

Practical Research on Virtual Vocal Performance Scene Enabled by Generative Artificial Intelligence

Jiayi Fang¹, Jing Zeng¹, Shuming Li^{1*}

¹ School of Music, Shandong University of Technology, Zibo, China

*Correspondence: Shuming Li

Received: 3 July 2026; **Accepted:** 28 July 2026; **Published:** 18 August 2026

Abstract:

With the rapid development of digital technology, generative artificial intelligence is gradually applied to music art creation by virtue of content generation and logical reasoning ability, which improves the problems of traditional vocal music performance scene solidification, single artistic conception expression and weak emotional atmosphere. Based on the relevant technical theories, this paper analyzes the artistic value of the traditional stage, and explores the application advantages of AI virtual scene in fitting the theme of the work, restoring the artistic conception of poetry and assisting emotional expression. The research takes the ancient poetry art song ‘Ding Feng Bo · Mo Hearing Lin Da Ye Sheng’ as the practical object, completes the whole process practice of AI script creation, lens generation and video synthesis, and builds a set of landing and reusable virtual vocal scene creation system. Practice has proved that generative AI can transform the text image, era background and emotional context of the work into immersive visual scenes, enrich the visual presentation of vocal music stage, and improve the singer’s work interpretation and stage interpretation level. This study adheres to the principle of ‘art-oriented, technology-oriented’, clarifies the auxiliary tool attributes of AI, and avoids the problem of aesthetic homogenization caused by technology abuse. This study expands the performance scope of vocal music art and provides a feasible reference for the second creation of vocal music and the innovation of AI-enabled vocal music performance.

Keywords:

Generative Artificial Intelligence, Virtual Vocal Music Performance Scene, Vocal Performance, Second Creation, Digital Empowerment

1. Introduction

Vocal music performance is a comprehensive art that integrates sound skills and emotional interpretation. ‘Sound and emotion’ is always its pursuit goal. Timbre texture, emotional communication and skill application together constitute the core dimension of vocal music emotional expression. With the rise of generative artificial intelligence represented by ChatGPT, Sora, DeepSeek and the expansion of application fields, artificial intelligence technology has also developed in the field of music. From emotional speech synthesis, voice conversion to intelligent performance

assistance, AI tools are continuously broadening the presentation boundary of traditional vocal music performance. However, the current integration of AI and vocal performance is still in the exploratory stage. How to effectively improve the emotional expression ability of artificial intelligence on the basis of adhering to the essence of vocal music performance art and relying on scientific scheme design has become a key issue to be solved urgently in the development of vocal music art in the digital age.

2. Development and Evolution of Generative Artificial Intelligence

Generative Artificial Intelligence (GenAI) is the core technology form formed by the evolution of artificial intelligence to the high-level stage. It is a new generation of artificial intelligence technology paradigm with intelligent creation and content generation as the core competence after perceptual intelligence and cognitive intelligence. From the academic point of view, generative artificial intelligence is based on deep learning theory and relies on large-scale pre-trained neural network model. Through supervised, semi-supervised and unsupervised training of massive, multi-source and multi-modal data, it can independently summarize and learn the internal distribution law and characteristic logic of data. [3] This technology does not need to manually preset fixed operation rules, and can independently produce new, original, logically complete and structured digital content according to the instructions, parameter settings and scene requirements input by users. Compared with the passive processing mode of traditional discriminant artificial intelligence, which can only identify, compare, classify and judge the stock data, generative artificial intelligence has realized the level leap of technical ability, and has the high-order ability of probability modeling, feature reorganization, original generation and scene generalization. It can independently complete the new creation of text, image, audio, video, code and other types of content. [4] At present, the core technology system supporting the landing iteration of generative artificial intelligence mainly includes generative adversarial networks, variational autoencoders, diffusion models, and Transformer pre-trained large models.

3. The Supporting Role of Scene Empowerment and Artificial Intelligence Virtual Scene in Vocal Music Performance

3.1. The importance of the performance scene to the singer

Singing is a comprehensive art that integrates hearing and vision, emotion and atmosphere. Many people mistakenly think that singing is all about singing, ignoring the key role of the performance scene. The performance scene composed of stage lighting, setting atmosphere, scene environment and audience atmosphere is an important carrier for singers to express their emotions, shape stage image and convey the core of works.

Performance scenes can stimulate stage expression and creative inspiration. The stage state of the singer is easily affected by the environment. The monotonous scene is easy to fall into rigid and rigid performance, and the performance scene that fits the theme can quickly awaken the singer's performance desire. The immersive stage scene will let the singer get rid of the state of mechanical singing, take the initiative to adjust the body movements and eye expressions with the rhythm of the scene, so that the singing is no longer limited to 'singing', upgraded to a complete stage performance with sound and emotion.

3.2. The help of virtual scene for vocal music performance

The virtual scene can accurately adapt to the theme of the song, render the emotional atmosphere, and help the singer to deeply integrate into the performance situation. Each vocal music work has a unique artistic conception and core story, but the traditional stage is limited by the construction cost and space conditions, and it is impossible to accurately restore the grand, delicate or fantastic exclusive scenes. There are often drawbacks such as the disconnection between the scene style and the song style, the thin stage atmosphere, and the difficulty of emotional expression.

The generative artificial intelligence has three core abilities: content generation, dialogue interaction and logical reasoning, which can effectively empower the innovative development of vocal music performance. At the level of content generation, AI can rely on the lyric image, creative background and emotional context of the work to generate scene copy, split-lens script, static picture and dynamic video suitable for the style of the work, build immersive virtual singing scene, and help the singer deepen the understanding and emotional input of the work. At the level of dialogue interaction, AI can rely on the narrative logic and character setting of vocal music works, combine the background of the times to generate script narrative and situational interaction content that fits the character's personality, enrich the level of scene plot, and enhance the drama authenticity and artistic expression of virtual scenes. At the level of logical reasoning, AI can systematically analyze the progressive relationship between the narrative structure and emotion of the work, sort out the emotional changes of the paragraphs, match the logic of the singing rhythm planning picture, provide scientific support for the scene production, and ensure that the visual narrative is highly compatible with the emotion of the work. [5]

4. The Generation Practice of Virtual Vocal Music Performance Scene

Based on the feasibility of generating artificial intelligence technology and the realistic construction demand of virtual vocal music scene, the complete transformation of vocal music works from text image to immersive audio-visual scene can be realized by relying on multimodal generation technology. Different from traditional vocal music performance forms, generative artificial intelligence can build an exclusive virtual stage adapted to vocal music narrative, emotion and work connotation through the digital creation logic of human-computer collaboration. Combined with this practical research, the complete construction process of virtual vocal music performance scene can be sorted out into three progressive links : track selection, text visual image conversion and video synthesis.

4.1. AI instructions generate scripts and static split screens

Combined with the singing and emotional progression of songs, the creation of vocal performance scripts can be completed by relying on structured AI prompt instructions. The system disassembles the singing paragraphs of the whole work, fits the rhythm of the sentences and the narrative of the lyrics, refines the environment, weather atmosphere, character status and emotional expression of different singing paragraphs, and restores the story context and artistic conception of the work. In the process of script creation, the style system of the repertoire is analyzed, the character modeling, scene and overall aesthetic tone are unified, and the consistency of the

narrative logic and visual style of the whole work is realized, so that the text script can fully serve the emotional expression needs.

After the complete script is formed, the generative artificial intelligence completes the lens disassembly and split lens prompt based on the composition logic, divides the independent lens according to the narrative order, and matches the exclusive scene, composition, depth of field, light and shadow tone and picture texture parameters for each lens. Combined with the details of the script, the positive prompt words suitable for the vocal music style are written to accurately limit the scene environment, character expression, picture quality, style and atmosphere tone. At the same time, the reverse prompt words are used to avoid the violation elements, and the modern style, exaggerated makeup, fuzzy distortion, abnormal composition and other defects that do not conform to the background of the work era and the artistic tone are eliminated, so as to ensure the professionalism and adaptability of the visual content from the source. Through the image generation tool, the standardized prompt words of each mirror number are imported, and the static sub-mirror screen is generated in batches. As the visual landing carrier of the text, the static lens intuitively presents the picture effect of each narrative node. It can not only verify the rationality of the previous script design, correct the problem that the picture does not match the artistic conception of vocal music performance, but also form a unified and standardized visual reference standard, lock the picture style, composition structure and detail texture for the next stage of video creation, and realize the seamless connection from the copy to the static image. [1]

In daily vocal music performance, AIGC digital intelligence tools are used to construct virtual vocal music performance scenes, including not only the visual depiction of characters in ancient poetry art songs, but also the image display of natural and social environment, as well as the image transformation of text story plots (fragments). This paper takes the ancient poetry creation song ‘Ding Feng Bo · Mo Hearing Lin Da Ye Sheng’ as an example to carry out a brief analysis. By inputting the corresponding instructions, AI can generate the singing script of the work, and sort out the core information of the film as a whole:

The core of the story is based on Su Shi’s “Ding Feng Bo · Mo Hearing the Sound of Leaves in the Forest,” restoring the rain scene in the Shahu Road of Huangzhou in the fifth year of Yuanfeng, and interpreting Su Shi’s broad-minded mood after the Wutai poetry case with the pictures of walking in the rain, remembering the past and sunny after the rain. Then he fixed the character standard. Su Shi’s appearance was square and round with a literati face. He was thin and had a sense of wind and frost. His temples were frosty and white. His jaws were black and white with goat-hu. His eyes were long and quiet, with vicissitudes and open-mindedness. Native skin, no beauty skin. Clothing props fixed green gray Northern Song Dynasty coarse linen cloth + brown braid + bamboo hat + yellow mud shoes, holding weathered old bamboo stick; no official robes, gold and jade accessories, embroidery patterns, clothing do old wear marks. The body shape is straight and relaxed, the walking is slow, and the waist and back are not rickets; the expression is indifferent, only showing a light smile, running, frowning, flustered, decadent, exaggerated emotions are strictly prohibited. [6]

4.2. Sub-mirror dynamic generation and complete video clip synthesis

The static split-lens picture can only achieve the stop-motion presentation of a single lens, and cannot match the coherent narrative rhythm and emotional flow of vocal performance. Relying on AI video generation technology and post-editing methods, the complete set of static lens material can be transformed into dynamic audio-visual content, and the results of text creation and static picture design can be inherited. Finally, the production of complete virtual vocal music performance video is completed, forming a full closed-loop creation system of ‘copywriting conception-static picture-dynamic filming’.

Based on the complete set of static sub-mirror images as the core visual reference, the standard prompt word system of each mirror is reused, and the single-lens dynamic material production is completed by relying on the AI video generation tool. The video generation stage strictly follows the established composition, tone, character image and scene atmosphere of the lens, only retains the environmental elements such as rain, mist, breeze and so on, which are suitable for the artistic conception of the work, maintains the stable and quiet film and television texture of the picture, avoids the redundant dynamic elements to interfere with the emotional expression of vocal music performance, ensures that the dynamic material is highly unified with the artistic style of static lens and text script, and realizes the coherence of the creative process. After the single-lens dynamic material is generated, the material integration clip is completed in combination with the narrative sequence and music rhythm of vocal music works. strictly follows the story sequence of the script to arrange the lens order, adjusts the length of single-lens playback according to the length of the phrase and the emotional ups and downs, and uses natural fade-in and fade-out transitions to connect different scene lenses for the emotional transition paragraphs of the work. Weaken the sense of fragmentation of the screen, connect fragmented lens materials, build a smooth and complete audio-visual narrative chain, and fit the progressive emotional logic of vocal performance. [2]

Finally, through the sound and picture alignment, color unity, image quality optimization to complete the finishing work. By using video editing software, the color temperature, light and shadow texture and film tone of the whole picture are unified, the color deviation and texture difference between different lenses are corrected, and the standardized video frame is standardized. Finally, with the help of the static sub-lens diagram of each mirror, the complete film of ‘Ding Fengbo · Mo Hearing the Sound of Lin Daye’ is output.

5. Conclusions

Traditional vocal music performance is mainly based on auditory presentation, which relies too much on the subjective interpretation of the singer. There are problems such as blurred scene images and difficulties in emotional substitution. It is difficult to fully show the literary artistic conception and dramatic tension of the work. This study combines the work ‘Ding Fengbo · Mo Hearing Lin Dayesheng’ to carry out a complete creative practice, and establishes a complete system covering text analysis, script creation, lens generation and video synthesis, and verifies the application value of AI in the second creation of vocal music. With the advantages of semantic analysis and generation, the generative AI can refine the background, narrative context and emotional tone of the work, and build an immersive virtual scene with unified style and rhythm adaptation. Compared with the traditional fixed stage setting, AI scene can dynamically adjust the atmosphere with emotional changes, visualize the artistic conception of lyrics and the background of the times, provide

singers with intuitive situational reference, and enrich the visual expression of vocal music stage. AI is only an auxiliary tool for artistic creation. It has the advantages of high efficiency, precision and standardization, but does not have the ability of humanistic aesthetics and emotional empathy. The unique artistic charm and cultural heritage of vocal music cannot be reproduced by AI, but AI helps us to realize the effective integration of digital technology and traditional vocal music art, innovate the second creation path of vocal music, broaden the boundary of stage performance, and provide practical paradigm for digital teaching of vocal music, stage innovation and traditional cultural communication. It also provides strong support for the digital innovation and development of classical vocal music art.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by the 2024 Shandong Province Graduate Education Quality Teaching Resource Project entitled “Development of a Case Database for Vocal Performance”, grant number SDYAL2024104.

References

- [1] Cao, W.H.; Cao, T.J. Content Generation of Virtual Reality Media Based on Generative Artificial Intelligence. *Publishing Science*, 2026, 34(2), 50-58.
- [2] Xiao, X.L. Generative Artificial Intelligence Empowering the Creation of Contextualized Learning Environments for Ancient Chinese Poetry in Lower Grades. *Chinese Language World*, 2026, 16, 17-18.
- [3] Fan, L.; Kai, X. Practical Applications of Artificial Intelligence-Assisted Artistic Creation and Design. Beijing: Chemical Industry Press, 2025.
- [4] Wang, X.L.; Jiang, X.X.; Wang, Y. A Comprehensive Guide to AI Practice. Beijing: Posts & Telecom Press, 2025.
- [5] Xie, J.Y. Research on the Generation of Artistic Conception in Musical Performance. Nanjing: Nanjing University of the Arts, 2015.
- [6] Zhou, Y.X. Analysis and Performance Interpretation of the Art Song “Ding Feng Bo”. Qufu: Qufu Normal University, 2025.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)