1 인 미디어 창작자를 위한 딥러닝 기반 작곡 어플리케이션

김보경*, 윤소지*, 이승희*, 임예진*, 유견아**, 임성현***
*덕성여자대학교 컴퓨터공학과
**덕성여자대학교 컴퓨터공학과 교수
***우아한형제들
bogyung9981@duksung.ac.kr, lauren9928@duksung.ac.kr, dswu181014@duksung.ac.kr,
lyejin4735@duksung.ac.kr, kyeonah@duksung.ac.kr, sunghyun.lim@gmail.com

Music Composition Application with Deep Learning for content creators

BoGyung Kim*, SoJi Yun*, SeungHee Lee*, YeJin Lim*, KyeonAh Yu**, SungHyun Lim***
*Dept. of Computer Engineering, Duksung Women's University
**Professor, Dept. of Computer Engineering, Duksung Women's University
***Woowahan Brothers

*: these authors contributed equally to this work.

요 약

1 인 미디어 산업의 성장으로 다양한 콘텐츠 제작의 증가와 함께 영상의 분위기를 좌우하는 BGM 의 수요도 급증하고 있다. 그러나 무료 음원은 한정되어 있으며 이미 많은 영상에 쓰여 시청자에게 흔한 느낌을 준다. 특히 MCN 에 소속되지 않은 콘텐츠 크리에이터들은 개성 있고 영상에 어울리는 음원 확보에 어려움을 겪고 있다. 본 연구는 이러한 콘텐츠 제작 환경을 개선하기 위해 창작자가 직접 녹음하거나 악보를 스캔해 자신만의 음원을 제작할 수 있는 웹 애플리케이션 '플랫'을 제안한다. 본 연구를 통해 콘텐츠 크리에이터들은 독창적이고 풍성한 콘텐츠를 만들 수 있으며, 음악적 숙련도와 관계없이 쉽게 음원을 만들 수 있어 작곡에 대한 접근성이 좋아질 것으로 보인다. 또한, 딥러닝을 활용해 음악을 창작함으로써 인공지능 작곡 분야를 활성화하고 디지털 음악 시장의 새로운 분야를 개척하는 데 이바지할 것으로 기대한다.

I. 서론

국내 1 인 미디어 시장은 오랜 기간에 걸쳐 꾸준히 성장해왔다. 2000 년대 초반 판도라 TV 와 아프리카 TV 가 등장하고 이러한 영상 플랫폼으로 수입을 벌어들이면서 콘텐츠 크리에이터 또한 하나의 직업으로 분류되기 시작했다. MCN 비즈니스가 활성화되고 1 인 미디어가 산업화한 것은 2015 년경으로, 현재 1 인 미디어 시장의 규모를 생각하면 성장세가 얼마나 가파른지 짐작할 수 있다[1]. 1 인 미디어 시장과 더불어 OTT 시장 또한 빠르게 성장해왔다. 그중 다양한 콘텐츠를 무료로 감상할 수 있는 유튜브는 독보적인 1 위를 기록하고 있고, 유튜브를 통해 1 인 미디어를 제공하는 '유튜버' 또한 늘고 있다.

콘텐츠 크리에이터들이 유튜브에 영상을 올리는데 있어 단순 콘텐츠와 영상 내용만 필요한 것은 아니다. 영상의 공백을 메워주고 청각적 풍성함을 제공함으로써 영상의 질을 높일 수 있는 배경음악 또한 매우 중요하다. 그러나 유튜브는 상업적 이윤을 창출할 수 있는 플랫폼으로 저작권에 매우 민감하여 영상에서 사용 가능한 무료 BGM 은 매우 한정적이다. 따라서 유튜브 영상들을 보다 보면 BGM 의 유사성을 느낄 수 있다. 따라서 본 연구에서는 이러한 유튜브 영상 속 BGM 의 한계를 넘어설 수 있는 웹 서비스를 제안한다.

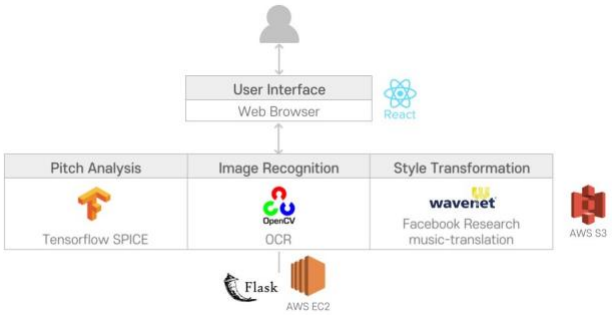
본 연구는 음악적 숙련도와 관계없이 콘텐츠 크리에이터가 직접 BGM 을 제작해 영상에 본인만의 독창적인 분위기를 담아낼 수 있고자 한다. 사용자가 흥얼거리는 음성 혹은 실제 악보를 wav 파일로 변환해 주고, 카메라 앱의 뷰티필터처럼 사용자가 원하는 테마를 입혀 곡의 분위기를 적용함으로써 쉽게 다양한 음악을 만들어 낼 수 있다.

이렇게 제작된 음악은 사용자의 유튜브 영상 속

BGM 으로 사용할 수 있으며, 본인 고유의 색을 입힌 영상을 제작해 영상의 질을 높인다.

II. 작곡 웹 어플리케이션

1) 시스템개요



[그림 1] 시스템 구성도
Fig 1. System Diagram

플랫의 시스템 구성도는 그림 1 과 같다. 음정 분석 모델은 Tensorflow JS 로 변환해 서비스에 내장하고, 악보 이미지 인식 모델은 Flask API 로 AWS EC2 에 배포하였다. 또한 악보 이미지나 음원과 같은 파일들은 AWS S3 를 사용해 저장·관리하였다. 웹 프론트는 React.js 를 사용해 구현하였으며 AWS Cloudfront 를 통해 정적 파일을 호스팅하였다.

2) 음정분석

음정 분석 기능으로는 Tensorflow SPICE 모델[2]을 사용해 음의 높낮이인 피치 추출 기능을 제공한다. 이 모델은 곡 1,000 개의 클립을 모아놓은 MIR-1k dataset 으로 훈련되었다. SPICE 모델의 형식에 맞게 16kHz 의 샘플링 속도와 단 하나의 채널(모노)로 오디오 파일을 입력값으로 넘겨주었다. 또한 output 인 피치 식별에 대한 모델의 신뢰도인 uncertainties 를 이용해 신뢰도가 낮은 모든 피치 추정치를 제거하고 나머지 피치 값들을 제공한다. 내부적으로 피치를 CQT(Constant Q-transform)를 함으로써 낮은 주파수 대역에서 더 높은 주파수 해상도를 가지게 하여 낮은 음질의 해상도를 향상시킨다.

본 연구에서는 최종적으로 산출한 미디 값을 바탕으로 사용자들에게 음의 높낮이를 볼 수 있는 단순화된 악보를 제공한다.

3) 악보 분석

악보 이미지 인식 기능으로는 OCR 기능을 제공한다. OCR(Optical Character Recognition)은 광학 문자 인식이라는 의미로, 장치로 종이에 인쇄되거나 손으로 쓴 문자를 컴퓨터에 입력하는 것을 말한다.

디지털 악보 이미지를 불러오면 윤곽선 검출을 통해 오선 이미지를 만든다. 해당 이미지에서 템플릿 매칭을 통해 오선, 조표(샵, 플랫), 음표(온음표, 2 분음표, 4 분음표, 8 분음표), 쉼표(온쉼표, 2 분 쉼표, 4 분 쉼표, 8 분 쉼표), 잇단음표 등을 인식한다. 이때 지정된 임곗값을 통해 식별 후 미디 파일을 생성한다.

본 연구에서는 OpenCV 로 구현한 오픈 소스인 Sheet Vision[3]에서 쉼표와 음표(8 분음표)를 추가하고 잇단음표를 인식하는 알고리즘을 더해 이미지 인식 기능의 정확도를 높였다. 전처리 과정을 통해 인식 시간을 줄이고, 사용자의 편의를 위해 미디 파일을 wav 파일로 변환한다.

4) 분위기 변환

분위기 변환 기능으로는 WaveNet[4] 기반으로 만들어진 AutoEncoder 를 사용한다. WaveNet 은 Audio 의 autoregressive 한 특징을 Convolution layer 의 변형을 통해 학습하여 audio 를 생성하는 모델이고, AutoEncoder 는 비지도 방식으로 훈련되는 인공 신경망이다.

본 연구에서는 A Universal Music Translation Network [5] 논문을 기반으로 개발된 Facebook Research music-translation[6] 오픈소스를 활용해 해당 기능을 수행한다. 연구의 목적은 그루브한, 신나는, 트렌디한, 차분한 분위기의 domain 을 input 으로 넣어 Data Augmentation 하고 WaveNet Encoder 로 학습한 뒤, domain 분류가 가능한 embedding 을 만들고 WaveNet decoder 로 산출하는 것이었다. 그러나 WaveNet 이 가지는 높은 비용과 긴 훈련 시간으로 인해 학습에 고사양 컴퓨팅 자원을 요구했고, 이는 본 연구에서 사용 가능한 자원의 범위를 벗어났다. 따라서 모델을 학습할 다른 방안을 찾고자 했고 이에 Cycle GAN 을 사용해 보고자 하였다.

Cycle GAN 은 unpaired image to image 학습 데이터로 훈련 가능하며, 허밍 파형의 CQT 스펙트로그램을 타겟 분위기 CQT 스펙트로그램으로 변환한다. 이후 해당 스펙트로그램을 WaveNet 을 이용해 파형으로 복원하고, 소리로 듣는다[7]. 본 연구에서는 Cycle GAN 모델을 생성하고자 GTZAN 데이터셋을 통해 신나는, 트렌디한, 차분한 분위기의 음원을 CQT 스펙트로그램으로 생성하였다. 구글 Machine Perception research 에서 수집한 허밍 데이터 또한 CQT 스펙트로그램을 생성하였으나, 해당 데이터는 30% 정확도로 품질이 낮아 이를 보완할 데이터를 구축해야 한다.

창작물을 풍성하게 할 수 있는 더 다양한 옵션을 제공하기 위해 분위기뿐만 아니라 악기 소리로도 음원을 변환할 수 있는 기능도 추가하였다. 미디를 wav 로 변환하기 위해서는 soundfont 가 필요한데, 다양한 악기의 soundfont 를 사용해 wav 로 변환하여 악기 소

리를 생성했다. 분위기와 악기 변환을 동시에 제공함으로써 템스가 하나 늘어난 만큼 사용자가 이해하기 쉽게 UI 를 개선해야 할 것으로 보인다.

III. 시스템 구현 및 테스트

본 연구는 앞서 설명한 바와 같이 크게 음정 분석, 분위기 변환, 악보 이미지 인식 부분으로 나뉘며 웹을 통해 기능을 제공한다. 서비스의 전체 흐름도는 아래 그림 2 와 같다.

[그림 2] 시스템 흐름도
Fig 2. System Flow Chart

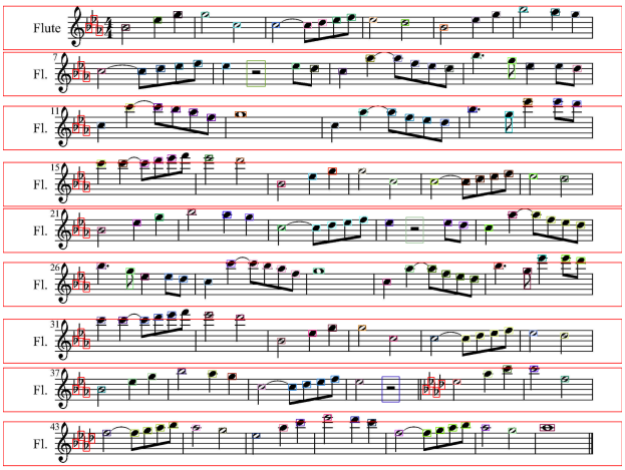
메인화면에서 음정 분석과 분위기 변환을 제공하는 작곡 서비스 혹은 악보 이미지 인식 서비스를 제공하는 악보 들어보기 페이지로 이동한다.

작곡 페이지로 들어가면 SPICE 모델을 불러오고 허밍을 녹음할 MediaStream API 인 getUserMedia 의 권한 설정을 한다. 사용자가 미디어의 사용을 허락하면 서비스 이용을 할 수 있다. 해당 페이지 UI 는 작곡 기능을 수행할 수 있는 “나만의 BGM 을 만들어보세요” 세션과 데모 세션으로 구성되어 있다. 작곡 세션은 사용자 편의성을 위해 STEP 1, 2, 3 으로 가이드라인을 제공한다. STEP 1 은 사용자가 허밍을 녹음하고 그림 3 같이 녹음한 것을 점그래프 악보로 단순화시킨다. STEP 2 는 녹음한 음원을 바탕으로 사용자가 원하는 다양한 스타일(그루브한, 신나는, 트렌디한, 차분한 분위기)을 적용할 수 있다. STEP 3 은 창작된 BGM 을 들어보고 다운로드하여 활용할 수 있다.

[그림 3] 작곡 페이지
Fig 3. Compose page

악보 들어보기 페이지는 악보를 볼 줄 모르는 사용자도 전문가의 악보를 업로드하고 들어 봄으로써 음악적 이해도를 높이고 창작에 영감을 받는 공간이다.

해당 페이지로 들어가면 악보 업로드 버튼을 눌러 디지털 악보를 업로드할 수 있다. PC 에 저장된 디지털 악보를 업로드하면 이미지 파일을 AWS s3 로 전송하고 악보 분석 Flask API 를 호출한다. API 가 호출되면, 서버에서 저장된 템플릿과 사용자가 업로드한 디지털 악보의 템플릿을 서로 매칭시킨 후 분석이 완료되면 음정과 박자를 피아노로 청각 화해 음원으로 변환한다. 완성된 음원은 s3 에 업로드하고 음원 주소를 API response 값으로 반환한다. API 가 음원 주소를 성공적으로 반환하면 audio 컴포넌트를 렌더링해 들어보고 다운로드 할 수 있다.



[그림 4] 악보 이미지 인식
Fig 4. Image recognition

IV. 결론

본 연구에서는 허밍만으로 손쉽게 BGM 을 생성할 수 있는 웹 솔루션을 개발하였다. 이 웹 애플리케이션은 음악적 숙련도와 관계없이 악상이 떠오르면 즉시 허밍을 해 쉽게 음원을 제작할 수 있고, 점그래프 악보를 보며 음의 높낮이를 확인할 수 있어 작곡에 대한 접근성을 향상하고 진입 장벽을 낮추었다. 이러한 접근으로 콘텐츠 크리에이터 뿐만 아니라 작곡에 관심 있는 모든 사람이 이용이 가능하다. 또한, 악보를 볼 줄 모르는 사용자를 위해 디지털 악보를 인식하여 음을 읽고 피아노 음향으로 변환시켜 들려준다. 이를 통해 전문가의 악상으로부터 음악적 이해도를 높이고 창작에 영감을 받을 수 있다. 이는 BGM 의 한계로 인해 콘텐츠 제작에 어려움을 겪는 콘텐츠 크리에이터가 독창적이고 풍성한 결과물을 만들 수 있도록 도움을 준다.

본 연구를 진행하면서 주요 기능들을 완성도 높게 구현하는 데에 몇 가지 한계에 부딪혔다. 첫 번째로 분위기 변환은 네트워크가 무거워 인풋 데이터 해상도 높이는 데 어려움이 있었고, 음을 생성하는 모델인 WaveNet 이 파형을 생성하는데 비용이 많이 들고, 모델을 훈련시키는데 고사양의 컴퓨팅 자원을 요구하여 활용도가 낮았다. 두 번째로 악보 이미지 분석에서 템플릿을 매칭할 때, 서비스 내부에 저장된 템플릿과 모두 비교 매칭하기 때문에 오선의 양에 비례해 시간이 소요된다.

향후에는 사용자가 만든 창작물에 대한 저작권 보호가 필요할뿐더러 창작물이 저작권에 위배가 되는가에 대한 판별이 필요하다. 본 논문은 사용자의 허밍 자체가 기존의 음원과 유사하지 않다는 가정하에 작성하였으나, 결과물이 실제로 저작권에 위반되지 않는지 확인하는 연구는 계속 진행 중이다. 현재로서는 저작권 위배 여부를 판별할 수 있는 명확한 기준이 존재하지 않아 QbSH (Query-by-Singing/Humming)[8] 시스템을 사용해 창작물과 이미 등록된 음원의 유사도를 추출하고자 한다. 또한 창작물에 대한 저작권을 보호할 방법으로는 창작물의 메타데이터를 가진 NFT 를 부여해 소유권을 입증하고자 한다[9][10].

- 본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다 -

참고문헌

[1] 유다정, “8 조원 '1 인 미디어' 산업, 5G 시대 혁신성장의 새로운 기회될까”, 디지털투데이, 2019 년 08 월 30 일자

[2] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi and M. Velimirović, "SPICE: Self-Supervised Pitch Estimation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1118-1128, 2020

[3] Cal Prett, Alex Reichenbach, “Sheet Vision”, Github, https://github.com/cal-pratt/SheetVision

[4] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, K. Simonyan, “Neural audio synthesis of musical notes with WaveNet autoencoders” ICML 17: Proceedings of the 34th International Conference on Machine Learning, Sydney NSW Australia, 2017, pp. 1068–1077

[5] N. Mor, L. Wold, A. Polyak and Y. Taigman, “A Universal Music Translation Network”, ICLR, New Orleans, Louisiana, United States, 2019, pp. 1-13

[6] Facebookresearch, “music-trainslation”, Github, https://github.com/facebookresearch/music-translation

[7] S. Huang, Q. Li, C. Anil, X. Bao, S. Oore, R. B. Grosse, “TIMBRETRON: A WAVENET(CYCLEGAN(CQT(AUDIO))) PIPELINE FOR MUSICAL TIMBRE TRANSFER”, ICLR, New Orleans, Louisiana, United States, 2019, pp. 1-17

[8] 박성주, 정광수, “다성음원 기반 QbSH 시스템을 위한 매칭엔진의 설계 및 구현”, 멀티미디어학회논문지, 제 15 권, 제 1 호, pp. 18-34, 2012

[9] 이지훈, 이영신, 남현우, “블록체인 기반의 영상저작물 저작권 보호 및 유동 모델 가이드라인 연구”, 한국디자인리서치학회, 6 권, 1 호, pp. 18-172, 2021

[10] 김현경, “저작권 관점에서 NFT 에 대한 미국의 관련 법적 검토”, 한국저작권보호원, 글로벌이슈리포트, pp. 1-15, 2021