사운드 디자인을 위한 K-POP 음악의 Wave-U-Net 및 주파수 분석을 통한 자동 Bass line 표기

Automatic transcription of bass line using Wave-U-Net and frequency analysis in K-POP music for Sound Design

주 저 자 : 최진솔(Choi, Jin Sol) 광운대학교 전자통신공학과

공동저자 : 김경철(Kim, Kyoung Chul) 광운대학교 전자통신공학과

교신저자 손채봉(Sohn, Chae-Bong) 광운대학교 전자통신공학과
cbsohn@kw.ac.kr

접수일자 2019. 8. 26. / 심사완료일자 2019. 9. 14. / 게재확정일자 2019. 9. 22.
본 논문은 중소벤처기업부에서 지원하는 2018년도 산학연협력 기술개발사업(No. S2658023)의 연구수행으로 인한 결과물임을 밝힙니다.

Abstract

In recent K-POP music, the melody line is important, but also how to design the sound that composes the song becomes an important factor. There are many various active research using artificial intelligence in the music field, such as automatically separating the desired audio from the mixture audio and composing music by themselves. In this paper, we use 'Wave-U-Net', which specializes in separating sound sources containing various sound sources, which are the characteristics of K-POP music or Electronic music, which are popular all over the world. The particular instrument here focuses on the 'Bass'. We train Wave-U-Net using bass characteristics that have significantly low frequency band than other instruments. And we analyzes the bass frequency values from random songs using Mel-Spectrogram. Lastly, we experiment to automatically analyze the chord progression of the random song. In order to make the best use of the frequency characteristics of bass instruments, experiments are conducted by comparing three methods 'Low Pass', 'Bass Boost' and 'Low Pass + Bass Boost' and we select the best model. Finally, based on the results of this experiments, we propose ideas that can be applied.

Keyword

Sound Design, Wave-U-Net, Frequency analysis, Weighted Average, K-POP music

---

요약

최근의 K-POP 음악은 멜로디 라인도 중요하지만 곡을 구성하는 사운드를 어떻게 디자인 하는지가 중요한 요소가 되었다. 자동으로 mixture audio에서 원하는 오디오를 분리해 내고, 작곡도 스스로 하는 등 음악 분야에서의 인공지능을 이용한 연구가 활발하게 진행되고 있다. 따라서 본 논문에서는 최근 전 세계적으로 인기를 끌고 있는 K-POP 음악 혹은 전자음악의 특징인 다양한 사운드소스가 섞인 음원을 각각의 사운드소스로 분리하는 것에 특화되어있는 Wave-U-Net을 활용하여 특정 악기를 분리한다. 여기서 특정 악기는 본 논문에서 베이스에 초점을 둔다. 다른 악기에 비해서 두드러지게 낮은 주파수 대역을 가지는 베이스 특성을 이용하여 Wave-U-Net을 통해 학습되어 임의의 곡에서 추출된 베이스 사운드소스 주파수 값을 Mel-spectrogram을 통해 분석하고, 곡의 코드 진행을 자동으로 분석해 주기 위한 실험을 진행한다. 베이스 악기의 주파수 특징을 최대한 활용하기 위해 Low pass, Bass boost, Low pass + Bass boost 3가지 방식을 적용하여 실험을 진행하여 비교하고, 각 방법으로 분석한 베이스 라인의 정확도를 분석하여 이를 원본 음원으로부터 추출한 베이스 라인과 비교하여 최적의 방법을 선정한다. 마지막으로 본 실험을 진행하며 도출한 결과를 바탕으로 적용할 수 있는 아이디어를 제시한다.

목차

1. 서론

1-1. 사운드 디자인

사운드 디자인은 음향의 다양한 요소들을 영화, 방송, 극장 등 사용되어지는 상황에 맞추어 원하는 분위기나 효과를 내기 위한 작업을 말한다. 따라서 사운드 디자인은 곡의 감정을 이끌어내고 분위기를 주도하는데 굉장히 중요한 역할을 한다.

1-2. 연구의 배경과 목적 및 방법

최근 K-POP 음악의 구성을 분석해 보면 대체적으로 Intro, Verse1, Build-up/Bridge, Drop/Chorus, Verse2, Build-up/Bridge, Drop/Chorus, Bridge, Drop/Chorus, Outro 형태의 진행으로 이루어진다. 이러한 형식을 갖춘 곡들은 보통 Bridge 파트를 제외하고는(모든 곡이 해당하는 것은 아니지만) 같은 코드의 반복을 기반으로 하여 각 파트에서 다른 악기를 사용하거나 박자감의 변화 등 전반적인 곡의 느낌에서 기승전결(起承轉結)이 느껴지게끔 사운드를 디자인하고 구성하는 것이 일반적이다. 보통 후렴구 파트인 Drop/Chorus는 곡의 전(轉)을 표현하기 때문에 가장 화려하거나, 풍성한 사운드를 표현한다. 즉 저주파 대역부터 고주파 대역을 아우르는 악기들로 채워진다. 이렇게 음악에서는 단순한 멜로디 라인도 중요하지만 전반적인 사운드의 디자인을 어떻게 구성하는지에 따라 곡의 완성도가 완전히 달라질 수 있다.

음악에서 가장 아랫단을 받쳐주며 곡의 전반적인 탄탄함을 느껴지게 하는 역할을 하는 악기인 베이스는 대부분 곡의 코드 진행에 근음(根音)을 연주한다. 따라서 베이스음을 듣는 것이 곡의 코드 진행을 파악하기 위한 가장 쉬운 방법 중 하나이다. 절대음감을 소유한 사람일 경우에는 곡을 듣고 바로 곡의 코드를 음계에서 찾아내는 것은 어려운 일이 아닐 것이다. 혹은 절대음감이 아니더라도 청음이 능숙하게 가능한 사람이라면 직접 음을 하나씩 짚어 가며 찾을 수 있기 때문에 이는 역시 크게 문제되지 않는다. 하지만 상대적인 음계의 차이를 인지하는 것에 어려움을 느끼는 사람이라면 악보가 존재하지 않는 곡을 연주하는 것은 사실상 불가능 할 수 있다.

따라서 본 논문에서는 베이스 라인 음 혹은 코드를 귀로만 듣고 정확한 음계를 찾아내는 것에 어려움을 느끼는 사람이 많은 것에 착안하여 Wave-U-Net을 통해 베이스 악기를 분리해 낸 뒤, 분리된 베이스 음계를 Low Pass, Bass Boost, Low Pass + Bass Boost 3가지 방식을 적용하여 각각의 주파수를 비교 분석하고 가장 정확한 분석을 하는 방법을 이용하여 자동으로 베이스 라인 또는 곡의 코드진행을 알려주기 위한 실험을 진행한다. 그리고 옥타브 또한 코드와 함께 출력되도록 한다. 그러나 추출된 음계만으로는 major 혹은 minor 등 코드의 디테일함을 표현하는 부분에서는 무리가 있기 때문에 이 부분은 생략하고, 근음(根音)만을 나타내는 것을 본 논문에서는 기본으로 한다.

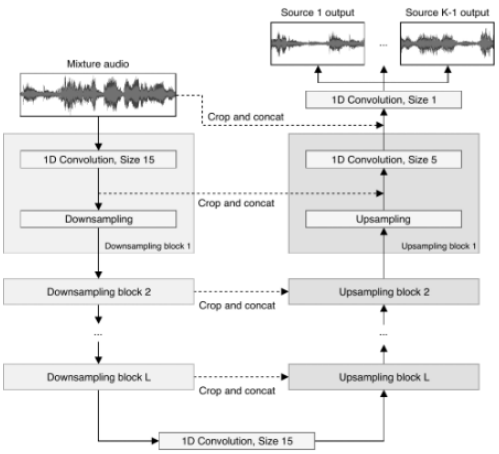
앞서 사운드 디자인은 곡의 감정을 이끌어 내고 분위기를 주도하는 것에 있어 굉장히 중요한 역할을 한다고 언급하였는데, 본 논문에서는 사운드소스의 분석에 따라 추출된 베이스 라인의 시각화를 통해 사운드 디자인적으로 어떤 의미와 가치를 지니고 또한 어떤 기능을 가질 수 있는지 그 연관성을 제시한다.

2. 이론적 배경

2-1. Wave-U-Net

Wave-U-Net은 다양한 사운드소스가 섞여 있는 음

원에서 각 성분의 분리에 최적화 되어있는 네트워크 중 하나이다. 이 네트워크에는 사운드소스 분리를 위한 다양한 모델이 있는데 특히 음성 분리에 대한 모델이 대부분을 차지한다. 그러나 본 논문에서는 multi-instrument source separation에 적합한 'Model 6'을 이용하여 학습을 진행한다. 단순하게 Low frequency 특징을 추출하여 실험을 진행하기에는 베이스 라인 외에도 드럼이나 다른 사운드소스 에서도 Low frequency를 가지는 사운드소스가 있기 때문에 잡음이 섞일 확률이 높아 정확도가 떨어질 것으로 판단되어 Wave-U-Net을 활용한다.



[그림 1] Wave-U-Net 구조

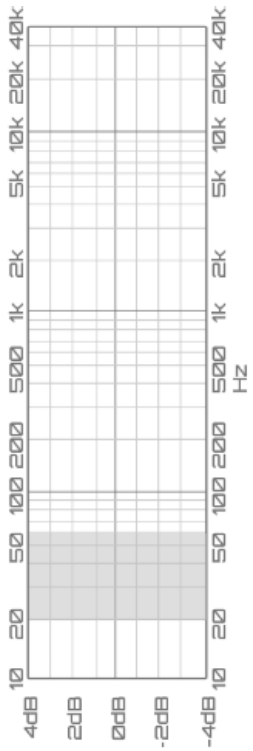
[그림 1]은 Wave-U-Net의 구조[1]를 보여준다. Wave-U-Net은 의료 영상 또는 이미지 segmentation에서 가장 많이 활용되는 U-Net에서 스펙트로그램(Spectrogram)을 이용한 time-domain 음원 분리 네트워크이다.

2-2. Frequency

사람의 귀로 들을 수 있는 음파의 주파수를 '인간의 가청 주파수'라고 하는데, 개인의 청력, 연령대 그리고 음의 크기에 따라 각기 다르지만, 보통 20Hz에서 20kHz 대역을 '가청 주파수대(Audio Frequency Band)'라고 한다. 그러나 본 논문에서는 베이스 악기를 이용한 실험이기 때문에 베이스 주파수 대역에만 초점을 맞추었다.

일반적으로 베이스 악기가 가지는 주파수 대역은 20 ~ 250Hz 인데, 주파수 대역에 따라 세부적으로 나누었을 때 Sub Bass, Bass 두 가지로 나눌 수 있다.

Sub Bass : 20 ~ 60 Hz[2]



[그림 2] Sub Bass 주파수 대역 범위

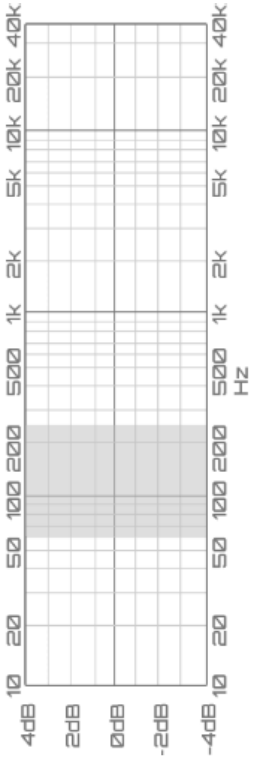
20 ~ 60 Hz 사이 주파수 대역의 베이스를 Sub Bass 라고 한다. Low frequency 대역을 차지하는 드럼의 kick 중에서도 low kick은 보통 60 ~ 80 Hz 영역에 속하기 때문에, Sub Bass 주파수 대역은 베이스 악기를 제외하고는 대부분의 악기가 잘 가지지 않는 주파수 대역이기도 하다.

Bass : 60 ~ 250 Hz[3]

1) Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation." arXiv preprint arXiv:1806.03185 (2018).

2) https://www.teachmeaudio.com/mixing/techniques/audio-spectrum/

3) https://www.teachmeaudio.com/mixing/techniques/audio-spectrum/



[그림 3] Bass 주파수 대역 범위

해당 주파수 영역은 소리의 두껍고 얇음을 결정한다.

최근 대부분의 음악에서 베이스는 90 ~ 200Hz 사이에 분포하고 있기에, 본 논문에서는 가청 주파수대역 시작인 20Hz에서 베이스 주파수 대역보다 조금 아래인 180Hz 까지를 이번 실험에서 측정할 주파수 영역으로 두고 실험을 진행한다.

아래의 [표 1][4]은 20 ~ 180 Hz 사이의 음계 별 해당 주파수이다. 알파벳 옆의 숫자는 해당 옥타브(Octave)를 의미한다.

[표 2] 코드별 해당 주파수 대역

Note	Frequency (Hz)
E0	20.60
F0	21.83
F#0/Gb0	23.12
G0	24.50
G#0/Ab0	25.96
A0	27.50
A#0/Bb0	29.14
B0	30.87
C1	32.70
C#1/Db1	34.65
D1	36.71
D#1/Eb1	38.89
E1	41.20
F1	43.65
F#1/Gb1	46.25
G1	49.00
G#1/Ab1	51.91
A1	55.00
A#1/Bb1	58.27
B1	61.74
C2	65.41
C#2/Db2	69.30
D2	73.42
D#2/Eb2	77.78
E2	82.41
F2	87.31
F#2/Gb2	92.50
G2	98.00
G#2/Ab2	103.83
A2	110.00
A#2/Bb2	116.54
B2	123.47
C3	130.81
C#3/Db3	138.59
D3	146.83
D#3/Eb3	155.56
E3	164.81
F3	174.61

2-3. Instruments

K-POP 음악 혹은 최근 대다수의 전자음악을 구성하는 사운드소스를 주파수 대역에 따라 대략적으로 분류를 하면 아래의 [표 2]와 같다.

[표 3] K-POP 혹은 최근 대부분의 전자음악을 구성하는 사운드 소스 별 주파수 대역 범위

Range	Instruments
Low	Bass, Drum(Kick), Percussion, Brass
Mid	Piano, String, Synth, Brass, Drum(Snare, Toms)
High	FX, String, Synth, Drum(Hi-hat, Cymbal)

실질적으로는 [표 2]에서 나오는 악기들이 표기한 해당 주파수대 에서만 있는 것은 아니고, 다양한 주파

4) http://pages.mtu.edu/~suits/notefreqs.html

수 대역을 가질 수 있다. 그러나 같은 주파수대역을 가진 여러 사운드소스가 동시에 나올 경우, 주파수 대역이 겹치게 되어 해당 사운드소스 음색의 고유한 특성이 변질될 수 있기 때문에 음악 제작 시 최대한 주파수 대역을 고르게 배치하도록 디자인 하고, Mixing 과정에서 각각의 사운드소스의 균형을 잡아 주는 작업을 진행하는 것이 필수적이다. 따라서 본 논문에서는 코드 혹은 음계 추출의 정확한 실험 결과를 도출하기 위하여 음계를 가지는 악기 중 최대한 다른 악기와의 주파수 대역의 겹침이 적은 베이스 악기를 바탕으로 실험을 진행한다.

3. 실험

3-1. Dataset

Wave-U-Net을 학습하기 위해 사용한 dataset은 MUSDB18(Music Separation Database)[5]이다. 이 dataset은 train과 test 폴더로 나누어져 있다. train에는 100곡이 포함되어있는데, 이 중 75곡은 랜덤하게 선정되고 나머지 25곡은 validation set으로 사용된다. test에는 50곡으로 구성되어 있다. 한 곡당 모든 사운드소스가 조합되어있는 원본 음원, Vocal, Drum, Bass 그리고 그 외 악기인 Other로 구성되어 있다. 모든 파일은 Stereo 타입 이며 44.1kHz로 인코딩 되어 있다.

3-2. Procedure(과정)

3-1.1 Wave-U-Net

Wave-U-Net은 입력된 음원에서 Vocal, Drum, Bass, Other 4가지 성분을 분리하도록 학습된다. 정확한 베이스 라인 추출을 위해 다양한 방법으로 학습을 진행한다. 학습을 진행하기 전에는 전체 성분을 학습하는 것 보다 최소한의 성분 분리를 위한 학습 결과가 더 좋을 것으로 예상하였으나, 학습 결과 기존의 방식 Bass, Drum, Vocal, Other 4가지 성분을 모두 분리하도록 학습시킨 방법이 훨씬 각 성분들 마다 깔끔하게 분리가 되는 것을 확인했다. 그러나 4가지 성분 모두 분리하는 것이 아니고 베이스만을 분리하기 위해 학습시킨 방법, 혹은 베이스 포함 그 외 다른 성분들을 포함하여 학습시킨 결과는 전반적으로 해당 성분에 잡음이 많이 섞여 분리가 깨끗하지 못했다. 따라서 본 논문에서는 모든 성분을 분리하는 기존의 방식을 이용한다.

3-1.2 4 Different Method

우선 Low Pass의 경우 베이스의 주파수대가 다른 악기에 비해 매우 낮은 특성을 이용하여 1kHz 이상의 주파수 대역은 모두 잘라내고, 1kHz 이하의 주파수 대역의 음원을 사용한다. Bass Boost는 주요 베이스 악기의 주파수대인 20 ~ 200Hz의 dB값을 다른 주파수대에 비해 높여줌으로써 베이스 악기 소리가 더욱 두드러져 들리게 하여 실험에 사용한다. 마지막으로 앞의 두 방법을 합쳐 1kHz 이하의 주파수 대역만을 남기고 여기서 20 ~ 200Hz 사이의 dB 값을 높여준 음원을 실험에 사용한다. 1kHz 이상의 주파수 대역을 잘라내고, Bass Boost 하는 과정은 FabFilter Pro-Q3[6]를 사용한다.

원본 음원과 전처리를 적용시킨 3가지 음원을 학습된 Wave-U-Net을 통해 베이스 라인을 분리한다.

3-2. 과정

실험 테스트를 위해 K-POP 음악 중 대중적이면서 후렴구가 가장 클라이맥스를 잘 표현하며, 베이스 악기가 후렴구에 선명하게 들리는 곡으로 선별하여 진행했다. 그리고 Bridge 파트를 제외하고 코드 진행이 거의 동일한 곡으로 선정했다. [표 3]은 앞서 제시한 코드별 해당 주파수([표 1])에서 최솟값과 최댓값을 설정한 것이고, 해당 범위 안의 주파수 값을 출력하게끔 지정해 두었다.

[표 4] 코드 별 주파수 대역 설정 값

Note	Min Frequency (Hz)	Max Frequency (Hz)
E0	20.025	21.215
F0	21.215	22.475
F#0/Gb0	22.475	23.810
G0	23.810	25.230
G#0/Ab0	25.230	26.730
A0	26.730	28.320

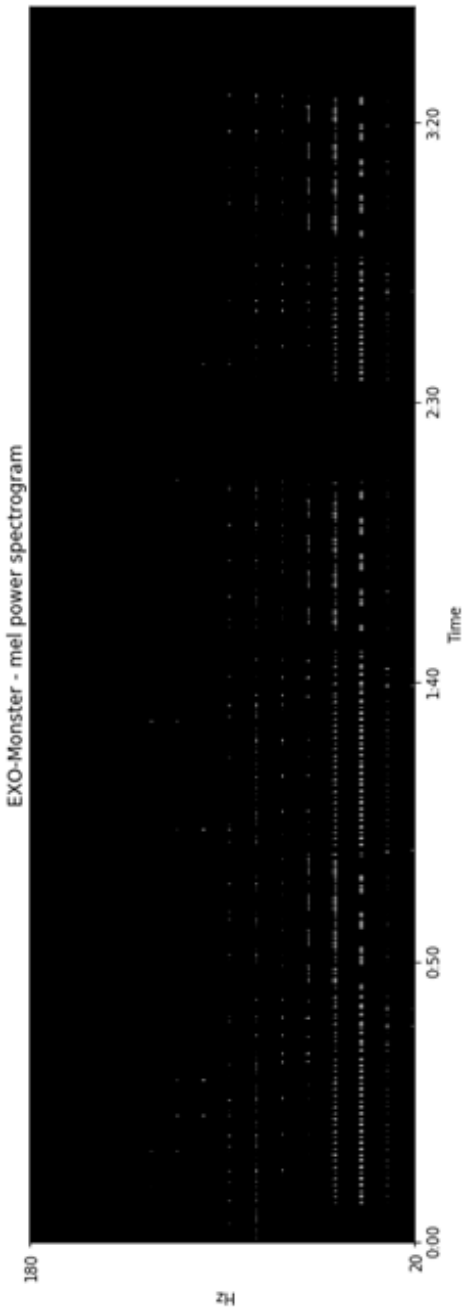
5) https://sigsep.github.io/datasets/musdb.html

6) https://www.fabfilter.com/products/pro-q-3-equalizer-plug-in

A#0/Bb0	28.320	30.005
B0	30.005	31.785
C1	31.785	33.675
C#1/Db1	33.675	35.680
D1	35.680	37.800
D#1/Eb1	37.800	40.045
E1	40.045	42.425
F1	42.425	44.950
F#1/Gb1	44.950	47.625
G1	47.625	50.455
G#1/Ab1	50.455	53.455
A1	53.455	56.635
A#1/Bb1	56.635	60.005
B1	60.005	63.575
C2	63.575	67.355
C#2/Db2	67.355	71.360
D2	71.360	75.600
D#2/Eb2	75.600	80.095
E2	80.095	84.860
F2	84.860	89.905
F#2/Gb2	89.905	95.250
G2	95.250	100.915
G#2/Ab2	100.915	106.915
A2	106.915	113.270
A#2/Bb2	113.270	120.005
B2	120.005	127.140
C3	127.140	134.700
C#3/Db3	134.700	142.710
D3	142.710	151.195
D#3/Eb3	151.195	160.185
E3	160.185	169.710
F3	169.710	179.805
Pause	0.000	0.000

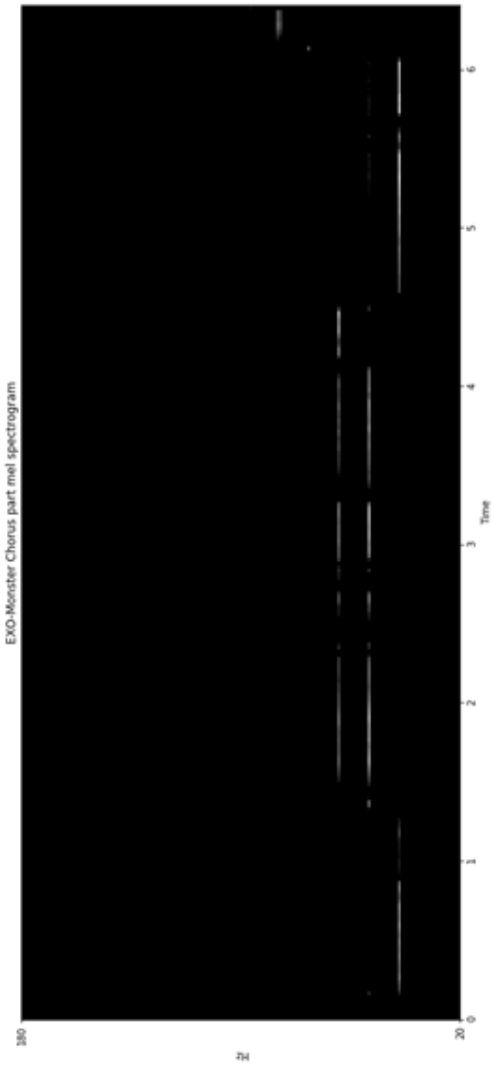
4. 실험결과

첫 번째 실험 결과는 Wave-U-Net을 활용하여 학습시킨 결과로 임의의 K-POP 음악의 원본을 적용했을 때 나온 결과이다. [그림 4]는 EXO의 곡인'Monster'를 적용했을 때 곡 전체에 대한 Mel Power Spectrogram이다.



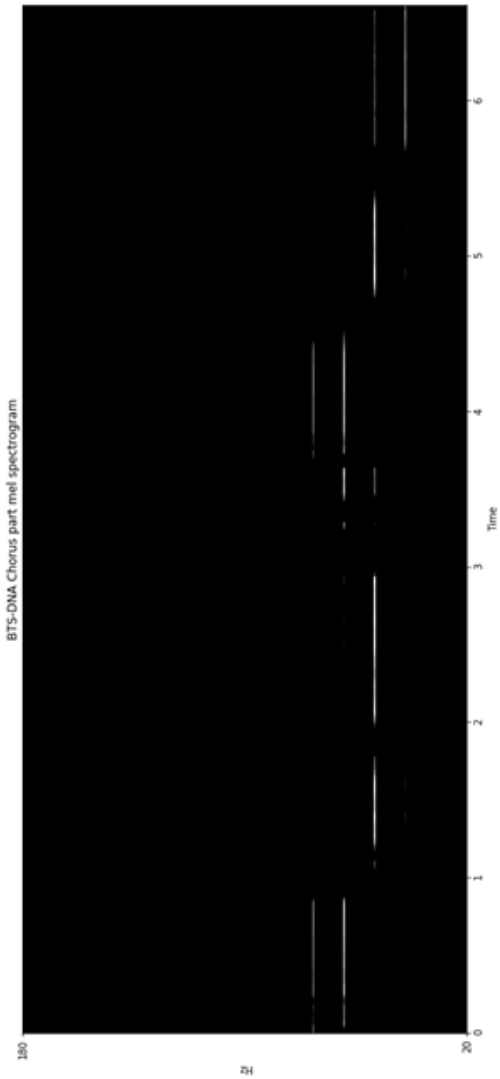
[그림 4] EXO - Monster Bass Mel Power Spectrogram

본 논문에서는 대부분의 곡들이 특정 부분(Bridge part)을 제외하고 같은 코드 진행을 한다고 앞서 언급하였기 때문에 곡에서 가장 하이라이트라고 할 수 있는 Chorus part만을 따로 잘라내어 분석했다.

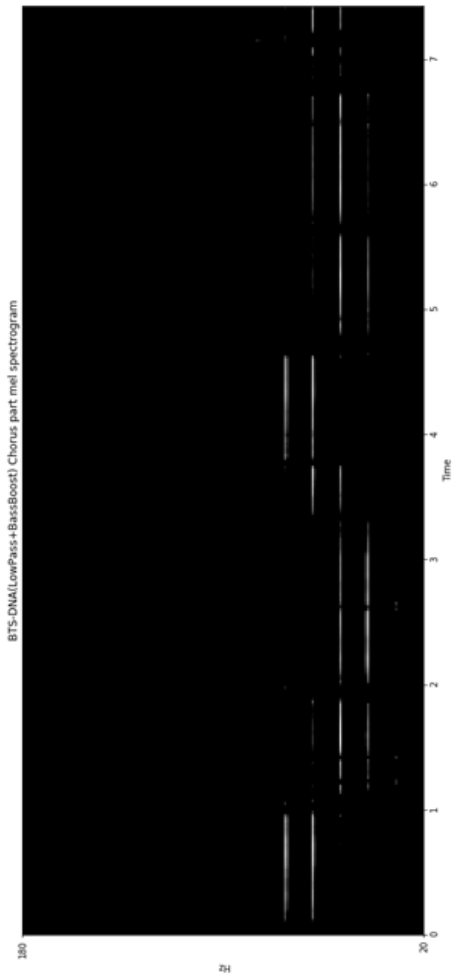


[그림 5] EXO - Monster의 Chorus 파트의 Bass Mel Power Spectrogram

[그림 5]에서 볼 수 있듯이, time domain에 따른 주파수 값의 dB 값이 밝게 보이는 것을 확인할 수 있다. 또 1초 초반 대에서 4초대 까지 이어지는 부분을 보면 선명한 색상으로 서로 다른 주파수대가 두 줄로 나오는 것을 확인할 수 있는데, 이는 더 높은 dB 값을 가지는 frequency대에 가중치를 부여하여 가중평균을 통해 계산한 결과, 곡의 베이스 라인을 제대로 출력하는 것을 확인했다. 원래 EXO의'Monster' Chorus part의 베이스 라인은 'F Bb / Bb C F F'인데 실험 결과 'F1 B1 / B1 C2 F1 F2'로 마지막 두 개의 F 코드는 옥타브가 다른 코드라는 것 까지 잘 표현하는 것을 확인할 수 있었다.



[그림 6] 방탄소년단(BTS) - DNA Chorus 파트 - Original mode



[그림 7] 방탄소년단(BTS) - DNA Chorus 파트 - Low Pass + Bass Boost mode

[그림 6]과 [그림 7]은 방탄소년단(BTS)의 곡인 'DNA'의 전주 부분의 베이스 파트만을 추출한 것이다. [그림 6]은 어떠한 전 처리를 하지 않은 원본 그대로의 곡을 통해 추출한 베이스의 Mel Power Spectrogram 이고, [그림 7]은 1kHz 이상의 frequency는 제거하고 1kHz 이하의 frequency 중에서도 20 ~ 200Hz 사이의 dB 값을 동일하게 올려준 후 추출한 베이스의 Mel Power Spectrogram이다.

원래 이 곡의 베이스 코드 라인은 'C# G# A B / C# G# F# B'인데, Original mode인 [그림 6]의 경우에 가중평균을 이용하여 베이스 라인을 추출하였을 때 'C2 A1 A1 B1 / C2 A1 F1 B1' 으로 나와 원곡의 코드에 반음이 올라가거나 혹은 내려가 약간의 오차가 발생하였다. 그러나 Low pass와 Bass Boost를 함께 적용한 [그림 7]의 경우에는 'C#2 G#1 A1 B1 / C#2 G#1 F#1 B1' 으로 정확하게 코드 진행을 표현하는 것을 확인했다. 그 외에도 나머지 두 방법인 Low Pass 만을 이용한 방법과, Bass Boost만을 이용한 방법은 곡에 따라 정확하게 코드를 확인할 수 있는 것도 있었지만, 약간의 오차를 출력하는 경우도 발생하는 것을 확인하였다. 따라서 본 실험을 통해 틀린 코드를 출력하는 횟수가 가장 낮은 Low Pass + Bass Boost mode가 다른 나머지 3가지 방법보다 훨씬 좋은 성능을 낸다는 것을 확인할 수 있었다.

5. 결론

본 실험에서 Wave-U-Net과 여러 가지 전 처리 방법을 통해 K-POP 음악의 베이스 라인 혹은 코드 진행을 추출하는 것을 실험했다. MUSDB18 dataset을 활용하여 multi-instrument separation을 위한 Model 6의 트레이닝을 진행하였고, Original mode, Low Pass mode, Bass Boost mode, Low Pass + Bass Boost mode 4가지 각기 다른 전 처리를 통하여 최종적으로 Low Pass + Bass Boost mode가 가장 정확한 코드를 추출하는 방법임을 실험을 통해 확인할 수 있었다.

그리고 추출된 결과를 통해서 대략적인 베이스 연주 형태 또한 확인할 수 있었다. 하나의 코드를 정확하게 끊어서 연주하는 방법은 출력 결과 코드와 코드 사이에 Pause 되는 구간을 설정해 두었기 때문에 Pause로 출력되는 추간에서 끊어짐을 확인할 수 있었다. 그리고 하나의 음과 다음 음이 자연스럽게 이어지는 형태의 경우에는 해당 코드 사이에 이어지는 코드가 출력됨을 확인하였다.

현재 이를 통해 프로그래밍 언어 중 python을 기반으로 하는 lilypond 사보 프로그램과 연동하기 위한 연구를 계속해서 진행 중이다. 추후 본 논문에서 추출한 결과 값을 보정하여 lilypond 사보 프로그램과의 연동을 성공적으로 진행할 시, 자동으로 베이스 악보를 그려주어 본 실험에 대한 결과를 구체적으로 시각화 할 수 있는 프로그램을 개발할 수 있을 것으로 예상한다.

더 나아가 매번 연주자에 따라 연주가 자유롭게 달라지는 재즈 음악의 경우, 화려한 베이스 솔로 부분에서 본 논문에서 진행한 방법을 활용한다면, 구하기 힘든 재즈 악보를 대신할 수 있다고 예상한다.

이를 통해 사운드소스의 분석에 따라 추출된 베이스라인의 시각화를 통해 곡의 전반적인 분위기를 결정할 수 있도록 하는 사운드 디자인적 관점에서의 연관성을 이끌어 낼 수 있었으며, 또한 이는 음악 제작의 접근성을 높여주는데 기여한다. 따라서 본 연구는 곡의 구성을 디자인하고 사운드를 디자인 하는 과정에 큰 역할을 할 수 있을 것으로 기대한다.

참고문헌

1. Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-u-net: A multi-scale neural network for end-to-end audio source separation." arXiv preprint arXiv:1806.03185 (2018).
2. Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks.
3. https://github.com/
4. http://pages.mtu.edu/~suits/notefreqs.html
5. https://sigsep.github.io/datasets/musdb.html
6. https://www.fabfilter.com/
7. https://www.teachmeaudio.com/