Proceeding of ICIHCS 2025 Symposium: The Dialogue Between Tradition and Innovation in Language Learning
DOI: 10.54254/2753-7064/2025.HT28341

Artificial Intelligence-based Music Evaluation: Progress, Challenges and Prospects

Boyuan Yao

Crescent School, Toronto, Canada
yaoboyuan0215@gmail.com

Abstract. Music is among the most expressive and subjective of art forms, which makes its evaluation uniquely difficult. Traditional rule-based methods that rely on objective features such as pitch, rhythm, and audio quality often fail to capture the richness of human perception, particularly when it comes to subjective qualities like creativity, emotion, and cultural context. With the rise of Artificial Intelligence, researchers have begun exploring new approaches for music evaluation that better align with human judgment. This paper surveys the three major strands of work in this emerging field: human-grounded datasets for preference learning, embedding- and distribution-based metrics, and learned predictors and foundation-model evaluators. Each category is examined in detail, with representative works introduced alongside their respective strengths and limitations. The paper then compares these approaches, identifies challenges common across the field, and discusses possible future directions. By analyzing existing research and future prospects, this paper highlights the potential of Artificial Intelligence (AI) to transform music evaluation into a more reliable, inclusive, and scalable process.

Keywords: Artificial intelligence, music evaluation, human-grounded datasets.

1. Introduction

Music is one of the most important art forms of human society, with some believing in its history to possibly predate spoken language [1]. While the main pursuit of music composition in past centuries has always been that of beauty or interest, the means of such creation has evolved alongside technological advancements. After the invention of the computer, the field of computer music quickly expanded, with people exploring fields such as algorithmic music and FM synthesis [2].

As one of the most significant technological breakthroughs of recent times, Artificial Intelligence (AI) has already had an incredible impact in day to day life. AI-generated content can be found in domains of written text, image, videos, websites, and it has reached the field of music as well. Instead of using rigid rule-based programs to compose music, various tools based on Transformers [3] and Diffusion [4] models are now available for consumer use, such as Suno and Stable Diffusion, capable of generating outputs based on text or audio prompts inputted by users. Beyond music generation, AI tools have also been built for purposes such as music recommendation, music transcription, music arrangement, etc.

DOI: 10.54254/2753-7064/2025.HT28341

Despite these innovative products and research, an area that requires further exploration is music evaluation. Indeed, just like many other art forms, music is highly subjective; while some might find a certain musical piece to be soothing and enjoyable, others might find it to be boring and uninspiring. Because of this, an objective, universal music evaluation tool would be massively helpful. Not only can it be used to categorize and sort the vastly expansive song libraries on current music platforms, but it can also be utilized for training music generation models as a feedback function.

Earlier attempts at music evaluation focus on rules-based methods such as evaluating objective audio quality, or musical features including pitch, harmony, rhythm, etc., assigning weights to each of them and calculating a comprehensive score. Such methods have obvious limitations: they fail to consider the subjective musicality of songs and human preferences. This led to researchers attempting to use Artificial Intelligence as means of music evaluation. Yao et al., for example, aimed to create a dataset of music and corresponding human ratings in various categories to aid the training process of AI music evaluation [5]. Another example is Prompting Audio-Language Models (PAM) for Audio Quality Assessment, which takes a different approach by directly leveraging large audio-language models, designing prompts to guide a pretrained model in evaluating aspects of audio quality.

This paper aims to examine current progress in the field of AI-based music evaluation. Furthermore, this paper will discuss the potential for improvements in each of these approaches, suggesting possible directions for future research in this area.

2. Application of AI models in music evaluation each work

Existing research work in the field of AI based music evaluation can be organized into a few different categories based on their goal and methods.

2.1. Human-grounded datasets for preference learning

Since it is widely established that preferences for music are highly subjective, one major approach in AI music evaluation is building large-scale datasets of music audio files and their corresponding human judgment scores, which could then be used as the foundation for training AI evaluators. These datasets capture human perceptions of music, including qualities like musical pleasantness, creativity, and naturalness, which rule-based metrics simply cannot reliably quantify.

SongEval is an influential attempt in this direction [5]. The authors collected more than 140 hours of music samples generated by models, which are then rated by 16 musically trained professionals. Other than the large size of the dataset, another notable feature for SongEval is that rather than focusing on a single holistic rating for a piece of music, SongEval utilizes multi-dimensional annotations, which includes overall coherence, memorability, naturalness of vocal breathing and phrasing, clarity of song structure, and overall musicality. This structured approach makes annotator ratings as consistent as possible, while enabling evaluators to learn richer representations of human perception. This dataset has become an important benchmark for researchers developing AI evaluators, as it directly links generated music to human-grounded scores.

ARMOR represents another key contribution, aimed at the “meta-evaluation” problem [6]. Rather than producing music using AI models, ARMOR provides a dataset of music samples paired with human judgments that is explicitly designed to test the validity of automatic metrics. This makes ARMOR unique: instead of being training data for models, the human judgments act as a ground truth against which the performance of objective evaluators (such as Fréchet Audio Distance) can be

DOI: 10.54254/2753-7064/2025.HT28341

judged. This design acknowledges benchmarking is not only necessary for generation but also evaluation itself, and human preferences should be considered the final standard.

Finally, Music Arena introduces a live, renewable method for crowdsourcing human evaluations for music [7]. Instead of static annotated datasets, it is an open platform where any listener can use to compare two text-generated music samples and vote for the one they prefer. This pairwise preference collection produces ongoing, scalable feedback aligned with real human listening habits. Unlike static benchmarks, Music Arena allows evaluators to adapt dynamically to new AI music generation models as they emerge.

Together, these datasets and platforms illustrate how human annotations serve as the cornerstone of AI music evaluation. They highlight the importance of preference learning and the practical trade-offs between static datasets, meta-evaluations, and live collection platforms.

2.2. Embedding- and distribution-based metrics

Another strand of research attempts to bypass the need for continuous, tedious human annotation by using pretrained embedding spaces and statistical distribution to approximate perceptual similarity. These metrics assume that, for most cases, if two music distributions are similar in a machine-learned embedding space, their perceived quality should also be similar to human ears.

Fréchet Audio Distance (FAD), introduced by Kilgour et al., adapts the widely used Fréchet Inception Distance (FID) from computer vision to the field of computer audio [8]. At its core, FAD compares the distributions of real and generated audio embeddings extracted from a pretrained model. Because it does not require per-sample human ratings like examples from the previous section, FAD enables scalable, automatic evaluation of generative systems. The metric has been widely applied in music and speech synthesis, becoming a de facto standard in generative audio evaluation. Despite its popularity, it is important to note that its reliability is heavily dependent on the choice of embedding model and the dataset used to estimate statistics. For instance, if the embedding model is trained on speech rather than music, FAD may poorly capture musical structure.

More recently, MAUVE Audio Divergence (MAD) has been proposed as an improvement over FAD [9]. Inspired by the MAUVE metric in natural language generation, MAD computes divergences between distributions of generated and real audio embeddings in a way that captures not just mean and covariance but also higher-order differences. Early results suggest MAD correlates more strongly with human judgments of musicality than FAD, especially for long-form generative music. Its downside, however, is computational complexity and sensitivity to embedding dimensionality, which makes it less straightforward to apply at scale.

Another line of work investigates how FAD should be applied to music specifically. One toolkit study explored different embedding extractors (e.g., CLAP, Jukebox-derived models) and different reference datasets to adapt FAD for generative music evaluation [10]. Their results demonstrated that FAD's correlation with human judgments varies widely depending on embedding choice, highlighting the fragility of purely embedding-based metrics. This suggests that while embedding metrics are attractive for scalability, they may struggle with generalization across genres and cultural contexts.

Overall, embedding-based approaches represent a scalable but imperfect alternative to human data. They capture broad distributional properties of music but remain sensitive to the embedding model and datasets, leaving open the question of whether they can truly align with subjective human judgment.

DOI: 10.54254/2753-7064/2025.HT28341

2.3. Learned predictors and foundation-model evaluators

A third approach is to train models to predict human judgments, or to leverage foundation models as evaluators. Unlike embedding metrics, these methods aim to learn human preference functions directly, either by regression, ranking, or prompting large-scale multimodal models.

MOSNet is a landmark example in this approach [11]. Originally proposed and developed for speech synthesis evaluation, it trains a deep neural network to predict Mean Opinion Scores (MOS) from raw audio features. By minimizing the error between predicted and human-provided MOS, MOSNet provides a direct, automated AI predictor of perceptual quality. This work has inspired many other studies in both speech and music evaluation. However, the reliability of MOSNet-like models depends on the scale and diversity of training data; without large annotated datasets, they risk overfitting to narrow conditions and generalizing poorly.

Building on this, industry-driven work such as DNSMOS extended MOS prediction into real-world noisy speech conditions [12]. These models trained on massive crowdsourced human ratings achieved production-level performance, making them suitable for deployment in audio enhancement systems. Although DNSMOS was not designed specifically for music, it demonstrates the potential of large-scale, data-driven MOS prediction models for audio, and similar architectures could be adapted to musical contexts in the future.

In parallel, models like NISQA moved beyond predicting a single MOS value to estimating multiple perceptual dimensions (e.g., coloration, discontinuity) [13]. Similar to human judgment based models like SongEval, this multi-dimensional approach makes evaluators more interpretable and better aligned with how humans perceive complex audio. Such methods, if adapted to music, could better capture different facets of musical quality like harmony, timbre, or emotional expressiveness.

More recently, foundation models have opened new directions. Prompting Audio-Language Models (PAM) introduces the idea of using large Audio-language Models (ALMs) as evaluators by prompting them with textual queries about musical qualities [14]. Because ALMs are trained on both audio and text, they can flexibly evaluate dimensions like creativity, style, or genre adherence without task-specific training. PAM shows promising correlation with human preferences, but also reveals challenges that are commonly seen in many large-scale AI models: foundation models may hallucinate, be biased toward training data distributions, or struggle with fine-grained subjective criteria.

Taken together, learned MOS predictors and prompted foundation model evaluators represent a rapidly evolving direction for AI music evaluation. They bring evaluators closer to modelling human perception directly, but raise issues of data scale, generalization, and reliability, especially as evaluators move from narrow MOS regression to broad, text-driven judgments

3. Comparison, challenges and future prospects

The three major categories of existing work in this field as described in the previous section each present their unique features and limitations.

3.1. Comparison

Table 1 provides the comparison of approaches in music evaluation.

DOI: 10.54254/2753-7064/2025.HT28341

Table 1. Comparison of approaches in music evaluation

Approach	Representative Work	Features	Limitations
Human-Grounded Datasets for Preference Learning	SongEval [5], ARMOR [6], Music Arena [7]	Ground truth directly tied to human judgments, better aligned with human preferences; multidimensional annotations; live, renewable crowd feedback	Data collection is slow and difficult to scale; potential annotator bias; limited cultural and genre diversity; pairwise crowdsourced data may be inconsistent
Embedding- and Distribution-Based Metrics	Fréchet Audio Distance (FAD) [8], MAD [9], FAD toolkit adaptations [10]	Scalable, automatic evaluation; no need for human annotations; widely researched and applied across music and speech	Highly dependent on embedding model choice; may lack optimization for musical structure; fragile generalization across genres and cultures
Embedding- and Distribution-Based Metrics	MOSNet [11], DNSMOS [12], NISQA [13], PAM [14]	Predict human opinion scores directly; can estimate multiple perceptual dimensions; foundation models allow flexible evaluation across creative qualities	Requires large annotated training datasets; risks of overfitting; foundation models prone to hallucination and bias; interpretability challenges for subjective criteria

3.2. Common challenges

Despite the obvious differences in the purposes and methodologies of each of the approaches mentioned above, they share challenges that are common for this field. These challenges highlight the fundamental difficulty of aligning machine-learned models to humans' subjective judgements of music.

3.2.1. Subjectivity and cultural dependence of human musical judgment

The most persistent challenge in this field is the inherently subjective nature of human music evaluation. Unlike other audio related tasks such as speech intelligibility or audio denoising, where there are clearer objective standards, musical quality is highly dependent on culture, context, and individual taste. For example, what may be considered "coherent structure" or "pleasant harmony" in Asian folk music or improvisational jazz may be interpreted very differently in a Western classical context. This subjectivity inevitably introduces bias in human-grounded datasets like SongEval and ARMOR, since annotators often share similar training backgrounds. It also limits embedding-based metrics such as FAD, which implicitly assume a universal representation of "musical similarity." Even foundation models like PAM, while flexible, are often trained on datasets that disproportionately reflect Western genres, risking cultural skew. Thus, the first and most pervasive challenge is how to capture the diversity of musical perception in a way that remains fair, representative, and consistent across audiences.

DOI: 10.54254/2753-7064/2025.HT28341

3.2.2. Scalability and data availability

Another major challenge is scaling reliable training datasets and evaluation methods. Human-grounded approaches like SongEval produce rich annotated data, but that requires months of expert labeling, which cannot feasibly provide sufficient amounts of training data considering the modern scale of large models. Even learned predictors such as MOSNet or DNSMOS require enormous annotated datasets for training. Embedding-based metrics partially solve scalability by automating evaluation, but they still depend on pretrained embedding extractors, which themselves require massive datasets to train. Furthermore, live evaluation platforms like Music Arena can solve the problem of quantity, at the cost of risking inconsistent quality in crowdsourced feedback. In all approaches, there is a fundamental bottleneck: high-quality evaluation requires either large human-labeled corpora or large-scale embedding models, both of which are costly and slow to obtain.

3.2.3. Reliability and generalizability

A third challenge lies in creating evaluation metrics that remain reliable and generalizable across genres, and styles. Many models perform well when evaluated on narrow benchmarks but fail in broader settings that are closer to real-world applications. For instance, MOS predictors trained on studio-quality data may perform poorly when assessing live recordings or experimental genres. Foundation model approaches like PAM are more flexible, but these have their unique issues: hallucination, inconsistency, and prompt sensitivity. This unreliability poses a practical risk: a model evaluated as “high quality” by current metrics might still sound incoherent or unmusical to listeners outside the evaluation dataset’s cultural or stylistic domain. Without stronger guarantees of generalization, the use case for AI evaluation methods are extremely limited, failing to track true listener perception.

3.3. Future prospects

While the challenges in AI-based music evaluation are substantial, they also point toward promising directions for future research and innovation. Addressing these issues could require combining technological advances with new ways of thinking about music and culture.

3.3.1. Hybrid human–AI evaluation

A promising future direction is the development of hybrid pipelines that combine scalable AI metrics with human calibration. Instead of fully replacing human annotators, models could use embedding-based metrics like FAD or MAD for fast, large-scale screening, while a smaller set of human judgments provides ground-truth alignment. Platforms such as Music Arena already move in this direction by continuously collecting listener judgments, which can potentially be used to recalibrate automatic metrics periodically. Future systems might take this further via “active learning,” where AI models can learn to selectively ask humans to annotate only the most uncertain or representative examples. This approach balances scalability with subjective accuracy, ensuring that evaluation remains grounded in human perception while reducing annotation costs.

Another major challenge is scaling reliable training datasets and evaluation methods. Human-grounded approaches like SongEval produce rich annotated data, but that requires months of expert labeling, which cannot feasibly provide sufficient amounts of training data considering the modern scale of large models. Even learned predictors such as MOSNet or DNSMOS require enormous annotated datasets for training. Embedding-based metrics partially solve scalability by automating

DOI: 10.54254/2753-7064/2025.HT28341

evaluation, but they still depend on pretrained embedding extractors, which themselves require massive datasets to train. Furthermore, live evaluation platforms like Music Arena can solve the problem of quantity, at the cost of risking inconsistent quality in crowdsourced feedback. In all approaches, there is a fundamental bottleneck: high-quality evaluation requires either large human-labeled corpora or large-scale embedding models, both of which are costly and slow to obtain.

3.3.2. Reliability and generalizability

Future research will also need to move beyond Western-centric benchmarks in order to better account for cultural diversity in music. Current datasets and embedding models often lack sufficient representation for non-Western music, which limits the global applicability of evaluators. One possible solution is training embedding models on cross-cultural datasets that span diverse genres, instruments, and styles. Another is designing evaluators that explicitly incorporate genre or context as part of the evaluation criteria, rather than assuming a universal measure of musical quality. For example, an AI evaluator might learn to assess improvisational creativity in jazz differently from harmonic stability in classical music. By embracing cultural and stylistic variation, AI evaluators can become more inclusive and reflective of real-world music listening experiences.

3.3.3. Foundation models as universal music critics

Finally, the rise of large multimodal foundation models opens the possibility of highly general, flexible evaluators that function more like “AI music critics.” Systems like PAM demonstrate how audio-language models can be prompted to give subjective judgments of qualities such as creativity, emotion, or genre adherence. As these models grow in scale and sophistication, they may be able to integrate multiple modalities—listening to the audio, analyzing lyrics, and contextualizing within cultural trends—providing richer and more holistic evaluations in an almost agentic way. However, for this to succeed, such models would have to become more explainable and controllable, ensuring that their judgments are not opaque or biased. If these challenges are overcome, foundation models could shift AI music evaluation from narrow score prediction to nuanced, critic-like assessments that mirror the complexity of human musical discourse with revolutionary efficiency and scale.

4. Conclusion

To conclude, this paper provides a detailed and comprehensive overview of the current status of research and products in the field of AI music evaluation. This paper examines existing studies, analyzing their unique values and current limitations, while also providing a comparison of the various approaches. This paper then identified common challenges shared by all the approaches, and suggests possibilities for future prospects in the field.

Looking ahead, the field is still in its early stages of development, but hybrid systems that integrate human feedback with scalable AI, culturally diverse datasets, and foundation models acting as “AI music critics” point toward a promising future. Ultimately, successful music evaluation will require bridging the gap between computational efficiency and the richness of human experience, creating tools that not only measure sound but also reflect the emotional, creative, and cultural essence of music.

DOI: 10.54254/2753-7064/2025.HT28341

References

[1] Fitch, W. (2013) Musical protolanguage: Darwin's theory of language evolution revisited. Frontiers in Psychology, 4, 1-15.
[2] Wojnar, A. (1963) An analysis and synthesis procedure for feedback FM systems. IRE Transactions on Audio, 11, 54-62.
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.
[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684-10695.
[5] Yao, J., Zhang, K., Li, M., Chen, H., Wang, T., Liu, X. and Zhao, Y. (2025) SongEval: A benchmark dataset for song aesthetics evaluation. arXiv preprint arXiv: 2505.10793. Retrieved from https: //arxiv.org/abs/2505.10793
[6] Wang, S., Bao, Z. and E, J. (2021) Armor: A benchmark for meta-evaluation of artificial music. Proceedings of the 29th ACM International Conference on Multimedia, 3182-3190.
[7] Kim, Y., Park, J., Choi, H., Lee, J., Chen, J. and Nam, J. (2025) Music Arena: Live evaluation for text-to-music. arXiv preprint arXiv: 2507.20900. Retrieved from https: //arxiv.org/abs/2507.20900
[8] Kilgour, K., Zuluaga, M., Roblek, D. and Sharifi, M. (2018) Fréchet audio distance: A metric for evaluating music enhancement algorithms. arXiv preprint arXiv: 1812.08466. Retrieved from https: //arxiv.org/abs/1812.08466
[9] Huang, Y., Li, J., Xu, K., Chen, S., Wang, R. and Zhang, Y. (2025) Aligning text-to-music evaluation with human preferences. arXiv preprint arXiv: 2503.16669. Retrieved from https: //arxiv.org/abs/2503.16669
[10] Gui, A., Li, T., Sun, Q., Wang, Y. and Yang, X. (2024) Adapting Fréchet audio distance for generative music evaluation. ICASSP 2024-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1061-1065. IEEE.
[11] Lo, C., Fu, S., Huang, W., Wang, X., Yamagishi, J., Yu, C. and Tsao, Y. (2019) MOSNet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv: 1904.08352. Retrieved from https: //arxiv.org/abs/1904.08352
[12] Reddy, C.K.A., Gopal, V. and Cutler, R. (2021) DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. ICASSP 2021-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6493-6497. IEEE.
[13] Mittag, G., Naderi, B., Möller, S., Ribeiro, F. and Cutler, R. (2021) NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. arXiv preprint arXiv: 2104.09494. Retrieved from https: //arxiv.org/abs/2104.09494
[14] Deshmukh, S., Park, H., Li, C., Wang, Y., Liu, J., Ma, C. and Li, H. (2024) PAM: Prompting audio-language models for audio quality assessment. arXiv preprint arXiv: 2402.00282. Retrieved from https: //arxiv.org/abs/2402.00282