Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 3, 2026

Creative Demands and Practical Limitations of Controllable Music Generation from the Human Perception Perspective

Xinyu Jiang

School of Johns Hopkins University, Maryland, Baltimore, 2120, USA

Abstract: Controllable music generation aims to allow creators to guide AI in music creation through intuitive means, but the existing systems are still far from achieving this goal. From the perspective of human perception, this study combines computer music education and practical experience to explore the actual demands of creators for a generation system in terms of timbre control, rhythm arrangement, emotional expression, and editability of results. The study found that current controllable music generation has three deficiencies: First, there is a semantic mismatch between the system's parameter settings and the auditory perception of creators, and the adjustable dimensions are difficult to align with intuitive judgments in music creation; Second, the response of the generation model to control instructions lacks stability and precision, and there are frequent deviations between the creative concept and the final output; Third, the existing tools do not support an iterative workflow of repeated listening and local modification in music production, making it difficult to further edit and refine the generated results. The above problems show that, in order to serve the creative needs of people in their work, controllable music generation should be based on their perception habits and working methods rather than technical performance optimisation.

Keywords: Controllable Music Generation, Human Perception, Human-Machine Collaboration, Music Creation, Generation Model.

1. Introduction

In recent years, generation models based on deep learning have shown remarkable results in the field of music creation. Algorithms can now automatically generate symbolic melodies and, at the end of the process, produce complete audio files of well-structured and stylish works. Around 2023, models such as Suno AI and MusicGen were gradually released to the public. The user enters a short text description, and in about a few dozen seconds, a full-fledged work with music and singing will be generated. With the improvement of accessibility, people now expect that the professional threshold for music creation will be relatively low.

Technology has developed, but the basic problem remains: Can the present system be applied in actual creation work? For creators, the good tools should not only offer a single-output audio but also be flexible enough to meet their various demands at different times in creation, such as setting the general direction, adjusting specific sections, and adding generated content to their own works. However, such tools are still lacking in terms of fine control: if users wish to modify a particular detail in the generated result, they have to regenerate the whole part and cannot make targeted modifications based on what has already been generated.

Most of the current research on controllable music generation is based on technical paths, and these generally introduce multi-dimensional control signals, such as text prompts, harmony sequences, and emotional labels. However, it has not been determined whether these control methods meet the operating requirements of auditory thinking [1]. Music is a phenomenon of sound in time, and its standards for evaluation vary from person to person. A system with "controllability" of technical indicators may not achieve the creative goals in actual operation. From the perspective of auditory perception, this paper investigates the discrepancy between the ideal requirements and practical limitations in controllable music generation, but does not offer technical solutions. It will specify the problems at different levels and offer a reference for modifying research ideas.

The core judgment of this article is as follows: The current predicament faced by controllable music generation does not stem from insufficient technical capabilities, but rather from the fact that the research approach is centered around models and optimized based on technical indicators, without taking into account the auditory perception patterns and actual working methods of the creators as the fundamental reference for system design. This article does not aim to provide technical solutions, but rather attempts to clarify the problem levels and offer references for the adjustment of research approaches.

2. The Perceptual Demands of Creators for Controllable Music Generation and Their Levels

To understand this deficiency, we first need to determine what kind of control creators actually require in their daily work. The control requirements for creators of music generation systems are distributed at all times during music creation, and one must know these demands before determining whether the current system is lacking [2].

The demands for control at the level of sound and texture are relatively low. In the production of traditional music, the shaping of sound colour occurs at all stages of creation, and at each step, various adjustable parameters are available, such as selecting a waveform for a synthesizer oscillator, setting the cutoff frequency of a filter, and adjusting the reverb time and compression ratio of an effect. Creators adjust these parameters to get closer to the sound they want in their minds, and hope that the generation system can also offer such fine

control over the sound colour, rather than using general text descriptions such as "warm" or "bright". However, the actual output of AI audio generation models often has metallic noise or unnatural fluctuations in dynamics. Metal artifacts often have a glass-like high-frequency sound, sharp sibilants, and a robotic sound colour; this is because the acoustic dimensions of the voice are uniformly processed during generation and cannot be adjusted independently. AI-generated audio has poor control over loudness and sound colour stability [3].

Creators need to organise the temporal structure and phrase organisation at this level. Music has been lauded for the rise and fall of phrases, the contrast of sections, and the design of an emotional arc. Creators generally have a structure in mind for how the prelude introduces, how the chorus raises, and how the coda concludes, and then add specific notes. However, most of the current generation models are token-by-token prediction methods that are essentially local-to-global processes. Block-by-block generation can ensure a smooth connection between adjacent measures, but it is difficult to maintain the overall large-scale structure of the music. The problem of disordered sections and a lack of return to the theme are especially prominent in long-term works.

The control demands at the emotional level are relatively high. Music moves people by expressing the emotions that artists wish to convey. Creators hope that the generation system will be able to understand and execute emotional-level instructions, rather than a simple switch between "happy" and "sad", but rather gradual changes and subtle shifts in emotion. However, text-to-music models are not very good at expressing emotions subtly, and the resulting output is generally too neutral and lacks a clear direction. Flexible collection of emotions is precisely the area that the current system has failed to cover.

Overall, the creators' demand for controllable music generation is not merely a matter of adjusting technical parameters in a single dimension, but rather a systematic requirement that encompasses three aspects: timbre texture, temporal structure, and emotional expression. These three aspects, from concrete to abstract, and from local to overall, form a complete control demand spectrum. The current systems vary in their performance across these three levels, but the common problem is that they can respond to certain technical instructions, but they cannot truly understand the musical intentions corresponding to these instructions in the creators' auditory experience.

3. Semantic Dislocation Between Parameter Setting and Auditory Perception

The main problem of controllable music generation is that the control parameters of the system do not align with the auditory perception of creators.

Most generation systems have preset control items: tempo, key, instrument configuration, track density. Although these parameters are clear from an engineering perspective, they do not directly correspond to the attributes that creators are interested in for listening. After hearing the generated result, if a creator feels that the push of the chorus is lacking and wishes to add a forward-moving sense, there is no option named "drive" or "forward momentum" in the system. He can only change the speed or volume and regenerate, then listen to the result. There is no stable correspondence between hearing and machine operation, and thus it cannot be realised. Research has shown that the current text-to-music models are mainly used as creative sketch tools and are still far from being production-ready. Integration of AI audio materials into the human creative process and the control problem have not been solved yet [4].

The reason is that the objects of perception for music are highly subjective and multi-faceted. The same piece of music can be performed with different attention to the flow of melody, colour of harmony, or rhythm. The preset control parameters of the generation system are often convenient for technical implementation but not based on the requirements of auditory perception. Thus, the adjustable parameters do not meet the creator's requirements for adjustment.

This problem is not restricted to this. The system has the right theoretical parameters, but in practice, it performs better than expected. The Audio Diffusion Model is a black box, essentially. Musical skills and semantic attributes are intertwined in a large network of millions of parameters. Users cannot express fine-grained changes, such as a slightly slower speed or a slightly lower pitch of the human voice, through text. Any modification at the text level may lead to the generation of the entire song instead of a local adjustment. Geek Park has reported that if users wish to adjust the ratio of instruments, they can only repeatedly modify the prompt and generate again [5]. The Unpredictability continuously damages people's trust in creators. If the operating results cannot be predicted reliably, it will be difficult to add a generation tool in the creative decision-making process. In human collaboration for music, a simple phrase such as "Make it softer here" can express a particular wish and is easily understood. However, this natural mode of communication has not yet been achieved in human-machine interaction.

The essence of semantic dislocation lies not in the insufficient precision of certain parameter settings, but in the fundamental mismatch between the underlying logic of the system design and the perceptual logic of the creator. The technical side strives for the clarity and quantifiability of parameters, while the creative side relies on the ambiguity and wholeness of auditory experience. The translation layer between these two logics has not yet been established. Creators are forced to repeatedly switch between technical language and musical language, and the inefficiency and unreliability of this conversion process constitute the primary obstacle to the generation of controllable music.

4. The Non-Editability of the Generated Results and Its Disruption to the Creative Process

The instability of the control signal is only half the problem; the other half is that even if the generated results are largely usable, it is extremely difficult for users to make any modifications. Iteration is the basic way of music creation. Through multiple listens, problems are identified and locally corrected to gradually achieve the desired sound in one's mind. MIDI editing in Logic Pro, cutting and splicing of audio clips, and drawing automation curves for effects are all part of the cycle of "listening-modifying-re-listening".

However, the current main generation systems do not use this logic. The simplest way to put it is that the generated results are almost unchangeable. Most end-to-end models map the input to a complete audio file, and users do not have permission to modify any part of it; they cannot drag the

duration of a chord, lower the volume of a section, or replace the sound of a certain instrument. If the generated content does not meet expectations, there is no other choice but to revise the prompt and regenerate; then, from a group of new results, select one that is relatively satisfactory. Thus, the creative process is reduced to a single selection after repeated modifications, and the creator's authority gradually declines.

The model has a problem; it is not good. The working methods of diffusion models and autoregressive models are to gradually converge from random noise to a complete output, and the intermediate process lacks musical readability; users have no entry point for intervention and adjustment [6]. The representation method of an audio diffusion model is far removed from that of human hearing, and musical skills and semantic attributes have all been compressed into a network of millions of parameters; therefore, it is difficult for creators to affect specific features. To some extent, these tools are closer to presenting the model than to expanding the expressiveness of the model.

This deficiency is more prominent in actual music creation. In the audio preparation for live performances, the synchronisation of sound and pictures must be accurate to the frame, the levels of the tracks need to be repeatedly adjusted, and changes in effects should closely follow changes in emotion. All the above operations require fine control of audio materials, and this is exactly what AI tools for "generate and finish" cannot do. Even if the model can generate a functional part of the material, it still needs to be exported to a digital audio workstation for further processing. Generation is a single step, and production is another; there is a gap between the two, and this gap has already reduced the value of AI tools in actual creation [7].

The issue of non-editability is, on the surface, a functional deficiency, but at its core, it is a conflict between the generation paradigm and the creative logic. The working methods of diffusion models and autoregressive models determine that they can only provide "finished products", while music creation is essentially a process of continuous modification and gradual approximation. When the tool can only give the final result and cannot intervene in the intermediate steps, it is serving not creation but selection. This also explains why professional creators generally hold a reserved attitude towards current AI music tools - they provide options rather than malleable materials.

5. The Subjective Differences in Auditory Judgment and the Lack of Evaluation Reference

In addition to lacking editing functions, controllable music generation also suffers from a serious problem: the absence of a standard evaluation index. Traditional music production at least still has technical indicators such as spectrum, phase and level that can be referenced, but whether the generated music is pleasant or not ultimately depends on individual hearing, and the standard varies from person to person, making it difficult to unify [8].

Aesthetics of music are inherently very subjective problems. The same generated result will be praised by some people for its sound-colour combination, but others will find it stiff and harsh; some will hear emotional tension, while others will find it empty and dull. As a result of the above division, it has become difficult to determine what is considered a good outcome. Therefore, the creators do not know what to do; even if they learn how to operate a certain system, they cannot be sure that their modifications will lead to a better listening experience.

Another reason why people are more likely to be positively or negatively influenced by AI music is identity cognition. The same piece of music receives a lower score when it is labelled as "generated by AI", and the same audio, if labelled as "created by humans", is often rated higher. This bias is not due to actual auditory experience but rather to an innate prejudice against machine-created works. Therefore, even if the generation quality of AI continues to improve, the acceptance by the audience will not rise at the same time. Therefore, the circumstances of the creators are relatively difficult. They cannot only handle the unpredictable technical problems but also face an identity label based on the source in the minds of the audience.

There is also a genuine defect in the limited scope of the training data itself. At present, most of the learning materials for music generation models are Western pop and classical music, and the proportion of materials from other cultures is very small [9]. When creators wish to create music in a particular style or use non-standard instruments, the models are relatively weak; therefore, the quality of the output cannot be guaranteed. The higher the demand for "universality" and "controllability" at the technical level, the more difficult it is to meet all the requirements of music. This data-level narrowing indicates that music AI lacks cultural diversity and also results in an uneven distribution of the "controllability" problem in practice.

The absence of evaluation references has led to a vicious cycle for controllable music generation: without a stable evaluation standard, it is impossible to determine the direction of control; without a clear direction, the control itself loses its meaning. This problem cannot simply be attributed to "aesthetic subjectivity", but should be understood as the neglect of evaluation dimensions in the design of the music generation system. When the system only provides parameters but no reference framework, the creator is actually searching for a way out in a maze without a coordinate system.

6. From "Feeling Feasible" to "Practically Usable": The Real Distance in Practice

What would be the deficiencies in an actual project? In some cases of music production, controllable music generation is still quite far from being "just okay" and "actually usable". The "Hearing the Stars" project is for autistic children with auditory sensitivity, and its volume, tone and rhythm will be adjusted according to the children's real-time responses. Based on the above reasons, a high-precision, high-speed controller that can smoothly change parameters in real time is required. Any delay or sudden change will reduce the children's interest [10]. The final design of the project is a pre-built sound asset library and real-time parameter mapping; a generative model for content creation in real time will not be used. It was not out of technical conservatism; rather, the existing system could not yet realise stable, high-precision control in real time. Stability and controllability are required for special group applications, and newness is not.

Large-scale live shows are similar. For both the Spark25 charity concert and the musical Hamilton, there are many strict constraints on the organisation of audio materials, such as uniform sound-field coverage, consistent phase, precise

synchronisation of audio and video, and quick switching in emergencies. Such work does not require creativity but rather accuracy: frame-accurate timing, dB-accurate dynamic range, and finely adjustable parameter curves at every control point. Generative models can offer a reference for the creative stage, but they do not meet the high-precision requirements of traditional editing. If the generated results cannot be easily incorporated into the existing production process [11], they will only serve as a source of semi-finished materials for professional producers and not be true creative tools.

These experiences point to the same problem. Most of the current music generation research aims to "create a whole work of music from scratch", but actual creation is rarely like that. Creators often edit, adapt, organise and enhance the original materials. If a generative model can only generate from scratch and cannot be used in an editing workflow, it will have limited practical value. Controllable AI models have lowered the threshold for music creation, and non-professionals can adjust macro parameters such as genre and mood to achieve good results. However, professional creators need more granular control than this coarse-grained control offers; they do not want a lower threshold but rather a tool that can help them further develop their ideas while maintaining fine-grained control.

The gap from "seeming feasible" to "actually usable" ultimately lies in the distance between "generating a piece of music" and "completing a work". The former is a one-time output action, while the latter is a continuous creative process. The current controllable music generation systems are designed with the assumption that users want the former, but professional creators need the latter. This mismatch of goals fundamentally limits the practical value of AI tools in music creation more than any single technical indicator deficiency.

7. Conclusion

The demand for professional creators precisely shows the main problem this paper addresses: there is a considerable gap between the creative requirements and the actual conditions of controllable music generation. The control required of creators includes several factors, such as the mood, structure and emotion of music, and these are based on the natural laws of creation in sound. However, the current system has the following deficiencies: the parameter settings do not align with perceptual experience; generation results cannot be locally modified; and a stable reference for evaluation criteria is unavailable. In summary, the main point of this article is that the research on controllable music generation requires a shift in perspective - from "what the model can control" to "what the creator needs to control", and from "what technology can generate" to "what creation requires". This shift is not a denial of technology, but a return to the fundamental goal of using technology to serve creation.

The reasons for these problems are not solely technical deficiencies but also stem from the setting of research ideas. Most current work focuses on the model itself—finding ways to make it learn more control signals and improve control accuracy. But it rarely asks a simple question: What are these control signals about for the creators? A system with good technical indicators may not be practical for use in creation. These are two different evaluation logics; the former is the model's performance assessment, and the latter is user experience.

In terms of teaching and applying computer music, to enable controllable generation to be realised in practice, the direction of research needs to be adjusted. The focus should be on the individual rather than the model, and the design starting point should be auditory perception laws and creative habits, not technical indicators. Move away from the idea of "full automation" and adopt "semi-automatic cooperation"; embed the generation function into the creator's workflow rather than replacing it. What the creator needs is not a tool for writing songs; he wants a partner in creation based on his ideas.

The discussion in this paper is based on the existing literature and limited practical observations, covers cases in different environments, and may not cover all kinds of creations. Controllability requirements in different environments vary, and more generalisations of these observations need to be made. The purpose of controllable generation is to empower human creativity with technology, not to have people conform to technology. It needs to develop algorithms and have an understanding of and respect for music creation.

References

[1] Dash, A., & Agres, K. (2024). AI-based affective music generation systems: A review of methods and challenges. ACM Computing Surveys, 56(11), 1–34.

[2] Briot, J. P., Hadjeres, G., & Pachet, F. D. (2020). Deep learning techniques for music generation (Vol. 1). Springer.

[3] Maman, B., Zeitler, J., Müller, M., & Bermano, A. H. (2024). Multi-aspect conditioning for diffusion-based music synthesis: Enhancing realism and acoustic control. IEEE Transactions on Audio, Speech and Language Processing, 33, 68–81.

[4] Ronchini, F., Comanducci, L., Perego, G., & Antonacci, F. (2025). PAGURI: A user experience study of creative interaction with text-to-music models. Electronics, 14(17), 3379.

[5] GeekPark. (2026, August 24). AI Music reaches the “what now?” moment: Why are Chinese large models taking the hardest path? https://w.geekpark.net/news/369273

[6] Hachana, R., & Rasheed, B. (2025). Probe-assisted fine-grained control for non-differentiable features in symbolic music generation. IEEE Access, 13, 28059–28070.

[7] Clester, I. J., & Freeman, J. (2023). Composing with generative systems in the digital audio workstation. In Joint Proceedings of the ACM IUI Workshops.

[8] Lerch, A., Arthur, C., Bryan-Kinns, N., Ford, C., Sun, Q., & Vinay, A. (2025). Survey on the evaluation of generative models in music. ACM Computing Surveys, 58(4), 1–36.

[9] Mehta, A., Chauhan, S., Djanibekov, A., Kulkarni, A., Xia, G., & Choudhury, M. (2025, April). Music for all: Representational bias and cross-cultural adaptability of music generation models. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 4569–4585).

[10] Zheng, B., Yang, A. H., Ruan, J., He, J., Li, X., Chen, Y. H., & Ma, X. (2026, May). Real-time language model jamming: A case study for live music accompaniment generation. In 2026 IEEE 32nd Real-Time and Embedded Technology and Applications Symposium (RTAS) (pp. 261–274). IEEE.

[11] Raposo, A. N., & Soares, V. N. (2025). Modular microservices architecture for generative music integration in digital audio workstations via VST plugin. Future Internet, 17(10), 469.