Accuracies in Algorithmic Predictors of Musical Emotion

Jackie Zhou [*1], **Cameron Anderson** [†1], **Michael Schutz** [‡1,2]
[1] *Department of Psychology, Neuroscience & Behaviour*; McMaster University, Hamilton, Ontario, Canada
[2] *School of the Arts*; McMaster University, Hamilton, Ontario, Canada

1 Introduction

Music Information Retrieval (MIR) is an interdisciplinary domain where researchers develop and deploy algorithms to shed light on musical structure. Notable applications include analyzing features in expansive music databases, computational composition, and categorizing tracks' genres, artists, and instrumentation. To automate musical analyses, MIR studies often employ diverse audio feature extraction. MIRToolbox is perhaps the most widely used, having received over 1300 citations in academic publications [1]. Past research has used MIRToolbox to extract features in diverse musical works [2-3]. This includes perceptual work using MIRToolbox to evaluate features' effects. For example, one study used MIRToolbox functions to assess how mode (specific grouping of notes that contribute an emotional aspect in music), and tempo affected perceived emotion in Western classical melodies [2]. Yet, despite its wide use, the accuracy of MIRToolbox algorithms remain underexplored. According to previous work [3], only one study has directly assessed the reliability of MIRToolbox feature extraction algorithms [4]. We address this gap by comparing automated analyses of timing and mode from MIRToolbox with manual analyses of classical works.

2 Methods

2.1 Stimulus Preparation

To evaluate algorithmic consistency in analyses of major and minor excerpts, we analyzed excerpts from Chopin's *Préludes* [5], which includes 12 major and 12 minor pieces. Chopin composed all pieces for piano, providing some level of timbal consistency across different interpretations. The corpus includes performances by prominent pianists, such as Friedrich Gulda, Vladimir Ashkenazy, Martha Argerich, and Pietro de Maria—enabling comparisons of multiple interpretations. We prepared musical excerpts with Amadeus Lite, capturing the first eight full measures of each prelude performance (appending partial lead-in measures where necessary) and included a two-second fade-out.

2.2 Feature Extraction and Comparison

We analyzed timing and mode in the initial eight measures of each excerpt, excluding the two-second fade-out. We followed methods outlined in [6], codifying timing features in attacks per second (i.e., attack rate). To assess mode, we consulted the respective scores.

We used functions from MIRToolbox and assessed mode and timing using default parameters. To evaluate mode and key, we encoded information from *mirkey*, and *mirmode*. *Mirkey* estimated the tonal center of each piece, using the highest coefficient from a key strength graph. *Mirmode* predicts mode as a number between -1 and +1 (positive values identify major keys and negative values minor keys). We calculated attack rate with two functions: *mironsets* estimates the number of note attacks in the given audio input; *mirlength* evaluates the duration in seconds of each eight-measure excerpt. We calculated the predicted attack rate by dividing the number of onsets in an excerpt by its duration in seconds.

We assessed algorithmic consistency by comparing manual analyses of mode and attack rate with MIRToolbox estimates. We compared multiple interpretations of each piece to assess reliability, clarifying predictions' resilience across variation in recording quality, performance environment and timbre.

3 Results

We compared attack rate to information previously tabulated by our team for related projects [7] and compared computed estimates of mode and key to the nominal information coded in the score (i.e., the notated key).

3.1 Mode

Figure 1 depicts where predictions of key and mode align with the nominal key and mode. Green points indicate predictions consistent with both nominal mode and key. Purple points indicate predictions aligning with the nominal *mode*, but not key (e.g., a prediction of "f# minor" for pieces nominally in "eb minor"). Orange points indicate predictions consistent with neither nominal mode nor nominal key. Across the four performers, 62.5% (60/96) of key predictions were consistent with the nominal key (Argerich: 15/24=62.5%; Ashkenazy: 17/24=70.8%; De Maria: 15/24=62.5%; Gulda: 13/24=54.2%).

Notably, across all four performance interpretations, three preludes—C, E, and F minor—were incorrectly predicted to be in different keys. This was the case for almost all unaligned key predictions except one. Chopin's A minor prelude was predicted as G major across all four interpretations. This was the only piece where the algorithm consistently predicted *the same incorrect key* across all interpretations.
Although only 62.5% of predictions were consistent with the nominal key, 84.4% of mode predictions aligned with the manually analyzed major/minor mode. Mode predictions of Argerich and Ashkenazy's performances were 88% accurate, whereas predictions of de Maria's performances were 83% accurate, and Gulda's 79% accurate. For each performer, the

[*] zhouz109@mcmaster.ca
[†] andersoc@mcmaster.ca
[‡] schutzm@mcmaster.ca

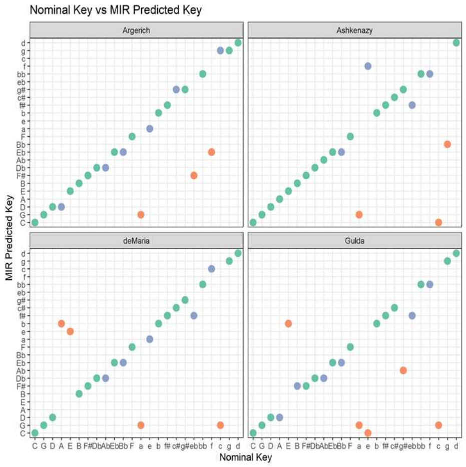

Figure 1: Mode and Key Comparison

algorithm incorrectly predicted mode in at least three, and at most five, excerpts out of 24.

3.2 Attack Rate

Figure 2 plots MIR attack rate predictions against previously extracted values. A Locally Estimated Scatterplot Smoothing curve shows MIR-predicted values begin to level off at higher attack rates, indicating worse estimation of timing in faster pieces. Despite this, automated and manual analyses of timing correlated strongly (*R* = 0.7, *p* < 0.01).

4 Discussion

Inconsistencies regarding timing information represent a technical challenge (i.e., a need for better onset detection), rather than a conceptual one. Consequently, we focus our discussion on discrepancies in mode estimates which raise numerous issues both technical and theoretical. Specifically, they raise questions regarding why multiple interpretations of the same performances differ in their assessment with *mirmode*. Although different performers will often use different tempi, a piece's modality is not generally thought to vary as a function of interpretation. Therefore, these results showing variable estimates of key and mode for different performances of the same composition suggest extraneous factors may affect the accuracy of estimates widely used in the music cognition literature as ground-truth (i.e., information assumed to be true) for perceptual experiments.

5 Conclusion

Ensuring accurate and consistent MIR algorithms is crucial as their convenience might not fully capture audio performance nuances in automated music analyses. We offer a new method for exploring tools which are widely used within the field of music cognition—yet may not be as accurate as would be assumed given their prominence.

Acknowledgments

We are grateful to Konrad Swierczek for his module which inspired part of this project. This work was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Canada Foundation for Innovation.

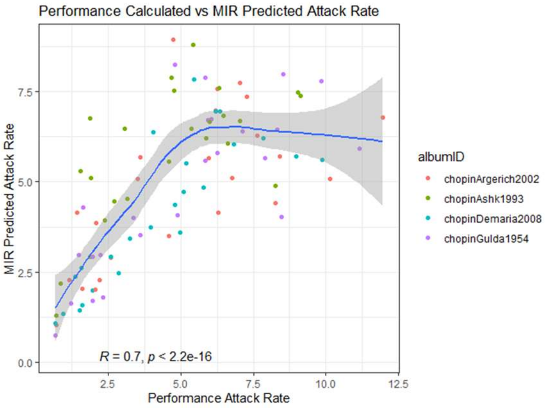

Figure 2: Attack Rate Comparison

References

[1] Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for music information retrieval. *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft Für Klassifikation EV, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*, 261–268.

[2] Quinto, L., & Thompson, W. F. (2013). Composers and performers have different capacities to manipulate arousal and valence. *Psychomusicology: Music, Mind & Brain,* **23**(3), 137–137.

[3] Lange, E. B., & Frieler, K. (2018). Challenges and opportunities of predicting musical emotions with perceptual and automatized features. *Music Perception: An Interdisciplinary Journal*, **36**(2), 217-242.

[4] Kumar, N., Kumar, R., & Bhattacharya, S. (2015, February). Testing reliability of Mirtoolbox. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS)* (pp. 710-717). IEEE.

[5] Chopin, F. (1839). *Preludes op. 28* (R. Pugno, Ed.). Schlesinger'sche Buch-und Musikhandlung.

[6] Poon, M., & Schutz, M. (2015). Cueing musical emotions: An empirical analysis of 24-piece sets by Bach and Chopin documents parallels with emotional speech. *Frontiers in Psychology*, **6**(November), 1–13.

[7] Anderson, C. J., & Schutz, M. (2022). Exploring Historic Changes in Musical Communication: Deconstructing Emotional Cues in Preludes by Bach and Chopin. *Psychology of Music*, **50**(5), 1424–1442.