논문 2023-18-27

Music Transformer 기반 음악 정보의 가중치 변형을 통한 멜로디 생성 모델 구현

(Implementation of Melody Generation Model Through Weight Adaptation of Music Information Based on Music Transformer)

조 승 아, 이 재 호*
(Seunga Cho, Jaeho Lee)

Abstract : In this paper, we propose a new model for the conditional generation of music, considering key and rhythm, fundamental elements of music. MIDI sheet music is converted into a WAV format, which is then transformed into a Mel Spectrogram using the Short-Time Fourier Transform (STFT). Using this information, key and rhythm details are classified by passing through two Convolutional Neural Networks (CNNs), and this information is again fed into the Music Transformer. The key and rhythm details are combined by differentially multiplying the weights and the embedding vectors of the MIDI events. Several experiments are conducted, including a process for determining the optimal weights. This research represents a new effort to integrate essential elements into music generation and explains the detailed structure and operating principles of the model, verifying its effects and potentials through experiments. In this study, the accuracy for rhythm classification reached 94.7%, the accuracy for key classification reached 92.1%, and the Negative Likelihood based on the weights of the embedding vector resulted in 3.01.

Keywords : Feature Extraction, Muisc Transformer, Short Time Fourier Transform, Convolutional Neural Network, Generation Model

Ⅰ. 서 론

음악은 그 자체로 인간의 감성과 감정을 자극하는 독특한 예술 형태이다. 그러나 이런 감성과 감정을 만들어내는 데에는 복잡한 음악 이론과 구조적인 요소가 작용한다. 이러한 요소들 중 키 [1]와 박자 [2]는 특히 음악의 기본적인 구조와 분위기를 결정하는데 중요한 역할을 한다.

키는 음악의 조화와 음악적 색채를 결정하는데 중요한 역할을 한다. 특정 키에서 연주되는 음악은 동일한 멜로디라도 다른 키에서 연주될 때와는 다른 감정과 분위기를 전달할 수 있다. 또한, 키 변화는 음악의 긴장감을 높이고, 리스너의 기대감을 조절하는데 사용될 수 있다. 따라서, 키는 음악이 전달하려는 메시지와 감정을 조절하는 중요한 도구이다.

박자는 음악의 리듬과 속도를 결정하는 역할을 한다. 박자는 음악의 템포를 설정하고, 음악의 부드러움이나 격렬함을 결정하는 데 중요한 역할을 한다. 음악의 박자는 리스너의 심장 박동과 호흡 패턴에 영향을 미치며, 이는 리스너의 심리적 상태와 감정 반응에 큰 영향을 미친다. 따라서, 박자는 음악이 리스너에게 미치는 심리적 영향을 조절하는 데 중요한 역할을 한다.

*Corresponding Author (izeho@duksung.ac.kr)
Received: Jun. 28, 2023, Revised: Aug. 1, 2023, Accepted: Sep. 9, 2023.
Seunga Cho: Duksung Women's University (B.S. Student)
Jaeho Lee: Duksung Women's University (Assist. Prof.)

ⓒ IEMEK J. Embed. Sys. Appl. 2023 Oct. 18(5) 217-223
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2023.18.5.217

기존 음악 생성 모델 [3]들은 LSTM (Long Short-Term Memory) [4], GAN (Generative Adversarial Network) [5], Transformer [6]와 같은 모델을 활용한 연구들이 진행되었지만, 그 대부분은 고수준의 특성에 초점을 맞추고, 주요 음악 이론을 고려하지 않았다. 이는 사용자의 음악적 선호나 의도를 제대로 반영하지 못하는 기계적이고 무감각한 음악을 만들어내는 결과를 초래하였다. 이러한 단점을 보완하기 위해서 본 연구에서는 새로운 접근법을 제안한다.

본 연구에서는 MIDI (Musical Instrument Digital Interface) [7] 형태의 악보를 WAV (Wavefrom Audio File Format) 형태로 변환한 후, STFT (Short-Time Fourier Transform) [8]를 이용하여 Mel Spectrogram으로 변환한다. 이렇게 변환된 Mel Spectrogram은 두 개의 CNN (Convolutional Neural Network) [9]를 통과하여 키와 박자 정보를 추출하며, 이 정보는 이후 Music Transformer [10]의 입력으로 사용된다. 이 과정에서 키와 박자 정보는 각 MIDI 이벤트의 임베딩 벡터와 가중치를 다르게 곱하여 합쳐진다. 이때 가중치는 표 1과 같이 임의의 4개의 가중치를 설정한 후 NLL (Negative Log Likelihood)을 비교하여 최적의 가중치를 찾아내었다. 이 과정을 통해 모델은 사용자의 의도를 더욱 정확하게 반영하며, 더욱 의미 있고 감성적인 음악을 만들어낼 수 있다.

본 연구의 동기는 음악 생성 모델이 더욱 진보하고 다양한 음악을 생성할 수 있도록 돕는 것이다. 이는 딥러닝과

인공지능이 음악 생성에 미치는 영향에 대한 새로운 이해를 제공하며, 이는 음악의 세계를 넓히는 데 중요한 역할을 한다. 본 연구에서 제안하는 모델은 이러한 목표를 향한 중요한 한 걸음이며, 이를 통해 음악 생성 모델의 가능성을 더욱 확장하고자 한다.

더욱이, 본 연구의 중요성은 개인화된 음악 생성 시스템에 있어서 두드러진다. 이러한 시스템은 사용자의 개별적인 선호도에 따라 특정 키, 박자, 그리고 기타 음악적 요소를 선택할 수 있는 기능을 제공한다. 이렇게 제공된 정보를 토대로, 시스템은 사용자 맞춤형의 음악을 생성한다. 본 연구에서 제안하는 조건부 음악 생성 모델은 이러한 개인화된 음악 생성 시스템에 적합하다. 사용자가 직접 원하는 키와 박자를 지정할 수 있게 함으로써, 사용자는 자신의 음악 취향을 보다 세밀하게 표현할 수 있으며, 이를 바탕으로 정확한 스타일의 음악을 생성할 수 있다.

이러한 연구 배경과 동기를 바탕으로, 본 연구는 조건부 음악 생성 모델을 향한 새로운 시도이다. MIDI에서 WAV로, 그리고 Mel spectrogram [11]으로의 변환 과정, 그리고 이를 바탕으로 한 키와 박자 분류, 이 모든 과정을 통해 본 연구는 음악 생성의 본질적 요소를 명확히 인지하고, 이를 음악 생성 과정에 명시적으로 통합하려는 시도를 보여준다.

본 논문은 2023년 한국 통신학회 "박자와 키 조건을 적용한 Conditional Music Transformer 모델에 관한 연구"를 선행 연구로 하여 음악 생성에 대한 방법론을 세부적으로 설명하고, 이에 대한 실험적 검증을 제공한다.

Ⅱ. 연구배경

조건부 음악 생성 분야는 다양한 연구로부터 성장해 왔다. 본 논문에서는 그 중 주요한 방법론들에 대해 검토한다.

초기에는 LSTM을 기반으로 한 음악 모델이 주로 활용되었다 [7]. 이런 접근법은 LSTM의 시퀀스 데이터 처리 능력을 활용하여 음악 패턴을 학습하고 생성하는데 효과적이었다. 그러나 복잡한 음악 구조나 장기적인 의존성을 완전히 처리하는 데는 한계가 있었다.

그 다음으로는 GAN을 활용한 음악 생성 연구가 주목받았다 [12]. 이 연구는 실제 음악 데이터를 기반으로 고품질의 새로운 음악을 생성하는 GAN의 능력을 보여주었다. 그러나 해당 연구는 음악 생성보다 음악 합성에 더욱 초점을 맞추고 있으며, MIDI 데이터를 활용하여 더욱 다양한 음악 데이터에 대한 접근성을 제공한 본 모델과 달리 Mel Spectrogram 데이터를 기반으로만 작동한다. 또한 단순한 CNN 구조를 활용하여 빠르게 학습 가능한 본 모델과 달리 복잡한 GAN 구조를 사용하는 한계를 가지고 있다.

이후에는 Transformer 모델을 활용한 음악 생성 연구가 등장하였다 [12]. 이 연구는 Transformer의 장기 의존성 처리 능력과 복잡한 음악 구조 학습 능력을 강조하였다. 그러나 Transformer 모델의 크기와 과적합 문제는 이 연구를 제한하였다.

다음으로, VAE (Variational AutoEncoder) [13]를 활용한 음악 생성 연구가 제안되었다 [14]. 이 연구는 잠재 공간의 표현력을 활용하여 음악을 생성하는 방법론을 제시하였다. 그러나 이 방법론은 음악의 구조적인 특성을 충분히 반영하지 못하는 한계를 가지고 있었다.

이후, Reinforcement Learning을 활용한 음악 생성 연구가 제안되었다 [15]. 이 연구는 보상 함수를 기반으로 음악 생성을 탐색하는 방법론을 제안하였다. 그러나 학습 과정이 복잡하고, 보상 함수 설계의 어려움으로 인해 이 방법의 실제 적용이 어렵다는 한계가 있었다.

또한, 신경망 자기 회귀 모델을 사용한 음악 생성 연구도 있었다 [16]. 이 방법론은 음표 간의 조건부 확률을 모델링하여 음악을 생성하였다. 그러나 이 방법은 복잡한 음악 구조를 학습하고 생성하는 데 어려움을 겪었다.

마지막으로, 어텐션 메커니즘 [17]을 사용한 음악 생성 연구가 제안되었다 [18]. 본 연구는 음악 생성 과정에서 중요한 정보에 주목하여 음악 구조를 학습하려고 시도하였다. 그러나 이 방법은 주로 단기적인 패턴에 집중하며, 장기적인 음악 구조를 완전히 반영하지 못했다.

이전 연구들이 각기 다른 방법론을 사용하여 음악 생성을 탐구했지만, 공통적으로 몇 가지 핵심적인 문제를 해결하지 못했다. 첫째, 대부분의 연구는 단기적인 음악 패턴만을 고려하며, 장기적인 음악 구조나 개별 음표 간의 복잡한 상호작용을 제대로 반영하지 못했다. 둘째, 기존 방법론들은 학습 과정이 복잡하고, 모델의 훈련에 많은 시간과 자원을 요구했다. 이는 모델의 실용성을 저하시키는 주요한 요인이었다.

본 연구에서 제안하는 새로운 모델은 이런 문제들을 극복하기 위해 설계되었다. 본 모델은 음악의 단기적 패턴뿐 아니라 장기적 구조도 학습할 수 있는 새로운 아키텍처를 가지고 있다. 또한, 본 모델은 효율적인 학습 과정을 제공하며, Attention구조를 활용한 Transformer를 기반으로하는 Music Transformer를 활용 하기 때문에 LSTM과 RNN에 비해 빠른 훈련 시간을 보장한다. 이런 특징은 음악 생성의 품질을 향상시키는 한편, 실제 음악 산업에 적용할 때의 실용성도 높인다. 이를 통해 본 연구는 조건부 음악 생성 분야의 발전에 기여하고자 한다.

Ⅲ. 모델 설계

1. 기술적 배경

1.1 STFT

STFT는 신호 분석의 중요한 기법 중 하나로, 시간에 따른 신호의 주파수 변화를 분석하는데 이용된다. 일반적인FT (Fourier Transform)은 시간 영역의 신호를 주파수 영역으로 변환하여 신호의 주파수 성분을 분석하는 데에 탁월하나, 그 과정에서 시간에 따른 주파수 변화에 대한 정보가 손실되는 단점이 있다.

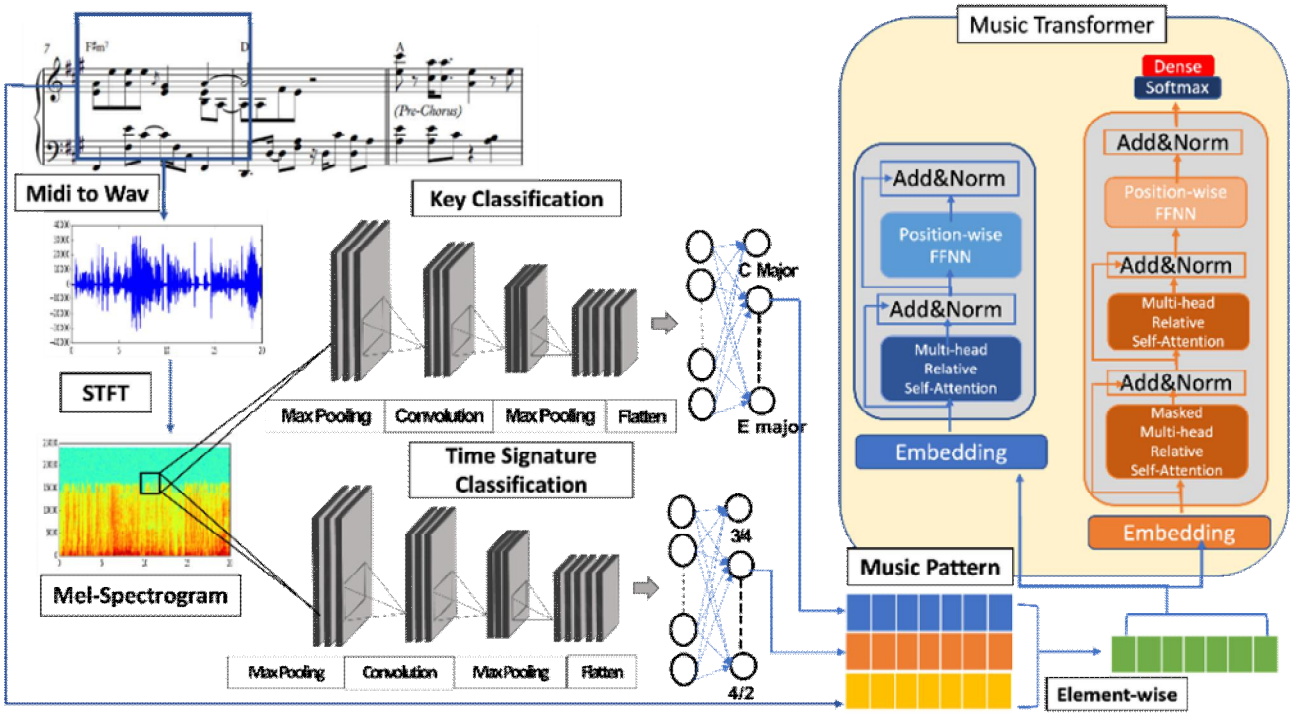


그림 1. 음악 특성 추출 및 생성 모델 시스템 구조
Fig. 1. Music Classification and Generation Model System Architecture

STFT는 이러한 문제를 해결하기 위해 신호를 작은 시간 구간으로 나눈 후, 각각의 구간에 대해 FT를 적용한다. 이렇게 함으로써 시간에 따른 주파수 변화에 대한 정보를 유지하면서 신호를 주파수 영역으로 분석할 수 있게 된다. 이러한 특성 때문에 STFT는 동적인 신호 분석에 주로 사용되며, 특히 음악과 같이 시간에 따라 주파수 성분이 변하는 신호의 분석에 적합하다.

음악 데이터 분석에 STFT를 적용하면, 음악의 시간에 따른 주파수 변화를 세밀하게 파악할 수 있다. 이 정보는 음악의 다양한 특성을 이해하고, 이를 기반으로 음악을 생성하는 데에 중요한 역할을 한다. 이 연구에서는 MIDI 데이터를 WAV 형식으로 변환한 후, STFT를 이용하여 Mel Spectrogram으로 변환하는 과정을 통해, 더욱 풍부한 음악적 특성을 추출하고자 한다. 이 과정을 통해 얻어진 음악 특성 정보는 이후의 음악 생성 과정에 활용되며, 이를 통해 보다 세밀하고 다양한 음악 생성이 가능해질 것으로 기대된다.

1.2 Music Transformer

Music Transformer는 Transformer 아키텍처를 기반으로 한 자기회귀 (auto-regressive) 음악 생성 모델이다. Transformer 아키텍처는 Attention is All You Need [6]에서 제안되었으며, 그 이후 다양한 자연어 처리 문제에서 뛰어난 성능을 보여주었다. Transformer의 핵심 구성 요소는 '어텐션 메커니즘 (attention mechanism)'으로, 이를 통해 모델은 입력 데이터의 다양한 부분에 가중치를 부여하며 중요한 정보를 추출할 수 있다.

Music Transformer는 이러한 Transformer 아키텍처를 음악 생성 문제에 적용한 것이다. Music Transformer는 '이벤트' 기반의 표현 방식을 사용하며, 각 이벤트는 음악의 각각의 요소 (예를 들어, 음표의 시작, 음표의 끝, 피치, 벨로시티 등)를 나타낸다. 이렇게 표현된 이벤트들은 순차적으로 배열되어 Music Transformer의 입력으로 사용된다.

Transformer의 '셀프-어텐션 (self-attention)' 메커니즘을 통해 Music Transformer는 각 이벤트가 그 이전의 모든 이벤트와 어떻게 관련되어 있는지를 학습한다. 이는 각 이벤트의 컨텍스트 (context)를 파악하는 데 중요한 역할을 하는데, 이는 음악이 시간에 따른 패턴과 구조를 가지고 있으며, 각 음표나 이벤트가 그 이전의 음표나 이벤트에 의존적이라는 음악의 특성 때문이다.

그러나 Music Transformer는 음악의 키나 박자와 같은 정보를 명시적으로 처리하지 않는다는 한계가 있다. 이러한 정보는 음악의 기본적인 특성을 결정하는 중요한 요소이며, 이를 고려하지 않으면 음악의 감정적 톤이나 리듬을 제어하는 데 어려움이 생긴다. 따라서 본 연구에서는 Music Transformer에 키와 박자 정보를 명시적으로 인코딩하여, 이를 모델의 입력으로 제공하는 방식을 제안한다.

2. 모델 시스템 개요

본 연구에서 제안하는 모델은 기존의 음악 생성 모델에 음악의 키와 박자라는 핵심 요소를 직접 고려하여 생성하는 새로운 접근 방식을 적용한 것이다. 이 모델의 설계는 크게 두 단계로 나눌 수 있다. 음악 특성 추출 단계와 음악 생성 단계이다. 그림 1은 음악 특성 추출 및 생성 모델에 관한

시스템 구성도이다.

음악 특성 추출 단계에서는 MIDI 형태의 악보를 WAV 형식으로 변환하며, 이를 STFT를 이용해 Mel spectrogram으로 변환한다. 이후, Mel spectrogram은 두 개의 CNN을 통과시켜서 박자와 키 분류를 진행한다. 이러한 과정은 실제로 듣는 음악의 본질적인 특성인 키와 박자를 효과적으로 추출하고, 이를 사용하여 음악의 세부적인 요소를 학습하는 데에 중요하다.

음악 생성 단계에서는 첫 번째 단계에서 추출된 Time Signature와 키 정보를 MIDI 이벤트에 대한 임베딩 벡터와 조절 가능한 가중치를 곱한 후, 이를 Music Transformer에 입력으로 제공한다. 이 가중치는 연구를 통해 찾아낸 적절한 값으로, 이를 통해 키와 박자 정보와 MIDI 이벤트 정보의 중요성을 균형 있게 반영할 수 있다. 이런 과정을 거친 정보는 Music Transformer에 입력되어, 새로운 음악 시퀀스를 생성한다.

따라서 본 모델은 음악의 키와 박자라는 중요한 요소를 포함하여, 음악의 본질적인 특성을 바탕으로 새로운 음악을 생성하는 방법을 제안한다. 이를 통해 본 모델은 기존의 음악 생성 모델이 가지고 있던 장르와 같은 고수준의 특징만을 고수한다는 제한 사항을 극복하고, 보다 다양한 스타일을 지닌 음악 생성이 가능하게 된다.

3. 데이터 구성 및 전처리

본 연구에서는 다양한 장르와 스타일을 포함하는 MIDI 악보 데이터를 사용하였다. 실험에 사용된 MIDI 데이터[19]는 AI-HUB를 통해 구축이 되었으며, 300개의 데이터를 가지고 실험이 진행되었다. MIDI는 디지털 형태로 음악을 표현하는 데이터 포맷으로, 각각의 음표, 음표가 연주되는 시점, 음표의 지속 시간 등 음악에 대한 많은 정보를 담고 있다. 하지만 MIDI는 기본적으로 이산적인 정보를 담고 있어, 연속적인 시간-주파수 영역의 정보를 제한적으로만 포착할 수 있다는 한계가 있다.

이러한 한계를 극복하기 위해 MIDI 데이터를 WAV 형식의 오디오 데이터로 변환하는 과정을 거친다. 이는 연속적인 시간-주파수 영역의 정보를 MIDI 데이터에 더욱 풍부하게 반영할 수 있게 한다. 이어서, 변환된 WAV 데이터를 STFT를 이용해 Mel spectrogram으로 변환한다. Mel spectrogram은 주파수 스펙트럼을 사람의 청각 체계에 더욱 가깝게 만든 것으로, 각 주파수 구간의 에너지를 시간에 따라 표현하는 2차원 이미지로 볼 수 있다. 그림 2는 데이터셋에 관한 전처리 과정이다.

본 연구에서는 키와 박자라는 음악의 본질적 요소를 고려한 조건부 음악 생성을 시도하였다. 키에 대해서는 music21 라이브러리를 이용하여 키를 추출한 후, 정수 ID로 레이블링 후 임베딩 레이어를 통해 임베딩 벡터로 변환을 하였다. 박자의 경우 pretty_midi 라이브러리를 활용하여 박자를 추출한 후 키와 마찬가지로 정수 ID로 레이블링 후 임베딩 레이어를 통과시켜 임베딩 형태로 변환하였다. 키와 박자에

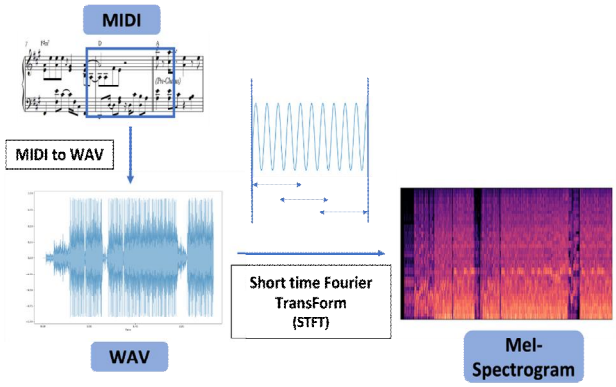


그림 2. 데이터셋 전처리 과정
Fig. 2. Data preprocessing process

대한 임베딩 벡터는 Music Transformer의 입력 임베딩 벡터의 크기를 고려하여 256차원으로 설정하였으며, 이 정보는 MIDI 이벤트의 임베딩 벡터와 결합되어, Music Transformer의 입력으로 제공된다.

키와 박자의 임베딩 벡터는 각각 별도의 가중치를 통해 MIDI 이벤트의 임베딩 벡터와 Element wise하게 결합되는데, 이 가중치는 학습 과정에서 최적화된다. 원곡의 비중을 높게 유지해야한다고 생각되어 표 1과 같이 임의로 가중치 4개를 설정하여 실험을 진행한 후 가장 NLL이 낮은 MIDI:키:박자 = 6:1:3의 가중치를 채택하였다. 이렇게 결합된 벡터는 Music Transformer에 입력되어 음악 생성 과정을 시작한다. 이렇게 하면 각 MIDI 이벤트가 특정 키와 박자에 따라 조건부로 생성되며, 따라서 생성된 음악은 키와 박자 정보를 보다 정확하게 반영할 수 있다.

따라서, 본 모델은 MIDI 데이터를 WAV 형식의 오디오 데이터로 변환하고, 이를 Mel spectrogram으로 전처리하는 방식을 통해 음악 생성에 필요한 다양한 특성을 학습한다. 그리고 이렇게 학습된 모델은 키와 박자라는 음악의 본질적 요소를 고려하여 음악을 생성하며, 이를 위해 키와 박자 정보를 임베딩 형태로 변환하여 Music Transformer의 입력으로 제공한다. 이는 음악 생성 모델이 키와 박자 정보를 직접적으로 반영하여 더욱 다양하고 풍부한 스타일의 음악을 생성할 수 있도록 하는 중요한 접근 방법이다.

4. 음악 특성 추출 모델

본 논문에서 음악 특성 추출을 위해 CNN을 채택하였다. CNN은 지역적 연결성 (Local connectivity)과 가중치 공유 (Weight sharing)의 특성으로 인해, 2차원 데이터에서 공간적 구조 정보를 효과적으로 학습하는 데 뛰어난 성능을 보인다. 이러한 특성은 Mel spectrogram과 같은 2차원 데이터 형태의 음악 정보에서도 효과를 발휘하며, 복잡하고 다양한 음악 특성의 패턴을 식별하는 데 도움을 준다.

본 논문에서는 두 개의 CNN 모델을 독립적으로 훈련하여 각각 음악의 키와 박자 정보를 추출한다. 키 분류를 위한 CNN 모델은 모든 음계의 메이저와 마이너 키를 포함한 총 24개의 클래스로 분류하는 작업을 수행한다. 이에 반해,

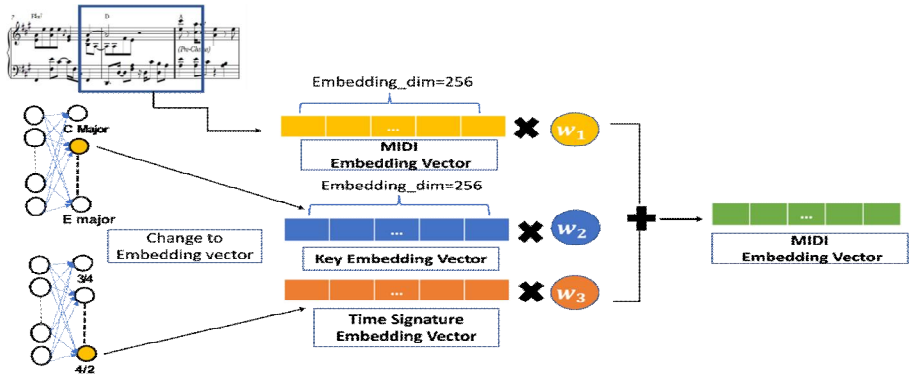


그림 3. 음악 특성 추출 모델 구성도
Fig. 3. Music Classification Model Architecture

박자 분류를 위한 CNN 모델은 일반적으로 사용되는 박자를 포함한 총 6개의 클래스로 분류하는 작업을 수행한다. 이는 1/4, 2/4, 3/4, 4/4, 6/8 박자를 포함하며, 음악의 리듬적 요소를 정확하게 추출하는 데 기여한다. 그림 3은 음악 특성 추출 모델에 관한 구성도이다.

각 CNN 모델의 출력은 음악의 키와 박자에 대한 특성 정보를 담고 있는 벡터로, 이 벡터들은 각각의 특성에 대해 실험 부분의 표 1에서 처럼 최적화된 가중치와 함께 MIDI 이벤트의 임베딩 벡터와 결합된다. 이 과정에서는 가중치 곱셈을 통해 각 특성 정보의 중요도를 조절하고, 이를 통해 보다 정확한 음악 정보 표현을 가능하게 한다.

이렇게 결합된 벡터는 이후 Music Transformer의 입력으로 사용되며, 이를 통해 키와 박자라는 핵심적인 음악적 요소를 음악 생성 과정에 직접적으로 반영한다. 이러한 접근 방식은 음악 생성 모델의 성능과 다양성을 향상시키는데 크게 기여하며, 더욱 실제적이고 다양한 음악 생성을 가능케 한다. 이는 본 연구가 제안하는 음악 특성 추출 모델의 가장 주목할 만한 특징이다.

또한, CNN 모델의 효과적인 학습을 위해, 합성곱 계층(Convolution layer)과 정규화 계층 (Normalization layer), 그리고 비선형 활성화 함수를 이용한 다층 구조를 사용하였다. 합성곱 계층은 지역적 연결성을 보장하며, 입력 데이터의 공간적 구조 정보를 유지하면서 특징을 추출한다. 정규화 계층은 네트워크의 학습 안정성을 보장하며, ReLU 활성화 함수는 모델의 비선형 특성을 증폭시킨다.

이러한 다층 구조는 CNN이 복잡한 패턴을 감지하는 능력을 향상시키며, 본 연구의 음악 특성 추출 모델에서 중요한 역할을 한다. 이렇게 구성된 모델은 MIDI 이벤트와 함께 음악의 키와 박자 정보를 더욱 정확하게 반영하여, 음악 생성의 질과 다양성을 높이는 데에 크게 기여한다.

5. 음악 생성 모델

이 연구의 음악 생성 모델은 Music Transformer를 기반으로 설계되었다. Music Transformer는 Transformer 모델을 활용한 음악 생성 모델로, 음악의 장기 종속성을 처리하는 능력을 강화한 것이 특징이다.

Music Transformer는 MIDI 이벤트를 높은 차원의 벡터로 인코딩하고, 이 인코딩된 벡터는 Transformer의 입력으로 사용된다. 이 모델은 자기회귀적 방식으로 작동하여, 현재의 음표를 생성하는데 이전에 생성된 음표들을 참조한다. 이 과정에서 Transformer의 Multi-Head Attention 메커니즘이 사용되며, 이를 통해 모델은 서로 멀리 떨어진 음표들 사이의 관계를 학습하고, 이를 이용하여 음악의 장기적인 구조를 생성하는 데 필요한 정보를 추출한다.

이 연구에서는 Music Transformer에 음악 특성 추출 모델에서 추출한 키와 박자 정보를 추가하는 과정을 통해, 기존의 모델에 비해 음악 생성의 정밀도와 다양성을 향상시킨다. 이를 위해, 추출된 키와 박자 정보는 각각 독립적인 임베딩 벡터로 변환되고, 이 임베딩 벡터들은 실험적으로 결정된 가중치를 곱한 후 원래의 MIDI 임베딩 벡터와 합쳐진다.

이렇게 결합된 벡터는 Music Transformer의 입력으로 제공되어, 키와 박자 정보가 음악 생성 과정에 직접적으로 반영된다. 이 과정은 생성된 음악의 감정적 톤과 리듬을 세밀하게 제어할 수 있게 하며, 이를 통해 사용자 맞춤형 음악을 생성하는 것이 가능해진다.

본 모델의 성능은 특정 지표를 통해 평가되며, 실험 결과에 따라 모델의 하이퍼파라미터는 조정된다. 본 모델을 통해 생성된 음악이 기존의 방법으로 생성된 음악에 비해 어떠한 차이를 보이는지, 그리고 이 차이가 어떻게 음악의 품질과 다양성에 영향을 미치는 지는지에 대해 심도 있게 탐구하였다.

실제 실험에서는 주어진 키와 박자에 대해 잘 반응하여, 다양한 음악을 생성하였다. 또한, 생성된 음악은 모두 각각의 키와 박자에 잘 맞는 형태를 보였다. 이는 본 모델이 키와 박자 정보를 정확하게 인식하고 이를 음악 생성에 적절히 반영하였음을 나타낸다.

이 연구를 통해, Music Transformer를 기반으로 한 음악 생성 모델이 음악의 키와 박자 정보를 성공적으로 활용하여, 다양하고 품질이 높은 음악을 생성할 수 있음이 확인되었다. 이런 성과는 음악 생성 AI의 성능 향상뿐만 아니라, 사용자 맞춤형 음악 추천, 영화나 게임 등 다양한 분야의 사운드트랙 제작 등에도 응용 가능할 것으로 기대된다.

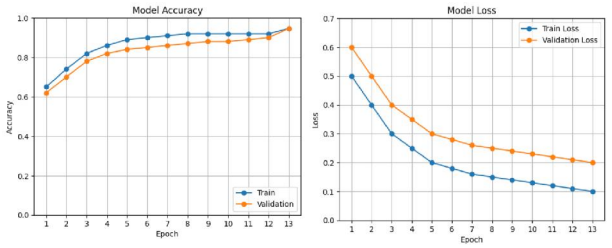


그림 4. 박자에 관한 정확도 및 손실도
Fig. 4. Accuracy and Loss for Time Signature

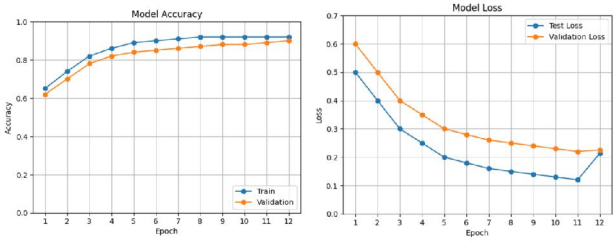


그림 5. 키에 관한 정확도 및 손실도
Fig. 5. Accuracy and Loss for Key

Ⅳ. 실 험

본 논문에서는 음악의 본질적인 요소인 키와 박자를 고려한 조건부 음악 생성을 위한 새로운 모델을 제안하였다. 초기 단계에서는 MIDI 형태의 악보를 WAV 형태로 변환하고, 이를 STFT를 이용하여 Mel Spectrogram으로 변환하였다. 이 과정을 통해 얻어진 정보는 두 개의 Convolutional Neural Networks (CNN)을 통과하며 키와 박자 정보를 추출하는 데 사용되었다.

분류를 위해 사용한 데이터셋은 300개의 원곡과 2700개의 유사곡으로 구성되었고, 이들은 각각 music21 라이브러리와 pretty_midi 라이브러리를 이용하여 키와 박자 정보를 추출 및 레이블링하였다. 데이터셋은 훈련, 검증, 테스트 셋으로 7:2:1의 비율로 분할하였다. CNN은 총 4개의 층으로 구성되었으며, 배치 크기는 32로 설정하였다. 이 CNN을 통해, 박자 분류에서는 13 에포크에서 94.7%의 정확도와 0.1739의 손실을, 키 분류에서는 92.1%의 정확도와 0.1739의 손실을 기록하였다. 그림 4는 박자에 관한 정확도 및 손실도이며, 그림 5는 키에 관한 정확도 및 손실도이다.

이렇게 분류된 키와 박자 정보는 Music Transformer의 입력으로 사용되었다. 각 정보에 대해 정수ID를 부여하고, 이를 차원이 256인 임베딩 벡터로 변환한 후, MIDI 이벤트의 임베딩 벡터와 함께 가중치를 다르게 설정하여 element-wise 합으로 결합하였다. 표 1은 각각의 가중치에 관하여 NLL의 결과를 나타낸 표이다. NLL은 모델의 예측 성능을 평가하는 손실 함수 중 하나로, 낮은 NLL 값은 모델의 예측이 실제 값에 가까울 때 나타난다.

가중치 비율이 8:1:1일 경우 NLL은 2.97, 7:2:1일 경우 3.01, 6:1:3일 경우 2.84, 6:2:2일 경우 2.89로 측정되었다. 이 중 가장 낮은 NLL 값을 보인 비율은 6:1:3으로, 이 설정을 통해 가장 높은 성능을 달성하였다. 이를 통해 입력된 MIDI의 키와 박자를 추출하고 이들의 가중치 비율을 조정하는 다양한 실험을 통해서 원곡과 감성 및 분위기가 차별화되는 멜로디를 자유롭게 생성할 수 있다.

표 1. 임베딩 벡터에 관한 가중치에 따른 NLL
Table 1. Negative Log-Likelihood (NLL) based on Weighted Embedding Vectors

Ratio of Weighted Embedding Vectors	NLL (Negative Log Likelihood)
8:1:1	2.97
7:2:1	3.01
6:1:3	2.84
6:2:2	2.89

표 2. 리스닝 테스트를 통한 모델 성능 평가
Table 2. Evaluation of model performance through listening tests

	Music Transformer	Our Model
Results	12	18

또한 음악 생성 모델에서 가장 많이 쓰이는 성능 평가 방법인 리스닝 테스트를 진행을 하였다. 리스닝 테스트는 30명을 상대로 진행이 되었다. 본 모델과 일반 Music Transformer를 이용하여 생성된 음악에 대하여 진행을 하였으며 표 2와 같이 30명중 18명이 일반 Music Transformer보다 본 모델을 이용하여 생성된 음악이 더욱 성능이 좋다고 판단하였다.

이렇게 음악의 키와 박자라는 본질적 요소를 음악 생성에 통합함으로써, 음악 생성 모델의 사용자 의도 반영을 실현하였으며, 그 효과와 가능성을 실험적으로 검증하였다.

Ⅴ. 결 론

본 연구는 조건부 음악 생성에 대한 새로운 접근법을 제안하며, 이를 통해 음악의 본질적 요소인 키와 박자를 포괄적으로 고려하는 것을 목표로 하였다. 악보를 WAV 형태로 변환하고 Mel Spectrogram으로 다시 변환하는 과정을 거친 뒤, 두 개의 CNN을 통과시켜 키와 박자 정보를 분류하는 과정을 통해 이 목표를 성취하였다. 이 분류 정보는 Music Transformer에 입력으로 사용되며, 이 과정에서 키와 박자 정보에 따라 MIDI 이벤트의 임베딩 벡터와 가중치가 조정된다. 이러한 과정을 통해 음악의 본질적 요소인 키와 박자를 반영하여 보다 자연스럽고 풍부한 음악 생성을 실현하였다.

결론적으로, 본 연구는 조건부 음악 생성에 대한 새로운 방법을 제공하며, 음악 생성 모델의 효과적인 활용과 음악 연구의 진전을 위한 새로운 기회를 제공한다. 이 논문이 제안하는 모델은 음악 연구자들과 개발자들이 보다 복잡하고 다양한 음악적 요소를 통합하여 음악 생성 모델을 구축하는 데 도움이 될 것이다. 이렇게 생성된 모델은 음악 산업에도 다양한 방식으로 적용될 수 있을 것이다. 예를 들어, 개인화된 음악 추천이나 생성, 음악 교육, 콘텐츠 제작 등에 활용될 수 있다.

References

[1] R. Kamien, "Music: An Appreciation," McGraw-Hill, 2007.

[2] J. London, "Hearing in Time: Psychological Aspects of Musical Meter," Oxford University Press, 2004.

[3] H. W. Dong, W. Y. Hsiao, L. C. Yang, Y. H. Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment," Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[4] S. Hochreiter, J. Schmidhuber, "Long Short-term Memory," Neural Computation, Vol. 9, No. 8, pp. 1735-1780, 1997.

[5] I. Goodfellow, "Generative Adversarial Nets," Advances in Neural Information Processing Systems 27, pp. 2672-2680, 2014.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin "Attention Is All You Need," Advances in Neural Information Processing Systems 30, pp. 5998-6008, 2017.

[7] A. R. Selfridge-Field, "Concepts of MIDI: Musical Instrument Digital Interface," Computing in Musicology, Vol. 8, pp. 3-5, 1992.

[8] L. Cohen, "Time-frequency Distributions-a Review," Proceedings of the IEEE, Vol. 77, No. 7, pp. 941-981, 1989.

[9] Y. LeCun, Y. Bengio, G. Hinton, "Deep Learning," Nature, Vol. 521, No. 7553, pp. 436-444, 2015.

[10] H. Huang, T. Cooijmans, A. Roberts, A. Courville, D. Eck, "Music Transformer: Generating Music with Long-Term Structure," Proceedings of ICLR, 2019.

[11] S. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28, No. 4, pp. 357-366, 1980.

[12] T. Qian, J. Kaunismaa, T. Chung, "cMelGAN: An Efficient Conditional Generative Model Based on Mel Spectrograms," arXiv preprint arXiv:2205.07319, 2022.

[13] D. P. Kingma, M. Welling, "Auto-Encoding Variational Bayes," in Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014. [Online]. Available: https://arxiv.org/abs/1312.6114.

[14] A. Roberts, J. Engel, C. Raffel, I Simon, C Hawthorne, "MusicVAE: Creating a Palette for Musical Scores wit h Machine Learning," Google AI Blog, 2018. [Online]. Available: https://ai.googleblog.com/2018/02/musicae-creating-palette-for-musical.html.

[15] N. Jaques, S. Gu, R. E. Turner, D. Eck"Tuning Recurrent Neural Networks with Reinforcement Learning," arXiv preprint arXiv:1611.02796, 2017.

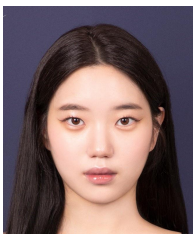
[16] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," arXiv preprint arXiv:1609.03499, 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems 30, pp. 5998-6008, 2017.

[18] H. Huang, T. Cooijmans, A. Roberts, A. Courville, D. Eck, "Music Transformer: Generating Music with Long-Term Structure," in Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019.

[19] http://www.aihub.or.kr/

Seunga Cho (조 승 아)



2023 Department of Software from Duksung Women's University (B.S.)

2023~Hana Financial Group Inc.

Field of Interests: mage Processing, CNN, Neural Network

Email: cat2995@naver.com

Jaeho Lee (이 재 호)

2005 Department of Electronics and Electronical Engineering from Korea University (M.S.)

2013 Department of Electronics and Electronical Engineering from Korea University (Ph. D.)

2020~Department of Software from Duksung Women's University (Assistant Prof.)

Career:

2013~2015 Future Standardization Lab., LG Electronics (Senior Engineer)

2015~2020 Department of Information and Communications Engineering from Seowon University (Assistant Prof.)

Field of Interessts: WPAN, MAC, Bluetooth, Wi-Fi, Localization, NLP, Machine Learning, Generative Model

Email: izeho@duksung.ac.kr