BEYOND THE SAMPLE ALONE: FAMILIARITY AND CONTEXT IN HUMAN EVALUATION OF GENERATED MUSIC

Anonymous Authors
Anonymous Affiliation
anonymous@ksmi.kr

ABSTRACT

Human evaluation of generated music is often modeled as a direct function of sample-level representations, yet it remains unclear how much of subjective judgment can actually be explained by the generated sample alone. To investigate how listeners evaluate generated music, we conducted a continuation-based listening test using 200 piano continuations generated from shared primers across ten classical composer styles, rated by 190 listeners on six rating dimensions and Familiarity. Sample-only predictions captured general rating trends at the aggregate level, but performed poorly for individual listeners and failed to generalize to unseen composers. Variance decomposition further revealed that the unexplained portion of ratings was largely attributable to individual listener differences. We therefore examined Familiarity as a key interpretable listener-side factor. While Familiarity showed a strong style-level baseline across composer contexts, its explanatory role for ratings was driven primarily by continuation-level familiarity, and it also modulated how relative musical change was evaluated. These results suggest that evaluation of generated music is layered: beyond signal-level properties, human judgments reflect structured listener-side responses that are not fully captured by coarse similarity or sample-only representations.

1. INTRODUCTION

Evaluating generated music remains a central challenge in Music Information Retrieval (MIR). Recent work has largely framed this problem as one of predicting human judgments from sample-level representations, whether through learned embeddings, per-sample metrics, or benchmark datasets with listener ratings [1–3]. This framing is useful for comparing systems at scale, but it also raises a basic question of how much of human evaluation can actually be explained by the generated sample alone.

This question is particularly important for music, where judgments are often context-dependent and listener-dependent. Prior studies in MIR have shown that human annotations can exhibit structured disagreement rather than pure noise [4–6]. Related research in music cognition further suggests that listener-dependent factors such as familiarity, expectation, and processing fluency can shape aesthetic response [7, 8]. Among these, familiarity is especially relevant because it provides an interpretable link between stylistic context and listener response. Yet familiarity has rarely been modeled in a structured way in generated music evaluation. We therefore treat familiarity as one listener-side explanatory variable and ask whether it helps account for part of the variation that sample-only representations leave unexplained.

In this paper, we address this gap through a continuation-based evaluation design. Instead of comparing unrelated clips, we construct a dataset around shared composer-style primers. This makes it possible to examine evaluation at three levels: across stylistic contexts, across continuations within the same context, and across listeners rating the same continuation. More specifically, we use this setting to connect three questions that are often studied separately: what sample-only prediction captures, what structured variation remains unexplained, and whether part of that variation can be interpreted through familiarity and context-relative change.

2. EXPERIMENTAL SETUP

We conducted a continuation-based listening experiment using piano continuations generated from shared primer excerpts. Ten four-bar piano primers were selected from MAESTRO dataset [9] to provide distinct stylistic baselines drawn from classical composers while avoiding highly recognizable excerpts. For each primer, 20 continuations were generated with a pre-trained MusicTransformer [10], yielding a final set of 200 continuations. This controlled shared-context design allowed us to distinguish broad style-level effects from continuation-level variation while minimizing inter-model variability and broader genre-level preference confounds.

Subjective ratings were collected through an online listening study with 190 participants. Each participant rated 20 samples in a balanced design such that two continuations were heard from each primer context. After listening, participants rated each sample on 7-point Likert scales for Overall, Melodiousness, Creativity, Rhythmicity, Coherence, Naturalness, and Familiarity. No information was provided about the use of AI generation.

For signal-based modeling, we extracted MERT [11] audio embeddings from the full sequence, primer, and continuation segments. These were projected into a shared low-dimensional space via PCA, yielding full-sequence

and relative-change (continuation minus primer) components for use in subsequent models. Analyses combined ridge regression for sample-only prediction with mixed-effects models for variance decomposition, familiarity decomposition, and interaction testing.

3. RESULTS

Human ratings were multidimensional, with a consistent evaluative structure across rating dimensions. Overall was broadly aligned with Melodiousness, Coherence, and Naturalness, while Creativity and Familiarity were more distinct from the main rating axis. This motivated treating Familiarity not simply as another rating dimension, but as a listener-side variable potentially explaining variation across ratings.

We first tested whether ratings could be predicted from the continuation sample alone. As shown in Figure 1, sample-only models captured part of the average rating tendency: predictions aligned moderately with aggregated sample ratings, but performance dropped substantially when the same predictions were compared with individual listener ratings. Generalization also deteriorated sharply under composer-holdout evaluation. Thus, sample-level representations recovered some broad ranking of continuations, but not the fuller structure of ratings across listeners and stylistic contexts.

To characterize what remained unexplained, we fit null mixed-effects models that decomposed variance across composer context, continuation identity within composer, listener identity, and residual variance. Variance decomposition showed that listener-level variance accounted for a substantial share of total variance across the six rating dimensions. In Overall ratings, for example, listener identity explained a larger share of variance than either composer context or continuation identity alone. Importantly, this should not be interpreted as evidence for Familiarity alone: listener-side variance may reflect multiple factors, including familiarity, preference, musical background, or response style. However, it does show that the unexplained portion of evaluation is structured rather than purely noisy.

We next examined how Familiarity itself was patterned across contexts. Familiarity showed a strong style-level baseline, and composer contexts that received higher Overall ratings also tended to be rated as more familiar. However, it was not reducible to composer context alone. Within the same stylistic context, continuations still varied substantially in perceived Familiarity, and the Familiarity–Overall relation differed across composers.

We then asked whether this variation in Familiarity helped explain the other ratings. As summarized in Table 1, adding raw Familiarity improved model fit substantially over the sample-only baseline, and decomposed models showed that continuation-level familiarity provided the clearest contribution across most rating dimensions. Thus, raw Familiarity improved prediction, but its contribution became more interpretable when decomposed into style-level and continuation-level components.

Finally, Familiarity also shaped how relative musical change was evaluated. Interaction analyses revealed that the relationship between relative-change axes and ratings depended on continuation-level familiarity. In some cases, the same change dimension tended to be evaluated more positively when familiarity was high than when it was low; in others, familiarity attenuated the strength of the change effect. These patterns suggest that Familiarity influences not only rating magnitude but also how listeners interpret the continuation relative to its preceding context.

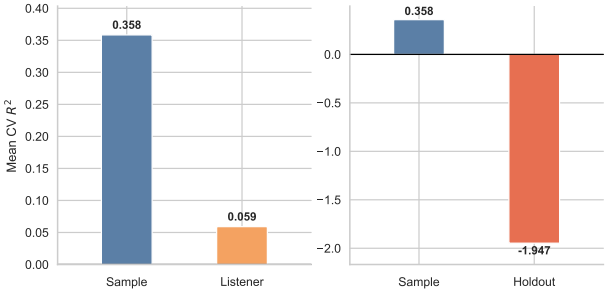

Figure 1. Performance of the sample-only prediction. Left: mean cross-validated R^2 at the sample and listener levels. Right: mean R^2 for sample-level CV and composer holdout.

Outcome	Sample	+Raw	+Decomp	Key component
Overall	0.09	0.26	0.27	continuation
Melodiousness	0.09	0.29	0.31	continuation
Creativity	0.03	0.03	0.07	listener
Rhythmicity	0.07	0.21	0.22	continuation
Coherence	0.11	0.31	0.31	continuation
Naturalness	0.07	0.24	0.24	continuation

Table 1. Marginal R^2 of sample-only, raw-Familiarity, and decomposed-Familiarity models across the six rating dimensions. The final column indicates the familiarity component showing the clearest unique contribution in nested component comparisons.

4. DISCUSSION AND CONCLUSION

Taken together, these findings suggest that generated music evaluation is layered. Sample-only representations capture part of the broad rating pattern, but a substantial portion of rating variation remains structured across contexts, continuations, and especially listeners. Familiarity provides one interpretable listener-side factor that explains part of this remaining variation.

Importantly, we do not argue that Familiarity exhaustively explains human evaluation. Rather, our results show that listener-side structure matters and that continuation-based designs provide a useful way to study it. For MIR, this implies that improving evaluation is not only a matter of learning better sample-level representations, but also of accounting for how listeners interpret musical events relative to stylistic context. This provides a useful first step toward context-sensitive and user-aware evaluation frameworks for generated music.

5. References

[1] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin, "Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[2] G. Ma, Y. Xia, J. Yao, H. Xue, H. Liu, S. Wang, H. Liu, and L. Xie, "The icassp 2026 automatic song aesthetics evaluation challenge," *arXiv preprint arXiv:2601.07237*, 2026.

[3] A. Lerch, C. Arthur, N. Bryan-Kinns, C. Ford, Q. Sun, and A. Vinay, "Survey on the evaluation of generative models in music," *ACM Computing Surveys*, vol. 58, no. 4, pp. 1–36, 2025.

[4] N. L. Masclef, A. Vaglio, and M. Moussallam, "User-centered evaluation of lyrics-to-audio alignment." in *ISMIR*, 2021, pp. 420–427.

[5] A. Flexer, T. Lallai, and K. Rašl, "On evaluation of inter-and intra-rater agreement in music recommendation," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, 2021.

[6] J. S. Gómez-Cañón, E. Cano, P. Herrera, and E. Gómez, "Joyful for you and tender for us: The influence of individual characteristics and language on emotion labeling and classification," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020, pp. 853–860.

[7] A. Korkor, V. Noreika, C. Di Bernardi Luft, and M. Pearce, "Relationships between surprise, liking, and error perception in musical listening." *Psychology of Aesthetics, Creativity, and the Arts*, 2025.

[8] E. Schubert, D. J. Hargreaves, and A. C. North, "A dynamically minimalist cognitive explanation of musical preference: is familiarity everything?" *Frontiers in psychology*, vol. 5, p. 38, 2014.

[9] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[10] H. C.-Z. Anna, V. Ashish, U. Jakob, S. Ian, H. Curtis, S. Noam, D. Monica, E. Douglas *et al.*, "Music transformer: Generating music with long-term structure," *arXiv preprint*, 2018.

[11] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.