Tracing Back Music Emotion Predictions To Sound Sources And Intuitive Perceptual Qualities

Shreyan Chowdhury, Verena Praher, Gerhard Widmer

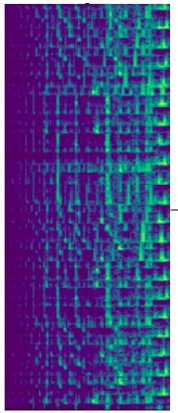
JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

SMC Reconnecting spaces 2021

$(\mathbf{X}_i, [y_i^{(a)}, y_i^{(v)}]) \in \mathcal{D}$

(DEAM)



Audio to Emotion Model

Arousal $\hat{y}^{(a)}$

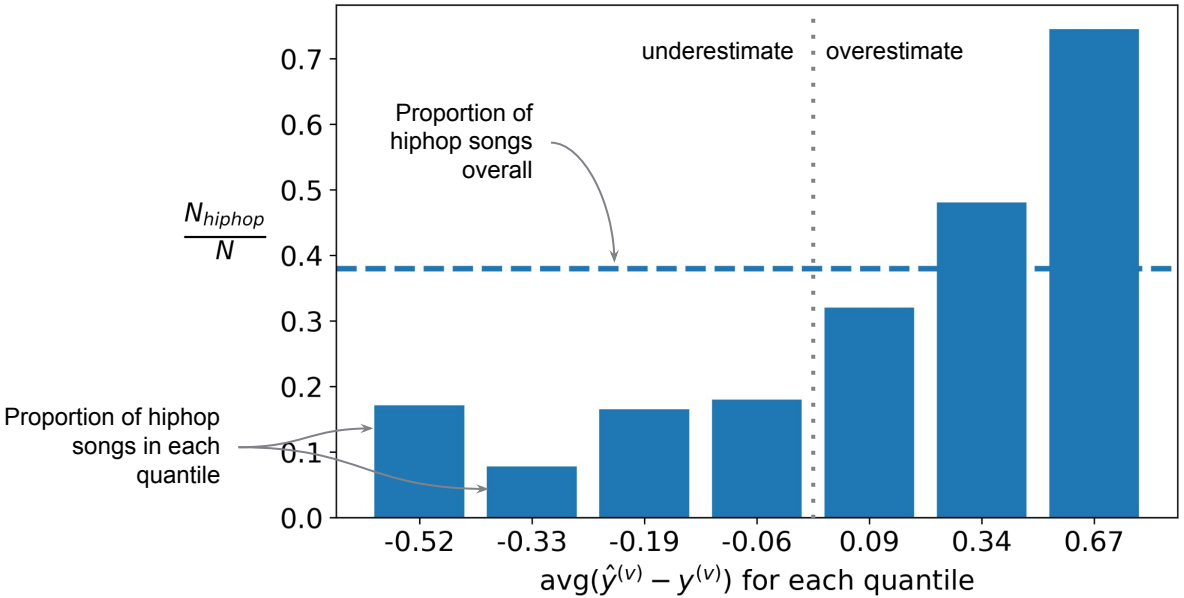
Valence $\hat{y}^{(v)}$

$\tilde{\mathbf{X}} \in \tilde{\mathcal{D}}$

(PMEmo)

$y^{(a)}, \hat{y}^{(a)}, y^{(v)}, \hat{y}^{(v)} \in [-1, 1]$

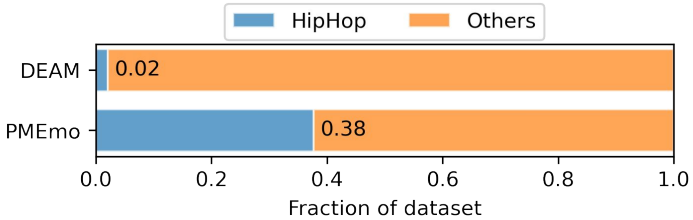
Valence Error Quantiles



Test Set: PMEmo

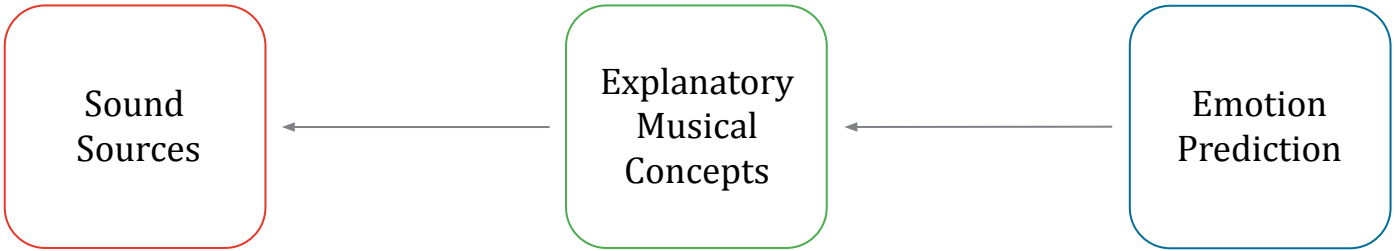
Peering Inside the Model and the Music

- Why the overestimation?

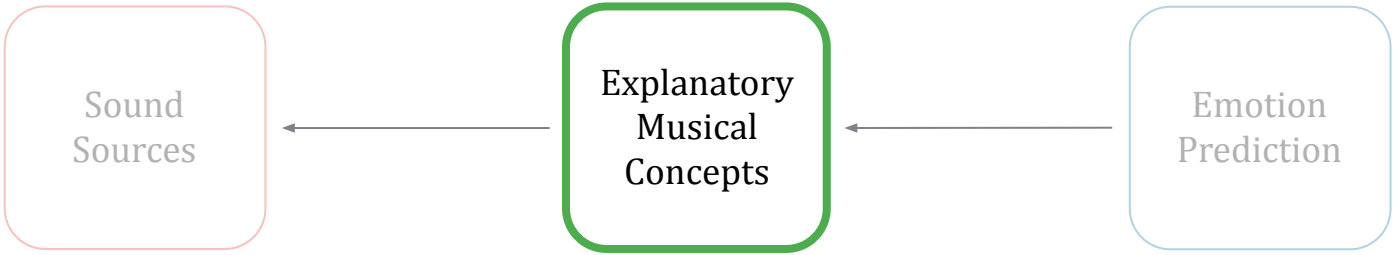


- What musical properties lead to this specific bias?
  - Two-level explanations
- Does intervention lead to expected change?
  - Retraining on balanced data

Two Levels of Explanations



Two Levels of Explanations



Explanatory Musical Concepts: Mid-level Features

Low-level features, such as pitch

Building blocks of musical signals

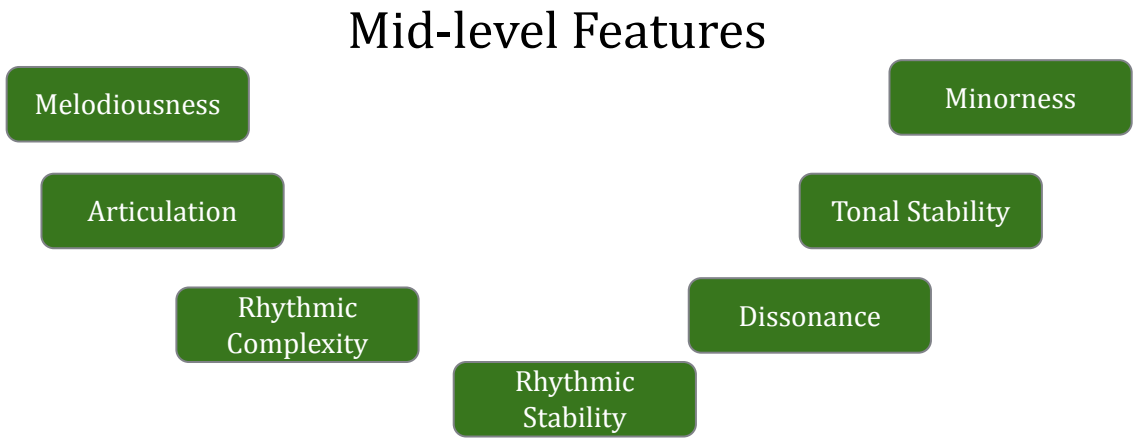
Mid-level Features

Perceptual and subjective, but make intuitive musical sense

High-level features, such as genre

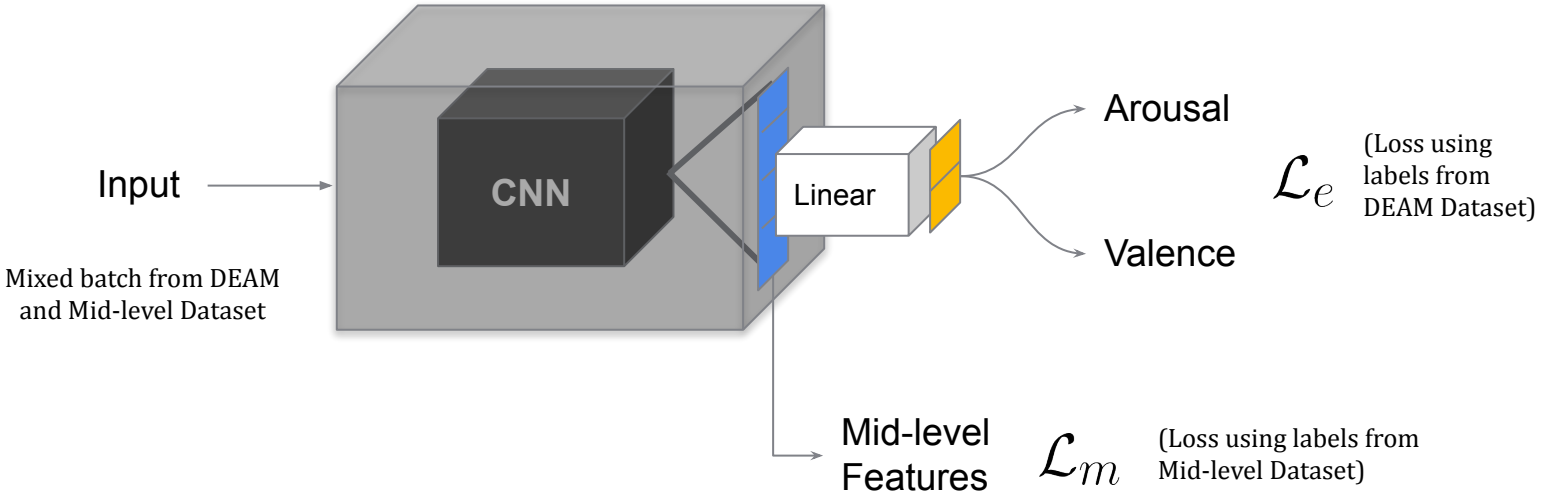
Subjective, abstract descriptions

Mid-level Features: What are they?



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

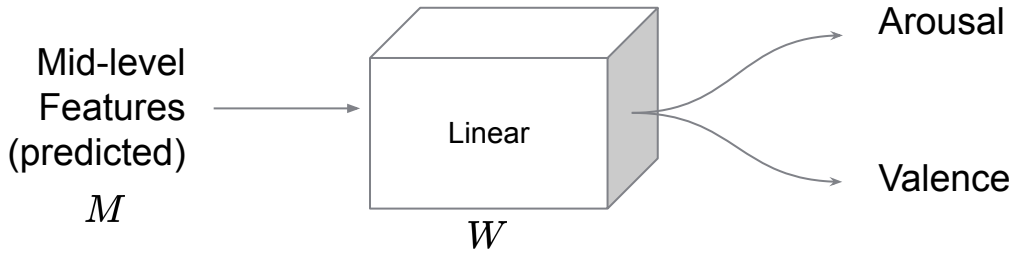
Training the Explainable Model



Joint multitask learning: backpropagate combined loss

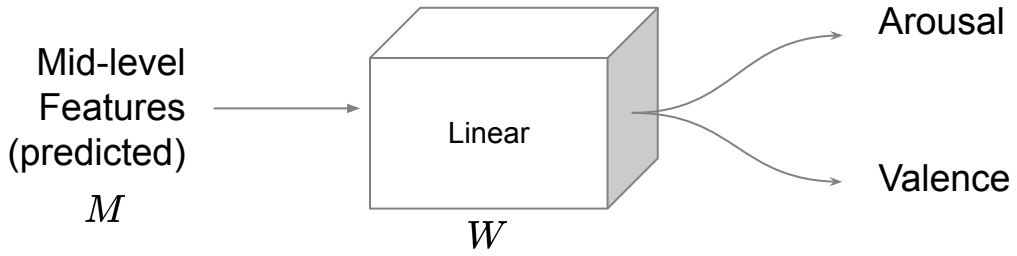
$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_e$$

Obtaining Mid-level Explanations



Effect: $\mathrm{E}_j^{(k)} = w_j^{(k)} m_j \quad k \in \{\text{arousal, valence}\}$

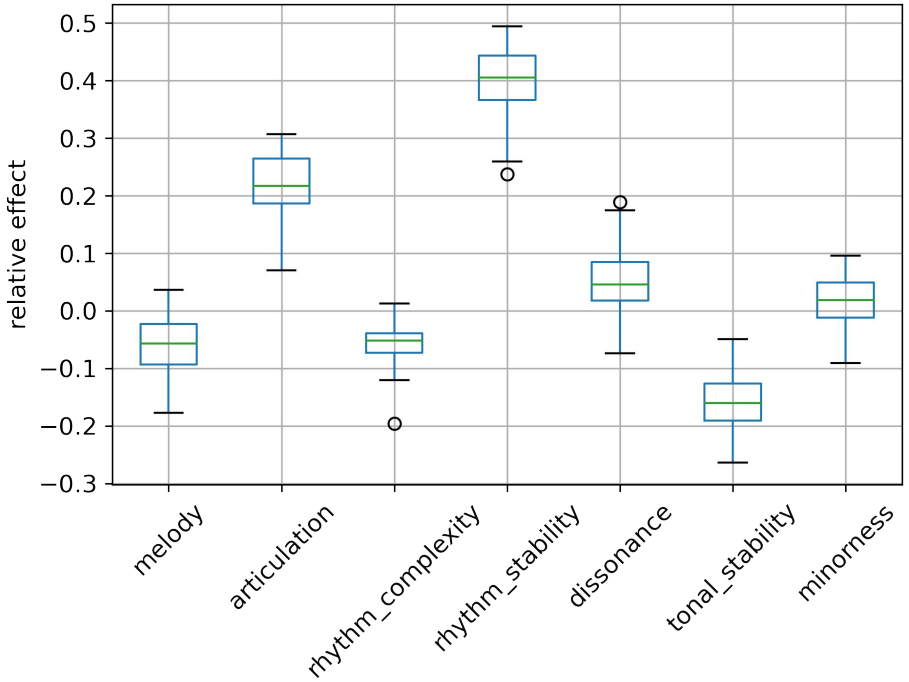
Obtaining Mid-level Explanations



Effect: $\mathrm{E}_j^{(k)} = w_j^{(k)} m_j \qquad k \in \{\text{arousal, valence}\}$

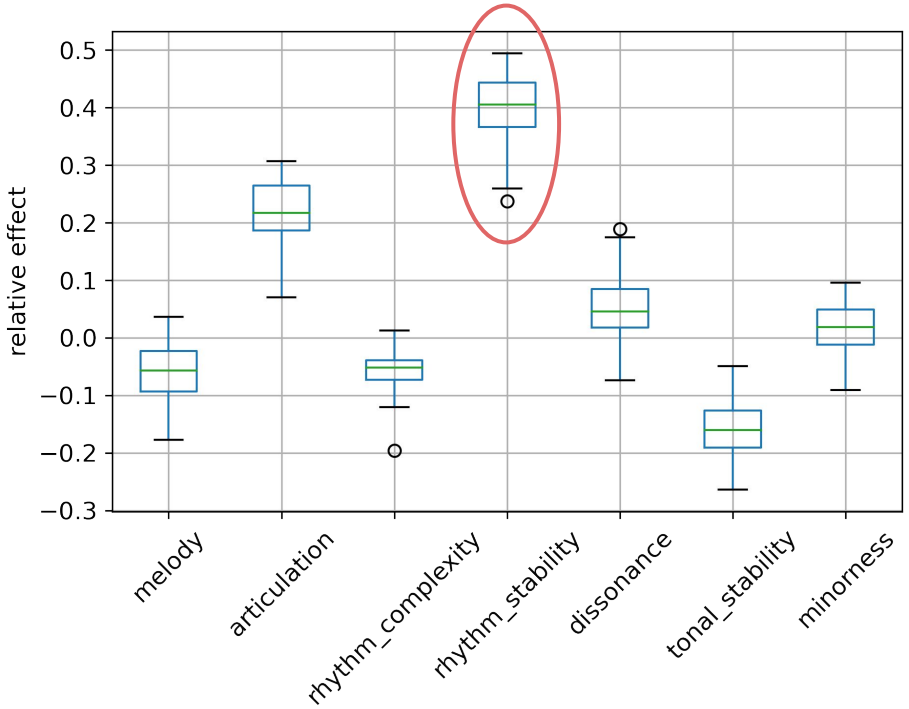
Relative Effect: $\mathrm{E}_{\mathrm{rel}\ j}^{(k)} = \dfrac{\left|\mathrm{E}_j^{(k)}\right|}{\sum_i \left|\mathrm{E}_i^{(k)}\right|}$

Mid-level Explanations for Valence in Hiphop



Relative effects of mid-level features on valence
for hiphop songs

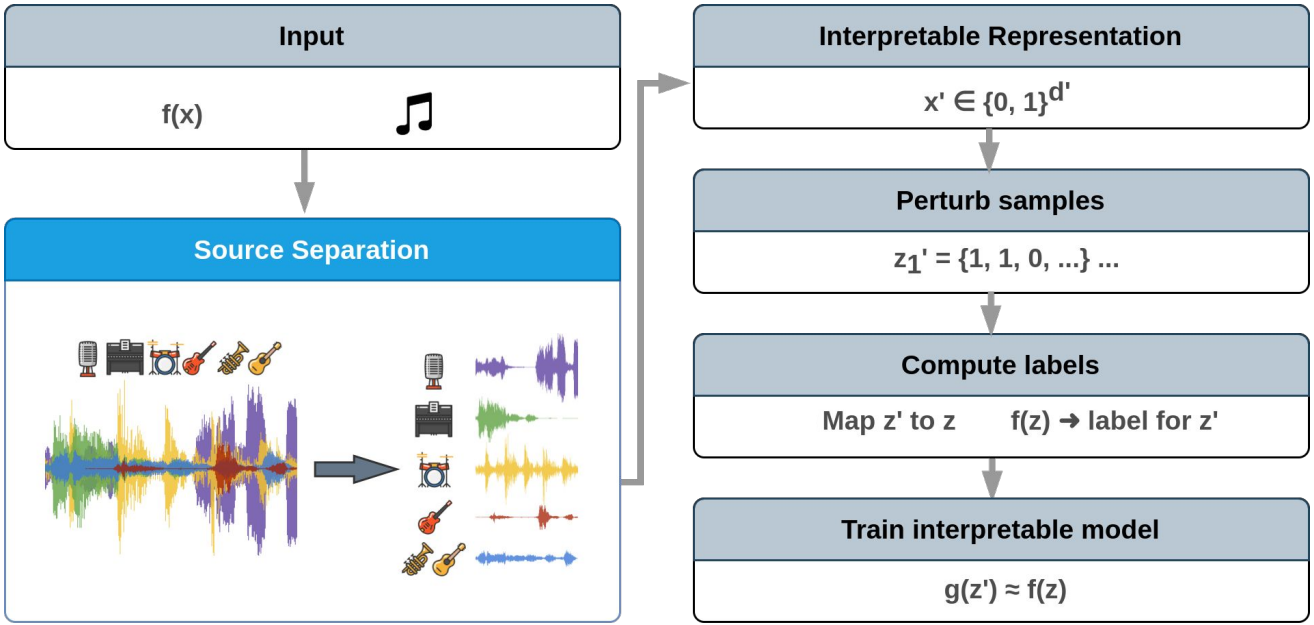
Mid-level Explanations for Valence in Hiphop



What musical component is the “cause” for high rhythm stability predictions?

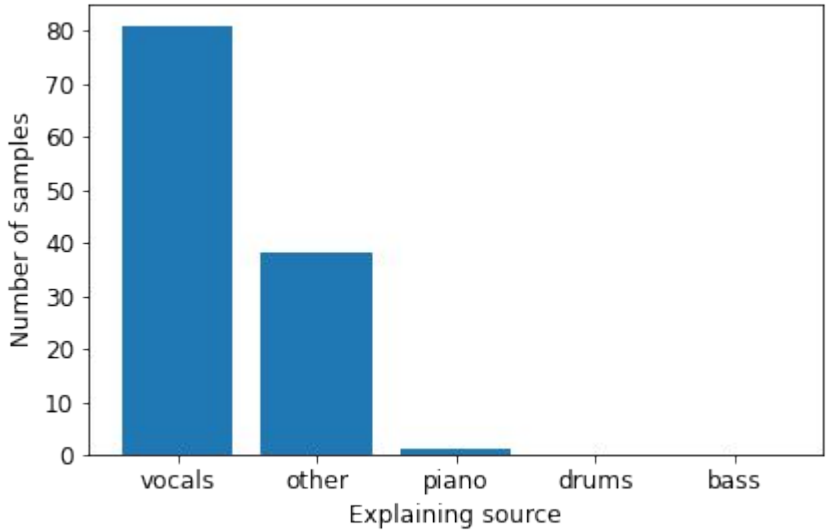

Sound Sources
Explanatory Musical Concepts
Emotion Prediction

Obtaining Sound Source Explanations: audioLIME

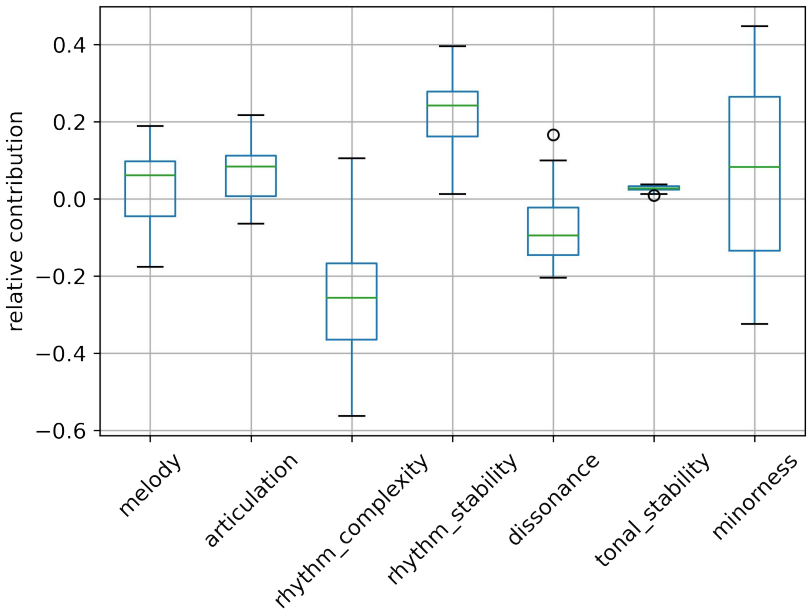


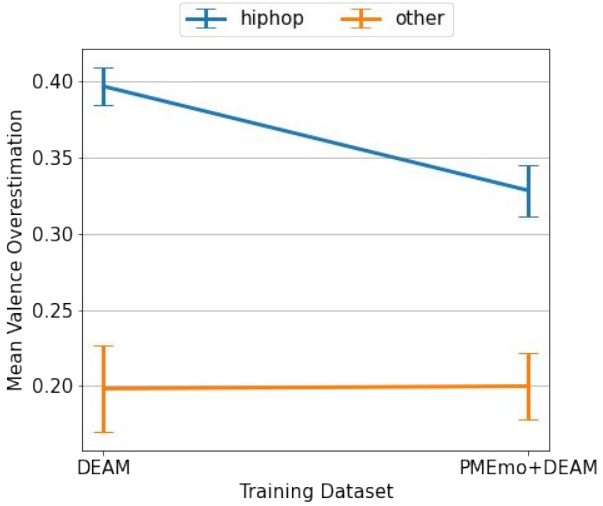
V. Haunschmid, E. Manilow, and G. Widmer, *audioLIME: Listenable Explanations Using Source Separation*, in 13th International Workshop on Machine Learning and Music, 2020

Sound Source Explanations for Rhythm Stability

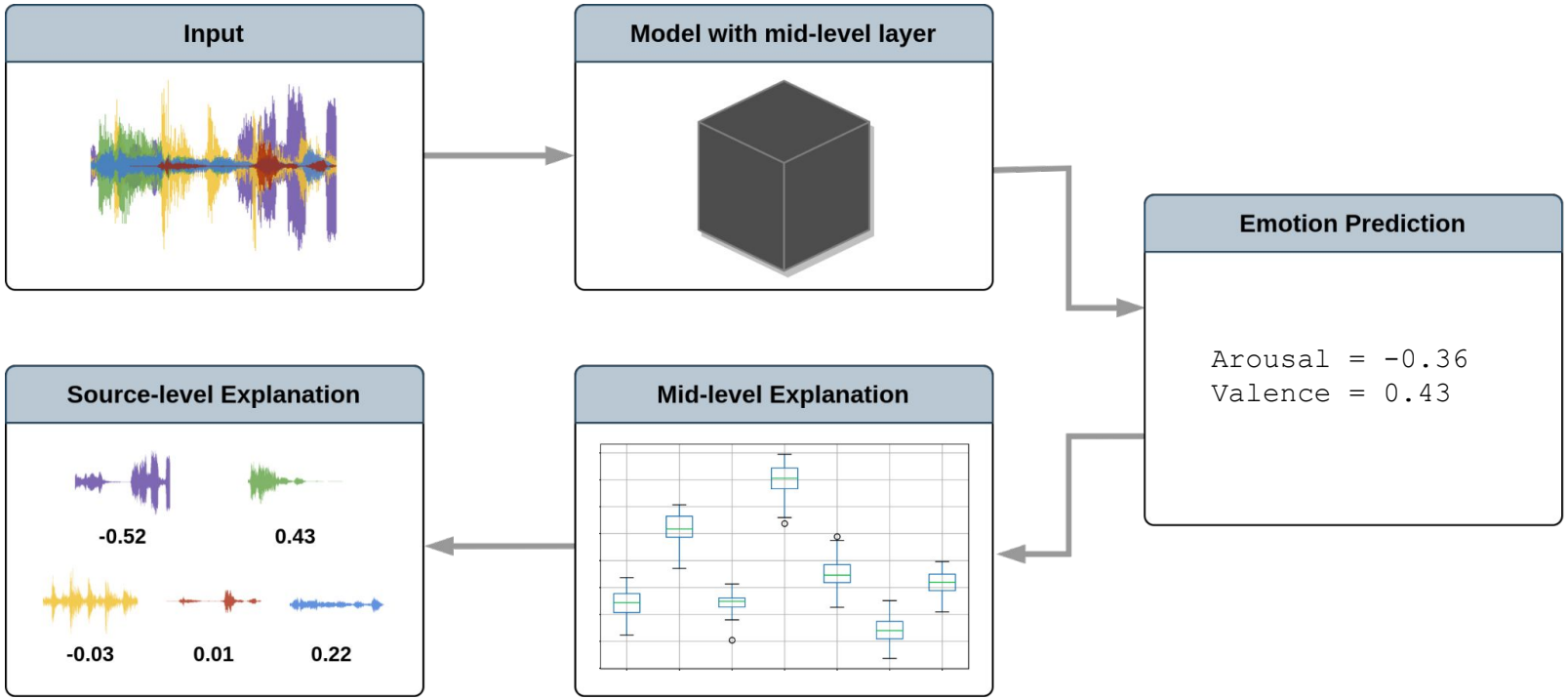


Retraining with Combined Training Set

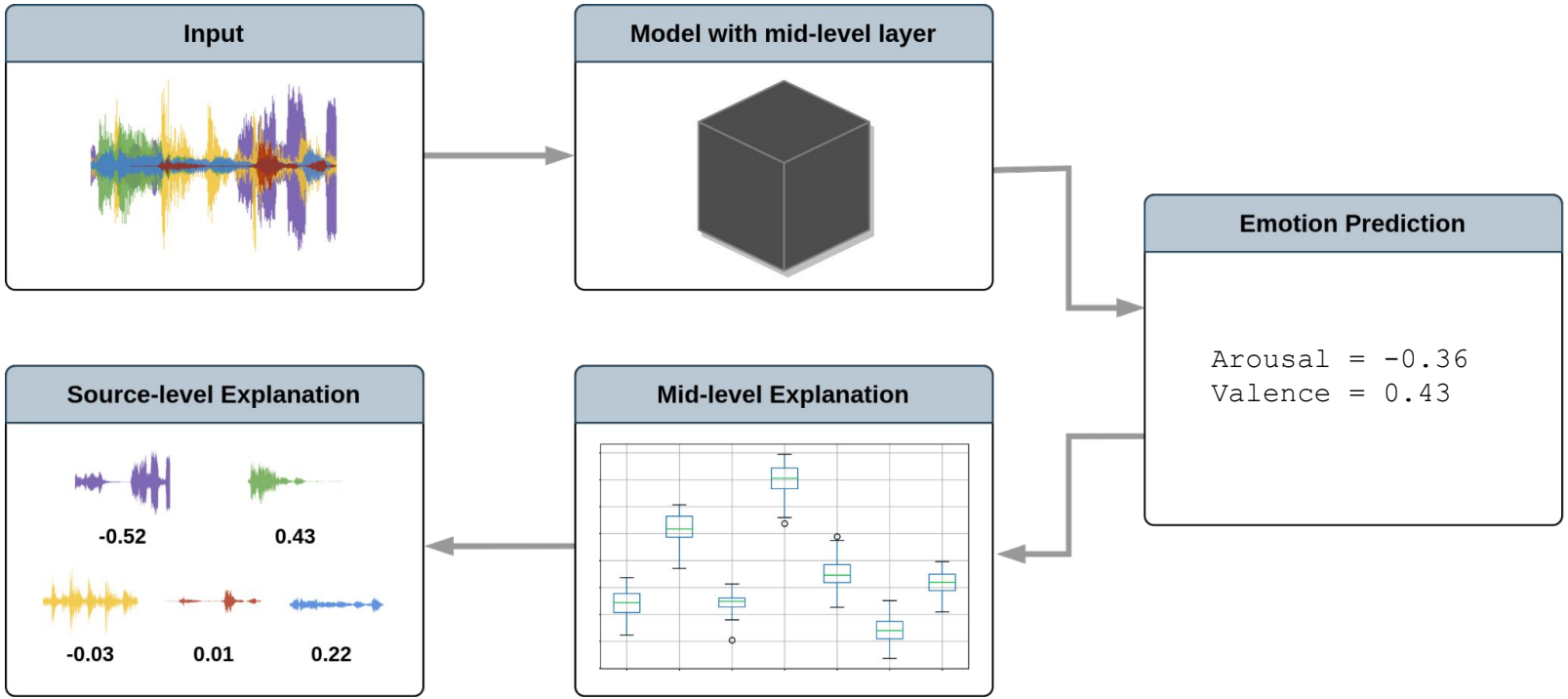




Schematic


Input
Model with mid-level layer
Emotion Prediction
Arousal = -0.36
Valence = 0.43
Source-level Explanation
-0.52
0.43
-0.03
0.01
0.22
Mid-level Explanation

Schematic



Thank you!