음악 파운데이션 모델
– 텍스트-오디오 및 오디오-MIDI 양방향 합성을 위한 통합 아키텍처 –

김민재, 장순철, 박성현, 조승현
이모션웨이브(주)
e-mail:csh@emotionwave.com

Music Foundation Model
– A Unified Architecture for Music Generation, Transcription, and Synthesis

Min-Jae Kim, Soon-Chul Jang, Seong-Hyeon Park, Seung-Hyun Cho
Emotionwave Inc.

요 약

본 논문은 텍스트 기반 음악 생성과 오디오-MIDI 양방향 변환을 통합한 음악 생성·분석 시스템을 제안한다. 제안 시스템은 섹션, 비트, 틱 수준으로 구성된 계층적 xLSTM 생성기를 통해 64차원 잠재 시퀀스를 생성하고, 이를 오디오 경로의 VAE 디코더 또는 악보 경로의 FiLM 네트워크를 통해 각각 파형과 기호 표현으로 변환한다. 역방향으로는 입력 오디오를 동결된 VAE 인코더를 통해 256차원 잠재 표현으로 변환하고, 학습된 코덱을 사용하여 MIDI 채보 및 재합성을 수행한다. 단일 통합 코덱 구조를 활용함으로써 별도의 전문 모델 없이도 텍스트 기반 음악 생성, MIDI 합성, 오디오 분석, 기호 편집 등 다양한 음악 제작 워크플로우를 지원할 수 있다.

1. 서론

음악 생성과 음악 분석은 오랫동안 별개의 연구 영역으로 발전해 왔다. 텍스트 조건 음악 생성 모델은 파형 공간에서 우수한 성능을 보이지만, 음악의 기호적 구조를 직접적으로 다루지는 않는다. 반면 채보 및 소스 분리 시스템은 오디오 신호를 분석할 수 있으나, 음악 형식이나 화성 구조와 같은 상위 수준의 음악 정보를 충분히 반영하기 어렵다. 실제 음악 제작 워크플로우에서는 이러한 도구들을 개별적으로 연결하고, 결과물을 수작업으로 통합해야 하는 번거로움이 존재한다.

본 연구의 핵심 관찰은 오디오와 악보가 동일한 음악 내용을 서로 다른 방식으로 표현한 것이라는 점이다. 두 표현을 모두 포괄할 수 있는 충분히 풍부한 잠재 공간을 구성할 수 있다면, 음악 생성과 분석을 하나의 통합된 시스템 안에서 처리할 수 있다. 최근 음악 VAE와 신경망 코덱 연구는 압축된 잠재 공간이 음향 정보를 효과적으로 보존하면서도 해석 가능한 구조를 유지할 수 있음을 보여주었다.

본 논문은 이러한 관점에서 텍스트 생성, 오디오 인코딩, MIDI 합성, 악보 생성을 통합적으로 수행할 수 있는 시스템 구조를 제안한다. 제안 시스템은 계층적 xLSTM을 사용하여 텍스트 입력을 64차원 잠재 시퀀스로 변환하고, 이를 오디오 경로의 VAE 디코더 또는 악보 경로의 FiLM 네트워크를 통해 각각 파형과 기호 표현으로 생성한다. 역방향에서는 오디오를 동결된 VAE 인코더로 256차원 잠재 표현으로 변환하고, 학습된 코덱을 통해 양방향 악보 변환을 지원한다.

2. 시스템 아키텍처

2.1 개요

제안 시스템은 생성 경로와 분석 경로로 구성된 두 가지 주요 파이프라인을 포함한다.

생성 경로에서는 동결된 텍스트 인코더인 Flan-T5가 입력 텍스트를 임베딩으로 변환한다. 이후 세 개의 자기회귀 xLSTM 모듈 [1]이 계층적으로 작동한다. 섹션 수준에서는 곡의 전체 구조, 예를 들어 인트로, 벌스, 코러스, 아웃트로를 결정한다. 비트 수준에서는 각 구절의 음악적 내용을 생성하며, 틱 수준에서는 비트당 12틱의 프레임 해상도로 세부적인 시간 정보를 생성한다. 이 과정을 통해 생성된 64차원 잠재 시퀀스는 두 개의 독립적인 디코딩 경로로 분기된다. 오디오 경로는 동결된 RAVE 스타일 VAE 디코더를 통해 파형을 생성하고, MIDI 경로는 FiLM 조건부 네트워크를 통해 피아노롤을 생성한다.

분석 경로에서는 입력 오디오가 동결된 VAE 인코더를 통해 256차원 잠재 표현으로 변환된다. 학습된 MidiAudioCodec은 이 256차원 표현을 두 가지 방식으로 처리한다. 필요한 경우 64차원 생성 공간으로 투영하거나, FiLM 조건부 네트워크를 통해 직접 악보 표현으로 변환한다. 이를 통해 사용자는 기존 오디오를 분석하고, 필요에 따라 기호적으로 편집한 뒤 다시 오디오로 합성할 수 있다.

2.2 계층적 xLSTM 생성

생성의 핵심은 세 단계로 구성된 xLSTM 계층이다 [1]. 각 단계는 독립적인 자기회귀 모듈로 구현되며, 이전 단계의 출력을 조건으로 사용한다. xLSTM은 행렬 기반 LSTM(mLSTM)과 스칼라 기반 LSTM(sLSTM)을 결합하여 긴 시퀀스에서도 메모리 효율적으로 작동한다.

SectionXLSTM은 곡의 전체 구조를 결정한다. 이 모듈은 12개의 mLSTM 계층(d = 512, 8 attention heads)으로 구성되며, Flan-T5-Large에서 추출한 1024차원 텍스트 임베딩을 입력으로 사용한다. 먼저 SectionCountPredictor가 전체 섹션 수, 일반적으로 2-4개를 예측하고, 이어서 각 섹션의 특성을 연속 분포인 혼합 가우시안 형태로 생성한다. 이렇게 생성된 섹션 잠재 표현은 다음 단계의 조건으로 사용된다.

BeatXLSTM은 각 섹션 내부의 음악적 내용을 구절 수준에서 전개한다. 이 모듈 역시 12개의 mLSTM 계층(d = 512, 8 attention heads)으로 구성되며, 섹션 잠재 표현과 텍스트 임베딩을 동시에 조건으로 사용한다. 결과적으로 각 섹션은 4-32개의 비트로 확장되고, 각 비트는 개별 음악 이벤트를 표현한다.

TickXLSTM은 가장 세밀한 시간 해상도에서 작동한다. 이 모듈은 비트당 12개의 틱을 생성하므로 약 16배 압축된 시간 표현 안에서 음표의 정확한 온셋과 지속 시간을 표현할 수 있다. TickXLSTM 또한 12개의 mLSTM 계층을 사용하며, 비트 잠재 표현과 텍스트 정보를 조건으로 한다.

모든 단계는 SwiGLU 활성화 함수 [3]와 RMSNorm 정규화 [4]를 적용하여 수치적 안정성을 높인다. 또한 드럼, 피아노, 베이스 등 악기 카테고리를 임베딩하여 각 단계의 입력에 주입함으로써, 모델이 악기별 특성을 인식할 수 있도록 한다. 추론 시 계산 복잡도는 단계당 O(1)이다. 이는 순환 상태 업데이트만으로 이전 정보를 유지하기 때문이며, 트랜스포머의 KV 캐시 기반 O(T) 접근 방식과 대비된다.

이러한 계층적 설계는 실제 음악가의 작곡 과정을 반영한다. 먼저 전체 형식을 구상하고, 다음으로 구절 수준의 내용을 구체화하며, 마지막으로 음표 수준의 세부 사항을 정교화하는 방식이다. 실제로 이러한 계층 구조는 여러 시간 척도에 걸쳐 음악적 일관성을 유지하는 데 유리하다.

2.3 MidiAudioCodec

MidiAudioCodec은 생성된 64차원 표현과 오디오/MIDI 표현 사이의 변환을 담당한다. 이 모듈은 다음과 같은 기본 연산을 지원한다.

MIDI에서 오디오로 변환하는 경로에서는 피아노롤, 즉 128개 음정에 대한 속도 값이 트랜스포머 인코더에 입력된다. 이 인코더는 전체 피아노롤을 512차원의 컨텍스트 벡터로 요약한다. 이후 FiLM(Feature-wise Linear Modulation) [5] 조건부 디코더가 이 컨텍스트를 사용하여 합성곱 네트워크의 특징 맵을 변조하고, 그 결과 64차원 잠재 표현을 생성한다. 마지막으로 동결된 VAE 디코더가 이 64차원 벡터를 오디오 파형으로 변환한다.

오디오에서 MIDI로 변환하는 역방향 경로도 지원된다. 입력 오디오는 먼저 동결된 VAE 인코더를 통해 256차원 표현으로 변환되고, FiLM 조건부 디코더는 이 256차원 표현을 피아노롤 로짓으로 디코딩한다.

텍스트 기반 생성 경로에서는 xLSTM이 이미 64차원 잠재 표현을 생성하므로, 이를 직접 VAE 디코더에 투영하여 오디오를 생성한다.

FiLM 조건부 구조를 선택한 이유는 실용적인 효율성에 있다. FiLM 기반 조건화는 크로스 어텐션 기반 조건화보다 메모리 효율적이며, 30분 이상의 긴 시퀀스를 4096프레임 윈도우 단위로 청킹하여 처리할 때도 안정적으로 작동한다. 또한 악기 카테고리(0-18: 드럼, 피아노, 베이스, 기타, 현악기, 금관악기, 목관악기, 신시사이저 등)와 타악기 여부를 나타내는 플래그를 임베딩하여 각 단계에 주입함으로써, 모델이 음색 정보를 명시적으로 인식할 수 있도록 한다.

2.4 생성 잠재 공간

xLSTM 계층의 출력인 64차원 잠재 시퀀스는 전체 시스템의 중심 표현이다. 이 차원 수는 경험적으로 결정되었다. 오디오 VAE의 256차원 표현보다 작기 때문에 생성 과정의 계산 효율성을 확보할 수 있으며, 동시에 음표의 온셋, 음량, 음색과 같은 주요 음악적 특징을 충분히 보존할 수 있다.

생성된 64차원 시퀀스는 오디오와 MIDI로 각각 변환된다. 첫째, 오디오 경로에서 MidiAudioCodec은 64차원 표현을 256차원 VAE 공간으로 투영하고, 동결된 VAE 디코더가 이를 파형으로 변환한다. 둘째, MIDI 경로에서는 동일한 64차원 표현이 FiLM 조건부 디코더를 통해 음정과 속도 정보를 포함한 피아노롤로 변환된다.

중요한 점은 두 경로가 동일한 64차원 소스에서 분기된다는 것이다. 따라서 생성된 오디오와 MIDI는 구조적으로 일관성을 가지며, 사용자는 오디오를 들으면서 이에 대응하는 악보를 동시에 확인할 수 있다.

2.5 추론

제안 시스템은 여러 실무적 음악 제작 워크플로우를 자연스럽게 지원한다.

텍스트 기반 음악 생성은 가장 기본적인 사용 방식이다. 사용자가 텍스트 프롬프트를 입력하면 세 단계의 xLSTM이 순차적으로 작동하여 64차원 잠재 표현을 생성한다. 이 잠재 표현은 동시에 오디오 경로의 VAE 디코더와 악보 경로의 FiLM 디코더로 전달되므로, 사용자는 생성된 음악과 그에 대응하는 기호 표현을 함께 얻을 수 있다.

기존 악보로부터 오디오를 생성하는 경로도 지원된다. 피아노롤 또는 MIDI 파일을 입력하면 트랜스포머 인코더가 이를 요약하고, FiLM 조건부 디코더가 64차원 잠재 표현으로 변환한 뒤, VAE 디코더가 오디오를 생성한다. 이를 통해 음악가는 악보를 먼저 작성한 후 이를 음원화할 수 있다.

오디오 분석 및 악보 추출은 역방향 워크플로우에 해당한다. 기존 오디오 파일을 VAE 인코더로 256차원 잠재 표현으로 변환한 후, FiLM 조건부 디코더가 이를 악보로 변환한다. 이는 오디오 채보(transcription) 작업을 수행하는 과정이다.

오디오 분석과 재합성도 가능하다. 오디오를 분석하여 악보로 변환한 뒤, 사용자가 음정 변경이나 박자 조정과 같은 기호적 편집을 수행하고, 편집된 악보를 다시 오디오로 변환할 수 있다. 이 과정에서는 별도의 채보 모델과 합성 모델이 필요하지 않으므로 워크플로우가 간결해진다.

잠재 공간 탐색 역시 지원된다. 생성된 64차원 잠재 시퀀스의 값을 조작함으로써 음악의 형식, 구절 구조, 미세한 타이밍 등을 체계적으로 변형할 수 있다. 이는 음악 편집을 위한 새로운 인터페이스로 활용될 수 있다.

긴 시퀀스는 별도의 처리 전략을 적용한다. 3분 이상의 음악, 즉 약 32K 프레임 이상의 시퀀스에 대해 xLSTM 계층은 순환 상태를 유지하면서 선형 시간 O(T)에 전체 시퀀스를 처리한다. 코덱 계층은 4096프레임 단위의 윈도우로 시퀀스를 나누어 처리함으로써 메모리 사용량을 선형적으로 유지한다.

3. 설계 원칙 및 제한 사항

3.1 설계 원칙

본 시스템의 핵심 설계는 두 가지 관찰에서 출발한다.

첫째, 오디오와 악보는 동일한 음악을 서로 다른 방식으로 표현한 것이다. 따라서 두 표현을 연결하는 통합 인터페이스를 구축하면, 사용자는 어떤 형식의 입력으로도 음악을 다룰 수 있다. 본 연구에서는 이를 양방향 코덱으로 구현하였다. 오디오 측에는 동결된 VAE 기반 256차원 표현이 있고, 생성 측에는 학습된 계층 기반 64차원 표현이 있으며, MidiAudioCodec은 이 두 표현을 매개한다.

둘째, 채보 모델, 합성 모델, 생성 모델과 같이 전문화된 모델을 각각 구축하기보다는 공통 표현을 중심으로 통합 구조를 설계하는 것이 더 간결하고 효율적이다. 이러한 설계를 통해 다음과 같은 장점을 얻을 수 있다.

먼저 생성 경로의 일관성이 확보된다. xLSTM이 생성하는 64차원 잠재 표현은 오디오 경로의 VAE와 MIDI 경로의 FiLM이라는 두 개의 독립적인 디코더로 분기된다. 두 경로가 동일한 소스에서 출발하므로 생성된 음악과 악보는 자동으로 정합성을 가진다. 별도의 모델을 사용했다면 이러한 일관성을 보장하기 위해 복잡한 후처리가 필요했을 것이다.

다음으로 분석 경로의 유연성이 확보된다. 입력 오디오는 VAE를 통해 256차원 표현으로 인코딩된 후 여러 방향으로 처리될 수 있다. FiLM 디코더를 통해 악보로 변환하거나, VAE 디코더를 통해 재합성하거나, 64차원 표현으로 투영하여 생성 경로와 연결할 수도 있다. 단일 표현 공간을 중심으로 설계했기 때문에 이러한 유연성이 자연스럽게 나타난다.

마지막으로 시스템의 경제성이 향상된다. 텍스트–음악, MIDI–음악, 음악–MIDI 변환마다 별도의 모델을 학습하는 대신, 학습된 코덱 계층이 모든 변환을 처리한다. 이는 계산 비용을 낮추고 시스템 유지보수를 단순화하며, 일관된 음악 표현을 보장하는 데에도 기여한다.

3.2 제한 사항과 과제

현재 시스템은 몇 가지 제약을 가진다.

첫째, 정량적 평가가 아직 충분하지 않다. Fréchet Audio Distance나 음표 수준 채보 정확도(F1)와 같은 표준 지표를 사용하여 기존 전문 모델들과 체계적으로 비교할 필요가 있다.

둘째, 모델은 생성과 기본적인 채보 작업에 최적화되어 있으나 고급 신호 처리 기능은 아직 다루지 않는다. 예를 들어 드럼과 베이스를 개별적으로 분리하는 소스 분리(source separation)는 현재 지원하지 않는다.

셋째, xLSTM 계층은 단성 음악 또는 가벼운 다성 음악을 주된 대상으로 가정한다. 많은 악기가 동시에 연주되는 복잡한 편성, 예를 들어 대규모 오케스트라 음악은 학습 데이터의 분포를 벗어날 수 있다.

넷째, 5분 이상의 매우 긴 작곡에 대해서는 아직 광범위한 테스트를 수행하지 않았다. 청크 단위 처리는 로컬 수준의 음악적 일관성을 유지하는 데 도움이 되지만, 곡 전체에 걸친 구조적 일관성, 예를 들어 형식의 전개나 클라이맥스의 위치는 약화될 수 있다.

4. 향후 방향

본 시스템은 다음과 같은 방향으로 확장될 수 있다.

첫째, 정량적 벤치마킹이 필요하다. 기존 전문 모델들과 음향 품질, 채보 정확도, 생성 결과의 음악적 일관성 등을 체계적으로 비교해야 한다.

둘째, 악기별 제어의 세밀도를 높일 수 있다. 현재는 악기 카테고리 수준의 조건만 사용하지만, 각 악기에 독립적인 잠재 코드를 부여하면 음악가가 악기별 생성 결과를 더 정밀하게 조정할 수 있을 것이다.

셋째, 사용자 연구가 필요하다. 실제 음악가들과의 협업을 통해 시스템의 실용성을 평가하고, 어떤 워크플로우가 가장 유용한지 분석해야 한다.

넷째, 잠재 공간의 구조를 더 깊이 이해할 필요가 있다. 현재 64차원 잠재 표현은 하나의 통합 벡터로 사용되지만, 이 중 어떤 차원이 음색, 멜로디, 리듬 등 특정 음악 요소와 관련되는지를 분리할 수 있다면(disentanglement), 사용자 제어는 더욱 직관적으로 이루어질 수 있다.

마지막으로 실제 제작 도구와의 통합이 중요하다. 제안 시스템을 DAW(Digital Audio Workstation) 플러그인으로 구현하거나 기존 음악 제작 워크플로우에 자연스럽게 통합할 수 있다면, 음악가들의 실질적인 활용 가능성이 높아질 것이다.

참고문헌

[1] B. Beck, B. Chronopoulou, D. Cichocki, H. Dey, D. Goldsborough, G. Gonçalves, “xLSTM: Extended Long Short-Term Memory,” arXiv preprint arXiv:2405.04517, May 2024.

[2] A. Caillon, P. Esling, “RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis,” arXiv preprint arXiv:2111.05011, 2021.

[3] N. Shazeer, “GLU Variants Improve Transformer,” arXiv preprint arXiv:2002.05202, 2020.

[4] B. Zhang, S. Sennrich, “Root Mean Square Layer Normalization,” Advances in Neural Information Processing Systems, 2019.

[5] E. Pérez, F. Strub, H. de Vries, V. Dumoulin, A. Courville, “FiLM: Visual Reasoning with a General Conditioning Layer,” Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6594-6602, 2018.