Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/119/2025.21641

A Survey of Text-to-Music Generation with Deep Learning

Jiaxing Dong[1,a,*]

[1]*School of Computer Science, Northwestern Polytechnical University, Chang'an Campus, 1 Dongxiang Road, Chang 'an District, Xi 'an City, Shaanxi, China*
a. 3034144091@qq.com
**corresponding author*

Abstract: The field of text-to-music generation has witnessed remarkable progress in recent years, fueled by advancements in deep learning techniques. This survey paper provides a comprehensive overview of the state-of-the-art in this domain, focusing on three key aspects: model architectures, music representations, and training strategies. We first delve into the prevalent architectures employed for music generation, including language models, diffusion models, and hybrid approaches, analyzing their strengths and limitations. Subsequently, we explore various music representations, ranging from quantized waveforms and spectral features to latent representations, discussing their impact on the quality and expressiveness of the generated music. Finally, we examine different training strategies, such as standard diffusion, rectified flow, and knowledge distillation, highlighting their effectiveness and efficiency in optimizing model performance. Through this in-depth analysis, we aim to provide researchers and practitioners with a clear understanding of the current landscape and future directions in text-to-music generation. We also discuss open challenges and opportunities for further research, paving the way for the development of more sophisticated and versatile systems capable of generating high-fidelity and expressive music from natural language descriptions.

Keywords: Text-to-music generation, deep learning, language models, diffusion models, music representation, waveform, spectrogram, latent representation.

1. Introduction

Music, as a core medium for emotional expression and cultural inheritance, has traditionally relied on the artistic intuition and experience of professional composers. However, traditional music creation faces challenges such as efficiency bottlenecks, style rigidity, and high technical thresholds, limiting its ability to meet the demand for personalized, real-time music in the digital age. Advances in deep learning, especially the use of Transformer models in natural language processing and diffusion models in image generation, have paved the way for new cross-modal generation tasks, including Text-to-Music Generation. This emerging field seeks to synthesize high-quality music directly from natural language instructions, aligning abstract text semantics with musical features like melody, rhythm, and timbre. The technological foundation for this direction rests on three key aspects: theoretical exploration of multimodal modeling, where music generation requires learning from both symbolic text and continuous audio signals; dual empowerment of data and computational power, with large-scale music datasets and GPU capabilities enabling the training of complex generative

DOI: 10.54254/2755-2721/119/2025.21641

models; and urgent demand from application scenarios, driven by use cases in game/film music automation, personalized recommendations, and assisting disabled individuals. Current technological trends include language models like MusicLM and MusicGen, which use a phased generation strategy, and diffusion models such as Riffusion and Noise2Music, which excel in audio fidelity and style diversity. Despite these advancements, gaps remain, including a semantic gap in fine-grained alignment between text descriptions and music's emotional nuances, a structural gap in modeling long-term dependencies, and an evaluation gap due to a lack of unified objective metrics for assessing music quality. Understanding these gaps and the differences between various architectures is crucial for overcoming technical bottlenecks and advancing creative AI, with this review offering a comprehensive overview of the field's evolution and potential for human-machine co-creation.

This paper focuses on deep learning-based text-to-music generation technologies, providing a systematic analysis of the key technical frameworks, challenges, and research progress in the field. The scope is defined to include end-to-end generation models, such as Transformer-based language models and diffusion models, with an emphasis on how these models extract temporal and structural features from text semantics for music generation. It excludes traditional symbolic music generation methods like MIDI sequence generation and rule-based techniques. The paper also examines three primary representations of music signals—quantized waveforms, spectral features (e.g., Mel-spectrograms), and latent space encoding (e.g., VAE, diffusion latent variables)—analyzing their impact on generation quality, computational efficiency, and controllability. Additionally, it explores cutting-edge training strategies like standard diffusion training, Rectified Flow, and knowledge distillation, highlighting their relationship to generation performance. The core goals include: deconstructing technological architectures (e.g., MusicLM and Riffusion) to reveal the complementary strengths of language models in structured music generation and diffusion models in high-fidelity audio synthesis; diagnosing the challenges of multi-granularity alignment, long-sequence modeling, and evaluation systems; and offering forward guidance on emerging directions like cross-modal learning, neural audio encoding, and human feedback reinforcement learning (RLHF). The paper aims to provide value by constructing a triadic "architecture-representationtraining" technical analysis system, offering decision support for technical selection based on quantitative analysis, and exploring the ecological implications of generative music technology on artistic production paradigms, including copyright ethics, creator collaboration, and human-machine interaction design.

2. Model architectures for text-to-music generation

2.1. Language Models (LMs)

Language models (LMs) are widely used in text-to-music generation by converting textual input into a music representation to generate corresponding audio content. Transformer-based models, such as MusicLM [1], MusicGen [2], and MeLoDy [3], are the main approaches, utilizing self-attention mechanisms to process text and music relationships, enabling the generation of structured music over long time spans. MusicLM leverages hierarchical encoding to map text into high-dimensional feature vectors for high-quality audio generation, though it demands significant computational resources. MusicGen processes structured audio features like Mel spectrograms to generate music matching the text description, and MeLoDy innovates with multimodal learning by incorporating both text and audio features. While LMs offer advantages such as long-range dependency modeling and parallel processing, they face challenges like high computational costs and instability during generation, particularly with complex musical structures. Recurrent Neural Networks (RNNs), including LSTM and GRU variants, were previously dominant in music generation, excelling in sequential modeling but struggling with long sequences, parallelization, and generation accuracy. LMs, especially

DOI: 10.54254/2755-2721/119/2025.21641

Transformer-based models, are more flexible and controllable, allowing users to specify music style or emotion. However, they rely on large-scale data for training, may lack diversity in generation, and face challenges in precise control over complex musical expressions. Figure 1 shows the architecture for text-to-music generation.

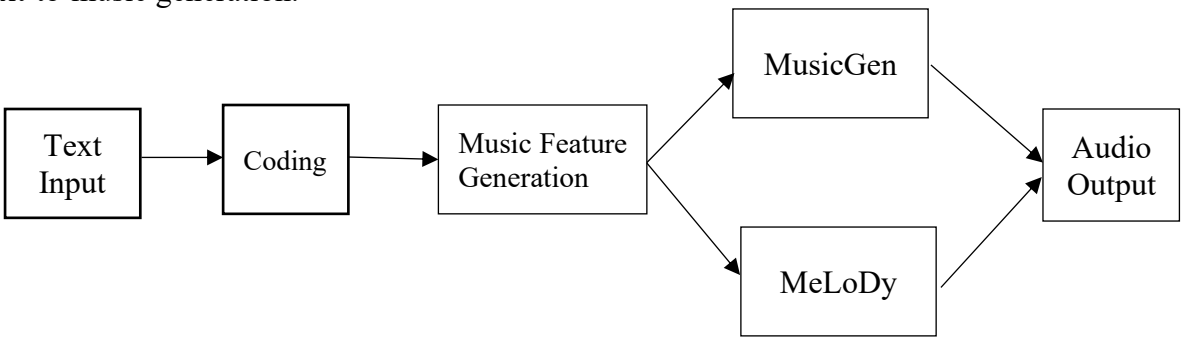

Figure 1: Architecture for Text-to-music Generation

2.2. Diffusion Models (DMs)

Diffusion Models (DMs) have advanced in generative modeling, particularly in image and audio generation, and have shown strong capabilities in text-to-music generation. These models work by progressively denoising data, consisting of a forward diffusion process where noise is added to real data until it becomes nearly pure noise, and a reverse denoising process where the model learns to recover the original data. Notable models like Denoising Diffusion Probabilistic Models (DDPM), Improved Denoising Diffusion Models (IDDPM) [4], and Score-based Generative Models enhance generation quality and diversity. Applications in text-to-music generation include Mousai [5], which generates music based on text descriptions, Noise2Music [6], which maps noise to audio space while preserving stylistic details, and Riffusion [7], which generates music from spectrograms with simple text inputs. Diffusion models are advantageous for producing complex, expressive music segments, and they generally outperform traditional models like GANs in generation quality, offering high diversity and training stability. However, their limitations include slow generation speed due to multiple denoising steps, high computational resource demands, and large data requirements, making real-time applications challenging.

2.3. Hybrid Models

Hybrid models combine different generative models, typically blending language models and diffusion models, to overcome the limitations of individual approaches and enhance the quality, controllability, and diversity of the generated output. By combining language models (such as Transformers), which understand and generate text descriptions, with diffusion models that excel in generating high-quality music, hybrid models leverage the strengths of both. The advantages of this approach include the ability of language models to extract features like emotion and style from natural language, guiding the diffusion model to generate music that aligns with these descriptions, while diffusion models ensure high sound quality and diversity. However, challenges remain in ensuring effective collaboration between the models and maintaining high-quality music generation. Other hybrid methods involve combining RNNs with diffusion models for better long-sequence modeling or using Variational Autoencoders (VAEs) alongside diffusion models to improve diversity and quality by utilizing latent space. The potential advantages of hybrid models include comprehensive benefits, such as balancing text understanding and high-quality generation, and improved control over the style, emotion, and structure of the music. On the other hand, challenges include coordination

DOI: 10.54254/2755-2721/119/2025.21641

issues between different models, the complexity of training with large labeled data, and slower generation speeds, particularly when incorporating diffusion models.

3. Music Representations

3.1. Quantized Waveform Representation

In music generation, waveform representation is a direct and accurate method of audio expression, capturing original sound information by converting audio signals into discrete digital points. This approach preserves detailed characteristics of the sound, including frequency, amplitude, and temporal features. Waveform quantization involves converting continuous analog signals into digital form through sampling and quantization techniques. Sampling divides the audio signal into discrete time points at a specific frequency, often 44.1 kHz or higher, while quantization maps the amplitude values to a finite digital range, with common depths being 16-bit or 24-bit. Quantization introduces errors, known as quantization noise, which may cause distortion, particularly at lower bit depths. The advantages of waveform representation include high fidelity, as it directly reflects the original audio signal and captures subtle variations, resulting in more realistic generated music, and rich detail expression, making it suitable for high-quality audio generation tasks. However, the approach also has limitations, such as high computational resource consumption due to the large storage and processing needs, especially for high-quality audio generation. Additionally, generating audio directly from waveforms is more challenging than from spectrograms or other features, requiring models to handle the complexity of continuous waveform generation, which places greater demands on computational capacity.

3.2. Spectral Feature Representation

Spectral feature representation converts audio signals into frequency-domain features, offering a more compact and efficient way to represent audio compared to waveform representation. By analyzing the frequency distribution of the audio signal, this method captures important characteristics of the sound. A spectrogram is created by decomposing the audio signal using the Short-Time Fourier Transform (STFT), displaying the frequency strength distribution over time, and is commonly used to represent the spectral information. The Mel-spectrogram, based on the Mel scale, simulates human auditory perception by transforming the audio signal into amplitude information across Mel frequency bands, making it particularly suitable for music generation tasks. The Mel-spectrogram's advantage lies in its ability to simulate human hearing more closely, resulting in more natural-sounding music. In music generation, spectral feature representations, particularly Mel-spectrograms, are widely used. For example, in deep learning-based music generation models, audio signals are typically converted into Mel-spectrograms for training and generation. These models can produce high-quality audio by learning spectrogram patterns, and they are also effective in modeling emotions and styles. Spectral features enable generation models to learn the patterns associated with specific emotional or stylistic characteristics, allowing the creation of music that matches given input text or emotional cues. The advantages of spectral feature representation include its compactness and efficiency, as spectrograms require fewer parameters than waveforms, reducing computational resource consumption. Additionally, since the frequency domain provides a stronger structural nature, it allows models to capture audio regularities more effectively. However, there are limitations, such as the loss of some audio quality during the transformation to spectral features, which can result in a lower fidelity compared to waveform representation. Additionally, reconstructing the original audio signal from a spectrogram, especially a Mel-spectrogram, is challenging. Although techniques like the Inverse Short-Time Fourier Transform (ISTFT) can be applied, the reconstructed audio may still suffer from detail loss and distortion.

DOI: 10.54254/2755-2721/119/2025.21641

3.3. Latent Representations

Latent representation refers to mapping audio data to a low-dimensional latent space, enabling more efficient processing and generation. Variational Autoencoders (VAEs) are one of the most widely used techniques for latent representation, compressing complex high-dimensional audio into a latent space and generating new audio from that space. A VAE consists of an encoder, which compresses the audio into a latent representation, and a decoder, which generates new audio from that latent space. The training objective of a VAE is to maximize the lower bound of the log-likelihood of the data, allowing the model to learn effective latent representations. In music generation, VAEs can produce diverse audio segments and facilitate smooth transitions by interpolating in the latent space, making them suitable for tasks like melody generation and style transfer. Other latent representation techniques include Generative Adversarial Networks (GANs) and autoregressive models. GANs use a competitive training process between a generative and a discriminative model to generate high-quality audio, while autoregressive models generate audio step-by-step by using the previous output as input for the next step. These techniques have widespread applications in audio generation. The main advantages of latent representations are dimensionality reduction, which simplifies the training process by mapping audio signals to a low-dimensional latent space, and the ability to generate diverse and flexible music, excelling in tasks like style transfer and audio transformation. However, latent representations also have limitations, such as lower generation precision compared to direct waveform generation, due to information loss during compression. Additionally, the structure of latent space is often difficult to interpret, leading to potential challenges in controlling the generated output and unpredictability in the results.

4. Training Strategies

4.1. Standard Diffusion Training

Diffusion models (DMs) have garnered significant attention for their success in generative tasks, particularly in generating images and audio. The standard diffusion training involves a two-step process: forward and reverse diffusion. In the forward diffusion process, noise is progressively added to the data, such as audio or images, with each diffusion step increasing the noise until the data becomes pure noise. This process is fixed and independent of the training data, and its main objective is to provide training data for the subsequent denoising process. The reverse diffusion process is where the model plays a crucial role. During training, the model learns to recover the original data from the noisy samples by progressively removing noise, essentially reversing the forward diffusion process. The model's goal in this phase is to restore the clear, structured data from the noisy input. The loss function commonly used in diffusion model training is based on noise removal, with the objective of minimizing the difference between the noise predicted by the model and the true noise at each time step. The mean squared error (MSE) is typically employed as the loss function, defined as:

$$L = E[\|\epsilon_\theta(x_t, t) - \epsilon\|^2]$$

where $\epsilon\theta(xt,t)$ represents the noise predicted by the model at time step t, ϵ is the real added noise, and xt is the data after t steps of diffusion. This loss function helps the diffusion model efficiently learn the denoising process, resulting in high-quality audio generation.

4.2. Flow-based Model Training

Flow-based models [8], [9] are a class of generative models that generate data through parameterized invertible transformations. These models map a simple latent distribution, such as a standard normal

DOI: 10.54254/2755-2721/119/2025.21641

distribution, to a complex data distribution, like audio data, using a series of invertible transformations. Unlike traditional Generative Adversarial Networks (GANs), which rely on a discriminator to assess data quality, flow-based models directly learn the data generation process. The architecture typically consists of multiple invertible transformations, each layer of which captures the complexity of the data distribution while ensuring that the mapping from latent space to data space remains invertible. By optimizing these transformations, the model generates samples that resemble real data. One of the key advantages of flow-based models is their fast generation speed, as the process is deterministic and does not require adversarial training like GANs. Additionally, flow-based models offer precise control over the generation process, leading to more stable and consistent output. Moreover, unlike GANs, flowbased models avoid the mode collapse problem, ensuring diversity in generated samples during training. These features have enabled flow-based models to perform well in audio generation tasks, such as music synthesis and speech generation.

4.3. Knowledge Distillation

Knowledge distillation [10] is a model compression technique designed to transfer the knowledge from a large, complex model to a smaller, more efficient model. This process enhances the efficiency of the smaller model without a significant reduction in performance. Knowledge distillation typically occurs in two stages: first, a powerful ”teacher model” is trained, and then the knowledge from the teacher is transferred to a smaller ”student model” by minimizing the difference between their outputs. The teacher model, being more complex and expressive, provides the student model with the ability to achieve similar performance despite having fewer computational resources. In the context of music generation, knowledge distillation can improve the performance of smaller generative models, making them more suitable for resource-constrained environments. For example, a large music generation model (such as a Transformer-based model) can be trained first, and its knowledge can then be distilled into a smaller model that retains the ability to generate high-quality music while significantly reducing computational overhead during inference.

5. Evaluation Metrics And Datasets

5.1. Objective Metrics for Music Generation

Objective metrics are used for quantitatively evaluating the quality of generated music and can typically be computed through automated methods, facilitating large-scale evaluations. Audio quality metrics, such as Frechet Audio Distance (FAD),´ are commonly used to measure the similarity between generated and real audio. FAD calculates the difference between the distributions of audio samples by transforming audio into feature representations, such as spectral features or Mel spectrograms, and comparing the feature distributions of generated and real audio. The process involves extracting features from audio using deep neural networks (e.g., VGGNet), calculating the mean and covariance of the feature space for both generated and real audio, and computing the Frechet distance between these two feature sets. Smaller distances indicate that the generated audio´ is more similar to the real audio. Musicality metrics, such as pitch accuracy and rhythm accuracy, evaluate the performance of generated audio in terms of melody, harmony, and rhythm. Pitch accuracy assesses whether the pitch of the notes in the generated audio matches the original notes by calculating the error between the fundamental frequency (F0) in the audio and the real notes. Rhythm accuracy evaluates whether the rhythm of the generated audio aligns with the rhythm pattern of the original music, typically analyzed by examining beats and rhythmic patterns, such as quarter notes and eighth notes. These musicality metrics help assess the quality, coherence, and expressiveness of generated music.

DOI: 10.54254/2755-2721/119/2025.21641

5.2. Subjective Evaluation Methods

Although objective metrics provide valuable quantitative evaluations, subjective evaluation methods still hold significant importance in the field of music generation. Subjective evaluations are typically performed through manual auditory tests, which can directly reflect human listeners' perceptions. Listening tests are conducted by having reviewers listen to the generated audio and score it based on various aspects, such as audio quality, musicality, emotional expression, and creativity. The evaluation dimensions include: Audio Quality, which refers to the clarity and naturalness of the audio; Musicality, which assesses the coherence of musical elements such as melody, harmony, and rhythm; Emotional Expression, which evaluates whether the generated audio conveys specific emotions like happiness or sadness; and Creativity, which measures the uniqueness and originality of the generated audio. These evaluation results provide direct feedback on the effectiveness of music generation models, guiding further improvements. However, subjective evaluation faces several challenges: Inconsistency of Evaluation Criteria, as different listeners may have different perceptions and scores for the same audio; Complexity of Evaluation, as the aesthetic standards of music are complex, making it difficult for reviewers to evaluate all dimensions of musicality accurately; and Time and Cost of Evaluation, as large-scale listening tests require significant time and resources, with potentially low repeatability. To overcome these challenges, researchers often design standardized evaluation processes, use multiple reviewers for scoring, or combine objective metrics with subjective evaluations for a more comprehensive assessment.

5.3. Common Datasets for Text-to-Music Generation

Generating high-quality music requires large training datasets. Below are some commonly used datasets for text-to-music generation research. The Million Song Dataset contains over one million songs and is widely used in music generation and recommendation system research. It provides metadata for songs, such as artist, album, and genre, which can be utilized for text-based music generation tasks. The FMA (Free Music Archive) is an open-source music dataset that includes various music genres and is suitable for music generation and analysis research. The FMA dataset provides complete audio files and associated tags for model training and evaluation. AudioSet is an audio dataset developed by Google that contains over two million audio clips. It is widely applied in audio classification, generation, and recognition tasks. The audio tags cover a variety of sound events, making it suitable for a wide range of audio generation tasks. These datasets provide rich training data for text-to-music generation, helping models learn the ability to generate music from natural language descriptions.

6. Conclusion

This paper reviews the latest developments in the field of text-to-music generation, focusing on several key aspects. Firstly, it examines model architectures, highlighting the application of language models, diffusion models, and hybrid models in music generation, along with their respective advantages and limitations. Secondly, it discusses music representation, exploring the impact of various methods such as quantized waveforms, spectral features, and latent representations on generation quality. The paper also addresses training strategies, detailing how techniques like standard diffusion training, corrected flow training, and knowledge distillation contribute to improved model performance.

DOI: 10.54254/2755-2721/119/2025.21641

References

[1] *Andrea Agostinelli, Timo I Denk, Zalan Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qing Huang, Aren Jansen, Adam Roberts, Marco´ Tagliasacchi, et al. Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325, 2023.*

[2] *Jiawei Zhang, He Yu, Tianyu Qin, Yi Zeng, Ruihua Li, Qiushi Liu, Ziyang Luo, George Zhang, Zhijie Liu, and Furu Wei Wu. Musicgen: Simple and controllable music generation with transformers. arXiv preprint arXiv:2306.05284, 2023.*

[3] *Gautam Mittal, Jesse Engel Yang, Curtis Hawthorne, Ian Simon, Wei Ping Shen, Mauro Verzetti, Antoine Caillon, Adam Roberts, Marco Tagliasacchi, and Douglas Eck. Jen-1: Text-guided universal music generation with diffusion models. arXiv preprint arXiv:2306.05284, 2023.*

[4] *Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. arXiv preprint arXiv:2102.09672, 2021.*

[5] *Andrey Pashevich, Anton Sorokin, Kundan Kumar, Branislav Raj, Carlos Cano de Oliveira, and Aaron Courville.´ Mousai: Text-conditioned music generation with long-context diffusion. In International Conference on Machine Learning, pages 8619–8628. PMLR, 2021.*

[6] *Qing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zheng Zhang, Zheng Zhang, Jia Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. arXiv preprint arXiv:2302.03917, 2023.*

[7] *Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation. 2022.*

[8] *Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. 2022.*

[9] *Patrick Esser, Shubhendu Kulal, Andreas Blattmann, Robin Entezari, Johannes Muller, Hamed Saini, Yelena Levi, Dustin Lorenz, Andreas Sauer, Florian¨ Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first International Conference on Machine Learning, 2024.*

[10] *Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512, 2022.*