Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Comment:

Suno: potential, prospects, and trends*

Jiaxing YU[1], Songruoyao WU[1], Guanting LU[1], Zijin LI[2], Li ZHOU[3], Kejun ZHANG‡[1,4]

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*Department of Music Artificial Intelligence and Music Information Technology, Central Conservatory of Music, Beijing 100031, China*

[3]*School of Arts and Communication, China University of Geosciences (Wuhan), Wuhan 430074, China*

[4]*Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314100, China*

E-mail: yujx@zju.edu.cn; wsry@zju.edu.cn; 3210105631@zju.edu.cn; lzijin@ccom.edu.cn; zhouli@cug.edu.cn; zhangkejun@zju.edu.cn

Received Apr. 17, 2024; Revision accepted May 24, 2024; Crosschecked June 7, 2024; Published online June 20, 2024

https://doi.org/10.1631/FITEE.2400299

Suno has attracted wide attention due to its impressive capabilities. It demonstrates technological advancements and opens up new possibilities for music composition, representing a milestone in the development of artificial intelligence (AI) music generation. In this paper, we first introduce the background and summarize the general technical framework of AI music generation, followed by an analysis of Suno's advantages and disadvantages. Finally, we discuss the future trends in Music and AI.

1 Introduction

In March 2024, Suno AI (https://www.suno.ai) released its latest AI music generation platform, Suno v3 (Suno) (Freyberg, 2024). Endowed with outstanding creative capabilities, Suno has quickly attracted widespread attention from Music and AI enthusiasts. It can generate personalized music based on text descriptions and make music composition simple and efficient, allowing everyone to experience the joy of creating music.

Suno's popularity is inseparable from the rapid development of AI music generation technologies. Currently, AI music generation contains two main categories: symbolic music generation (Huang CZA et al., 2019; Huang YS and Yang, 2020; Hsiao et al., 2021) and audio music generation (Agostinelli et al., 2023; Copet et al., 2023; Huang QQ et al., 2023). Symbolic music generation usually represents music as sequences of musical events and uses time-series models to predict future musical events based on preceding ones. It exhibits characteristics such as efficient information encoding, understanding of complex musical structures, and editability. Audio music generation often uses deep learning models to process music in the audio format, showcasing characteristics such as richness of information, authenticity of performance, and diversity of generation.

Exploring the general technical framework of AI music generation reveals that it primarily comprises three modules: a lyric generation model, a voice synthesis model, and a music generation model. First, the lyric generation model generates matching lyrics based on the text descriptions provided by users. Subsequently, the voice synthesis model converts these lyrics into singing voices with appropriate melodies. Finally, the music generation model generates corresponding accompaniments from the text descriptions and the singing voices.

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 62272409), the Key R&D Program of Zhejiang Province, China (No. 2022C03126), and the Ministry of Culture and Tourism of China (No. 2022DMKLB001)

ORCID: Kejun ZHANG, https://orcid.org/0000-0003-4592-1818

Based on this general technical framework, Suno realizes the entire creative process from text to song, fully demonstrating its advantages such as integrated lyric–melody–song creation, simplified and diverse music composition, and cross-cultural communication and integration.

Suno's success leads to a new trend in AI music generation. A surge of emerging AI music generation platforms has sprung up, such as SkyMusic (https://www.tiangong.cn) released by Kunlun Tech and Stable Audio (https://www.stableaudio.com) released by Stability AI. The emergence of such platforms is likely to spark a revolution in Music and AI and may have profound impacts. In the remainder of this paper, we will introduce the background of AI music generation, summarize the general technical framework, and then discuss Suno's potential, prospects, and future trends in Music and AI.

2 Background of AI music generation

In recent years, the advancements in neural networks have significantly contributed to the progress of AI music generation, which comprises two main categories: symbolic music generation and audio music generation. Fig. 1 shows the timeline of AI music generation.

2.1 Symbolic music generation

Symbolic music generation considers music pieces as sequences of text-like information, converting musical elements, such as pitch, position, duration, and velocity, into symbols for processing, modeling, and recreation. It primarily uses time-series models (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017) to effectively capture the temporal structure and handle long-distance dependencies between musical elements, thus generating stable and long-structured music pieces.

Symbolic music generation exhibits the following characteristics:

1. Efficient information encoding: The model can process and generate music more efficiently by converting musical elements into symbols (Huang CZA et al., 2019; Huang YS and Yang, 2020; Ren et al., 2020; Hsiao et al., 2021; Zeng et al., 2021).

2. Understanding of complex musical structures: Using advanced time-series models, symbolic music generation can understand and generate music pieces with complex musical structures (Wu J et al., 2020; Zou et al., 2022; Wu XD et al., 2024).

3. Editability: A significant advantage of symbolic music generation is that the generated music pieces are easy to edit. Since musical elements are converted into symbols, users can adjust them more flexibly.

Symbolic music generation can achieve high-quality music generation with low computing costs. This unique advantage ensures that its contributions to music generation will further be studied.

2.2 Audio music generation

Audio music generation uses deep learning models to process, analyze, and generate music in the audio format. With the rapid development of

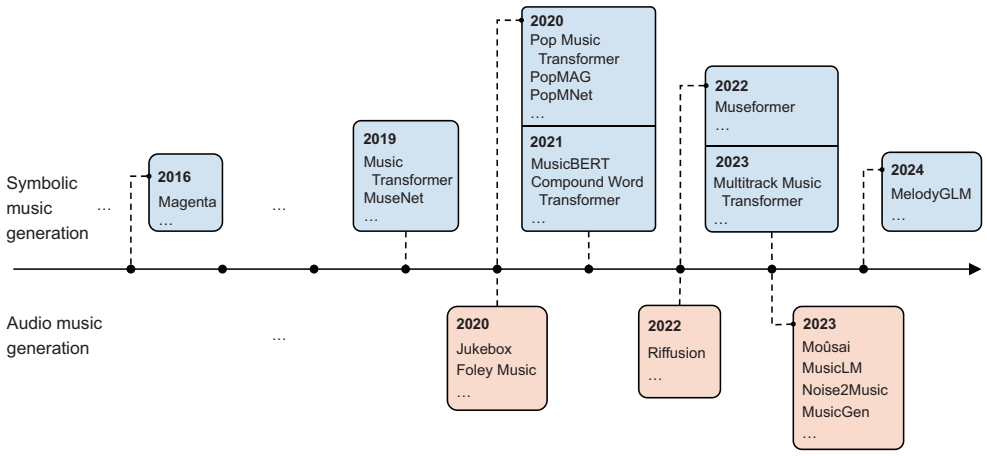

Fig. 1 Timeline of AI music generation, including symbolic music generation and audio music generation

audio generation (Kreuk et al., 2023; Liu et al., 2023), it evolves into more complex and creative applications, particularly speech generation and audio music generation. Speech generation converts text into speech by simulating human speech from the perspectives of tone, rhythm, and emotion, with standard models including FastSpeech (Ren et al., 2021) and SpeechT5 (Ao et al., 2022). Audio music generation generates music based on text descriptions, such as Jukebox (Dhariwal et al., 2020), Riffusion (Coldewey, 2022), MusicLM (Agostinelli et al., 2023), Noise2Music (Huang QQ et al., 2023), and MusicGen (Copet et al., 2023).

Audio music generation exhibits the following characteristics:

1. Richness of information: By directly processing audio data, it can accurately capture the details of music (Agostinelli et al., 2023; Copet et al., 2023), such as timbre, pitch, and rhythm.

2. Authenticity of performance: It can understand the complex musical information and achieve more authentic performance (Ren et al., 2021; Ao et al., 2022), such as the mixed sounds of different instruments.

3. Diversity of generation: Audio music generation supports diverse music generation using text descriptions that contain different styles and emotions (Dhariwal et al., 2020; Agostinelli et al., 2023; Copet et al., 2023).

3 General technical framework of AI music generation

The general technical framework of AI music generation primarily consists of three modules (Fig. 2): a lyric generation model, a voice synthesis model, and a music generation model. The three modules are detailed below:

1. Lyric generation model: A lyric generation model employs advanced natural language processing technologies, enabling a deep understanding of complex user instructions and emotional expressions. This module is featured by its high flexibility and innovation, which allow the model to quickly generate lyrics that match the text descriptions provided by users.

2. Voice synthesis model: A voice synthesis model transforms lyrics into natural singing voices. It not only preserves the meaning and emotions of the lyrics but also produces singing voices that closely match human performances. Taking Bark (https://github.com/suno-ai/bark) as an example, it adopts a multi-stage encoding approach to process user's inputs by converting the original text into tokens, which are then used to generate corresponding audio through a GPT (Brown et al., 2020) framework. Bark effectively decouples the training process, allowing each stage to be trained independently.

3. Music generation model: The overall architecture of audio music generation is shown in Fig. 3. The model takes music references (such as singing voices) as the input and generates corresponding music as the output. It enhances robustness to multi-batch, high-noise training data by continuously optimizing and iterating between audio and latent space, thereby improving the accuracy and reliability of music generation. Additionally, conditions, such as text descriptions and music references, can ensure that the generated music conforms to specific preferences.

We believe that the general technical framework of AI music generation described above lays the foundation for Suno, which aims to convert text descriptions into singing voices with specific emotions and styles, and to create harmonious accompaniments that match the descriptions and singing voices. It enables a complete creative process from text to an entire song.

4 Potential and prospects

As mentioned, Suno shows great potential in AI music generation. Its main advantages include:

1. Integrated lyric–melody–song creation: Existing AI music generation technologies, such as MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023), typically focus on specific aspects of song creation, such as text-to-music, and do not encompass the entire creative process. In contrast, Suno builds a multi-modal music composition platform that integrates the latest advancements in lyric generation, voice synthesis, and music generation. This fusion of innovative technologies creates new possibilities for the interaction between texts and songs, offering users a richer and more profound experience.

2. Simplified and diverse music composition: Suno makes simplified and diverse music creation possible. Unlike existing music generation

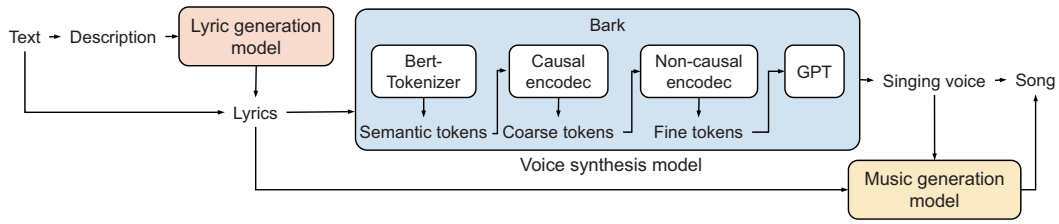

Fig. 2 General framework of AI music generation, including the lyric generation model, the voice synthesis model, and the music generation model

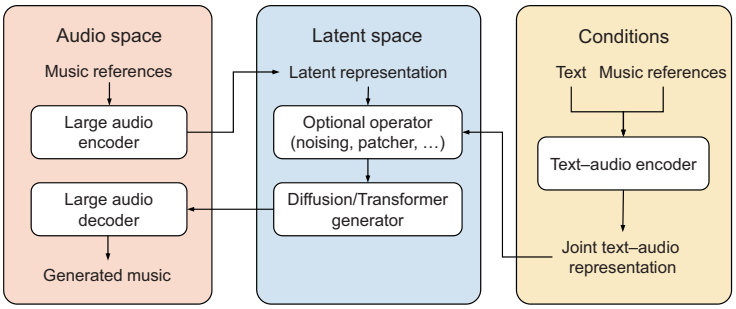

Fig. 3 Overall architecture of music generation models that take music references (such as singing voices) as the input and output the generated music

technologies that require complex text inputs with musical knowledge, Suno stands out by providing music composition solutions based on simple text descriptions. This unique way of creation provides a new way for music lovers to express themselves.

3. Cross-cultural communication and integration: Suno effectively promotes cross-cultural music communication and integration. As one of the most widely used AI music generation platforms, it breaks down geographical and cultural boundaries and facilitates the blending of musical elements from different cultures, opening up new paths for the development of global music culture.

Suno is changing the creative methods of amateur and professional creators. It greatly broadens the possibilities of musical expression, enabling more people to participate. Similar to ChatGPT's impact on various fields (Zhou et al., 2023), Suno is becoming a revolutionary force in the music industry, and its influence may extend beyond creation, profoundly impacting more fields such as music cognition, music industry, and music aesthetics.

5 Limitations

Although Suno has demonstrated strong capabilities in understanding and generation of music, it still has several limitations (Fig. 4):

1. Long-term coherence: Suno may face issues establishing coherence between different sections of longer songs. This highlights the challenge of long-term dependencies encountered by the Transformer (Vaswani et al., 2017) architecture, which is still one of the hot topics in academic research (Al-Rfou et al., 2019; Dai et al., 2019).

2. Fine-grained creation: Suno excels in basic song generation but still struggles to meet more detailed creative demands. For instance, when users attempt to mix multiple music styles or set tempos (in beats per minute, or BPM), Suno often fails to do so perfectly. To address this challenge, one potential solution is to implement a conversational generation platform that allows users to edit generated music pieces through dialog.

3. Multilingual support: Suno exhibits variations in its creative capabilities when dealing with different languages. For example, the model shows inadequacies in understanding the semantics and cultural context of Chinese lyrics. This challenge stems from the complexity of the language and the culture, requiring the model to parse literal meanings and capture subtle cultural nuances, which are crucial for song composition.

4. Content safety and legal risks: Suno may

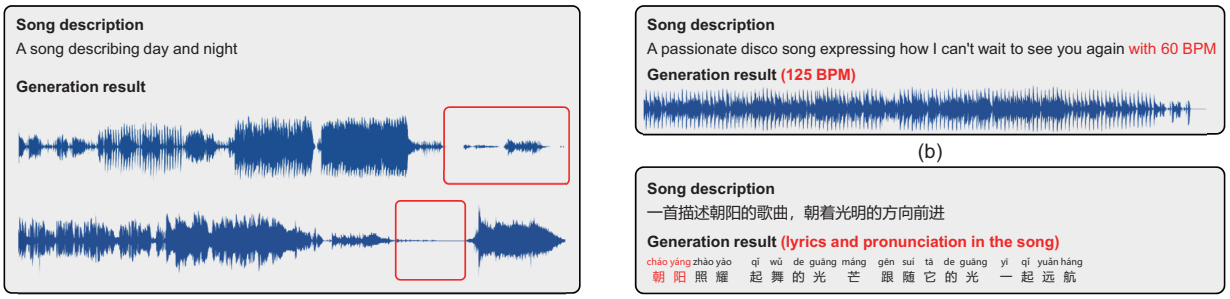

Fig. 4 Examples of Suno's limitations: (a) long-term coherence—Suno may face problems with continuity between different parts of longer songs; (b) fine-grained creation—the tempo of the generated song is 125 BPM, which does not match the description; (c) multilingual support—Suno is not good at dealing with Chinese polyphonic characters; for example, "朝阳" here is sung as "cháo yáng" by Suno, but the correct pronunciation is "zhāo yáng"

generate content involving sensitive issues or cultural conflicts without clear content restrictions. To reduce these potential risks, technologies such as reinforcement learning from human feedback (RLHF) can train specific reward models (e.g., for content safety) from human preference data and optimize the generation model, which has been adopted by existing studies (Ouyang et al., 2022; Touvron et al., 2023) and is also applicable for Suno.

6 Discussions and conclusions

Suno's emergence has already sparked discussions about Music and AI. Here, we discuss several future trends in Music and AI:

1. Music creation paradigms: With advancements in tools like Suno, Music and AI is moving toward more interactive composition experiences. This new paradigm enables music generation based on user descriptions and refines the creation process through dialog (Yuan et al., 2024). In the future, users can express their musical preferences through simple descriptions, making the creative process more flexible. Meanwhile, real-time feedback and dialog allow users to engage more deeply in music generation, ensuring that the generated music meets their emotional and aesthetic requirements.

2. Music cognition: The application of AI music technologies will profoundly impact our cognition of music. The involvement of AI expands our understanding of music. By analyzing a large number of music pieces, AI can discover previously overlooked patterns and trends in music creation, thereby advancing the development of music theories (Yu et al., 2022; O'Boyle, 2023).

3. Music industry and ecosystem: The further development of emerging AI technologies will revolutionize the music industry and its ecosystem. Many aspects of music, such as composition, distribution, and consumption, will become more efficient, personalized, and diverse. It will also lower the threshold for music generation, making it easier for creators to enter the field, thereby increasing the diversity of musical works. These changes not only foster new ways of music distribution and consumption, but also promote effective management of music copyrights.

4. Music aesthetic evaluation: Suno leads to discussions on the aesthetic value of AI-generated music. Formulating relevant evaluation standards can help us better understand and appreciate them. It may challenge traditional musical aesthetic theories, prompting us to think about how to improve aesthetic value in technological progress and gain a deeper understanding of music as a complex form of cultural and emotional expression.

In summary, as a pioneer in the field of Music and AI, Suno integrates numerous cutting-edge technologies to promote the advancement of AI music generation. We believe that AI music generation platforms will revolutionize traditional ways of creating music and give rise to new applications, ultimately forming a new trend in Music and AI.

Contributors

Jiaxing YU, Songruoyao WU, Guanting LU, and Kejun ZHANG drafted the paper. Zijin LI and Li ZHOU helped

organize the paper. Kejun ZHANG revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Agostinelli A, Denk TI, Borsos Z, et al., 2023. MusicLM: generating music from text. https://arxiv.org/abs/2301.11325

Al-Rfou R, Choe D, Constant N, et al., 2019. Character-level language modeling with deeper self-attention. 33rd AAAI Conf on Artificial Intelligence, p.3159-3166. https://doi.org/10.1609/AAAI.V33I01.33013159

Ao JY, Wang R, Zhou L, et al., 2022. SpeechT5: unified-modal encoder-decoder pre-training for spoken language processing. Proc 60th Annual Meeting of the Association for Computational Linguistics, p.5723-5738. https://doi.org/10.18653/V1/2022.ACL-LONG.393

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34th Int Conf on Neural Information Processing Systems, Article 159.

Coldewey D, 2022. Try Riffusion, an AI Model That Composes Music by Visualizing It. https://techcrunch.com/2022/12/15/try-riffusion-an-ai-model-that-composes-music-by-visualizing-it/ [Accessed on Apr. 6, 2024].

Copet J, Kreuk F, Gat I, et al., 2023. Simple and controllable music generation. Proc 37th Int Conf on Neural Information Processing Systems, Article 2066.

Dai ZH, Yang ZL, Yang YM, et al., 2019. Transformer-XL: attentive language models beyond a fixed-length context. Proc 57th Conf of the Association for Computational Linguistics, p.2978-2988. https://doi.org/10.18653/V1/P19-1285

Dhariwal P, Jun H, Payne C, et al., 2020. Jukebox: a generative model for music. https://arxiv.org/abs/2005.00341

Freyberg K, 2024. Introducing v3. https://www.suno.ai/blog/v3 [Accessed on Apr. 6, 2024].

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hsiao WY, Liu JY, Yeh YC, et al., 2021. Compound Word Transformer: learning to compose full-song music over dynamic directed hypergraphs. 35th AAAI Conf on Artificial Intelligence, p.178-186. https://doi.org/10.1609/AAAI.V35I1.16091

Huang CZA, Vaswani A, Uszkoreit J, et al., 2019. Music Transformer: generating music with long-term structure. 7th Int Conf on Learning Representations.

Huang QQ, Park DS, Wang T, et al., 2023. Noise2Music: text-conditioned music generation with diffusion models. https://arxiv.org/abs/2302.03917

Huang YS, Yang YH, 2020. Pop Music Transformer: beat-based modeling and generation of expressive pop piano compositions. Proc 28th ACM Int Conf on Multimedia, p.1180-1188. https://doi.org/10.1145/3394171.3413671

Kreuk F, Synnaeve G, Polyak A, et al., 2023. AudioGen: textually guided audio generation. 11th Int Conf on Learning Representations.

Liu HH, Chen ZH, Yuan Y, et al., 2023. AudioLDM: text-to-audio generation with latent diffusion models. Proc 40th Int Conf on Machine Learning, p.21450-21474.

O'Boyle M, 2023. (Re)Discovering Music Theory: AI Algorithm Learns the Rules of Musical Composition and Provides a Framework for Knowledge Discovery. https://csl.illinois.edu/news-and-media/rediscovering-music-theory-ai-algorithm-learns-the-rules-of-musical-composition-and-provides-a-framework-for-knowledge-discovery [Accessed on Apr. 6, 2024].

Ouyang L, Wu J, Jiang X, et al., 2022. Training language models to follow instructions with human feedback. Proc 36th Int Conf on Neural Information Processing Systems, Article 2011.

Ren Y, He JZ, Tan X, et al., 2020. PopMAG: pop music accompaniment generation. Proc 28th ACM Int Conf on Multimedia, p.1198-1206. https://doi.org/10.1145/3394171.3413721

Ren Y, Hu CX, Tan X, et al., 2021. FastSpeech 2: fast and high-quality end-to-end text to speech. 9th Int Conf on Learning Representations.

Touvron H, Martin L, Stone K, et al., 2023. Llama 2: open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.

Wu J, Liu XG, Hu XL, et al., 2020. PopMNet: generating structured pop music melodies using neural networks. *Artif Intell*, 286:103303. https://doi.org/10.1016/J.ARTINT.2020.103303

Wu XD, Huang ZJ, Zhang KJ, et al., 2024. MelodyGLM: multi-task pre-training for symbolic melody generation. https://arxiv.org/abs/2309.10738

Yu HZ, Varshney LR, Taube H, et al., 2022. (Re)Discovering laws of music theory using information lattice learning. *IEEE BITS Inform Theory Mag*, 2(1):58-75. https://doi.org/10.1109/MBITS.2022.3205288

Yuan RB, Lin HF, Wang Y, et al., 2024. ChatMusician: understanding and generating music intrinsically with LLM. https://arxiv.org/abs/2402.16153

Zeng ML, Tan X, Wang R, et al., 2021. MusicBERT: symbolic music understanding with large-scale pre-training. Findings of the Association for Computational Linguistics, p.791-800. https://doi.org/10.18653/V1/2021.FINDINGS-ACL.70

Zhou J, Ke P, Qiu XP, et al., 2023. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, early access. https://doi.org/10.1631/FITEE.2300089

Zou Y, Zou P, Zhao Y, et al., 2022. MELONS: generating melody with long-term structure using transformers and structure graph. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.191-195.