(IJ-11) Generating Music for Musical Scenes: Practice of AIGC Technology in Musical Creation for Entrepreneurship

Wei Li (Wish Li, Li Li) wei_li_1@alumni.brown.edu

Huiyue Gao gaohuiyue1029@163.com

Jiaqi Xu 18817371519@163.com

ABSTRACT

Musical theatre is a comprehensive art form that combines various elements such as music, dance, and drama to tell a story through dialogue, singing, and dancing. The overall musical theatre market in China is still in its early stage of development, with more and more musicians and playwrights attempting to create original musicals. However, due to the drawbacks of long creative cycles and high production costs, it is difficult for fledgling musical theatre companies to overcome these challenges and maintain a regular performance schedule, resulting in many failures in their entrepreneurial endeavors. One of the difficulties is creating different scene scores based on different musical theatre scenes. This challenge not only involves a lot of work and time, but also often leaves creators with insufficient inspiration based solely on the scenes, making it difficult to compose. In this paper, we propose an innovative application of AIGC (Artificial Intelligence Generated Content) technology in musical theatre creation, using cross modal machine learning to generate corresponding scores for different scenes, and through experimentation, adapting the commercially performed play "Blind Date is Cool, Being Serious is Fool?" into a musical to verify the feasibility of this approach, and analyzing the impact of different scenes on music generation. AIGC technology can serve as an auxiliary tool for musical theatre creation, saving creative time, improving the efficiency of theatre companies, and helping early-stage musical theatre teams shorten production cycles, thus assisting entrepreneurs in achieving success in their endeavors. More about this research can be found on my website www.liliwish.com.

INTRODUCTION

1. Musical Theatre

Musical theatre is a comprehensive art form that integrates various expressions such as music, dance, drama, and more, including plot, characters, dialogue, music, and dance, with the aim of telling a story or expressing a theme through various artistic forms. Originating in late 19th century America, musical theatre was initially known as "operetta" and was a light-hearted, upbeat form of musical drama with catchy tunes and lively dancing. Over time, as the art form developed, musical theatre became more complex and diversified, incorporating a wider range of musical and dance styles, and it has not only thrived in the United States, but also gained popularity worldwide.

The creation of a musical theatre production is a complex process that involves multiple stages and steps. Generally speaking, the creative process of a musical theatre can be divided into the following three aspects: (1). Scriptwriting: The story and plot of the musical are written by scriptwriters, who need to design the plot, characters' personalities, and actions, and also consider the integration of songs and dances. (2). Music composition: The music of the musical is composed by composers and lyricists working together. Composers need to create music based on the plot's emotions and atmosphere, while lyricists need to write suitable lyrics for the music. (3). Choreography: The dance of the musical is designed and choreographed by choreographers. Choreographers need to consider the coordination of dance with music and plot, as well as aspects such as rhythm, form, and style of the dance.

Musicals can also be adapted from plays, TV shows, movies, comics, and other sources. Among them, adapting plays into musicals has become a popular entrepreneurial direction for many emerging theater companies. By adapting a well-received and commercially successful play into a musical, it can greatly save the time of scriptwriting. Plays that have already been successfully performed commercially have a certain box office foundation and audience base, which can also help increase the box office of the adapted musical. The experiment in this article is to adapt a commercially performed play "Blind Date is Cool, Being Serious is Fool?”

2. Music In Musical

The music in a musical can be broadly categorized into two types: "songs with lyrics and melodies" and "scene underscore." Songs with lyrics and melodies are the music that requires actors to sing, such as the song "Memory" in the musical "Cats." Scene underscore, on the other hand, is music without lyrics that serves as background music to connect scenes and enhance the overall storytelling, often appearing between different songs.

The focus of this article is on the process of creating scene underscore, and does not discuss the process of lyric and melody composition. Scene underscore does not require the involvement of a lyricist, as it is solely composed by the composer, without the need to consider the relationship between lyrics and melodies, making it more suitable for utilizing AIGC (Artificial Intelligence Generated Content) technology. The creation of scene underscore in musicals requires different underscore for different scenes, while also emphasizing the continuity of the overall musical style, and ensuring smooth transitions between songs, plot, and story. The experimental scene underscore in this article does not consider the continuity of the overall musical style, and focuses only on individual scenes for music composition. The disadvantage of such an approach is the lack of overall continuity, but the advantage is that it allows composers to break free from the conventional thinking patterns of the entire musical, and explore more diverse inspirations.

3. The Challenges of Creating Background Underscore

In the process of creating a musical, composers often face challenges due to time constraints, as the creative team typically has tight deadlines. Composers often have limited time to generate sufficient inspiration, as the average duration of a musical can range from 1.5 to 3 hours [1], and the entire musical score needs to be composed within this timeframe. Additionally, musical theatre scoring typically involves arranging and composing for a variety of instruments, which further adds to the complexity of musical theatre scoring. Furthermore, the creation of a musical also requires showcasing the unique personality of the production, with subjective elements, making the underscore distinct and recognizable.

4. Motivation

This article utilizes AIGC technology as an auxiliary tool, capable of generating underscore with harmonious performance for multiple instruments simultaneously. This not only provides composers with more creative inspiration but also helps them identify suitable directions for their musical composition. By making secondary adjustments to the generated underscore, composers can highlight their own personal style. The application of AIGC technology can effectively assist emerging theater companies in shortening the creative cycle of productions, reducing time costs, and contributing to the success of their entrepreneurial endeavors.

RELATED WORK

1. AIGC (Artificial Intelligence Generated Content)

The popularity of ChatGPT has sparked immense interest in related artificial intelligence technologies. ChatGPT, along with other AIGC technologies, falls under the realm of artificial intelligence-generated content. AIGC is a generation technique that uses AI models to create digital content such as images, music, and natural language. The process of AIGC generation involves extracting and understanding intent information from human-provided instructions, and then generating content based on the knowledge and intent information. For example, OpenAI's MuseNet [2] model can generate songs in 10 different music genres and 15 different music styles based on user inputs. It can also imitate classical music styles, such as composing music in the style of Mozart, as well as imitate popular music styles, such as creating music in the style of Lady Gaga [2]. The recent popular Midjouney is capable of generating images and other content based on textual descriptions provided by users [3]. These are all applications of AIGC technology.

The foundation of AIGC technology is the transformer architecture [4]. Prior to the emergence of the transformer architecture, generative models followed different developmental paths in various fields, but the transformer architecture [4] served as a crossroads for these diverse paths. In 2017, the Transformer was introduced by Vaswani et al. for natural language processing (NLP) tasks. In the NLP field, many renowned large-scale language models, such as BERT and GPT, utilize the transformer architecture as their primary building block. In the CV field, Vision Transformer (ViT) [5] has further pushed the boundaries of the transformer architecture, laying the foundation for

generating music from images. Overall, the emergence of transformer-based models has revolutionized AI generation and provided possibilities for large-scale training.

2. Generating Music for Scenes

Generating music for scenes can also be abstracted as generating music from images, as scenes are spatial concepts in musical theater that share similarities with the essence of images. Therefore, generating music from images is a research direction within AIGC technology, specifically falling under the category of cross-modal machine learning, which also includes other research directions such as multi-modal machine learning and large-scale models.

Modalities refer to the different forms or manifestations of things. The world around us contains various modalities, such as visual objects, auditory sounds, tactile textures, and olfactory smells, among others. In cross-modal machine learning, the goal is to establish mapping relationships between different modalities, so that models trained on cross-modal data can generate output in a different modality based on input from one modality. For example, generating images from text input. In the context of this discussion, the application is generating music from images.

Algorithms and applications that directly generate music from images are not very common. Typically, an indirect approach is used to achieve music generation from images through a multistep process involving image-to-text-to-music generation. Specifically, cross-modal techniques are used to first extract information from images and generate textual descriptions of the images, which are then used to generate music.

3. Image to Text

There are several approaches proposed for text-music generation. Text-Music Generation [6] utilizes a deep cross-modal correlation learning architecture that employs intermodal canonical correlation analysis to measure the similarity of temporal structures between audio and lyrics. Another approach, JTAV [7], integrates textual, acoustic, and visual information using cross modal fusion and attentive pooling techniques to better understand social media content. In contrast, [8]

combines various types of music-related information such as playlists-track interactions and genre metadata, and aligns their latent representations to model unique music pieces. The CLIP Interrogator, created by pharma psychotic, is a tool for both artists and prompt engineers. It leverages the power of OpenAI's CLIP models [9] to test a given image against a variety of artists, mediums, and styles.

4. Text to Music

Rather than referring to it as text-to-music, it would be more accurate to call it description-tomusic. This is because most text-to-music techniques aim to generate melody based on lyrics. The goal of the technology proposed in this paper, however, is to generate music based on textual descriptions of images, which is fundamentally different from the former. Therefore, description-to-music is a more appropriate term.

MusicLM [11] is a mature technology for description-to-music. Specifically, MusicLM treats conditional music generation as a hierarchical sequence-to-sequence modeling task and generates coherent music. This model is capable of generating high-fidelity music from text descriptions, such as "a calming violin melody backed by a distorted guitar riff". MusicLM: Generating Music From Text [10] Mubert AI is a platform which can make text-to-music engine available for all content creators. It will generate a suitable track with duration up to 25 minutes which is good enough for music generation based on different scenes.[11]

METHOD

In this paper, we utilize AIGC technology to first transform scenes from a musical into text descriptions using "image-to-text" techniques, and then generate music using "description-tomusic(text-to-music)" techniques. Specifically, the entire process is illustrated in Figure 1, which involves four steps: scene selection, image-to-text conversion, text pruning, and text-to-music conversion.

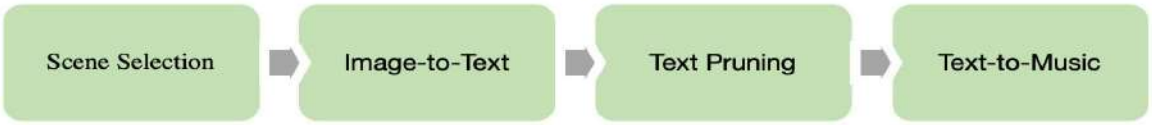

Figure 1

1. **Scene Selection**

First, the scenes from the musical are sliced, and among all the sliced scenes, representative scenes are selected as inputs for the model. The selection of scenes is crucial for music generation, as it can greatly impact the resulting music. The optimal scene selection should meet the following criteria:

CTIRERIA 1: No characters present: Scenes with characters should be avoided during scene selection, because different actors may perform the same musical, but the scenes remain unchanged. This eliminates the variable of characters' impact on music generation. The influence of characters on scene-based music generation will also be analyzed in subsequent experiments.

CTIRERIA 2: Clear visuals: Scenes with clear visuals should be chosen, avoiding scenes that are too bright, too dark, or blurry etc.

CTIRERIA 3: Simplistic visuals: Scenes with minimal objects should be chosen, avoiding scenes with too many objects. If a scene contains too many objects and is overly complex, it may result in overly complex text descriptions during the "image-to-text" step, which could hinder the "description-to-music" process in later stages.

2. **Image-to-Text**

In the experiment, we utilized CLIP Interrogator [12] to accomplish the process of image-to-text.

3. **Text Pruning**

Optimize the obtained text by pruning it. The text obtained from the previous step "Image-toText" is usually too lengthy and contains errors. It needs to be pruned through human intervention to make it suitable for the input of the next step "Text-to-Music", while also complying with the author's intention.

4. **Text-to-Music**

In the experiment, we utilized Mubert text-to-music to generate music from the input image. Then we convert the mp3 file into music score.

EXPERIMENT

1. Experiment

The experiment adapted an original stage play "Blind Date is Cool, Being Serious is Fool?" into an original musical. This is an original stage play from a startup theater company in Beijing, China. The play consists of 3 parts, each telling a different love story where people's life paths are changed by blind dates. The play has already had 4 successful commercial performances in Beijing, China, with considerable box office results.

Figure 2

To adapt this play into a musical, it needs to be done based on different scenes. The play consists of 3 acts, so there are at least 3 fundamental scenes. We will demonstrate the specific process of experimentation using one of the minimalist scenes from Act 1 as an example.

This scene does not contain any characters, the visuals are clear, and the composition is simple, which meets the requirements for input scenes in our AIGC approach.

When inputting this scene into the "Image-to-Text" model, CLIP Interrogator, the generated text description is as follows:

a table with two chairs and a vase with flowers on it, set design, tables and chairs, chairs and tables, still life photo of a backdrop, detailed set design, white tablecloth, inspired by Peter Brook, marina abramovic, hundreds of chairs and tables, romantic ambiente, containing tables and walls, inspired by Carrie Mae Weems

The original description was too complex and contained some errors. Following the author's subjective intent, the text description has been adjusted as follows:

a table with two chairs and a vase with flowers on it.

Inputting this scene into the Mubert text-to-music engine, an MP3 format music is generated and the MP3 audio file is converted to sheet music through a scoring software shown in Figure 2.

2. Experimental Analysis:

2.1 The impact of text pruning on music generation.

The most important manual intervention in the above experiment process is text pruning. In the Image-to-Text stage, the generated textual descriptions are too complex to be used as input for the existing Text-to-Music AI technology. Therefore, text pruning is necessary.

Text pruning is a subjective process that requires creators to adjust the text based on their own preferences. Typically, the generated text in the Image-to-Text stage is complex because existing AI algorithms tend to describe all information in the image in detail. The more information described, the better the algorithm performs. However, in the field of "generating music based on scenes", more detailed descriptions do not necessarily correspond to the creator's intentions.
Proper text pruning can help creators optimize the text according to their subjective artistic ideas.

Text pruning is not only influenced by the creator's subjective intentions, but also by the development of "text-to-music" technology. Since this technology has certain restrictions on the number of characters in the input text, the Mubert engine selected in this paper requires the text to be less than 200 characters. If the text exceeds 200 characters, the program will generate an error. Therefore, overly complex descriptions cannot be used as input for "text-to-music" and music cannot be generated.

2.2 The Impact of Different Text Descriptions on Music Generation.

We compare the differences in music generated from the following two text descriptions:

Text A!*a table with two chairs and a vase with flowers on it.*

Text B!a beautiful table with two beautiful chairs and a beautiful vase with beautiful flowers on it.

Here are two texts with an obvious difference of 4 "beautiful" adjectives in text B. We inputted both texts into the music generation engine Mubert, and the resulting music scores are shown in Figure 3 for Text A and Figure 4 for Text B.

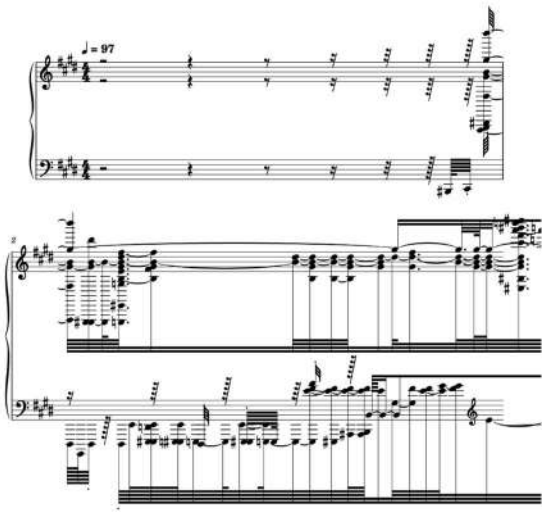

Figure 3.

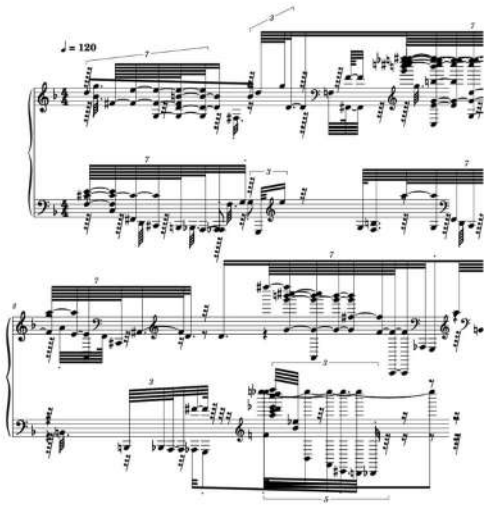

Figure 4.

The differences in the generated music are all produced by the algorithm itself without any manual intervention. The key in Figure 3 is E major while the key in Figure 4 is F major, and the melody are totally different. These differences can provide composers with more inspiration, but also require creators to be more cautious in the "manual pruning" stage, as different descriptions can produce completely different effects.

CONCLUSION & FUTURE WORK

This paper innovatively applies AIGC technology to the creation of musicals, effectively helping fledgling theater companies to quickly create music for scenes, providing more musical inspiration for creators, saving time and costs in musical creation, and increasing the possibility of entrepreneurial success. The feasibility of this paper's proposed method has been verified through experiments, by adapting a previously staged play into a musical. The proposed method relies on two algorithms: image-to-text generation and text-to-music generation. Currently, these two algorithms have limitations in accuracy and may require manual modifications. However, with the continuous development and maturity of AIGC technology, it is believed that these limitations will be gradually addressed.

In future exploration, the direction proposed in this paper can be continued by trying different AIGC algorithms to generate diverse types of music. Alternatively, a reverse approach can be explored by generating scenes based on music. This can provide broader perspectives and possibilities for musical creation.

More about this research can be found on my website www.liliwish.com.

FUNDING

This article supported by “the Fundamental Research Funds for the Central Universities”, Academic-level student project from The Central Academy of Drama, NO.YNXS2212 "Multimodal Generation and Display with Music#

REFERENCE

[1]. Musical theatre (2023) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Musical_theatre.

[2]. Christine Payne, "MuseNet," OpenAI, openai.com/blog/musenet, 2019.

[3]. Midjourney. Available at: https://www.midjourney.com.

[4]. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[5]. R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv:1910.13461, 2019.

A.Dosovitskiy,L.Beyer,A.Kolesnikov,D.Weissenborn,X.Zhai,T.Unterthiner,M.Dehghani,M. Minderer,G.Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[6]. Y. Yu, S. Tang, F. Raposo, and L. Chen, "Deep cross-modal correlation learning for audio and lyrics in music retrieval," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 15, no. 1, pp. 1–16, 2019.

[7]. H. Liang, H. Wang, J. Wang, S. You, Z. Sun, J.-M. Wei, and Z. Yang, "Jtav: Jointly learning social media content representation by fusing textual, acoustic, and visual features," arXiv preprint arXiv:1806.01483, 2018.

[8]. A.Ferraro,X.Favory,K.Drossos,Y.Kim,andD.Bogdanov, "Enrichedmusicrepresentationswithmultiplecross-modal contrastive learning," IEEE Signal Processing Letters, vol. 28, pp. 733–737, 2021.

[9]. Clip: Connecting text and images (no date) CLIP: Connecting text and images. Available at: https://openai.com/research/clip.

[10]. Agostinelli, A. et al. (2023) MusicLM: Generating music from text, arXiv.org. Available at: https://arxiv.org/abs/2301.11325.

[11]. Thousands of staff-picked royalty-free music tracks for streaming, videos, podcasts, commercial use and online content (no date) Mubert. Available at: https://mubert.com/.

[12]. Pharmapsychotic (no date) Pharmapsychotic/clip-interrogator: Image to prompt with blip and clip, GitHub. Available at: https://github.com/pharmapsychotic/clip-interrogator.