An Affordable, Controllable, and Modular System for AI Music Generation

Xingjian Du

December 15, 2024

Abstract

This report details the development of an affordable, controllable, and modular system for AI music generation. Inspired by Fordist principles of modular manufacturing, our pipeline enables text-to-music generation using independent, specialized components. Each module performs a distinct task, ensuring interpretability, flexibility, and maintainability. The proposed system leverages state-of-the-art AI techniques, optimizing resources and delivering high-quality, customizable music outputs.

1 Introduction

AI-driven music generation has emerged as a transformative technology with applications spanning entertainment, gaming, and multimedia content creation. Existing end-to-end systems, while effective, lack interpretability and control, often limiting their practical utility.

Our project proposes a modular pipeline for text-to-music generation. The system starts with a user-provided text prompt, which is transformed into lyrics using a Large Language Model (LLM). These lyrics guide parallel processes for melody generation and singing voice synthesis. The final product integrates vocals and accompaniment through a mixing stage, offering unparalleled control and flexibility compared to monolithic models.

2 Methodology

2.1 Pipeline Overview

The music generation pipeline is structured into four distinct stages:

1. **Text Prompt Stage:** The system accepts user input describing the desired song's concept, mood, or theme. This text is processed to generate structured lyrics.

2. **Lyric and Melody Generation:** Lyrics generated by the LLM are paired with matching melodies using a MIDI generator, ensuring coherence between textual and musical elements. The MIDI generator employs probabilistic sequence models and harmonization rules to create melodies that align with lyrical rhythm.

3. **Voice Synthesis:** Synthesized lyrics and melodies are combined in DiffSinger to produce a natural-sounding singing voice. Advanced AI techniques, including DiT (Diffusion Transformer) for velocity field prediction and Conv2NeXT for text processing, enable precise synchronization and audio clarity. The training incorporates masked audio modeling and classifier-free guidance to balance creativity and fidelity.

4. **Accompaniment and Mixing:** The vocal track is used to generate instrumental accompaniment via OpenSingSong. The accompaniment generation leverages pretrained chord recognition models and diffusion-based generative models, ensuring the instrumental layer complements the vocal melody. Both components are combined in the final mixing stage with dynamic equalization and reverb application to create a professionally polished output.

2.2 Advantages of the Modular Approach

- **Interpretability and Control:** Each module has a well-defined function, allowing outputs to be monitored and adjusted at every stage.

- **Flexibility:** Modules can be independently upgraded or replaced, adapting the system to diverse musical styles or languages. For instance, the MIDI generator can switch to style-specific models for jazz or classical compositions.

- **Maintainability:** Modularization simplifies debugging, quality control, and development cycles by isolating issues to specific components.

- **Resource Efficiency:** Independent modules optimize computational and training resources without retraining the entire system. For example, voice synthesis uses lightweight training datasets compared to end-to-end models.

2.3 Data Curation and Training

Singing Voice Synthesis: The M4Singer dataset, containing 29.77 hours of high-quality Mandarin singing data, was used. Special markers were employed to indicate notes for synchronization, such as pitch encoding with specific tonal annotations. Training utilized masked audio modeling to improve generalization and classifier-free guidance to enhance control over style and expression.

Accompaniment Generation: A dataset of 3,000 Mandarin songs was curated, separating vocal and instrumental tracks using advanced source separation methods like Demucs and Band-Split RoFormer. Noise and distortion were intentionally introduced during embedding extraction to improve robustness against recording artifacts. Chord recognition models were fine-tuned on these datasets to provide musically coherent accompaniment prompts.

3 Results and Discussion

The modular pipeline produced high-quality music outputs with several notable advantages:

- **Customizability:** Users could adjust lyrics, melody, and mixing parameters, tailoring outputs to specific needs. For example, the melody's tempo and key could be dynamically changed during generation.
- **Efficiency:** Training and inference times were significantly reduced compared to end-to-end approaches. Accompaniment generation achieved near-real-time performance by leveraging optimized chord-based prompting.
- **Scalability:** The system easily adapted to new styles and languages by fine-tuning specific modules. Initial experiments demonstrated successful transfer to English singing using curated datasets.

However, some challenges were identified:

- High-pitch vocals exhibited hissing noise, which required vocoder fine-tuning. Integrating neural vocoders designed specifically for singing synthesis is proposed as a solution.
- Lack of comprehensive evaluation metrics for singing voice synthesis and accompaniment quality. Preliminary user studies and perceptual listening tests are planned to address this gap.

Proposed solutions include developing English singing datasets, fine-tuning language models for duration information, and incorporating pretrained chord recognition models to enhance accompaniment generation quality.

4 Conclusion

This project demonstrates the potential of modular systems in AI music generation, offering a controllable, efficient, and scalable alternative to end-to-end models. Future work will focus on expanding linguistic and stylistic capabilities, refining individual components, and establishing robust evaluation metrics. Additionally, integrating neural vocoders and advanced harmonization techniques is expected to further enhance the system's performance.

Acknowledgments

We thank our collaborators and funding sources for their support in realizing this project.

References

1. Zhang, Lichao, et al. *M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus.* Advances in Neural Information Processing Systems, vol. 35, pp. 6914–6926, 2022.
2. Peebles, William, and Xie, Saining. *Scalable diffusion models with transformers.* Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023.

3. Feng, Jianwei, et al. *Conv2NeXt: Reconsidering Conv NeXt Network Design for Image Recognition.* 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT), pp. 53–60, 2022, IEEE.

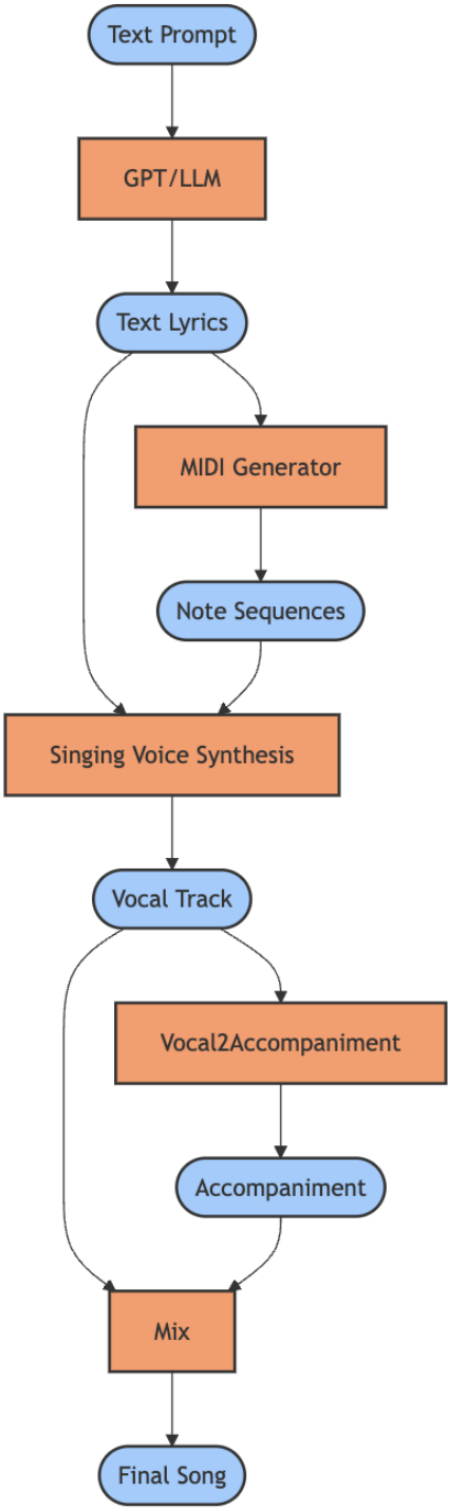

Figure 1: Overview of the Modular AI Music Generation Pipeline. Starting with a user-provided text prompt, the system generates structured lyrics using a Large Language Model (LLM). These lyrics are processed in two parallel branches: one generates a melody via a MIDI generator, while the other creates a synchronized singing voice using a singing voice synthesis module (e.g., DiffSinger). The vocal track, combined with melody, serves as input to a vocal-to-accompaniment system (e.g., OpenSingSong) for generating instrumental accompaniment. Finally, all components—vocals, melody, and accompaniment—are mixed to produce the final song.