iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70529

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

From Text to Tune: An End-to-End AI Pipeline for Automated Music Composition

Shivhar Dhulshette, Arya Shende, Rohit Ingole, Prof. Varsha Kulkarni
JSPM's Imperial College of Engineering and Research

Abstract: ***We present an integrated system that generates complete musical compositions from textual inputs. Our approach leverages three distinct pretrained models: a text-to-text generation model to produce song lyrics, a text-to-speech (TTS) model to vocalize the lyrics, and a text-to-audio music generator to synthesize complementary background music. Finally, the system overlays the synthesized voice with the background music to produce a final song. Experimental demonstrations indicate that this modular pipeline can generate coherent vocal renditions accompanied by music that supports the song's mood and style. This work highlights a flexible framework that may be extended to a range of creative and interactive music applications.***
Keywords: AI-generated music, text-to-speech, music generation, lyric synthesis, audio mixing, modular pipeline

I. INTRODUCTION

Recent advances in natural language processing and audio synthesis have opened new possibilities for automated creative content generation. In music, various models have been developed to generate either instrumental tracks or speech; however, the integration of these components to produce a complete musical piece remains underexplored. Our work bridges this gap by combining state-of-the-art text-to-text, TTS, and music generation models into a single pipeline. Given a textual prompt that indicates a song's topic or style, the system generates lyrics, synthesizes the vocals, creates an instrumental background, and finally merges these components to produce a finished song.

The modular design of our pipeline enables independent refinement of each component. This work is particularly significant for creative industries, where rapid prototyping of musical ideas or adaptive soundtrack generation for multimedia applications is highly desirable.

II. BACKGROUND AND RELATED WORK

Generative models in the audio domain have rapidly evolved, with recent works such as MusicLM focusing on high-fidelity music synthesis from text descriptions. Similar to these approaches, our system draws on:

Text-to-Text Generation: Transformer-based models like T5 have proven effective for creative text generation (Raffel et al., 2020).

Text-to-Speech Synthesis: Models such as Bark deliver expressive and natural-sounding vocal synthesis.

Music Generation: Diffusion and autoregressive methods (e.g., MusicGen) have achieved notable progress in synthesizing diverse musical backgrounds.

Audio Combination: Techniques for audio mixing (using libraries like pydub) allow us to blend separately generated tracks.

Prior research in conditional audio synthesis has focused on generating isolated components of music. Text-to-speech models, such as those based on deep generative methods, have shown impressive performance in generating natural-sounding speech. Similarly, text-to-music models have been explored to create background tracks or instrumentals based on textual descriptions. Meanwhile, large language models have achieved notable success in generating coherent lyrical content. Unlike previous works that target a single aspect of music synthesis, our pipeline unifies these approaches to generate an end-to-end musical composition.

Recent advancements in transformer-based text generation (e.g., LaMini-Flan-T5) and diffusion or autoregressive models for audio synthesis (e.g., MusicGen and Bark) have provided the necessary building blocks for our system. Our approach integrates these diverse methods to generate a complete audio experience from a simple text prompt.

Our work differentiates itself by integrating these components into a single automated pipeline that not only generates creative content but also produces a complete musical output.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. Overview

Our system comprises three primary modules:

Lyric Generation: A text-to-text generation model produces song lyrics based on a given prompt.
Voice Synthesis: A TTS model converts the generated lyrics into a vocal track.
Background Music Generation: A dedicated music generation model synthesizes instrumental audio based on a description.
Audio Mixing: A postprocessing module overlays the vocal track with background music, adjusting relative volumes to ensure clarity and balance.
The high-level architecture is illustrated in below Figure

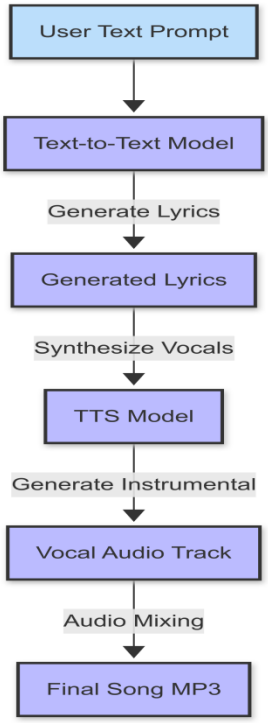


Figure 1. System Architecture Diagram

B. Module Descriptions

Lyric Generation

The system initiates the process by accepting a user-defined song topic or style. A pretrained text-to-text generation model (e.g., LaMini-Flan-T5) processes the prompt to produce coherent and creative song lyrics. The generated text is formatted with musical symbols to evoke a lyrical style, ensuring that the output is both human-readable and musically expressive.

Voice Synthesis

The generated lyrics are then converted to an audio waveform using a state-of-the-art TTS model (Bark). The TTS module synthesizes a natural vocal rendition of the lyrics. By relying on pretrained neural models, the system ensures that the synthesized speech preserves the prosodic and emotional cues intended by the lyrical content.

Background Music Generation

Simultaneously, a text-to-audio music generation model (MusicGen) produces a background instrumental track. This module is conditioned on a static text prompt (e.g., “Background music for the given song”) to create instrumental accompaniment that complements the lyrical content. The generated music maintains a consistent style and mood suitable for the intended genre.

Audio Mixing

Finally, the audio mixing module overlays the synthesized vocal track with the generated background music. Using an audio processing library, the system adjusts volume levels—typically reducing the background music’s amplitude—to ensure that the vocals remain prominent. The final composite is then exported in a standard audio format (MP3), suitable for playback and distribution.

C. User Interface

We utilize the Gradio library to provide a web-based interface that allows users to input a song topic, view generated lyrics, and download the final MP3 file. This interactive front end streamlines the user experience and encourages creative experimentation.

D. Workflow Diagram

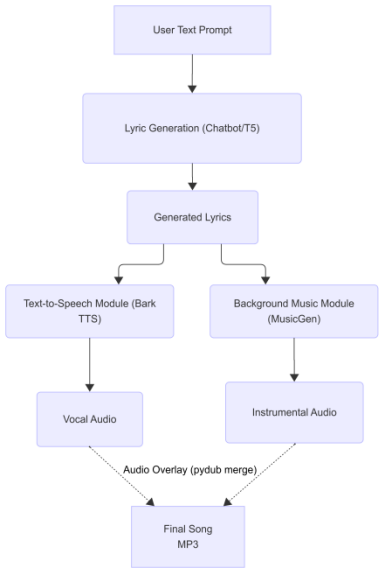


Figure 2.Workflow of our system illustrating the overall workflow

IV. IMPLEMENTATION AND EXPERIMENTAL EVALUATION

A. Implementation Details

Our pipeline is implemented using Python with several key libraries:
Hugging Face Transformers: Utilized for text-to-text generation.
Bark: Employed for high-fidelity TTS synthesis.
MusicGen: Used for generating background music.
Pydub and SciPy: Applied for audio processing and mixing.
Gradio: Provides a web-based user interface to facilitate interactive usage.
The system is modular, allowing each component to be independently updated or replaced. For instance, the lyric generation module may be swapped with a more advanced model in future iterations without affecting the TTS or music generation components.

B. Experimental Results

Preliminary tests indicate that the pipeline produces coherent songs that adhere to the style specified by the input prompt. Qualitative evaluations based on user feedback highlight the following observations:
Lyric Quality: The generated lyrics are contextually appropriate and exhibit creative variability.
Vocal Naturalness: The synthesized vocals are clear and exhibit appropriate intonation.
Musical Accompaniment: The instrumental background provides a complementary auditory experience without overpowering the vocal track.
Overall Coherence: The final composite song reflects a balanced integration of vocals and background music, offering an end-to-end solution for AI-generated musical content.
While the system performs well under typical conditions, certain limitations were observed. In some cases, the background music may not perfectly align with the emotional tone of the lyrics, suggesting that future work should explore more sophisticated conditioning techniques that jointly optimize both vocal and instrumental components.

V. DISCUSSION AND FUTURE WORK

Our work demonstrates the feasibility of an end-to-end pipeline that integrates multiple AI-driven modules to generate complete musical compositions. The modular architecture provides flexibility and extensibility:

Enhanced Conditioning: Future iterations could incorporate adaptive conditioning where the background music is dynamically generated based on specific lyrical cues.

Improved Vocal Synthesis: Integration of prosody control and emotion recognition may further enhance the expressiveness of the synthesized vocals.

User Feedback Loop: Incorporating a user feedback mechanism could allow iterative refinement of the generated song, leading to personalized music creation.

Real-time Generation: Optimizing processing times to enable near real-time song generation would expand the system's applications in interactive media and live performances.

Overall, the proposed system paves the way for creative applications in automated music production, digital art, and interactive entertainment.

VI. CONCLUSION

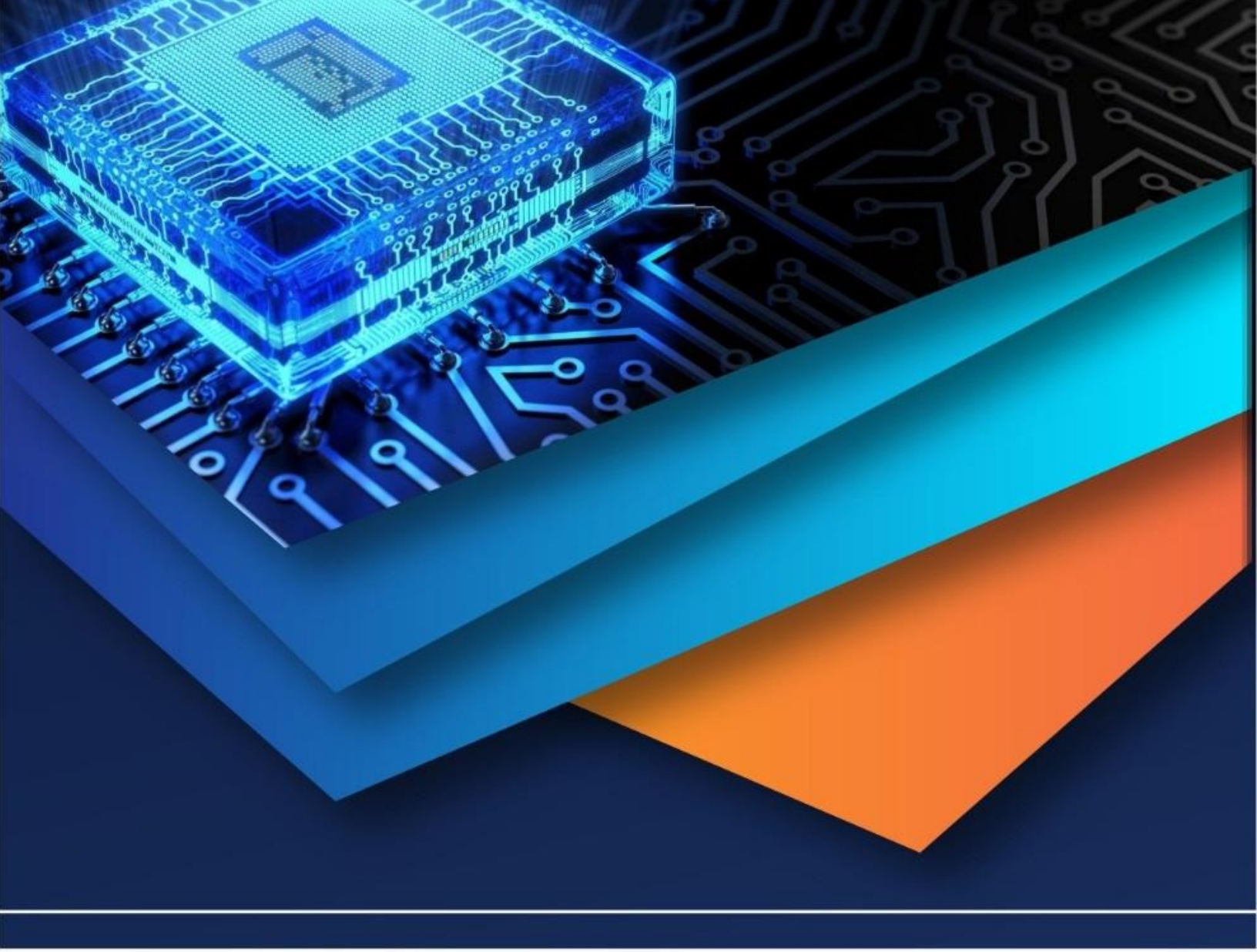
We have introduced a modular pipeline that unifies text-to-text generation, TTS, and music generation to create a complete AI-generated song. Our system demonstrates promising qualitative results by producing coherent lyrical content, natural vocal renditions, and complementary instrumental backgrounds.

Future research will focus on refining the conditioning mechanisms and integrating user feedback to further enhance the musical quality.

This work contributes a flexible framework for future explorations in AI-driven music generation.

REFERENCES

[1] Borsos, C., et al. (2022). AudioLM: A Framework for Audio Generation. Retrieved from arXiv:2202.11446.
[2] Dhariwal, P., et al. (2020). Jukebox: A Generative Model for Music. OpenAI. Retrieved from https://openai.com/blog/jukebox/.
[3] Forsgren, A., &Martiros, M. (2022). Riffusion: Text-to-Audio Diffusion for Music Generation. Retrieved from https://github.com/riffusion/riffusion-app.
[4] Hugging Face. (2023). Transformers Documentation. Retrieved from https://huggingface.co/docs/transformers/index.
[5] MBZUAI. (2023). LaMini-Flan-T5-248M. Retrieved from https://huggingface.co/MBZUAI/LaMini-Flan-T5-248M.
[6] Suno AI. (2023). Bark: High-Fidelity Text-to-Speech Synthesis. Retrieved from https://github.com/suno-ai/bark.
[7] Facebook AI. (2023). MusicGen: Generative Music from Text. Retrieved from https://github.com/facebookresearch/audiocraft.
[8] Gradio. (2023). Gradio: Build Machine Learning Demos and Web Apps. Retrieved from https://gradio.app/.
[9] The SciPy community. (2023). SciPy Documentation. Retrieved from https://docs.scipy.org/.
[10] Pydub. (2023). Pydub Documentation. Retrieved from https://github.com/jiaaro/pydub.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY