음악 생성형 AI 기술 동향

Trends in Generative Music AI Technologies

박지현 (J.H. Park, juhyun@etri.re.kr) 지능형콘텐츠인식연구실 책임연구원
김혜미 (H.M. Kim, miya0404@etri.re.kr) 지능형콘텐츠인식연구실 책임연구원
김정현 (J.H. Kim, bonobono@etri.re.kr) 지능형콘텐츠인식연구실 책임연구원

ABSTRACT

Recent advances in artificial intelligence (AI) are driving structural changes in music creation and research. Moving beyond traditional approaches that focus on models optimized for individual functions or specific tasks, the field is increasingly shifting toward foundation models capable of accommodating diverse inputs and representations while supporting the creative process as a whole. This study reviews the background and trajectory of this transition and explores how generative music AI is evolving from a simple automation tool to a partner that augments human creative capabilities. In particular, it synthesizes the technical and conceptual significance of key transitions, including the expansion of musical representations, the generalization of conditional generation and control, and the integration of model architectures. Moreover, it analyzes patterns of technical evolution and their practical implications by examining representative generative music AI models.

KEYWORDS Generative Music AI, Human-AI Collaborative Music Creation, Multimodal Music Generation, Music Foundation Models

I. 서론

지난 10여 년간 AI 기술의 급격한 발전으로 텍스트, 이미지, 비디오 등 다양한 콘텐츠 영역에서 창작 편의성이 크게 향상되었다. 특히, 2022년 이후 대규모 언어 모델(LLM: Large Language Model)과 디퓨전 기반 생성 모델[1]의 발전은 인간의 창작 활동을 AI가 실질적으로 보조하거나 일부 대체할 수 있음을 보여주었으며, 최근에는 음악 영역으로 확장되어 인간 중심적 예술인 음악의 창작을 자동화하거나 보조하는 기술로 주목받고 있다.

그림 1은 인간의 창의성과 AI의 학습 능력이 융합되어 새로운 음악 창작 패러다임을 형성하는 과정을 개념적으로 나타낸 것이다. 기존의 음악 창작에서 인간은 감정과 예술성을, 기술적 도구는 음악 편집의 편의성을 제공하거나 품질을 향상시키는 등의 기능을 제공하였으나, 현재는 음악 생성형 AI가 인간이 담당했던 창작 과정까지 일부 담당하면서 AI

* DOI: https://doi.org/10.22648/ETRI.2026.J.410209
* This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Development of Core Technologies for Copyright Verification of AI-Generated and Deepfake Music, Project Number: RS-2025-02216483).

OPEN 공공누리 출처표시 상업적이용금지 변경금지 공공저작물 자유이용허락
본 저작물은 공공누리 제4유형
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.
©2026 한국전자통신연구원

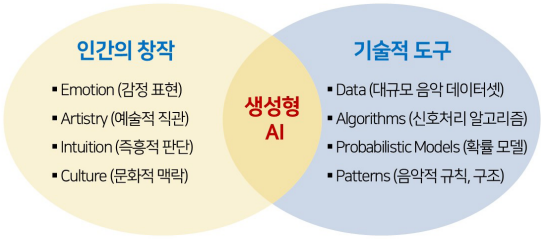


그림 1 인간-AI 창작 융합 개념도

보조 음악 창작이 이루어지고 있다.

음악은 언어와 달리 시간적 구조, 화성적 제약, 감정 표현이 동시에 요구되는 고차원 데이터로, 이를 기계적으로 생성하기 위해서는 시계열 모델링, 음색 표현, 의미 정렬 문제가 함께 해결되어야 한다. 초기 음악 생성 연구에서는 규칙 기반이나 확률 기반 접근이 주를 이루었으나, 2010년대 중반 이후 RNN[2], LSTM[3], Transformer[4] 등 딥러닝 기반 시퀀스 모델이 등장하면서 음악 패턴을 데이터로부터 학습하는 방식이 확산되었다. 이러한 심볼릭(Symbolic) 기반 모델은 멜로디와 코드 구조를 효과적으로 학습할 수 있었지만, 악기 음색이나 공간감, 보컬의 미묘한 표현과 같은 실제 음향적 특성을 충분히 재현하는 데는 한계를 보였다.

심볼릭 기반 모델의 기술적 한계를 해소하기 위해 2020년대 초반부터 오디오를 직접 생성하는 접근이 본격화되었다. 뉴럴 코덱(Neural Codec)[5]과 잠재 오디오 토큰(Latent Audio Token)[6] 기술의 도입은 음악을 압축된 토큰 형태로 표현하고, 이를 생성 모델이 직접 다룰 수 있게 하였으며, 이를 계기로 음악 생성 연구의 중심은 심볼릭 도메인에서 오디오 도메인으로 이동하였다[7,8]. 오디오를 직접 모델링하는 접근을 통해 단순한 멜로디와 코드 패턴의 조합을 넘어 음색, 감정 등 인간적 감각을 재현하는 등 보다 사실적인 음악 생성이 가능해졌고, 최근에는 완곡 단위의 음악을 생성하는 상용 서비스로까지 발전하고 있다[9].

이처럼 음악 생성형 AI가 발전하고 있지만, 한편으로는 텍스트–음악 의미 정렬의 한계, 장기 시퀀스에서의 구조적 일관성 유지, 생성 음악의 품질을 객관적으로 평가하기 위한 기준 부재 등 기술적 과제도 여전히 존재한다. 이러한 한계를 해결하려는 시도로 음악 생성형 AI는 단순한 오디오 신호 합성을 넘어, 인간의 창작 과정을 수학적 또는 확률적으로 모델링하려는 방향으로 확장되고 있으며[7], 음악의 문법 구조와 감정 표현을 동시에 학습하는 단계로 진화하고 있다.

이에 본고에서는 음악 생성형 AI의 기술적 구조와 발전 단계를 체계적으로 정리하고, 주요 기술 동향과 연구 흐름을 살펴보고자 한다.

II. 음악 생성형 AI의 발전 단계

지난 20여 년간의 음악 생성형 AI는 데이터 표현, 모델 구조, 학습 방식의 발전에 따라 여러 세대로 구분될 수 있다. 초기의 규칙 기반 작곡 시스템에서 출발하여, 딥러닝 기반 시퀀스 모델, 그리고 오디오 직접 생성을 거쳐 현재는 대규모 멀티모달 생성 모델로 발전하고 있다.

1. [1세대] 규칙 및 확률 기반 작곡 시스템

2000년대 초반까지의 음악 생성 연구는 규칙 또는 확률 모델에 기반하였다. 이 시기의 시스템은 사람이 정한 음악 규칙과 패턴을 이용해, 그 틀 안에서 자동으로 곡을 만드는 방식이었다.

대표적인 접근으로는 Markov Chain Composer [10], HMM(Hidden Markov Model) 기반 멜로디 생성기[11], Constraint 기반 음악 생성기[12] 등이 있다. 이들은 짧고 단순한 멜로디는 만들 수 있었지만, 음

악의 긴 흐름이나 감정의 변화를 표현하는 데는 한계가 있었다.

2. [2세대] 딥러닝 기반 심볼릭 생성기

2010년대 중반 이후, 딥러닝의 등장은 음악 생성 패러다임을 근본적으로 바꾸었다. RNN[2], LSTM[3], Transformer[4] 구조가 등장하면서 모델이 음악의 규칙을 직접 데이터로부터 학습할 수 있게 되었다.

대표 연구로는 구글 마젠타[13]의 MelodyRNN[14], PerformanceRNN[15], Music Transformer[16], OpenAI의 MuseNet[17], 야마하의 DeepComposer 등이 있다. 이들은 MIDI 형태의 데이터를 입력받아 음의 순서와 관계를 확률적으로 예측하는 방식으로 음악을 생성했다.

3. [3세대] 오디오 직접 생성 모델

2020년대 초반부터는 음악을 기호(Symbol)가 아닌 소리(Audio)로 직접 생성하려는 접근이 시도되었다. 핵심은 뉴럴 코덱[5]과 잠재 오디오 토큰[6] 기술로, 이를 통해 모델은 실제 음향 신호를 언어처럼 토큰화해 다루게 되었다.

대표 연구 및 모델은 OpenAI Jukebox[18], 구글의 MusicLM[19], 메타의 MusicGen[20] 등이 있다. 이 시기부터 음악 생성형 AI가 단순히 작곡을 돕는 수준을 넘어 보컬, 악기, 공간감을 함께 표현하는 완전한 곡을 만들 수 있게 되었다.

4. [4세대] 대규모 멀티모달 생성 모델

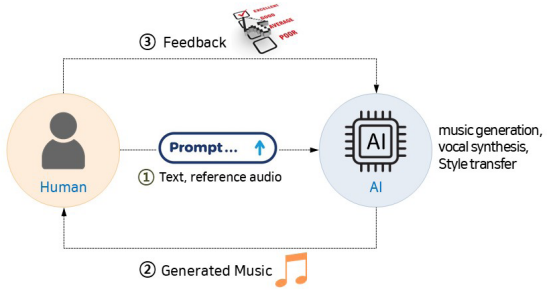
2023년 이후에는 음악 생성이 멀티모달(Multimodal) 영역으로 확장되었다. 텍스트, 영상, 보컬, 감정 등 다양한 입력을 함께 처리하는 모델들이 등장했으며, 상용화 서비스도 폭발적으로 증가했다.

대표 사례로는 3분 이상 완곡을 생성할 수 있는 Stability AI의 Stable Audio 2.0[21], 실시간 음악 생성이 가능한 통합 프레임워크인 메타의 Audiocraft[22], 텍스트 프롬프트 입력으로 완곡을 생성하는 Suno[9] 등이 있다. 이 시기 모델들은 단순한 기술 시연 단계를 넘어, 창작 도구로서의 완성도와 산업적 신뢰성을 동시에 갖추기 시작했다.

앞서 기술한 음악 생성형 AI의 발전 단계별 기술적 특징을 표 1에 정리하였다. 음악 생성형 AI의 핵심은 단순히 자동으로 음악을 만드는 기술이 아니라, 사람과 AI가 함께 음악을 만들어가는 새로운 방

표 1 음악 생성형 AI 발전단계별 기술적 특징

구분	1세대	2세대	3세대	4세대
시기	~2010	2015~2020	2020~2023	2023~현재
핵심 기술	규칙·확률 기반	RNN, Transformer	Neural Codec, Audio Tokens	Diffusion, Multimodal
대표 모델	Markov Composer	Music Transformer, MuseNet	Jukebox, MusicGen, MusicLM	Stable Audio, Suno, Udio
특징	명시적 규칙, 단순 구조	심볼릭 기반, 구조적 학습	오디오 직접 생성, 사실적 음향	완곡 생성, 상용화, 저작권 체계 정립
한계점	창의성 부족, 감정 표현 불가	음색·공간감 표현 불가	높은 계산량, 프롬프트와의 일치 한계	평가 기준 미비, 감정 일관성 부족



출처 게티이미지뱅크, 무단 전재 및 재배포 금지

그림 2 인간-AI 협력적 음악 창작 과정

식에 있다. 사람은 감정, 아이디어, 의도 같은 창작의 핵심 요소를 제공하고, 이를 텍스트와 멜로디 형태의 프롬프트로 AI에 전달한다. AI는 이 정보를 바탕으로 음악의 패턴과 구조를 학습해 사람이 상상한 아이디어를 실제 소리로 표현된 음악으로 바꾼다. 이렇게 만들어진 결과물은 완성된 곡이라기보다 사람이 다시 들어보고 수정할 수 있는 음악 초안에 가깝다. 사람은 AI가 만든 음악을 들으면서 자신이 의도한 감정, 리듬, 분위기가 잘 담겼는지 살펴본다. 만족스럽지 않다면 프롬프트를 바꾸거나 일부 구간을 편집해 AI가 더 나은 결과를 만들 수 있도록 피드백을 준다. 이런 식의 반복적인 상호작용은 AI가 점점 더 사람의 의도를 잘 이해하고 반영하도록 돕는다. 결국 음악 생성형 AI는 사람을 대신하는 존재가 아니라, 함께 음악을 만들어가는 창작 파트너로 볼 수 있다. 그림 2는 인간-AI 협력적 음악 창작 과정을 나타낸다.

음악 생성형 AI는 이제 단순히 음을 예측하는 모델을 넘어, 인간의 감정과 맥락을 이해하며 표현하는 인공 창작자의 단계로 진입하고 있다. 향후에는 AI가 단순히 음악을 잘 만드는 것에 그치지 않고, 법적 · 윤리적으로 신뢰할 수 있으며, 사람이 함께 참여해 창작할 수 있는 방향으로 기술적 발전이 진행될 것으로 전망된다[23].

III. 음악 생성형 AI 기술 구성요소

음악 생성형 AI는 단일 모델이 아닌 여러 AI 구성요소가 단계적으로 결합된 복합 시스템이다. 텍스트, 멜로디 등 다양한 입력을 해석하고, 이를 시간적/주파수적 구조로 변환한 뒤, 최종적으로 인간이 감상할 수 있는 오디오로 복원하는 과정을 거친다. 이 장에서는 음악 생성형 AI를 구성하는 5가지 핵심 기술 요소들을 설명한다.

1. 데이터 표현

AI가 음악을 이해하고 생성하기 위해서는 소리를 수학적 형태로 표현하는 과정이 필요하며, 그 방식에 따라 모델이 학습할 수 있는 정보 범위가 달라진다. 음악 생성형 AI에서 주로 사용되는 데이터 표현 방식은 다음과 같이 구분된다.

1.1 심볼릭(Symbolic) 표현

심볼릭 표현은 음악을 음높이(Pitch), 길이(Duration), 세기(Velocity), 코드(Chord), 템포(Tempo) 등 악보와 유사한 기호 수준으로 표현하는 방식이다. 이러한 표현은 화성 진행이나 리듬 구조와 같은 음악적 규칙을 학습하는 데 적합하며, 작곡 단계에서 효과적으로 활용된다. 그러나 실제 음향의 질감이나 음색, 공간감은 포함하지 못하므로, 생성 결과를 실제 소리로 변환하기 위해서는 별도의 합성 과정이 필요하다.

1.2 스펙트럼(Spectrum) 기반 표현

스펙트럼 기반 표현은 오디오를 시간-주파수(Time-Frequency) 영역으로 변환한 후, 소리의 에너지를 시각적으로 표현한 것이다. 이 방식은 음색과 공간적 특성, 발음 등 소리의 물리적 특징을 포착할

수 있으며, 음악을 일종의 이미지로 다루기 때문에 CNN이나 비전 트랜스포머(Vision Transformer)와 같은 구조를 적용하기 쉽다[24]. 다만, 스펙트로그램을 다시 오디오 파형으로 복원하려면 별도의 보코더(Vocoder)가 필요하며, 이 과정에서 고주파 정보나 세부 질감이 손실될 수 있다[25].

1.3 잠재 오디오 토큰(Latent Audio Token)

최근에는 뉴럴 코덱을 활용해 오디오를 압축된 토큰 시퀀스로 변환하는 방식이 널리 사용되고 있다. 예를 들어, SoundStream[5]이나 EnCodec[6]과 같은 모델은 긴 오디오를 짧은 숫자 코드로 압축해 소리를 일종의 언어처럼 다룰 수 있게 한다. 이렇게 만들어진 잠재 오디오 토큰은 음악의 음색, 보컬의 질감, 잔향 등 오디오의 사실적인 특징을 보존하면서도 효율적인 학습을 가능하게 한다.

2. 모델 구조

음악 생성형 AI의 모델 구조는 시간에 따른 음악 데이터를 어떻게 예측하느냐에 따라 구분할 수 있다.

2.1 Autoregressive 모델

Autoregressive 모델은 이전까지 생성된 음들을 기반으로 다음 음을 하나씩 예측하는 순차적 방식이다[1]. 텍스트를 생성하는 GPT 계열과 유사한 구조로, 짧은 구간에서는 음의 흐름과 구조를 잘 유지할 수 있다. 하지만, 생성 속도가 느리고 긴 시퀀스에서는 오류가 누적되기 쉽다.

2.2 Non-Autoregressive 모델

Non-Autoregressive 모델은 전체 시퀀스를 병렬로 예측하여 속도와 효율성을 높인다[26]. 이들은 트랜스포머나 컨포머(Conformer) 구조[27]를 사용하되, 학습 중 노이즈 마스킹을 통해 문맥을 복원하는 형태로 훈련된다. 이 방식은 음악을 빠르게 생성할 수 있는 것이 장점이나, 전역적인 문맥 정보를 효과적으로 반영하는 데 한계가 있다.

2.3 디퓨전(Diffusion) 기반 모델

디퓨전 모델은 이미지 생성에서 우수한 결과를 보인 확률적 노이즈 제거 구조를 오디오 도메인에 적용한 것이다[28]. 이 구조는 초기에는 무작위 잡음으로 시작해 점진적으로 진짜 소리를 복원하는 과정에서 음악의 질감과 감정을 자연스럽게 재현한다. 특히, 병렬 생성이 가능하고 긴 시간 구간을 안정적으로 생성할 수 있다는 장점이 있다. 반면에 전체 오디오의 긴 구조적 패턴을 명시적으로 모델링하지 않으므로 시간상으로 먼 구간 사이의 일관성이 낮다는 한계가 있다.

2.4 하이브리드(Hybrid) 모델

최근에는 Autoregressive 모델과 디퓨전 모델을 결합한 하이브리드 구조도 연구되고 있다. 예를 들어, Autoregressive 모델로 곡의 큰 구조(벌스, 후렴 등)를 먼저 생성하고, 디퓨전 모델로 세부 음향과 공간감을 다듬는 식이다[7]. 이 방식은 곡 전체 구성과 음질을 모두 고려할 수 있지만, 모델이 복잡해지고 학습 비용이 많이 든다.

3. 입력 조건 제어

음악 생성형 AI의 핵심은 사용자의 의도에 맞는 음악을 만들어내는 것에 있다. 즉, 텍스트, 멜로디, 장르, 보컬 등의 조건을 입력받고 이에 맞는 음악을 생성한다. 입력 조건 제어는 입력의 형태와 목적에 따라 구분할 수 있다.

3.1 텍스트 기반 제어(Text-to-Music)

자연어로 된 설명을 입력받아 그 의미에 맞는 음악을 생성하는 방식으로[29], 예를 들어 “잔잔한 피아노 발라드”나 “빠른 템포의 록 밴드 연주”와 같은 문장을 프롬프트로 입력하면, 모델은 문장 속 감정, 악기, 리듬 정보를 분석하여 대응되는 음악을 만든다. 텍스트 조건은 보통 CLAP(Contrastive Language-Audio Pretraining)[6], 또는 자체적인 멀티모달 임베딩 네트워크를 통해 벡터화되어, 트랜스포머 기반 모델의 인코더나 디코더에 조건으로 입력된다[30].

3.2 음악 기반 제어(Music-to-Music)

기존의 멜로디, 코드, 리듬, 혹은 악기 트랙을 입력으로 받아 이를 변형하거나 확장해 새로운 음악을 생성하는 방식이다. 예를 들어 사용자가 짧은 기타 리프를 입력하면, 모델은 이를 기반으로 드럼과 베이스를 추가해 완성된 곡으로 발전시킨다[8].

3.3 다중 조건 제어(Multi-Condition Control)

텍스트, 멜로디, 장르, 감정, 악기 구성 등 여러 형태의 조건을 동시에 결합하는 방식이다. 예를 들어 “잔잔한 피아노 선율로 시작해 감정이 점차 고조되는 영화 음악”이라는 프롬프트를 입력하면, 모델은 텍스트의 감정적 흐름과 음악적 진행을 함께 반영해 시간상으로 변하는 구조를 가진 음악을 생성할 수 있다. 이 방식은 음악의 다층적 특성을 포괄적으로 표현할 수 있다는 장점이 있어, 최근 Suno[9]나 Udio[31]와 같은 상용 서비스에서 활발히 활용되고 있다.

3.4 계층적 조건 결합(Hierarchical Conditioning)

하나의 조건을 모든 단계에 동일하게 적용하지 않고, 모델의 단계별로 다른 수준의 조건을 반영하는 방법이다[32]. 예를 들어 초반 단계에서는 곡의 전반적 분위기나 장르, 감정 등 거시적인 조건을 적용하고, 후반 단계에서는 악기 종류, 공간감, 음색 같은 미세한 요소를 반영한다. 조건 결합은 모델이 언제, 어떤 조건을, 어느 정도 비중으로 반영할지를 결정하는 의미적 융합의 과정이라고 할 수 있다. 이 영역은 음악 생성형 AI의 감정 표현력과 프롬프트 해석력을 결정하는 핵심 기술로 자리 잡고 있으며, 최근에는 사용자 피드백이나 시간적 감정선을 실시간으로 반영하는 동적 조건 제어 연구로 확장되고 있다.

4. 학습 방법

음악 생성형 AI는 음악 데이터의 복잡성과 생성 과정의 다양성으로 인해 단일 학습 방식보다는 여러 학습 방식을 결합한 형태로 발전해 왔다. 특히, 음악의 전체 구조와 세부 음향 특성을 동시에 반영하기 위해 다양한 접근방법이 함께 활용되고 있다.

4.1 조건부 학습(Conditional Training)

조건부 학습은 텍스트, 멜로디, 감정, 장르 등 입력 조건과 생성 음악 간의 대응 관계를 모델이 직접 학습하도록 하는 방식이다. 이를 통해 사용자의 의도를 반영한 음악 생성이 가능해지며, 텍스트-음악 의미 정렬과 감정 제어 성능을 향상시킨다. 최근에는 다중 조건을 동시에 입력받아 학습하는 방식이 일반화되고 있으며, 조건 간 상호작용의 효과적 반영을 위해 크로스 어텐션(Cross Attention) 구조[4]가 활용되고 있다.

4.2 계층적 학습(Hierarchical Training)

계층적 학습은 음악을 서로 다른 시간 해상도의 구조로 분해하여 학습하는 방식이다[33]. 상위 계

층에서는 곡의 전반적인 구조, 섹션 구성, 긴 호흡의 흐름을 학습하고, 하위 계층에서는 리듬, 음색, 미세한 음향 변화를 학습한다. 이러한 접근은 장기 시퀀스에서의 구조적 일관성 문제를 완화하는 데 이바지하며, 최근 완곡 단위 음악 생성을 가능하게 한 핵심 학습 전략 중 하나로 평가된다.

4.3 사전학습 및 미세조정(Pre-training & Fine-tuning)

대규모 음악 데이터로 사전학습을 수행한 후, 특정 장르나 용도에 맞게 미세조정을 하는 방식은 음악 생성형 AI에서도 중요한 학습 전략으로 자리 잡았다. 사전학습 단계에서는 일반적인 음악 패턴과 음향 특성을 학습하고, 미세조정 단계에서는 스타일, 언어, 보컬 특성 등 세부 요소를 조정함으로써 생성 결과의 제어성과 품질을 동시에 향상시킨다[34].

4.4 보상 기반 미세조정(RLHF for Music Generation)

최근에는 텍스트 생성 모델에서 활용되는 인간 피드백 강화학습(RLHF: Reinforcement Learning from Human Feedback) 기법[35]이 음악 생성에도 적용되고 있다[36]. 이 접근은 단순히 데이터에 기반한 지도학습을 넘어, 사용자의 주관적 선호나 평가를 직접 반영하여 모델을 조정한다는 점에서 의미가 크다. 이 접근은 아직 연구 초기 단계지만, 향후 AI-사용자 상호작용 기반 창작으로 발전할 가능성이 높다.

음악 생성형 AI의 학습은 “무엇을, 어떤 순서로, 어떤 목표로 학습할 것인가”에 대한 종합적 설계 과정이라 할 수 있다. 조건부 학습은 모델의 방향성을 결정하고, 계층적 학습은 구조적 일관성을 보장하며, 사전학습 및 미세조정은 모델의 표현력을 확장하고, 보상 기반 미세조정은 감성적 완성도를 더한다. 이 네 가지 축은 상호 보완적으로 작용하여, AI가 단순한 소리 생성기를 넘어 의미를 이해하고 감정을 표현하는 창작자형 모델로 진화하는 기반이 되고 있다.

5. 디코딩과 후처리

디코딩(Decoding)과 후처리(Post-processing)는 음악 생성형 AI 시스템의 마지막 단계로, 모델이 생성한 스펙트로그램이나 잠재 오디오 토큰과 같은 중간 표현을 실제로 사람이 감상할 수 있는 오디오 신호로 변환하고 품질을 향상시키는 역할을 한다. 이 단계는 생성 모델의 성능을 최종 사용자 경험으로 연결하는 핵심 과정으로, 결과물의 음질과 현실감을 크게 좌우한다.

5.1 디코딩(Decoding)

디코딩 과정은 모델이 생성한 스펙트럼 표현이나 잠재 오디오 토큰을 시간 영역의 오디오 파형으로 복원하는 단계이다. 스펙트럼 기반 모델에서는 보코더를 사용하여 주파수 영역 정보를 파형으로 변환하는데, 최근에는 WaveNet[37], HiFi-GAN[38] 등 딥러닝 기반 보코더가 주로 활용되고 있다.

뉴럴 코덱 기반 모델의 경우, 디코더는 잠재 오디오 토큰 시퀀스를 직접 고해상도 오디오로 복원한다. 이 방식은 음색, 공간감, 잔향 같은 세밀한 요소까지 정밀하게 재현할 수 있으며, 생성 과정 전반을 End-to-End로 통합할 수 있다는 장점이 있다. 디코딩 단계의 품질은 전체 음악 생성 결과의 자연스러움과 현실감에 직접적인 영향을 미친다.

5.2 후처리(Post-processing)

후처리는 디코딩된 오디오의 품질을 상용 수준

으로 끌어올리기 위한 과정으로, 정규화(Normalization), 이퀄라이징(Equalization), 다이내믹 레인지 조정, 공간감 부여(Reverberation) 등이 포함된다. 이러한 처리는 음량 균형을 맞추고, 청취 환경에 적합한 사운드를 구현하는 데 필수적이다.

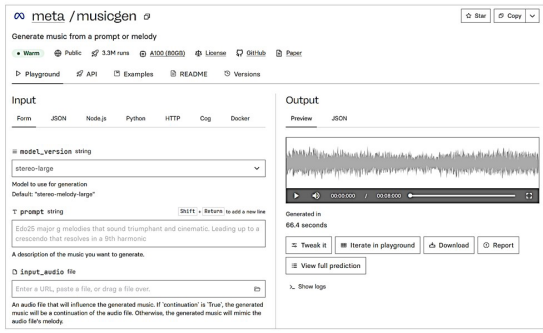
전통적으로 후처리는 규칙 기반 신호 처리 기법에 의존해 왔으나, 최근에는 후처리 과정 자체를 학습 기반으로 통합하려는 시도도 이루어지고 있다. 예를 들어 디코딩 단계와 후처리 단계를 하나의 신경망으로 결합함으로써, 생성된 음악의 음질을 자동으로 보정하고 스타일 일관성을 유지하려는 접근이 연구되고 있다[5,6]. 이러한 접근방법은 AI 작곡 결과물이 인공적 느낌에서 벗어나 진짜 스튜디오에서 제작된 음악처럼 들리도록 만드는 핵심 요소로 주목받고 있다.

IV. 대표적인 음악 생성형 AI 모델

이 장에서는 앞서 살펴본 음악 생성형 AI의 핵심 기술 요소들이 실제 모델에서 어떻게 구현되고 있는지를 대표적인 음악 생성형 AI 사례를 통해 분석한다. 각 모델은 데이터 표현 방식, 모델 구조, 입력 조건 제어, 학습 전략 측면에서 서로 다른 접근을 취하고 있으며, 이를 통해 음악 생성 기술의 발전 방향을 확인할 수 있다.

1. MusicGen

MusicGen[20]은 메타가 2023년에 공개한 AI 음악 생성 모델로, 문장이나 멜로디를 입력하면 그에 어울리는 음악을 만들어 준다. 이 모델은 EnCodec 기반 뉴럴 오디오 코덱[6]으로 오디오를 압축하여 음악 토큰(Music Token)으로 표현하고, 트랜스포머를 이용해 이 토큰 시퀀스를 직접 예측한다. 즉, 언어



출처 Reprinted from Replicate.com. https://replicate.com/meta/musicgen

그림 3 Replicate.com의 MusicGen 데모

모델이 단어를 예측하듯 음악 모델이 음향 토큰을 예측하는 구조이다.

그림 3의 MusicGen 데모에서 보이는 바와 같이 MusicGen은 입력으로 텍스트와 참조 멜로디를 받을 수 있는데, 참조 멜로디는 MIDI 또는 오디오 형태로 입력할 수 있다. 텍스트 임베딩은 CLAP[6]과 유사한 크로스모달 인코더로 변환되어 음악 생성의 조건으로 사용되며, 멜로디 입력은 기존 음악과 조화를 유지하는 방향으로 모델을 제약한다. 이 모델의 특징 중 하나는 단일 스테이지(Single-Stage) 구조라는 것이다. 이전의 Jukebox[18]나 MusicLM[19]처럼 여러 해상도의 모델을 단계적으로 쌓지 않고, 하나의 트랜스포머가 오디오 토큰 전체를 직접 예측한다. 그 덕분에 계산 효율이 높고, 실시간 음악 생성이 가능한 수준의 속도를 갖는다.

MusicGen은 공개 이후 학계와 업계에서 조건부 음악 생성의 사실상 표준 베이스라인이 되었다. 보컬 통합, 악기 스타일 파인튜닝, 감정 제어 모델 등 다양한 후속 연구가 MusicGen의 구조를 기반으로 발전했다. 다만, 기본 모델은 여전히 반주 중심으로 보컬을 포함한 완곡을 자연스럽게 생성하는 것에 제약이 있고 긴 곡 구조나 감정의 세밀한 변화 표현이 제한적이다.

2. LLambada

LLambada는 입력된 보컬 오디오와 텍스트 프롬프트를 조건으로 받아 고품질 반주 오디오를 생성하는 조건부 생성 모델로 SongGen-AI에 의해 개발되었다. 해당 모델은 GitHub를 통해 소스코드가 공개되었으며, 이를 기반으로 한 상용 서비스는 현재 개발 중이다[39].

LLambada의 주요 특징은 오디오-텍스트 정렬 모델과 2단계 생성 파이프라인을 통합한 구조에 있다. 이 모델은 반주를 직접 생성하는 대신, 생성 과정을 의미론적 생성 단계와 음향 생성 단계로 분리하여 효율성과 제어 가능성을 향상시켰다.

의미론적 생성 단계(Semantic Generation Stage)에서는 반주의 고수준 구조와 리듬을 나타내는 의미론적 토큰을 생성한다. 입력 보컬 오디오는 MERT 인코더[40]를 통해 음악적 · 감정적 특징을 반영한 임베딩으로 변환되며, 텍스트 프롬프트는 CLAP 모델[6]을 사용해 오디오-텍스트 공간에 정렬된 임베딩으로 변환된다. 이후 트랜스포머 기반 언어 모델이 이들 정보를 조건으로 받아 반주의 의미론적 코드 시퀀스를 Autoregressive 방식으로 생성한다.

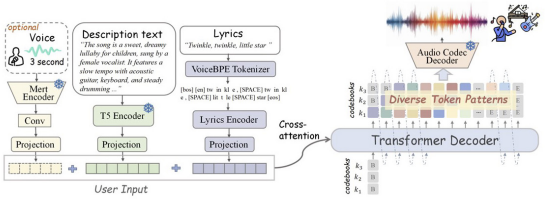
음향 생성 단계(Acoustic Generation Stage)에서는 의미론적 토큰과 저수준 음향 토큰을 입력으로 받아 고해상도 오디오 파형을 생성한다. 이 과정에서 Encodec 디코더[6]가 사용되며, 오디오 파형을 직접 예측하지 않고 뉴럴 오디오 코덱 기반 토큰을 활용하여 계산 효율을 높인다.

LLambada는 CLAP[6]을 통해 텍스트 프롬프트에 포함된 장르, 악기, 분위기 정보를 효과적으로 반영할 수 있으며, 보컬로부터 추출한 리듬과 화음 정보와 결합하여 자연스럽고 조화로운 반주를 생성한다. 이처럼 LLambada는 보컬과 텍스트 정보를 통합해 음악의 구조를 설계한 뒤 이를 실제 오디오로 변환하는 방식으로 작동하는 모델이다.

3. SongGen

SongGen은 문장으로 노래를 만드는 AI를 목표로 한 모델[8]로 코드가 GitHub에 공개되었다[41]. 그림 4는 SongGen의 개념도를 보이고 있는데, 기존의 가사 → 멜로디 → 보컬 → 반주의 단계적으로 이어지는 멀티스테이지 접근을 통합하여 하나의 Autoregressive 트랜스포머 모델이 가사, 보컬, 반주를 동시에 생성하는 단일 구조를 제안했다. 이 모델은 텍스트 설명, 가사, 3초 정도의 참조 음성을 조건으로 받아들인다. 텍스트는 곡의 분위기나 장르를, 가사는 언어적 내용을, 참조 음성은 보컬의 음색 정보를 제공하며, 세 조건이 합쳐져 하나의 토큰 시퀀스로 변환된다. 트랜스포머는 이를 언어 모델처럼 처리하여 보컬과 반주를 동시에 갖춘 음악적 문장을 예측한다.

SongGen은 출력 방식을 Mixed Mode와 Dual-Track Mode로 구분한다. Mixed Mode는 보컬과 반주가 합성된 음악을 생성하는 방식으로 원하는 결과물을 빠르고 간단하게 생성할 수 있지만 후처리 여지는 적다. Dual-Track Mode는 보컬과 반주를 별도 트랙으로 생성하여 믹싱이나 편집에 유연하다. 특히, Dual-Track Mode는 실제 작곡 워크플로우와



출처 Reprinted with permission from Z. Liu et al., "SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation," in Proc. Int. Conf. Mach. Learn., (Vancouver, Canada), July 2025, pp. 38351-38364.

그림 4 SongGen 개념도

유사하므로 음악 제작 도구와 결합이 용이하다.

SongGen의 가장 큰 의의는 가사, 보컬 정보까지 통합적으로 처리함으로써 Text-to-Song의 완전한 자동화를 오픈소스로 구현했다는 점이다. 다만, 보컬의 자연스러운 억양이나 장기적 구조의 일관성, 감정 표현력 등은 여전히 개선 과제로 보인다.

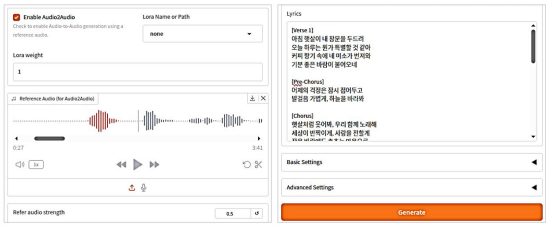


그림 5 ACE-Step의 음악 생성 예시

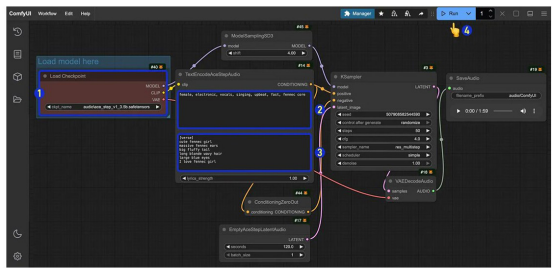
4. ACE-Step

ACE-Step은 최근 AI 음악 생성 분야에서 가장 주목받는 오픈소스 기반의 대형 음악 생성 모델이다[9]. 이 모델은 단순히 문장을 음악으로 바꾸는 AI를 만드는 것을 넘어, 다양한 음악 관련 작업(작곡, 보컬 합성, 가사 동기화 등)을 빠르고 안정적으로 수행할 수 있는 파운데이션 모델을 구축하는 것을 목표로 한다. 기존의 음악 생성 모델들은 보통 빠르게 만들면 품질이 떨어지고, 품질을 높이면 속도가 느려지는 문제를 겪었다. ACE-Step은 이 속도와 음악적 일관성 간의 균형을 잡는 데 초점을 맞추고 있다.

ACE-Step은 디퓨전 모델과 딥 압축 오토인코더(DCAE: Deep Compression Auto Encoder)[42]를 함께 사용하여 기존의 대규모 언어 모델 기반 시스템보다 약 15배 빠르게 음악을 생성한다. 저자에 따르면 NVIDIA A100 GPU 환경에서 약 4분짜리 음악을 20초 안에 생성할 수 있다고 한다. 기존의 일반적인 확산 모델은 긴 곡을 만들 때 흐름이 깨지거나 반복이 어색해지는 문제가 있었지만, ACE-Step은 멜로디, 리듬, 하모니 간의 구조적 연결성을 유지하여 긴 곡에서도 자연스럽고 완성도 높은 결과를 생성하며, 가사-음정 정렬을 모델 내부에서 자동으로 수행한다.

이러한 성능은 4개 핵심 기술의 통합을 통해 구현된다. (1) 확산 기반 생성(Diffusion-Based Generation)은 빠른 합성과 세밀한 음향 제어를 가능하게 한다. (2) 딥 압축 오토인코더(DCAE)는 오디오를 압축된 잠재 코드로 표현해 계산 효율과 음질을 동시에 확보한다. (3) 경량 선형 트랜스포머(Lightweight Linear Transformer)는 긴 음악 시퀀스를 효율적으로 처리하여 곡 전체의 흐름을 안정적으로 유지한다. (4) REPA(Representation-Predictor-Adapter) 구조는 MERT[40]와 mHuBERT[43]와 같은 표현 학습 기반 인코더를 활용하여 오디오와 텍스트 간 의미 정렬을 수행함으로써 감정적으로 일관된 음악 생성을 지원한다.

ACE-Step은 또한 텍스트 기반 스타일 및 구조 제어, 특정 구간의 선택적 재생성, 가사 수정, 스타일 변형 등 다양한 고수준 제어 기능을 제공하여 창작자의 개입 가능성을 크게 확장하는데, 그림 5에 보이는 바와 같이 공개된 소스 코드에 포함되어 있는 음악 생성 데모 페이지를 통해 이를 확인해 볼 수 있다. Apache 2.0 라이선스로 공개되어 상업적 활용이 가능하며, ComfyUI[44] 및 웹 UI와의 연동,



출처 Reprinted from ACE-Step GitHub. https://github.com/ace-step/ACE-Step

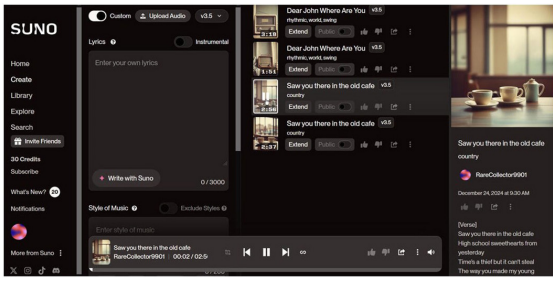
그림 6 ACE-Step에서의 ComfyUI 사용 예시

LoRA[45] 및 ControlNet[46] 기반 미세조정을 통해 커스터마이징도 용이하다. 그림 6은 ComfyUI를 활용한 예시를 보인다. 아직 감정 표현이나 보컬 자연스러움 측면에서는 추가 연구가 필요하지만, 속도, 구조, 제어를 균형 있게 결합한 차세대 연구형 음악 생성 파운데이션 모델로서 높은 연구적 가치를 지닌다.

5. Suno

Suno는 사용자가 입력한 텍스트 프롬프트를 기반으로 보컬, 가사, 악기 구성을 포함한 완전한 노래를 생성하는 End-to-End 음악 생성형 AI 시스템이다[9]. 그림 7은 SUNO를 이용한 음악 생성 예로써, 단순한 반주나 멜로디 합성을 넘어 가사와 감정 표현이 결합된 완성된 곡을 자동으로 생성하는 점이 핵심적 차별점이다.

Suno의 내부 구조와 학습 방식은 공개되지 않았으나 여러 AI 기술을 통합한 하이브리드 스택 구조로 작동하는 것으로 추정된다. 이러한 구조는 음악의 장기적 구조 일관성과 고음질 음향 품질을 동시에 달성하기 위한 설계로 이해할 수 있다. Suno의 음악 생성 과정은 프롬프트 해석 및 구조 설계, 핵심 음악 및 보컬 생성, 후처리 및 마스터링의 세 단계로 구성된다[47].



출처 Reprinted from L. Whitney, "How to Generate Your Own Music with the AI-Powered Suno," ZDNet, 2024. 12. 26. https://www.zdnet.com/article/how-to-generate-your-own-music-with-the-ai-powered-suno/

그림 7 SUNO를 이용한 음악 생성 예시

프롬프트 해석 및 구조 설계 단계에서는 사용자가 입력한 텍스트를 장르, 분위기, 템포, 주제 등의 음악적 속성으로 분석하고, [Verse], [Chorus], [Bridge]와 같은 구조 태그를 인식하여 곡의 전체 구성과 전환 패턴을 설계한다. 이 과정에는 대규모 언어 모델 또는 트랜스포머 기반 시퀀싱 모델이 활용되어 장기적인 음악 구조 형성을 담당한다.

핵심 음악 및 보컬 생성 단계에서는 트랜스포머와 디퓨전 모델을 결합한 하이브리드 생성 방식이 사용된다. 트랜스포머는 멜로디, 화성, 리듬 등 음악적 패턴을 예측해 곡의 구조를 형성하고, 디퓨전 모델은 이를 실제 오디오로 변환하며 사운드 질감을 조정한다. 이후 가사에 맞춰 멜로디 라인을 생성하고, AI 보컬 합성기를 통해 자연스러운 발음과 감정 표현을 갖춘 보컬을 합성한다.

마지막으로 후처리 단계에서는 생성된 오디오의 음량과 공간감을 조정해 상용 음원 수준의 완성도를 확보한다. 이러한 통합 파이프라인을 통해 Suno는 한 줄의 프롬프트만으로 완성된 곡을 생성하는 End-to-End 음악 생성 경험을 제공한다.

V. 결론

지난 10여 년간 음악 생성형 AI는 단순한 음표 조합을 넘어, 인간의 감정과 언어, 음향적 감각을 포괄하는 복합적 창작 시스템으로 발전해왔다. RNN과 LSTM 기반의 초기 작곡 알고리즘이 음악의 형식을 학습하는 단계였다면, 뉴럴 코덱과 디퓨전 모델을 결합한 최신 시스템은 음악의 질감과 감정을 재현하는 수준에 이르렀다. 특히, 2023년 이후 공개된 MusicGen, SongGen, LLambada, ACE-Step, Suno 등은 텍스트, 보컬, 음향을 통합적으로 다루며 AI가 음

악의 구조를 이해하고 설계하는 능력을 보여주었다.

음악 생성형 AI의 발전은 세 가지 기술적 축을 중심으로 이루어졌다. 첫째, 표현 단위가 심볼릭에서 오디오로 확장되면서 음색, 공간감, 다이내믹 등 물리적 청각 특성을 직접 모델링할 수 있게 되었다. 둘째, 텍스트, 멜로디, 장르, 감정, 악기 구성 등 다중 조건을 결합하는 조건부 제어 기술이 정교화되며 음악적 일관성과 표현력이 동시에 향상되었다. 셋째, 트랜스포머, 디퓨전 모델, 뉴럴 코덱, 보컬 합성 기술이 하나의 파이프라인으로 통합되며 텍스트 한 줄로 완곡을 생성하는 End-to-End 시스템이 가능해졌다.

이러한 진보는 기술적 성과를 넘어 음악 창작의 패러다임 전환을 촉발하고 있다. AI는 단순한 작곡 보조 도구를 넘어 공동 창작자로서의 가능성을 보이고 있으며, 그간 높은 숙련도를 요구했던 전문 음악 제작 영역이 AI 플랫폼을 통해 대중적 창작 생태계로 빠르게 재편되고 있다. 한편 텍스트와 음악 간 감정 정합성, 장기 시퀀스에서의 구조적 일관성, 스타일 모방 및 보컬 합성과 관련된 저작권, 퍼블리시티권 문제 등은 여전히 해결이 필요한 과제로 남아 있다.

용어해설

디퓨전 모델(Diffusion Model) 무작위 잡음에서 시작해 점진적으로 의미 있는 데이터를 복원하는 방식으로 결과물을 생성하는 확률 기반 생성 모델

심볼릭(Symbolic) 표현 음악을 실제 소리가 아닌 음높이, 길이, 세기, 코드 등 기호 정보로 표현하는 음악 데이터 표현 방식

뉴럴 코덱(Neural Codec) 딥러닝으로 오디오를 고도로 압축하면서도 고음질로 복원할 수 있도록 하고, 소리를 AI가 다루기 용이한 토큰 형태로 변환하는 기술

잠재 오디오 토큰(Latent Audio Token) 소리 데이터를 AI가 처리하기 쉽도록 핵심 특징만 뽑아 압축한 디지털 음악 조각

Autoregressive 모델 이전에 생성된 결과를 기반으로 다음 요소를 순차적으로 예측하여 음악을 생성하는 모델 구조

Non-Autoregressive 모델 전체 시퀀스를 병렬적으로 예측하여 빠른 음악 생성을 가능하게 하는 모델 구조

멀티모달(Multimodal) 텍스트, 오디오, 이미지 등 서로 다른 형태의 데이터를 동시에 처리하는 방식

미세조정(Fine-tuning) 사전학습된 모델을 특정 목적이나 스타일에 맞게 추가로 조정하는 학습 과정

RLHF(Reinforcement Learning from Human Feedback) 사람의 평가나 선호를 보상 신호로 활용해 모델을 개선하는 강화학습 방법

보코더(Vocoder) AI 음악 생성의 마지막 단계에서 모델이 생성한 추상적인 수치 데이터를 사람이 들을 수 있는 오디오 파형으로 복원해내는 오디오 합성 장치

파운데이션 모델(Foundation Model) 대규모 데이터로 학습되어 다양한 작업에 공통적으로 활용 가능한 범용 AI 모델

DCAE(Deep Compression Autoencoder) 오디오를 고도로 압축된 잠재 표현으로 변환해 계산 효율과 음질을 동시에 확보하는 신경망 기반 압축 기술

참고문헌

[1] J. Ho et al., “Denoising Diffusion Probabilistic Models,” in Proc. Int. Conf. Neural Inf. Process. Syst., (Vancouver, Canada), Dec. 2020, pp. 6840-6851.

[2] T. Mikolov et al., “Recurrent Neural Network Based Language Model,” in Proc. Interspeech, (Chiba, Japan), Sep. 2010, pp. 1045-1048.

[3] H. Sak et al., “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,” in Proc. Interspeech, (Singapore), Sep. 2014, pp. 338-342.

[4] A. Vaswani et al., “Attention Is All You Need,” in Proc. Int. Conf. Neural Inf. Process. Syst., (Long Beach, CA, USA), Dec. 2017, pp. 5998-6008.

[5] N. Zeghidour et al., “SoundStream: An End-to-End Neural Audio Codec,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, 2022, pp. 495-507.

[6] A. Defossez et al., “High Fidelity Neural Audio Compression,” Trans. Mach. Learn. Res., 2023.

[7] P. Dhariwal et al., “Jukebox: A Generative Model for Music,” arXiv preprint, 2020. doi: 10.48550/arXiv.2005.00341

[8] J. Copet et al., “Simple and Controllable Music Generation,” in Proc. Int. Conf. Neural Inf. Process. Syst., (New Orleans, LA, USA), Dec. 2023, pp. 47704-47720.

[9] https://suno.com

[10] D. Conklin, “Multiple viewpoint systems for music prediction,” J. New Music Res., vol. 24, no. 1, 1995, pp. 51-73.

[11] M. Allan et al., “Harmonising Chorales by Probabilistic Inference,” in Proc. Int. Conf. Neural Inf. Process. Syst., (Vancouver, Canada), Dec. 2004, pp. 25-32.

[12] T. Anders et al., “Constraint Programming Systems for Modeling Music Theories and Composition,” ACM Comput. Surv., vol. 43, no. 4, pp. 1-38, 2011.

[13] https://magenta.withgoogle.com/

[14] https://github.com/magenta/magenta/blob/main/magenta/models/melody_rnn/README.md

[15] https://magenta.withgoogle.com/performance-rnn

[16] C. Hawthorne et al., “Enabling Factorized Piano Music Modeling and Generation with the Music Transformer,” in Proc. Int. Conf. Learn. Represent., (New Orleans, LA, USA), May 2019.

[17] C. Payne, “MuseNet,” OpenAI Blog, 2019. https://openai.com/index/musenet/

[18] https://openai.com/index/jukebox/

[19] https://musiclm.com/

[20] https://musicgen.com/

[21] https://stableaudio.com/

[22] https://audiocraft.metademolab.com/

[23] J. de Berardinis et al., “Towards Responsible AI Music: an Investigation of Trustworthy Features for Creative Systems,” arXiv preprint, 2025. doi: 10.48550/arXiv.2503.18814

[24] Y. Gong et al., “AST: Audio Spectrogram Transformer,” in Proc. Interspeech, (Brno, Czech Republic), Aug. 2021, pp. 571-575.

[25] P. Hsu et al., “Towards Robust Neural Vocoding for Speech Generation: A Survey,” arXiv preprint, 2019. doi: 10.48550/arXiv.1912.02461

[26] J. Lee et al., “Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement,” in Proc. Conf. Empir. Methods Nat. Lang. Process., (Brussels, Belgium), Oct. 2018, pp. 1173-1182.

[27] A. Gulati et al., “Conformer: Convolution-augmented Transformer for Speech Recognition,” in Proc. Interspeech, (Shanghai, China), Oct. 2020, pp. 5036-5040.

[28] R. Mittal et al., “Symbolic Music Generation with Diffusion Models,” in Proc. Int. Soc. Music Inf. Retrieval Conf., Nov. 2021, pp. 468-475.

[29] A. Agostinelli et al., “MusicLM: Generating Music From Text,” arXiv preprint, 2023. doi: 10.48550/arXiv.2301.11325

[30] H. Liu et al., “AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, 2024, pp. 2871-2883.

[31] https://www.udio.com/

[32] Z. Wang et al., “Whole-song hierarchical generation of symbolic music using cascaded diffusion models,” in Proc. Int. Conf. Learn. Represent., (Vienna, Austria), May 2024.

[33] Y. Yi et al., “PerceiverS: A multi-scale perceiver with effective segmentation for long-term expressive symbolic music generation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 33, 2025, pp. 3975-3987.

[34] Y. Wang et al., “NotaGen: Advancing musicality in symbolic music generation with large language model training paradigms,” in Proc. Int. Joint Conf. Artif. Intell., (Montreal, Canada), Aug. 2025, pp. 10207-10215.

[35] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proc. Conf. Neural Inf. Process. Syst., (New Orleans, LA, USA), Dec. 2022.

[36] G. Cideron et al., “MusicRL: Aligning music generation to human preferences,” in Proc. Int. Conf. Mach. Learn., (Vienna, Austria), July, 2024, pp. 8968-8984.

[37] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” in Proc. ISCA Speech Synth. Workshop, (Sunnyvale, CA, USA), Sep. 2016.

[38] J. Kong et al., “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” in Proc. Int. Conf. Neural Inf. Process. Syst., (Vancouver, Canada), Dec. 2020, pp. 17022-17033.

[39] https://github.com/SongGen-AI/LLambada

[40] Y. Li et al., “MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training,” in Proc. Int. Conf. Learn. Represent., (Vienna, Austria), May 2024.

[41] https://github.com/LiuZH-19/SongGen

[42] J. Chen et al., “Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models,” in Proc. Int. Conf. Learn. Represent., (Singapore), May 2025.

[43] M. Zanon Boito et al., “mHuBERT-147: A Compact Multilingual HuBERT Model,” in Proc. Interspeech, (Kos, Greece), Sep. 2024.

[44] https://github.com/Comfy-Org/ComfyUI

[45] E.J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” in Proc. Int. Conf. Learn. Represent., Apr. 2022.

[46] L. Zhang et al., “Adding Conditional Control to Text-to-Image Diffusion Models,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., (Paris, France), Oct. 2023. pp. 3813-3824.

[47] L. Whitney, “How to Generate Your Own Music with the AI-Powered Suno,” ZDNet, 2024. 12. 26. https://www.zdnet.com/article/how-to-generate-your-own-music-with-the-ai-powered-suno/