Creating and Evaluating Challenge in Generative Music Education Systems

Filippo Carnovalini [1], and **Geraint A. Wiggins**[1,2]
[1]Vrije Universiteit Brussel
Pleinlaan 9, 1050 Brussel, Belgium
[2]Queen Mary University of London
Mile End Road, London E1 4NS, UK
filippo.carnovalini@vub.be

Abstract

Providing personalized learning paths increases the effectiveness of education. In the context of music education, personalization is often limited by the use of standard curricula for all students, as preparing or selecting ad-hoc learning material is non trivial and expensive. Computationally Creative systems could potentially help create personalized material, but to do so it would be necessary for them to be able to evaluate internally the degree of difficulty of the generated music. Automatic difficulty estimation has been proposed by Music Information Retrieval Research, but those descriptions of difficulty are however absolute and not personalized. We believe instead that different students will find different pieces more or less challenging than others. We propose therefore to use information theory metrics to estimate the level of *challenge* based on prior experience of the learners, by employing Idyoms, a system for the cognitive modeling of music. In two experiments we show that the system does model differences in levels of preparation when facing difficult pieces, and that it can account for different learning paths leading to different levels of challenge based on the appropriateness of the pieces known to the student, showing this approach could enable a Creative system to estimate the perceived difficulty of newly generated pieces and therefore improve the personalization of music education.

Introduction

Providing a personalized learning experience can significantly increase the effectiveness of education (Bloom 1984). Music teaching is an activity that offers some personalization afforded by the presence of personal teachers during one-to-one lessons. However, the degree of personalization is limited by the use of standard teaching materials and fixed curricula of pieces to learn, that typically apply to all students in a school (or in some cases even fixed by regional or national level curricula) (Carnovalini, Roda, and Wiggins 2023; Carnovalini, Espírito Santo, and Wiggins 2025).

This work is funded by the European Union, HORIZON-MSCA-2022-PF-01 Project ID 101108690 (CALIOPE). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the REA can be held responsible for them.

We propose that Computational Creativity can help achieve a higher degree of personalization. We have witnessed notable progress in music generation systems, and we posit that such systems could be employed for the creation of personalized exercises. This represents an interesting intersection of requirements that is underexplored in Computational Creativity, as the system would need to create sheet music that is a) not too hard for the student's current skill level, but also b) challenging enough to have educational benefits, and finally c) pleasant to play and listen to. The sum of a) and b) can be described as laying in the Zone of Proximal Development (Chaiklin 2003), which can be seen as an ideal area of a conceptual space that, *for a specific student* has the ideal amount of difficulty to enable learning without triggering frustration. It is worth noting that no two learners are the same, and therefore the shape of this area will always be different even for two students in the same level or class, as one may find certain skills easier or harder than the other (for example, in the same class of beginner pianists one may be faster in playing melodies, but another may be better at playing chords in the left hand). The need for personalization requires the system to be able to create novel pieces each time, which of course must also show sufficient quality, suggesting that this is indeed a prime use case for Computational Creativity, whereas Generative AI is currently capable of creating music, but is almost impossible to steer to creating pieces that match these strict educational requirements (Carnovalini et al. 2026).

There has been research in Music Information Retrieval systems to estimate and/or classify the difficulty of a musical piece, given the sheet music (Ramoneda et al. 2022), which could ideally be used as an internal evaluation for a creative system. However, this work aims to estimate difficulty as an absolute value, i.e., describing how hard a piece is compared to other pieces, *for most people*. We argue that to obtain proper personalization, a system for evaluating difficulty should try and make a cognitive model of the learner. In this study, we explore a system for the estimation of *challenge*, defined as *how difficult a certain piece will be to a certain person at a certain stage in their development* (or in other words, how much they will struggle to learn it). Exploring how effective such a measure is will enable further exploration in a creative system that generates music taking *challenge* into account.

A Cognitive Model

Idyoms

For the current investigation, we utilize Idyoms[1], a Julia implementation of IDyOM (Information Dynamics of Music), a cognitive model developed by Marcus Pearce (Pearce 2005) for the statistical modeling of the perception of melodic sequences. Idyoms is able to extract statistical patterns based on example sequences by creating a variable-order Markov model (Conklin and Witten 1995) and is then able to estimate the likelihood of unseen sequences by means of the Prediction by Partial Matching (PPM) algorithm (Cleary and Witten 1984), which combines the likelihoods of the different orders. The system also allows the definition of different *viewpoints*, i.e., different feature dimensions to be used for the estimates (e.g., pitch and duration of notes). The model consists of two components, which may be used separately or together. The Short-Term Model (STM) bases its prediction only on elements that are within the current sequence (i.e., within the same song/piece), incrementally updating the statistics as the sequence is being processed. The Long-Term Model (LTM), in contrast, bases its statistics on a corpus of previously-analyzed sequences, offering the option to update statistics as the new sequence is being processed or to use just the prior knowledge. In both cases, the model allows computation and analysis of the Information Content of each event in a sequence, which is an indication of how unexpected that event is based on the elements that precede it. It is computed as follows:

$$h(x) = -log_2(p(x)) \tag{1}$$

where $p(x)$ is the probability of the event according to PPM. Events with low probability will have high information content and vice versa. Information content has been shown to model aspects of human perception and cognition of musical sequences very closely (Pearce 2005; Pearce and Wiggins 2006; 2012; Hansen and Pearce 2014).

Information Content, Difficulty, and Struggle

Using Information Content or the related metric, Entropy, for difficulty estimation is an established idea (Ramoneda et al. 2024; Sébastien et al. 2012; Verstraeten 2025). These metrics can give an indication of how structurally complex and varied a piece of music is, and these aspects are related to the perceived difficulty in performing that piece. However, research that has used these measures has typically done so by computing the metrics only within a single piece (equivalent to Idyoms' STM approach) and/or using only zeroth-order probability estimations, which therefore do not capture the dynamic structure of the music. Both these two limitations are reasonable for the goal of describing the variance of the piece's elements (e.g., how spread are the pitches, or how varied are the rhythmic figures), but fall short when it comes to describe how difficult a piece will be to a certain learner based on their learning path, i.e., what we call *challenge*.

In order to model this personalized feature, we investigate the following approach: we expose the *Long*-term model of Idyoms to pieces that have already been studied by the learner, and compute the (average) Information Content of the new piece that is to be studied. As mentioned above, Information Content is correlated with surprise and unexpectedness. Our rationale is that, by modeling the unexpectedness of the new piece, we also model how much the learner will struggle to learn new sequences and techniques, that are not already part of their musical and technical vocabulary, which is roughly modeled by the LTM.

Idyoms has been shown effective as a model of the perception of music. Here, however, we are using it as a model of performance. We believe that the overall statistical modeling that Idyoms is based upon will also work as a model for the cognitive load that a musician must process when performing a new piece. While the present experiments are some of the first to test this hypothesis, we posit that expecting this well-established cognitive model of perception to highly correlate with performance is not an unreasonable assumption. The results of the following experiments corroborate this hypothesis, although at this stage it remains to be tested in future work with human participants.

Experiments

Using Information Content to Model Progression

Method. Our first experiment[2] is intended to assess the viability of general ideas outlined in the previous section. We model a student learning through a progressive exercise book, *Mikrokosmos*, a well-known piano book written by Béla Bartók (Bartók 1987), which contains pieces of progressive difficulty going from beginner to piano virtuoso. The book is available in digital format in the Mikrokosmos-Difficulty Dataset (MKD)[3], where the pieces are divided into three difficulty categories: *easy*, *medium*, and *hard*. Our experiment works as follows: we trained the Long-term model (LTM) of Idyoms with only the *easy* category and the second using both the *easy* and *medium* subsets of the dataset using the combination of Pitch and duration of the note events as the data representation[4]. We only used the right hand information in this experiment, in order to keep the alphabet size limited. We then predict the information content of the hard pieces based on the two learned models. The idea is to try to model a student that tried to approach the hard pieces only studying the easy pieces, and one who went through both easy and medium pieces before approaching the hard ones. The hypothesis is that the level of challenge (and therefore the average Information Content) is significantly lower for the latter, as the student we model has a better preparation for the hard pieces.

Results and Discussion. Figure 1 shows box plots detailing the distribution of the mean information content of the

[1] https://github.com/nick-harley/Idyoms

[2] The code for both experiments is available at: www.github.com/project-caliope/caliope_experiments

[3] https://zenodo.org/records/6092709

[4] *Viewpoint* in IDyOM and Idyoms terminology (Conklin and Witten 1995; Pearce 2005).

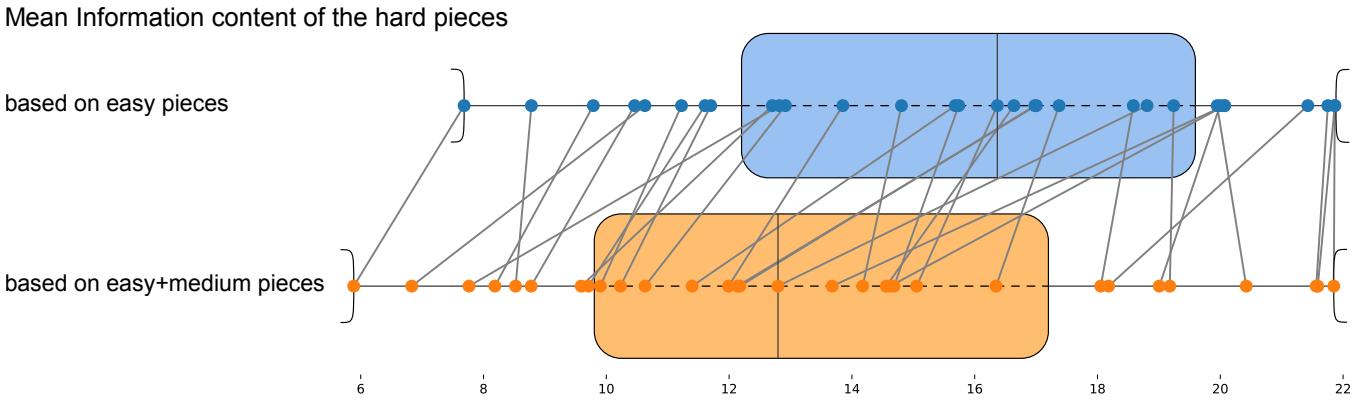

Figure 1: Paired box plot of the mean information content of hard pieces predicted by an Idyoms model trained on easy pieces and one trained on easy and medium pieces, with lines pairing values for the single pieces.

hard pieces, as predicted by the two models. The distribution is almost equally spread over a large range, but what is more interesting than the distribution itself is to see how the Mean Information Content changed for each piece based the model having been exposed to solely easy pieces or easy and medium pieces. To test that, we performed a paired nonparametric test (the Wilcoxon signed-rank test, graphically represented as the dots and lines in Figure 1), which confirmed the differences are significant ($T = 6.0$; $p = 1.3\text{e-}8(^{***})$; two-sided alternative rejected).

This first result indicates that Information Content can indeed model different levels of preparation, suggesting that it is indeed possible to give a personalized account of difficulty, or challenge. It is however possible to argue that this experiment offered little personalization, as both models came from the same dataset and have a non negligible overlap, and the results are only the effect of having more data in the easy+medium category. We address these reasonable points in the following section.

Personalization of Learning Paths

Method. In order to test how much our model can account for different learning paths, we performed a second experiment. We imagine two learners wanting to approach the same difficult piano piece, Alban Berg's *Piano Sonata* (Op. 1). This sonata is a challenging piece that does not rely on classical tonal theory, but is instead partly inspired by Schoenberg's (of whom Berg was a composition student) serialist ideas. The first student we envision goes through a more classical curriculum: we model this student with an Idyoms long-term model trained on the entirety of the Mikrocosmos dataset. The second, on the other hand, decides to train on fewer pieces, but ones which are more relevant to Berg's sonata. We collected a small dataset of 11 serialist pieces by Arnold Schoenberg and Anton Webern, on which we train a different instance of Idyoms' LTM. The first student has seen more pieces (some of which require high skill to master), but our hypothesis is that the second student, being more used to the relevant style of music, will struggle less on the piece, resulting in lower Information Content.

Results and Discussion In the previous experiment we compared the *Mean* Information Content of various pieces. Since we only analyze one piece in this second experiment, we compare the distribution of the Information Content of each event in the sonata, as predicted by the two models. Figure 2 shows the two distributions as violin plots. A Kruskal–Wallis H test confirms that the distributions are significantly different ($H = 15.82$; $p = 0.00007(^{***})$).

This result confirms that our model can account for different learning paths in a personalized manner, and that it is not merely an indication of having more data in the learned model. While in general adding more data to a model is expected to lower the Information Content, the relevance of the learned pieces is more important than the quantity of data. This is the expected behavior for the model: students who study more will generally struggle less, but the pieces they studied must be coherent with the piece they want to prepare to lower its challenge level.

Conclusions

We presented a problem for which Computational Creativity can offer help, namely the creation of personalized music exercises, and an approach to the estimation of *challenge*, defined as how difficult a certain piece will be to a certain person at a certain stage in their development, a key feature of such a creative system. This concept can become part of the internal evaluation for a Computational Creativity system creating music meant for learners, but could also help in recommending existing study material or automatically constructing personalized learning paths. Through two experiments we demonstrate that this approach is promising in providing a more personalized view on difficulty than what previously proposed.

There are however some limitations of the current study, which call for further work. Firstly, we have used a measure (Information Content) that is known to be significant for perception, but has not been validated for performance. The results suggest that indeed it could be appropriate for modeling performance cognition as well, but this remains to be tested.

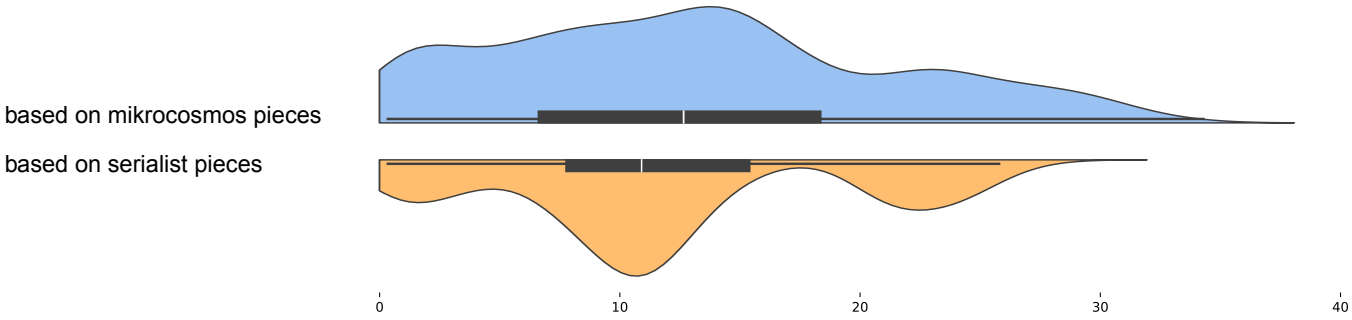

Figure 2: The distribution of the information content of all Berg's sonata's notes, as computed by a Idyoms model trained on all the MKD dataset and one trained on just serialist music.

The experiments could be repeated with more data and by using the information related to both hands, and possibly including additional information in our viewpoints such as fingering annotations or dynamics. It would also be interesting to model more closely existing piano curricula from music schools, to see how well this approach can relate to real-world scenarios, and finally, once a higher degree of maturity is reached, it would need to be validated with actual learners. Moreover, while the approach is promising, further work is needed to make it viable within a creative system so that it can be used by the final users, namely piano teachers and music students.

Despite the limitations, we believe the proposed approach opens interesting perspectives for AI-supported music education and for Computational Creativity alike, as it requires creating music that is novel and of quality under a personalized and educational point of view, rather than being generically a new piece of satisfactory quality.

Author Contributions

FC ideated the study and performed the experiments. GW provided supervision. Both collaborated in the experiment design and in the writing and revising of this article.

Acknowledgments

We wish to thank Louis Verstraeten for the prior work that opened the way for this study, Nick for the help with Idyoms, and Sam for his suggestion about serialist music, and all of the CCLab for help and support.

References

Bartók, B. 1987. *Mikrokosmos, Vol. 4*. Boosey & Hawkes.

Bloom, B. S. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13(6):4–16.

Carnovalini, F.; Scofield, S.; Verstraeten, L.; and Wiggins, G. 2026. The possibilities of personalized music education through exercise generation [version 1; peer review: 1 approved with reservations, 1 not approved]. *Open Research Europe* 6(108).

Carnovalini, F.; Espírito Santo, L.; and Wiggins, G. A. 2025. Personalized Music Education: A Systematic Review of AI Generation Methods. *IEEE Access* 13:207433–207447.

Carnovalini, F.; Roda, A.; and Wiggins, G. 2023. Interactive Generation of Musical Corpora for Piano Education: Opportunities and Open Challenges. In *Proceedings of the 15th International Conference on Computer Supported Education - Volume 1: CSME*, 412.

Chaiklin, S. 2003. The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context* 1(2):39–64. Publisher: Cambridge, UK: Cambridge University Press.

Cleary, J., and Witten, I. 1984. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications* 32(4):396–402.

Conklin, D., and Witten, I. H. 1995. Multiple viewpoint systems for music prediction. *Journal of New Music Research* 24:51–73.

Hansen, N. C., and Pearce, M. T. 2014. Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology* 5(1052).

Pearce, M. T., and Wiggins, G. A. 2006. Expectation in melody: The influence of context and learning. *Music Perception* 23(5):377–405.

Pearce, M. T., and Wiggins, G. A. 2012. Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science* 4(4):625–652.

Pearce, M. T. 2005. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. Dissertation, Department of Computing, City University, London, London,UK.

Ramoneda, P.; Tamer, N. C.; Eremenko, V.; Serra, X.; and Miron, M. 2022. Score Difficulty Analysis for Piano Performance Education based on Fingering. In *ICASSP 2022*, 201–205. Singapore, Singapore: IEEE.

Ramoneda, P.; Eremenko, V.; D'Hooge, A.; Parada-Cabaleiro, E.; and Serra, X. 2024. Towards Explainable and Interpretable Musical Difficulty Estimation: A Parameter-efficient Approach. In *ISMIR 2024*.

Sébastien, V.; Ralambondrainy, H.; Sébastien, O.; and Conruyt, N. 2012. Score analyzer: Automatically determining scores difficulty level for instrumental e-learning. In *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 571–576.

Verstraeten, L. 2025. Difficulty Analysis and Generation of Piano Exercises. Master's thesis, Vrije Universiteit Brussel, Brussels, Belgium.