Commentary on "A Computational Approach to the Detection and Prediction of (Ir)Regularity in Children's Folk Songs"

CARLOS EDUARDO CANCINO-CHACÓN
Johannes Kepler University Linz

ABSTRACT: This commentary on "A Computational Approach to the Detection and Prediction of (Ir)Regularity in Children's Folk Songs" by Mihelač, Povoh, and Wiggins reflects on the use of methods of (statistical) expectation to analyze musical structure and regularity, including potential biases of such methods, and provides some perspectives on leveraging information theoretic models of musical expectation to design cognitively plausible computational listeners of music.

ADVANCEMENTS in computational models have opened the possibility to explore hypotheses about aspects of human cognition in a controlled and reproducible way. Coupled with cognitive theories such as *predictive coding* (Clark, 2013; Friston, 2003), the power of (computational) statistical models can be leveraged to design systems that can replicate the behavior of humans, and thus, allow us to study hypotheses about the human experience. Computational models of musical understanding can allow us to test specific hypotheses about the underlying processes that are involved in the way humans compose, perform and/or listen to (or more broadly, experience) music.

In their paper, Mihelač et al. (2021) propose a method for classifying folk melodies according to their regularity using the information dynamics of music (IDyOM) model (Pearce, 2005; Wiggins et al., 2012), a well-known statistical model of musical expectation. Mihelač et al. show that such a model can be used to discover complexity in music that is considered to be "simple," such as European children's folk music. In this commentary, I would like to take a step back and reflect on the concepts of regularity and irregularity (or as Mihelač et al. use in their paper, "(ir)regularity") in musical structure, and provide a perspective on the use of data-driven statistical models to analyze musical structure.

The contributions of this paper are twofold: 1) I would like to present a critical view of the concept of regularity in musical structure, in particular connected to data-driven statistical models; and 2) I would like to present a personal vision of necessary aspects/components for comprehensive cognitively-plausible data-driven models of musical experience.[1]

Before starting a more in-depth discussion, for the sake of transparency, I would like to state that I am a computer scientist working primarily on the field of music information retrieval, and thus, this commentary comes from the point of view of someone working on developing models that capture aspects of the (human) musical experience, rather than directly studying the cognitive processes involved in music perception. Furthermore, I am a supporter of the predictive coding framework, and thus, I believe that information theoretic-derived methods can be good candidate models for modeling cognitive processes. While the question of whether a truly complete computational phenomenology that captures to all aspects of the human experience is possible (or even desirable) is a debate that goes far beyond the scope of this paper (see (Harlan, 1984) for an old-fashioned and optimistic account or (Ramstead et al., 2022) for a more recent discussion), we should ponder on the question of what are the limitations of current computational and statistical approaches to model aspects of our experience of music.

The rest of this commentary is structured as follows: I will first discuss potential issues with the concept of regularity and algorithmic bias in statistical models of music. Afterwards, I present an

integrated perspective on cognitively-plausible computational listeners. Finally, the last Section concludes this commentary.

THE GHOST IN THE SHELL[2]

"All models are wrong, but some are useful", often attributed to statistician George Box,[3] is a famous aphorism in statistics that recognizes the limitations of scientific models of describing complex (real) phenomena, while still acknowledging their usefulness. But when do the limitations of a model get in the way of its usefulness? Mihelač et al. argue that in computational models can offer *"a more objective analysis of music"* compared to traditional empirical approaches involving listeners, due to the subjective nature of musical experience. While I partially agree with this sentiment, I think it is important to be cautious with insights derived from statistical models of music. In this section, I want to focus on two aspects: 1) the concept of regularity in the context of statistical models of music and 2) the issue of representing/encoding music for computational models. In the rest of this section, I will focus on data-driven (statistical) models of music, since these models are the basis of many current cognitively-plausible models, but most of the discussion also applies to other kind of models.

Irregularity, Colonialism and Algorithmic Bias

While the concepts of regularity and irregularity are useful in many contexts, we should be careful with their application to the characterization of cultural objects like music, because of the baggage that they carry. In particular, the concept of irregularity as *"deviations"* from the established (typically Eurocentric) norms has many colonial implications. Examples of this issue can be seen throughout scholar music traditions of the 19th and 20th centuries, from Schenkerian analysis to Theodor Adorno's views on jazz (Adorno & Daniel, 1989), or European descriptions of non-Western traditions (e.g., 19th century descriptions of Afro-Brazilian music and dance, see Fryer, 2000 cited by Naveda, 2011). For a more in-depth discussion, see (Ewell, 2020) and references therein. Mihelač et al. avoid some of these issues by focusing on regularity in terms of *syntactic structure* with periodic dominant musical patterns and the relationships between these patterns. With this focus, structures that may seem syntactically regular in one piece might be perceived as irregular in another.

Still, issues with the concepts of regularity and irregularity become more complicated in the context of computational models when we consider *algorithmic bias*, systematic algorithmic errors that produce unintended and "unfair" outcomes. (Mitchell et al., 2021) identify two kinds of algorithmic biases: *statistical* and *societal*. Statistical bias refers to the mismatch in the distribution between the sample used to train[4] the model and the distribution of the real world. A typical example of this kind of bias occurs in facial recognition systems, where white males are disproportionately overrepresented in the training dataset, resulting in models that are worse at recognizing people of color (Introna & Wood, 2004; Van Noorden, 2020). In models of music experience, this kind of bias occurs when training models with datasets that overrepresent Western music (and in the context of Western classical music, Austro-German music in particular). For example, in the paper by Mihelač et al. the authors aimed for a good distribution of European folk music and included examples from 22 different countries. However, 124 of the 736 examples were from Germany, reflecting the widespread difficulty that researchers in the field have in achieving adequate representation, due in part to the limited availability of repositories containing repertoire from many parts of the world.

Societal bias, on the other hand, occurs when the model implicitly learns *social biases*[5] (e.g., gender gap, racism, etc.), even if the training dataset does reflect the real-world distribution of the data. Examples of this kind of bias have been shown, e.g., in deep learning models of natural language processing, where some state-of-the-art models have found a strong association of negative stereotypes like Muslims and terrorism (Solaiman et al., 2019). In the case of models of music, this kind of bias occurs in music recommendation systems, where the performance of these models has been shown to differ between groups of users depending on their characteristics (e.g., gender, race, ethnicity, age, etc.) (Melchiorre et al., 2021).[6]

A partial solution to some of these problems would be strictly delimiting the scope/applicability of the models–assuming that models *trained* on a particular kind of music would not generalize well enough (i.e., be applicable) to other kinds of music to make confident predictions. While this is a

somewhat unsatisfactory solution, there is a drive to develop interpretable machine learning models, and this effort is present in music research communities (Praher et al., 2021).

Representing Music in Computational Models

Computers do not "perceive" or "process" music in the same way that humans do. Typically, computers read/parse music in formats that encode just part of the information that constitutes the entire human experience of music. (Babbitt, 1965) proposed that music is a cultural construct that exists in 3 representational domains: the *graphemic* domain (the score), the *acoustic* domain (the performance, or physical realization of the score) and the *auditory* domain (the perception of the performance). These representational domains can be linked to the different roles in Kendall & Carterette's (1990) metabolic model of musical expression: we can consider musical expression as a communication process between a composer (graphemic), whose ideas are transformed into an acoustic signal by a performer (acoustic) and finally perceived and ultimately interpreted as a musical idea by the listener (auditory). More recent conceptualizations of this process describe it as dynamic and multi-directional, with the three domains often in constant interaction (e.g. Maes et al. (2014)). As argued by Wiggins et al. (2010), music as a whole cannot be effectively studied from the standpoint of pure audio analysis (i.e., acoustic domain), nor from that of pure music theory (i.e., the graphemic domain).

However, the complexity of handling all representational domains adds many layers of difficulty in developing models, and researchers need to trade-off ecological validity of the musical representation and the interpretability (and ultimate usefulness) of the models. Research on computational musicology tends to rely on symbolic representations of the music (i.e., machine readable version of the musical content, like MIDI, MusicXML or MEI), and these representations have some limits. For example, the way musical pitch is represented in the MIDI standard was designed to represent Western equal tempered music, and might not be the most appropriate way to represent micro-tonal music traditions like Turkish or Greek folk music. Furthermore, there is the general issue of quantization in music representations (not only in pitch, but in time, etc.), where complex aspects of music are discretized into "simpler" categories/scales, which usually tend to fit concepts of Western music traditions (e.g., equal temperament, isochronous beat grids, etc.) (Lenchitz, 2021). Still, there are efforts addressing some of these issues, like the work of the Music Encoding community[7], and work on multi-modal modeling of music (Simonetta et al., 2019).

TOWARDS DESIGNING COGNITIVELY-PLAUSIBLE COMPUTATIONAL LISTENERS

According to the predictive coding paradigm (Clark, 2013; Friston, 2003; Friston & Kiebel, 2009), the human brain is essentially a prediction machine, aiming to minimize the discrepancy between the organism's expectancies and imminent events. In this light, probabilistic and information theoretic models of music expectation have been shown to be adequate frameworks for developing cognitively-plausible models of music cognition (Huron, 2006; Temperley, 2007, 2019; Wiggins et al., 2010). We should, however, not wholesale discard non-cognitively plausible approaches. I am sure that many listeners of music have had their experience of music improved by learning facts about the structure of the music that would not be necessarily evident by just listening to the music (i.e., without any prior musical training). As a personal example, I collaborated with Olivier Lartillot in preparing a live visualization of Contrapunctus XIV, which is part of J. S. Bach's The Art of the Fugue BWV 1080 in a concert by the Danish String Quartet.[8] Working on this visualization helped me better understand the structure of the music and helped me to connect in an emotional level to what I used to think was a very cerebral and cold piece. Models like David Meredith's, focusing on music analysis using point-set compression (Meredith, 2016) or deep learning models of music classification have proven to be useful, and in some cases, are the state-of-the-art for many MIR tasks. In the rest of this section, I will discuss three aspects which I believe are important for designing comprehensive cognitively-plausible data-driven models of musical experience:

1. **Sequential Modeling and Considering Musical Performance.** Music is experienced sequentially by humans. Therefore, cognitively-plausible models of musical experience should be sequential, instead of just being static models that can process an entire piece

simultaneously (Widmer, 2017). Furthermore, as discussed above in "Representing Music in Computational Models", music is not only the sequence of elements in a score, but is experienced by humans through performance. While many models of expressive performance do not aim to be cognitively-plausible (Cancino-Chacón et al., 2018), there has been some success linking changes in expressive dimensions (such as timing and dynamics) to musical expectations. For example, (Gingras et al., 2016) showed that expressive timing can explain some aspects of musical structure and perceived tension, and (Cancino-Chacón, Grachten, Sears, et al., 2017) showed that using expectation features leads improves predictions of expressive tempo and dynamics. Both approaches use IDyOM, similar to the method proposed by Mihelač et al. to compute note (or chord) level information-theoretic features (Shannon entropy and information content).

2. **Embodied Music Cognition.** Including aspects from multiple representation domains should also include a component that explicitly models the embodied experience of music (Leman, 2008). As an example of this, (Cancino Chacón et al., 2014) showed that using restricted Boltzmann machines, a family of probabilistic neural networks, and a MIDI-like (i.e., a non-embodied) music representation it is possible to reproduce human perception of tonality in a probe-tone like setting. This model, however, could not capture octave equivalence, which is an important part of tonality in most (tonal) music traditions. (Agres et al., 2015) showed that using the same framework, but adding a more psychoacoustically-plausible representation of musical pitch (as a distribution of harmonics) leads to better prediction of tonal expectations, including octave equivalence.

3. **Long Term Musical Structure and Enculturation.** While there has been a lot of progress in sequential statistical models in many areas like natural language processing, modeling the long term structure of music is still a challenging task. This is partially due to the complex hierarchical nature of music, where the meaning of structural element (e.g., notes, motifs, etc.) are determined by context that might span several minutes in the past. Many of the state-of-the-art approaches for modeling musical structure take decidedly non-cognitively-plausible approaches and leverage the power of deep latent representations (Roberts et al., 2018; Wei & Xia, 2021). Unfortunately, with these kind of models it is hard to disentangle the contribution of the different aspects/elements of music. On the side of explainable cognitively-plausible models, IDyOM has an explicit long-term and a short term model. IDyOM is, however, a model of melodic expectation, and extending it for polyphonic music in a general way is not a trivial task.

Furthermore, there is the issue of *enculturation*, referring to the musical expectations that humans develop by being exposed to specific music traditions. Since this is a very important aspect of comparing regularity across music of different countries, Mihelač et al. used IDyOM's long-term model to capture the differences of each country. (Cancino-Chacón, Grachten, & Agres, 2017) used recurrent neural networks (RNNs) to model acoustic expectations of music directly from audio (using psychoacoustically-plausible frequency domain representations of audio) and then used the expectations to induce tonal knowledge in a similar manner to listeners. In this study, the RNNs were trained either on performances of Bach's Well Tempered Clavier (WTC) or all the albums by the Beatles. The results showed that models trained on the WTC performed more similarly to expert listeners (in Western classical music). The issue of enculturation is not only limited to the "score" domain, but also applies to the performance (both for performers and for listeners). This might be a reason why many listeners find Glen Gould's interpretations so transgressive.[9] I believe that comprehensive cognitively-plausible models of musical experience should include a component that models the "musical background" of listeners, and we should then consider the insights gained from such a model only in context of that background, without simply generalizing the findings to all music listeners.

CONCLUSIONS

In this commentary, I have presented some perspectives and potential issues on computational models of musical experience. New computational models and data availability have allowed for developing cognitively-plausible models of musical experience. We should, however, be careful of the scope of the models whenever we get insights from them.

As a final (meta-)commentary, I think it is very important to emphasize the need for more interdisciplinary research involving music cognition, musicology/music theory and computer science, not only in the development of computational models of musical experience. It is only this way that we will be able to pass the limitations of each field.

ACKNOWLEDGMENTS

NOTES

[1] Instead of just referring to musical listening, in this commentary I will use the term musical experience, which acknowledges other embodied non-auditory aspects (e.g., visual, sensorimotor, etc.) that contribute to our understanding/enjoyment of music (see Section 4.4.3 in Leman,2008).

[2] This is a reference to some of the philosophical questions explored in Shirow Masamune's highly influential work.

[3] A version of this aphorism is used as a section title in (Box, 1979).

[4] In the machine learning literature, *training* a model refers to the process of inferring (or learning) optimal parameters/settings of the model directly from data.

[5] In this case, social biases refers to the sociological use of the term bias, rather than the statistical one.

[6] Note that recommender systems do not typically aim to be cognitively plausible models of musical experience (Lex et al., 2021).

[7] https://music-encoding.org/about/

[8] The concert can be streamed on YouTube (the visualization can be seen sometimes on the left side of the stage): https://www.youtube.com/watch?v=S4UVJybA6ZQ&t=7263s

[9] See e.g., the heated discussions of Gould's performance of the first movement of Mozart's A major Sonata K 331 on YouTube.

REFERENCES

Adorno, T. W., & Daniel, J. O. (1989). On Jazz. *Discourse*, 12(1), 45–69.   Retrieved from http://www.jstor.org/stable/41389140

Agres, K., Cancino-Chacón, C., Grachten, M., & Lattner, S. (2015). Harmonics co-occurrences bootstrap pitch and tonality perception in music: Evidence from a statistical unsupervised learning model. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*. Pasadena, CA, USA.

Babbit, M. (1965). The Use of Computers in Musicological Research. *Perspectives of New Music*, 3(2), 74–83. https://doi.org/10.2307/832505

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building (Technical Summary Report No. ADA070213). Wisconsin University-Madison Mathematics Research Center. Wisconsin. USA

Cancino Chacón, C., Lattner, S., & Grachten, M. (2014). Developing Tonal Perception Through Unsupervised Learning. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, (pp. 195–200). Taipei, Taiwan

Cancino-Chacón, C., Grachten, M., & Agres, K. (2017). From Bach to The Beatles: The Simulation of Human Tonal Expectation Using Ecologically-Trained Predictive Models. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. Suzhou, China

Cancino-Chacón, C., Grachten, M., Goebl, W., & Widmer, G. (2018). Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities*, 5, 25. https://doi.org/10.3389/fdigh.2018.00025

Cancino-Chacón, C., Grachten, M., Sears, D. R. W., & Widmer, G. (2017). What Were You Expecting? Using Expectancy Features to Predict Expressive Performances of Classical Piano Music. In *Proceedings of the 10th International Workshop on Machine Learning and Music (MML 2017)*. Barcelona, Spain

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), (pp. 181–204). https://doi.org/10.1017/S0140525X12000477

Ewell, P. A. (2020). Music Theory and the White Racial Frame. *Music Theory Online*, 26(2). https://doi.org/10.30535/mto.26.2.4

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9), (pp. 1325–1352). https://doi.org/10.1016/j.neunet.2003.06.005

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,* 364(1521), (pp. 1211–1221). https://doi.org/10.1098/rstb.2008.0300

Fryer, P. (2000). *Rhythms of Resistance: African Musical Heritage in Brazil*. London, UK: Pluto.

Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G., & McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, 42(4), (pp. 594–609). https://doi.org/10.1037/xhp0000141

Harlan, R. M. (1984). Towards a Computational Phenomenology. *Man and World*, 17, (pp. 261–277). https://doi.org/10.1007/BF01250453

Huron, D. (2006). *Sweet Anticipation*. MIT Press. https://doi.org/10.7551/mitpress/6575.001.0001

Introna, L., & Wood, D. (2004). Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems. *Surveillance & Society*, 2(2/3), (pp. 177–198). https://doi.org/10.24908/ss.v2i2/3.3373

Kendall, R. A., & Carterette, E. C. (1990). The Communication of Musical Expression. *Music Perception*, 8(2), (pp. 129–163). https://doi.org/10.2307/40285493

Leman, M. (2008). *Embodied Music Cognition and Mediation Technology*. MIT Press. https://doi.org/10.7551/mitpress/7476.001.0001

Lenchitz, J. (2021). Reconsidering Quantization in MIR. *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*. Online

Lex, E., Kowald, D., Seitlinger, P., Tran, T. N. T., Felfernig, A., & Schedl, M. (2021). Psychology-informed Recommender Systems. *Foundations and Trends in Information Retrieval*, 15(2), (pp. 134–242). https://doi.org/10.1561/1500000090

Maes, P.-J., Leman, M., Palmer, C., & Wanderley, M. (2014). Action-based effects on music perception. *Frontiers in Psychology*, 4, 1008. https://doi.org/10.3389/fpsyg.2013.01008

Melchiorre, A. B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., & Schedl, M. (2021). Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5), 102666. https://doi.org/10.1016/j.ipm.2021.102666

Meredith, D. (2016). Analysing Music with Point-Set Compression Algorithms. In D. Meredith (Ed.), *Computational Music Analysis* (pp. 335–366). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-25931-4_13

Mihelač, L., Povoh, J., & Wiggins, G. A. (2021). A Computational Approach to the Detection and Prediction of (Ir)Regularity in Children's Folk Songs. *Empirical Musicology Review*, 16(2), 205-230. https://doi.org/10.18061/emr.v16i2.8245

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Naveda, L. (2011). Gestures in Samba: A cross-modal analysis of dance and music from the Afro-Brazilian culture. Unpublished doctoral dissertation. University of Ghent, Belgium. https://doi.org/10.21825/af.v24i1.18036

Pearce, M. T. (2005). The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition. Unpublished doctoral dissertation. City University London, UK.

Praher, V., Prinz, K., Flexer, A., & Widmer, G. (2021). On the Veracity of Local, Model-Agnostic Explanations in Audio Classification: Targeted Investigations with Adversarial Examples. IN *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, Online. (pp. 531–538).

Ramstead, M. J. D., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G., Smith, R., Dumas, G., Lutz, A., Friston, K., & Constant, A. (2022). From Generative Models to Generative Passages: A Computational Approach to (Neuro) Phenomenology. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00604-y

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. Stockholm, Sweden.

Simonetta, F., Ntalampiras, S., & Avanzini, F. (2019). Multimodal Music Information Processing and Retrieval: Survey and Future Challenges. In *Proceedings of the 2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, (pp. 10–18). Milan, Italy. https://doi.org/10.1109/MMRP.2019.00012

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. (2019). Release Strategies and the Social Impacts of Language Models. *CoRR*, abs/1908.09203. Retrieved from http://arxiv.org/abs/1908.09203

Temperley,   D.   (2007).   *Music   and   Probability*.   MIT   Press. https://doi.org/10.7551/mitpress/4807.001.0001

Temperley, D. (2019). Uniform Information Density in Music. *Music Theory Online*, 25(3). https://doi.org/10.30535/mto.25.2.5

Van Noorden, R. (2020). The Ethical Questions That Haunt Facial-Recognition. *Nature*, 587, (pp. 354–358). https://doi.org/10.1038/d41586-020-03187-3

Wei, S., & Xia, G. (2021). Learning Long-Term Music Representations via Hierarchical Contextual Constraints. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*. Online.

Widmer, G. (2017). Getting Closer to the Essence of Music: The Con Espressione Manifesto. *ACM Transactions on Intelligent Systems and Technology*, 8(2), (pp. 1–13). https://doi.org/10.1145/2899004

Wiggins, G. A., Müllensiefen, D., & Pearce, M. T. (2010). On the non-existence of music: Why music theory is a figment of the imagination. *Musicae Scientiae*, 14(1), (pp. 231–255). https://doi.org/10.1177/10298649100140S110

Wiggins, G. A., Pearce, M. T., & Müllensiefen, D. (2012). Computational Modeling of Music Cognition and Musical Creativity. In R. T. Dean (Ed.), *The Oxford Handbook of Computer Music*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199792030.013.0019