# Evaluating Extracted Musical Features with Versions


Konrad Swierczek[1], Michael Schutz[1,2]

swierckj@mcmaster.ca schutz@mcmaster.ca

[1] Department of Psychology, Neuroscience & Behaviour, McMaster University

[2] School of the Arts, McMaster University


MAPLE Lab — Music Acoustics Perception LEarning

Visit our website at **maplelab.net**

konradswierczek.ca

## Background

- Music Information Retrieval (MIR) is frequently used in academia and industry to analyze and classify digital music files
  - **Applications:** Recommendation Systems, Emotion Analysis, Genre Classification, Acoustic Fingerprinting, Music Generation

- Despite widespread use, little testing of MIR tools has been conducted
  - Evaluation is difficult: lack of ground truth and labelled data

**How can we evaluate the accuracy of subjective features?**

- In classical music, structural features like mode are unchanged while interpretive features like tempo are different in each performance

## Method

ESSENTIA

librosa

*MIRtoolbox* 1.8.1


| | Should Vary | Should Not Vary |
|---|---|---|
| Spectral Features | Spectral Centroid | Mode |
| Temporal Features | Tempo | Number of Onsets |


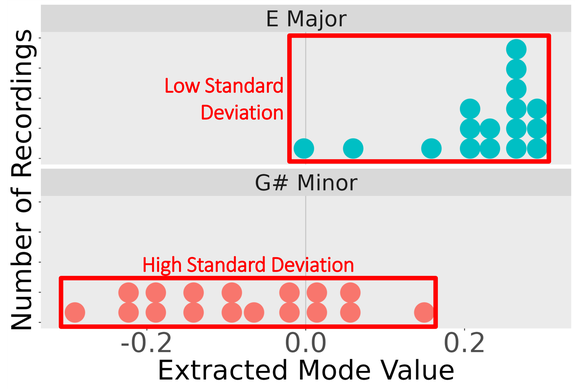


**Fig. 2**: Prelude with lowest (E Major) and highest (G# Major) extracted mode standard deviation. Each dot represents a version. Positive mode values indicate major mode, negative minor.

## Results

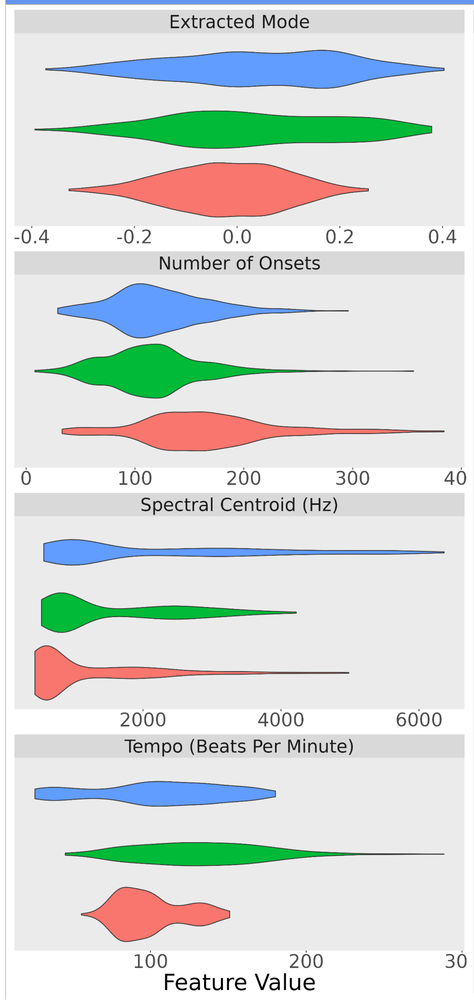


**Fig. 2**: Continuous distributions of raw feature values. Each plot includes all 384 (24 preludes x 16 versions) audio files. Features are shown in original units.

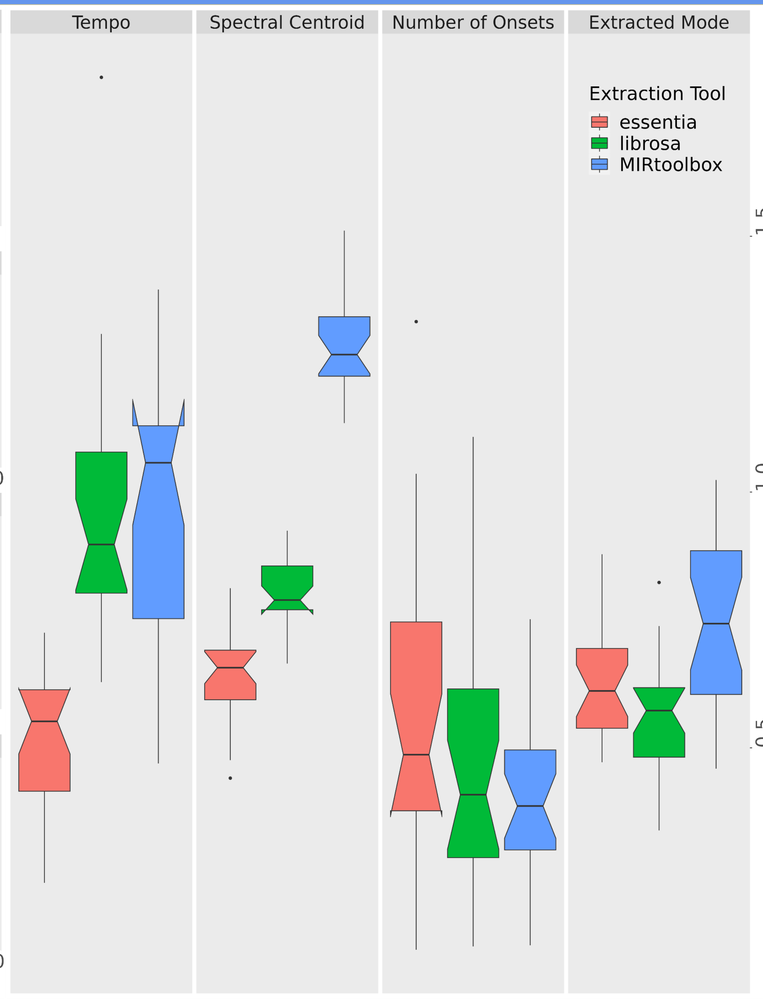


**Fig. 3**: Standard deviations of standardized feature values for each extraction tool. Each boxplot includes 24 data points, one for each prelude, calculated using the 16 versions. Higher values indicated a greater degree of variability while lower values indicate greater consistency between versions.

## Summary

- We propose a method for evaluating MIR features that does not rely on ground truth
- Mode extraction is more variable than number of onsets
- Apart from the number of onsets, MIRtoolbox is more variable than other tools
- These analyses can help inform decisions when selecting a tool or algorithm

## Selected References


Gómez, E. (2006). Tonal Description of Polyphonic Audio for Music Content Processing. *INFORMS Journal on Computing*, *18*(3), 294–304.

Kahneman, D., Sibony, O., & Sunstein, C. (2021). Noise: A Flaw in Human Judgment (p. 454). William Collins.

Kumar, N., Kumar, R., & Bhattacharya, S. (2015). Testing reliability of Mirtoolbox. *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 710 717.

Moffat, D., Ronan, D., & Reiss, J. D. (2015). *An Evaluation of Audio Feature Extraction Toolboxes*.

Sturm, B. L. (2016). The "Horse" Inside: Seeking Causes Behind the Behaviors of Music Content Analysis Systems. *Computers in Entertainment*, *14*(2), 1–32.


## Acknowledgments


Ministry of Ontario Research and Innovation

Canada Foundation for Innovation / Fondation canadienne pour l'innovation

NSERC CRSNG

Social Sciences and Humanities Research Council of Canada / Conseil de recherches en sciences humaines du Canada

Canada