제어 중심, 프롬프트 기반, 루프 특화: 2024- 2025 최신 음악 생성 모델 비교 연구

장소영, 이재호
덕성여자대학교

thdud030101@duksung.ac.kr, izeho@duksung.ac.kr

Control-Centric, Prompt-Driven, and Loop-Specialized: A Comparative Survey of 2024- 2025 Music Generation Models

Soyoung Jang, Jaeho Lee
Duksung Women's Univ.

요 약

본 논문은 2024 년 말부터 2025 년 상반기 발표된 MusiCoT, YuE, LoopGen, Symbolic RG Diffusion, DiffRhythm 다섯 가지 음악 생성 모델을 입력 모달리티, 아키텍처, 제어 메커니즘, 출력 용도의 네 가지 축으로 분류/비교하였다. CoT 기반 및 규칙 가이던스 모델(MusiCoT, Symbolic RG Diffusion)은 높은 구조 일관성과 규칙 준수율을 달성한 반면, 프롬프트/텍스트 기반 모델(YuE, DiffRhythm)은 사용자 친화적 워크플로우와 장편 합성에 강점을 보였다. LoopGen 은 학습 없이도 매끄러운 무한 루프를 신속히 생성하여 특수 목적 사운드 디자인에 적합함을 확인하였다. 각 모델은 FAD, SDR, MOS, SSI, LSS 등 객관적·주관적·구조 일관성 지표에서 상이한 성능 프로파일을 보였으며, 평가 지표 간 비교의 필요성을 시사하였다. 향후 연구에서는 벤치마크 환경 및 평가 지표의 표준화, 실시간/경량화 기법, 사용자 상호작용을 반영하는 멀티모달 제어 방안 개발이 요구된다.

Ⅰ. 서 론

최근 음악 생성 분야는 딥러닝 기술의 발전과 함께 새로운 전환점을 맞이하고 있다. 과거에는 순환 신경망(RNN)이나 생성적 적대 신경망(GAN) 계열의 모델들이 주로 사용되며 제한된 길이와 일관성 문제를 겪었으나, Transformer 기반 접근법이 장기 의존성 문제를 상당 부분 해소하면서 더욱 자연스럽고 구조화된 음악 합성이 가능해졌다. 동시에, 확산(diffusion) 모델이 잠재 공간에서 반복적으로 노이즈를 제거하는 메커니즘을 통해 고품질 오디오 생성 능력을 입증하였고, 대규모 언어 모델(LLM)을 활용해 가사나 스타일 키워드로 곡을 자동으로 합성하는 발전이 이루어졌다.

한편, 생각의 사슬(chain-of-thought; CoT) 프롬프트 및 비미분 규칙(non-differentiable rule)과 같은 제어 중심(control-centric) 기법은 모델 내부에 음악 구조나 화성 진행을 명시적으로 설계하고, 제약할 수 있도록 하여, 단순한 음향 생성 단계를 넘어 '기획된 음악 구조'의 구현 가능성을 높였다. 또한, LoopGen[1]과 같은 특수 목적 모델은 학습 없이도 원형 패딩(circular padding)을 적용해 무한 반복 루프(loopable audio)를 실시간으로 생성함으로써, 게임 배경음악이나 사운드 디자인 분야에서의 활용 가능성을 확대하고 있다.

본 논문에서는 2024 년 말부터 2025 년 상반기까지 발표된 다섯 가지 최신 음악 생성 모델—MusiCoT[2], YuE[3], LoopGen, Symbolic RG Diffusion[4], DiffRhythm[5]—을 대상으로, 입력 모달리티(input modality), 아키텍처 백본(backbone), 제어 메커니즘(control mechanism), 출력 용도(output use)의 네 가지 분류 축을 제안한다. 이를 바탕으로 각 모델의 핵심 아이디어와 성능 지표, 한계점을 비교·분석하고, 표준화된 평가 지표와 실시간·경량화, 인터랙티브 멀티모달 제어 등 향후 연구 방향을 제시함으로써, 음악 생성 연구의 현재 지형도를 조망하고자 한다.

Ⅱ. 음악 생성 모델 분류 체계

본 논문에서 다루는 다섯 가지 모델은 크게 세 가지 범주로 분류할 수 있다. 첫 번째 범주는 제어 중심 모델로, MusiCoT 와 Symbolic RG Diffusion 이 속한다. 이들 모델은 CLAP 임베딩 기반의 CoT 프롬프팅이나 non-differentiable rule 을 생성 과정에 직접 주입함으로써, 단순한 음파 생성 단계를 넘어 음악 구조와 화성 진행을 사전 설계하거나 엄격히 제약할 수 있도록 설계되었다. 이러한 접근은 생성된 음악이 기획된 형태와 일치할 뿐 아니라, 구조 분석 및 재활용에도 유리한 결과를 제공한다.

두 번째 범주는 프롬프트/텍스트 기반 모델(prompt-driven)로, YuE 와 DiffRhythm 이 이 분야를 대표한다. 이들 모델은 각각 LLaMA2 기반 파운데이션 언어 모델과 VAE-잠재공간 기반의 DiT Latent Diffusion 을 백본으로 삼아, 가사나 스타일 키워드만으로 수 분 길이의 완성도 높은 트랙을 자동으로 합성한다. 특히 YuE 는 최대 5 분 분량의 가사에서 음악으로의 생성을, DiffRhythm 은 보컬과 반주를 동시에 생성하는 초고속 파이프라인을 구현하여, 사용자 친화적인 텍스트 중심 워크플로우를 제공한다.

마지막으로 특수 목적 모델(specialized loop generation)로 분류되는 LoopGen 은 MAGNeT non-autoregressive 아키텍처에 circular padding 기법을 결합하여, 추가 학습 과정 없이도 주어진 시드 패턴(seed loop)을 무한 루프로 확장할 수 있도록 설계되었다. 이 방식은 게임 배경음악이나 인터랙티브 사운드 디자인에 필수적인 이음새 없는 루프(seamless loop) 제작에 최적화되어 있으며, 즉시 사용 가능한 경량/고속 솔루션이라는 강점을 지닌다.

이처럼 입력 모달리티, 아키텍처, 제어 메커니즘, 출력 용도 네 가지 축을 중심으로 모델을 분류함으로써, 서로 다른 설계 철학과 활용 목적을 명확히 구분할 수 있다.

Ⅲ. 음악 생성 모델 핵심 기술

MusiCoT 는 CLAP 임베딩을 기반으로 한 CoT 프롬프팅을 통해 전체 음악 구조를 먼저 설계한 뒤, Autoregressive 오디오 토큰 생성 과정을 거쳐 최종 음원을 합성하는 모델이다. 이 과정에서 사용자나 개발자가 의도한 테마, 전개 순서, 반복 패턴 등을 프롬프트로 명시할 수 있으며, 실제 실험 결과 구조 준수율(Structure Adherence Rate)이 87%에 달하고 프레셰 오디오 거리(Fréchet Audio Distance, FAD) 1.8, 평균 의견 점수(MOS) 4.1/5 의 우수한 성능을 달성했다. CoT 방식을 활용함으로써 생성 결과의 일관성과 분석 용이성을 대폭 향상시킨 것이 큰 특징이다.

YuE 는 LLaMA2 기반의 대규모 파운데이션 언어 모델을 음악 생성에 적용한 대표 사례다. 입력으로 가사만 받으면, 모델 내부에서 음절 단위 임베딩과 구조적 조건부 인코딩 과정을 거쳐 최대 5 분 길이의 트랙을 생성한다. 텍스트– 오디오 구조 일치도(SSI) 0.72, FAD 2.0, MOS 3.9/5 를 기록하여, 장편 음악 생성 시에도 가사– 멜로디 간 높은 정합성을 유지한다. 별도의 오디오 백본 없이도 자연스러운 편곡과 전개를 실현한 점이 주목된다.

LoopGen 은 학습이 필요 없는 non-autoregressive 루프 생성 모델로, MAGNeT 구조에 circular padding 을 적용하여 주어진 seed pattern 을 무한히 반복할 수 있는 오디오를 빠르게 합성한다. 별도의 추가 학습 없이도 루프 이음새 점수(Loop Seamlessness Score)를 30% 이상 개선하였으며, ABX 전환 테스트에서 45%의 정확도를 보여 원본과 구별하기 어려운 매끄러운 루프를 생성한다. 실시간 게임 사운드나 배경음 제작에 즉시 활용 가능한 솔루션이다.

Symbolic RG Diffusion 모델은 잠재 확산(Latent Diffusion) 프레임워크에 Stochastic Control Guidance 기법을 플러그인 방식으로 결합한 symbolic music generation 시스템이다. 입력으로는 피아노롤 기보와 외부 규칙 함수를 함께 사용하며, 이를 통해 음표 수준 정확도(Note-Level Accuracy) 92%와 규칙 준수율(Rule Compliance Rate) 95%를 달성하였다. 기보 수준에서 세부 음표 선택과 리듬 패턴을 엄격히 제어할 수 있어, 작곡가의 의도를 반영한 높은 정확도의 심볼릭 출력(symbolic output)을 생성하는 데 강점을 보인다.

DiffRhythm 은 VAE 잠재공간을 활용한 DiT 1111ssLatent Diffusion 모델로, 가사 텍스트와 스타일 프롬프트를 동시에 입력받아 보컬과 반주를 동시 생성하는 초고속 파이프라인을 제공한다. 최대 4 분 45 초 분량의 트랙을 불과 10 초 만에 합성할 수 있으며, 신호 대 왜곡비(SDR) 12.5 dB, FAD 2.2, MOS 4.0/5 의 성능을 기록했다. 특히 멀티스템(source-separated) 방식으로 보컬과 반주를 동시에 생성함으로써, 별도 분리나 후처리 없이 완전한 멀티트랙 음악을 즉시 활용할 수 있다는 점이 돋보인다.

Ⅳ. 음악 생성 모델 비교 분석

다섯 가지 모델을 네 가지 분류 축(입력 모달리티, 아키텍처, 제어 메커니즘, 출력 용도)과 주요 성능 지표(FAD, MOS, SDR, LSS, SSI 등) 관점에서 비교해 보면, 각 그룹 간에 명확한 장·단점이 드러난다. 제어 중심 모델인 MusiCoT 와 Symbolic RG Diffusion 은 구조 일관성 및 규칙 준수 측면에서 뛰어난 성능을 보였으나, 생성 단계에서 프롬프트 설계나 규칙 정의에 추가적인 수작업이 필요하다는 부담이 있다. 반면 프롬프트/텍스트 기반 모델인 YuE 와 DiffRhythm 은 가사나 스타일 키워드만으로도 장편 트랙을 자동 합성하는 편의성이 강점이지만, 구조적 일관성(SSI) 면에서는 제어 중심 모델에 비해 소폭 낮은 수치를 보였다. LoopGen 은 오직 루프 생성 목적에 최적화되어 즉시 사용 가능한 매끄러운 루프를 매우 짧은 시간 내에 산출할 수 있다는 점이 독보적이지만, 긴 호흡의 곡 제작이나 복잡한 구조 설계 기능은 제공하지 않는다.

성능 지표별로 살펴보면, 프레셰 오디오 거리(FAD)와 신호 대 왜곡비(SDR) 같은 객관적 측정치는 DiffRhythm 과 MusiCoT 가 비교적 낮은 값을 기록하여 고품질 오디오 합성에 강점을 나타냈으며, 평균 의견 점수(MOS) 역시 CoT 기반 및 잠재 확산(Latent Diffusion) 기반 모델이 평균 4 점대 초중반을 유지했다. 반면 LoopGen 의 루프 이음새 점수(LSS)와 ABX 정확도 지표는 루프 매끄러움 측면에서 원본과 구별이 거의 불가능할 정도로 우수했으나, 주관적 음악 감상 경험을 평가하는 MOS 는 보고되지 않아 장편 음악 품질에 대한 직접 비교는 제한적이다. Symbolic RG Diffusion 은 기보(symbolic) 생성 정확도와 규칙 준수율이 90% 이상으로 가장 높아, 악보 제작용 자동화 시스템에 적합하다. 요약하자면, 구조적 엄격성을 중시하는 응용에서는 제어 중심 모델이, 사용자 친화적 텍스트 워크플로우가 필요할 때는 프롬프트 기반 모델이, 그리고 반복 가능한 루프가 핵심인 경우에는 LoopGen 이 최적의 선택지가 된다.

Ⅴ. 결론

현재 음악 생성 모델들은 각자 특화된 장점에도 불구하고 몇 가지 공통적인 한계점을 지니고 있다. 첫째, 평가 지표의 표준화가 미흡하여 FAD, MOS 등 다양한 지표가 사용되지만 동일 환경에서의 체계적 비교가 부족하다. 둘째, 생성 속도와 연산 자원 간 균형 문제가 있어 실제 서비스를 위한 모델 경량화 및 온디바이스(on-device) 추론 기술이 필요하다. 셋째, 생성 과정에서 사용자의 취향과 수정 요청을 반영하는 인터랙티브 인터페이스 연구가 아직 초기 단계에 머물러 있다.

이러한 한계를 극복하기 위해 다음과 같은 연구 방향이 제시된다. 우선 공통 벤치마크를 통한 평가 지표 통일과 모델 간 공정한 비교가 필요하다. 또한 지식 증류, 양자화, 점진적 프롬프트 튜닝 등을 활용한 경량화로 실시간 생성 및 모바일 환경 지원을 강화해야 한다. 텍스트, 시각적 신호, 제스처 등 다양한 입력을 통해 생성 과정을 동적으로 조정하는 인터랙티브 멀티모달 제어 시스템 개발도 중요하다. 마지막으로, 심볼릭과 오디오 도메인을 융합하여 악보 수준의 정확도와 오디오 품질을 동시에 만족시키는 하이브리드 모델이 음악 생성 연구의 중요한 과제로 부상할 것이다.

참 고 문 헌

[1] Marincione, Davide, et al. "LoopGen: Training-Free Loopable Music Generation." arXiv preprint arXiv:2504.04466 (2025).

[2] Lam, Max WY, et al. "Analyzable chain-of-musical-thought prompting for high-fidelity music generation." arXiv preprint arXiv:2503.19611 (2025).

[3] Yuan, Ruibin, et al. "YuE: Scaling Open Foundation Models for Long-Form Music Generation." arXiv preprint arXiv:2503.08638 (2025).

[4] Huang, Yujia, et al. "Symbolic music generation with non-differentiable rule guided diffusion." arXiv preprint arXiv:2402.14285 (2024).

[5] Ning, Ziqian, et al. "DiffRhythm: Blazingly Fast and Embarrassingly Simple End-to-End Full-Length Song Generation with Latent Diffusion." arXiv preprint arXiv:2503.01183 (2025).