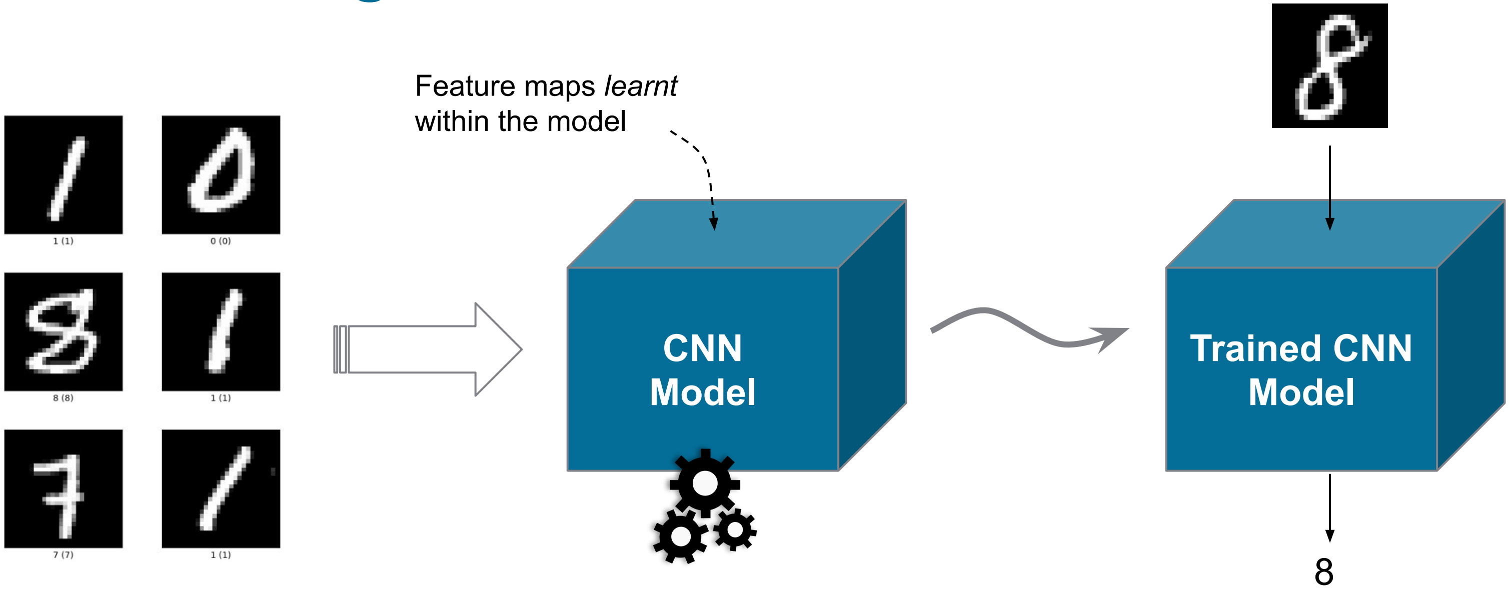


Model training
(optimizing model parameters to minimize *loss*, which is a function of the data and model parameters)

Machine Learning Refresher



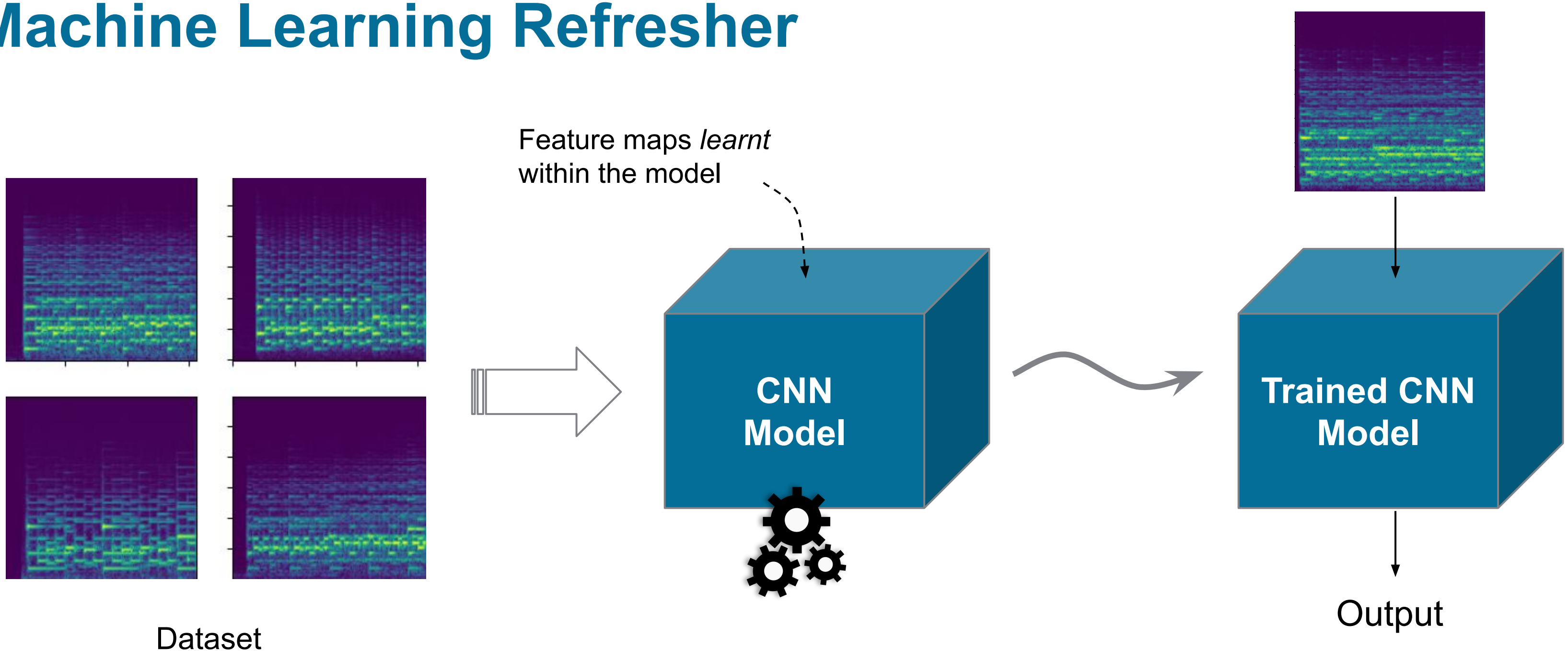
Machine Learning Refresher



Dataset

CNN: Convolutional Neural Network

Machine Learning Refresher



Typical Features for Music Content Analysis

- Time-domain features
  - Amplitude
  - Energy
  - Zero-crossing rate
- Frequency-domain features
  - Spectral centroid
  - Spectral flux
  - Mel-frequency cepstral coefficients
  - Spectral peaks
- Mixed features
  - Onset
  - Pitch
  - Tempo
  - Beats

Low-level features

The Semantic Gap

Low-level features

High-level features
(e.g. emotion)

Unambiguously defined and objectively verifiable

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

Features to Bridge the Gap?

Low-level features

Unambiguously defined and objectively verifiable

Mid-level Features

Perceptual and subjective, but make intuitive musical sense

(everything in between)

High-level features
(e.g. emotion)

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

Features to Bridge the Gap?

Mid-level Features

Perceptual and subjective, but make intuitive musical sense

- Melodiousness
- Articulation
- Rhythmic Complexity
- Rhythmic Stability
- Dissonance
- Tonal Stability
- Minorness

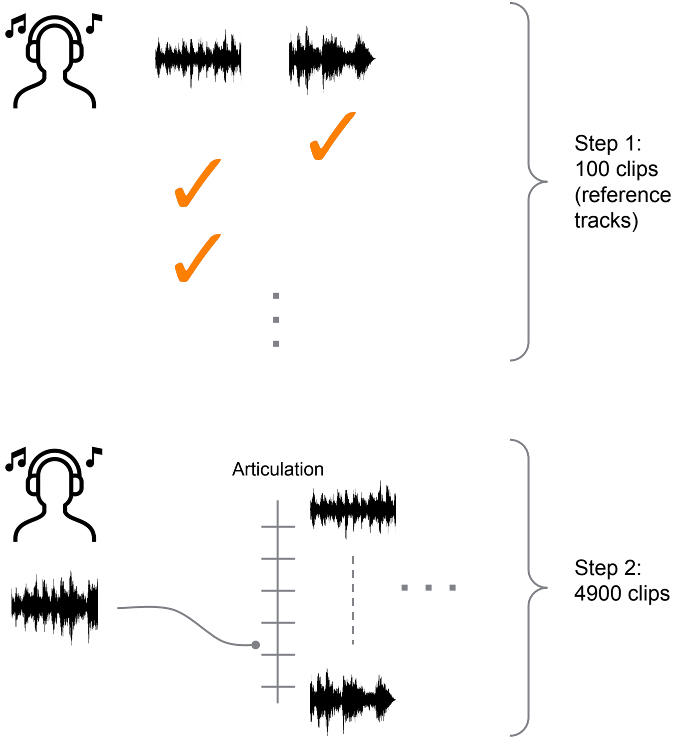
A. Aljanaki and M. Soleymani, *A Data-driven Approach to Mid-level Perceptual Musical Feature Modeling*, ISMIR 2018

Why Mid-level Features?

- Better representations of musical concepts
  - Unaffected by recording artefacts
  - Closer to human perception
- Better handle on search and retrieval
- Add interpretability/explainability to high-level concept models
- May improve prediction accuracy

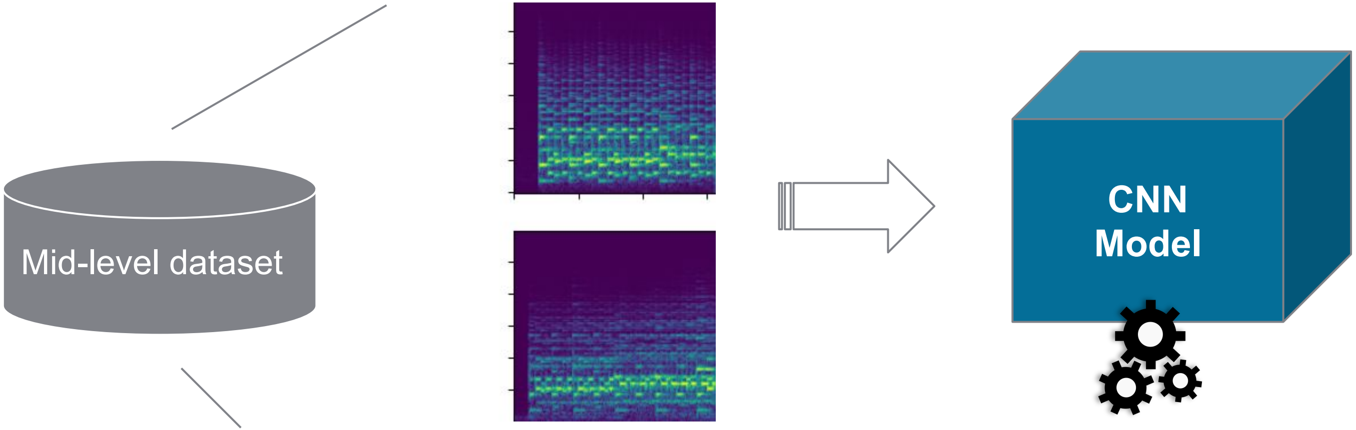
Mid-level Features through Data

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	Imagine marching along with the music. Which is easier to march along with?
Rhythmic Complexity	Is it difficult to repeat by tapping? Is it difficult to find the meter? Does the rhythm have many layers?
Dissonance	Which excerpt has noisier timbre? Has more dissonant intervals (tritones, seconds, etc.)?
Tonal Stability	Where is it easier to determine the tonic and key? In which excerpt are there more modulations?
Modality ('Minorness')	Imagine accompanying this song with chords. Which song would have more minor chords?



A. Aljanaki and M. Soleymani, *A Data-driven Approach to Mid-level Perceptual Musical Feature Modeling*, ISMIR 2018

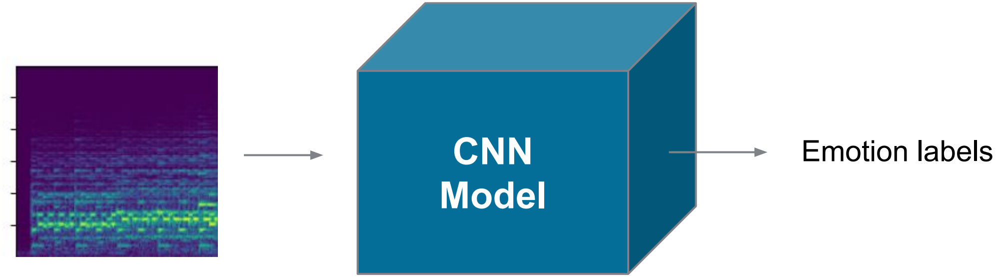
Learning to Predict Mid-level Features



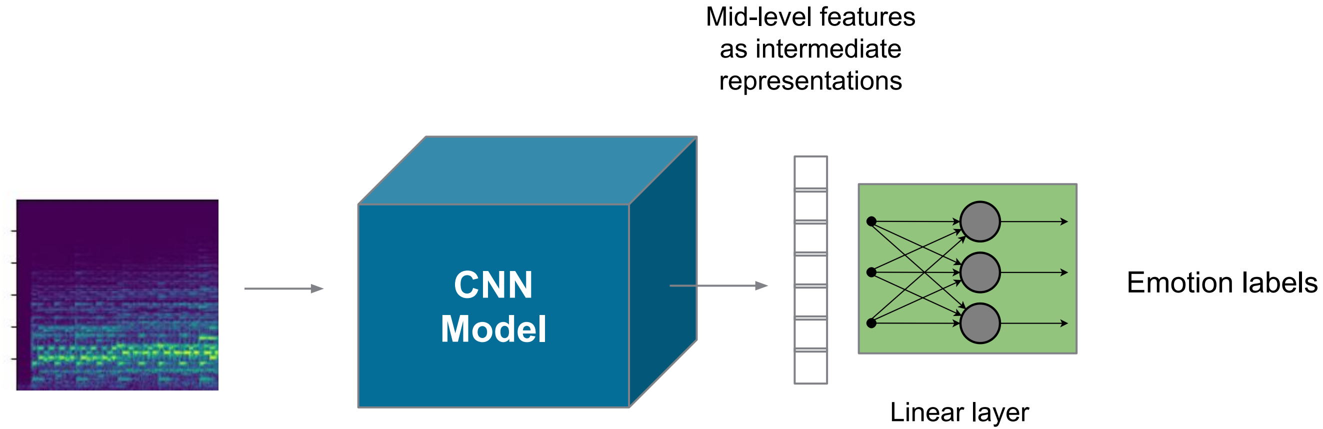
	song_id	melody	articulation	rhythm_complexity	rhythm_stability	dissonance	atonality	mode
2	1.0	8.8	4.0	4.0	6.7	3.4	7.4	6.0
3	2.0	9.0	2.8	2.3	8.0	2.2	7.6	3.8
4	3.0	7.6	8.0	6.3	7.7	3.2	6.6	4.8
5	4.0	7.2	3.2	5.0	6.3	2.4	7.8	6.2
6	5.0	8.0	3.8	4.8	6.5	2.6	7.2	6.2

S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "*Towards Explainable Music Emotion Recognition: The Route via Mid-level Features*," in Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019

Mid-level Features for Explainable Emotion Recognition

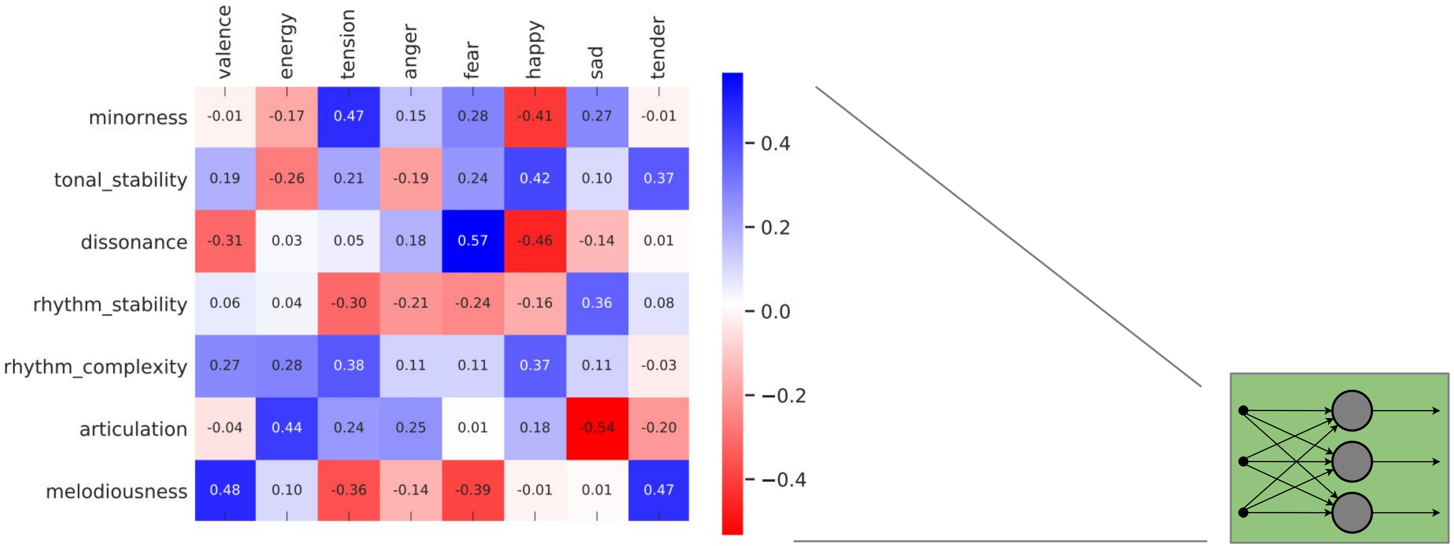


Mid-level Features for Explainable Emotion Recognition



Training labels: both mid-level and emotion annotations

Mid-level Features for Explainable Emotion Recognition



Learned weights of the linear layer

S. Chowdhury, A. Vall, V. Haunschmid, and G. Widmer, "*Towards Explainable Music Emotion Recognition: The Route via Mid-level Features*," in Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019

JOHANNES KEPLER UNIVERSITY LINZ

Part 2: Emotion in WTC

Performance Aspect of Music Emotion

“Singing, with intimate sentiment”
“Singing and expressive”

Beethoven - Piano Sonata No.30

Research Questions

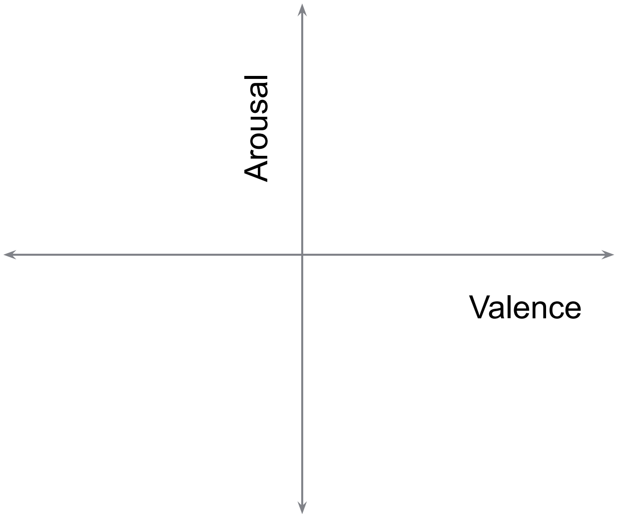
- Modeling perceived emotion in Bach's *Well Tempered Clavier Book 1.*

  - Comparison of feature sets:
    - Low-level audio features
    - Score-based features
    - Mid-level features
    - Emotion features
  - In each feature set, which features are the most important?
  - Which feature set best explains variation of arousal and valence
    - *between pieces*?
    - *between different performances of the same piece*?

JOHANNES KEPLER
UNIVERSITY LINZ

Research Questions

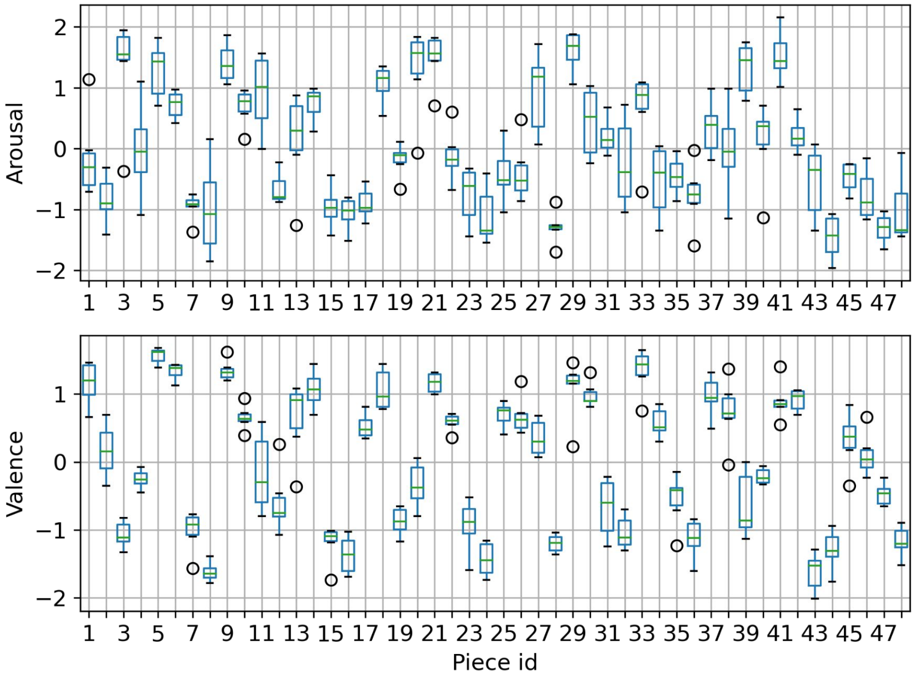
- Modeling perceived <u>emotion</u> in Bach's *Well Tempered Clavier Book 1.*



Data – WTC Recordings and Emotion Ratings

- 288 performances of the WTC
  (48 pieces played by 6 different pianists)
  - Glenn Gould
  - Friedrich Gulda
  - Angela Hewitt
  - Sviatoslav Richter
  - András Schiff
  - Rosalyn Tureck
- First 8 bars
- Arousal (0 to 100) and valence (–5 to +5) ratings by University students
- Each track rated by 29 participants

Distribution of Mean Emotion Ratings by Piece



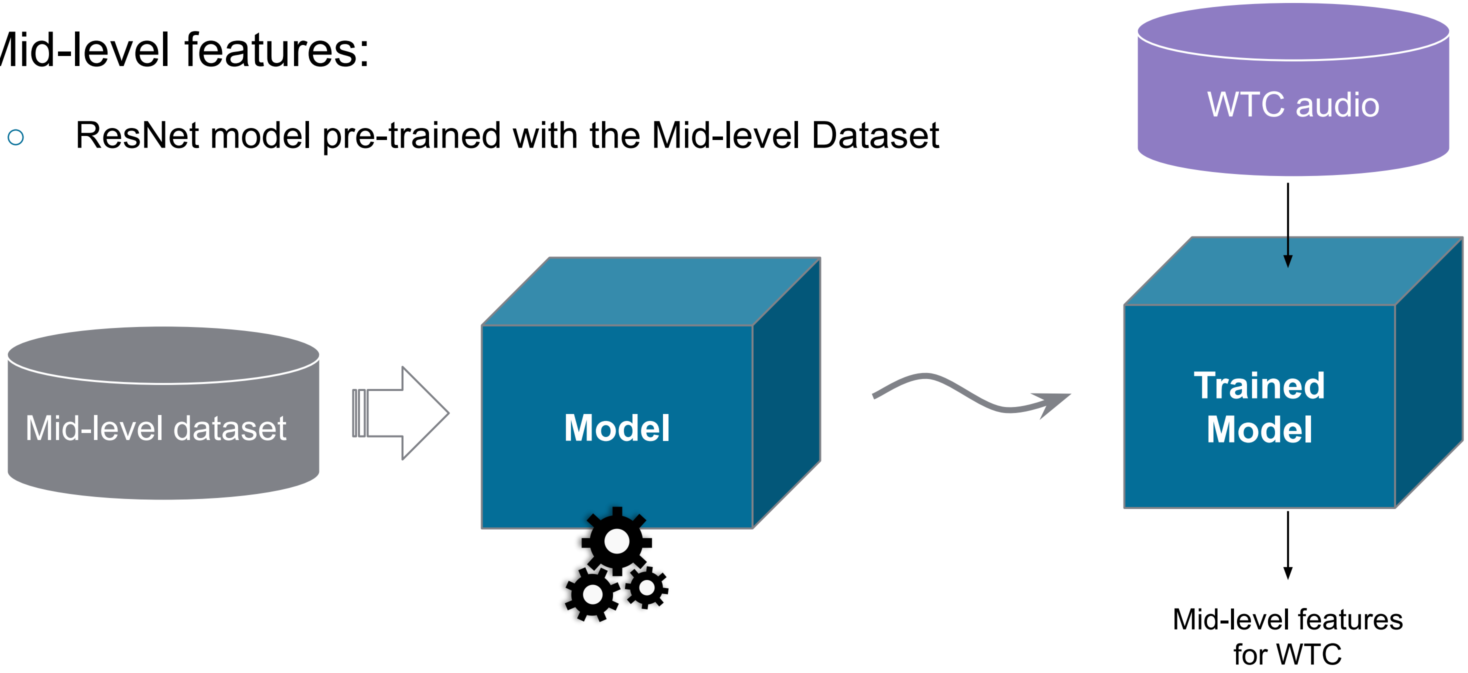
JOHANNES KEPLER UNIVERSITY LINZ

Feature Extraction

- Low-level audio features:
  - Essentia/Librosa
  - 11 features + mean and standard deviation for each feature across a clip
  - Time domain, frequency-domain, and mixed domain features
  - Loudness, onset rate, pitch salience, spectral centroid, tempo, etc.
- Score-based features:
  - Computed from sheet music
  - Onset density, pitch density, mode, key strength, inter onset interval.

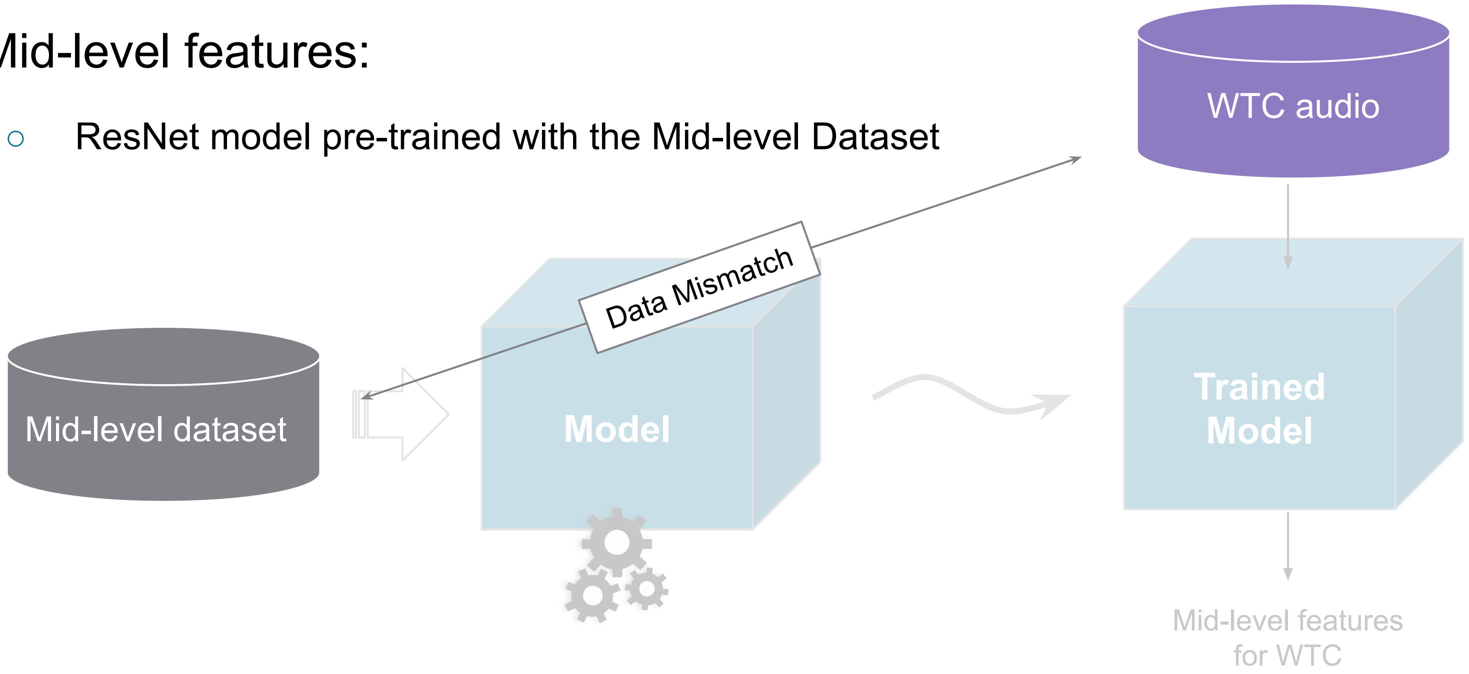
Feature Extraction

- Mid-level features:
  - ResNet model pre-trained with the Mid-level Dataset



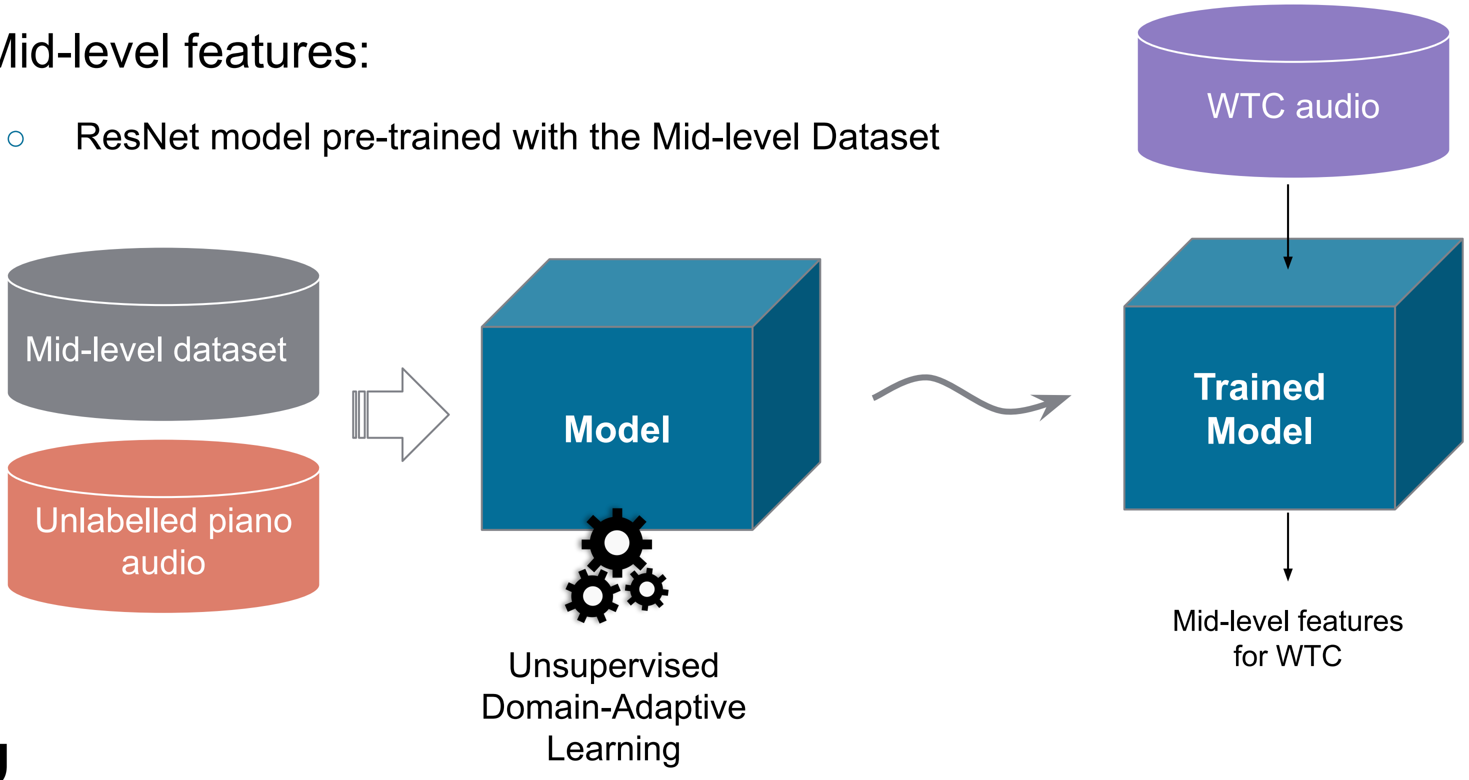
Feature Extraction

- Mid-level features:
  - ResNet model pre-trained with the Mid-level Dataset



Feature Extraction

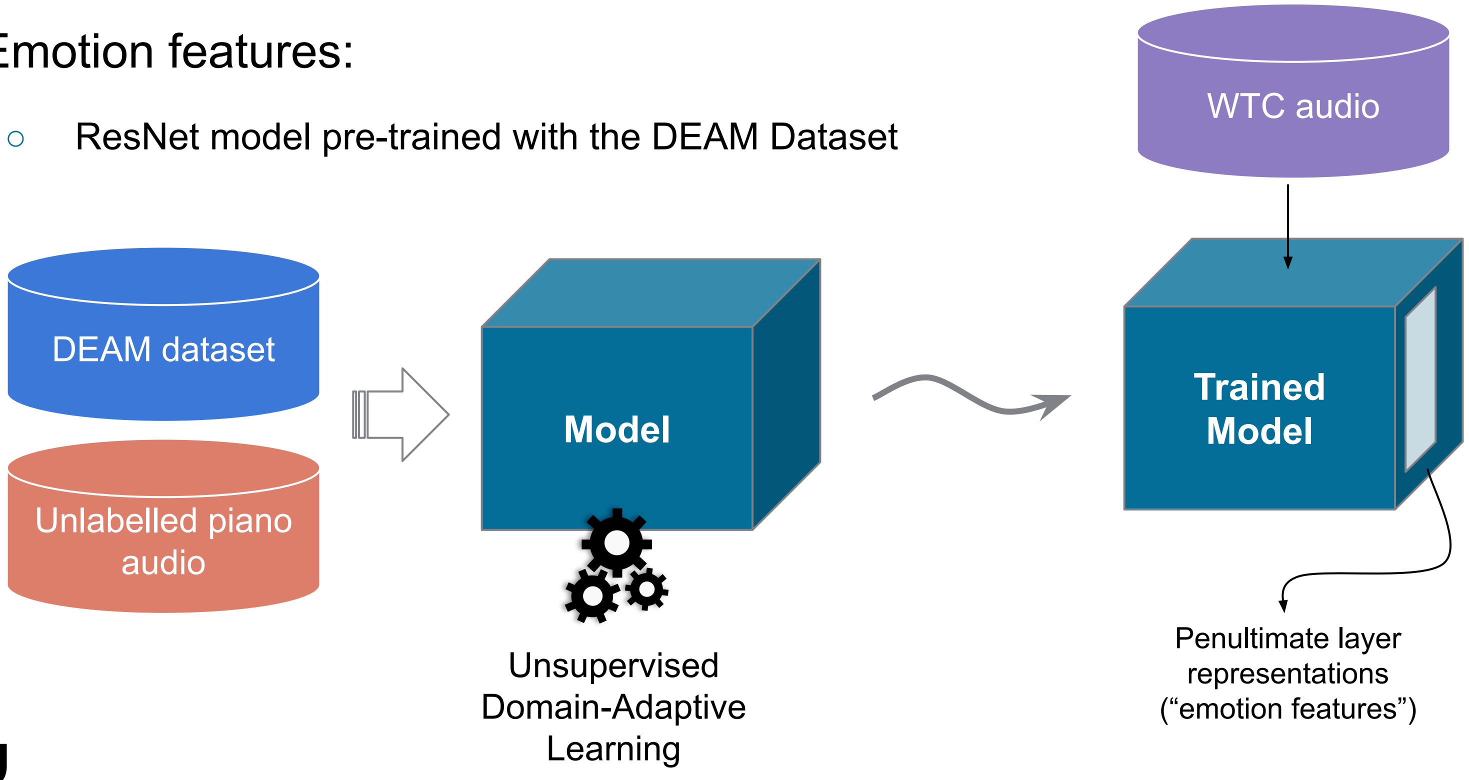
- Mid-level features:
  - ResNet model pre-trained with the Mid-level Dataset



JOHANNES KEPLER UNIVERSITY LINZ

Feature Extraction

- Emotion features:
  - ResNet model pre-trained with the DEAM Dataset



JOHANNES KEPLER UNIVERSITY LINZ

Feature Extraction

- Emotion features:
  - ResNet model pre-trained with the DEAM Dataset
  - Extract penultimate layer representations for WTC
  - 512 features per clip
  - Perform Principal Component Analysis to reduce feature space
  - Obtain 9 components that explain 98% of variance
  - “DEAMResNet” features

Feature Comparison

- Ordinary least squares fitting
- Regression metrics:
    - Adjusted R2 score, Root mean squared error (RMSE), Pearson’s correlation coefficient
    - Fraction of variance unexplained
- Feature importance metric:
    - T-statistic $t = \frac{\hat{\beta}}{\mathrm{SE}\left(\hat{\beta}\right)}$
- Mixed model regression metric:
    - Fraction of residual variance explained

Performance on only Gulda's Recording

	Arousal			Valence		
	$\tilde{R}^2$	RMSE	Corr	$\tilde{R}^2$	RMSE	Corr
Mid-level	0.84	0.36	0.93	**0.79**	0.42	0.91
DEAMResNet	**0.91**	0.27	0.96	0.69	0.50	0.86
Low-level	0.86	0.29	0.96	0.67	0.45	0.89
Score	0.31	0.74	0.67	0.61	0.55	0.83
B&S (exp 3)	0.48	-	-	0.75	-	-

A. Battcock and M. Schutz, "Acoustically expressing affect," Music Perception: An Interdisciplinary Journal, vol. 37, pp. 66–91, 2019

Performance on the Complete Dataset

Fitting

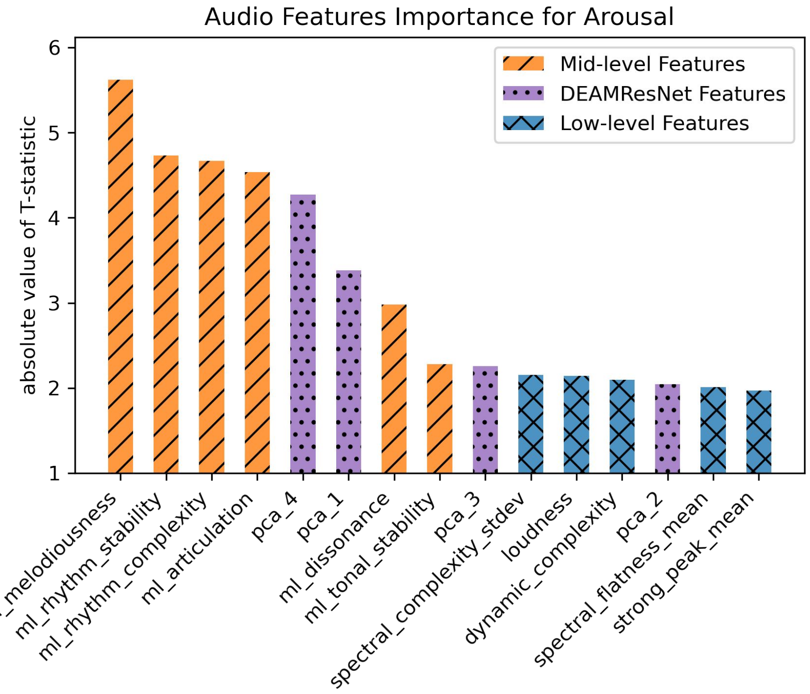
Feature Set	Arousal $\tilde{R}^2$	Arousal RMSE	Arousal Corr	Valence $\tilde{R}^2$	Valence RMSE	Valence Corr
Mid-level	0.68	0.56	0.83	0.63	0.60	0.80
DEAMResNet	**0.70**	0.54	0.84	0.42	0.72	0.69
Low-level	0.62	0.59	0.81	0.41	0.74	0.67
Score	0.41	0.75	0.65	**0.75**	0.49	0.87

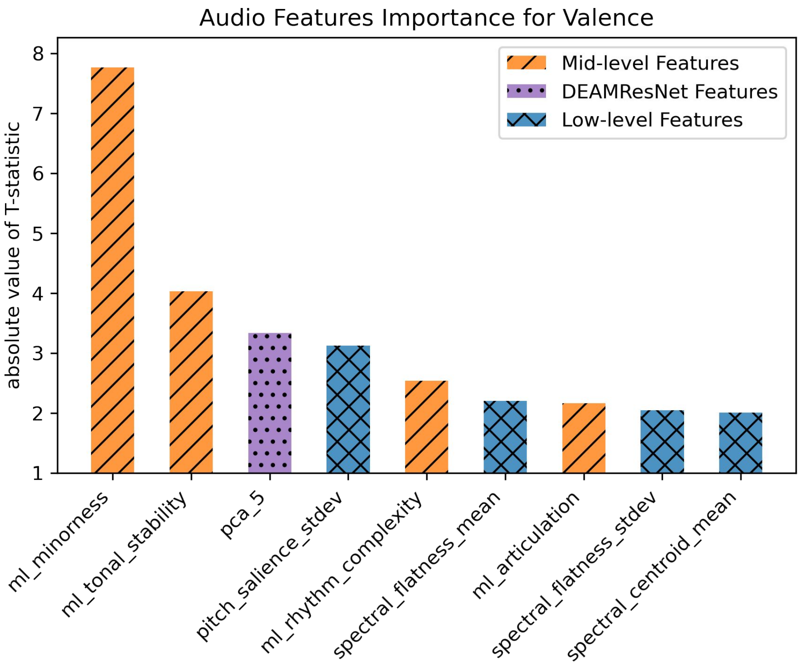
Generalization

Feature Set	Piece-wise A	Piece-wise V	Pianist-wise A	Pianist-wise V	LOO A	LOO 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**

JOHANNES KEPLER UNIVERSITY LINZ

Feature Importance among Audio-based Features





JOHANNES KEPLER UNIVERSITY LINZ

Testing Piece-wise Variation

$$E_{\text{random}} = \frac{\text{Var}_{\text{random}}}{\text{Var}_{\text{random}} + \text{Var}_{\text{residual}}}$$

Linear mixed models

All features
+
Random intercept
(piece id)

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**

Fraction of residual variance explained by the random effect of “piece id”.

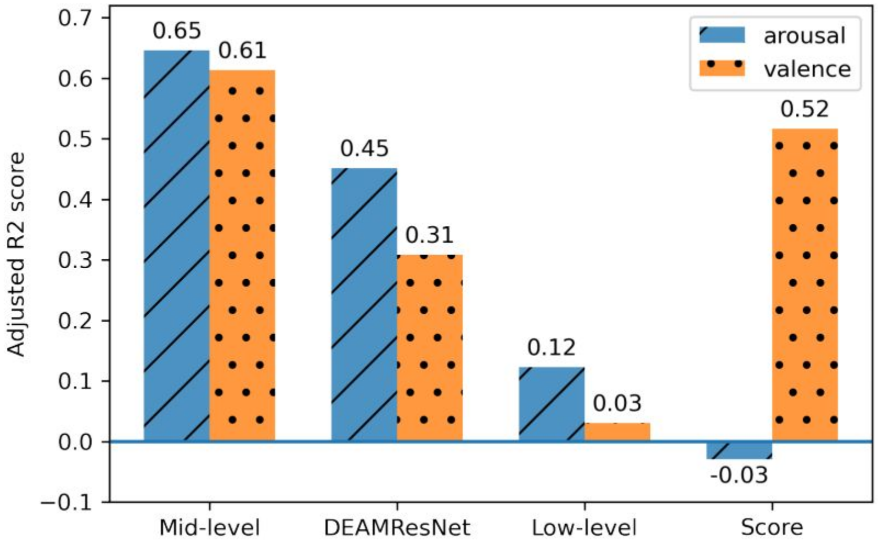
Testing Performance-wise Variation

- Overall means (of arousal or valence) are almost identical for all pianists
  - Linear mixed models cannot be used
- Train on 47 pieces and test on the remaining piece
- Metric: Fraction of Variance Unexplained

	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%)

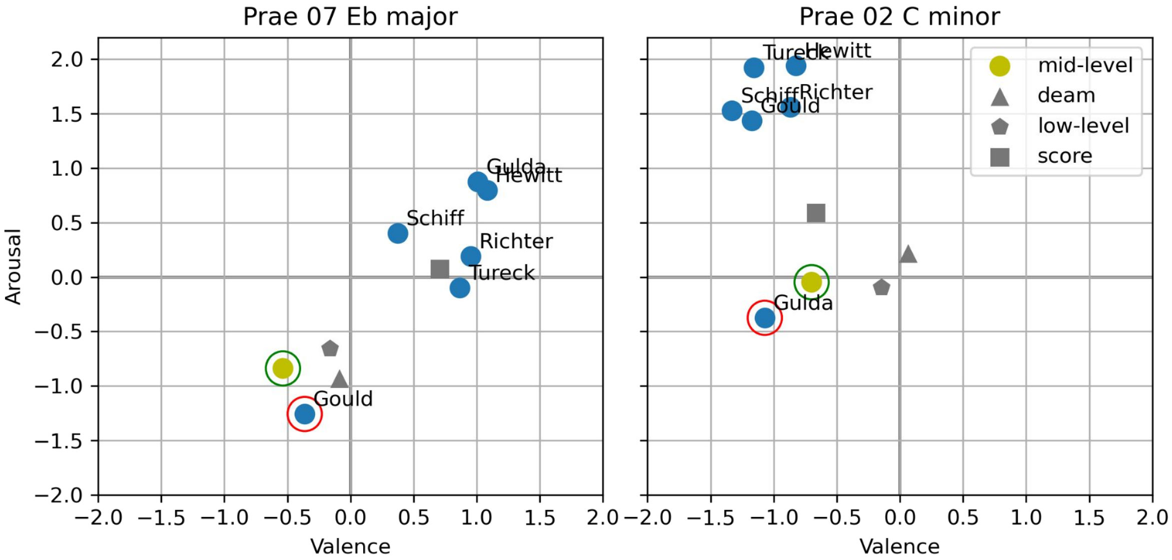
Generalizing Power: Predicting Emotion of Outliers

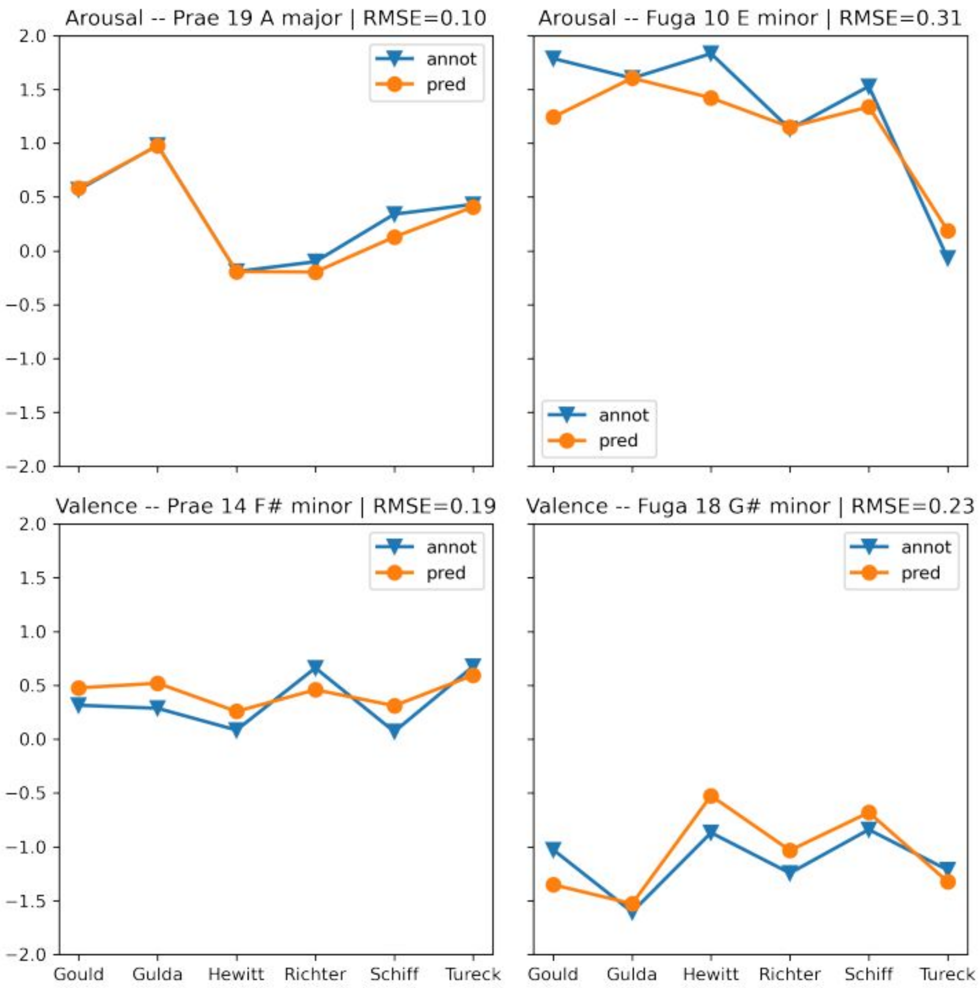
- Held out data: 48 outlier performances (one for each piece)



JOHANNES KEPLER
UNIVERSITY LINZ

Generalizing Power: Predicting Emotion of Outliers





JOHANNES KEPLER
UNIVERSITY LINZ

Thank you!